Brasileiras em Processamento de Linguagem Natural

PROCESSAMENTO DE LINGUAGEM NATURAL

Conceitos, Técnicas e Aplicações em Português

Organizado por:
Helena de Medeiros Caseli
Maria das Graças Volpe Nunes

ISBN: 978-65-02-00178-3

4ª Edição | 2026

Este livro é o produto de um sonho coletivo do grupo

Brasileiras em Processamento de Linguagem Natural (BPLN)

© 2026 by Helena de Medeiros Caseli, Maria das Graças Volpe Nunes, organizadoras.
Editoração eletrônica: Helena de Medeiros Caseli

Licença CC BY-NC-ND 3.0 BR

Ficha Catalográfica: Marina P. Freitas CRB-08/ 6069

Dados internacionais de Catalogação-na-Publicação (CIP)

Processamento de Linguagem Natural: Conceitos, Técnicas e Aplicações em Português / organização de Helena de Medeiros Caseli e Maria das Graças Volpe Nunes. — 4. ed. - São Carlos : BPLN, 2026. Internet: brasileiraspln.com/livro-pln-4ed/
1314 p.

Em 3 volumes.

ISBN: 978-65-02-00178-3

1. Processamento de Linguagem Natural 2. Inteligência Artificial. 3. I. Título.

Processamento de Linguagem Natural

Conceitos, Técnicas e Aplicações em Português

Organizado por:
Helena de Medeiros Caseli
Maria das Graças Volpe Nunes

4ª Edição
Abril 2026

https://brasileiraspln.ufscar.br/livro-pln-4ed/

Prefácio da 4a. edição

O Processamento de Linguagem Natural (PLN) surgiu praticamente ao mesmo tempo que os computadores, por volta da década de 1940, já que a tradução automática entre línguas foi um dos primeiros problemas submetidos aos primeiros computadores[1]. No Brasil, as pesquisas em PLN começaram timidamente ainda na década de 1970, entre acadêmicos interessados em Inteligência Artificial (IA). Em 1984, foi realizada a primeira edição do Simpósio Brasileiro de Inteligência Artificial[2] (SBIA), em Porto Alegre, RS, e boa parte dos trabalhos apresentados nesse evento eram da área de PLN. Naquela época, os sistemas propostos para português eram muito simples, sendo apenas estudos de caso bem elementares, como os que visavam encontrar respostas para perguntas do tipo Qual?, O quê?, Onde? e Quando?. Demorou mais uma década até que houvesse uma massa crítica de cientistas brasileiros dedicados ao processamento computacional do português.

No início da década de 1990, destacavam-se pesquisadores de PLN em diversos centros acadêmicos, como cientistas da computação e linguistas da UFRGS, PUCRS e da Unisinos, no RS; linguistas da PUC-Rio, no RJ; cientistas da computação na Unicamp, em SP, e na UFPE, em PE. Também nessa época, em 1993, nascia o NILC (Núcleo Interinstitucional de Linguística Computacional), no ICMC-USP/São Carlos[3], um grupo pioneiro em PLN, que teve como desafio inicial produzir o primeiro revisor ortográfico e gramatical comercial para o português do Brasil. Para levar a cabo essa iniciativa, foi necessário compor uma equipe de informatas e linguistas que se tornaram agentes multiplicadores do PLN no Brasil.

Em 1993 foi realizado o primeiro evento exclusivo de PLN dedicado ao português ibérico e brasileiro, em Lisboa. Tal evento se tornou o *International Conference on the Computational Processing of Portuguese* (PROPOR)[4], realizado alternativamente no Brasil e em Portugal, com 17 edições realizadas até essa data (2026). No âmbito nacional, em 2003 foi criado o Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana (STIL)[5], que abriga diferentes eventos satélites tratando de temas específicos para o processamento da língua, em particular o Português, com 15 edições realizadas até essa data (início de 2026). Em 2007, foi criada a Comissão Especial de PLN (CEPLN)[6] da Sociedade Brasileira de Computação. A comunidade cresceu, inicialmente graças à formação acadêmica de profissionais e pesquisadores, posteriormente também em consequência da maior demanda de empresas tecnológicas e de outras naturezas, atraindo profissionais com diferentes formações. Em 2020, foi criado o grupo Brasileiras em PLN (BPLN)[7], responsável pela idealização e edição deste livro.

[1] Os primeiros computadores também foram massivamente usados para realizar cálculos numéricos complexos, mas a tradução automática foi a primeira aplicação simbólica da computação.

[2] O SBIA é hoje o BRACIS. Para saber mais, consulte http://comissoes.sbc.org.br/ce-ia/pg/historico/?sec=SBIA-1984.

[3] http://www.nilc.icmc.usp.br/

[4] https://sites.google.com/view/ce-pln/eventos/propor

[5] https://sites.google.com/view/ce-pln/eventos/stil

[6] http://comissoes.sbc.org.br/ce-pln/

[7] https://brasileiraspln.com/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

https://brasileiraspln.ufscar.br/livro-pln-4ed/

Já há algum tempo as aplicações de PLN saíram dos laboratórios acadêmicos e entraram definitivamente no nosso cotidiano. O desenvolvimento de diversas áreas da computação de forma integrada (hardware, software, tratamento de grandes volumes de dados, aprendizado de máquina e aprendizado profundo) impulsionou o desenvolvimento do processamento da língua para uma nova esfera. Em novembro de 2022, o mundo conheceu o ChatGPT, um *chatbot* capaz de responder a qualquer pergunta ou solicitação (escrevendo códigos de programas, p. ex.), em língua natural, incluindo o português. Além de surpreender pelo seu desempenho linguístico, ele acendeu um sinal de alerta para a comunidade de IA e PLN, bem como para vários setores da sociedade. Várias indagações surgiram: Será que a IA está prestes a atingir a meta de criar sistemas que possam vir a substituir o homem em todas as tarefas? Quais os riscos que as diferentes sociedades correm se o mercado de trabalho for dominado por sistemas inteligentes? E se apenas poucas empresas dominarem o mercado de sistemas inteligentes? E o que pode acontecer com países não desenvolvidos? A conscientização desse cenário fez crescer o interesse pelas atividades da comunidade que desenvolve sistemas de IA, incluindo a de PLN, já que o ChatGPT é uma aplicação típica dessa área.

Durante todas essas décadas, a comunidade brasileira de PLN dedicou-se a construir recursos linguístico-computacionais e também aplicações variadas para o português. No início, eram interfaces simples que entendiam perguntas feitas em linguagem natural. Seguiram-se ferramentas de auxílio à escrita, tradutores, sumarizadores, classificadores de documentos, reconhecedores e sintetizadores de fala, entre outros. Atualmente, muitas aplicações envolvem os conteúdos das redes sociais, sendo que a língua tem sido processada também com o objetivo de conhecer a pessoa que faz uso dela, seja para lhe oferecer produtos e serviços, seja para analisar tendências e personalidades. Entram em cena os reconhecedores de opiniões e sentimentos, e de notícias falsas. Essa nova tendência de descobrir o que está implícito ou pode ser inferido da linguagem trouxe problemas muito desafiadores para a área. Vários domínios de conhecimento e atuação estão fortemente presentes nessa revolução tecnológica: educação, saúde, direito, bancos e financeiras, comunicação, comércio, segurança, governança, entre outros. Tanto por serem afetados por ela como por ajudarem a construir novas soluções.

Se antes os especialistas em PLN tinham a tarefa de explicitar e codificar conhecimento linguístico, hoje seu maior desafio é preparar dados/exemplos linguísticos para servirem de entrada aos algoritmos no treinamento de modelos computacionais capazes de transformar dados em conhecimento ou ações. E isso não é pouco. Para enriquecer os dados brutos de modo que o conhecimento adquirido seja o mais completo e correto possível, ou que a ação gerada seja a mais adequada, é preciso saber escolhê-los, prepará-los e eventualmente anotá-los com informações de várias naturezas: morfológicas, sintáticas, semânticas, extralinguísticas etc. É a essa tarefa que muitos especialistas em PLN de todo o mundo têm se dedicado ultimamente. Outro papel importante está nas avaliações de qualidade dos recursos e produtos construídos, o entendimento do problema, a identificação das limitações e a proposta de melhorias. É crucial que as bases do desenvolvimento dessas tecnologias sejam fortes, bem informadas, consistentes e éticas. A linguagem nos define, e ela faz agora parte dos nossos artefatos, a complexidade deste cenário ainda não é bem compreendida. Com este livro, agora na quarta edição, composto por três volumes, com dez novos capítulos, os 117 autores e autoras (39 a mais do que na terceira edição) pretendem contribuir para que o desenvolvimento da tecnologia que envolve a linguagem humana seja bem fundamentado, consciente e ético.

https://brasileiraspln.ufscar.br/livro-pln-4ed/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Sobre este livro

Um dos grandes desafios de escrever um livro sobre uma área tecnológica, como o PLN, é fazê-lo de tal forma que ele não fique obsoleto pouco tempo depois de seu lançamento. E esse risco é, de fato, muito grande. Cada vez mais as tecnologias não apenas avançam rapidamente como podem ser totalmente substituídas por outras, tornando-se descartáveis. Ainda assim, há sempre um núcleo de conhecimento que se mantém relevante ao longo do tempo e que é indispensável para quem quer conhecer ou atuar nessa área. Por outro lado, por ser uma área tecnológica, é indispensável que o livro também apresente o que há de mais atual.

Este livro aborda conhecimentos, técnicas e aplicações de PLN em português que são relevantes atualmente ou historicamente. Afinal, para chegar até aqui, um longo caminho de construção de conhecimento foi percorrido – ainda que possa parecer que cada nova tecnologia seja sempre uma nova invenção.

Para contemplar história, estado da arte e não obsolescência, optou-se por criar um livro na forma digital, de acesso gratuito e dinâmico, ou seja, temos atualizado seu conteúdo a cada nova edição. Mais do que apenas atualizá-lo com novos conhecimentos, o livro tem sido ampliado com novos capítulos a cada nova edição. As possibilidades de expansão são inúmeras, já que há uma demanda crescente por novas aplicações que envolvem PLN.

Na primeira edição, procuramos introduzir os principais conceitos e tarefas da área, algumas das aplicações mais relevantes hoje em dia, particularmente em português do Brasil, e algumas das principais técnicas usadas para o desenvolvimento dessas tarefas e aplicações. Os temas de todos os capítulos tinham potencial para serem aprofundados em edições futuras, o que de fato aconteceu.

Na segunda edição, revisada e ampliada, o conteúdo foi estendido com seis novos capítulos organizados em doze partes. Foram acrescentadas duas novas partes: Interação e Avaliação, dois temas importantíssimos para o PLN.

A terceira edição foi estendida com seis novos capítulos e também trouxe material suplementar em alguns deles, com vídeos dos autores e slides de apresentação do conteúdo do capítulo.

Nesta quarta edição, com o crescimento constante do livro, decidimos dividir os 46 capítulos em 3 volumes. O Volume 1 é dedicado a 21 capítulos introdutórios e conceituais da área de PLN: Introdução, Fala, Palavras, Estrutura, Significado, Discurso, Geração e Interação, Dados, Avaliação e Desafios e Perspectivas. O Volume 2 é composto por 7 capítulos totalmente dedicados aos modelos de linguagem (*Language Models*): conceituação, técnicas, treinamento, avaliação, aplicação. Já o Volume 3 concentra 18 capítulos que descrevem as aplicações variadas na área de PLN. Devido a essa nova organização, vários capítulos de edições anteriores receberam nova numeração.

Esperamos que essa nova apresentação propicie aos leitores de diferentes perfis uma navegação mais confortável e eficaz. Em cada volume, há uma seção (Sobre o Volume) trazendo informações específicas sobre o mesmo.

Esta edição foi estendida com dez novos capítulos, sendo um no Volume 1, cinco no Volume 2 e quatro no Volume 3. Outros nove capítulos foram atualizados, sendo dois

https://brasileiraspln.ufscar.br/livro-pln-4ed/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

deles – capítulo Sumarização Automática e capítulo Complexidade Textual e suas Tarefas Relacionadas – significativamente estendidos com novas seções, com novos coautores. Finalmente, a dez capítulos foram anexados vídeos sobre seus temas.

Ressaltamos que sempre é possível navegar de um capítulo a outro, de um volume a outro, e que não disponibilizamos versões impressas, mas que o arquivo pdf de cada capítulo está acessível logo abaixo de seu título.

Este livro deve continuar crescendo com novas contribuições, na forma de novos capítulos, novos recursos, ou novas versões dos capítulos já existentes. Por isso, atente-se sempre para a data de publicação dos conteúdos. A fim de abranger temas variados nessa área, contamos com um número cada vez maior de autores colaboradores. A **quarta edição** conta com 118 autores, sendo eles professores universitários das principais universidades brasileiras e também do exterior, pesquisadores brasileiros que atuam no exterior, alunos de pós-graduação em Computação e em Linguística, profissionais de PLN que atuam em empresas. A variedade de autoria e a forma colaborativa de construção de conteúdos, sob diferentes visões (computacional e linguística), dão a este livro uma característica única, uma personalidade. E essa personalidade, que tanto nos orgulha, faz deste livro uma história do PLN escrita a muitas mãos.

Com este livro, as autoras e os autores pretendem oferecer um corpo de conhecimento suficientemente abrangente sobre o PLN em português do Brasil – ainda que não esgote todos os temas da área – que atenda expectativas de leitores com diferentes perfis. O livro pode ser usado como material didático para disciplinas de cursos de Computação ou Linguística, ou áreas afins, servindo como fonte de conhecimento teórico e prático para quem está aprendendo essa ciência. Pode servir de referência a profissionais que fazem uso do PLN para criar soluções tecnológicas, servindo como fonte de técnicas e métodos para desenvolvimento de sistemas tecnológicos. Também acadêmicos e curiosos, de um modo geral, nele encontrarão uma visão geral da área de PLN, e como ela pode interagir com outras áreas de conhecimento. Em Como ler este livro, apresentamos em detalhes os volumes desta edição.

Agradecemos o apoio da NAACL RAF 2021, uma iniciativa da NAACL[8] (*North American Chapter of the Association for Computational Linguistics*) para apoiar projetos da comunidade de PLN na América Latina.

[8]https://naacl.org/

https://brasileiraspln.ufscar.br/livro-pln-4ed/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Como citar

Caseli, H.M.; Nunes, M.G.V. (org.) **Processamento de Linguagem Natural: Conceitos, Técnicas e Aplicações em Português**. 4 ed. BPLN, 2026. ISBN 978-65-02-00178-3. Disponível em: https://brasileiraspln.ufscar.br/livro-pln-4ed.

Note 1: Bibref

```
@Book{BPLN_livro_4ed:2026,
   title     = {Processamento de Linguagem Natural: Conceitos,
                Técnicas e Aplicações em Português},
   editor    = {Caseli, H. M. and Nunes, M. G. V.},
   year      = {2026},
   publisher = {BPLN},
   edition   = {4},
   isbn      = {978-65-02-00178-3},
   url       = {https://brasileiraspln.ufscar.br/livro-pln-4ed}
}
```

https://brasileiraspln.ufscar.br/livro-pln-4ed/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Como ler este livro

O objetivo dos autores deste livro (117 nesta edição) é oferecer aos leitores conhecimento sobre PLN, ao mesmo tempo, introdutório e abrangente, de modo a atender às expectativas de leitores com diferentes perfis. O livro pode ser usado como material didático para disciplinas de IA e PLN de cursos de computação ou linguística, ou áreas afins, e também pode servir de referência a profissionais que fazem uso do PLN para criar soluções tecnológicas.

A fim de oferecer conhecimento sobre temas variados nessa área, foi importante contar com um grande número de especialistas brasileiros[9], a quem as organizadoras agradecem profundamente.

Esta quarta edição propõe uma nova organização do conteúdo do livro. Agora ele está dividido em três volumes, conforme indicado abaixo. Na seção Sobre o Volume, de cada volume, o leitor encontrará mais detalhes sobre seu conteúdo.

O Volume 1 abriga os capítulos que introduzem a área, seus principais conceitos, um pouco de história do PLN brasileiro, conceitos da Linguística, as divisões usuais das tarefas do PLN (sintaxe, semântica, geração), o processamento da fala, o tema da avaliação de sistemas de PLN,o papel dos *corpora*, entre outros temas gerais. Esses capítulos podem ser lidos de forma sequencial por leitores iniciantes, já que estão em ordem crescente de complexidade; ou, dependendo do perfil e do interesse do leitor, podem ser lidos em qualquer ordem, uma vez que são independentes entre si.

O Volume 2 traz capítulos exclusivamente sobre os modelos de linguagem (*Language Models*), dada sua relevância atual na área de IA e PLN. Os capítulos introduzem os conceitos comuns à área de Inteligência Artificial, apresentam as técnicas atuais, discutem a preparação de dados, o treinamento na prática, técnicas de mitigação de problemas, a avaliação dos modelos, bem como uma discussão aprofundada sobre os problemas no uso desses modelos; tudo isso com foco na língua portuguesa. O leitor interessado nesta área pode explorá-la integralmente e também complementá-la com capítulos relacionados em outros volumes, como aqueles do Volume 1 sobre Geração, Avaliação, Diálogo, *Dataset* e *Corpus*, além das diversas aplicações do Volume 3 que utilizam esses modelos.

O Volume 3 é dedicado às aplicações em geral. São trabalhos de PLN aplicados a tarefas recorrentes no PLN (classificação textual, sumarização, tradução, entre outras) e domínios do conhecimento (Direito e Saúde, entre outras). Trata-se de um volume que deve receber muitas contribuições ao longo do tempo.

Complementam esse livro as informações Sobre as(os) autoras(es).

https://brasileiraspln.ufscar.br/livro-pln-4ed/

[9] A lista completa das autoras e autores deste livro encontra-se em: **Sobre as/os autoras/es**

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Processamento de Linguagem Natural

Conceitos, Técnicas e Aplicações em Português

Organizado por:
Helena de Medeiros Caseli
Maria das Graças Volpe Nunes

4ª Edição
Volume 1: Conceitos e Técnicas
Abril 2026

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

Conteúdo

Volume 1: Conceitos e Técnicas 1
Sobre o Volume 1 . . . 1
Como citar . . . 2

I Introdução 4

1 O que é PLN? 5
1.1 Introdução . . . 5
1.2 A língua . . . 8
1.3 Um breve histórico do PLN . . . 9
1.4 Os Paradigmas de PLN . . . 12
1.5 Vale a pena relembrar . . . 13

II Fala 15

2 Texto ou fala? 16
2.1 Histórico e panorama da área . . . 16
2.2 Aspectos teóricos fundamentais . . . 18
2.2.1 A estrutura da língua falada . . . 19
2.2.2 Reconhecimento de fala . . . 28
2.2.3 Síntese de fala . . . 39
2.2.4 Considerações finais . . . 41

3 Recursos para o processamento de fala 43
3.1 Introdução . . . 43
3.2 Recursos para síntese de fala . . . 46
3.3 Recursos para segmentação prosódica . . . 49
3.4 Recursos para reconhecimento automático de fala . . . 52
3.5 Recursos para reconhecimento de emoções . . . 55
3.6 Recursos para predição de pontuação no cenário de ASR . . . 58
3.6.1 Ilustrando o uso do ASR Whisper na predição da pontuação em português . . . 58
3.6.2 Descrição do Corpus MuPe e seu *dataset* de teste . . . 60
3.6.3 *Corpora* usados nos trabalhos de predição da pontuação para ASRs . 61
3.7 Considerações finais . . . 64
Agradecimentos . . . 64

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

III Palavras **65**

4 Sequência de caracteres e palavras **66**
4.1 Conceitos básicos da morfologia . . . 68
4.1.1 Morfema . . . 68
4.1.2 *Token* e *Type* . . . 71
4.1.3 Lexema, Lexia e Lema . . . 71
4.1.4 Léxico e Gramática . . . 72
4.1.5 Léxico comum e Léxico especializado . . . 73
4.1.6 Palavras funcionais e palavras lexicais . . . 74
4.1.7 Processos de formação das palavras . . . 74
4.1.8 Morfologia e morfossintaxe . . . 75
4.2 O processamento morfológico em PLN . . . 75
4.2.1 Sentenciação . . . 76
4.2.2 Tokenização . . . 77
4.2.3 Tokenização em Subpalavras . . . 79
4.2.4 Normalização . . . 80
4.2.5 PoS *tagging* . . . 81
4.2.6 Anotação de atributos morfológicos . . . 83
4.3 Ferramentas e recursos para o processamento morfológico . . . 84
4.3.1 Ferramentas para o português . . . 84
4.3.2 Recursos para o português . . . 87
4.3.3 PortiLexicon-UD: um recurso para processamento léxico em português 89
4.4 Considerações finais . . . 95

5 Expressões multipalavras **97**
5.1 Introdução . . . 97
5.2 Hora de dar nome aos bois . . . 98
5.2.1 As pedras fundamentais . . . 98
5.2.2 Definições de MWE: uma pedra no nosso sapato . . . 99
5.2.3 Separando o joio do trigo . . . 101
5.2.4 MWEs para dar e vender . . . 102
5.2.5 Um osso duro de roer . . . 104
5.3 Tarefas: botando a mão na massa . . . 105
5.4 Recursos: uma joia rara . . . 107
5.4.1 Léxicos de MWEs . . . 107
5.4.2 *Corpora* de MWEs . . . 110
5.4.3 *Corpora* de MWEs em português . . . 111
5.5 Avaliação: colocando os pingos nos 'i's . . . 112
5.6 Até onde chegamos e para onde vamos . . . 114
Criação e manutenção de recursos . . . 114
Diálogo entre projetos e equipes . . . 115
Diversidade linguística . . . 115
Descrição aprimorada e ampliada de MWEs . . . 116
Processamento cognitivo de MWEs . . . 116
5.7 Ao infinito e além . . . 116
Agradecimentos . . . 117

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

IV Estrutura **118**

6 A ordem e a função das palavras em uma sentença **119**
6.1 Introdução . . . 119
6.2 Reflexões Iniciais . . . 121
6.3 Noções Básicas de Sintaxe . . . 122
6.4 Tipos de representação . . . 124
6.4.1 Colchetes . . . 124
6.4.2 Árvores . . . 125
6.4.3 Setas . . . 126
6.4.4 Parênteses . . . 126
6.4.5 Indentação . . . 127
6.5 Sintaxe de constituência . . . 127
6.5.1 Possibilidades de interpretação e ambiguidades sintáticas . . . 129
6.6 Sintaxe de dependência . . . 131
6.6.1 Núcleo e dependente . . . 132
6.6.2 A representação da sintaxe de dependência . . . 133
6.6.3 Projetos de anotação multilingue: Universal Dependencies . . . 136
6.7 Qual é melhor: constituência ou dependência? . . . 140
6.8 Fronteiras da sintaxe . . . 141
6.8.1 Sintaxe e Morfologia . . . 141
6.8.2 Sintaxe e Semântica . . . 142
6.8.3 Sintaxe e Pragmática . . . 143
6.8.4 Sintaxe e Discurso . . . 143
6.9 Considerações finais . . . 144

7 Ferramentas e recursos para o processamento sintático **145**
7.1 Introdução . . . 145
7.2 Tipos de *parsing* . . . 145
7.3 Recursos e ferramentas para o português . . . 147
7.3.1 *Corpora* . . . 148
7.3.2 *Parsers* . . . 149
7.4 Visualização, anotação e edição de *treebanks* . . . 159
7.4.1 Árvores de constituência . . . 159
7.4.2 Árvores de dependência . . . 160
7.4.3 Anotação de *corpus* em múltiplos níveis . . . 161
7.5 Considerações finais . . . 163

V Significado **164**

8 E o significado? **165**

9 Semântica com técnicas simbólicas **171**
9.1 Bases de Conhecimento Semântico . . . 173
9.1.1 *Wordnets* . . . 174
9.1.2 FrameNet . . . 178
9.1.3 ConceptNet . . . 183
9.2 Considerações finais . . . 189

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

10 Representação vetorial e semântica distribucional **191**
10.1 Representação Vetorial . . . 193
10.1.1 Matriz termo-documento . . . 195
10.1.2 Matriz termo-contexto . . . 196
10.1.3 Calculando a similaridade entre vetores . . . 197
10.2 Vetores esparsos . . . 198
10.2.1 Atribuindo pesos aos termos da matriz termo-documento com TF-IDF 199
10.2.2 Atribuindo pesos aos termos da matriz termo-contexto com PMI . . . 200
10.2.3 Reduzindo a dimensionalidade com LSA . . . 202
10.3 Vetores densos estáticos . . . 204
10.3.1 Word2Vec . . . 206
10.3.2 Fasttext . . . 209
10.3.3 Glove . . . 211
10.4 Considerações finais . . . 213

VI Discurso **214**

11 Modelos discursivos **215**
11.1 Introdução . . . 215
11.2 Modelos de relações discursivas . . . 217
11.2.1 O modelo RST (Mann e Thompson, 1988) . . . 220
11.2.2 CST: Cross-document Structure Theory, 2000 . . . 222
11.3 Recursos e aplicações para o português brasileiro . . . 225
11.4 Considerações finais . . . 226

12 Resolução de correferência **228**
12.1 Introdução . . . 228
12.2 Resolução de Correferência . . . 228
12.2.1 Referentes . . . 229
12.2.2 Relações Semânticas Referenciais . . . 231
12.2.3 Correferência, Anáfora e Catáfora . . . 232
12.2.4 Referências Endofóricas e Exofóricas . . . 234
12.2.5 Correferência, coerência e coesão . . . 234
12.3 Abordagens Computacionais para Resolução de Correferência . . . 234
12.3.1 Modelos Baseados em Regras . . . 235
12.3.2 Modelos Baseados em Aprendizado de Máquina . . . 235
12.3.3 Entity-Mention . . . 238
12.3.4 Mention-Ranking . . . 239
12.3.5 Antecedent-Tree . . . 240
12.3.6 Modelos Voltados à Língua Portuguesa . . . 241
12.4 Avaliação da Tarefa de Resolução de Correferência . . . 242
12.4.1 Métricas de Avaliação . . . 243
12.5 Aplicações . . . 246

VII Geração e Interação **248**

13 Geração de linguagem natural **249**

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

13.1 Introdução . . . 249
13.1.1 Geração de Linguagem Natural . . . 251
13.1.2 Seleção de Conteúdo e Realização Textual . . . 252
13.1.3 Avaliação de sistemas de GLN . . . 252
13.1.4 Arquiteturas de GLN . . . 253
13.2 Arquitetura Baseada em Templates . . . 253
13.2.1 Aplicações . . . 254
13.2.2 Exemplo – Sistema Rosie . . . 254
13.2.3 Limitações . . . 255
13.3 Arquitetura *pipeline* . . . 256
13.3.1 Aplicações . . . 258
13.3.2 Exemplo . . . 258
13.3.3 Limitações . . . 263
13.4 Arquitetura Ponta-a-Ponta . . . 263
13.4.1 Aplicações . . . 265
13.4.2 Exemplo – sistema CoronaReporter . . . 266
13.4.3 Limitações . . . 267
13.5 Arquiteturas Pré-treinadas . . . 268
13.5.1 Aplicações . . . 269
13.5.2 Exemplo . . . 269
13.5.3 Limitações . . . 270
13.6 Conclusão . . . 271

14 Perguntas e Respostas **273**
14.1 Introdução . . . 273
14.2 Classificação de Sistemas de Perguntas e Respostas . . . 275
14.2.1 Tipo de pergunta . . . 275
14.2.2 Tipo de Fonte de Conhecimento . . . 277
14.2.3 Domínio de conhecimento . . . 281
14.3 Abordagens . . . 282
14.3.1 Abordagem modular com documentos . . . 283
14.3.2 Abordagem modular com grafo de conhecimento . . . 287
14.3.3 Abordagem *end-to-end* . . . 289
14.4 Métodos de avaliação . . . 291
14.5 Outras tarefas relacionadas a PR . . . 293
14.5.1 Comunidades de Perguntas e Respostas - *Community* QA . . . 293
14.5.2 PR de múltiplas etapas *Multihopping* QA . . . 294
14.5.3 PR com dados visuais . . . 294
14.6 Sistemas de PR para o Português . . . 294
14.7 Considerações finais . . . 295

15 Diálogo e Interatividade **297**
15.1 Introdução: Dialogando . . . 297
15.2 Conceitos: O que forma um diálogo? . . . 299
15.3 Fenômenos: O que ocorre em uma conversa? . . . 302
15.3.1 Incrementalidade . . . 302
15.3.2 Troca de Turnos e Estrutura Sequencial . . . 304
15.3.3 Disfluências e Reparos . . . 305

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

15.3.4 Construção de Base Comum e de Referências . . . 308
15.3.5 Sinais Multimodais . . . 309
15.3.6 Análises de diálogo em português . . . 310
15.4 Modelos e Tarefas: Implementando um processo complexo . . . 310
15.4.1 Modelos . . . 311
15.4.2 Tarefas . . . 314
15.5 Avaliação: Verificando a qualidade de um sistema . . . 315
Considerações éticas . . . 318
15.6 Recursos: Trabalhando com dados e sistemas interativos . . . 318
15.6.1 Atuando de Forma Responsável . . . 318
15.6.2 *Corpora* . . . 319
15.6.3 Ferramentas . . . 322
15.6.4 Para Saber Mais . . . 324
15.7 Conclusão: Mantendo o valor humano em foco . . . 324
15.8 Exercícios . . . 326
Agradecimentos . . . 328

VIII Dados 329

16 Conjunto de dados, *dataset* e *corpus* 330
16.1 Introdução . . . 330
16.1.1 *Dataset* ou *corpus* anotado? . . . 331
16.1.2 Anotação linguística . . . 333
16.2 *Datasets* pra quê? . . . 335
16.2.1 Sobre a importância dos dados . . . 337
16.3 Características de um bom *dataset* linguístico . . . 338
16.4 Por onde começar? . . . 339
16.4.1 Definição do problema ou tarefa . . . 340
16.4.2 Conjunto de etiquetas e instruções: o esquema de anotação . . . 340
16.4.3 Escolha do *corpus* . . . 342
16.4.4 Codificação . . . 343
16.4.5 Anotação: sabedoria de especialistas ou sabedoria da multidão? . . . 345
16.4.6 Ferramentas de anotação . . . 347
16.4.7 Estratégias de anotação . . . 348
16.4.8 Formas de avaliação . . . 349
16.5 Procedimentos e estratégias de anotação e revisão . . . 349
16.5.1 Palavras, regras e padrões linguísticos na construção de um *corpus* padrão ouro . . . 349
16.5.2 Revisão de anotação . . . 352
16.5.3 Até onde precisamos rever? . . . 354
16.6 Como avaliar a qualidade do *dataset*? . . . 354
16.6.1 Concordância entre-anotadores . . . 356
16.6.2 Quando a discordância não é erro: divergências legítimas . . . 357
16.6.3 Avaliação intrínseca . . . 358
16.6.4 Avaliação extrínseca . . . 360
16.7 Em resumo... . . . 362
Agradecimentos . . . 363

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

17 Aprendizado Transdutivo em PLN **364**
17.1 Introdução . . . 364
17.2 E o que seria esse tal de aprendizado transdutivo? . . . 365
17.2.1 Transdução como forma de aprendizado . . . 367
17.3 Aplicações do Aprendizado Transdutivo em PLN . . . 369
17.3.1 Métodos baseados em propagação de rótulos . . . 369
17.3.2 Adaptação de modelos supervisionados . . . 372
17.3.3 Avaliação dos dados anotados . . . 377
17.4 Benefícios e Limitações . . . 378

IX Avaliação **379**

18 Avaliação de tecnologias de linguagem **380**
18.1 Introdução: Avaliando . . . 380
18.2 Contexto: Por onde começar? . . . 382
18.2.1 Um pouco de história . . . 382
18.2.2 Formalização de problemas e de modelos . . . 384
18.2.3 Caracterização de tarefas de PLN . . . 385
18.3 Paradigmas: Tipos de avaliação . . . 387
Avaliação Conjunta e *Leaderboards* . . . 390
18.4 Procedimentos: Como avaliar? . . . 392
18.4.1 Hipóteses e Experimentos . . . 392
18.4.2 Referências . . . 393
18.4.3 Mensurações . . . 394
18.4.4 Partição de dados . . . 395
18.4.5 Validação Cruzada . . . 395
18.4.6 Comparação de Performance . . . 397
18.4.7 Testes de Significância Estatística . . . 397
18.4.8 *Baselines* . . . 398
18.4.9 Ablação e Substituição . . . 399
18.4.10 Análise de Dados . . . 400
18.4.11 Análise de Erro . . . 400
18.4.12 Avaliação Humana . . . 401
18.5 Métricas: Medindo a performance . . . 402
18.5.1 Usando métricas de forma responsável . . . 402
18.5.2 Métricas comuns em PLN . . . 404
18.6 Uso Responsável e Boas Práticas . . . 405
18.6.1 Como reportar resultados? . . . 405
18.6.2 Como tornar os resultados reprodutíveis? . . . 408
18.6.3 Como testar e documentar um sistema? . . . 409
18.6.4 Como organizar experimentos? . . . 410
18.6.5 Atuando com ética . . . 410
18.7 Para Concluir . . . 410
18.8 Exercícios . . . 412
Agradecimentos . . . 414

19 Avaliação conjunta em português **415**

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

19.1 Apresentação . . . 415
19.2 Avaliação conjunta: o que é e para que serve . . . 416
19.2.1 Características e ingredientes de uma avaliação conjunta . . . 418
19.3 Por que participar de uma avaliação conjunta . . . 419
19.4 Como implementar uma avaliação conjunta . . . 420
19.4.1 Cartografia do problema e da tarefa . . . 420
19.4.2 Criação dos recursos . . . 421
19.4.3 Medidas de avaliação . . . 422
19.5 Como avaliar uma avaliação conjunta . . . 422
19.5.1 Críticas e limitações . . . 422
19.6 Morfolimpíadas: análise morfológica (2003-2004) . . . 424
19.7 Primeiro e Segundo HAREM: avaliação de entidades, relações e expressões temporais (2005 e 2007-2008) . . . 425
19.7.1 Reconhecimento de entidades mencionadas . . . 425
19.7.2 Identificação, classificação e normalização de expressões temporais . 426
19.7.3 ReRelEM - relações entre entidades . . . 427
19.8 CLEF: Recuperação de Informação e QA (2003-2009) . . . 428
19.8.1 CLEF clássico . . . 428
19.8.2 Resposta automática a perguntas (QA@CLEF) . . . 429
19.8.3 GeoCLEF . . . 429
19.8.4 GikiP: Informação geográfica (2008) . . . 430
19.8.5 GikiCLEF . . . 430
19.9 Págico: RI complexa sobre culturas de língua portuguesa (2012) . . . 430
19.10ASSIN e ASSIN 2: similaridade semântica e inferência (2016 e 2019) . . . 432
19.11IberLEF 2019: reconhecimento de entidades e de relações . . . 433
19.12IDPT: Identificação de ironia (2021) . . . 434
19.13ABSAPT: Avaliação de mineração de opiniões por aspecto (2022) . . . 434
19.14DIP: Desafio de Identificação de Personagens (2022-2023) . . . 435
19.15Resumo das avaliações conjuntas . . . 435
19.16Considerações finais e lições aprendidas . . . 436
Agradecimentos . . . 437

X Desafios e Perspectivas 438

20 Questões éticas em IA e PLN 439
20.1 Ética em IA . . . 439
20.2 Ética em PLN . . . 445
20.3 Modelos de língua como fonte de conhecimento? . . . 446

21 E agora, PLN? 448
21.1 Desafios e perspectivas para o PLN-Português . . . 448
21.2 Há limites para o PLN? . . . 451

Referências 453

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Apêndices **536**

A Alguns pressupostos para o processamento de fala **536**
A.1 Alguns pressupostos: estatística, probabilidade, teoria da informação 536
A.1.1 Probabilidade 537
A.1.2 Variáveis aleatórias 537
A.1.3 Média e Variância 538
A.1.4 Covariância e Correlação 538
A.1.5 Vetores aleatórios e distribuições multivariadas 539
A.1.6 Algumas distribuições úteis 539
A.1.7 Teoria da Estimação e Teste de Significância 540
A.1.8 Teoria da Informação, Entropia e Informação Mútua 541

B Diretrizes de avaliação **543**
B.1 Diretrizes de avaliação para um modelo de cinco estágios 543

C EAGLES **547**
C.1 EAGLES 547

Sobre as/os autoras/es **549**
Helena de Medeiros Caseli 549
Maria das Graças Volpe Nunes 549
Adriana Silvina Pagano 549
Aline Aver Vanin 549
Aline Villavicencio 550
Amanda Pontes Rassi 550
Ana Clara Souza Pagano 550
Arnaldo Candido Junior 550
Brielen Madureira 550
Camila de Araújo Azevedo 551
Carlos Ramisch 551
Cláudia Freitas 551
Daniela Barreiro Claro 551
Dante Augusto Barone 552
Diana Santos 552
Edresson Casanova 552
Eduardo Cortes 552
Elisa Terumi Rubel Schneider 552
Eloize Rossi Marques Seno 553
Evandro Fonseca 553
Flaviane Romani Fernandes Svartman 553
Heliana Mello 553
Ivandré Paraboni 554
Jackson Wilke da Cruz Souza 554
Laila Mota 554
Lucelene Lopes 554
Maria José Bocorny Finatto 554
Mariza Ferro 555
Marli Quadros Leite 555

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Nataly Leopoldina Patti da Silva 555
Norton Trevisan Roman 555
Paula Christina Figueira Cardoso 555
Priscila Osório Côrtes 556
Renata Ramisch 556
Renata Vieira 556
Ricardo Marcacini 556
Roana Rodrigues 556
Sandra Maria Aluísio 557
Solange Rezende 557
Tayane Soares 557
Thiago Castro Ferreira 557
Valéria de Paiva 558
Vinícius G. Santos 558
Vládia Pinheiro 558

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Volume 1: Conceitos e Técnicas

Sobre o Volume 1

Este volume destina-se a introduzir os principais conceitos, tarefas, técnicas e métodos do PLN. Ele deve fornecer uma boa visão geral da área, registrando marcos importantes desde seu início. O leitor iniciante se beneficiará de uma leitura sequencial dos capítulos (exceção feita à Parte 2, sobre Fala, que é bastante independente das demais).

A **Parte 1 (Introdução)** introduz os principais conceitos e traz um breve histórico do PLN, no Capítulo 1, e é leitura obrigatória para o leitor não familiarizado com PLN.

A **Parte 2 (Fala)**, ao contrário do restante do livro, que trata de processamento de texto, apresenta a área de processamento de fala: seus principais conceitos (Capítulo 2), e técnicas, recursos e aplicações (Capítulo 3). Além dos capítulos deste Volume, o leitor também encontrará no Capítulo Classificação de Áudio aplicada à Saúde do Volume 3 uma aplicação do processamento da fala para a classificação de áudios visando a detecção de problemas respiratórios. O leitor, ainda que interessado apenas na síntese ou reconhecimento de fala, se beneficiará dos demais capítulos deste livro para complementar conceitos comuns, como a anotação e o uso de *corpus*, o processamento da fala transcrita em texto, entre outros.

Na **Parte 3 (Palavras)**, a primeira fase do processamento textual é discutida. O Capítulo 4 dedica-se a desvendar a morfologia, que estuda as palavras isoladamente, suas partes (morfemas), seus processos de derivação e composição, bem como partes importantes do processamento automático como a tokenização e a atribuição das categorias das palavras (*part-of-speech tagging*). Se o leitor desejar aprofundar o estudo sobre a importância das palavras para a complexidade textual, sugere-se a leitura do Capítulo Complexidade Textual e suas Tarefas Relacionadas. Se ao leitor interessa o processamento de expressões idiomáticas, o Capítulo 5 explora amplamente o mundo das expressões multipalavras, que traz grandes desafios à medida que a combinação dos componentes dessas expressões está relacionada à cultura de uso da língua.

Crescendo em complexidade, a **Parte 4 (Estrutura)** considera a ordem das palavras numa sentença, buscando extrair seus papéis na organização sintática parcial ou total da sentença. Com tal conhecimento, o processamento da língua alcança um novo patamar, e a partir dele já é possível realizar várias tarefas de PLN, como *parsing* parcial ou total, e viabilizar várias aplicações. O Capítulo 6 fornece toda a conceitualização de sintaxe, os principais tipos de análise, suas diferenças, vantagens e desvantagens. O Capítulo 7 mostra as diferentes ferramentas computacionais para o processamento sintático, em especial, as que são dedicadas ao português. As Partes 3 e 4 são indispensáveis para o leitor, estudante ou profissional, que pretende atuar na área de PLN, pesquisando ou implementando sistemas.

A **Parte 5 (Significado)** promove um salto significativo para a complexidade do PLN: trata dos conceitos, modelos e técnicas relativos à apreensão do sentido implicado pela língua escrita. Isso pode ocorrer pelo uso de teorias e modelos simbólicos ou não simbólicos. O Capítulo 8 introduz toda a complexidade da semântica da língua. Os Capítulos 9 e

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

10 mostram as diferentes abordagens (simbólica e estatística, respectivamente) para o tratamento do sentido. A leitura desta parte é indispensável para quem quer conhecer de que forma o PLN busca apreender o significado das expressões linguísticas num texto.

Questões discursivas e retóricas implicadas pelo texto são tratadas na **Parte 6 (Discurso)** deste volume. No Capítulo 11, o leitor encontra os principais modelos discursivos para PLN; no Capítulo 12, um fenômeno muito frequente, e clássico no PLN, é tratado em detalhes: como resolver as correferências discursivas presentes em textos.

A **Parte 7 (Geração e Interação)** trata das diferentes arquiteturas de sistemas de geração de linguagem natural no Capítulo 13, e cobre, nos Capítulos 14 e 15, dois tipos de sistemas clássicos de PLN que se tornaram muito populares com o comércio eletrônico e, mais recentemente, com os agentes conversacionais: são os sistemas de perguntas e respostas, mais conhecidos na sua denominação em inglês – *Question Answering* – e os sistemas de diálogos.

A **Parte 8 (Dados)** trata da escolha e da preparação dos dados que alimentam os algoritmos e os métodos e critérios de avaliação dos sistemas criados. O Capítulo 16 aborda tudo o que está envolvido na construção e na anotação de *datasets* ou *corpus*, bem como seu papel no treinamento de modelos de aprendizado de máquina. Para o leitor interessado em saber como os dados são usados no treinamento dos grandes modelos de linguagem (*Large Language Models*, LLMs), sugere-se a leitura do Capítulo O papel dos dados no pré-treinamento de Grandes Modelos de Linguagem. E o Capítulo 17 discute o aprendizado transdutivo, uma alternativa ao aprendizado indutivo, como forma de incorporar grandes quantidades de dados não rotulados e, com isso, reduzir custos e aumentar a eficiência no treinamento de modelos.

A **Parte 9 (Avaliação)** inclui o Capítulo 18, que apresenta um panorama dos métodos de avaliação comumente usados para medir, analisar e comparar o desempenho de sistemas de PLN, e o Capítulo 19, que aborda o tema da avaliação conjunta (*shared tasks*) e oferece um amplo panorama das avaliações conjuntas promovidas para a língua portuguesa: leitura obrigatória para todos os desenvolvedores de sistemas de PLN, sejam pesquisadores ou não. No escopo dos LLMs, sugere-se a leitura do Capítulo Avaliação de Grandes Modelos de Linguagem, que complementa a discussão desse importante tópico, que é a avaliação de sistemas tecnológicos.

Antes de finalizar o volume, a **Parte 10 (Desafios e Perspectivas)** discute algumas questões éticas (Capítulo 20) que a IA, em geral, e o PLN, em particular, têm provocado, pela forma como novas tecnologias têm sido criadas e usadas recentemente. Para aprofundar ainda mais esta discussão, sugere-se a leitura do Capítulo Responsabilidade no desenvolvimento e uso de tecnologias de linguagem baseadas em IA que discute responsabilidade no desenvolvimento dos grandes modelos de linguagem. Finalmente, o último Capítulo (21) discorre sobre algumas perspectivas para o PLN do português.

Como citar

Caseli, H.M.; Nunes, M.G.V. (org.) **Processamento de Linguagem Natural: Conceitos, Técnicas e Aplicações em Português**. 4 ed. BPLN, 2026. v. 1. ISBN 978-65-02-00148-6. Disponível em: https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Note 1: Bibref

```
@Book{BPLN_livro_4ed_vol1:2026,
   title     = {Processamento de Linguagem Natural: Conceitos,
                Técnicas e Aplicações em Português},
   editor    = {Caseli, H. M. and Nunes, M. G. V.},
   year      = {2026},
   publisher = {BPLN},
   edition   = {4},
   volume    = {1},
   isbn      = {978-65-02-00148-6},
   url       = {https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1}
}
```

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Parte I

Introdução

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Capítulo 1

O que é PLN?

Helena de Medeiros Caseli
Maria das Graças Volpe Nunes
Adriana Pagano

Publicado em: 26/09/2023
Atualizado em: 16/04/2026

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

1.1 Introdução

O Processamento de Linguagem Natural (PLN) é um campo de pesquisa que tem como objetivo investigar e propor métodos e sistemas de processamento computacional da linguagem humana. O adjetivo "Natural", na sigla, se refere às línguas faladas pelos humanos, distinguindo-as das demais linguagens (matemáticas, visuais, gestuais, de programação etc.). No decorrer deste livro, os termos "língua", "linguagem humana" e "linguagem natural" serão usados indistintamente; já "linguagem" pode eventualmente se referir a qualquer tipo de linguagem. Na área da Ciência da Computação, PLN está ligado à área de Inteligência Artificial (IA) e também está intrinsecamente relacionada à Linguística Computacional.

Para deixar mais claro o que entendemos por PLN, vamos esclarecer o que se faz nessa área. De modo geral, em PLN buscam-se soluções para problemas computacionais, ou seja, tarefas, sistemas, aplicações ou programas, que requerem o tratamento computacional de uma língua (português, inglês etc.), seja escrita (texto) ou falada (fala). Línguas como as de sinais também têm sido alvo de estudos da área. Cada modo tem suas especificidades. No caso da fala, as características que a distinguem da língua escrita são relacionadas a questões da produção (síntese) e recepção (reconhecimento) do som. Recursos da fala, como a entonação, o volume, o sotaque, podem tanto dificultar o reconhecimento ou a síntese, como também facilitar o reconhecimento de sentimentos ou intenções do falante. Qualquer que seja o modo, fala, escrita, línguas orais e línguas de sinais compartilham a dificuldade maior em PLN: a apreensão do significado de uma expressão linguística. Isso vai ficar claro no decorrer deste livro.

O PLN se divide em duas grandes subáreas: **Interpretação (ou Compreensão) de Linguagem Natural** – NLU (do inglês, *Natural Language Understanding*), e **Geração de Linguagem Natural** – NLG (do inglês, *Natural Language Generation*)[1].

Situa-se em NLU tudo o que diz respeito ao processamento que visa à análise e à interpretação da língua. Por análise, entende-se a segmentação e classificação dos componentes linguísticos (e.g. palavras e suas classes morfológicas e gramaticais, seus traços semânticos

[1]Embora algumas siglas, como NLP (*Natural Language Processing*) e AI (*Artificial Intelligence*) tenham sido traduzidas e sejam amplamente utilizadas em português, as siglas NLU e NLG são utilizadas, em textos em português, em sua grafia inglesa.

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

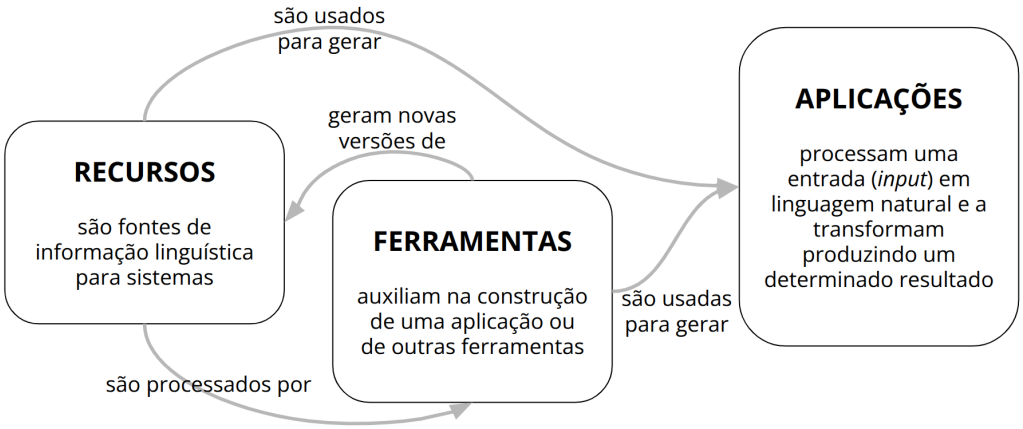
ou ontológicos etc.). Já interpretação se refere à tentativa de apreender significados construídos pelo ser humano. Numa interação com um *chatbot*, por exemplo, a interpretação ocorre quando o sistema processa um texto do usuário para descobrir o que ele – o sistema – deve fazer a seguir: se fornecer uma resposta ou executar uma ação. Logo ficará claro que respostas mais ou menos bem-sucedidas do sistema para o significado tencionado pelo humano podem ser suficientes para muitas aplicações, e que o completo alinhamento entre o significado tencionado pelo humano e aquele interpretado pela máquina não deve ser parte das nossas expectativas.

Em NLG, por outro lado, o objetivo é a geração de linguagem natural. Um exemplo de NLG é a geração de respostas ao usuário dos *chatbots*. Para o sistema, isso significa decidir o que responder e como apresentar essa resposta ao usuário. O ChatGPT[2] é um exemplo de sucesso: é capaz de gerar língua de forma tão ou mais fluente quanto muitos humanos.

É importante esclarecer, desde já, alguns conceitos amplamente usados no decorrer deste livro. Eles dizem respeito à classificação de alguns sistemas de PLN quanto ao seu uso. Esses conceitos são: aplicações, recursos e ferramentas.

Primeiramente, é relevante observar como esses conceitos se relacionam entre si. A Figura 1.1 esquematiza essa dinâmica.

Figura 1.1: Relacionamento entre conceitos



Como vemos na Figura 1.1, em PLN as ferramentas auxiliam na construção de uma aplicação, que pode ser um sistema computacional (desktop, web) ou um aplicativo. As aplicações fornecem um resultado ao usuário tendo uma entrada (*input*) ou saída (*output*) em linguagem natural. Aplicações fazem uso de ferramentas ou conjuntos de ferramentas, conhecidos como "*toolkits*". Também necessitam recursos, os quais fornecem informações linguísticas necessárias para que as aplicações consigam processar a língua da maneira adequada.

É importante notar que a denominação utilizada – aplicação, recurso ou ferramenta – é imprecisa e depende do uso. Por exemplo, um corretor ortográfico pode ser uma aplicação a ser usada de forma autônoma ou um passo intermediário para uma aplicação de correção de redações; um tradutor automático pode ser uma aplicação em si, com uma interface para colocar um texto de entrada e obter um texto de saída, mas também pode ser usado como ferramenta para traduzir um *corpus* de uma língua para outra, visando

[2] https://chat.openai.com/

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

a criação de recursos em línguas de comunidades tecnologicamente menos desenvolvidas; um sumarizador automático pode ser usado para criar resumos para um usuário qualquer, mas também pode ser usado por um buscador da web como passo intermediário para um sistema de recuperação de informação; um dicionário é um recurso, mas também pode ser usado como um aplicativo para consulta; um modelo de língua pode se transformar num *chatbot*, e assim por diante. Os conceitos são caracterizados e exemplificados no Quadro 1.1.

Quadro 1.1: Exemplos de aplicações, recursos e ferramentas

Conceito	**Caracterização**	**Exemplos**
Aplicações	processam uma entrada (*input*) em linguagem natural e a transformam produzindo um determinado resultado	- tradutor automático - corretor ortográfico ou gramatical - assistentes virtuais/*chatbots* - sumarizador automático - sistemas de recomendação em sites de e-commerce ou entretenimento - sistemas de auxílio à escrita - sistemas de classificação textual - sistemas de recuperação de informação - sistemas de detecção de *fake news*
Recursos	são fontes de informação linguística para sistemas	- léxicos (listas de palavras com informações associadas) da língua em geral ou de terminologia de domínio - dicionários monolíngues ou bilíngues - *corpus* (*datasets* linguísticos) anotados manual ou automaticamente (para referência, teste ou treinamento de algoritmos de aprendizado de máquina) - listas de frequências de palavras - (mais estruturados) taxonomias, ontologias, redes de sinônimos e antônimos - em formato matemático: modelos de língua estatísticos (probabilidades) ou neurais (pesos) para informar qual palavra deve ser a próxima num dado contexto
Ferramentas	auxiliam na construção de uma aplicação ou até de outras ferramentas	- segmentadores textuais em *tokens* (tokenizador), sentenças, parágrafos - *stemmers* (extratores de raiz de uma palavra (e.g. “corr” de “correr”) - lematizadores (e.g. “correr” de “corri”) - etiquetadores morfossintáticos (PoS *taggers*) para classe de palavras (verbo, substantivo, adjetivo, artigo, preposição, advérbio etc.) - etiquetadores semânticos, analisadores sintáticos parciais (*chunkers*) e completos (*parsers*) - concordanciadores - interfaces de anotação de *corpus* - componentes em kits de ferramentas (*toolkits*), como o NLTK[3]

Neste livro iremos aumentar gradativamente a complexidade do tratamento da língua no PLN, com foco no português brasileiro. Antes de iniciar esta trajetória, a Seção 1.2 apresenta nosso objeto de pesquisa, a língua. Em seguida, a Seção 1.4 introduz os principais paradigmas do PLN, que serão retomados em diversos momentos neste livro. Por fim, a Seção 1.5 destaca os principais pontos apresentados no capítulo.

[3] https://www.nltk.org

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

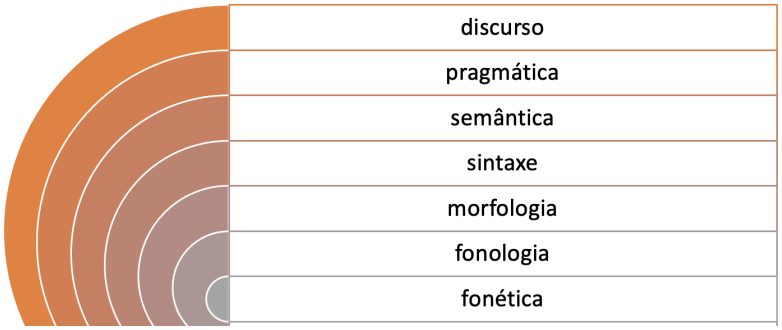
1.2 A língua

A capacidade de usarmos a linguagem para representar nossa realidade e nos comunicar é algo que distingue o ser humano dos outros seres vivos. Poder criar significados, expressar-se e ser compreendida é um dos grandes avanços no desenvolvimento de uma criança. Nos primeiros anos de vida, um bebê vai adquirindo a habilidade de se expressar em sua língua materna. Anos depois, normalmente a criança adquire a capacidade de utilizar símbolos para registrar aquilo que ela deseja por meio da língua escrita. A língua, como um sistema de construção de representações do mundo e comunicação, sobretudo no modo escrito, é o foco deste livro.

Ao longo do livro, nosso foco predominante será a língua escrita[4], ou seja, sequências de caracteres representados de forma grafológica, os quais constroem significados para nós humanos. Em PLN, chamamos a língua escrita de **texto**, para distingui-la da linguagem oral, que é chamada de **fala**. Portanto, apesar de a linguística reconhecer que existem textos escritos e textos falados, em PLN a palavra **texto** se refere principalmente ao texto escrito. Em relação à língua, neste livro, os exemplos estão em **português brasileiro**, embora muitas das técnicas descritas aqui possam ser aplicadas a outros idiomas.

A linguagem humana organiza-se em diferentes dimensões. A Figura 1.2 mostra uma representação das subáreas que estudam o sistema linguístico.

Figura 1.2: Representação das subáreas de estudo da linguagem



Na Figura 1.2, a língua é representada por meio de círculos concêntricos, sendo cada um deles objeto de estudo de uma subárea dos estudos linguísticos. No núcleo, os sons e sua organização são estudados pela **fonética** e pela **fonologia**. Envolvendo a estrutura sonora, temos o estudo de como os morfemas se organizam para formar palavras, que é objeto de estudo da **morfologia**. Envolvendo a morfologia, temos o estudo de como as palavras se organizam em estruturas para formar sintagmas e orações, objeto de estudo da **sintaxe**. No círculo envolvendo a sintaxe, temos a **semântica**, que estuda o significado de palavras e frases, enquanto a **pragmática** enfoca como as orações são utilizadas na interação para fins comunicativos específicos. Já **discurso** é uma denominação que abrange os estudos com foco no texto como um todo, podendo se referir à análise das relações entre frases ou partes de um texto, ou das etapas na estrutura de um texto.

Cada língua tem suas especificidades que determinam, por exemplo, desde como os caracteres podem ser combinados para compor uma palavra (uma sequência válida que tenha significado naquela língua) até regras que definem a estrutura (sintaxe) dessa língua. No decorrer deste livro, serão abordados os desafios do PLN em cada uma dessas subáreas.

[4]Com exceção de alguns capítulos que tratam de processamento da língua falada.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Contudo, é importante que fique claro que as estratégias computacionais usadas para o processamento da linguagem muitas vezes utilizam conhecimentos de várias subáreas ao mesmo tempo. Por exemplo, no processamento morfossintático realizado por um etiquetador (*tagger*), informações morfológicas e sintáticas são consideradas para se determinar a categoria gramatical (*part-of-speech*, PoS) de uma palavra.

1.3 Um breve histórico do PLN

A pesquisa em PLN teve início na primeira metada do século XX combinando ideias da linguística e da computação. Naquela época, havia duas visões conflitantes de como o processamento automático da língua deveria acontecer: o racionalismo e o empirismo. O racionalismo advoga que parte significativa do conhecimento humano não vem dos sentidos, mas é herdada geneticamente. Noam Chomsky é um dos advogados desta vertente com suas convicções sobre a linguagem inata e a gramática universal. O empirismo, por outro lado, acredita que a mente não é dotada de princípios e procedimentos pré-determinados, mas sim de capacidade para realizar operações gerais de associação, reconhecimento de padrões e generalizações. Para essa linha, não há conhecimento inato e o aprendizado decorre da experiência. Segundo Freitas (2022), "para o empirismo, somos bons em generalizar, mesmo com dados esparsos [...] e é isso que nos possibilita viver em um mundo repleto de incertezas". O racionalismo está associado ao paradigma simbólico enquanto o empirismo, ao estatístico e neural. Esses paradigmas serão apresentados na Seção 1.4. Foi o empirismo que passou a dominar as pesquisas em PLN neste século.

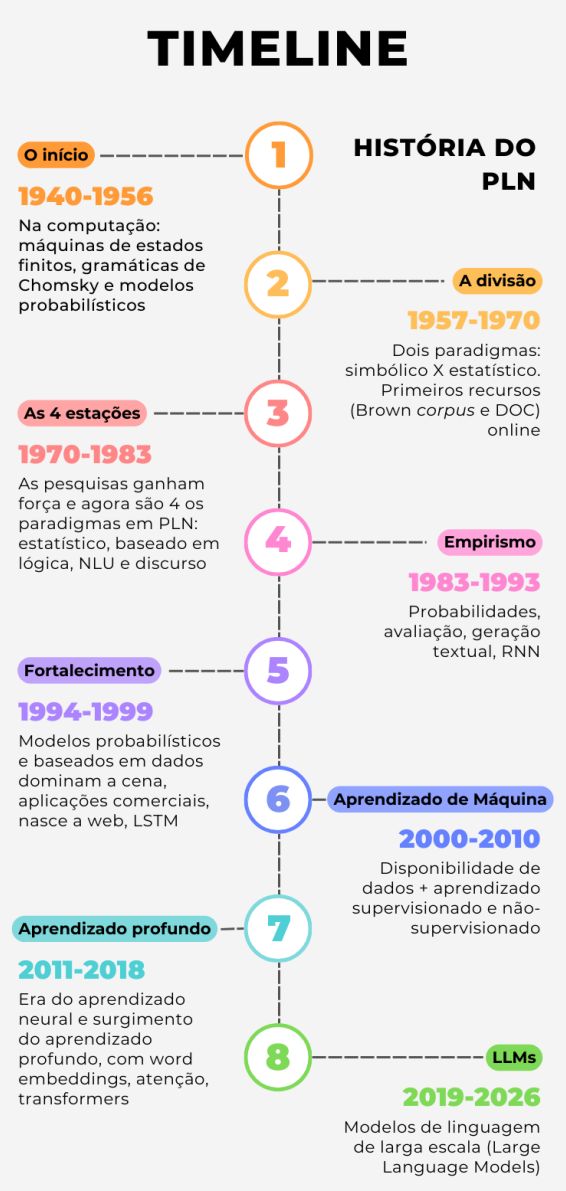
O empirismo – e consequentemente o aprendizado de máquina estatístico e neural – baseia-se no uso de *corpora* para estudar e formalizar fenômenos, verificar e validar hipóteses e evidências linguísticas. Um exemplo desta linha de pensamento é a famosa frase de Firth (1957): "You shall know a word by the company it keeps", que é a fundamentação por trás de modelos de representação textual que definem o significado de uma palavra com base no seu contexto.

A Figura 1.3 traz a linha do tempo do PLN com alguns marcos importantes. A história do PLN teve início por volta de 1940. As primeiras ideias para criação da área surgiram logo após a 2ª Guerra Mundial (Jurafsky; Martin, 2008). Na computação, esta época foi marcada por avanços com as máquinas de estados finitos, os modelos probabilísticos e as gramáticas de Chomsky.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Figura 1.3: Timeline do PLN



https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

No final de 1950 até por volta de 1970, o PLN se dividiu em dois paradigmas: o simbólico e o estatístico. Nesta época, o paradigma simbólico se baseava nas linguagens formais de Chomsky e nas estratégias de *parsing top-down* e *bottom-up*, em programação dinâmica e nos trabalhos de raciocínio lógico da IA (sistemas de Q&A baseados em casamento de padrão e buscas por palavras-chave). Já o paradigma estatístico, ainda engatinhava, com os métodos bayesianos que eram usados, por exemplo, para reconhecimento ótico de caracteres e atribuição de autoria. Também foi na década de 1960 que surgiram os primeiros recursos online: o Brown Corpus, para o inglês, e o DOC (Dictionary on Computer) para o chinês.

A partir de 1970 as pesquisas em PLN ganham força e podem ser organizadas em quatro paradigmas: estatístico, baseado em lógica, NLU e discurso. No paradigma estatístico surgem novos algoritmos para reconhecimento de fala, principalmente cadeias de Markov (HMM), e a metáfora do canal ruidoso (*noisy channel*) e decodificação que seriam a fundamentação para diversos avanços no PLN (ver os Capítulos 2 e Tradução Automática). No paradigma lógico, os trabalhos da época se baseavam em gramáticas funcionais. Trabalhos de interpretação textual (NLU) também surgiram naquela época com os primeiros programas para entendimento da linguagem (*input* em língua natural) com foco na semântica, no discurso, papéis semânticos e suas representações. Por fim, também foram propostos trabalhos que focavam no discurso para problemas como a resolução de correferência (ver Capítulo 12).

As décadas de 1980 e 1990 marcam o estabelecimento do paradigma estatístico como a estratégia hegemônica. Nesse período, modelos que haviam perdido popularidade, como as máquinas de estados finitos e modelos probabilísticos (da IBM), retornam com força total e são a base para avanços como os da tradução automática estatística (ver Capítulo Tradução Automática). Uma vez que sistemas de PLN passam a ser produzidos e consumidos, cresce o foco na forma de avaliação dos modelos, com definição de procedimentos e métricas. Tamém foi nessa época que o aprendizado neural deu seus primeiros passos no PLN e uma das arquiteturas neurais muito utilizadas surgiu nesse período (1985): as RNNs (Recurrent Neural Networks).

A década de 1990 se encerra com a consagração da estatística como paradigma dominante no PLN. Probabilidades passam a ser incorporadas em algoritmos para *parsing*, *part-of-speech tagging*, resolução de correferência e processamento de discurso. As metodologias de avaliação emprestadas do reconhecimento de fala e recuperação de informação passam a ser usadas em diveras aplicações de PLN. Os avanços na velocidade e na capacidade de memória dos computadores permitem a exploração comercial de aplicações de PLN (e.g. reconhecimento de fala e correção gramatical). O surgimento da web enfatiza a necessidade por recuperação (Capítulo Recuperação de Informação) e extração (Capítulo Extração de Informação) de informação baseadas em linguagem. Por fim, outra arquitetura neural muito usada no início do aprendizado neural em PLN também é proposta nesse período (1997): as LSMTs (Long Short-Term Memory).

A primeira década deste século (2000-2010) viu a disponibilidade cada vez maior de grandes quantidades de dados que, combinados com o aprendizado supervisionado e máquinas cada vez mais poderosas, foram usados para gerar soluções para diversos problemas de PLN. Algoritmos de AM como SVM (Support Vector Machines) e técnicas de máxima verossimilhança (*maximum entropy*) se tornaram bastante usuais na geração de modelos de PLN. Também foi nesse período que o aprendizado não-supervisionado ganhou atenção renovada em aplicações como a tradução estatística (ver Capítulo Tradução Automática) e a modelagem de tópicos.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Podemos dizer que a era que se seguiu, que é a que estamos no momento, é a era do aprendizado neural e profundo onde surgiram as técnicas que garantiram uma revolução no PLN como as *word embeddings* (e.g. word2vec em 2013), as arquiteturas encoder-decoder (em 2014), o modelo de atenção (em 2017) e os primeiros modelos de linguagem pré-treinados (e.g. BERT e GPT a partir de 2018). A partir de 2019 passamos a viver na era dos modelos de linguagem de larga escala e a eles dedicamos um volume inteiro nesta edição do livro: o Volume 2.

No Brasil, as primeiras pesquisas em PLN datam da década de 1970. Em 1984 ocorreu a primeira edição da principal conferência de inteligência artificial do Brasil, o BRACIS (antes conhecido como SBIA), na qual haviam vários trabalhos de PLN. Na década de 1990 surgiram grupos de pesquisa de PLN em instituições como a UFRGS, a PUCRS, a Unisinos, a PUC-Rio, a Unicamp, a UFPE e o ICMC-USP. Foi em 1993 que o NILC (Núcleo Interinstitucional de Linguística Computacional) foi criado a partir de um projeto audacioso de desenvolvimento do primeiro revisor ortográfico e gramatical para o português do Brasil.

Atualmente, a principal conferência dedicada ao PLN no Brasil é o STIL[5], criado em 2003 (como TIL) por pesquisadores do NILC. O PROPOR, que existe desde 1993, é outra importante conferência da área para quem se dedica ao PLN do português e suas variantes. Em 2007 o PLN passou a ter representação junto à Sociedade Brasileira de Computação por meio da Comissão Especial de PLN[6] (CEPLN). Em 2020, o grupo Brasileiras em PLN[7] (BPLN) foi criado para unir mulheres brasileiras atuantes ou interessadas no PLN. Foi em 2023 que a primeira edição deste livro, uma iniciativa inovadora do BPLN, foi lançada.

1.4 Os Paradigmas de PLN

Até a década de 1980, o PLN se baseava no que chamamos de **paradigma simbólico**, segundo o qual todo conhecimento sobre a língua é expresso explicitamente em formalismos como léxicos, regras, linguagens lógicas etc., ou seja, formas compreensíveis ao humano. Por exemplo, é possível escrever regras que determinem que, em português, há concordância entre o gênero gramatical atribuído a um substantivo e o gênero atribuído ao adjetivo que o acompanha. Assim, exemplos como "abacaxi maduro" serão considerados corretos de acordo com essa regra, enquanto que outros, como "abacaxi madura", não.

No início dos anos 1990, as máquinas ganharam mais capacidade de memória e processamento, e diversos algoritmos de aprendizado de máquina foram propostos dando origem ao que chamamos de **paradigma estatístico**. Grandes conjuntos de textos (também chamados de ***corpus***) passaram a ser usados como fonte de conhecimento para "ensinar" as máquinas. Por exemplo, fenômenos como a concordância entre substantivo e adjetivo, mencionada anteriormente, passaram a ser aprendidos a partir de exemplos de ocorrência no *corpus* como: "tomate estragado", "kiwi maduro", "gergelim preto". A língua é, então, representada em modelos probabilísticos aprendidos a partir da frequência de ocorrência. Regras explícitas ou implícitas (percursos em árvores, por exemplo) são criadas com base em probabilidades calculadas a partir dos exemplos. Esses modelos são usados para classificar, resumir, traduzir ou gerar novos textos. Uma vez que esses modelos são aprendidos a partir de dados reais, eles têm uma grande chance de serem bons modelos da língua.

[5] https://sol.sbc.org.br/index.php/stil/issue/archive
[6] https://www.sbc.org.br/processamento-de-linguagem-natural/
[7] https://brasileiraspln.com/

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

A tradução automática (Capítulo Tradução Automática) foi a aplicação de PLN que deu notoriedade a esse paradigma estatístico, que era o mais aplicado até a década de 2010.

O tempo passou e as máquinas continuam ganhando poder de memória e processamento, o que possibilita que grandes quantidades de dados sejam processadas por estruturas (arquiteturas e algoritmos) bastante complexas, como as Redes Neurais Profundas (conhecidas em inglês como *deep learning*). No momento da escrita deste capítulo, o **paradigma neural** é o mais adotado para tarefas de PLN. Da mesma forma que o paradigma estatístico, as redes neurais também se baseiam em grandes volumes de dados para aprender um modelo; contudo, a forma como esse aprendizado é realizado é diferente, uma vez que envolve várias camadas de unidades de processamento para reconhecer os padrões recorrentes (para saber mais, sugere-se a leitura do Capítulo Modelos de linguagem). Assim, enquanto em outras técnicas de aprendizado de máquina (*machine learning)* tradicional (*shallow* ou baseado em *features*) os algoritmos especificam como o aprendizado deve ocorrer, no *deep learning*, devido à complexidade das arquiteturas compostas por diversas camadas de processamento, não é possível saber exatamente com base em quê o modelo foi aprendido. Além disso, diferentemente do paradigma simbólico, no paradigma neural, o conhecimento da língua é dado por valores numéricos, e não por símbolos ou regras. Dessa forma, o conhecimento linguístico ou a parte do código que tenha produzido um determinado comportamento são praticamente irrecuperáveis, tornando o código opaco, e seu efeito, não previsível (não determinístico).

Nesse sentido, pode-se notar que o PLN tem acompanhado a evolução de paradigmas da IA: simbólico, estatístico e neural. Porém, diante da insuficiência de uma única abordagem, ganham espaço os **paradigmas híbridos**, que combinam principalmente o simbólico com um dos demais, garantindo, assim, alguma explicitação do conhecimento, consequentemente, alguma explicabilidade dos passos seguidos pelos algoritmos.

Além da IA, o PLN tem intersecção com diversos campos de pesquisa e de aplicação no mercado de trabalho como mineração de textos, recuperação de informação (Capítulo Recuperação de Informação) e ciência de dados. Na atualidade, todas as aplicações computacionais que processam texto são passíveis de utilizar em maior ou menor grau as técnicas de PLN.

1.5 Vale a pena relembrar

Antes de passarmos para os próximos capítulos deste livro, seguem algumas considerações importantes:

- **Diferentes abordagens podem ser aplicadas no PLN,** desde aquelas associadas ao paradigma mais tradicional (o simbólico) àquelas possibilitadas por paradigmas mais recentes, como o estatístico e o neural.
- **Todas as estratégias automáticas para processamento da língua têm limitações**. Assim, o que define a escolha da melhor estratégia são diversos fatores como: apoio de especialistas (necessário para o paradigma simbólico), poder computacional (um limitante para o paradigma neural) e a disponibilidade de recursos linguísticos em grande quantidade (necessária para as abordagens baseadas em *corpus*).
- **A maioria das estratégias processa caracteres e não unidades linguísticas**. Muitas estratégias geram modelos com base em coocorrência e contexto de ocorrência de palavras e frases, ou seja, são abordagens baseadas em padrões de caracteres. Um modelo neural, por exemplo, não sabe que "casa" pode significar o lugar onde

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

"alguém" mora. Desse modo, podemos dizer que as estratégias usadas na maioria das aplicações do PLN não aprendem a língua, mas apenas aprendem a reproduzir e, às vezes, extrapolar (generalizar) o que aprenderam em um *corpus* de treinamento.

- Muita atenção tem sido dada aos algoritmos de aprendizado de máquina e às arquiteturas neurais, mas **nem tanta atenção assim tem sido dada aos formalismos de representação semântica**. Como vimos, nos diversos anos de pesquisa e desenvolvimento em PLN, as estratégias e abordagens vão e vêm, mas a linguagem natural é muito mais complexa de se aprender e compreender do que uma simples contagem de frequências e coocorrências. Assim, apesar de muito esforço sendo empregado na investigação e evolução de métodos neurais, o conhecimento linguístico e de uso da língua ainda não foi completamente representado/capturado por nenhum dos métodos atuais. Por isso, o processamento completo e adequado da língua só será possível com formalismos de representação híbridos, que incorporam também estruturas semânticas explícitas, e, portanto, mais robustos do que os que são usados hoje por métodos estatísticos e neurais.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Parte II

Fala

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Capítulo 2

Texto ou fala?

Camila de Araújo Azevedo
Heliana Ribeiro de Mello
Priscila Osório Côrtes

Publicado em: 26/09/2023

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

2.1 Histórico e panorama da área

O processamento da língua falada depende de uma vasta gama de conhecimentos que inclui acústica, fonologia, fonética, linguística geral, semântica, sintaxe, pragmática, estruturas discursivas, entre outras. Para além disso, outros conhecimentos mais comuns à ciência da computação, à engenharia elétrica, à matemática e, até mesmo à psicologia, também são necessários. Neste contexto, este capítulo visa oferecer um panorama da área e das habilidades e métodos mais conhecidos no universo do processamento computacional da língua falada.

Desde os primórdios do surgimento da interação falada na espécie humana até os dias de hoje – e podemos afirmar com tranquilidade, que assim também será no futuro imaginável –, a fala tem sido o principal instrumento para a troca de informações e de coesão social (Rizzolatti; Arbib, 1998). É através da fala[1] que expressamos nossas emoções, a nossa atitude em relação a fatos e eventos, bem como negociamos ideias e ações. A capacidade linguística nos diferencia de outras espécies, mas é a fala, e o que ela nos proporciona, que nos identifica como humanos. Estima-se que a fala tenha surgido na filogênese humana há cerca de 60 mil anos, enquanto a escrita, que é uma tecnologia desenvolvida pelos humanos, surgiu provavelmente há cerca de 10 mil anos. A chamada "dupla articulação" presente na linguagem humana é uma habilidade exclusiva da nossa espécie. Ela se caracteriza por ser a articulação entre unidades significativas (morfemas) e fonemas, que são elementos finitos que se combinam de forma variada, criando infinitas possibilidades de morfemas[2]. A língua falada é hoje expandida para além do domínio da interação face-a-face para meios como a telefonia, a televisão, a interação via computadores. Os aplicativos para interações multimodais imagem/som ganharam uma dimensão inimaginável com a eclosão da pandemia do Sars-Cov-19 em 2020, demonstrando claramente a preferência dos humanos pela interação via fala.

Tal preferência também se reflete na interação homem-máquina e, apesar de ainda estarmos distantes de um mundo em que homens e máquinas interagem majoritariamente

[1] Ressaltamos que "fala" neste contexto inclui a comunicação gestual em língua de sinais. As línguas de sinais, como LIBRAS, para o português brasileiro, são línguas naturais, com a mesma riqueza e complexidade que as línguas que se utilizam de sinal sonoro.

[2] Cf. a discussão sobre Ciências da Fala, em Barbosa, Plínio A. (2020) Ciências da fala. In: *Speech Sciences Entries*. Speech Prosody Studies Group. Disponível em: https://gepf.falar.org/entries/1.

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

através da verbalização oral, já temos aplicações que nos permitem interagir com as máquinas através de comandos orais no contexto doméstico, comercial e computacional.

Em sua fase inicial, o processamento de língua falada em português era bastante limitado devido à falta de recursos computacionais e técnicas apropriadas. As primeiras abordagens eram baseadas em regras gramaticais e modelos acústicos simples. No entanto, com o avanço da tecnologia e o aumento do poder computacional, novas técnicas e abordagens foram desenvolvidas, resultando em avanços significativos nessa área.

A partir da década de 1990, técnicas baseadas em estatística começaram a ganhar popularidade. Esses modelos estatísticos utilizam algoritmos de aprendizado de máquina, como as redes neurais artificiais, para melhorar o desempenho do processamento de língua falada em português. Com a disponibilidade de grandes quantidades de dados de fala e avanços em hardware e software, os sistemas de reconhecimento de fala começaram a se tornar mais precisos e eficientes.

Outro marco importante no processamento de língua falada em português foi a introdução dos sistemas de síntese de fala (Seção 2.2.3). Esses sistemas permitem que um computador gere fala humana a partir de texto escrito em português. Inicialmente, a síntese de fala em português era baseada em técnicas concatenativas, que envolviam a gravação de segmentos de fala de um locutor humano e a concatenação desses segmentos para gerar a fala sintetizada. A concatenação refere-se ao processo de unir ou combinar várias partes ou segmentos de fala para formar uma sequência contínua ou mais longa de palavras ou frases. Com o tempo, surgiram abordagens baseadas em síntese de formantes (na fala, um formante é uma ressonância específica ou pico de intensidade em um espectrograma de som. Os formantes são associados à forma e ao posicionamento da cavidade oral, da faringe e da língua durante a produção de sons da fala, especialmente as vogais) e síntese de fala concatenativa com modelos estatísticos, proporcionando uma qualidade de síntese cada vez melhor.

Avanços mais recentes no processamento da fala em português estão relacionados ao uso de modelos de linguagem neural (Capítulo Modelos de linguagem), como os modelos de transformação de sequência a sequência (Seq2Seq) e as redes neurais convolucionais (CNNs) e recorrentes (RNNs). Esses modelos têm oferecido resultados impressionantes em várias tarefas de processamento de língua falada, como reconhecimento automático de fala, tradução automática de fala e resumo automático de áudio.

Além disso, com o advento dos assistentes virtuais e sistemas de processamento de linguagem natural, a interação por meio da fala em português tornou-se cada vez mais comum. Empresas de tecnologia estão investindo em pesquisas e desenvolvimento para melhorar a compreensão e a resposta dos sistemas de processamento de língua falada em português, a fim de proporcionar uma experiência mais natural e intuitiva aos usuários.

Para que se alcancem bons resultados no processamento computacional da fala é preciso que haja *datasets* e *corpora* de fala[3] de alta qualidade. Tem havido um esforço considerável da comunidade de pesquisadores para a compilação de dados dessa natureza. Para o português brasileiro, destaca-se o recente *corpus* CORAA ASR v. 1.1 (Corpus de Áudios Anotados)[4] voltado para tarefas de reconhecimento de fala (Candido Junior et al., 2021), que é apresentado no Capítulo 3.

Os sons da fala podem ser digitalizados e processados usando-se algoritmos tanto para **reconhecimento de fala** (transcrição de formas de onda em texto) quanto para **síntese**

[3]Para saber mais sobre *datasets* e *corpora*, sugere-se a leitura do Capítulo 16.

[4]https://github.com/nilc-nlp/CORAA e https://sites.google.com/view/tarsila-c4ai/coraa-versions?pli=1

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

de fala (conversão de texto em formas de onda). O processo de digitalização da fala envolve a conversão do sinal analógico das ondas sonoras em um formato digital que pode ser armazenado e manipulado por um computador. Isso é normalmente feito usando-se um conversor analógico-digital (CAD), que amostra, isto é, faz uma amostragem da onda sonora em intervalos regulares e converte cada amostra em um número binário. Uma vez que o sinal da fala tenha sido digitalizado, ele pode ser processado usando-se várias técnicas, como filtragem, compressão e análise.

Um sistema computacional para a língua falada necessita de capacidades tanto de reconhecimento quanto de síntese de fala. Entretanto, esses dois componentes não são suficientes para a construção de um sistema útil. Um componente de compreensão e diálogo é necessário para a interação com o usuário; o conhecimento de domínio é necessário para guiar a interpretação da fala pelo sistema e permitir que ele determine a ação apropriada. Para todos esses componentes, há uma série de desafios, que incluem robustez, flexibilidade, facilidade de integração e eficiência de engenharia.

2.2 Aspectos teóricos fundamentais

A língua falada é utilizada para diversas funções que se estabelecem entre falantes e ouvintes. A produção e a percepção são ambos elementos importantes na cadeia da fala. A fala se inicia com uma intenção (volição) de comunicação no cérebro do falante, o qual ativa movimentos musculares para a produção de sons. O ouvinte, por sua vez, recebe os sinais sonoros em seu sistema auditivo, processando-os para transformá-los em sinais neurológicos que o cérebro pode compreender. O falante monitora e controla continuamente os órgãos vocais ao receber a sua própria fala como feedback (Moore, 2007).

Considerando os componentes universais da comunicação verbal, a interação falante/ouvinte é tecida a partir de vários elementos distintos. Como dito, o processo de produção da fala começa com a mensagem semântica na mente de uma pessoa a ser transmitida ao ouvinte através da fala. O equivalente computacional ao processo de formulação da mensagem é a semântica da aplicação que cria o conceito a ser expresso. Após a criação da mensagem, o próximo passo é convertê-la em uma sequência de palavras. Cada palavra consiste em uma sequência de fonemas e respectivos alofones (realizações fonéticas correlacionadas do fonema) que correspondem à pronúncia das palavras. Cada frase também contém um padrão prosódico que denota a duração de cada fonema, entonação da frase e volume dos sons. Uma vez que o sistema de linguagem finaliza o mapeamento, o falante executa uma série de sinais neuromusculares. Os comandos neuromusculares realizam o mapeamento articulatório para controlar as cordas vocais, lábios, mandíbula, língua e véu palatino, produzindo assim a sequência sonora como saída final. O processo de compreensão da fala funciona na ordem inversa. Primeiro, o sinal é enviado para a cóclea no ouvido interno, que realiza a análise de frequência como um banco de filtros. Em seguida, um processo de transdução neural converte o sinal espectral em sinais de atividade no nervo auditivo, correspondendo aproximadamente a um componente de extração de recursos. Atualmente, ainda não está claro como a atividade neural é mapeada no sistema de linguagem e como a compreensão da mensagem é alcançada no cérebro.

Os sinais de fala são compostos de padrões sonoros analógicos que servem como base para uma representação discreta e simbólica da linguagem falada – fonemas, sílabas e palavras. A produção e interpretação desses sons são regidas pela sintaxe, semântica e estrutura informacional da língua falada. Neste capítulo, adotamos uma abordagem de

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

baixo para cima para introduzir os conceitos básicos, começando pelos sons e passando pela fonética e fonologia, chegando até as sílabas e palavras.

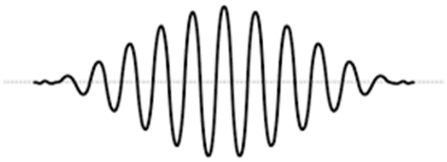
2.2.1 A estrutura da língua falada

Nesta seção, revisamos brevemente os sistemas de produção e percepção de fala humana. Esperamos que, algum dia, a pesquisa em linguagem falada nos permita construir um sistema de computador tão bom quanto o nosso próprio sistema de produção e compreensão de fala.

2.2.1.1 Som

O som é uma onda de pressão longitudinal formada por compressões e rarefações das moléculas de ar, em uma direção paralela àquela da aplicação de energia. Compressões são zonas onde as moléculas de ar foram forçadas pela aplicação de energia a uma configuração mais apertada do que o normal, e rarefações são zonas onde as moléculas de ar estão menos densamente empacotadas. As configurações alternadas de compressão e rarefação de moléculas de ar ao longo do caminho de uma fonte de energia são às vezes descritas pelo gráfico de uma onda senoidal. A forma básica de uma curva senoidal (Figura 2.1) é de uma onda suave, que se repete ao longo de um eixo horizontal. Ela se assemelha a uma série de montanhas e vales, subindo e descendo de forma suave. Neste tipo de representação, as cristas da curva senoidal correspondem a momentos de compressão máxima e os vales correspondem a momentos de rarefação máxima.

Figura 2.1: Curva senoidal



2.2.1.2 Produção de Fala[5]

Aqui revisamos os sistemas básicos de produção de fala humana, que influenciaram a pesquisa em codificação, síntese e reconhecimento de fala.[6]

2.2.1.2.1 Articuladores

A fala é produzida por ondas de pressão de ar que emanam da boca e das narinas de um falante.[7] Na maioria das línguas do mundo, o inventário de fonemas pode ser dividido em duas classes básicas:

- **Consoantes** – articuladas na presença de constrições na garganta ou obstruções na boca (língua, dentes, lábios) enquanto falamos;
- **Vogais** – articuladas sem grandes constrições e obstruções.

[5]Recomendamos o PRAAT como software para análise da fala: https://www.fon.hum.uva.nl/praat/manual/Intro.html
[6]Toda a seção sobre produção da fala é orientada pelo livro de Huang et al. (2001).
[7]Cf. O site Fonética e Fonologia:https://fonologia.org/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

Os sons podem ser subdivididos ainda mais em subgrupos com base em certas propriedades articulatórias. Essas propriedades derivam da anatomia de alguns articuladores importantes e dos locais onde eles tocam as fronteiras do trato vocal humano. Além disso, um grande número de músculos contribui para a posição e o movimento dos articuladores. Nós nos restringimos a apenas uma visão esquemática dos principais articuladores. Os componentes principais do aparelho de produção da fala são os pulmões, traquéia, laringe (órgão de produção de voz), cavidade faríngea (garganta), cavidade oral e nasal. As cavidades faríngea e oral são geralmente referidas como o trato vocal, e a cavidade nasal como o trato nasal. O aparelho de produção de fala humano consiste em:

- **Pulmões:** fonte de ar durante a fala;
- **Cordas vocais (laringe):** quando as pregas vocais são mantidas próximas uma da outra e oscilam uma contra a outra durante um som da fala, o som é categorizado como sonoro. Por exemplo, /b d g/. Quando as pregas são muito soltas ou tensas para vibrar periodicamente, o som é categorizado como surdo. Por exemplo, /p t k/. O local onde as pregas vocais se unem é chamado de glote;
- **Véu palatino (palato mole):** atua como uma válvula, abrindo para permitir a passagem de ar (e, portanto, ressonância) através da cavidade nasal. Sons produzidos com a aba aberta incluem /m/ e /n/;
- Palato duro: uma superfície relativamente dura e longa no teto dentro da boca; quando a língua é colocada contra ela, permite a articulação de consoantes, como o λ em alho /aλu/;
- **Língua:** articulador flexível, afastado do palato para vogais, colocado próximo ou sobre o palato ou outras superfícies duras para articulação de consoantes;
- **Dentes:** outro local de articulação usado para segurar a língua para certas consoantes, como /t d/;
- **Lábios:** podem ser arredondados ou espalhados para afetar a qualidade das vogais, e completamente fechados para interromper o fluxo de ar oral em certas consoantes /p b m/.

2.2.1.2.2 O Mecanismo de Sonorização

A distinção mais fundamental entre os tipos de som na fala é a distinção sonoro/surdo. Sons sonoros, incluindo vogais, têm em sua estrutura temporal e de frequência um padrão regular que sons surdos, como a consoante /s/, não possuem. Sons sonoros geralmente têm mais energia. O que no mecanismo de produção de fala cria essa distinção fundamental? Como já dito na Seção 2.2.1.2.1, quando as pregas vocais vibram durante a articulação do fonema, o fonema é considerado sonoro; caso contrário, é surdo. Vogais são sonoras durante toda a sua duração. Os timbres distintos de vogais são criados usando a língua e os lábios para moldar a principal cavidade de ressonância oral de maneiras diferentes. As pregas vocais vibram em taxas mais lentas ou mais rápidas, desde tão baixas quanto 60 ciclos por segundo (Hz) para um homem de tamanho grande, até 300 Hz ou mais para uma mulher ou criança pequena. A taxa de ciclagem (abertura e fechamento) das pregas vocais na laringe durante a fonação de sons sonoros é chamada de frequência fundamental(f0). Isso ocorre porque ela estabelece a linha de base periódica para todos os harmônicos de frequência mais alta contribuídos pelas cavidades de ressonância faríngea e oral. A frequência fundamental também contribui mais do que qualquer outro fator único para a percepção de altura (o aumento e queda semelhante à música das tonalidades de voz) na fala.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Uma vez que a onda glotal é periódica, consistindo na frequência fundamental (f0) e em um número de harmônicos (múltiplos integrais de f0), ela pode ser analisada como uma soma de ondas senoidais. As ressonâncias do trato vocal (acima da glote) são excitadas pela energia glotal. Vamos supor, para simplicidade, que o trato vocal seja um tubo reto de área transversal uniforme, fechado na extremidade da glote e aberto nos lábios. Quando a forma do trato vocal muda, as ressonâncias também mudam. Harmônicos próximos às ressonâncias são enfatizados, e, na fala, as ressonâncias das cavidades que são típicas de configurações articulatórias particulares (por exemplo, os diferentes timbres vocálicos) são chamadas de formantes. As vogais em uma forma de onda de fala real podem ser visualizadas a partir de várias perspectivas diferentes, por exemplo, enfatizando uma visão em seção transversal das respostas harmônicas em um único momento ou, por outro lado, uma visão de longo prazo da evolução da trajetória dos formantes ao longo do tempo.

2.2.1.3 Percepção da Fala

Existem dois componentes principais no sistema de percepção auditiva: os órgãos auditivos periféricos (orelhas) e o sistema nervoso auditivo (cérebro). A orelha (ouvido externo) capta um sinal de pressão acústica, processa-o, transformando-o primeiro em um padrão de vibração mecânica na membrana basilar e depois representando o padrão por uma série de pulsos a serem transmitidos pelo nervo auditivo. A informação perceptual é extraída em vários estágios do sistema nervoso auditivo. Nesta seção, focamos principalmente nos órgãos auditivos.

2.2.1.3.1 Fisiologia do Ouvido

O ouvido humano tem três partes: o ouvido externo, o ouvido médio e o ouvido interno. O ouvido externo consiste na parte visível externa e no canal auditivo externo, que forma um tubo ao longo do qual o som viaja. Esse tubo tem cerca de 2,5 cm de comprimento e é coberto pelo tímpano na extremidade distante. Quando variações na pressão do ar alcançam o tímpano do exterior, ele vibra e transmite as vibrações aos ossos adjacentes do seu lado oposto. A vibração do tímpano está na mesma frequência (compressão e rarefação alternadas) que a onda de pressão sonora que chega. O ouvido médio é um espaço ou cavidade cheia de ar com cerca de 1,3 cm de largura e volume de cerca de 6 cm^3. O ar viaja pela abertura (quando aberta) que conecta a cavidade com o nariz e a garganta. Há, ainda, a janela oval, que é uma pequena membrana na interface óssea com o ouvido interno (cóclea). Uma vez que as paredes da cóclea são ósseas, a energia é transferida por ação mecânica do estribo para uma impressão na membrana que se estende sobre a janela oval.

A estrutura relevante do ouvido interno para a percepção sonora é a cóclea, que se comunica diretamente com o nervo auditivo, conduzindo uma representação do som para o cérebro. A cóclea é um tubo espiralado de cerca de 3,5 cm de comprimento, que se enrola cerca de 2,6 vezes. A espiral é dividida, principalmente pela membrana basilar que corre longitudinalmente, em duas câmaras preenchidas de líquido. A cóclea pode ser considerada grosseiramente como um banco de filtros, cujas saídas são ordenadas por localização, de modo que uma transformação de frequência local é realizada. Os filtros mais próximos da base da cóclea respondem às frequências mais altas, e aqueles mais próximos do ápice respondem às mais baixas.

Em psicoacústica, faz-se uma distinção básica entre os atributos perceptuais de um som, especialmente de um som de fala, e as propriedades físicas mensuráveis que o caracterizam.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Cada um dos atributos perceptuais, conforme listado a seguir, parece ter uma forte correlação com uma propriedade física principal, mas a conexão é complexa, porque outras propriedades físicas do som podem afetar a percepção de maneiras complexas.

O Quadro 2.1 traz a relação entre atributos perceptuais e físicos do som.

Quadro 2.1: Relação entre atributos perceptuais e físicos do som

Quantidade física	**Qualidade perceptual**
Intensidade	Volume
Frequência fundamental	Tom
Forma espectral	Timbre
Tempo de início/fim	Temporização
Diferença de fase na audição binaural	Localização

Embora sons com níveis de intensidade maiores geralmente soem mais altos, a sensibilidade do ouvido varia com a frequência e a qualidade do som. Uma divergência fundamental entre as qualidades físicas e perceptuais é o fenômeno da percepção de igualdade de intensidade não uniforme de tons de frequências variadas. Em geral, tons de diferentes alturas têm diferentes níveis percebidos de volume. Há uma relativa insensibilidade do ouvido a sons de baixa frequência em níveis de intensidade moderados a baixos. A sensibilidade auditiva atinge um máximo em torno de 4 kHz, que está próximo da primeira frequência de ressonância do canal auditivo externo, e atinge outro pico em torno de 13 kHz, a frequência da segunda ressonância.

A altura está, de fato, mais intimamente relacionada com a frequência fundamental. Quanto maior a frequência fundamental, maior a altura que percebemos. No entanto, a discriminação entre duas alturas depende da frequência da altura inferior. A altura percebida mudará à medida que a intensidade aumentar e a frequência for mantida constante.

Em um exemplo da não identidade de efeitos acústicos e perceptuais, foi observado experimentalmente que, quando o ouvido é exposto a dois ou mais tons diferentes, é comum que um tom possa mascarar os outros. O mascaramento provavelmente é mais bem explicado como um deslocamento ascendente no limiar auditivo do tom mais fraco pelo tom mais alto. Tons puros, sons complexos, bandas estreitas e amplas de ruído mostram diferenças em sua capacidade de mascarar outros sons. Em geral, tons puros, próximos em frequência, se mascaram mais do que tons amplamente separados em frequência. Um tom puro mascara tons de frequência mais alta com mais eficácia do que tons de frequência mais baixa. Quanto maior a intensidade do tom de mascaramento, mais ampla é a faixa de frequências que ele pode mascarar. O mascaramento, no contexto da fala e da audição, pode ter um impacto significativo, causando dificuldade de compreensão e reduzindo a inteligibilidade, além de aumentar o esforço de escuta. O mascaramento pode afetar o reconhecimento automático de fala aumentando a taxa de erros, levando à perda de partes importantes do discurso (perda de contexto) e dificultando a separação de vozes.

A escuta binaural melhora muito nossa capacidade de sentir a direção da fonte de som. A atenção à localização está principalmente focada na discriminação lateral ou de lado a lado. As pistas de tempo e intensidade têm diferentes impactos para frequências baixas e altas, respectivamente. Sons de baixa frequência são lateralizados principalmente com base na diferença interaural de tempo, enquanto sons de alta frequência são localizados principalmente com base na diferença interaural de intensidade.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Finalmente, uma questão perceptual interessante é a questão da qualidade de voz distinta. O discurso de pessoas diferentes soa diferente. Em parte, isso se deve a fatores óbvios, como diferenças na frequência fundamental característica causada, por exemplo, pela maior massa e comprimento das pregas vocais masculinas adultas em comparação com as femininas. Mas existem efeitos mais sutis também.

Em psicoacústica, o conceito de timbre (de um som ou instrumento) é definido como o atributo da sensação auditiva pelo qual um sujeito pode julgar que dois sons apresentados de maneira semelhante, com a mesma intensidade e altura, são diferentes. Em outras palavras, quando todas as diferenças facilmente mensuráveis são controladas, a percepção restante de diferença é atribuída ao timbre. Isso é mais facilmente ouvido na música, onde a mesma nota na mesma oitava, tocada por igual tempo, por exemplo, em um violino, soa diferente de uma flauta. O timbre de um som depende de muitas variáveis físicas, incluindo a distribuição de energia espectral do som, o envelope temporal, a taxa e profundidade de modulação de amplitude ou frequência e o grau de inarmonia de seus harmônicos.

2.2.1.3.2 Análise de Frequência

Pesquisadores têm realizado trabalhos experimentais psicoacústicos para derivar escalas de frequência que tentam modelar a resposta natural do sistema perceptual humano, uma vez que a cóclea do ouvido interno atua como um analisador de espectro. O complexo mecanismo do ouvido interno e do nervo auditivo implica que os atributos perceptuais de sons em diferentes frequências podem não ser completamente simples ou lineares por natureza. É bem conhecido que a altura musical ocidental é descrita em oitavas e semitons. A altura musical percebida de tons complexos é basicamente proporcional ao logaritmo da frequência. Para tons complexos, a diferença perceptível para frequência é essencialmente constante na escala de oitavas/semitons. As escalas de altura musical são usadas em pesquisas prosódicas (sobre a geração de contorno de entonação da fala).

2.2.1.4 Panorama genérico dos níveis de análise da fala

A fala, diferentemente da escrita, não é uma tecnologia desenvolvida pelos humanos. É algo bem mais complexo e antigo, sendo hoje considerada, por alguns, como uma dotação genética e, por outros, como o produto de diferentes processos cognitivos e corpóreos.

A fala humana pode ser definida genericamente como o processo de expressar pensamentos, ideias e emoções por meio da produção de sons articulados. É uma forma de comunicação específica dos seres humanos e é fundamental para a interação social e o desenvolvimento das sociedades.

A caracterização da fala humana envolve vários aspectos tais como:

- **Produção de sons articulados:** A fala envolve a produção de sons através da coordenação dos órgãos articulatórios, como a língua, os lábios, os dentes e a glote. Esses órgãos são responsáveis por modificar a corrente de ar expirada pelos pulmões para produzir os diferentes sons da fala.
- **Sistema linguístico (linguagem):** A fala é mediada pela linguagem, que é um sistema de símbolos e regras que permite a comunicação entre os indivíduos. A linguagem compreende elementos fonéticos (sons), fonológicos (padrões de som), morfológicos (estrutura das palavras), sintáticos (ordem das palavras), semânticos (significado das palavras) e pragmáticos (uso da linguagem em contextos específicos).

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

- **Expressão de pensamentos e emoções:** A fala humana permite expressar uma ampla gama de pensamentos, ideias e emoções. Além da transmissão de informações, a fala também é utilizada para expressar sentimentos, intenções, opiniões e experiências pessoais.
- **Comunicação social:** A fala é um meio de interação social fundamental. Por meio da fala, os indivíduos podem se comunicar, compartilhar informações, estabelecer conexões emocionais, resolver problemas e coordenar atividades em grupo.
- **Aquisição:** A habilidade de falar é adquirida ao longo do desenvolvimento humano. As crianças passam por um processo de aprendizado da fala, no qual adquirem as habilidades motoras necessárias para articular os sons e aprendem as regras e estruturas da linguagem de seu ambiente.

É importante ressaltar que a fala humana é altamente diversa e comporta variações entre diferentes idiomas, culturas e indivíduos. Além disso, a fala também pode ser afetada por condições clínicas, como distúrbios da fala e da linguagem.

Diferentemente do que acontece para a escrita, o processamento computacional da fala não parte do encadeamento simbólico de grafemas organizados em itens lexicais e suas supra-estruturas sintáticas. É preciso converter o sinal sonoro em símbolos passíveis de análise por um sistema computacional, ou seja, as ondas sonoras precisam ser convertidas em bits processáveis computacionalmente. Ademais, a fala não pode prescindir de um nível analítico comumente ignorado pelas análises da escrita: a pragmática e, mais especificamente o seu nível prosódico e suas correspondências na estruturação informacional. Neste capítulo não há a possibilidade de explorarmos este assunto com a profundidade que ele merece, portanto recomendamos ao leitor recorrer a leituras específicas para se inteirar sobre isso.

Nas próximas subseções, faremos um apanhado genérico sobre o nível analítico mínimo, o fonético-fonológico.

2.2.1.4.1 Fonética e fonologia

Agora discutiremos as noções de fonética e fonologia básicas necessárias para o processamento da linguagem falada. Fonética refere-se ao estudo dos sons da fala, sua produção, classificação e transcrição. Fonologia é o estudo da distribuição e padrões dos sons da fala em uma língua e das suas regras implícitas.

Ao linguista Ferdinand de Saussure (1857-1913) atribui-se a observação de que a relação entre um sinal e o objeto significado por ele é arbitrária. Assim, um mesmo conceito é arbitrariamente expresso em línguas diferentes: usamos [pɛ] em português para nos referirmos ao mesmo conceito que em inglês foneticamente seria [fʊt] . Para a fonética, isso significa que os sons da fala não têm um significado intrínseco e devem ser distribuídos aleatoriamente no léxico.

Os sons são apenas um conjunto de efeitos arbitrários disponibilizados pela anatomia vocal humana. Assim como as impressões digitais, a anatomia vocal de cada falante é única, o que resulta em vocalizações também únicas. No entanto, a comunicação linguística é baseada na comunalidade de formas no nível perceptual. Para permitir a discussão das semelhanças, os pesquisadores identificaram certas características gerais dos sons da fala que são adequadas para a descrição e classificação das palavras nos dicionários. Eles também adotaram vários sistemas de notação para representar o subconjunto de fenômenos fonéticos que são cruciais para o significado.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Na ciência da fala, o termo fonema é usado para denotar qualquer uma das unidades mínimas de som da fala em uma língua que podem servir para distinguir uma palavra de outra. O termo fone é utilizado para denotar a realização acústica de um fonema. Há duas classes de fonemas: vogais e consoantes (Seção 2.2.1.2.1).

As vogais são definidas fonologicamente com base em três características principais: qualidade, altura e tensão.

A qualidade vocálica refere-se à diferença perceptível entre os diferentes sons de vogais. Ela é determinada principalmente pela posição da língua e pelos formatos das cavidades oral e faríngea durante a produção da vogal. Por exemplo, as vogais podem ser classificadas como "abertas" ou "fechadas", dependendo da posição da mandíbula e da abertura da boca. Por exemplo, /a/ é uma vogal aberta e /i/ é uma vogal fechada.

A altura vocálica se refere à posição vertical da língua em relação ao palato durante a produção da vogal. As vogais podem ser classificadas como "alta", "média" ou "baixa" com base na posição da língua. Por exemplo, a vogal /i/ em "pique" é considerada alta, enquanto a vogal /a/ em "casa" é considerada baixa.

A tensão vocálica se refere à tensão muscular envolvida na produção da vogal. As vogais podem ser classificadas como "tensas" ou "frouxas". Vogais tensas são produzidas com maior tensão muscular e duração, enquanto vogais frouxas são produzidas com menos tensão muscular e têm uma duração mais curta. No português brasileiro não se considera que haja essa diferenciação. No português europeu, dependendo do dialeto, seriam encontradas vogais tensas como o /ɔ/ em "corta" ou "porta", e vogais frouxas como o /i/ em "pia" ou "fria".

Essas características fonológicas das vogais são usadas para distinguir as palavras em um determinado idioma. As diferenças na qualidade, altura e tensão vocálicas são consideradas contrastivas e podem levar a diferentes significados das palavras. Por exemplo, as palavras "bela"/ ˈbɛlɐ/ e "bola"/ ˈbɔlɐ/ são distinguidas pela qualidade vocálica dos fonemas /ɛ/ e /ɔ/ respectivamente.

A forma e a posição da língua na cavidade oral não formam uma obstrução significativa do fluxo de ar durante a articulação das vogais. No entanto, variações no posicionamento da língua conferem a cada vogal seu caráter distintivo, alterando a ressonância, assim como diferentes tamanhos e formas de garrafas produzem efeitos acústicos diferentes quando são golpeadas. A energia primária que entra nas cavidades faríngea e oral na produção das vogais vibra na frequência fundamental. As principais ressonâncias das cavidades oral e faríngea para as vogais são chamadas de f1 e f2 - primeiro e segundo formantes, respectivamente. Eles são determinados pelo posicionamento da língua e pela forma do trato oral nas vogais e determinam o timbre ou a qualidade característica da vogal.

As consoantes, por outro lado, são definidas fonologicamente como sons produzidos por meio de um bloqueio ou estreitamento parcial ou completo do trato vocal, que resulta na turbulência do ar que passa através do ponto de obstrução.

Existem diferentes tipos de consoantes, classificadas de acordo com o ponto e modo de articulação, e também com a presença ou ausência de vozeamento (consoantes surdas e sonoras) (Seção 2.2.1.2.1). Por exemplo, as consoantes, quanto ao ponto de articulação, podem ser bilabiais, alveolares, palatais ou velares, além de serem oclusivas, fricativas, aproximantes ou nasais, quanto ao modo de articulação, entre outras classificações possíveis.

Um exemplo de um par de consoantes contrastivas no português seria /p/ e /b/. Ambas são consoantes oclusivas bilabiais, produzidas bloqueando completamente o fluxo de ar nos lábios (para /p/) ou. além disso, vibrando as cordas vocais enquanto bloqueiam o fluxo

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

de ar (para /b/).

2.2.1.4.2 O Alofone: Som e Contexto

As tabelas que representam vogais e consoantes fornecem símbolos abstratos para os fonemas[8] – principais distinções sonoras. As unidades fonêmicas devem estar correlacionadas com distinções de significado potencial. Por exemplo, a mudança criada ao manter a língua alta e à frente (/i/) em comparação à posição diretamente abaixo (frontal) para /e/, no contexto consonantal /m _ w/, corresponde a uma importante distinção de significado no léxico do português: mil /miw/ vs. meu /mew/. Esta distinção de significado, condicionada por um par de sons bastante similares, em um contexto idêntico, justifica a inclusão de /i/ e /e/ como distinções logicamente separadas. No entanto, um dos sons fundamentais que distingue significados é muitas vezes modificado de forma sistemática por seus vizinhos fonéticos. O processo pelo qual sons vizinhos influenciam um ao outro é chamado de coarticulação. As variações na realização fonética de um fonema, resultantes dos processos coarticulatórios, são chamadas de alofones. As diferenças alofônicas são sempre categóricas, ou seja, podem ser entendidas e denotadas por meio de um pequeno número delimitado de símbolos ou diacríticos nos símbolos fonêmicos básicos.

2.2.1.4.3 Taxa de articulação e Coarticulação

Além dos alofones, existem outras variações na fala para as quais não é possível delimitar um pequeno conjunto de categorias estabelecidas de variação. Essas variações são graduais, existindo ao longo de uma escala para cada dimensão relevante, com falantes distribuídos de maneira ampla. Falantes individuais podem variar suas taxas de acordo com o conteúdo e contexto de sua fala, e também pode haver grandes diferenças entre os falantes de uma dada língua. Alguns falantes podem fazer pausas frequentes, enquanto outros podem falar muitas palavras por minuto com quase nenhuma pausa entre enunciados. Nas taxas mais rápidas, é menos provável que os alvos de formantes sejam completamente alcançados. Além disso, alofones individuais podem se fundir ou desaparecer completamente (por exemplo, possibilidades do dialeto mineirês no enunciado “você sabe se esse ônibus passa na Savassi”, passível de realização, em representação ortográfica, como “cêsasessonspasansavas”)

2.2.1.4.4 Sílabas e Palavras

Os fonemas são como tijolos em uma construção. Para contribuir para o significado de uma língua, eles devem ser organizados em extensões coesas mais longas, e as unidades formadas devem ser combinadas em padrões característicos para ter significado, como sílabas e palavras.

A sílaba, uma unidade intermediária, é considerada como interposta entre os fonemas e o nível da palavra. O conceito de sílaba é complexo, com implicações tanto para a produção quanto para a percepção da fala. Aqui trataremos a sílaba como uma unidade perceptual. Em português, as sílabas geralmente são centradas em torno de vogais. Por exemplo, numa palavra como “casa” /ka.za/, há duas sílabas porque há duas vogais. Para dividir completamente uma palavra em sílabas, é necessário fazer julgamentos de afiliação consonantal (tomando as vogais como pico da sílaba). A questão de saber se tais julgamentos devem ser baseados em critérios articulatórios ou perceptuais, e como podem ser

[8] Consultar a tabela da Associação Internacional de Fonética disponível em https://www.internationalphoneticassociation.org/content/full-ipa-chart e no Apêndice A.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

rigorosamente aplicados, ainda não está resolvida. Os núcleos das sílabas podem ser considerados picos de sonoridade (seções de alta amplitude). Esses picos de sonoridade têm vizinhanças afiliadas de sonoridade estritamente não crescente. Para a diferenciação dos níveis de sonoridade, pode-se utilizar uma escala de sonoridade, classificando consoantes ao longo de um continuum de oclusivas, africadas, fricativas e aproximantes. Portanto, em uma palavra como "verbal", a silabificação seria "ver-bal", mas não "ve-rbal", porque colocar a aproximante /r/ antes da oclusiva /b/ na segunda sílaba violaria o requisito de sonoridade não crescente em direção à sílaba.

As sílabas são consideradas pelos fonólogos como tendo uma estrutura interna, e vale a pena conhecer os termos atribuídos às partes dessa estrutura. Considere uma sílaba como "trans" /trans/, por exemplo. Ela consiste em um pico vocálico, chamado de núcleo, cercado pelos outros sons em suas posições características. O elemento inicial de uma sílaba é o ataque - preenchido por consoantes. O ataque é um elemento opcional - há sílabas sem ataque, por exemplo, em uma palavra como "as". A rima consiste da combinação do núcleo com consoantes finais, a coda, se estas estiverem presentes. Em alguns tratamentos, a última consoante em um *cluster* de final de sílaba pertenceria a um apêndice e não à coda. Assim, em "trans", teríamos /tr/ em ataque e /ans/ em rima; a rima é formada pelo núcleo, que é /a/, e pela coda que é /ns/. A sílaba é às vezes considerada o domínio primário da coarticulação, ou seja, os sons dentro de uma sílaba influenciam mais a realização uns dos outros do que os mesmos sons se estiverem separados por uma fronteira de sílaba.

O conceito de palavra parece intuitivamente óbvio para a maioria dos falantes de línguas indo-europeias. A palavra pode ser definida, de forma geral, como: um item lexical, com um significado aceito em uma determinada comunidade de fala, e que tem a liberdade de combinação sintática permitida pela sua classe (substantivo, verbo etc.).

Na fala, há um problema de segmentação das palavras: elas se fundem, a menos que sejam afetadas por uma disfluência (problema não intencional de produção de fala) ou pela pausa deliberada (silêncio) por alguma razão estrutural ou comunicativa. Isso é surpreendente para muitas pessoas, porque a alfabetização condicionou os falantes/leitores de línguas indo-europeias a esperar um espaço em branco entre as palavras na página impressa. Mas na fala, apenas algumas pausas verdadeiras (o equivalente sonoro de um espaço em branco entre sinais gráficos na escrita) podem estar presentes. Portanto, o que parece para o olho do leitor como "você sabe se esse ônibus passa na Savassi" na escrita, soaria para o ouvido, se simplesmente usarmos letras para representar seus sons correspondentes no dialeto mineirês, como "cêsasessonspasansavas" (Seção 2.2.1.4.3) – não há pausas nesse enunciado. Frequentemente, o que encontramos na fala, são quebras prosódicas, que podem ser de natureza não-terminal – indicando unidades entoacionais em um enunciado e representadas por /, e quebras terminais, indicando a conclusão de um enunciado e representadas por //. Assim, dependendo da constituição informacional, uma sequência de palavras como: "não deu a altura que a Mari marcou lá", pode ser enunciada com propósitos ilocucionários distintos como as seguintes configurações, dentre outras:

> não deu a altura que a Mari marcou lá // um enunciado, com uma unidade entoacional;
> não // deu a altura que a Mari marcou lá // dois enunciados, com uma unidade entoacional cada;
> não // deu a altura / que a Mari marcou / lá // dois enunciados, um com uma unidade entoacional e o outro com três unidades entoacionais.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Certos fatos sobre a estrutura das palavras e as suas possibilidades de combinação são evidentes para a maioria dos falantes nativos e foram confirmados por décadas de pesquisa linguística. Alguns desses fatos descrevem as relações entre as palavras quando consideradas isoladamente, outros dizem respeito a grupos de palavras relacionadas que parecem intuitivamente similares ao longo de alguma dimensão de forma ou significado - essas propriedades são chamadas de paradigmáticas. As propriedades paradigmáticas das palavras incluem a sua classe gramatical, a sua morfologia flexional e derivacional e a sua estrutura em compostos. Outras propriedades das palavras dizem respeito ao seu comportamento e distribuição quando combinadas para fins comunicativos em enunciados – essas propriedades são chamadas de sintagmáticas.

2.2.2 Reconhecimento de fala

A tarefa de reconhecimento de fala, também conhecida como ASR (do inglês, *automatic speech recognition*), consiste na transformação do sinal acústico de um trecho de fala em um trecho de texto (Figura 2.2).

Figura 2.2: Conversão do áudio da fala em sua transcrição textual.



Essa tarefa tem diversas aplicações, mas a mais difundida é no uso de assistentes de voz, também conhecidos como assistentes virtuais. Os assistentes, comumente embutidos em celulares, como o próprio nome revela, foram criados para ajudar as pessoas em tarefas corriqueiras, como enviar mensagens, fazer ligações, agendar compromissos etc. Para que a ajuda dos assistentes "valha a pena", eles devem interagir com o humano da forma mais natural, isto é, por meio da fala. Para que isso aconteça, o assistente precisa, antes de tudo, compreender a fala do humano. A primeira etapa dessa compreensão[9] envolve o reconhecimento da fala, ou a sua conversão em texto.

No processamento da fala, assim como em diversas aplicações de PLN na atualidade, também concluiu-se ao longo do tempo que os modelos de aprendizado profundo, baseados em dados, são os que geram melhores resultados. Essa abordagem se baseia em grandes quantidades de dados, a partir dos quais a rede neural conseguirá aprender, isto é, identificar padrões e ajustar os pesos dos neurônios. No caso do reconhecimento de fala, os dados são *corpora* de áudio e texto, isto é, para cada trecho de áudio produzido por humanos, em geral uma sentença ou enunciado, deve haver uma transcrição correspondente, para que o modelo consiga associar uma coisa à outra. A seguir, falaremos mais sobre como devem ser esses dados, e sobre aspectos fundamentais do reconhecimento de fala.

[9] A compreensão da fala envolve também a tarefa chamada de NLU, *Natural Language Understanding*, na qual aspectos relacionados ao sentido do enunciado são processados pelo computador e convertidos em algum tipo de ação. No caso dos assistentes de voz, um enunciado como "Escrever um e-mail", dito pelo usuário, será convertido na ação de escrita do e-mail - provavelmente algum aplicativo de e-mail será aberto ou o assistente perguntará para o usuário quem seria o destinatário ou qual o texto do e-mail etc. Essa tarefa não faz parte do que hoje está no escopo da maioria dos modelos de reconhecimento de fala, mas é a etapa que a segue.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

2.2.2.1 Coleta de dados

Os dados, que são o ponto de partida para o treinamento de uma rede neural, devem ser os mais representativos possíveis para a língua falada que se deseja processar. O que isso quer dizer? Da mesma forma como acontece com humanos, a rede neural aprende a partir do que é mostrado a ela, e ela aprende melhor o que for mostrado mais vezes. Nesse sentido, essa seção aborda alguns pontos muito importantes na coleta dos dados: propósito, público-alvo, variações de fala e contexto.

No caso do reconhecimento de fala, é ideal que se tenha em mente para qual **tipo de produto** o modelo de ASR será usado. Tomando novamente como exemplo os assistentes virtuais, seu objetivo principal é o reconhecimento correto de comandos de voz. Dessa forma, os dados para o treinamento da rede neural deverão conter também[10] comandos de voz, instâncias primordiais da interação de usuários com assistentes. É claro que é possível construir um reconhecedor de fala "geral", isto é, que não esteja destinado a um tipo específico de aplicação, mas que visa a reconhecer qualquer tipo de fala que for dado como entrada, seja um diálogo com um *chatbot*, seja uma conversa entre amigos. No entanto, a acurácia de um modelo "geral" tenderá a ser bem inferior à de um modelo específico, uma vez que a fala espontânea encontrada em conversas entre amigos possui muitas particularidades que dificultam o reconhecimento, tais como sobreposição de fala, ruídos de ambiente e fala menos articulada.

Os dados também precisam representar o **usuário-alvo**. Com relação a assistentes de voz, os usuários costumam ser pessoas portadoras de celulares, o que hoje em dia significa "praticamente todo mundo". Mas, pensando bem, talvez nem tanto crianças abaixo de 12 anos ou idosos com mais de 70. Dessa forma, as gravações que compõem o *corpus* de treinamento precisam ser feitas por todo tipo de usuário, mas especialmente por adolescentes e adultos de uma faixa etária entre 12 e 70 anos, em igual proporção de homens e mulheres. Se um modelo for treinado apenas com crianças do gênero feminino, por exemplo, ele será excelente em reconhecer a fala de crianças do gênero feminino, mas provavelmente bem ruim em reconhecer a fala de senhores de 70 anos.

Outro ponto ao qual devemos nos atentar no momento de coleta de dados é a **representatividade dialetal**. Da mesma forma que o modelo precisa ver áudios produzidos tanto por homens quanto por mulheres, adolescentes e idosos, ele também precisa ver áudios de usuários de Caucaia (CE) e de Uruguaiana (RS), por exemplo, localidades nas quais o português falado difere consideravelmente no âmbito fonético, principalmente. Se o modelo for treinado com dados de usuários da mesma variedade dialetal, ele será bom em reconhecer a fala desses usuários, mas não tão bom em reconhecer a fala de usuários de outras regiões. Nesse sentido, vale mencionar que enquanto as variações de fala encontradas nas variantes do português brasileiro e europeu – ou mesmo nos diferentes sotaques e pronúncias dentro do próprio Brasil – têm um grande impacto no PLN da fala, esse impacto no PLN de texto é bem menor.

Finalmente, é preciso também levar em consideração a **forma como a gravação foi feita**. Idealmente, para o produto assistente de voz, as gravações que comporão o *corpus* de treinamento deverão também ter sido feitas utilizando-se o gravador do celular, inclusive com os ruídos de fundo típicos do contexto de uso final da aplicação. As pessoas utilizam o celular na rua, dentro de carros, em casa, em restaurantes, onde há ruídos de conversas, trânsito, música etc., mas muito raramente em estúdios com isolamento acústico perfeito.

[10] Digo "também", porque se os dados contiverem apenas comandos de voz, a aplicação ficará muito restrita. Falaremos disso mais adiante.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Portanto, é preciso mostrar à rede neural uma parcela significativa de áudios com esses tipos de ruído[11].

Em resumo, os dados do treinamento de uma rede neural precisam ser representativos da interação ou contexto de uso, tanto no conteúdo e formato do texto, quanto na forma de gravação, e do perfil de usuário que se quer atingir.

Talvez o leitor esteja se perguntando onde é possível encontrar dados tão peculiares. De fato, esse é um grande desafio da tarefa de reconhecimento de fala, senão o maior. Em se tratando do português, assim como faltam recursos para outras tarefas de PLN, faltam também *corpora* de áudio e texto suficientemente grandes que estejam disponíveis de forma gratuita. Há alguns recursos grátis na internet, como o Mozilla Common Voice (sentenças lidas, em sua maioria)[12] e o LibriVox (audiolivros)[13], mas, infelizmente, eles são insuficientes em termos do número de horas de gravação para se treinar um modelo *end-to-end* do zero. Em geral, o treinamento de uma rede neural para o reconhecimento de fala requer milhares de horas[14]. Fica aqui um convite aos recém-chegados à área para investir na coleta de dados para o português brasileiro.

Para lidar com essa questão da disponibilidade de dados, existem algumas técnicas. Uma técnica bastante usada é a de aumento de dados (*data augmentation*)[15]. Essa estratégia não é restrita ao reconhecimento de fala, mas, no caso desta tarefa, se refere ao aumento dos dados com base em manipulações dos dados já existentes. Um número de gravações do *corpus* de treinamento pode, por exemplo, sofrer adição de ruídos diversos, como os mencionados anteriormente. Suponhamos que o *corpus* de treinamento seja composto por 100 horas de gravação. Podemos, por exemplo, separar 20% dos áudios e adicionar cinco tipos de ruídos a eles, de modo que teremos ao final 200 áudios diferentes (100 áudios iniciais + 100 gerados por manipulação). Assim, os dados resultantes serão diferentes entre si, mas não haverá o trabalho de se criar novos dados do zero. Há outras técnicas para se melhorar a acurácia de um modelo, das quais falaremos na Seção 2.2.2.5.

2.2.2.2 Pré-processamento do texto: limpeza e formatação

Uma vez coletados os dados de texto e fala para formar o *corpus* paralelo de treinamento, é necessário formatá-los para que possam servir de entrada para a rede neural. Essa seção descreve o processo de limpeza e formatação do texto correspondente à transcrição dos áudios. Idealmente, não deve haver muitos erros de digitação ou grafia nas transcrições, para que a rede não aprenda errado. Em outras palavras, a saída de um reconhecedor não deve conter erros de grafia, por isso não seria bom treinar um modelo com um *corpus* no qual o *token* "tambem" ocorresse um número igual ou superior de vezes que sua versão correta, "também". Se esse fosse o caso, o modelo aprenderia que o *chunk* acústico [tẽbẽj] [tẽ corresponderia a "tambem", e, por conseguinte, a saída do modelo conteria o *typo*

[11] É claro que áudios com ruídos muito intensos atrapalham consideravelmente o reconhecimento de fala e não devem ser considerados válidos ou representativos do uso, e por isso, não devem ser incluídos no treinamento.

[12] https://commonvoice.mozilla.org/en/datasets

[13] https://librivox.org/search?primary_key=52&search_category=language&search_page=1&search_form=get_results

[14] O número de horas depende da arquitetura de rede neural utilizada. Estamos considerando aqui o estado da arte, que são modelos *end-to-end*. Modelos híbridos conseguem ser bem treinados com bem menos horas. Falaremos mais disso na seção sobre arquiteturas.

[15] Na realidade, há mais de uma técnica de aumento de dados. Uma delas, bastante utilizada em treinamentos para ASR, é a *spec augmentation* (Park et al., 2019), que, apesar de ter "*augmentation*" no nome, não aumenta os dados, mas faz edições nos dados já existentes.

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

"tambem". Por isso, é importante fazer um levantamento desse tipo de erro no *corpus* de treinamento, por exemplo, contrastando a lista de palavras do *corpus* com uma lista-referência da língua para a qual a aplicação está sendo desenvolvida[16].

Depois de levantados os erros, é preciso corrigi-los de alguma forma caso sejam muito frequentes. Isso é muito comum em dados coletados na internet ou que não passaram por um processo rigoroso de transcrição e revisão. Outra forma de lidar com esse problema dos *typos*, caso não se queira investir tempo na limpeza dos dados, é implementar um módulo de pós-processamento que corrige grafias incorretas, mas isso pode trazer desvantagens, como um possível aumento na latência (tempo corrente entre a fala do usuário e o reconhecimento do texto, crucial em aplicações como a dos assistentes de voz).

Finalmente, talvez seja necessário normalizar o texto antes do treinamento[17]. As técnicas de normalização são as mesmas utilizadas em processamento de texto (Capítulo 4), por isso não vamos repeti-las aqui. Vale apenas dizer que atualmente existem modelos de reconhecimento de fala *end-to-end*, isto é, que têm como entrada o texto não normalizado, minimamente manipulado, e como saída, a transcrição também já normalizada inversamente, da forma exata como deve aparecer para o usuário. No entanto, para se obter uma acurácia boa em modelos *end-to-end*, é necessária uma quantidade muito grande de dados, o que é inviável de se obter para muitos pesquisadores e empresas, por isso não se deve descartar a normalização.

2.2.2.3 Pré-processamento do áudio

Depois da limpeza do texto, é preciso "limpar" os áudios. Áudios distorcidos[18] devem ser removidos e também aqueles cuja duração é muito discrepante da duração da maioria. Mais uma vez, isso só é necessário caso o número de áudios *outliers* seja muito grande. Um caso ou outro não vai atrapalhar a aprendizado. Por fim, os áudios e a transcrição devem ser segmentados e alinhados de alguma forma, caso já não estejam assim. Essa segmentação e alinhamento são importantes para garantir que a rede possa aprender a partir de dados que sejam os mais específicos e corretos possíveis.

Conforme mencionado anteriormente, o reconhecimento de fala é feito atualmente por meio de redes neurais, mas, qualquer que seja a arquitetura utilizada (veremos as principais na próxima seção), a primeira etapa envolve processamento de sinais. O primeiro passo é sempre a conversão do sinal analógico para digital. A isso se segue a extração de informações do sinal, que serão os elementos de entrada para a rede neural (combinados ao texto)[19].

2.2.2.3.1 Conversão analógico-digital

Como explicado na Seção 2.2.1, o sinal acústico da fala nada mais é que o resultado da vibração das pregas vocais pela passagem do ar. O ar que respiramos passa pelas cordas

[16]Vale lembrar que um sistema de reconhecimento de fala é dependente do idioma, isto é, um modelo bem treinado em dados do português não conseguirá fazer um bom reconhecimento do francês. O mais próximo que se pode chegar de modelos independentes de língua seria por meio de *transfer learning*, que é uma técnica de aprendizado de máquina que "aproveita" um treinamento prévio como ponto de partida para um treinamento com outros dados. No *transfer learning*, os pesos da rede neural não são iniciados em 1, mas já ajustados com base no treinamento anterior.

[17]Há dois tipos de normalização de texto, a direta, conhecida apenas como "normalização", e a inversa, conhecida como "ITN (*inverse text normalization*)", da qual falaremos mais adiante.

[18]Nos referimos a arquivos de áudio em que ocorre distorção do sinal e perceptualmente ouve-se um chiado alto que compromete a compreensão da fala.

[19]Exceto no modelo wav2vec (Baevski et al., 2020), que será descrito mais adiante.

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

vocais e causa sua vibração, gerando ondas sonoras, que passam pela faringe e laringe até atingir a cavidade bucal. Nela, as ressonâncias geradas pela vibração das pregas encontram obstáculos e são por eles modificadas e, finalmente, liberadas com a abertura da boca (e pelo nariz, no caso de nasais), quando falamos. Os "obstáculos" mencionados são as diferentes posições que os nossos articuladores assumem[20]. Dessa forma, o nosso aparato vocálico atua como um filtro para as frequências originais emitidas pela glote, e o que ouvimos é o que passou pelo filtro. Essas frequências filtradas são captadas por microfones como ondas analógicas, que precisam ser digitalizadas para serem processadas por um sistema de reconhecimento de fala.

A conversão do sinal envolve dois processos: a **amostragem** e a **quantização**[21]. A amostragem é a seleção das amostras de amplitude do sinal acústico que serão medidas para se representar digitalmente a onda. Deve-se selecionar, no mínimo, duas amostras por ciclo, uma correspondente ao pico, e a outra, ao vale da onda. O número de amostras por segundo corresponde à taxa de amostragem. Qualquer que seja a taxa de amostragem, a máxima frequência registrada em 1 segundo será sempre a metade do número de amostras em 1 segundo, uma vez que uma repetição da onda deverá ter, pelo menos, duas medições para ser minimamente registrada. A taxa de amostragem para gravações de fala deve ser de no mínimo 20 kHz (vinte mil medições por segundo), uma vez que a maioria das informações relevantes para a fala estão abaixo de 10 kHz.

A quantização é a representação desses valores de amplitude em inteiros pelo computador. As representações mais comuns para um sinal acústico são de 8 ou 16 bits. Quanto maior o número de bits que podem ser alocados para representar uma medição de amplitude, melhor será a representação digital da onda, uma vez que mais pontos de amplitude poderão ser armazenados.

2.2.2.3.2 Janelamento

Pelo fato de ser gerado de maneira irregular (vibrações da glote), o sinal de fala é um sinal não-estacionário, isto é, não mantém suas propriedades constantes por mais de 100 ms. No entanto, entre 5 e 100 ms, as propriedades se mantêm relativamente constantes, e o sinal se assemelha a um sinal estacionário[22]. Por isso, para representar um sinal com duração de vários segundos ou até minutos, utiliza-se o método de janelamento[23]. Esse método consiste na fragmentação do sinal em pequenas janelas de tempo de modo que o início da próxima janela ocorra cerca de alguns milissegundos após o início da anterior[24]. Para que não haja cortes abruptos na representação da amplitude do sinal entre uma janela e outra, costuma-se aplicar a função Hamming em cada janela. Essa função aproxima de zero os valores de amplitude nas extremidades das janelas.

2.2.2.3.3 Extração de informações das frequências do sinal

Uma vez separado em janelas, é preciso extrair as informações das frequências do sinal digital, pois é nas frequências que residem os correlatos dos fones (a informação que nos

[20] Em linguística, a vogal [i], por exemplo, costuma ser descrita como "anterior alta não-arredondada". Cada um desses adjetivos se refere a um aspecto da articulação do [i]. "Anterior" e "alta" se referem ao posicionamento da língua, e "não-arredondada", à configuração dos lábios.

[21] Recomendamos a leitura de (Johnson, 2011) para saber mais.

[22] Para saber mais, consulte (Rabiner; Juang, 1993).

[23] Para representar um sinal estacionário, como o ruído branco, bastaria apenas uma janela, uma vez que esse sinal é constante ao longo do tempo.

[24] Segundo Fayek (2016), costuma-se usar intervalos de 20 a 40 ms para a janela, e 10 ms de deslocamento (50-60% de sobreposição entre duas janelas consecutivas).

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

permite identificar diferentes fones)[25]. São informações de frequência e pressão que servirão de entrada para a modelagem da fala. Há mais de um método de extração dessas informações, mas o mais comum atualmente é a Transformada Discreta de Fourier (DFT), computado pelo algoritmo FFT (*Fast Fourier Transform*). Esse método é aplicado a cada janela, tendo como entrada a amplitude do sinal em um dado intervalo de tempo, e, como saída, informações de frequência e pressão para cada janela.

Depois de extraídas, as informações das frequências do sinal são convertidas para a escala mel (Stevens, 1937), uma escala de frequência baseada na percepção humana do sinal acústico. Nosso ouvido é mais sensível a mudanças sutis de amplitude nas frequências mais baixas e menos sensível a mudanças nas frequências mais altas. Dessa forma, a escala mel agrupa as frequências com base em filtros logaritmicamente diferentes, isto é, as frequências mais baixas possuem mais agrupamentos de menos faixas de frequência, cada, e as frequências mais altas possuem menos agrupamentos com mais faixas de frequência em cada um. Assim, as frequências mais baixas são representadas em mais detalhes do que as mais altas.

As janelas de sinal digitalizado e representado na forma de frequências na escala mel são transformadas em vetores, que servirão de entrada para a rede neural de reconhecimento de fala, como veremos adiante.

2.2.2.4 Modelos de reconhecimento

O problema de reconhecimento de fala é um problema de classificação de sequências. A entrada é um sinal contínuo, o sinal acústico, que deve ser primeiro filtrado para que a fala seja separada do ruído[26], e digitalizado. Assim, o sinal é transformado em uma sequência de unidades discretas, como vimos na seção anterior. Essa sequência de unidades será classificada como outra sequência, que será a saída do processo. A sequência de saída é, na maioria dos casos, palavras.

No caso da conversão de fala em texto, a diferença de tamanho entre a sequência de entrada da rede neural, vetores com *features* acústicas, e a de saída, palavras, costuma ser muito grande. Lembre-se de que o áudio foi digitalizado e, com a extração das informações de frequência, vetorizado. Cada vetor corresponde a uma janela de 10 ms, como vimos na Seção 2.2.2.3.2, então, para uma sentença de 10 s, com 5 palavras, teríamos 100 vetores. Para minimizar essa discrepância, realiza-se um *subamostragem*, processo de redução do número de vetores do *input*.

Até alguns anos atrás, empregavam-se modelos estatísticos híbridos para resolver o problema do reconhecimento de fala. As arquiteturas utilizadas continham módulos que eram treinados de maneira independente. Os módulos eram o modelo acústico (AM), o modelo de língua (LM) e um modelo lexical com um dicionário de pronúncias. Os modelos conhecidos como HMM (*Hidden Markov Model*) foram amplamente utilizados com relativo sucesso nas tarefas de ASR. No entanto, essas arquiteturas trabalhavam com modelos de linguagem baseados em n-gramas[27] e assumiam independência entre as probabilidades de ocorrência dos fones, e, por isso, não eram eficazes em processar informações de longa distância[28]. Hoje, as arquiteturas do tipo *encoder-decoder* são as mais utilizadas em ASR.

[25] Lembre-se das aulas de fonética acústica e dos formantes que caracterizam cada vogal. Para saber mais, consulte (Johnson, 2011).

[26] Atualmente, as redes neurais conseguem aprender qual parte de um áudio contém fala.

[27] N-gramas representam contextos muito restritos, na maioria das vezes.

[28] Outra desvantagem dos modelos híbridos é o fato de o LM precisar alocar muita memória (100 GB) para que se obtenham bons resultados no modelo. Isso se torna proibitivo em se tratando de ASR

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

Os modelos HMM que geravam melhores resultados eram baseados numa arquitetura de máquina de estados finitos, em que cada estado corresponde a uma parte de um fone. Por exemplo, para o fone [a], gerava-se um HMM com três estados: o primeiro representando o início do fone [a], o segundo representando a parte mais estável do fone, e o último, o final do fone. Dessa forma, os modelos eram treinados para todos os fones da língua. Para tratar o problema mencionado anteriormente de ausência de contexto, treinava-se modelos com grupos de três fones seguidos (trifones). Os melhores modelos eram agrupados no módulo do modelo acústico. A saída do modelo acústico, por sua vez, era interpolada com um dicionário de pronúncias. O último passo era a combinação da saída do módulo lexical com um modelo de língua, que continha n-gramas e suas probabilidades de ocorrência. O Quadro 2.2 demonstra esse processo:

Quadro 2.2: Modelos de reconhecimento

AM + Léxico	**LM**
ingressos para a próxima seção ->	ingressos para a próxima **sessão**

Na primeira coluna do Quadro 2.2, temos a saída do modelo acústico e do léxico, que é "ingressos para a próxima seção". Note-se que a palavra "seção", um dos homófonos conhecidos do português, não está escrita da maneira correta. Nesse contexto, a grafia correta seria "sessão". A adição de um modelo de língua, treinado com uma quantidade suficiente de dados, é capaz de acertar a grafia correta para esse contexto, com base nas relações entre as palavras.

Como o treinamento do modelo acústico HMM era baseado nos fones, era necessário balancear os dados de treinamento foneticamente. Isto é, a distribuição dos fones nos dados deveria refletir a sua proporção na língua falada[29]. A consoante [l], por exemplo, um dos fones mais frequentes do português brasileiro, deveria ocorrer mais vezes nos dados de treinamento do que sua parente [lh], menos comum.

Uma arquitetura parecida com as híbridas, chamada CTC (*Connectionist Temporal Classification*), configura a forma mais simples de executar a tarefa de reconhecimento de fala. Assim como os modelos acústicos dos modelos híbridos, o CTC atribui *labels* (classes, dentre as possíveis letras do alfabeto) a cada *frame* de atributos acústicos e depois elimina as letras duplicadas seguidas uma da outra. O principal problema do CTC é a sua "falta de memória", isto é, ele considera todas as saídas independentes umas das outras, e computa a saída para um instante t apenas com base na entrada desse mesmo instante t.

Mais recentemente, começou-se a empregar redes neurais recorrentes na tarefa de ASR. Basicamente, essas redes, chamadas de RNN, tinham a vantagem de armazenar informação desde o início da sequência, ou no nosso caso, da sentença, configurando uma forma de "memória"[30]. A computação dentro de uma unidade da rede leva em consideração a saída da unidade da etapa anterior bem como a saída do próprio neurônio na etapa atual. As RNN-T (T de *Transducer*) são a combinação do CTC, enquanto modelo acústico, com um

embarcados em dispositivos.

[29] Alcaim et al. (1992) descrevem uma metodologia para gerar listas foneticamente balanceadas. Um procedimento semelhante, baseado no cálculo do qui quadrado, pode ser aplicado ao balanceamento de *corpora*.

[30] RNN têm a desvantagem dos *vanishing gradients* e precisam ser combinadas com técnicas de *gating*, como LSTM e GRUs. Na prática, RNNs são séries de LSTMs.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

predictor que faria as vezes de modelo de língua e reavaliaria a saída do CTC, gerando uma nova saída, levando em consideração o contexto.

Outra opção muito usada são os Transformers com *self-attention*. De forma resumida, diferentemente das RNN, nos Transformers, os vetores de entrada e de saída têm o mesmo tamanho e cada bloco de atenção tem acesso às entradas dos blocos anteriores. Assim, cada entrada é comparada com as demais para que a saída mais provável seja gerada. Os Transformers são eficazes em modelar contextos mais distantes, mas menos eficazes em contextos de curta distância.

Atualmente, tanto RNN-T quanto Transformers são técnicas bastante utilizadas em ASR. No entanto, alguns estudos mais recentes apontam outras soluções como ainda melhores. Gulati et al. (2020) mostram resultados competitivos com o uso de *Conformers*, arquitetura que une as redes convolucionais (CNN) com os Transformers (daí o nome "*conformer*"). Na combinação CNN + Transformers, as limitações de ambas arquiteturas são suavizadas, porque o que é deficiente em uma é o ponto forte da outra. Os Transformers são melhores em contextos mais globais, e as CNN, em contextos mais locais.

Nas arquiteturas de *encoder-decoder*, o "*encoding*" pode assumir diferentes unidades, como fones, sílabas ou grafemas. No entanto, os resultados mais competitivos em ASR utilizam *wordpieces* como as menores unidades codificadas. *Wordpieces*, ou *subwords*, são exatamente o que os nomes indicam: partes de palavras (Capítulo 4). Mas não devem ser confundidos com morfemas! Diferentemente dos morfemas, as *wordpieces* não carregam nenhum significado necessariamente[31]. Elas podem ser geradas de maneira empírica por diferentes algoritmos (WordPieceModel, *byte pair encoding* (BPE) e outros) e constituem um vocabulário induzido a partir de dados de texto. A segmentação das palavras da língua em unidades menores é, de certa forma, arbitrária (sua geração envolve etapas "*greedy*"), embora se baseie na frequência com que essas unidades aparecem no *corpus*. Por exemplo, em um *corpus* formado apenas por sentenças com verbos no infinitivo, é de se esperar que um vocabulário induzido a partir dele contenha alguma *wordpiece* que termine em "-ar", como **tar_** (o "*underscore*" após a *string* representa final de palavra). Dessa forma, caso o modelo se depare com o neologismo "deletar", considerando que ele não esteve presente no *corpus* de treinamento, o modelo conseguirá gerá-lo concatenando a *wordpiece* "tar_" com outras *wordpieces* (talvez "de_", de "deixar, derrubar", "le_" de "ler, levar", e "tar_").

A abordagem de *wordpieces* como unidade de modelagem se mostrou melhor do que a de grafemas no que diz respeito especialmente às palavras OOV (*out-of-vocabulary*), como neologismos, nomes próprios, palavras estrangeiras e termos da moda. Nos modelos híbridos, os *frames* acústicos eram mapeados para fones e depois era necessária uma interpolação com um dicionário de pronúncia para gerar as palavras. Nos modelos *end-to-end*, em que se busca eliminar essa última etapa, *wordpieces* têm gerado resultados melhores pelo fato de trazerem em si uma espécie de contexto. Na maioria das línguas, incluindo o português, um grafema isolado pode ser associado a mais de uma pronúncia, como é o caso de "r" ("rato" e "caro"). Ao contrário, o fone [h] de "rato" não ocorrerá na *wordpiece* "_ro". Os grafemas e o léxico de pronúncia funcionam bem para palavras conhecidas da língua, mas deixam a desejar quando se deparam com palavras que não estão no dicionário.

Mais recentemente, em 2019, uma arquitetura bastante promissora foi proposta pela Facebook AI, o *encoder* wav2vec[32]. Baseado no word2vec (Capítulo 10) do processamento de texto, a ideia do wav2vec é obter representações vetoriais diretamente a partir do áudio

[31] Apenas no caso de uma *wordpiece* coincidir com um morfema, ela intrinsecamente terá significado, mas isso não é levado em consideração pelo modelo.

[32] https://ai.facebook.com/blog/wav2vec-state-of-the-art-speech-recognition-through-self-supervision/.

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

puro, isto é, eliminando a etapa de extração de atributos acústicos e a necessidade de se treinar com áudios transcritos. Por meio de duas redes convolucionais sucessivas, o modelo transforma áudio digitalizado em vetores e aprende distinguindo trechos reais de áudio de trechos modificados por ele mesmo. O wav2vec é uma arquitetura de aprendizado autossupervisionada (*self-supervised learning*) que aprende a predizer trechos de áudio. Esse modelo depois pode ser combinado com outras redes neurais usadas em ASR. A grande vantagem dessa abordagem é que ela resolve o principal problema da tarefa de reconhecimento de fala: a falta de dados de áudio e texto, especialmente para *low resource languages*, para as quais a oferta de dados é baixíssima ou até mesmo inexistente. Mesmo para línguas como o inglês, bem representado em termos de dados para processamento de fala, o wav2vec é bastante eficiente, porque precisa de 100 vezes menos horas de áudio de treinamento do que as arquiteturas *end-to-end* que vimos acima (Baevski et al., 2020).

2.2.2.5 Etapas adicionais

Devido à escassez de dados de fala anotados disponíveis e à necessidade que os modelos *end-to-end* têm de muitos dados, várias técnicas vêm sendo experimentadas para que seja possível contornar essa questão. Uma técnica bastante conhecida é o *shallow fusion* (Williams et al., 2018). Nessa técnica, um modelo de língua, treinado a partir de *corpora* de textos, é adicionado ao *pipeline* de treinamento. Esse LM externo, como é chamado, é eficaz em completar sequências de palavras, então sua contribuição se dá na reavaliação de dado segmento para um *chunk* mais provável. Suponhamos que o modelo de ASR treinado com áudio e texto tenha gerado a seguinte saída "essas ideias como-as com sebo". O recálculo da hipótese pelo LM externo provavelmente chegaria em "essas ideias como as concebo", que é um trecho mais provável de ocorrer, dada a semântica das palavras envolvidas.

Há muitas outras técnicas de aprendizado de máquina que podem ser usadas, e combinadas, para aprimorar o resultado de um sistema de reconhecimento de fala. Há quem recorra à síntese de áudio para resolver o problema da falta de dados, por exemplo.

2.2.2.6 Pós-processamento

Uma última etapa do *pipeline* de um sistema de ASR costuma ser a normalização inversa, mais conhecida como ITN (*Inverse Text Normalization*)[33]. O que ocorre nessa etapa é a conversão de *strings* que foram transcritas da forma como foram faladas (domínio falado) em símbolos (domínio escrito). Isso se aplica principalmente a números, unidades de medida, moedas, contas de matemática, números romanos, tudo que envolve uma simbologia diferente de "palavra". Normalmente, a hipótese gerada pelo ASR não contém caracteres numéricos nem outros símbolos, como "ª" ou "&", então, para que um número de telefone, por exemplo, seja mais legível para o usuário final da aplicação, se faz a normalização inversa. Dessa forma, um número como "noventa e nove nove nove nove nove nove nove nove nove", formato gerado na hipótese, pode ser convertido para (99) 9999-9999. O mesmo costuma acontecer com quantias monetárias, como "dois milhões e quinhentos mil reais", que pode ser transformado em "R$ 2.500.000" ou "R$ 2500000", a depender da convenção adotada.

[33] Isso depende bastante de qual aplicação se está desenvolvendo. Em se tratando de aplicações para fins comerciais, é extremamente comum haver a normalização inversa, por questões de experiência do usuário.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Os módulos de ITN podem ser feitos por meio de regras escritas por especialistas ou podem ser redes neurais. Recentemente, começou-se a migrar para os ITNs neurais, como indica o artigo da Amazon AWS AI de 2021 (Sunkara et al., 2021). Um ITN baseado em regras funciona segundo um modelo de transdutor de estados finitos (FST), semelhante à máquina de estados finitos mencionada anteriormente na explicação dos HMMs.

2.2.2.7 Métricas de avaliação

A acurácia de um modelo de ASR costuma ser medida em termos de taxa de erro de palavras e de sentenças. As métricas mais utilizadas são a *Word Error Rate* (WER) e a *Sentence Error Rate* (SER). A WER é calculada com base na soma de deleções, substituições e inserções dividida pelo total de palavras da referência e multiplicada por 100. Veja o exemplo abaixo, retirado da base de teste do LibriVox:

> **Referência:** A virtude é comunicável. O vício é contagioso. Os governos fracos fazem fortes os ambiciosos e insurgentes. Atividade sem juízo é mais ruinosa que a preguiça.
> **Hipótese :** A virtude é comunicável. O vício é contagioso. Os governos fracos fazem fortes os ambiciosos e insurgentes. **Atividades** sem juízo é mais **ruidosa** que a preguiça.
> **Avaliação:** S S

O trecho da referência é a transcrição manual do áudio, e o trecho da hipótese é a saída gerada por um sistema de ASR. Os segmentos sublinhados são aqueles cujo reconhecimento automático errou. Enquanto a referência era "atividade", no singular, a hipótese gerada foi "atividades", no plural; enquanto a referência era "ruinosa", a hipótese foi "ruidosa". Esses são exemplos de erros de substituição e a WER desse trecho é dada por 2/26 * 100 = 7,69%, em que 2 é a soma das substituições e 26 é o total de palavras do trecho.

Em geral, calcula-se um valor único de WER, para um dado conjunto de teste, para se avaliar o desempenho de um modelo. Atualmente, os melhores modelos atingem um valor de WER inferior a 5% sem técnicas de *fine-tuning* e *shallow fusion*.

A métrica SER é referente à computação do número de sentenças com pelo menos um erro. Portanto, para um conjunto de teste com 100 sentenças, das quais dez apresentaram um ou mais erros de inserção, deleção ou substituição, a taxa de SER será de 10%. Por ser mais detalhada e dar uma ideia melhor do desempenho de um modelo, a WER costuma ser mais utilizada do que a SER. A SER é indicada para casos em que se queira medir o desempenho de um normalizador inverso, por exemplo, em que o número de *tokens* de uma sentença não normalizada para uma normalizada não nos diz muito. Por exemplo, a sentença "Você me deve cinco reais", quando normalizada inversamente, gera "Você me deve R$5,00", a depender da convenção adotada pelo ITN. Digamos que a saída de um ITN para essa sentença seja "Você me deve R$ 5,00". Se computarmos o WER, obteremos 2/4 * 100 = 50%. Nesse caso, o WER não nos diria muito sobre a eficácia do ITN. Por isso é mais interessante computarmos a SER e sabermos qual a porcentagem de sentenças do conjunto de teste que apresentaram algum erro de normalização.

Como bem apontou Jurafsky; Martin (2023), talvez fosse interessante criar uma métrica que levasse em consideração a relevância das palavras da sentença, atribuindo um peso maior às palavras mais relevantes, que são, em geral, palavras de conteúdo, como verbos e nomes (Capítulo 4). Por exemplo, uma sentença como "Mande um beijo para a Juliana" reconhecida por um ASR como "Mande um beijo pra Juliana" seria muito menos

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

problemática para todos os efeitos do que uma saída como “Mande um beijo para a Júlia”. Embora o WER da segunda sentença (16,6%) seja menor do que o da primeira (50%), a primeira hipótese é muito mais fiel ao conteúdo da sentença. Em muitas aplicações, o ASR é o primeiro passo de um *pipeline* de PLN que envolve a atribuição da sentença a uma intenção do falante e depois realiza uma ação. Nesse caso, enviar um beijo para a pessoa errada pode ter sérias consequências.

2.2.2.8 Desafios em reconhecimento de fala

Mesmo quando um modelo atinge uma acurácia de quase cem por cento de acerto no reconhecimento das palavras, há ainda alguns erros bastante difíceis de corrigir. Os casos que apresentamos aqui valem para o português brasileiro. É possível que se apliquem a outras línguas em situações parecidas, mas o que será apresentado se baseia nas observações com relação ao português do Brasil. Esses problemas estão relacionados aos artigos “a” e “o”, vogais átonas, na maioria das vezes, quando ocorrem no fim de uma palavra seguidas da mesma vogal também em posição átona, como no Exemplo 2.1.

Exemplo 2.1:

> Mande um beijo **para a** Amanda

Quando falamos espontaneamente, ou até mesmo numa fala colaborativa, cujo “interlocutor” é um assistente virtual, situação em que tendemos a falar de um modo mais monitorado e articulado, as vogais em sequência são pronunciadas de forma contínua, numa mesma corrente de ar. Não costumamos fazer pausas (chamadas de *glottal stops*) entre uma vogal e outra nessas situações. No Exemplo 2.1, o [a] final de “para” se junta ao [a] do artigo “a” e ambos podem ser interpretados pelos modelos como sendo apenas um único fone [a], como ilustrado em Exemplo 2.2.

Exemplo 2.2:

> Mande um beijo par**a a** Amanda

Embora a diferença de duração entre um caso e outro seja de apenas alguns milissegundos, nem sempre o modelo consegue fazer a segmentação correta. Dessa forma, é possível que um modelo reconheça “Mande um beijo **para** Amanda” em vez do esperado. Isso não quer dizer que os modelos nunca irão acertar o trecho “para a”. Como mostrado nas seções anteriores, há outros fatores que não apenas a correspondência grafema-fone em jogo no reconhecimento de fala (por exemplo, a distribuição das palavras na língua dada pelo LM).

Exemplo 2.3:

> Quero instalar o WhatsApp

Algo semelhante poderia acontecer com Exemplo 2.3, em que os modelos podem ter dificuldade em reconhecer o artigo “o” pelo fato de a vogal [o] átona ser bastante próxima em qualidade da semivogal de “wa” em “WhatsApp” e de ambas serem produzidas em coarticulação. É possível que uma saída para a transcrição automática dessa sentença fosse “Quero instalar WhatsApp”.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Esses dois exemplos têm outro ponto em comum: ambas as possibilidades são bastante banais e frequentes na língua. Tanto "para" quanto "para a" são formas muito usadas em qualquer contexto. O mesmo vale para "instalar WhatsApp" e "instalar o WhatsApp". As duas formas são muito comuns. Isso dificulta a resolução do problema por meio de uma interpolação com um modelo de língua, por exemplo, uma vez que as formas com e sem artigo provavelmente serão bem próximas em probabilidade de ocorrência.

Outro caso de semelhança fonética que confunde um modelo de ASR é o par "no/do" (e suas variações). Pelo fato de as duas preposições poderem ocorrer nos mesmos contextos e ainda serem formadas de apenas dois fones muito parecidos, a sua distinção não é trivial para o modelo. Desse modo, uma sentença como "vou buscar um trabalho **na** escola" pode facilmente ser reconhecida como "vou buscar um trabalho **da** escola". É claro que isso depende também do quão articulada a fala é e também da qualidade do áudio, e da presença ou ausência de ruído.

Todos os casos relatados nesta seção não constituem, a priori, erros graves de reconhecimento de fala, uma vez que não alteram o significado das sentenças em questão de maneira drástica. Apesar disso, como dito na Seção 2.2.2.7, a principal métrica utilizada na avaliação de um modelo de ASR não faz nenhum tipo de discriminação entre as palavras, e considera todas de igual peso. Embora a princípio um pouco injusta, essa prática se explica pelo fato de que seria necessário algum trabalho etiquetador para identificar as palavras relevantes nas sentenças. Talvez a classificação binária entre palavras de conteúdo versus palavras gramaticais (Capítulo 4) não fosse suficiente para todos os casos. Poderia haver, por exemplo, algum caso em que "na" e "da" trouxessem uma distinção decisiva de significado. Talvez por isso ainda seja mais viável manter todas as palavras com o mesmo status durante a avaliação.

2.2.3 Síntese de fala

Síntese de fala é o processo de conversão de texto ortográfico para áudio. Nos sistemas de conversão texto-fala ocorre um mapeamento de sequências de letras para formas de ondas sonoras.

Comumente utilizado por softwares de acessibilidade, módulos de atendimento automático e assistentes virtuais, os sistemas de conversão texto-fala têm suas unidades acústicas segmentadas e concatenadas conforme informações de transcrição fonética do texto que se deseja sintetizar, transformando então aquela sentença em sinal acústico.

Um TTS (do inglês, *text-to-speech*) pode ser dividido em duas etapas: a primeira, chamada de análise do texto, onde o texto de entrada é normalizado e transcrito da forma ortográfica para a fonológica; e a segunda, síntese do sinal, onde ocorre a concatenação das unidades fonológicas e a inserção da prosódia. Vamos detalhar cada uma destas etapas a seguir.

2.2.3.1 Análise do texto

Na etapa de Análise do texto o objetivo é decodificar o texto de entrada e prepará-lo para ser convertido em áudio. Essa etapa, também conhecida como pré-processamento, pode ser dividida em outras duas tarefas: a normalização, que expande o texto de entrada para a sua forma literal; e a segunda, que converte o texto já expandido para fonemas, ou representações de pronúncia, e o entrega para a etapa seguinte.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

2.2.3.1.1 Normalização

Ao receber o texto a ser sintetizado o sistema de TTS, nesse primeiro estágio, a tarefa é normalizar a sentença de entrada. Nesta etapa normalizar significa substituir elementos do texto como números e abreviaturas, por palavras ou sequência de palavras escritas por extenso. Exemplos são apresentados no Quadro 2.3.

Quadro 2.3: Exemplos de normalização

Texto de entrada	Texto normalizado
1990	mil novecentos e noventa
68,3%	sessenta e oito vírgula três por cento
Av.	avenida
km²	quilômetros quadrados
2:45PM	Duas e quarenta e cinco da tarde

Algumas classes de normalização têm mais problemas do que outras. As siglas, por exemplo, podem ser lidas letra por letra, como "OMS", ou como uma única palavra, no caso dos acrônimos, como em "USP", ou ainda serem expandidas como em "SP – São Paulo". No português ainda temos o caso do gênero gramatical para casos como dos algarismos 1 e 2, que podem ser expandidos como um/uma e dois/duas, a depender da palavra que vem a seguir. Exemplos são apresentados no Quadro 2.4.

Quadro 2.4: Exemplos de normalização para algarismos

Texto de entrada	Texto expandido
1 mesa	uma mesa
1 copo	um copo
2 dias	dois dias

Para a etapa de normalização, a divisão por categoria pode tornar a organização do trabalho mais fácil e/ou intuitiva. Assim, dá-se conta de casos a serem normalizados por categorias ou classes de fenômenos linguísticos de acordo com suas ocorrências na língua. Algumas categorias de normalização no português brasileiro são apresentadas no Quadro 2.5.

Quadro 2.5: Categorias de normalização no português brasileiro

Categoria	Exemplos
Abreviaturas	"Av." "ltda." "Sra."
Siglas	"SP" "USP" "IR"
Números cardinais	"102" "1.500" "18,20"
Números ordinais	"2º" "3ª"
Datas	"08/07/2014" "1º de maio de 1.886"
Valores monetários	"R$ 32,50" "US$ 150,00" "R$ 149,6 bilhões"

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Links/URLs	"www.google.com"
Porcentagem	"12,8%" "21,3%"
Unidades de medida	"3cm" "8,515 mi Km²" "100km/h" 45

A tarefa de normalização do texto pode ser feita com a utilização de duas diferentes técnicas: (1) É possível optar por desenvolvê-la por meio de regras: muitas vezes utilizando-se de expressões regulares, tais regras são descritas de modo a analisar o texto *token* a *token* e buscar padrões compatíveis no texto. Uma vez que um padrão do texto dá match com uma regra descrita, a regra cuida de substituir o *token* em questão por seu correspondente por extenso. Modelos de TTS mais robustos contam com sistemas como o Kestral de (Ebden; Sproat, 2014) que também é baseado em regras, mas primeiro classifica e analisa cada entrada do texto e depois produz um novo texto usando uma gramática de verbalização. O normalizador desenvolvido com base em regras tem a vantagem de não depender de dados de treinamento anotados, mas as regras podem se tornar complexas e frágeis, além de carecer de escritores especializados para mantê-las.

Há também normalizadores baseados em redes neurais (2) chamados de modelo codificador-decodificador, que demonstram melhor funcionamento se comparados aos normalizadores baseados em regras, mas que demandam grandes conjuntos de dados anotados.

Além das etapas aqui apresentadas, a síntese de fala ainda passa pela etapa de conversão grafema-fonema, treinamento da voz e validação do modelo treinado. A etapa de conversão grafema-fonema para o português brasileiro é comumente realizada com uso de regras descritas de modo a mapear as letras do alfabeto para o som correspondente a ela, de acordo com o contexto em que tal letra aparece. Já os treinamentos do modelo de voz, por muitos anos feitos por meio de métodos estatísticos (*Hidden Markov Models* – HMMs), hoje são comumente realizados com o uso de redes neurais, método conhecido como Tacotron2 integrado à LPCnet. A avaliação de qualidade e acurácia desses modelos é feita por meio de uma medida numérica baseada na opinião pessoal de humanos, o Mean Opinion Score (MOS) é uma classificação de qualidade de voz. O teste consiste em humanos falantes nativos do idioma ouvirem e atribuírem uma nota entre 1 (ruim) e 5 (excelente) para áudios sintetizados a partir do modelo a ser avaliado. A média das notas atribuídas aos áudios sintéticos passam a ser a nota da avaliação do modelo. Ainda muito dependentes da impressão dos avaliadores humanos, a acurácia dos modelos assim treinados ainda não é mensurada numericamente, ou seja, com avaliações automáticas e objetivas, o que torna a validação das tecnologias hoje empregadas na área bastante dependentes das percepções dos avaliadores.

2.2.4 Considerações finais

Neste capítulo, vimos um pouco sobre a história do processamento de fala, sobre as características da língua falada e sobre as principais tarefas da área de processamento de fala, que são o reconhecimento automático e a síntese de fala. Esperamos ter conseguido demonstrar no que o processamento de fala difere do processamento de texto e quais são os seus principais desafios. De maneira semelhante ao que ocorre no processamento de texto, há carência de dados de qualidade para o processamento do português brasileiro em comparação com o cenário do processamento do inglês. Atualmente, os modelos de reconhecimento de fala *end-to-end*, que são o estado da arte, necessitam de uma quantidade muito grande de dados para que seja obtida uma qualidade de ponta. Os modelos de

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

síntese, por sua vez, necessitam de menos horas de fala, porém a qualidade das gravações precisa ser impecável e há a necessidade de se gravar a mesma pessoa, o que aponta para um custo elevado, tanto financeiro quanto de tempo.

Conforme demonstrado na Seção 2.2.2.1, em se tratando de ASR, é necessário considerar variações dialetais, tanto de pronúncia quanto de vocabulário e sintaxe, durante o treinamento dos modelos. Os dados precisam ser suficientemente variados e representativos de cada variedade a fim de que um sistema genérico o bastante para dada língua seja desenvolvido. Isso não ocorre no processamento de texto nas mesmas proporções. Especialmente quando comparamos o português europeu com o brasileiro no que diz respeito ao reconhecimento, e também à síntese de fala, por serem variedades muito diversas, especialmente foneticamente, seria preciso construir sistemas de ASR separados para processar as duas línguas. No processamento de texto, diferentemente, pelo fato de a língua escrita ser mais conservadora, as duas variedades se aproximam, embora cada uma continue tendo suas peculiaridades de grafia, vocabulário e sintaxe. O impacto da distância entre as variedades se torna mais evidente na síntese de fala, uma vez que um sistema desenvolvido para o português europeu não seria bem aceito por falantes brasileiros residentes no Brasil. Basta pensar no quão estranho seria utilizar um assistente virtual que falasse português europeu. Apesar de todas essas considerações, vemos despontar, nos últimos meses, modelos de reconhecimento e de síntese de fala treinados com várias línguas. Os estudos de Yang et al. (2023), Pratap et al. (2020a) e Saeki et al. (2023) explicam como essas técnicas funcionam. Esse tópico é bastante interessante e será objeto de uma próxima edição deste capítulo.

Além do reconhecimento e da síntese de fala, há várias outras tarefas na área de processamento de fala. Podemos elencar aqui as seguintes: clonagem de voz, detecção de palavras-chave, identificação de falantes, diarização da fala, entre outras. O Capítulo 3 deste livro tratará de recursos para o desenvolvimento dessas e de outras tarefas do processamento de fala e também apresentará uma breve descrição de cada uma.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Capítulo 3

Recursos para o processamento de fala

Edresson Casanova
Vinícius G. Santos
Flaviane R. Fernandes Svartman
Marli Quadros Leite
Arnaldo Candido Jr.
Ricardo M. Marcacini
Solange O. Rezende
Sandra Maria Aluísio

Publicado em: 26/09/2023
Atualizado em: 16/04/2026

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

3.1 Introdução

Até a metade de 2020, o português brasileiro (PB) possuía apenas algumas dezenas de horas de dados de fala públicos ou abertos para pesquisas acadêmicas, disponíveis para treinar modelos para os sistemas mais comuns, que são os reconhecedores automáticos de fala (em inglês, *Automatic Speech Recognition* ou ASR) e os sintetizadores de fala (em inglês, *Text-to-Speech Synthesis* ou TTS). Havia um grande contraste com a língua inglesa, cujos recursos eram **maiores tanto em número de horas quanto em número de locutores** e, assim, mais adequados à aplicação de métodos de aprendizado profundo de máquina, chamados de *deep learning*, em inglês.

Para o treinamento de modelos de reconhecimento de fala, havia aproximadamente 60 horas, divididas em quatro pequenos conjuntos de dados de fala lida (em inglês, *read speech*), isto é, uma fala preparada para ser lida, em contraste com a fala espontânea: (1) o Common Voice Corpus versão 5.1 (da Mozilla)[1] (2) o *dataset* Sid, (3) o VoxForge e (4) o LapsBM 1.4[2]. Para o treinamento de modelos de síntese de fala, havia um conjunto de dados de um único locutor com 10 horas e 28 minutos de fala, chamado TTS-Portuguese Corpus[3].

A fala espontânea possui fenômenos que tornam o seu reconhecimento mais complexo do que o da fala lida, como as pausas preenchidas e as disfluências de edição. Exemplos de projetos que tratam da fala lida são o Librivox[4], que distribui os livros de domínio público em formato de áudio. Estes áudios têm sido usados em vários projetos para criação de recursos para processamento de fala em inglês como o LibriSpeech ASR Corpus[5] e

[1] https://commonvoice.mozilla.org/pt/
[2] O *dataset* Sid, o VoxForge e o LapsBM 1.4 estão disponíveis em: https://igormq.github.io/datasets/
[3] https://github.com/Edresson/TTS-Portuguese-Corpus
[4] https://librivox.org/pages/about-librivox/
[5] https://www.openslr.org/12

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

o LibriTTS[6], ambos alocados no repositório *Open Speech and Language Resources*. O LibriSpeech é um grande *corpus* de fala lida em inglês, com 1.000 horas, destinado a pesquisas de reconhecimento automático de fala. O LibriTTS é um *corpus* multilocutor derivado do LibriSpeech para pesquisas em síntese de fala, com 585 horas.

Nesse cenário de escassez de dados públicos de fala em PB para treinamento de sistemas de processamento de fala, foi concebido, em agosto de 2020, o projeto TaRSila[7] do Center for Artificial Intelligence[8] da Universidade de São Paulo, financiado pela IBM e FAPESP. O projeto TaRSila visa a aumentar os conjuntos de dados de fala em PB tanto para treinamento de sistemas como também para pesquisas linguísticas nas seguintes **tarefas** do processamento de fala:

1. reconhecimento automático de fala espontânea, que transcreve automaticamente a fala espontânea, por exemplo, diálogos, entrevistas e conversas informais;
2. síntese de fala multilocutor expressiva, que gera vozes de diferentes locutores/falantes de maneira próxima a um falante humano, a partir de um texto;
3. clonagem de voz. A clonagem de voz engloba duas grandes tarefas do processamento de fala: a síntese de fala e a conversão de voz. O objetivo da clonagem é copiar a voz de um locutor e gerar novas amostras de áudio utilizando-se de características da voz do locutor. Existem diferentes métodos de clonagem, sendo os mais interessantes os de síntese de fala multilocutor *zero-shot* e os métodos de conversão de voz *zero-shot* que conseguem clonar a voz de um locutor utilizando apenas alguns segundos de fala;
4. modelagem de tópicos a partir das transcrições dos áudios, que é útil para organizar, resumir e visualizar o conteúdo de vídeos em tópicos similares;
5. predição da pontuação de segmentos de fala e predicação da capitalização, isto é, quais palavras devem ser escritas com a primeira letra maiúscula. Essas tarefas são importantes para facilitar o entendimento humano do texto de uma transcrição automática. Também são importantes para o encadeamento de sistemas, quando o ASR é usado antes de um sistema de tradução automática ou de um sistema de reconhecimento de entidades nomeadas (em inglês, *Name Entity Recognition* ou NER). Exemplos de entidades nomeadas são nomes de pessoas, de lugares, e de organizações, entre outras;
6. segmentação prosódica da fala espontânea em unidades prosódicas terminais e não terminais, que transmitem a ideia de conclusão do enunciado e a ideia de não conclusão, respectivamente. Ajuda a analisar o conteúdo de uma sequência falada, dados os vários sentidos possíveis, e facilita a criação de conjuntos de dados para treinamento de ASRs para fala espontânea, por indicar o limite de uma sequência de fala e, consequentemente, auxiliar na predição da pontuação (tarefa citada acima);
7. reconhecimento de emoções a partir da fala (em inglês, *Speech Emotion Recognition* ou SER). Reconhece o estado emocional de um locutor a partir de sua fala e é útil para muitas aplicações, como o desenvolvimento de ferramentas de diagnóstico para terapeutas, assistentes de voz e ferramentas para análise de comunicações em call centers.

Além das sete tarefas acima em estudo no TaRSila, o livro sobre Processamento de

[6] https://www.openslr.org/60/
[7] https://sites.google.com/view/tarsila-c4ai/
[8] http://c4ai.inova.usp.br/

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Fala[9] (Bäckström et al., 2022) apresenta outras tarefas típicas, como o reconhecimento e verificação de locutor, a restauração de fala e a diarização:

- reconhecimento de locutor e verificação de locutor, que se referem, respectivamente, à identificação do locutor (quem está falando?) e à verificação se o locutor é quem afirma ser;
- a restauração (ou aprimoramento) da fala refere-se à melhoria de uma gravação de um sinal de fala para, por exemplo, remover o ruído de fundo ou o efeito da acústica do ambiente;
- a diarização da fala é o processo de segmentar uma conversa de vários falantes em segmentos contínuos de um único falante.

O livro coloca o reconhecimento de emoções em áudios (citada acima) no grupo das tarefas de análise paralinguística, dado que extrai do sinal de áudio informação não linguística (diferente da pontuação, por exemplo) e não relacionada à identidade de um locutor, como fazem as tarefas de reconhecimento e verificação de locutor.

Neste capítulo, apresentamos os recursos de fala criados nos três primeiros anos do projeto TaRSila para ilustrar várias das tarefas da área de processamento de fala, acima elencadas, que são definidas e exemplificadas em cada seção. Nesse percurso, fazemos um contraste com a língua inglesa que possui mais recursos para cada tarefa, citando os recursos disponibilizados na literatura tanto para o inglês como para o português.

Os vários recursos desenvolvidos no TaRSila têm o prefixo CORAA (CORpus de Áudios Anotados), que é um grande *corpus* multipropósito do português brasileiro no qual os arquivos de áudios estão alinhados com transcrições que foram (ou estão sendo) manualmente validadas para cada tarefa estudada no TaRSila.

O alinhamento de um trecho de áudio com a transcrição correspondente indica o tempo de início e o tempo final do trecho (também chamado de minutagem do áudio) (veja a Figura 3.1), formando pares usados no aprendizado supervisionado de um modelo de reconhecimento de fala. O reconhecedor Whisper[10] da OpenAI, por exemplo, foi treinado dessa forma. O uso de pares áudio-transcrição para treinamento de reconhecedores de fala não é a única abordagem para a tarefa, que também pode ser feita por aprendizado não supervisionado, como é o caso, por exemplo, do wav2vec-U[11]. Essa é uma abordagem na qual o aprendizado dispensa a necessidade de transcrições, e ocorre apenas por meio de áudio. Em todo caso, para a avaliação do desempenho de um reconhecedor, é importante que haja os pares áudio-transcrição.

Figura 3.1: Excerto do inquérito SP_D2_255 do NURC-SP com dois trechos (text) e indicação do tempo de início (xmin) e tempo final (xmax) em segundos no formato TextGrid. A anotação "..." refere-se a uma pausa silenciosa.

```
intervals [1]:
    xmin = 0
    xmax = 0.3575589499895925
    text = "..."
intervals [2]:
    xmin = 0.3575589499895925
    xmax = 5.103648343063277
    text = "bem nós gostaríamos de começar esta nossa conversa falando sobre transportes e viagens"
```

[9] https://speechprocessingbook.aalto.fi/.
[10] https://github.com/openai/whisper/
[11] https://github.com/facebookresearch/fairseq/tree/main/examples/wav2vec/unsupervised

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

Começamos apresentando, na Seção 3.2 (*corpora* para TTS), o TTS-Portuguese Corpus, *corpus* para treinamento de modelos de síntese de fala, criado e disponibilizado no início de 2020 com a fala de um único locutor. Esse *corpus* permitiu avançar pesquisas sobre síntese de fala, conversão de voz e uma abordagem de aumento de dados para treinar modelos de reconhecedores de fala em cenários de baixos recursos de dados. Na Seção 3.3 (com enfoque em anotação prosódica), apresentamos o *corpus* CORAA NURC-SP (foco na versão mínima), que contém 334 horas de fala espontânea e fala preparada de falantes de São Paulo, capital, divididas em uma parte com áudios e transcrições manuais não alinhadas originalmente (47 inquéritos) e outra parte de áudios somente (328 inquéritos). Na Seção 3.4 (*corpora* para ASR), apresentamos os *corpora* CORAA ASR, CORAA Certas Palavras, CORAA MuPe e CORAA NURC-SP Audio Corpus (com enfoque em ASR), uma coleção de *corpora* com enfoque em fala espontânea que juntos totalizam aproximadamente 963 horas. Na Seção 3.5 (recursos para reconhecimento de emoções), apresentamos o CORAA SER versão 1.0, composto por aproximadamente 50 minutos de segmentos de áudio rotulados em três classes: neutro, não neutro feminino e não neutro masculino, sendo que a classe neutra representa segmentos de áudio sem estado emocional bem definido e as classes não neutras representam segmentos associados a um dos estados emocionais primários da fala do locutor. Finalmente, na Seção 3.6 (predição e pontuação), revisitamos o *corpus* do CORAA MuPe, com 300 horas de áudios de histórias de vida e transcrições com pontuação, que foi cedido ao TaRSila em um convênio entre o Instituto de Ciências Matemáticas e de Computação da Universidade de São Paulo (ICMC-USP), Universidade Federal de Goiás (UFG) e Museu da Pessoa, usado para a avaliação da tarefa de predição de pontuação do ASR Whisper da OpenAI. Na Seção 3.7 finalizamos o capítulo com a revisão dos tópicos cobertos aqui.

3.2 Recursos para síntese de fala

Os sistemas de síntese de fala receberam muita atenção a partir de 2017 devido ao grande avanço proporcionado pela aplicação do *deep learning* (Goodfellow et al., 2016) nessa área, que permitiu a popularização e aprimoramento de assistentes virtuais, como Apple Siri (Gruber, 2009), Amazon Alexa (Purington et al., 2017) e Google Home (Dempsey, 2017).

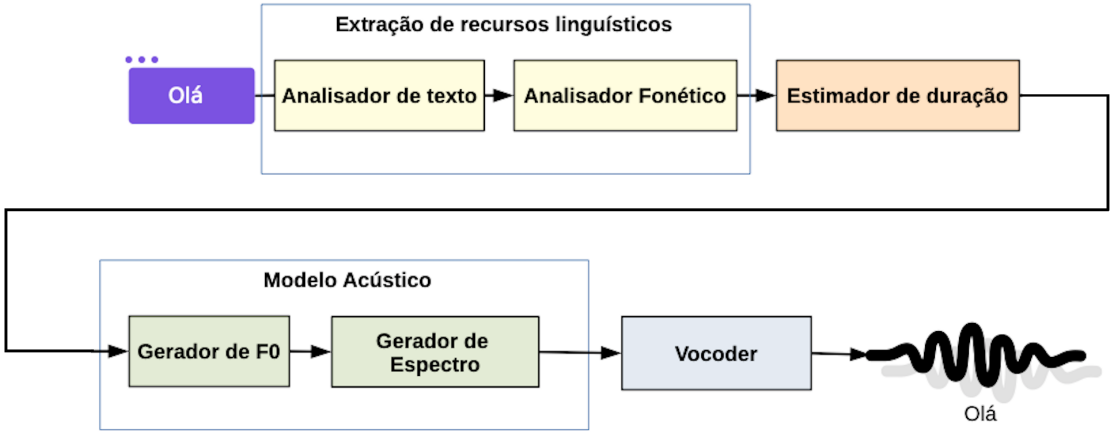
Os sistemas tradicionais de síntese de fala foram muito utilizados até 2010. Entretanto, a partir dessa data, a síntese de fala baseada em redes neurais tornou-se gradualmente o método dominante e alcançou uma qualidade de áudio superior aos sistemas tradicionais (Tan et al., 2021). De acordo com Tachibana et al. (2017), os sistemas tradicionais de síntese de fala não são fáceis de se desenvolver, devido ao fato de serem compostos por muitos módulos específicos, tais como um analisador de texto, um analisador fonético, um estimador de duração, um modelo acústico e um *vocoder*. A Figura 3.2 apresenta um diagrama com os principais componentes de um sistema de síntese de fala tradicional (Casanova et al., 2021). Vários trabalhos na literatura exploraram essa abordagem clássica (Braude et al., 2013; Charpentier; Stella, 1986; Klatt, 1980; Siddhi et al., 2017; Teixeira et al., 2003; Tokuda et al., 2000; Wang; Georgila, 2011; Ze et al., 2013).

O advento do *deep learning* permitiu a integração dos módulos específicos dos sistemas de síntese de fala tradicionais em um único modelo. Apesar dos modelos baseados em *deep learning* serem às vezes criticados devido à dificuldade de interpretá-los, vários sistemas de síntese de fala baseado em *deep learning* (Kim et al., 2020, 2021; Kyle et al., 2017; Ping et al., 2017; Shen et al., 2018; Tachibana et al., 2017; Valle et al., 2020; Wang et al., 2023a; Wang et al., 2017) demonstraram a capacidade de sintetizar fala com uma qualidade muito

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Figura 3.2: Os principais componentes de um sistema de síntese de fala tradicional.



Fonte: Adaptado de (Casanova et al., 2021, fig. 14.1, p. 184)

promissora, superior, inclusive, à dos sistemas tradicionais.

Modelos baseados em *deep learning* requerem uma quantidade maior de dados para treinamento, portanto, idiomas com poucos recursos disponíveis ficam prejudicados. Por esse motivo, a maioria dos modelos de síntese de fala atuais são projetados para o inglês (Kim et al., 2020, 2021; Ping et al., 2017; Shen et al., 2018; Valle et al., 2020; Wang et al., 2023a), que é um idioma com muitos recursos disponíveis publicamente.

Para o inglês existem vários *corpora* que podem ser utilizados para treinar modelos de síntese de fala baseados em *deep learning*, por exemplo, os *corpora* VCTK (Veaux et al., 2017), LJ Speech (Ito, 2017), LibriTTS (Zen et al., 2019) e LibriTTS-R (Koizumi et al., 2023a).

O *corpus* VCTK (Veaux et al., 2017) é composto por 44 horas de fala de 108 locutores, sendo 61 do sexo feminino e 47 do sexo masculino. Além disso, o *corpus* inclui amostras de 11 variedades linguísticas do inglês, sendo elas: britânico, americano, canadense, neozelandês, sul-africano, australiano, escocês, norte-irlandês, irlandês, indiano e galês. A taxa de amostragem dos áudios presentes nesse *corpus* é de 48 kHz. O *corpus* LJ Speech (Ito, 2017) foi derivado de audiolivros e tem aproximadamente 24 horas de fala de uma locutora profissional em uma taxa amostragem de 24 kHz. LJ Speech é um dos *corpora* abertos mais populares para síntese de fala de um único locutor. O *corpus* LibriTTS (Zen et al., 2019) também foi derivado de audiolivros e possui 585 horas de fala de 2456 locutores, sendo 1185 do sexo feminino e 1271 do sexo masculino. A taxa de amostragem dos áudios presentes nesse *corpus* é de 24 kHz. Por outro lado, o *corpus* LibriTTS-R (Koizumi et al., 2023a) foi criado com a aplicação do modelo de restauração de fala (em inglês, *speech restoration*) Miipher (Koizumi et al., 2023b) no *dataset* LibriTTS. As amostras do LibriTTS-R são idênticas às do LibriTTS, com apenas a qualidade de som melhorada. Os resultados dos experimentos mostraram que os modelos de síntese de fala treinados com o LibriTTS-R apresentaram qualidade significativamente melhor em comparação com os modelos treinados no LibriTTS.

Para a língua portuguesa, até meados em 2019, não havia nenhum *corpus* disponível publicamente com quantidade de horas e qualidade de áudio suficientes para treinar modelos de síntese de fala baseados em *deep learning*. Embora um *corpus* para síntese de fala em português europeu tenha sido disponibilizado em 2001, pelo fato de ele ter duração de

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

aproximadamente 100 minutos apenas (Teixeira et al., 2001), não é possível utilizá-lo no treinamento de modelos baseados em *deep learning*.

Para suprir essa carência de dados para síntese de fala no português brasileiro, em 2019, a coleta do *corpus* TTS-Portuguese Corpus (Casanova, 2019; Casanova et al., 2022, 2022) foi iniciada (Casanova, 2019). Posteriormente, em 2020, o *corpus* foi tornado público (Casanova et al., 2022) e os detalhes de sua compilação foram publicados em um artigo (Casanova et al., 2022).

Para a construção do TTS-Portuguese Corpus, foram utilizados textos de domínio público. Inicialmente, buscando alcançar um vocabulário amplo, extraíram-se todos os artigos das seções de destaques da Wikipédia (da época em que foi compilado) para todas as áreas do conhecimento. Após essa extração, os artigos foram segmentados em sentenças (considerando-se a pontuação textual). Durante as gravações, o locutor recebeu sentenças desse conjunto, que foram escolhidas de forma aleatória. Além disso, foram utilizados os 20 conjuntos de sentenças foneticamente balanceadas, cada conjunto contendo 10 sentenças, propostas por Seara (1994). Por fim, para aumentar o número de perguntas e introduzir um discurso mais expressivo, foram ainda utilizadas frases do Chatterbot-corpus[12], um *corpus* criado originalmente para a construção de *chatbots*. Desse modo, o TTS-Portuguese Corpus possui um vocabulário amplo com palavras de diversas áreas. Além disso, também possui uma representação de fala expressiva com o uso de perguntas e respostas de um conjunto de dados de *chatbot*.

A gravação do TTS-Portuguese Corpus foi realizada por um locutor masculino, nativo do português brasileiro, não profissional, em ambiente silencioso, mas sem isolamento acústico devido às dificuldades de acesso a estúdio de gravação. Todos os áudios foram gravados com frequência de amostragem de 48 kHz e resolução de 32 bits. No *corpus*, cada arquivo de áudio possui sua respectiva transcrição textual (a transcrição fonética não foi fornecida). O TTS-Portuguese Corpus consiste em um total de 71358 palavras faladas, 13311 palavras únicas, resultando em 3632 arquivos de áudio e totalizando 10 horas e 28 minutos de fala. Os arquivos de áudio variam em duração de 0.67 a 50.08 segundos (Casanova et al., 2022).

Em paralelo com o TTS-Portuguese Corpus, foram lançados dois conjuntos de dados para reconhecimento automático de fala do português, com boa qualidade. O primeiro, o CETUC (Alencar; Alcaim, 2008), disponibilizado publicamente por Quintanilha et al. (2020), tem aproximadamente 145 horas de fala de 100 locutores. Nesse *corpus*, cada locutor pronunciou mil sentenças foneticamente balanceadas extraídas de textos jornalísticos; em média, 1,45 horas gravadas por locutor. Já o segundo, o *corpus* Multilingual LibriSpeech (MLS) (Pratap et al., 2020b), é derivado dos audiolivros LibriVox e abrange 8 idiomas, incluindo o português. Para o português, os autores disponibilizaram aproximadamente 130 horas de fala provenientes de 54 locutores, uma média de 2.40 horas de fala por locutor. Embora a qualidade de ambos os *corpora* seja boa, os dois foram disponibilizados com uma taxa de amostragem de 16 kHz e não possuem pontuação em seus textos, dificultando a aplicação desses *corpora* para síntese de fala. Além disso, a quantidade de fala de cada locutor nos dois *corpora* é baixa, o que torna difícil criar um conjunto de dados com um vocabulário grande o suficiente para síntese de fala de um único locutor.

Além disso, mais recentemente, o *corpus* CML-TTS (Oliveira et al., 2023) foi proposto. O CML-TTS é baseado no *corpus* Multilingual LibriSpeech (MLS) e foi adaptado para treinamento de modelos de síntese de fala. O CML-TTS é composto por audiolivros em sete

[12] https://github.com/gunthercox/chatterbot-corpus/

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

idiomas: holandês, francês, alemão, italiano, português, polonês e espanhol. Os autores recriaram o *corpus* MLS mantendo a pontuação e os áudios com uma taxa de amostragem de 24 kHz. Amostras que não atendiam às especificações descritas anteriormente foram descartadas. Para o português, após a filtragem, os autores obtiveram aproximadamente 69 horas de fala, provenientes de 31 homens e 17 mulheres.

A Tabela 3.1 apresenta as estatísticas dos recursos disponíveis para síntese de fala do inglês e do português.

Tabela 3.1: Estatísticas dos recursos disponíveis para síntese de fala nas línguas inglesa e portuguesa.

Corpora	Idioma	Duração	Falantes	Disponibilizado
VCTK	inglês	44 horas	108	2017
LJ Speech	inglês	24 horas	1	2017
LibriTTS	inglês	585 horas	2456	2019
LibriTTS-R	inglês	585 horas	2456	2023
TTS-Portuguese Corpus	português	10.4 horas	1	2019
CETUC	português	145 horas	100	2020
MLS	português	130 horas	54	2020
CML-TTS	português	69 horas	48	2023

3.3 Recursos para segmentação prosódica

Existem vários estudos na literatura de processamento de fala com foco na detecção de fronteiras prosódicas nas línguas naturais (Ananthakrishnan; Narayanan, 2008; Huang et al., 2008; Jeon; Liu, 2009; e.g. Wightman; Ostendorf, 1991). Para o inglês, entre os recursos frequentemente utilizados em aplicações que consideram fronteiras prosódicas, podemos citar o Santa Barbara Corpus of Spoken American English (Du Bois et al., 2000--2005) e o Boston University Radio Speech Corpus (Ostendorf et al., 1995). O primeiro contém ≈20 horas de fala espontânea de gêneros variados, transcritas e segmentadas manualmente em unidades entoacionais final e não final (Du Bois et al., 1992). Já o segundo contém 10 horas de notícias de rádio, das quais 3,5 horas estão prosodicamente anotadas de acordo com o sistema ToBI (Beckman et al., 2005).

Para o português brasileiro, trabalhos desenvolvidos no âmbito do projeto C-ORAL–Brasil avançam os estudos para a detecção automática de fronteiras prosódicas na fala espontânea a partir de parâmetros fonético-acústicos e fronteiras identificadas perceptualmente por anotadores treinados (Raso et al., 2020; Teixeira et al., 2018; Teixeira, 2022; Teixeira; Mittman, 2018). Os estudos utilizam excertos de fala monológica masculina (8–24 minutos de áudio e 1339–3697 palavras), provenientes dos *corpora* anotados C-ORAL–Brasil I e II (Mello et al., no prelo; Raso; Mello, 2012a). No âmbito do projeto TaRSila, o CORAA NURC-SP, que é preparado tanto para viabilizar estudos linguísticos quanto o processamento computacional, possui pelo menos 40 horas prosodicamente anotadas. O *corpus* é dividido em três subcorpus, incluindo: (a) CORAA NURC-SP Audio Corpus, focado em ASR; (b) CORAA NURC-SP Mínimo, voltado para estudos prosódicos; e (c) CORAA NURC-SP CATNA, áudios com transcrição não alinhada. Mais detalhes podem ser encontrados no página oficial do projeto[13].

[13] http://tarsila.icmc.usp.br:8080/nurc/home

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

O CORAA NURC-SP toma como base dados provenientes do projeto acadêmico NURC–Norma Urbana Linguística Culta, que foi iniciado em 1969 com o objetivo de documentar e estudar a língua portuguesa falada por pessoas com ensino superior completo, denominadas 'cultas', de cinco capitais brasileiras: Recife, Salvador, Rio de Janeiro, São Paulo e Porto Alegre. O projeto resultou num grande *corpus* (aprox. 1.570 horas, 2.356 falantes) reunido ao longo dos anos 1970 e 1980 (Castilho, 1990).

Como em todas as capitais, o NURC-São Paulo (NURC-SP)[14] reúne mais de 300 horas de gravação, apresentando falantes com nível superior; nascidos e criados na cidade; filhos de falantes nativos de português; igualmente divididos em homens e mulheres; e distribuídos em três faixas etárias (25–35, 36–55 e 56 anos em diante). As gravações foram realizadas em três situações, gerando diferentes gêneros discursivos: palestras/aulas em contexto formal proferidas por um locutor (EF); diálogos entre documentadores e um participante (DID); e diálogos entre dois participantes mediados por documentadores (D2).

O *corpus* do projeto NURC tem sido amplamente utilizado para estudar vários aspectos da língua falada, tendo se tornado um dos *corpora* mais influentes da linguística brasileira. A maioria dos estudos deriva de transcrições de pequenos *subcorpora* compartilhados por pesquisadores que trabalham em cada capital (Castilho, 1990, 2021), aqui referidos como *corpus* mínimo. Assim, a contraparte de áudio era normalmente desconsiderada devido à dificuldade de acesso às fitas magnéticas de rolo nas quais as gravações foram feitas. Recentemente, um protocolo para digitalizar, anotar, armazenar e divulgar o material do acervo do NURC-Recife, o NURC Digital (Oliviera Jr., 2016), foi desenvolvido e completamente implementado. Inspirados nesse protocolo, desenvolvemos, no âmbito do projeto TaRSila, um processo para o alinhamento texto-fala do Corpus Mínimo do NURC-SP.

Embora os procedimentos que orientam o processamento do NURC-SP sejam baseados no protocolo do NURC Digital, eles incorporam sistemas de processamento de fala que incluem, por exemplo, um reconhecedor automático de fala atual (Whisper[15]), um alinhador forçado áudio-transcrição baseado em síntese de fala (aeneas[16]) e alinhadores fonéticos automáticos (Batista et al., 2022; Kruse; Barbosa, 2021) usados em conjunto com um método para a segmentação automática de fala baseada em prosódia (Biron et al., 2021).

A versão CORAA do NURC-SP é composta por 375 inquéritos (aprox. 334 horas de gravação), dos quais alguns já tinham transcrições — mas, até então, não alinhadas ao áudio — e a grande maioria é composta apenas de áudio. No âmbito do TaRSila, o NURC-SP foi dividido em três *subcorpora* de trabalho:

- o Corpus Mínimo (21 gravações + transcrições), que está sendo utilizado para avaliar atualmente os métodos de processamento de todo o acervo;
- o Corpus de Áudios e Transcrições Não Alinhados (26 gravações + transcrições), que está sendo segmentado automaticamente em unidades prosódicas pelo método de Biron et al. (2021), adaptado ao português brasileiro, e validado manualmente; e
- o Corpus de Áudios (328 gravações sem transcrição), que vem sendo transcrito automaticamente pelo ASR Whisper, modelo treinado em 680 mil horas de dados multilíngues coletados da web. Embora outros modelos de reconhecedores automáticos tenham sido avaliados em uma amostra representativa do NURC-SP Corpus Mínimo (Gris et al., 2022), o modelo Whisper foi escolhido, pois, além de realizar

[14] https://nurc.fflch.usp.br/
[15] https://github.com/openai/whisper/
[16] https://www.readbeyond.it/aeneas/

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

uma transcrição de qualidade, consegue colocar pontuações, facilitando a leitura da transcrição automática. As transcrições estão sendo manualmente validadas para corrigir erros do reconhecedor automático.

Entre esses conjuntos de dados, o Corpus Mínimo é o conjunto que se encontra completamente processado (Santos et al., 2022). Ele está disponível publicamente no repositório do Portulan Clarin[17] e compreende 21 arquivos de áudio e transcrições multiníveis (≈18 horas, ≈155 mil palavras) alinhadas ao áudio de acordo com unidades prosódicas linguisticamente motivadas abrangendo os três gêneros textuais especificados anteriormente (DID, EF, D2). O conceito de unidade prosódica que utilizamos aqui está fundamentado nos princípios do método de segmentação prosódica do C-ORAL–BRASIL (Raso; Mello, 2012a). Portanto, no fluxo da fala, podemos reconhecer fronteiras de unidades com valores terminais ou não terminais. **Quebras prosódicas terminais** (TB, *terminal break*) marcam sequências terminadas, ou seja, comunicam a conclusão do enunciado, formando a menor unidade pragmaticamente autônoma da fala, enquanto **quebras prosódicas não terminais** (NTB, *non-terminal break*) sinalizam uma unidade prosódica não autônoma e cuja informação não está concluída dentro de um mesmo enunciado. A identificação das quebras prosódicas é baseada principalmente na relevância perceptiva (auditiva) das pistas prosódicas, mas também na inspeção visual da síntese do sinal acústico fornecida pelo Praat (Boersma; Weenink, 2023). As principais pistas para uma quebra prosódica no português brasileiro são a inserção de pausas e mudanças relacionadas à frequência fundamental e à duração (Raso et al., 2020; Serra, 2009).

As transcrições multiníveis consistem nas seguintes camadas de intervalo anotadas no programa de análise de fala Praat. Veja a Figura 3.3 para uma ilustração:

- 2 camadas (TB-, NTB-) nas quais a fala de cada locutor (-L1, -L2) e documentador (-DOC1, -DOC2) é segmentada em unidades prosódicas e transcrita de acordo com normas adaptadas do Projeto NURC.
- 1 camada (LA) para a fala transcrita e segmentada de qualquer locutor aleatório.
- 1 camada para comentários (COM) acerca do áudio e da anotação.
- 1 camada contendo a versão normalizada (-NORMAL) da transcrição de todas as camadas TB e LA.
- 1 camada contendo a pontuação (-PONTO) que finaliza cada TB.

O processamento do Corpus Mínimo do NURC-SP envolveu várias etapas. Em primeiro lugar, os anotadores foram treinados no uso do software Praat e na aplicação das diretrizes de anotação. Paralelamente, foram feitos o alinhamento automático entre o áudio e a transcrição original, usando o aeneas, e a preparação dos arquivos de alinhamento para anotação, que inclui uma revisão ortográfica ampla num editor de texto. Em seguida, foram realizados testes de confiabilidade entre avaliadores para avaliar a segmentação prosódica, com um valor de kappa (Capítulo 16) acima de 0.8 como critério. Na sequência, procedeu-se à anotação, que envolveu: (i) a revisão da transcrição original de acordo com as diretrizes adaptadas do projeto NURC, (ii) a correção do alinhamento automático texto-fala e (iii) a segmentação da fala em unidades prosódicas. Após a conclusão da anotação, os arquivos anotados passaram por uma inspeção realizada por um especialista, em busca de desvios significativos das diretrizes de anotação. Em seguida, a ortografia foi verificada e o texto foi normalizado, a fim de tornar o conjunto de dados adequado para o processamento

[17]Disponível em https://hdl.handle.net/21.11129/0000-000F-73CA-C sob a licença CC BY-NC-ND 4.0.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Figura 3.3: Excerto de SP_EF_153 com cinco camadas anotadas no Praat.



de linguagem natural. Por fim, foi realizada a anotação da pontuação seguindo as normas ortográficas do português brasileiro.

A relevância de um *corpus* de português brasileiro processado e anotado prosodicamente está no fato de que a delimitação de fronteiras prosódicas melhora o desempenho de sistemas de processamento de línguas naturais (e.g. Chen; Hasegawa-Johnson, 2004; Lin et al., 2016, 2019; Ludusan et al., 2015; Yang et al., 2011b) e é *input* para a predição de pontuações automáticas. Além disso, é possível usar tal *corpus* como um conjunto de referência para o treinamento de sistemas automáticos de reconhecimento de fala espontânea, detecção de sotaques e *parsing* e, assim, alavancar o desenvolvimento de métodos de processamento de fala do português brasileiro e viabilizar novos estudos linguísticos, dada a sua futura disponibilização integral num portal web que permitirá pesquisas específicas.

3.4 Recursos para reconhecimento automático de fala

A seguir são apresentados grandes *corpora* para a criação de sistemas de reconhecimento de fala voltadas para a língua portuguesa[18]. Observa-se que muitos recursos são multilíngues, porém a Tabela 3.2 detalha especificamente as estatísticas para a língua portuguesa. Entre os recursos apresentados, existe uma preponderância um pouco maior da variedade brasileira nos recursos existentes. Porém, a variedade europeia também é contemplada.

Tabela 3.2: Estatísticas aproximadas dos principais recursos disponíveis para ASR na língua portuguesa.

Corpora	Horas	Áudios	Falantes	Licença	Lançamento
CORAA ASR 1.1	289	402.000	1.700	CC BY-NC-ND	2022
CORAA MuPe	365	317.746	289	CC BY-NC-ND	2025
CORAA NURC-SP Audio Corpus	239	177.224	401	CC BY-NC-ND	2024
CORAA Certas Palavras	70	42.486	190	CC BY-NC	2026
Mozilla Common Voice 24	228	196.195	3.804	CC-0	2025
MultiLingual LibriSpeech	285	40.000	68	CC BY	2020
MultiLingual TeDx Corpus	164	93.000	-	CC BY-NC-ND	2021

[18] Convém observar que alguns valores são estimativas dos respectivos autores. Muitos projetos estão em atividade, e os valores apresentados devem aumentar com passar do tempo.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

CORAA ASR[19] (Candido Junior et al., 2022) é um *corpus* para reconhecimento automático de fala que contém também fala espontânea, um tópico pouco pesquisado em projetos similares. Esse *corpus* faz parte do *corpus* multi-tarefa CORAA e está inserido no projeto TaRSila. O CORAA ASR é a junção de cinco projetos independentes: (1) ALIP (Gonçalves, 2019); (2) C-ORAL–Brasil I (Raso; Mello, 2012a); (3) NURC-Recife (Oliviera Jr., 2016); (4) SP-2010 (Mendes; Oushiro, 2012); (5) TeDx Talks. Os quatro primeiros projetos foram originalmente criados para análises linguísticas e adaptados para a tarefa de reconhecimento automático de fala. O último é composto de áudios cedidos pela organização TED (The Eletronic Development) para a tarefa de reconhecimento e não deve ser confundido com o *corpus* oficial TeDx Talks Brazil, detalhado a seguir, pois existem diferenças entre os áudios disponibilizados. A fala espontânea é mais difícil de ser reconhecida do que a fala preparada, mais comum nos outros projetos, devido à presença mais frequente de fenômenos como pausas preenchidas, hesitações e revisões.

O Corpus CORAA MuPe (Evaldo Leal et al., 2025) é um conjunto de 300 horas de histórias de vida que foi cedido ao projeto TaRSila em um convênio de colaboração iniciado em dezembro de 2022 entre o MuPe, o ICMC-USP e a UFG. O objetivo do convênio é o estudo e o desenvolvimento de modelos de ASR, de métodos de segmentação automática de transcrição e modelagem de tópicos baseada nas transcrições de vídeos. Mais detalhes sobre esse *corpus* serão apresentados na Seção 3.6.2.

CORAA NURC-SP Audio Corpus (Lima et al., 2024) é um grande *corpus* de fala espontânea com foco em ASR, o primeiro no sotaque paulistano. Esse *corpus* foi adaptado para a tarefa de ASR a partir dos dados do projeto NURC-SP, conforme apresentado na Seção 3.3. É importante ressaltar que o Audio Corpus não deve ser confundido com outros lançamentos do NURC-SP, como o CATNA e o Corpus Mínimo. O processamento do *corpus* seguiu passos semelhantes ao CORAA ASR, visto que no CORAA ASR são incluídos áudios do NURC-Recife projeto irmão do NURC-SP.

O data CORRA Certas Palavras (Araújo, 2026) contém audios cedidos pelo programa de radio Certas Palavras, que foi ao ar entre as décadas de 1980 e 1990. Esse *corpus* possui dois locutores principais, Jorge Marques de Vasconcellos e Claudiney José Ferreira. Isso faz com que o *corpus* seja ideal para tarefas de síntese de fala baseada em fala espontânea. Porém, devido ao razoável número de entrevistados estimados no período (cerca de 190), optamos por listá-lo juntamente com recursos para reconhecimento de voz, visto que o Certas Palavras pode ser utilizado para diferentes tarefas.

O *corpus* Common Voice[20] (Ardila et al., 2019) é um projeto de uso aberto criado pela Fundação Mozilla, responsável pelo navegador Firefox. O projeto é uma resposta à carência de recursos para várias línguas, incluindo o português. No projeto, os usuários podem ao mesmo tempo contribuir para o crescimento da base e acessar áudios de outras pessoas. A proposta de criação de uma grande base colaborativa segue a mesma linha de outros projetos de sucesso em diferentes áreas de aplicação, tais como a Wikipédia e projetos de código aberto de modo geral. Para colaborar com o projeto, os usuários podem doar áudios em suas próprias vozes e revisar doações de outros usuários. O projeto conta com ferramentas para a coleta, a validação e a internacionalização (adequação a diferentes idiomas). A licença de uso permissiva desse projeto permite a exploração do *corpus* inclusive com fins comerciais. Na versão 13, o *subcorpus* para a língua portuguesa conta com 197 horas de áudios e transcrições, das quais 151 foram validadas.

[19] https://github.com/nilc-nlp/CORAA
[20] https://commonvoice.mozilla.org/pt

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

O *corpus* MultiLingual LibriSpeech[21] (MLS) (Pratap et al., 2020b) foi pensado pelos seus autores tanto para aplicações em síntese quanto em reconhecimento de fala, devido a isso, sendo descrito aqui e na Seção 3.2. Especificamente para a tarefa de reconhecimento, pode ser combinado com outros recursos, visto que possui relativamente poucos falantes (locutores de audiolivros). Cabe aqui comentar que a aplicação de *corpora* para síntese em reconhecimento não é exclusividade do MLS; outros recursos como o *corpus* CETUC (Alencar; Alcaim, 2008) também são relevantes em ASR. Na prática, todos os recursos mencionados na Seção 3.2 podem ser efetivamente usados na tarefa de reconhecimento. Tais recursos são compostos por áudios mais limpos, geralmente em qualidade de estúdio. Por conta disso, modelos construídos unicamente sobre esse tipo de áudio são apropriados apenas para reconhecimento de fala em cenários com pouco ruído. Para contornar essa característica, o projetista pode injetar ruídos nos áudios ou combiná-los com áudios de outros projetos em diferentes níveis de qualidade.

O MultiLingual TeDx Corpus[22] (Salesky et al., 2021) foi proposto para permitir pesquisas nas áreas de reconhecimento automático da fala e tradução da fala para texto[23]. O recurso é composto por palestras sobre os mais variados assuntos, sendo gerenciado no escopo do projeto TEDx, vinculado ao grupo TED (Technology, Entertainment and Design). No caso da língua portuguesa, também existem traduções das transcrições para as línguas inglesa e espanhola. Além disso, áudios em espanhol e francês também contam com traduções para o português.

Existem outras bases para tarefa de ASR que também valem a pena ser citadas. Entre elas, o Multilingual Spoken Corpus[24] (Mazumder et al., 2021) é uma base de palavras faladas em 50 idiomas e contém um recorte de cerca de 1 segundo dos áudios do Common Voice, totalizando 58 horas de áudio em português. Diferentemente das outras bases discutidas até o momento, os áudios desse *corpus* são compostos de palavras soltas, em vez de enunciados completos. Esse tipo de *corpus* se destina ao treinamento de sistemas de reconhecimento em domínios específicos (por exemplo, teleatendimento bancário). Entre as bases menores, pode-se destacar os *corpora* LapsBM, Sidney, VoxForge, três *corpora* que totalizam, aproximadamente, 4, 1 e 1 horas , respectivamente, levantados por Quintanilha et al. (2020) e disponíveis para download na página do pesquisador[25].

Por fim, algumas das bases não são voltadas a ASR, mas a tarefas relacionadas, como tradução de fala para texto. O *corpus* CoVoST (Wang et al., 2020a, 2020b) é um recorte da base Common Voice, mas com foco em tradução de fala para texto. Na versão 2, cerca de 17 horas são disponibilizadas para o português com as respectivas traduções para o inglês. O *dataset* Vox Populi[26] (Wang et al., 2021) é uma iniciativa da empresa Meta com foco principal no treinamento semi-supervisionado e não-supervisionado de modelos de aprendizado de máquina. A base contém transcrições para algumas línguas, mas o português não é contemplado. Ao todo, 17.500 horas de áudio foram disponibilizadas para o idioma.

[21] http://www.openslr.org/94/
[22] http://www.openslr.org/100
[23] Na tradução de fala para texto, o idioma entre o áudio original e a transcrição são diferentes.
[24] https://mlcommons.org/en/multilingual-spoken-words/
[25] https://igormq.github.io/datasets/
[26] https://github.com/facebookresearch/voxpopuli

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

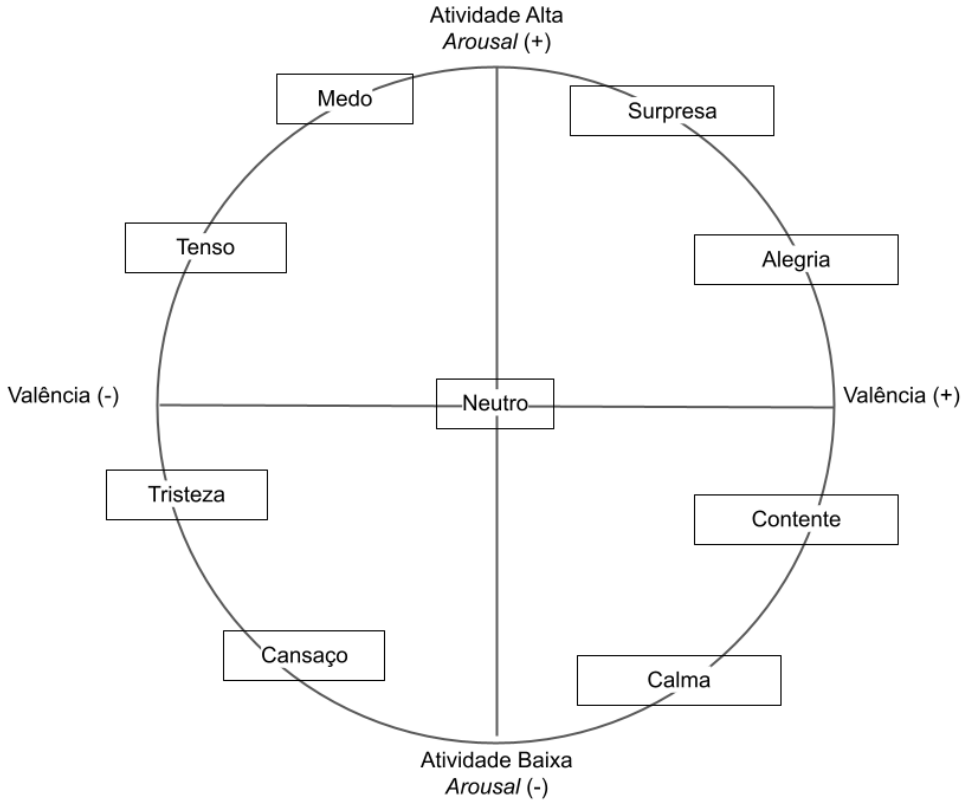
©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

3.5 Recursos para reconhecimento de emoções

O reconhecimento de emoções a partir da fala é uma área de estudo promissora que visa compreender as emoções expressas vocalmente pelos indivíduos (Akçay; Oğuz, 2020; El Ayadi et al., 2011; Singh; Goel, 2022). Uma das teorias mais clássicas nesse campo é a Teoria das Emoções Básicas de Ekman (Ekman, 1992), que descreve a existência de seis emoções primárias: alegria, tristeza, raiva, medo, surpresa e aversão.

Recentemente, outras teorias e modelos têm sido propostos, obtendo-se um espectro mais detalhado de emoções. Nesse sentido, o Modelo Circumplexo de Russel (Posner et al., 2005) oferece uma perspectiva complementar, ao representar as emoções em um espaço bidimensional, com eixos de valência (positivo/negativo) e intensidade (ativa/passiva), conforme apresentado de forma simplificada na Figura 3.4. Reconhecer emoções na fala tem muitas aplicações práticas, como a análise de atendimento ao cliente, apoio na avaliação do estado emocional de indivíduos durante terapias, e o desenvolvimento de assistentes virtuais mais empáticos, o que ajuda a desenvolver técnicas mais eficientes para interação humano-computador (Wani et al., 2021).

Figura 3.4: Modelo Circumplexo de Russell (simplificado) para representar as emoções em um espaço bidimensional, com eixos de valência (positivo/negativo) e intensidade (ativa/passiva).



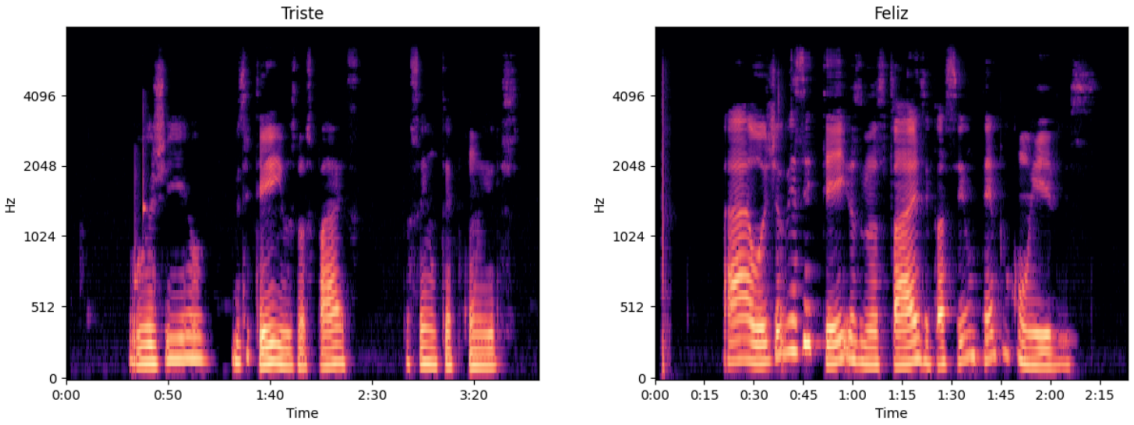
Há três grandes desafios em tarefas de reconhecimento de emoções a partir da fala. O primeiro desafio consiste em representar a fala de forma computacionalmente viável, transformando o sinal acústico em representações que contenham características relevantes

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

para identificar as emoções contidas no sinal. Nesse contexto, uma estratégia tradicional é processar o sinal de áudio para identificar características prosódicas (Rao et al., 2013), como duração e intensidade da fala. Por exemplo, nota-se maior intensidade vocal na emoção "alegria", enquanto a "tristeza" costuma ter intensidade vocal reduzida. Outra estratégia é o reconhecimento de emoções por meio de espectrogramas (Özseven, 2018), uma representação visual do espectro de frequência de um sinal de áudio ao longo do tempo, obtidas por meio da aplicação da transformada de *Fourier* em janelas de áudio. Na Figura 3.5, há a ilustração de dois espectrogramas que representam o sinal acústico do texto "hoje eu visitei os meus pais e passei um tempo com eles", falado por uma mesma pessoa. No lado esquerdo, o trecho foi falado com a emoção "triste". No lado direito, a mesma sentença foi falada com a emoção "alegre". Os modelos de aprendizado de máquina exploram as características extraídas dos espectrogramas para aprender a diferenciar as categorias de emoção.

Figura 3.5: Exemplo de espectrogramas para o trecho "hoje eu visitei os meus pais e passei um tempo com eles". (Esquerdo) Espectrograma da fala triste. (Direito) Espectrograma do mesmo trecho, mas com a emoção de alegria.



O segundo desafio está relacionado à disponibilidade de *corpora* anotados para a tarefa de reconhecimento de emoções. Esses *corpora* são fundamentais para o treinamento de métodos de aprendizado de máquina. Por fim, o terceiro desafio envolve escolha e parametrização do método de aprendizado de máquina visando a geração de modelos eficientes para reconhecimento de emoções.

No âmbito do projeto TaRSila, há uma frente de trabalho denominada SER (*Speech Emotion Recognition*) que visa enfrentar os desafios mencionados anteriormente, com foco específico no reconhecimento de emoções na fala em português. Um diferencial importante deste projeto é o desenvolvimento de abordagens que lidam com fala espontânea, que apresenta desafios adicionais em comparação com a fala preparada. Enquanto a fala preparada envolve cenários planejados ou ensaiados, na qual o indivíduo tem tempo para estruturar suas ideias e escolher suas palavras antes de expressá-las, a fala espontânea ocorre de forma mais imediata, como conversas informais e discussões em grupo, contendo hesitações, pausas, repetições, ruídos e interrupções. Vale ressaltar que a fala espontânea pode expressar emoções de forma mais autêntica, sem ensaios ou autocontrole geralmente presentes na fala preparada.

Uma das etapas cruciais desse projeto foi a preparação do *corpus* CORAA-SER, que consiste em aproximadamente 1 hora de áudio de fala espontânea, anotado com presença

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

ou ausência de emoção, envolvendo homens e mulheres. O *corpus* foi obtido a partir de anotações paralinguísticas de outro *corpus* denominado C-ORAL–BRASIL I, um *corpus* de referência do português brasileiro falado informal (Raso; Mello, 2012b). A primeira versão do CORAA-SER está disponível publicamente[27]. Com o CORAA-SER, já foi possível explorar diferentes técnicas de representação e métodos de aprendizado de máquina para identificar padrões emocionais na fala espontânea em português. Uma visão geral com os resultados de diferentes trabalhos e grupos de pesquisa foram sumarizados por Marcacini et al. (2022).

O CORAA-SER possui segmentos de áudio rotulados em três categorias: neutro (491 áudios), não-neutro-feminino (89 áudios) e não-neutro-masculino (45 áudios). Também são disponibilizadas duas versões pré-processadas dos áudios:

- Características prosódicas: foram disponibilizadas características físicas da fala, como entonação, ritmos, tom, tempo, intensidade etc. Esse tipo de pré-processamento é tradicionalmente utilizado em métodos tradicionais de reconhecimento de emoções. No total, 56 características prosódicas foram disponibilizadas.
- Características do Wav2Vec: foi utilizado um modelo Wav2Vec (Baevski et al., 2020) pré-treinado para extração de características do áudio. Essas características podem ser usadas para treinar um classificador de reconhecimento de emoções na fala.

Entre os resultados mais recentes, incluindo os resultados obtidos no CORAA-SER, vale destacar o desempenho promissor de modelos estado da arte para reconhecimento de emoções na fala, especialmente baseados em técnicas de *deep learning* e *transfer learning* (Chen; Rudnicky, 2023; Gauy; Finger, 2022; Lope; Graña, 2023; Wagner et al., 2023). No contexto do *deep learning*, arquiteturas como redes neurais convolucionais (CNNs), redes neurais recorrentes (RNNs) e Transformers têm sido amplamente aplicadas, devido à sua capacidade de aprender representações intermediárias a partir dos segmentos de áudios para a tarefa de reconhecimento de emoções. Já *transfer learning* é uma abordagem geralmente usada em conjunto com *deep learning* para o reconhecimento de emoções, permitindo utilizar modelos pré-treinados em grandes *corpora* de áudio. Esses modelos pré-treinados são geralmente usados em tarefas de reconhecimento de fala. A ideia é explorar conhecimento prévio adquirido por esses modelos e especializá-lo para uma nova tarefa, como o reconhecimento de emoções. Essa etapa é denominada de ajuste fino e, em geral, depende de um *corpus* anotado, o que aumenta a importância de projetos como o CORAA-SER do TaRSila.

Para finalizar, vale destacar que a tarefa de reconhecimento de emoções a partir da fala ainda possui muitos desafios relacionados à representação computacional da fala, disponibilidade de *corpora* anotados e escolha de métodos de aprendizado de máquina adequados para esta tarefa. No âmbito do projeto TaRSila, a frente de trabalho SER tem buscado superar esses desafios, com ênfase na fala espontânea em português. A criação do *corpus* CORAA-SER foi um passo relevante nesse processo, pois já permitiu a exploração de algumas técnicas pela comunidade (Marcacini et al., 2022). As direções futuras e oportunidades de pesquisa neste tema são promissoras. Muitos pesquisadores estão investigando métodos de *transfer learning* para reconhecimento de emoções, baseado em conhecimento prévio de modelos pré-treinados para fala como o Wav2Vec (Baevski et al., 2020) e HuBERT (Hsu et al., 2021). Essas abordagens têm demonstrado um potencial promissor para melhorar a precisão e a eficiência do reconhecimento de emoções em áudio. Também

[27]CORAA-SER v1: https://github.com/rmarcacini/ser-coraa-pt-br

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

devemos destacar a importância dos trabalhos que ainda exploram características prosódicas, uma vez que relacionar características de duração, intensidade, *pitch* e entonação com diferentes categorias de emoção fornecem maior interpretabilidade no reconhecimento de emoções a partir da fala.

3.6 Recursos para predição de pontuação no cenário de ASR

A saída de sistemas ASR convencionais é uma das principais fontes de dados que requerem capitalização e pontuação, pois é feita de uma sequência de palavras somente. Exemplos de ASR convencionais comerciais são o Google Cloud Speech-to-Text, Microsoft Azure Speech Services, IBM Watson Speech to Text, e SpeechMatics. Quando a saída é um texto escrito para ser lido em voz alta, isto é, um discurso, a tarefa é chamada de restauração da pontuação original, e para a fala conversacional/espontânea a tarefa é chamada de predição da pontuação (Păiş; Tufiş, 2022).

Apresentamos nesta seção, o *dataset* de teste do Corpus CORAA MuPe, balanceado por sexo, com histórias de vida de homens e mulheres, que foi criado para avaliar a tarefa de predição de pontuação no contexto de reconhecedores automáticos de fala, usando como reconhecedor o Whisper da OpenAi (Radford et al., 2022).

Também trazemos um resumo dos trabalhos em predição da pontuação em fala preparada e espontânea, descrevendo suas abordagens e *datasets*. Vamos contrastar primeiramente a diferença entre a saída de ASRs convencionais e de ASRs que além de transcreverem automaticamente um segmento de fala espontânea, conseguem fazer a predição da pontuação e capitalização como é o caso do Whisper.

3.6.1 Ilustrando o uso do ASR Whisper na predição da pontuação em português

O trecho do Quadro 3.1 é de uma história de vida do MuPe[28] que apresenta seis turnos de uma entrevista (P = pergunta, R = resposta) em que os sinais de pontuação foram removidos da transcrição e as primeiras palavras de cada oração são apresentadas em letras minúsculas. Esse trecho ilustra a saída de um ASR convencional que, embora não tenha erros na transcrição de palavras, ajuda a enfatizar o quanto a ausência de pontuação pode dificultar a compreensão do texto quando se torna longo.

Quadro 3.1: Trecho de uma história de vida do MuPe para ilustrar o formato de saída de um ASR convencional

P – tá e sua mãe ela fazia o que
R – a minha mãe ela trabalhou mais de 30 anos numa tecelagem aqui em são paulo chamada guilherme jorge que fica lá na vila formosa ela é auxiliar de tecelagem a vida inteira trabalhou nisso e cuidava da casa
P – mas ela veio de onde onde ela nasceu
R – a minha mãe nasceu na cidade de no sertão de pernambuco chama lá a cidade chama bodocó pernambuco
P – e quando que ela veio para são paulo
R – ela veio para são paulo no final dos anos 40 porque a situação em pernambuco estava no campo estava muito difícil havia seca então ela e três duas irmãs e um irmão vieram pra são paulo início para trabalhar na casa de uma tia dela no bairro da penha que tinha uma

[28] https://museudapessoa.org/historia-detalhe/?id=7136&download_integra_text_pdf/

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

pensão aí depois cada um foi como a maioria dos nordestinos chega fica na casa dos familiares e depois vai arrumando emprego aí vai arrumando sua vida ou seja minha mãe e meu pai também vieram pra são paulo porque lá em minas não havia trabalho e ele como tinha essa vontade de trabalhar acredito eu ele antes de completar 18 anos ele fugiu de casa veio para são paulo e ele era o caçula do primeiro casamento da minha vó

A saída do Quadro 3.2 foi gerada pelo ASR Whisper para o mesmo trecho de áudio relacionado a mesma história de vida do MuPe mostrada no Quadro 3.1. As entidades nomeadas aparecem em negrito, para facilitar a análise. Whisper não é um ASR convencional. Ele foi foi treinado pela empresa de pesquisa em Inteligência Artificial OpenAI usando um grande conjunto de dados multilíngues coletados da web. Ele tem uma nova arquitetura multitarefas, isto é, ele é treinado para prever diversas tarefas de processamento de fala ao mesmo tempo: (i) detecção de atividade de voz, que instrui o modelo a funcionar apenas quando há uma linguagem humana específica e ser robusto ao lidar com ruído/música de fundo; (ii) tradução da fala para o inglês e (iii) reconhecimento de fala multilíngue com pontuação. Das várias pontuações inseridas na transcrição, Whisper não é capaz de gerar ponto e vírgula e dois pontos.

Quadro 3.2: História de vida do MuPe gerada pelo ASR Whisper

P – E sua mãe, ela fazia o quê?
R – A minha mãe trabalhou mais de 30 anos numa tecelagem aqui em **São Paulo**, chamada **Guilherme Jorge**, que fica lá na **Vila Formosa**. Era auxiliar de tecelagem e a vida inteira trabalhou nisso e cuidava da casa.
P – mas ela veio de onde, de onde ela nasceu.
R – A minha mãe nasceu na cidade de um sertão de **Pernambuco**, lá a cidade chama-se **Bodocó**, em **Pernambuco**.
P – E quando ela veio para **São Paulo**...
R – Ela veio para **São Paulo** no final dos anos 40, porque a situação em **Pernambuco** estava no campo, estava muito difícil, havia seca, então ela e as duas irmãs e o irmão vieram para **São Paulo**, início para trabalhar na casa de uma tia dela, no bairro da **Penha**, que tinha uma pensão. Aí depois cada um foi, como a maioria dos nordestinos, chega, fica na casa dos familiares, depois vai arrumando emprego, aí vai arrumando sua vida. Ou seja, minha mãe e meu pai também vieram para **São Paulo**, porque lá em **Minas** não havia trabalho. E ele, como tinha essa vontade de trabalhar, acredito eu, antes de completar 18 anos ele fugiu de casa, veio para **São Paulo**. Ele era o caçula do primeiro casamento da minha avó.

Quando a saída do ASR Whisper é comparada com a transcrição manual pontuada e capitalizada (mostrada no Quadro 3.3; entidades nomeadas em negrito), notamos que o Whisper gera orações mais curtas e, portanto, mais sentenças (11) do que a transcrição manual (6). No entanto, a capitalização utilizada é muito semelhante à manual (ver Tabela 3.3). Quanto à capitalização, ela é usada principalmente para entidades nomeadas relacionadas (EN) a cidades, estados e regiões.

Quadro 3.3: História de vida do MuPe gerada por transcrição manual

P – Tá, e sua mãe, ela fazia o que?
R – A minha mãe, ela trabalhou mais de 30 anos numa tecelagem aqui em **São Paulo**, chamada **Guilherme Jorge**, que fica lá na **Vila Formosa**, ela é auxiliar de tecelagem, a vida inteira trabalhou nisso e cuidava da casa.
P – Mas ela veio de onde? Onde ela nasceu?

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

R – A minha mãe nasceu na cidade de no **Sertão de Pernambuco**, chama lá, a cidade chama **Bodocó**, **Pernambuco**.
P – E quando que ela veio para **São Paulo**?
R – Ela veio para **São Paulo** no final dos anos 40, porque a situação em **Pernambuco** estava no campo, estava muito difícil, havia seca, então, ela e três, duas irmãs e um irmão vieram pra **São Paulo**, início para trabalhar na casa de uma tia dela, no bairro da **Penha**, que tinha uma pensão, aí, depois cada um foi como a maioria dos Nordestinos, chega, fica na casa dos familiares e depois vai arrumando emprego, aí vai arrumando sua vida,ou seja, minha mãe e meu pai também vieram pra **São Paulo**, porque lá em **Minas** não havia trabalho, e ele como tinha essa vontade de trabalhar acredito eu, ele antes de completar 18 anos, ele fugiu de casa, veio para **São Paulo** e ele era o caçula do primeiro casamento da minha vó.

Tabela 3.3: Comparação da pontuação e capitalização da saída do Whisper com uma transcrição manual de um trecho de uma história de vida do MuPe.

	Saída do Whisper	Transcrição Manual
Vírgula	24	31
Reticências	1	0
Ponto Final	9	3
Ponto de Interrogação	1	4
Capitalização no início de sentenças	10	7
Capitalização de EN de Lugares	14	14
Outras Capitalizações	0	1

3.6.2 Descrição do Corpus MuPe e seu *dataset* de teste

O MuPe é um museu virtual que visa contar e preservar as histórias de vida das pessoas e incentiva a participação de pessoas de diferentes idades, sexos, raças e profissões. Fundado em 1991, o MuPe contém atualmente um rico e extenso acervo digital de narrativas de fala espontânea em português, chamadas de histórias de vida que são contadas pelas próprias pessoas ou por terceiros.

As narrativas são gravadas de três formas: (i) na sede do Museu, em estúdio – gravadas em vídeo e coletadas por entrevistadores especializados na metodologia de história de vida, (ii) enviadas via internet pelo Programa Conte sua História ou (iii) via Museu que Anda, programa em que as narrativas de pessoas fora da sede são gravadas em cabines itinerantes. Cada entrevista constitui uma unidade do acervo que é formada pela gravação em áudio ou vídeo da entrevista, a transcrição e edição de cada narrativa, acompanhada de fotos e documentos enviados pelas pessoas que contam suas narrativas de vida.

Depois de gravadas, as histórias de vida coletadas pelo MuPe são transcritas e revisadas. As transcrições possuem anotações de risos, palmas, assobios, fala emocionada, pausas, entre outros, utilizando parênteses. Além disso, as expansões de acrônimos são anotadas usando colchetes. A transcrição é segmentada em enunciados, com pontuação, usando sete sinais de pontuação (ver Tabela 3.5). Os turnos são indicados pelos rótulos P/1 (e P/2) e R seguidos da transcrição do turno, onde P/i (i = 1 ou 2) indica o entrevistador (1 ou 2 entrevistadores) e R o entrevistado. No entanto, como as pausas preenchidas e disfluências de edição (por exemplo, revisões e repetições) comuns na fala espontânea não são anotadas, a transcrição do MuPe pode ser chamada de transcrição textual adaptada.

O *dataset* de teste é composto por 10 narrativas de vida retiradas do projeto Ponto de Cultura da plataforma MuPe. O Corpus MuPe contém 280 narrativas de vida, sendo

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

a maioria delas com conteúdo transcrito completo; algumas poucas apresentam somente um resumo. A Tabela 3.4 mostra as estatísticas do *dataset* de teste, dividido em duas amostras: narrativas masculinas e femininas. O *dataset* de teste do MuPe é composto por 1.349 turnos e totaliza aproximadamente 17 horas. Ele está disponibilizado publicamente[29], com os links dos áudios e as transcrições manuais anonimizadas, cujos nomes completos dos entrevistados (ou dos familiares) masculinos foram trocados por "João" e os femininos por "Maria", para não haver grande perda de material original e ainda atender ao requisito do convênio para disponibilização pública. O nome da família foi trocado por "nome da família".

Tabela 3.4: Estatísticas do *dataset* de teste do MuPe. O tamanho médio do turno e o tamanho médio da sentença são calculados em palavras, sem contar a pontuação. Consideramos como sentença os segmentos terminados em ponto de interrogação, ponto de exclamação e ponto final. O número de *tokens* inclui pontuação.

	Amostra Masculina	Amostra Feminina	Total
Duração do Áudio	8:06:21 h	8:42:13	16:48:34 h
# Turnos	834	515	1.349
Tamanho Médio do Turno	85.26 ± 169.44	138.80 ± 325.46	105.11 ± 240.72
# Sentenças	4.100	4.640	8.740
Tamanho Médio da Sentença	17.16 ± 27.47	14.57 ± 14.64	15.79 ± 21.67
# *Tokens*	83.953	79.377	163.330

Tabela 3.5: Distribuição das classes de pontuação nas amostras do MuPe.

	Amostra Masculina (#)	Amostra Feminina (#)	Total # (%)
Reticências	257	110	367 (1,56%)
Ponto de Exclamação	45	172	217 (0,92%)
Ponto Final	3.247	4.002	7.249 (30,8%)
Ponto de Interrogação	808	466	1.274 (5,41%)
Vírgula	8.383	6.571	14.954 (63,5%)
Ponto e Vírgula	62	21	83 (0,35%)
Dois Pontos	293	358	651 (2,76%)
Total			23.521

3.6.3 *Corpora* usados nos trabalhos de predição da pontuação para ASRs

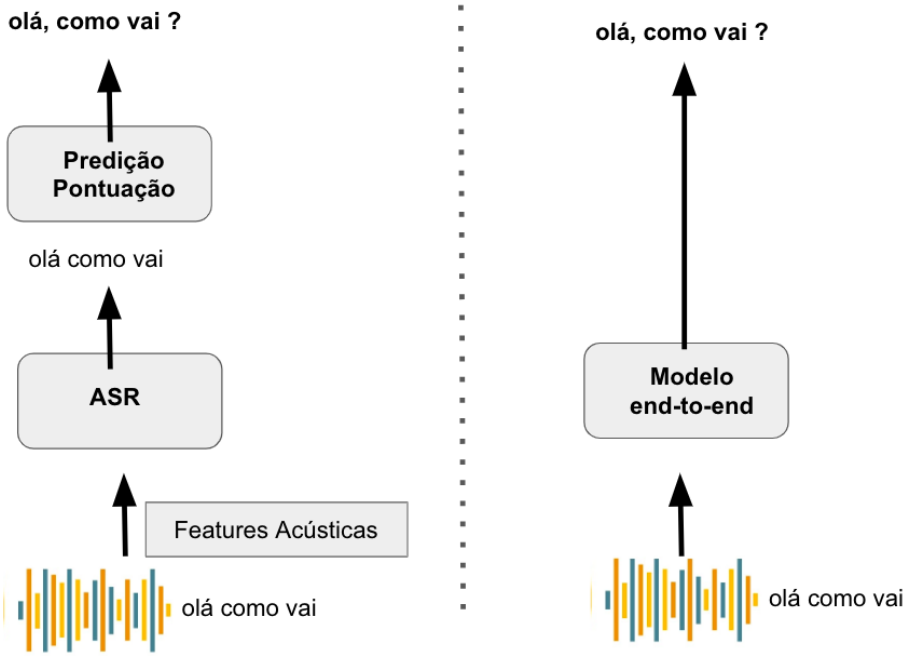
A abordagem dominante na literatura, conhecida como abordagem em cascata, é treinar um modelo de reconhecimento de fala (ou usar um pronto) e um modelo de predição de pontuação separadamente, e, em seguida, colocá-los em cascata, ou seja, inserir marcas de pontuação na transcrição gerada pelo ASR como uma etapa de pós-processamento. Para cada *token* na saída do ASR, as *features* acústicas do áudio são obtidas e são usadas como entrada para o módulo de predição de pontuação (Figura 3.6, esquerda).

[29] https://github.com/nilc-nlp/asr-punctuation-evaluation

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

Figura 3.6: Abordagem de um sistema em cascata (esquerda) e abordagem de sistemas *end-to-end* (direita).



Fonte: Adaptado de (Nozaki et al., 2022, fig. 1, p. 1812)

Em relação aos recursos, os três tipos de recursos usados para predição e restauração de sinais de pontuação são lexicais, prosódicos e a combinação de recursos prosódicos e lexicais.

Abordagens lexicais recentes na literatura para restauração de pontuação usam redes neurais profundas. As abordagens variam desde o uso de word embeddings pré-treinados, mecanismo de atenção, abordagens baseadas em Transformers treinadas em grandes *corpora* de texto, seja usando apenas o modelo BERT pré-treinado (Devlin et al., 2019) ou realizando uma comparação de modelos diferentes baseados em Transformers. Quanto às *features* prosódicas elas são baseadas nas durações das pausas, início e duração das palavras que são seguidas por uma pontuação, F0 (pitch) e duração de fonemas, além de *features* mais complexas baseadas no modelo wav2vec[30].

Os métodos usam as métricas precisão (P), revocação (R) e F1-score (F1) para avaliar o desempenho dos quatro rótulos (Vírgula, Ponto Final, Ponto de Interrogação e O – sem sinal de pontuação), propostos no IWSLT 2011 (*International Workshop on Spoken Language Translation* 2011).

No artigo de Gris et al. (2023), foram revisados cinco trabalhos da literatura para uma comparação com o uso do Whisper no *dataset* de teste do MuPe[31]. A Tabela 3.6 mostra um resumo dos trabalhos avaliados, com indicação do *corpus* usado na avaliação e seu tamanho em número de enunciados/duração em horas; os conjuntos de teste do IWSLT 2011 são apresentados em número de palavras para transcrições manuais (Referência) e transcrições automáticas (realizadas por um ASR), nesta ordem (Che et al., 2016).

[30] https://github.com/pytorch/fairseq/blob/master/examples/wav2vec
[31] https://github.com/nilc-nlp/asr-punctuation-evaluation

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Tabela 3.6: Resumo dos trabalhos apresentados por Gris et al. (2023).

Trabalhos	Língua(s)	Test Set	Duração/Tam.
Alam et al. (2020)	inglês, bengali	IWSLT 2011 Corpus	12.626/12.822 pal.
Yi; Tao (2019)	inglês	IWSLT 2011 Corpus	12.626/12.822 pal.
Zelasko et al. (2018)	inglês	Fisher Corpus	1.100 enunciados
Sunkara et al. (2020)	inglês	Fisher Corpus	42 h
Nozaki et al. (2022)	inglês, japonês	MuST-C Corpus	2.641 enunciados
Gris et al. (2023)	português	Corpus CORAA MuPe	16:48:34 h

Alam et al. (2020) avaliaram vários modelos de língua para o inglês (BERT, RoBERTa, ALBERT, DistilBERT) e modelos multilíngues para o bengali (mBERT, XLM-RoBERTa) disponíveis no repositório Hugging Face[32]. Avaliaram o desempenho dos quatro rótulos: Vírgula, Ponto, Pergunta e O (sem sinal de pontuação seguido), nos dois conjuntos de teste do IWSLT 2011.

Yi; Tao (2019) propuseram um modelo baseado em auto-atenção usando embeddings de palavra e de fala, respectivamente Glove (Pennington et al., 2014) e Speech2Vec (Chung; Glass, 2018), resolvendo o problema de dependência dos dados de fala alinhados com sua transcrição, pois para muitas línguas há carência destes recursos. Como Alam et al. (2020), os autores também avaliaram seu modelo no conjunto de dados do IWSLT 2011, mas os resultados de Alam et al. (2020) ainda são melhores do que os de Yi; Tao (2019) (exceto para ponto de interrogação), provavelmente devido ao fato de Alam et al. (2020) usar uma técnica de aumento de dados, que melhora o desempenho em dados com ruídos, e um modelo baseado em Transformers.

Zelasko et al. (2018) e Sunkara et al. (2020) avaliaram a previsão de pontuação na fala espontânea. Zelasko et al. (2018) reforçam que o problema da tarefa de previsão de pontuação para fala espontânea é a falta de conjuntos de dados de referência. Eles avaliaram dois modelos de redes neurais profundas: um baseado em *Convolutional Neural Nets* (CNN) e outro baseado em redes *Long Short-Term Memory* Bidirecionais (Bi-LSTM). Os modelos são treinados no Fisher Corpus (Cieri et al., 2004), que inclui anotação de pontuação e capitalização. O conjunto de dados de treinamento consiste em 348 horas de conversação e os conjuntos de desenvolvimento e teste cada um contém cerca de 42 horas.

Sunkara et al. (2020) propõem uma nova estrutura de fusão multimodal de embeddings lexicais e acústicos para previsão de pontuação em fala espontânea chamada arquitetura de aprendizagem semissupervisionada multimodal (MuSe). Embora os resultados de Sunkara et al. (2020) não sejam diretamente comparáveis com os de Zelasko et al. (2018) no Fisher Corpus, pois as divisões dos conjuntos de dados são diferentes, Sunkara et al. (2020) obtiveram melhor desempenho em todas as classes de pontuação.

Nozaki et al. (2022) propuseram um modelo *end-to-end* para reconhecimento de fala com pontuação (Figura 3.6, direita). Eles usaram dois conjuntos de dados de idiomas diferentes: o MuST-C, um *corpus* multilíngue[33] (Di Gangi et al., 2019) que foi usado como o conjunto de dados em inglês e o JCALL, um conjunto de dados fechado que consiste de gravações de áudio de conversas, foi usado como o conjunto de dados em japonês.

A OpenAI lançou, em setembro de 2022, o Whisper ASR. Embora também seja um

[32] https://huggingface.co/docs/transformers/index

[33] MuST-C inclui gravações de áudio de TED Talks em inglês alinhadas no nível da frase com suas transcrições e traduções manuais.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

modelo de ASR *end-to-end* semelhante à abordagem de Nozaki et al. (2022), ele tem duas diferenças importantes: é de código aberto e foi treinado em um *dataset* grande e multilíngue. Whisper é um ASR capaz de incluir pontuação e capitalização nas transcrições (Radford et al., 2022), embora seja somente capaz de gerar 5 tipos de pontuações: reticências, ponto final, vírgulas, ponto de interrogação e ponto de exclamação. O conjunto de dados de teste do MuPe possui duas pontuações a mais (ponto e vírgula e dois pontos).

3.7 Considerações finais

Apresentamos neste capítulo os recursos de processamento de fala com foco em reconhecimento de fala, síntese de fala, reconhecimento de emoções, análise prosódica, e predição de pontuação em ASR. O capítulo apresentou diversos recursos desenvolvidos no âmbito do projeto TaRSila. O TaRSila, juntamente com outros projetos citados neste livro, tem o objetivo de alavancar as pesquisas no processamento da fala da língua portuguesa, permitindo que o idioma saia da faixa considerada como línguas de baixos recursos, e trazendo diversidade nos tipos de dados disponibilizados, como a inclusão de recursos para tratar a fala espontânea.

Além dos trabalhos descritos no capítulo, vários pesquisadores do projeto TaRSila trabalham com a área de processamento de áudio em português aplicado na área de saúde. O leitor encontra no Capítulo Classificação de Áudio aplicada à Saúde modelos e técnicas de classificação de áudio para a voz e fala de pacientes com problemas respiratórios.

Agradecimentos

Em primeiro lugar agradecemos aos bolsistas do projeto TaRSila que foram incansáveis nas revisões das transcrições automáticas, no treinamento e teste dos modelos para vários sistemas de processamento de fala. Este trabalho faz parte de um Acordo de Transferência de Tecnologia entre Museu da Pessoa (MuPe), Instituto de Ciências Matemáticas e de Computação da Universidade de São Paulo (ICMC/USP) e Universidade Federal de Goiás. Este trabalho foi realizado no Centro de Inteligência Artificial (C4AI-USP), com apoio da Fundação de Amparo à Pesquisa do Estado de São Paulo (bolsa FAPESP nº 2019/07665-4) e da IBM Corporation. Agradecemos também o apoio do Centro de Excelência em Inteligência Artificial (CEIA) financiado pela Fundação do Estado de Goiás (bolsa FAPEG nº 201910267000527), à Fundação de Apoio à Universidade de São Paulo (FUSP) e ao Conselho Nacional de Desenvolvimento Científico e Tecnológico (bolsa PQ CNPq, processo 304961/2021-3). Este projeto também foi apoiado pelo Ministério da Ciência, Tecnologia e Inovação, com recursos da Lei nº 8.248, de 23 de outubro de 1991, no âmbito do PPI-SOFTEX, coordenado pela Softex e publicado Residência no TIC 13, DOU 01245.010222/2022-44.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Parte III

Palavras

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Capítulo 4

Sequência de caracteres e palavras

Morfologia e morfossintaxe

Maria José Bocorny Finatto
Helena de Medeiros Caseli
Lucelene Lopes
Amanda Rassi

Publicado em: 26/09/2023

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

Neste capítulo, trataremos de algo que parece simples, mas não é: identificar a unidade mínima quando tratamos, computacionalmente, a língua. Essa delimitação não é consensual entre pesquisadores e profissionais de PLN. E, mesmo em linguística, há sempre controvérsias e necessidade de pontos de referência para se definir, por exemplo, o que seja uma palavra ou mesmo uma frase.

As subáreas especializadas dos estudos linguísticos entendem como **unidade mínima de processamento** diferentes elementos conforme seus focos e pontos de vista. A fonologia (Capítulo 2), por exemplo, considera o fonema como a menor unidade sonora e distintiva de uma língua. Se tomarmos o exemplo do que diferencia as palavras "sábia" (mulher inteligente), "sabiá" (pássaro) e "sabia" (verbo "saber"), percebemos que a sílaba tônica é o diferencial, especialmente quando pensamos em sons e fala e não em escrita.

Já a morfologia considera o **morfema** como a menor unidade dotada de significado na língua. Nessa perspectiva, temos os "pedacinhos" de palavras e seus valores, como seria o caso da marca de diminutivo "-inho", que assinala o masculino e o singular em "menininho", ou o segmento "-ei" no verbo "comprei", que marca um modo-tempo (pretérito perfeito do modo indicativo) e também um número-pessoa (primeira pessoa do singular).

Assim, conforme o ponto de vista de quem analisa, uma palavra pode ser feita de sons e de sílabas tônicas e/ou composta de vários segmentos gráficos menores. Podemos, ainda, considerar segmentos ou pedaços mais abrangentes, conforme o critério que utilizamos. Um exemplo nessa linha seria a palavra "guarda-pó", que pode ser considerada como uma palavra só ou a junção de duas palavras. Outro caso ilustrativo é "escova de dente", que, para alguns, é a união de três palavras, e, para outros, é uma palavra só, mesmo que não tenha hífen. Além dessas questões, também é controverso tratar das abreviaturas, siglas, interjeições, dos modos de escrita diferenciados nas redes sociais, com internetês, hashtags, emojis, símbolos e outras peculiaridades.

Fazendo um paralelo, podemos entender que, de modo geral, os modelos de PLN trabalham as **palavras** como unidade primária de processamento. Vejamos, por exemplo, o caso da frase no Exemplo 4.1.

Exemplo 4.1:

> Jacinta Maria comprou uma cadeira em São Paulo ontem e pagou 25 reais por ela.

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Na frase do Exemplo 4.1 são 15 palavras se considerarmos que é palavra toda a sequência de caracteres separada por um espaço em branco. Mas se pode pensar que Jacinta+Maria e São+Paulo são palavras compostas e que, talvez, o número 25 não seja bem uma palavra, não? A resposta será: depende do critério que você usar e da finalidade que tem ou busca com essa referência de unidade e/ou partes.

Ao fazer o processamento computacional de textos escritos, a definição de que tipo de unidade de processamento se quer buscar/estudar parece estar atrelada às necessidades da tarefa ou trabalho pretendidos. Geralmente, considera-se que uma palavra é, simplesmente, uma unidade grafológica delimitada, nas línguas europeias, entre espaços em branco na representação gráfica, ou entre um espaço em branco e um sinal de pontuação[1]. Essa é uma definição bastante concreta, e bastante prática. No entanto, ao pensarmos em nossos modelos computacionais e suas aplicações no mundo, é importante nos aprofundarmos um pouco mais na conceituação do que é uma palavra e nas possibilidades de processamento e implicações das decisões tomadas no pré-processamento dos *corpora*.

Segundo Cabré (1999, p. 20), as palavras são as unidades de referência da realidade empregadas pelos falantes. De acordo com essa definição, as palavras compõem a dimensão linguística mais estreitamente ligada ao mundo real. Ainda segundo a mesma autora, o **léxico** consiste no conjunto das palavras de uma língua e dos padrões que possibilitam a criatividade do falante. As palavras e, principalmente, as associações infinitas e imprevisíveis que os seres humanos são capazes de traçar entre elas, constituem a manifestação mais concreta e mais produtiva da língua. Assim, é importante que, ao processar dados textuais para gerar modelos computacionais, nos recordemos sempre de que não estamos simplesmente organizando um conjunto de caracteres ou ordenando uma representação ortográfica formal, mas que estamos trabalhando com recursos linguísticos que representam a experiência humana.

Em seguida, devemos considerar também o conceito de **palavra computacional**, que se refere a uma unidade linguística que foi adaptada ou criada especificamente para facilitar seu processamento por máquinas. Isso pode envolver a manipulação de palavras, frases ou até mesmo caracteres de maneira que seja mais conveniente para algoritmos e sistemas de PLN lidarem com elas.

A necessidade dessas palavras computacionais surge devido às complexidades do processamento de linguagem natural por máquinas. A linguagem humana é rica e ambígua, cheia de nuances e variações que podem ser difíceis de interpretar e analisar automaticamente. Portanto, ao transformar palavras em formas mais padronizadas ou simplificadas, os sistemas de PLN podem executar tarefas como análise gramatical, extração de informações (Capítulo Extração de Informação) e tradução (Capítulo Tradução Automática) com maior eficácia. Dependendo do objetivo da tarefa ou aplicação de PLN, é possível definir quais rotinas de pré-processamento são mais produtivas para criar as palavras computacionais, ou seja, pode-se remover ou não acentos ortográficos, espaços em branco em itens como “fim de semana”, ou hífens como em “guarda-chuva”.

No mesmo sentido, dependendo dos propósitos da tarefa, um modelo de aprendizado de máquina pode precisar definir expressões multipalavras (Capítulo 5) (e.g. “nem que a vaca tussa”, “deus me livre” ou “ciência de dados”), hashtags (e.g. “#euamopuzzles”), URLs e outros compostos como palavras (ou unidades lexicais) únicas, como se fossem representadas sem espaços: “nemqueavacatussa”, “deusmelivre”, “ciênciadedados” e “euamopuzzles”.

[1] Considerando línguas como o português nas quais o processo de tokenização (apresentado na Seção 4.2) pode ser baseado na presença de espaços em branco delimitadores das palavras. Em outras línguas, como o alemão ou o chinês, por exemplo, esse processo de tokenização pode ser um pouco mais complexo.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Essas e outras estratégias similares podem facilitar o processamento automático, mas devem considerar as necessidades de cada aplicação computacional.

Este capítulo apresenta conceitos (Seção 4.1) relacionados a essas unidades mínimas de processamento e que devem ser considerados quando lidamos com textos. Em seguida, são apresentadas as tarefas relacionadas ao processamento morfológico dos textos (Seção 4.2). E depois são indicadas as ferramentas e os recursos disponíveis para o português, exemplificando com uma aplicação prática de um recurso específico para o português brasileiro (Seção 4.3). Por fim, fazemos uma retomada dos principais tópicos discutidos, apresentando as considerações finais e os planos para uma futura versão deste capítulo (Seção 4.4).

4.1 Conceitos básicos da morfologia

Antes de vermos como identificar e tratar computacionalmente as unidades mínimas de processamento, cabe definir alguns conceitos linguísticos básicos necessários. Nesta seção, definimos os conceitos de Morfema (e todos os seus tipos) (Seção 4.1.1), *Token* e *Type* (Seção 4.1.2), Lexema, Lexia e Lema (Seção 4.1.3), Léxico e Gramática (Seção 4.1.4), Léxico comum e especializado (Seção 4.1.5), além de palavras funcionais e lexicais (Seção 4.1.6). Essa seção traz, ainda, informações sobre os processos de formação das palavras (Seção 4.1.7) e morfologia e morfossintaxe (Seção 4.1.8).

4.1.1 Morfema

O objeto principal de estudo da morfologia é o **morfema**, definido linguisticamente como "unidade mínima significativa". Segundo essa definição, o morfema é a menor unidade linguística dotada de significado, considerando que há outras unidades linguísticas que também possuem significado, como a palavra, o sintagma, a frase, a oração, o período, o texto etc. Além disso, o **morfema** é considerado como "dotado de significado" por oposição ao fonema, que é o objeto de estudo da fonética e da fonologia, e, na verdade, é a menor unidade de análise linguística, porém não possui significado em si, mas tem a função de estabelecer diferença de significado entre uma palavra outra. Em outras palavras, o **fonema** é apenas uma unidade linguística distintiva (não significativa), pois diferencia as palavras por meio de seus sons (e.g. "faca" e "vaca" são duas palavras diferentes que se distinguem pelo fonema inicial "\f" ou "\v"). O mesmo vale para o trio "sábia", "sabia" e "sabiá", lembrando que estamos no território dos sons e não da escrita. Para saber mais sobre o processamento da fala, sugere-se a leitura de Capítulo 2 e Capítulo 3.

Em uma explicação simples, podemos dizer que os morfemas são os pedacinhos que se juntam para formar as palavras. E esses "pedacinhos" podem ser de vários tipos: **desinência**, **raiz**, **radical**, **afixo**, **vogal temática** e **tema**. Por exemplo, podemos juntar o radical "experiment" com a vogal temática "a" com a desinência verbal "ri" e com a desinência verbal "a" para formar a palavra "experimentaria". Esses quatro "pedacinhos" são chamados de morfemas e eles possuem significados:

(i) "experiment" significa o conceito lexical de "prova, ensaio, tentativa";
(ii) "a" significa que é um verbo da primeira conjugação;
(iii) "ri" significa que esse verbo está flexionado no tempo futuro do pretérito do modo indicativo;

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

(iv) "a" significa que esse verbo está flexionado na terceira pessoa do singular.

A seguir apresentamos brevemente uma definição e exemplos dos vários tipos de morfemas em português.

4.1.1.1 Desinência

Desinências são os morfemas que geralmente ficam no final da palavra e podem marcar gênero e número (no caso dos substantivos e adjetivos) ou marcar número, pessoa, tempo e modo (no caso dos verbos). Por isso as desinências podem ser classificadas em: nominais ou verbais.

Em português, o substantivo "meninas" é formado pelo radical "menin" + duas **desinências nominais**: "a", que indica feminino, e "s", que indica plural. Assim, dizemos que a palavra "meninas" está flexionada no feminino plural. Já o verbo "adotássemos" é formado pelo radical "adot" + a vogal temática "a" + duas **desinências verbais**: "sse", que indica modo e tempo (pretérito imperfeito do subjuntivo), e "mos", que indica número e pessoa (primeira pessoa do plural). Assim, dizemos que o verbo "adotássemos" está flexionado na primeira pessoa do plural do pretérito imperfeito do subjuntivo.

4.1.1.2 Raiz e radical

Na linguística teórica, existe uma diferença conceitual entre **raiz** e **radical**. Embora a definição de ambos os conceitos ressalte que são os constituintes da palavra que contêm significado lexical, no caso da **raiz**, ela não inclui afixos derivacionais ou flexionais (e.g. "beb" é a raiz de "beber", "beberemos", "bebendo", "bebida", "bebidinhas" e tantas outras formas flexionadas). No caso do **radical**, ele não inclui afixos de flexão, mas pode incluir afixos derivacionais (e.g. "beb" é o radical de "beber", "beberemos" e "bebendo", mas "bebid" seria o radical correto de "bebida" e de "bebidinhas"). Essa distinção, em termos linguísticos, é muito sutil e geralmente as aplicações de PLN assumem os dois termos como sinônimos.

A raiz ou radical é o morfema nuclear de uma palavra, ou seja, aquele constituinte básico que expressa sua base significativa, que designa o significado lexical da palavra. Portanto, ele é o componente comum a todas as palavras que pertencem à mesma família. Por exemplo, em português, "menino", "meninas", "meninada", "meninice" e outras possuem a mesma raiz ou radical "menin".

4.1.1.3 Afixo

Os **afixos** são os morfemas lexicais que se juntam com o radical ou com o tema para formar uma nova palavra, neste caso chamada de palavra derivada. A inserção de um afixo ao radical de uma palavra pode mudar-lhe o sentido ou adicionar-lhe uma ideia secundária ou ainda mudar sua classe gramatical.

Em português, os afixos podem ser de três tipos:

1. **prefixos**, quando são inseridos antes do radical (ex: "**des**matar", "**i**mortal", "**anti**oxidante", "**re**fazer";
2. **infixos**, quando são inseridos no meio de um radical, mas são bem raros; e
3. **sufixos**, quando são anexados ao final do radical (ex: "ativa**mente**", "imagina**ção**", "cresci**mento**").

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Apesar de os sufixos e as desinências serem morfemas acoplados ao final da palavra, eles não devem ser confundidos, pois os primeiros criam novas palavras derivadas, a partir de um processo de formação de palavras chamado de derivação. Já a adição de desinências não cria novas palavras, apenas flexiona a palavra existente em uma nova forma flexionada.

4.1.1.4 Vogal temática

Vogal temática é o nome dado às vogais que aparecem imediatamente após o radical da palavra, mas não representam seu gênero. Em português, as vogais temáticas podem ser de dois tipos:

1. **nominais**, que podem ser "**a**" (ex: "atlet**a**", "colega", "dentist**a**"), "**e**" (ex: "agent**e**", "recorrent**e**", "alegr**e**") ou "**o**" (ex: "pássar**o**", "crocodil**o**", "dezembr**o**"); e
2. **verbais**, que indicam as 3 conjugações verbais: "**a**", para verbos da primeira conjugação (ex: "and**a**r", "passe**a**r", "fal**a**r"), "**e**" ou "**o**", para verbos da segunda conjugação (ex: "escrev**e**r", "l**e**r", "faz**e**r", "p**ô**r", "rep**o**r", "comp**o**r") ou "**i**", para verbos da terceira conjugação (ex: "ag**i**r", "assum**i**r", "part**i**r").

As vogais temáticas não devem ser confundidas com as desinências nominais, por exemplo, que indicam gênero em "menin**o**" e "menin**a**". Neste caso, "o" e "a" são desinências nominais porque esses morfemas marcam os gêneros masculino e feminino, respectivamente. Novamente, cabe mencionar que há uma boa e extensa discussão sobre a natureza e funcionamento das vogais temáticas em linguística. Para quem quiser uma visão aprofundada, vale consultar o trabalho de Santana (2019), uma tese de doutorado sobre vogais temáticas.

4.1.1.5 Tema

Tema é a forma lexical que se cria quando se juntam dois morfemas: o radical e a vogal temática. Por exemplo, a partir da combinação do radical "crianç" com a vogal temática nominal "a", forma-se o tema "criança". Embora seja a junção de dois tipos de morfemas, o tema também é considerado como um morfema.

A forma lexical assumida pelo **tema** coincide com as formas de lexema, lexia e lema, que são as formas de entrada dos verbetes em dicionários, e que serão explicadas na Seção 4.1.3.

Ressalte-se, no entanto, que os termos **tema** e **lema**, na Linguística Textual, representam ideias completamente diferentes desses conceitos da morfologia.

4.1.1.6 Considerações até então

Por fim, vale dizer que todos esses tipos de morfemas explicados nas subseções anteriores podem ser agrupados em duas categorias: (i) **morfemas lexicais**, que representam a família semântica de determinada palavra, ou seja, a raiz, o radical e o tema; (ii) **morfemas gramaticais**, que inserem alguma informação à palavra existente, ou seja, as desinências, os afixos e as vogais temáticas.

Há ainda um tipo de morfema (ou fonema, dependendo da abordagem) que não foi explorado aqui porque não é relevante para os estudos de PLN, que são as vogais e as consoantes de ligação. Elas não possuem significado, mas, por vezes, são inseridas entre um radical e uma desinência ou um afixo, por uma motivação fonológica.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

4.1.2 *Token* e *Type*

Token é um termo que significa qualquer sequência de caracteres à qual se atribui um valor. Nas línguas europeias, a sequência consiste em caracteres delimitados por espaços gráficos, sendo que a tokenização é ajustada para separar sinais de pontuação. Mas, na grande maioria das línguas, a tokenização não opera por espaços gráficos. Diante dessa definição, é comum associarmos *token* à palavra escrita. Nesse sentido, a quantidade de palavras + sinais de pontuação de uma sentença equivale à quantidade de *tokens*, por exemplo, a frase do Exemplo 4.2 contém 12 *tokens*, já que, em PLN, os sinais de pontuação (vírgula e ponto final) também são considerados *tokens*[2].

Exemplo 4.2:

> Eu sempre viajo para Campinas, para Salvador e para Belém.

Type, por sua vez, refere-se aos *tokens* únicos encontrados numa frase ou texto. Retomando a frase do Exemplo 4.2, encontramos 10 *types* ("eu", "sempre", "viajo", "para", "Campinas", ",", "Salvador", "e", "Belém" e "."). Nessa sentença, a palavra "para" ocorre três vezes, então ela é contada 3 vezes como *token*, mas apenas 1 vez como *type*.

A proporção *token*/*type* (divisão da quantidade de *tokens* pela quantidade de *types*) é um importante indicativo da riqueza lexical de um texto; ou seja, ela indica qual a diversidade de palavras existentes em um *corpus*, excluindo suas repetições. Mas, nessa medida, apenas as formas de palavras (as palavras diferentes) e o número total de palavras são contados. Isto é, sinais de pontuação não são considerados.

4.1.3 Lexema, Lexia e Lema

Lexema é sinônimo de unidade lexical, o que implica características de som, forma e significado. Por exemplo, "comprei" é um lexema cuja representação fonética é [kõpr' ej] ; morfologicamente, é um verbo flexionado na primeira pessoa do singular, no pretérito perfeito do modo indicativo. Seu significado é o que encontramos nos dicionários: adquirir (algo, produto, serviço etc.) em troca de pagamento.

Vale assinalar que, nos estudos do léxico do Brasil, temos também o termo técnico **lexia,** que corresponde à realização concreta de um **lexema.** Por exemplo, um **lexema** – que seria uma forma em abstrato, como "árvore" – pode acontecer sob a forma de uma **lexia** como "árvores". **Lexia** é, nessa perspectiva, uma "forma que um **lexema** assume no discurso. Exemplo: 'O dia está claro.' Temos aí quatro lexias" (Biderman, 1978, p. 130). A **lexia** realiza-se no discurso e/ou texto e se distingue do **lexema**, que se situa ao nível do sistema abstrato que é a língua. Em resumo, o **lexema** é uma representação conceitual enquanto a **lexia** é a unidade linguística materializada no discurso.

Como você pode perceber, em linguística, temos vários termos para designar, algumas vezes, uma mesma noção. Por isso, um termo como **palavra** também equivale a **vocábulo**. Se quiser saber mais sobre essas diferentes concepções linguísticas, no âmbito dos estudos do léxico, vale dar uma olhada na parte introdutória do trabalho de Sarmento (2019).

Por sua vez, **lema** é a representação das propriedades sintático-semânticas de um item lexical. Isso significa que, a partir de um lema, é possível saber quais argumentos a ele se relacionam. Por exemplo, "comprar" é um verbo que seleciona dois argumentos: um

[2]Para isso, em muitos modelos, é necessário uma etapa a mais de processamento para isolar o sinal de pontuação da palavra anterior, já que graficamente eles costumam estar juntos.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

sujeito e um objeto. Esses dois argumentos são necessários para que a estrutura na qual ele está inserido seja gramatical, ou seja, aceita e compreendida pelos falantes. Além disso, é por meio do lema que se pode acessar seu significado: "comprar" remete a uma ação que envolve uma moeda e a obtenção de algo. Nesse sentido, o lema pode ser considerado uma parte do lexema.

A palavra, na forma de lema, é também a forma de entrada dos verbetes em um dicionário, tendo-se em mente também uma categoria de palavra. Por isso, temos "filósofo" como o lema dos substantivos "filósofo", "filósofos", "filósofa" e "filósofas" e temos "filosofar" como o lema de "filosofei", "filosofamos", "filosofemos" e todas as demais flexões do verbo.

4.1.4 Léxico e Gramática

> Se é verdade que não existe língua sem gramática, mais verdade ainda é que sem léxico não há língua. As palavras são a matéria-prima com que construímos nossas ações de linguagem.
>
> (Antunes, 2017, p. 27)

A afirmação de Antunes (2017) na epígrafe diz muito e coloca em relação os elementos que estruturam e fazem funcionar uma língua. Um conjunto de regras sistemáticas servem para definir o que é considerado certo ou normal em uma língua. Por exemplo, segundo a regra de ordem de palavras, usamos, em um discurso normal, não poético, a frase "ele leu o livro" e não "o leu ele livro". Regras como essa apontam para **gramática**, enquanto o **léxico** corresponde ao conjunto de palavras de uma língua.

Léxicos contêm as palavras de uma língua juntamente com as definições morfossintáticas (Seção 4.1.8) possíveis para cada uma das palavras. Geralmente cada palavra do léxico tem associada a ela uma ou mais triplas com sua categoria gramatical, também chamado de PoS (*Part-of-Speech*), seu lema e suas características morfológicas, também chamadas de *features*.

As categorias gramaticais podem variar segundo os critérios da representação que será adotada, podendo seguir um entre diversos padrões. Porém, para o português é usual definir as categorias gramaticais: **substantivos**, **adjetivos**, **nomes próprios**, **numerais**, **pronomes**, **preposições**, **conjunções**, **advérbios** e **verbos**. Dependendo dos critérios escolhidos, pode-se incluir outras categorias como **artigos** ou **determinantes**. Pode-se também promover divisões, como por exemplo dividir as **conjunções** em **conjunções coordenativas** e **conjunções subordinativas**, ou ainda **verbos** em **verbos auxiliares** e **verbos plenos**. A escolha do conjunto de categorias possíveis é a primeira decisão para a construção de um léxico. Um exemplo de categorias adotadas para o português é apresentado na Seção 4.3.3.

Igualmente, a escolha das características morfológicas que serão consideradas e seus valores possíveis é também uma decisão importante que deve ser tomada. É usual que a definição de categorias de PoS e características morfológicas seja acompanhada de um conjunto de etiquetas (em inglês, *tags*) que serão usadas para representar as informações associadas a cada palavra do léxico. Por exemplo, uma entrada de um léxico para a palavra "elas" pode conter a categoria gramatical pronome, o lema "ele" e características de pronome pessoal na terceira pessoa do plural e gênero feminino. Neste exemplo, trata-se de uma palavra que só possui uma possível tripla PoS, lema e *features*. No entanto, é bastante comum encontrarmos palavras que possuem diversas triplas possíveis de informações associadas, como por exemplo, a palavra "casas" que pode ser:

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

- um verbo, com o lema "casar", no presente do indicativo, na segunda pessoa do singular;
- um substantivo, com o lema "casa", que é do gênero feminino e está no plural.

Cabe salientar que pela própria natureza das línguas, por mais completo que um léxico possa ser, sempre é possível ter palavras da língua ausentes do léxico.

4.1.5 Léxico comum e Léxico especializado

O **léxico comum** corresponde ao conjunto de palavras de uma língua que não têm um "conceito técnico-científico" bem determinado, historicamente construído, atrelado a ela. Em contrapartida há o **léxico especializado**, no qual a palavra assume um significado específico/especial em relação a um sistema de conceitos específico, que geralmente corresponde a uma área de conhecimento, ciência ou especialidade. É esse ambiente "especializado" que definirá se ela pode ser entendida como uma terminologia "técnica" (**termo**) ou uma palavra comum. Um item que a gente lê e diz que é um termo é, por exemplo, "ferritina", enquanto "caderno" parece um protótipo de palavra comum, do léxico comum. Novamente, pode-se pensar que a categorização ou classificação são referências e que sempre pode haver algo que parece um meio-termo.

E há ainda termos técnicos que passam a funcionar como palavras, na língua comum, e vice-versa. Um exemplo interessante é o caso de "criança", que no âmbito jurídico corresponde a uma pessoa que tem até doze anos de idade, conforme vemos no Estatuto da Criança e do Adolescente (ECA)[3] do Brasil. Isto é, nesse "cenário especializado" do ECA, a palavra "criança" assume status de terminologia, pois corresponde a um conceito específico, oposto a outros. Já a palavra "acetona", como sinônimo de "removedor de esmalte de unhas", é algo que fez o caminho inverso, uma vez que passou de termo do léxico especializado a palavra do léxico comum.

Veja como fica o termo "DNA", que é uma sigla para um termo "técnico" em inglês, nome de um ácido, que é usado em diferentes situações e parece circular entre o ambiente técnico Exemplo 4.4 e o ambiente da linguagem comum Exemplo 4.3, do nosso dia a dia – pois passou a corresponder a um nome de um exame para confirmação de paternidade.

Exemplo 4.3:

> No programa de TV, Joana disse que ia jogar na cara do ex-namorado um **DNA**. E disse que ele, depois, ia ver que o filho que ele renegou tem um DNA de gente de bem.

Exemplo 4.4:

> O **DNA** (ácido desoxirribonucleico) é um tipo de ácido nucleico que possui papel fundamental na hereditariedade, sendo considerado o portador da mensagem genética.[4]

Vale mencionar, também, que para aplicações que envolvem mais do que um idioma, como a Tradução Automática (Capítulo Tradução Automática), os léxicos são bilíngues

[3] https://www.planalto.gov.br/ccivil_03/leis/l8069.htm
[4] Fonte: https://www.biologianet.com/biologia-celular/dna.htm

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

(ou multilíngues) especificando não apenas as palavras que compõem os léxicos dos vários idiomas, mas também o mapeamento (paralelismo) existente entre palavras de um e outro(s) idioma(s).

4.1.6 Palavras funcionais e palavras lexicais

As **palavras funcionais/gramaticais** e as **palavras lexicais** são outra dualidade, também complexa, que podemos tentar "resolver" ou melhor, entender, pensando em classificá-las. As palavras funcionais/gramaticais ficam em uma **classe fechada**. Já as palavras lexicais ficam em outro grupo ou tipo, pensando que correspondem a uma **classe aberta**. A classe fechada é assim pensada porque tem um número finito de componentes. A classe aberta, por outro lado, acomoda um número bem maior de componentes, pois é uma classe que tem a ver com a capacidade de as pessoas criarem palavras novas.

Podemos pensar que as preposições do português são as mesmas desde sempre; não criamos muitas. Já os adjetivos e os substantivos não param de nos surpreender, pois parece que há uma inventividade envolvida em nomes e qualificativos, como o adjetivo, que também pode ser substantivo "cloroquiner". Essa nova palavra surgiu no contexto da Pandemia de Covid-19, em 2020, no Brasil.

Pensar em conjuntos também nos ajuda a entender essa diferença entre funcional/gramatical e lexical. Mas sempre poderemos pensar que uma palavra como "não" é uma palavra lexical, se o critério para classificar for "palavra que tem um sentido" em si mesma. Via de regra, algumas classes de palavras são sempre consideradas como de classe aberta, como os verbos, os adjetivos e os substantivos, enquanto outras classes são sempre definidas como de classe fechada, tais como os artigos (determinantes), as preposições e as conjunções. Outras classes, como os advérbios ou os pronomes, por exemplo, podem ser considerados palavras lexicais ou funcionais, dependendo de suas subclassificações.

4.1.7 Processos de formação das palavras

Existem dois tipos de processos usados para a formação de novas palavras: (i) por derivação e (ii) por composição. São mecanismos linguísticos que permitem criar novas palavras a partir de unidades já existentes na língua.

A **derivação** é um processo pelo qual novas palavras são criadas adicionando afixos (prefixos, sufixos, infixos etc.) à raiz ou radical. Esses afixos podem alterar o significado, a classe gramatical (substantivo, adjetivo, verbo etc.) ou outros aspectos da palavra base. Por exemplo, considere o substantivo "amigo". Se adicionarmos o sufixo "-ável" a ele, obtemos o adjetivo "amigável". Nesse caso, o sufixo altera o sentido da palavra e também sua classe gramatical.

Existem cinco tipos de derivação:

1. **prefixal**, quando se adiciona um prefixo ao radical;
2. **sufixal**, quando se adiciona um sufixo ao radical, como no exemplo acima;
3. **parassintética**, quando se adiciona ao mesmo tempo um prefixo e um sufixo ao radical, como no caso de "desmatamento", que é derivado de "mata";
4. **imprópria**, quando muda a categoria gramatical da palavra, mas sem alterar sua forma, como no caso de "Ela tem um andar lento", em que "andar" originalmente é um verbo, mas passa a ser um substantivo nesse contexto; e
5. **regressiva**, quando se suprime uma desinência de um verbo para formar um substantivo, como é o caso de "choro", que é derivado de "chorar".

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Já a **composição** é um processo em que novas palavras são formadas combinando duas ou mais palavras independentes, ou dois radicais, para criar uma nova palavra com um significado diferente. As palavras compostas podem ser formadas por substantivos, adjetivos, verbos, advérbios e outras classes gramaticais. Além disso, elas podem ser escritas juntas, separadas por hífen ou até mesmo separadas sem qualquer marcação, dependendo da língua e das convenções ortográficas. São exemplos de palavras formadas por composição: "girassol" ("gira" + "sol"), "planalto" ("plano" + "alto"), "guarda-chuva" ("guarda" + "-" + "chuva").

Existem 2 tipos de composição:

1. por **justaposição**, em que uma nova palavra é formada a partir da união de dois ou mais radicais, sem apresentar alterações nos seus sons, ou seja, sem alterações fonéticas, como em "cachorro-quente" ("cachorro" + "-" + "quente"), "passatempo" ("passa" + "tempo"), "guarda-chuva" ("guarda" + "-" + "chuva"); e
2. por **aglutinação**, em que as palavras também são formadas pela união de dois ou mais radicais, porém sofrem alterações, como "vinagre" ("vinho" + "acre"), "embora" ("em" + "boa" + "hora") e "fidalgo" ("filho" + "de" + "algo").

Ambos os processos de derivação e de composição são fundamentais para a expansão do vocabulário e a expressão de nuances semânticas na linguagem.

4.1.8 Morfologia e morfossintaxe

Por fim, mas não menos importante, é necessário definir o escopo de estudo do que chamamos de morfologia, pois o seu objeto de estudo muitas vezes se intersecta com o objeto de outra área da linguística, a chamada Sintaxe, que será explorada no Capítulo 4. Na fronteira entre a morfologia e a sintaxe, está a morfossintaxe. Na prática, essas três áreas estão intimamente ligadas e relacionadas, mas, para fins didáticos, distinguimos esses termos a partir de seus objetos de estudo.

A **morfologia** é o ramo da linguística que se concentra no estudo dos morfemas, que são os "pedacinhos" significativos que formam as palavras. Assim, a morfologia examina como eles se combinam nos processos de flexão e de formação de palavras. Em PLN, a morfologia cuida também da classificação dos atributos morfológicos (ou *features* morfológicas), tais como os traços de gênero, número, modo, tempo, pessoa, voz, caso, entre outros.

Já a **morfossintaxe** examina como as escolhas morfológicas (como flexões verbais e concordância nominal) afetam a organização das palavras em uma sentença e como essas escolhas influenciam a estrutura sintática. Em outras palavras, ela categoriza as palavras em diferentes classes de palavras (ou categorias gramaticais) a partir da observação de seus atributos morfológicos. Em PLN, as classes de palavras são chamadas de *part-of-speech* ou PoS e a tarefa de atribuição de etiquetas de PoS nos textos será explicada na Seção 4.2.5.

Em resumo, a morfologia lida com a estrutura interna das palavras e os morfemas que as compõem, enquanto a morfossintaxe explora como as escolhas morfológicas afetam a estrutura das frases. Ambas as áreas são consideradas neste capítulo.

4.2 O processamento morfológico em PLN

Após definir conceitos necessários da área de morfologia, demonstraremos como tratar esse nível de análise linguística no Processamento de Linguagem Natural. Para desenvolver

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

praticamente qualquer aplicação de PLN, é necessário realizar fases/etapas que convencionamos chamar de **pré-processamento**. Nesse pré-processamento, algumas tarefas usuais são: segmentação do texto em sentenças (sentenciação), separação de palavras (tokenização), tokenização em subpalavras (vetorização de *subtokens*), normalização de palavras (lematização e radicalização), entre outras.

Além das etapas do pré-processamento, também podem ser realizadas tarefas de processamento do conteúdo dos textos, como a etiquetagem morfossintática das palavras em relação às suas classes gramaticais (tarefa de PoS *tagging*) e a anotação automática de seus atributos morfológicos (tarefa de anotação de *feats* ou *features* morfológicas), que também serão exploradas nesta seção.

4.2.1 Sentenciação

A sentenciação (ou sentenciamento) é o processo de segmentação do texto em sentenças, ou seja, é o processo de identificação de unidades textuais de processamento onde se definem os limites de cada sentença. A denominação de **detecção de limite de sentença** é frequentemente utilizada como sinônimo da segmentação de sentenças, pois o problema se limita a descobrir onde cada sentença termina (Hapke et al., 2019). Este processo é naturalmente complexo, pois a ambiguidade das línguas torna impossível ter sempre certeza de onde termina uma sentença (Read et al., 2012).

No caso do português escrito, as técnicas usuais se valem da busca de pontuações delimitadoras como “.”, “!” e “?”. Note-se que no processamento de textos falados, ou mesmo em algumas línguas onde a delimitação de sentenças não é feita por pontuação, o processo de segmentação de sentenças se torna ainda mais difícil.

A detecção do limite de sentença no português não tem como desafio identificar as pontuações delimitadoras, pois esse é usualmente um conjunto finito e conhecido (“.”, “!” e “?”, “...”). O desafio é desambiguar essas ocorrências com outros usos dos mesmos caracteres. Um exemplo disto é o caso das abreviações. Por exemplo, na sentença do Exemplo 4.5 temos duas ocorrências do caractere “.”.

Exemplo 4.5:

> Fui à clínica do Dr. Nilo.

Na primeira ocorrência, o “.” é utilizado como indicador da abreviação da palavra “Doutor” e, na segunda, como delimitador do fim da sentença. Note-se que a sentença do Exemplo 4.6 é uma sentença perfeitamente aceitável e, neste caso, o “.” está sendo utilizado duplamente como indicador de abreviação e fim de sentença.

Exemplo 4.6:

> Fui à clínica do Dr.

Outro caso comum de ambiguidade no uso de pontuações delimitadoras é encontrado em numerais. Em português, utiliza-se também o caractere “.” como delimitador de milhar em um número, enquanto em inglês ele é utilizado como separador de decimais. Algumas vezes ambos os usos aparecem, ainda que erroneamente, em sentenças em português, como no Exemplo 4.7.

Exemplo 4.7:

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

> A venda de 25.000 ações fez o índice de rentabilidade baixar para 0.5%, segundo a BOVESPA.

Um problema semelhante acontece com algumas definições matemáticas dentro de uma sentença, como quando se utiliza o caractere "!" para definir o fatorial de um número, como no Exemplo 4.8.

Exemplo 4.8:

> As permutações de cinco elementos podem ser calculadas como 5!, que é o fatorial de cinco.

Em todos estes casos, torna-se difícil detectar quando o caractere está sendo utilizado com função de fim de sentença ou não.

Por essas razões, o problema de segmentação automática de textos, ainda que explorado desde o início pela área de PLN, é bastante desafiador e ainda está em aberto. Atualmente utilizam-se três tipos de abordagens computacionais para resolvê-lo:

- **Abordagens baseadas em regras**, onde são definidos padrões de fim de sentença através de regras que podem incluir, por exemplo, heurísticas, abreviações usuais, expressões regulares para números e URLs. Este é, em geral, o método implementado em segmentadores de sentenças disponíveis em pacotes como o NLTK[5] (Bird; Loper, 2004).
- **Abordagens baseadas em aprendizado de máquina supervisionado**, ou seja, modelos computacionais treinados sobre conjuntos anotados (*gold standard*, veja Capítulo 16) onde o desafio é desenvolver um conjunto de treino de tamanho e características relevantes para os textos que se pretende sentenciar.
- **Abordagens baseadas em aprendizado de máquina não supervisionado**, ou seja, modelos computacionais treinados sobre conjuntos não anotados, mas que são suficientemente grandes e representativos para que se possa construir um modelo de linguagem adequado.

O problema de segmentação de sentenças é extremamente importante, pois, por ser uma etapa inicial do pré-processamento, os problemas não resolvidos nessa etapa tendem a prejudicar as etapas posteriores. Em comparação com outras tarefas de PLN, a segmentação de sentenças costuma receber menos atenção do que deveria, tanto no desenvolvimento de pesquisas, quanto na implementação de casos práticos.

Vale esclarecer que a sentenciação é uma tarefa de pré-processamento que tem relação com a morfologia porque, apesar de sua unidade de análise ser a sentença, os casos de ambiguidade (e, portanto, segmentação incorreta) têm a ver com a delimitação das palavras, que é o objeto de estudo da morfologia.

Na Seção 4.3.1, indicaremos alguns sentenciadores (ou também chamados *sentencizers*) disponíveis para o português.

4.2.2 Tokenização

A separação em unidades linguísticas mínimas é denominada **tokenização** (em inglês, *tokenization*) e, como já mencionado anteriormente, no caso do português é feita partindo

[5] https://www.nltk.org

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

da separação das palavras através de delimitadores. Neste caso, faz-se necessário identificar os limites das palavras através de caracteres delimitadores como espaços em branco ou símbolos de pontuação como “,”, “:”, “;”, “-” e “.”. Novamente, aqui é necessário se atentar para casos específicos como “,”, “-” e “.” que não devem ser separados dos demais caracteres que vêm antes ou depois. Por exemplo, a sentença do Exemplo 4.9 possui 11 *tokens*: “Além”, “disso”, “,”, “a”, “produção”, “será”, “descontinuada”, “em”, “8,3”, “%” e “.”, sendo que 8,3 deve ser considerado um *token* único e não 3 *tokens* separados.

Exemplo 4.9:

> Além disso, a produção será descontinuada em 8,3%.

Outra tarefa frequente da tokenização é a separação de **palavras contraídas**, por exemplo, a palavra “da” é separada em dois *tokens*: “de”+“a” e a palavra “nelas” é separada nos *tokens* “em”+“elas”. Essa tarefa é necessária para diversas aplicações e, em muitos casos, é um processo simples, pois a palavra contraída não é ambígua. No entanto, em alguns casos, pode ser necessário um processo de desambiguação, como nas palavras “pelo” (que pode ser “por”+“o” ou o substantivo “pelo”) e “consigo” (que pode ser tokenizada em “com”+“si” ou corresponder à conjugação do verbo “conseguir”). Esse é ainda um dos desafios da tokenização em português. A Figura 4.1 mostra o resultado da tokenização usando uma das ferramentas atualmente disponíveis para o português[6] para a sentença do Exemplo 4.10.

Exemplo 4.10:

> A raça Lulu da Pomerânia solta pelos em abundância pela casa.

Figura 4.1: Exemplo de tokenização com acertos e erros de descontração de palavras

```
<p> <s> A raça Lulu de_ a Pomerânia solta por_ os em abundância por_ a casa ./ </s> </p>
```

Na Figura 4.1 observamos que “da” e “pela” foram corretamente descontraídas como “de_ a” e “por_ a” respectivamente. Já o substantivo “pelos” foi incorretamente separado em “por_ os”. Problemas de descontração indevida, ou da falta dela quando necessária, geram problemas para a análise sintática, como veremos no Capítulo 4.

Além desses casos, há também outros que envolvem a decisão de separar ou não **palavras hifenizadas**. Por exemplo, usualmente não se separa palavras como “sexta-feira”, que correspondem a um substantivo único. No entanto, é usual tokenizar palavras como “sinto-me” em três *tokens*: o verbo “sinto”, o sinal de pontuação “-” e o pronome “me”. Mais uma vez a aplicação pretendida é que definirá o que deve ser feito ou não.

A tokenização, assim como o sentenciamento, é um processo que pode ser resolvido com estratégias baseadas em regras ou que utilizam as mesmas abordagens de aprendizado de máquina supervisionadas e não supervisionadas citadas anteriormente. A complexidade do processo de tokenização é, no entanto, menor que a do processo de sentenciação, pois o processo pode ser auxiliado pela existência de recursos léxicos que facilitam bastante a tarefa de identificar os limites possíveis da maioria dos *tokens* a serem separados.

Na Seção 4.3.1, apresentaremos os tokenizadores (ou também chamados *tokenizers*) disponíveis para o português.

[6] https://portulanclarin.net/workbench/lx-tokenizer/

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

4.2.3 Tokenização em Subpalavras

Outro conceito relacionado ao de unidade de processamento que se tornou bastante popular nos últimos tempos (principalmente com o surgimento das arquiteturas neurais para processamento da língua) é o de **subpalavra** (em inglês, *subword*). As aplicações recentes de modelos de linguagem baseados em redes neurais tornaram bastante comum a quebra de palavras em porções eventualmente menores, as subpalavras. Esse processo de **tokenização em subpalavras** tem por objetivo reduzir o vocabulário de trabalho de um modelo de linguagem a um tamanho finito, mas que possa ser usado para representar textos onde o número de *types* (quantidade de *tokens* distintos) seja potencialmente infinito. Dessa forma, pode-se utilizar um conjunto de treino que possua um vocabulário finito e que seja capaz de ser aplicado a qualquer texto[7].

Por exemplo, um modelo de linguagem pode ser projetado para ter um vocabulário de trabalho com 30.000 *types*, ou seja, um índice numérico de 1 a 30.000 único para cada *type* a ser representado de forma que se possa definir um vetor de 30.000 posições, em que cada posição corresponde a um dos *types* considerados. Essa abordagem é necessária para utilizar representações vetoriais como, por exemplo, as representações word2vec (Mikolov et al., 2013a) e GloVe (Pennington et al., 2014) a serem tratadas no Capítulo 10. Apesar dos vocabulários de trabalho serem usualmente grandes, esses vocabulários são insuficientes para representar textos em português associando uma possível palavra da língua a cada *type*, já que o português possui mais de 800.000 palavras, sem contar novas palavras que podem ser criadas, como o exemplo "cloroquiner" citado na Seção 4.1.6.

A abordagem de tokenização em subpalavras consiste em codificar diretamente algumas palavras mais comuns, como "de", "fazer", "são" e "feliz". No entanto, palavras mais raras, como "desfazer" ou "felizmente" podem ficar fora do vocabulário de trabalho (OOV do termo em inglês *out-of-vocabulary*) e portanto serem representadas como combinações de subpalavras, respectivamente: "de" + "s" + "fazer" e "feliz" + "mente".

Dessa forma, uma subpalavra pode ser uma sequência qualquer de caracteres ou podem ser sequências com algum significado linguístico como prefixos e sufixos (e.g., "mente" comumente designa advérbios, como "felizmente"), mas até letras únicas de forma que sempre seja possível representar qualquer palavra pela composição de subpalavras pertencentes ao vocabulário (e.g., a subpalavra "s", que é uma letra única, pode ser utilizada para, junto com as subpalavras "de" e "fazer", ser articulada para representar "de" + "s" + "fazer").

A escolha do vocabulário de trabalho, ou seja, a escolha das subpalavras que o compõem, pode ser feita utilizando diversas técnicas, mas três algoritmos são frequentemente utilizados: BPE (*Byte-Pair Encoding*) (Sennrich et al., 2016), Word-Piece (Schuster; Nakajima, 2012) e Unigram (Kudo, 2018). O primeiro, BPE, é inspirado em técnicas de compreensão de dados e busca representar como subpalavras os *tokens* mais frequentes. O segundo, Word-Piece, é utilizado pelo BERT (Devlin et al., 2019) e busca reduzir o tamanho do vocabulário através da escolha de subpalavras que possam ser utilizadas em um número maior de *tokens*. O terceiro, Unigram, também foca na redução do vocabulário e inicializa o treinamento com um vocabulário grande de caracteres, subpalavras e palavras, e vai reduzindo esse vocabulário mantendo somente os itens mais relevantes até alcançar o

[7] Geralmente, um modelo computacional é treinado em um conjunto de dados e usado em outros conjuntos diferentes. Nesse sentido, palavras que nunca foram vistas no conjunto de dados de treinamento são consideradas palavras desconhecidas (*unknown*) nos novos conjuntos. A existência de uma palavra desconhecida faz com que o modelo não saiba como tratá-la, o que pode impactar (e até mesmo inviabilizar) o processamento.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

tamanho de vocabulário desejado.

4.2.4 Normalização

Por sua vez, a **normalização** é a tarefa que converte as palavras para alguma forma padrão. São exemplos de normalização: conversão de versões abreviadas de palavras (e.g., conversão de "vc" para "você"), conversão para caracteres minúsculos (e.g., convertendo "Você" para "você"), lematização (e.g., estabelecendo que "somos" é uma conjugação do verbo "ser") e radicalização (e.g., estabelecendo que "retrabalho" tem o radical "trabalho"[8] precedido do prefixo "re"). De acordo com a abordagem utilizada, diferentes tipos de normalização podem ser necessários para tratar de maneira mais eficiente o processamento textual.

Na verdade, o propósito do tratamento computacional tem muita influência nos tipos de normalização que precisam ser feitos. Em alguns casos, certas informações no texto a ser processado podem ser vistas como relevantes, enquanto outras como apenas ruído. Por exemplo, quando temos como propósito identificar o sentido geral de um texto, a conversão de abreviaturas tende a auxiliar a identificação do conteúdo considerando, por exemplo, "PLN" e "Processamento de linguagem natural" como equivalentes. Por outro lado, com o propósito de identificar entidades nomeadas, a conversão para caracteres minúsculos pode dificultar o processamento.

A tarefa de **conversão de abreviações** é usualmente baseada em listas predefinidas de abreviações comuns que podem ser utilizadas em um processo de busca e substituição. No entanto, alguns cuidados usuais devem ser tomados para que somente *tokens* completos e com o sentido apropriado sejam substituídos. Por exemplo, uma abreviação usual em textos de mensagens é a representação "rs" como abreviação de "risos", mas a substituição desta abreviação, que está correta na sentença do Exemplo 4.11, ficaria completamente errada na sentença do Exemplo 4.12, pois poderia resultar em "Os youtube**risos** do RS tem muito sotaque." que substituiria erroneamente o final da palavra "youtubers".

Exemplo 4.11:

> Eu já sabia... rs.

Exemplo 4.12:

> Os youtubers do RS tem muito sotaque.

A tarefa de **conversão para caracteres minúsculos** é menos delicada, podendo ser facilmente implementada por simples processamento da representação individual dos caracteres. No entanto, mesmo nesse caso é necessário tomar certas precauções com nomes próprios e outras representações onde existe semântica associada ao uso de maiúsculas e minúsculas. Por exemplo, se fizermos a substituição de maiúsculas por minúsculas da frase do Exemplo 4.12, podemos erroneamente descaracterizar a denominação do estado do Rio Grande do Sul ("RS"), tornando o texto produzido ambíguo com a abreviação "rs". Apesar disso, o processo de conversão para caracteres minúsculos é uma tarefa usual, pois permite a consulta a recursos linguísticos de uma forma mais eficiente.

[8] Vale lembrar que o conceito de "radical" não inclui a vogal temática. No exemplo de "trabalho", o radical correto seria "trabalh" e o "o" seria uma vogal temática, portanto a forma "trabalho", na verdade, é o tema (e não o radical), porém essa distinção é muito específica e desnecessária para o PLN. Assim, a tarefa de radicalização, em PLN, pode considerar ora o radical ou a raiz, ora o tema.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

A tarefa de **lematização** envolve, frequentemente, a consulta a recursos linguísticos que possuem a definição de lemas e morfologia das palavras, como, por exemplo, um léxico da língua (Seção 4.3.2). O grande desafio desta tarefa é a desambiguação sintática das palavras que, segundo seu uso, podem ter lemas distintos. Por exemplo, na sentença do Exemplo 4.13 a palavra "casa" na sua primeira ocorrência terá como lema o verbo "casar", enquanto a segunda ocorrência terá como lema o substantivo "casa". Nesses casos é importante desambiguar morfossintaticamente as palavras (Lopes et al., 2023).

Exemplo 4.13:

> Quem casa, quer casa.

Outra abordagem utilizada é através de padrões, a fim de trazer a palavra para sua forma canônica (e.g., trazer substantivos para o masculino singular e todas as flexões do verbo para sua forma no infinitivo) (Bertaglia; Nunes, 2016). Note-se que essa abordagem requer sofisticações para tratar palavras que não têm comportamento regular. Por exemplo, se a palavra "meninas" pode ser trazida corretamente ao lema substituindo a terminação "as" por "o" resultando no lema "menino", a palavra "casas" seria erroneamente lematizada para o lema "caso" se utilizássemos o mesmo princípio.

A tarefa de **radicalização** tem o propósito de converter lexemas para seus radicais. Uma particular vantagem deste tipo de tarefa é uniformizar e diminuir o vocabulário, ainda que possam levar à perda de informação. No entanto, em algumas aplicações, a busca dos radicais de uma palavra pode auxiliar, inclusive, no estabelecimento de subpalavras que foi descrito na Seção 4.2.3. Por exemplo, as palavras "certo", "certidão", "incerto", "certamente", "certificação", "certeiro" e "incerteza" possuem o mesmo radical "cert" e, portanto, tornam "cert" um bom candidato a subpalavra. Algoritmos de radicalização (*stemming*, em inglês) podem ser encontrados em bibliotecas usuais da área de PLN, como NLTK (Bird; Loper, 2004), que oferecem opções para várias línguas, inclusive para o português.

Na Seção 4.3.1, apresentaremos algumas ferramentas disponíveis para o português que fazem a normalização dos textos, tais como lematizadores, radicalizadores, *stemmers* e outros.

4.2.5 PoS *tagging*

O PoS *tagging*, também conhecido como etiquetagem morfossintática, é uma técnica fundamental na área de PLN que envolve a atribuição de etiquetas gramaticais a cada palavra em um texto, com base na sua classe gramatical e em suas características morfológicas. Essas etiquetas ajudam a identificar a função sintática e morfológica das palavras em uma sentença, o que é crucial para a posterior análise sintática, mas também é útil para outras aplicações, como tradução automática (Capítulo Tradução Automática), análise de sentimentos, geração de resumos, entre outras.

As classes de palavras são universais e valem para a grande maioria das línguas naturais, incluindo o português. São elas: substantivos, verbos, adjetivos, advérbios, pronomes, numerais, artigos, conjunções, preposições e interjeições. Porém, as etiquetas de PoS que cada modelo de anotação define podem ser diferentes, assim como pode haver diferentes níveis de granularidade das etiquetas. Nesse sentido, um *tagger* (nome dado a uma ferramenta computacional que realizada a etiquetagem morfossintática) pode usar apenas as etiquetas de granularidade grossa (em inglês, *coarse tags* ou UPoS), que correspondem às

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

classes de palavras acima, enquanto outros *taggers* podem usar um conjunto de etiquetas de granularidade fina (em inglês, *fine-grained tags* ou XPoS). Por exemplo, os adjetivos costumam ser associados à *coarse tag* ADJ, mas também podem ser etiquetados como JJ (para adjetivos primitivos), JJR (para adjetivos comparativos) ou JJS (para adjetivos superlativos). Horsmann; Zesch (2016) propõem uma abordagem que combina os dois níveis de anotação para aumentar a precisão do *tagger*.

A Figura 4.2 mostra um exemplo de etiquetagem morfossintática realizada pelo *parser* LXUTagger[9].

Figura 4.2: Exemplo de anotação de PoS com *coarse tags*



Cada palavra é analisada quanto à sua forma e função na sentença. É necessário fazer essa análise dependente de contexto porque existem palavras polissêmicas, ambíguas, homônimas etc. Dentro do contexto, é possível, por exemplo, distinguir o substantivo "ajuda" (e.g. "Ele recebeu *ajuda* para executar o trabalho.") do verbo "ajuda" (e.g. "Sempre que precisa, ele *ajuda* a comunidade.").

Várias abordagens são possíveis e usuais para esse tipo de tarefa, desde a anotação manual das etiquetas e posterior treinamento de modelo, até o uso de redes neurais. As mais conhecidas ainda hoje são as baseadas em regras (Brill, 1992), em redes neurais artificiais (Schmid, 1994), as abordagens estocásticas (Hall, 2003) e as híbridas (Altunyurt et al., 2006; Zin, 2009).

A seguir, explicamos brevemente algumas dessas abordagens, sumarizando o trabalho de Zewdu; Yitagesu (2022), que fizeram recentemente uma revisão sistemática da literatura sobre as técnicas usadas para etiquetagem de PoS. Ressaltamos, no entanto, que, para uma visão mais aprofundada, recomendamos ler o artigo original.

- **Abordagem baseada em regras** – Utiliza regras criadas manualmente para atribuir *tags* às palavras de uma sentença. Essas regras podem ser: (i) construídas por especialistas linguistas e, assim, depender de características linguísticas, como informações lexicais, morfológicas e sintáticas, ou (ii) inferidas via aprendizado de máquina, a partir de uma grande quantidade de dados (Capítulo 16), dispensando regras especializadas. A primeira abordagem (geração manual de regras) é demorada e propensa a erros, enquanto a segunda (com uso de aprendizado de máquina em um *corpus* anotado) é mais eficiente, embora ainda exija *expertise* linguística.
- **Hidden Markov Models (HMM)** – É uma das abordagens mais amplamente utilizadas para PoS *taggers* que usam modelos estocásticos (Kumawat; Jain, 2015; Zin, 2009). Nessa abordagem, o *tagger* passa de um estado a outro por meio de um estado oculto. Esse estado oculto não é diretamente observável, mas a saída dependente do estado oculto é visível. O algoritmo de Viterbi é um método bastante conhecido para identificar a sequência mais provável de *tags* T={t1, t2, t3… tn} para cada palavra em uma sentença W={w1, w2, w3…wn} ao usar um modelo oculto de Markov.
- **Aprendizado de máquina baseado em *features*** – Os algoritmos de Aprendizado de Máquina mais comuns usados para PoS *taggers* são as redes neurais, *Naïve Bayes*,

[9] https://portulanclarin.net/workbench/lx-utagger/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

HMM, *Support Vector Machine* (SVM), *Conditional Random Field* (CRF), Brill e TnT. Para maiores informações sobre como cada um deles é usado para a tarefa de PoS *tagging*, ver Zewdu; Yitagesu (2022).

- **Aprendizado profundo** – É uma abordagem intensiva em dados baseada no uso de redes neurais artificiais em vários níveis. Nessa abordagem é usual realizar um pré-processamento dos dados (Dhumal Deshmukh; Kiwelekar, 2020). A saída do pré-processamento é, então, usada como entrada para a primeira camada da rede neural. A partir dessa entrada pré-processada, a rede neural utiliza-se de um modelo de linguagem pré-treinado para atribuir a etiqueta PoS mais provável a cada palavra. Cabe salientar que o treinamento desse modelo de linguagem se faz através de um conjunto de treino que alimenta a rede neural repetidas vezes recalculando os pesos de cada camada da rede. Os métodos sequenciais de aprendizado profundo mais comuns para etiquetagem de PoS são: FNN, MLP, GRU, CNN, RNN, LSTM e BLSTM. Não pretendemos detalhar todos eles aqui, mas indicamos o trabalho de Zewdu; Yitagesu (2022), em que todos eles são descritos em pormenores.

Após a etapa de PoS *tagging* para cada uma das palavras de uma sentença, elas passam então para outra etapa do processamento que também pertence à área da Morfologia, que é a atribuição das *features* (ou atributos) morfológicos, e que serão explorados nas próximas seções.

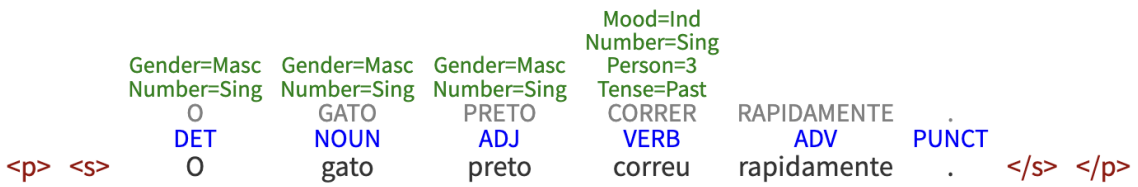
4.2.6 Anotação de atributos morfológicos

A anotação de atributos morfológicos, também conhecida como atribuição de *features* morfológicas (ou somente *feats*), é uma importante tarefa que envolve a marcação ou identificação de informações específicas sobre as características gramaticais e morfológicas de palavras em um texto. Esses atributos morfológicos incluem características como número, gênero, modo, tempo, pessoa e outras informações semelhantes.

O objetivo principal dessa anotação é capturar e codificar informações gramaticais relevantes de maneira estruturada, tornando possível para algoritmos de PLN e de AM entenderem e processarem corretamente a estrutura e as relações linguísticas presentes em um texto. Por exemplo, considere a palavra "falávamos". A anotação de atributos morfológicos, nesse caso, pode envolver a marcação de tempo (pretérito imperfeito), modo (indicativo), número (plural), pessoa (primeira) e forma verbal (finita).

No exemplo da Figura 4.2 – "O gato preto correu rapidamente" –, a anotação das *features* morfológicas seria conforme apresentado na Figura 4.3, gerada pela ferramenta LX-USuite.[10]

Figura 4.3: Exemplo de anotação de *features* morfológicas



Cabe a cada *tagger* (ou etiquetador morfológico) definir seu próprio conjunto de *features* ou características morfológicas que devem ser anotadas ou extraídas, o

[10] https://portulanclarin.net/workbench/lx-usuite/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

que é chamado de *tagset* (ou conjunto de etiquetas). Por exemplo, o UDPipe, que será abordado na Seção 4.3.1, identifica, para verbos, todas as *features* citadas anteriormente: para o verbo "comprei", lemma = "comprar" e *features* = "VERB _ Mood=Ind|Number=Sing|Person=1|Tense=Past|VerbForm=Fin".

A quantidade de etiquetas, bem como seus nomes, pode variar bastante de um *tagger* para outro, e essa falta de uniformização constitui um grande desafio para o PLN. De acordo com Fonseca et al. (2015b), o Penn Treebank do inglês, por exemplo, conta com 45 *tags* incluindo sinais de pontuação, enquanto o CLAWS5 possui 62 e o CLAWS7 possui 137 *tags*, sendo que os dois últimos usaram o mesmo *corpus* (British National Corpus).

Se o objetivo da tarefa for treinar um etiquetador morfossintático, a anotação pode ser feita manualmente por especialistas humanos que analisam e selecionam os atributos morfológicos para as palavras em um *corpus* de treinamento (Capítulo 16). Posteriormente, modelos de PLN e AM podem ser treinados usando esses dados anotados para automatizar a anotação de novos textos e auxiliar em várias tarefas e aplicações. Por outro lado, já existem também abordagens computacionais semelhantes às mencionadas na Seção 4.2.5, que também podem ser usadas para etiquetar automaticamente as palavras em relação às suas características morfológicas.

Além das tarefas de processamento aqui apresentadas, vale mencionar que há ainda outras, como a análise sintática automática (tarefa de *parsing*), a segmentação dos constituintes sintáticos dentro da frase (tarefa de *chunking*), a extração ou anotação de entidades nomeadas (tarefa de *Named Entity Recognition*) etc. O que nos interessa neste Capítulo são as tarefas ligadas ao processamento morfológico e morfossintático, portanto nos limitamos às mencionadas nesta seção.

4.3 Ferramentas e recursos para o processamento morfológico

Esta seção apresenta ferramentas computacionais e recursos disponíveis para o português, que processam a língua no nível morfológico e morfossintático. Para cada uma das tarefas de (pré-)processamento, existem ferramentas específicas que podem ser usadas para a análise automática do português. Dividimos esta seção em ferramentas (Seção 4.3.1) e recursos (Seção 4.3.2). Ao final da seção, selecionamos um dos recursos para explicá-lo de forma mais aprofundada, a saber, o PortiLexicon (Seção 4.3.3).

4.3.1 Ferramentas para o português

Existem ferramentas que têm como foco tarefas específicas de PLN. Outras, como NLTK[11] e spaCy[12], são módulos mais completos que implementam submódulos e funcionalidades diversas. Apresentamos a seguir ferramentas disponíveis para o processamento de textos, separadas por tarefa.

Tokenizadores

- O NLTK contém o `nltk.tokenize`[13], que é um submódulo que implementa diferentes classes e funções para a tarefa de tokenização, como: tokenização por espaço em branco, baseada em pontuação, considerando expressões multipalavra (Capítulo 5),

[11] https://www.nltk.org
[12] https://spacy.io/
[13] https://www.nltk.org/api/nltk.tokenize.html

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

entre outros. Também são implementadas funcionalidades para a personalização da tokenização, por exemplo, baseada em expressões regulares definidas pela pessoa usuária.

- O **tokenizador do spaCy**[14] (ou *tokenizer*) segmenta o texto em palavras e sinais de pontuação. Isso é feito aplicando regras específicas para cada idioma; normalmente é criado automaticamente quando uma subclasse *Language* é inicializada e lê suas configurações, como pontuação e regras de casos especiais. As exceções do *tokenizer* definem casos especiais como "don't" em inglês, que precisa ser dividido em dois *tokens*: {ORTH: "do"} e {ORTH: "n't", NORM: "not"}. Os prefixos, sufixos e infixos definem principalmente regras de pontuação – por exemplo, quando separar pontos (no final de uma frase) e quando manter o ponto pertencendo ao *token* (como nos casos de "Dr.", "25.000" e o fatorial "5!", apresentados na Seção 4.2.1).

Lematizadores

- O **lematizador do spaCy**[15] (ou *lemmatizer*) converte as palavras para suas formas básicas (lemas) usando regras baseadas em *tags* de classe gramatical ou tabelas de pesquisa. Diferentes subclasses de linguagem podem implementar seus próprios componentes lematizadores por meio de abordagens específicas para cada língua. Isso torna mais fácil personalizar como os lemas devem ser atribuídos ao seu *pipeline*.

Stemmers

- O submódulo `nltk.stem` implementa diferentes abordagens de *stemming*, incluindo:
  - O **RSLP**[16] (Removedor de Sufixos da Língua Portuguesa), inicialmente proposto por Orengo; Huyck (2001) como um algoritmo que realiza as seguintes etapas: redução do plural (supressão do "-s"), redução de feminino (supressão do "-a"), redução de advérbio (remoção do sufixo "-mente"), redução dos sufixos de aumentativo ("-ão") e diminutivo ("-inho"), redução de sufixos nominais (como "-mento", "-ção", "-ncia"), redução de desinências verbais, remoção de vogais e de acentos gráficos.
  - O **SnowballStemmer**, baseado no algoritmo proposto por Porter (1980) para o inglês e estendido para a linguagem Snowball[17] de processamento de *strings* para criação de algoritmos de *stemming* para diversas línguas, também está disponível para o português.
  - Uma classe para a personalização de *stemming* baseada em expressões regulares.

Sentenciadores

- **Punkt** é uma ferramenta para a segmentação de sentenças em português, ou seja, é um sentenciador. O NLTK inclui um modelo do Punkt[18], que foi treinado inicialmente a partir do *corpus* Floresta Sintá(c)tica (Afonso et al., 2002), mas pode ser retreinado e avaliado em outros *corpora*, lembrando que é mais rápido carregar um modelo treinado do que treiná-lo novamente.

[14] https://spacy.io/api/tokenizer
[15] https://spacy.io/api/lemmatizer
[16] https://www.inf.ufrgs.br//~viviane/rslp/index.htm
[17] https://snowballstem.org/
[18] https://www.nltk.org/howto/portuguese_en.html#sentence-segmentation

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

- No spaCy, a sentenciação é feita conforme a anotação de dependência é executada, por exemplo, por uma instância da classe *DependencyParser*. No entanto, caso um *parser* não seja incluído no *pipeline*, o **sentenciador do spaCy**[19] (ou *sentencizer*), que permite a detecção dos limites das sentenças com base em regras, pode ser incluído como componente no *pipeline*. O *sentencizer* é bem simples para permitir lógica de detecção de limite de frase personalizada que não requer análise de dependência, possibilitando a implementação de uma estratégia mais simples e baseada em regras que não requer o carregamento de um modelo estatístico.

Normalizadores

- A **Enelvo**[20] é uma biblioteca que pode ser usada para normalizar textos em português (Bertaglia; Nunes, 2017). Conforme apresentado na Seção 4.2.4, a etapa de normalização pode envolver várias microtarefas. No caso da Enelvo, ela corrige abreviações, gírias, erros ortográficos, capitaliza letras no começo das frases, de nomes próprios e acrônimos. Ela também possui uma função própria para remover pontuações e emojis.

Taggers

São chamadas de *taggers* as ferramentas de etiquetagem morfossintática e morfológica, ou seja, as ferramentas computacionais que atribuem automaticamente o PoS e as etiquetas morfológicas para cada palavra em uma frase. Antes de mencionar os *taggers* propriamente ditos para o português, convém citar o trabalho de Gonçalves et al. (2020), no qual os autores avaliaram várias ferramentas de etiquetagem morfossintática para o português, como: FreeLing[21], NLTK[22], OpenNLP[23], NLPyPort[24], PolyGlot[25], spaCy[26], StanfordNLP[27], TreeTagger[28] e LinguaKit[29]. Algumas delas fazem apenas o processamento do português europeu, mas a maioria processa as duas variantes do português.

Além dos PoS *taggers* indicados no parágrafo acima, também estão disponíveis para o português:

- O **Porttagger** é um etiquetador morfossintático (PoS *tagger*) multigênero para o português brasileiro treinado com um conjunto de dados variados, composto por notícias jornalísticas (de vários domínios), textos acadêmicos (do domínio de óleo e gás) e de conteúdo gerado por usuário (tweets do domínio do mercado financeiro), considerados *gold standard*, pois foram revisados por humanos. As etiquetas utilizadas são as 17 etiquetas do modelo gramatical Universal Dependencies[30] (Marneffe et al., 2021). O etiquetador realiza a classificação com base em um modelo de linguagem (o BERTimbau (Souza et al., 2020)), que foi implementado usando a biblioteca

[19] https://spacy.io/api/sentencizer
[20] https://github.com/thalesbertaglia/enelvo ou https://thalesbertaglia.com/enelvo/sobre/
[21] https://nlp.lsi.upc.edu/freeling/index.php/
[22] https://www.nltk.org
[23] https://opennlp.apache.org/
[24] https://github.com/NLP-CISUC/NLPyPort
[25] https://draquet.github.io/PolyGlot/
[26] https://spacy.io/
[27] https://stanfordnlp.github.io/stanfordnlp/
[28] https://www.cis.uni-muenchen.de//~schmid/tools/TreeTagger/
[29] https://github.com/citiususc/Linguakit
[30] https://universaldependencies.org

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

transformers[31]. O Porttagger possui uma acurácia do estado da arte, superando os resultados anteriores para a língua portuguesa. Para os textos jornalísticos, por exemplo, atinge valores acima de 99% de acurácia (Silva et al., 2023). O sistema está disponível para uso do público em geral[32].

- O ***tagger* do UDPipe**[33] consiste em vários modelos MorphoDiTa[34] que anotam PoS *tags* e/ou lemas. Por padrão, esse *tagger* costuma usar apenas um modelo, que gera todas as *tags* disponíveis (*coarse tags*, *fine-grained*, *features* morfológicas e lema). Porém, o desempenho melhora se um modelo apenas etiquetar as palavras (com UPoS, XPoS e *features* morfológicas), enquanto o outro modelo realiza a lematização. Também é possível especificar a quantidade de modelos que se quer utilizar. Vale ressaltar que o conjunto de etiquetas do UDPipe *tagger* é o mesmo das Universal Dependencies (Marneffe et al., 2021) e o formato de anotação dos dados é o mesmo do CONLL-U. Além do *tagger*, o UDPipe também dispõe de tokenizador, lematizador e *parser*, todos configuráveis para uso em modelos disponíveis, também sendo possível treinar novos modelos.
- O ***tagger* do Stanza**[35] também realiza anotação morfossintática e morfológica. No primeiro tipo, ele etiqueta as palavras em dois níveis: de granularidade grossa (UPoS) e de granularidade fina (XPoS). No segundo tipo, ele etiqueta as palavras a partir de *features* morfológicas (*UFeats*). O Stanza é um pacote Python para análise e processamento das línguas naturais que possui, além do *tagger*, várias outras ferramentas de PLN, como tokenizador, lematizador, *parser* de dependência, reconhecedor de entidades nomeadas e outras. O Stanza dispõe de modelos pré-treinados, inclusive com *treebanks* da UD.
- O ***tagger* do spaCy**[36] é um componente do *pipeline* que prediz as etiquetas de PoS a partir de qualquer *tagset*. Isso é muito útil porque permite que se use como dado de treinamento qualquer *corpus* anotado, sem precisar converter as etiquetas de um modelo para outro. A predição é feita a partir de um *pipeline* treinado com modelos estatísticos, que permitem ao spaCy fazer previsões sobre qual *tag* ou etiqueta provavelmente se aplica num dado contexto. Um componente treinado inclui dados binários que são produzidos mostrando a um sistema exemplos suficientes para que ele faça previsões generalizáveis para todo o idioma – por exemplo, uma palavra que vem depois de "o" em português provavelmente é um substantivo. Para a atribuição das *features* morfológicas, o spaCy dispõe de dois módulos: um estatístico, que atribui as *tags* de classe gramatical de granularidade grossa (UPoS); e um baseado em regras, principalmente para o caso de línguas com morfologia mais simples, como é o caso do inglês.

4.3.2 Recursos para o português

Nesta seção, indicaremos alguns recursos lexicais (principalmente *corpora*) anotados com informação morfológica e morfossintática disponíveis para o português. Não pretendemos aqui explorar todos os recursos disponíveis, mas apenas mencionar alguns exemplos, considerando que o Capítulo 16 aborda *corpus* e *datasets*. Ao final da seção, exploraremos

[31] https://github.com/huggingface/transformers
[32] https://huggingface.co/spaces/Emanuel/porttagger
[33] https://cran.r-project.org/web/packages/udpipe/vignettes/udpipe-train.html
[34] https://ufal.mff.cuni.cz/morphodita
[35] https://stanfordnlp.github.io/stanza/pos.html
[36] https://spacy.io/api/tagger

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

de forma aprofundada apenas um deles, a fim de exemplificar e deixar claros todos os conceitos e tarefas apresentados neste capítulo.

Os *corpora* mais amplamente conhecidos na comunidade do português e que contenham anotação de PoS são o Tycho Brahe (Namiuti, 2004), o Mac-Morpho e o Bosque. O primeiro deles é um *corpus* do português histórico europeu, que compilou textos literários antigos (do século XIV ao XIX). Por serem de um gênero e domínio muito específicos, além de conter palavras e estruturas arcaicas, ele não é um *corpus* muito representativo do português contemporâneo. Os outros dois (Mac-Morpho e Bosque) serão apresentados na sequência desta seção. Além deles, indicaremos também outros *corpora* mais recentes do português que contenham anotação morfossintática. Alguns deles serão retomados no Capítulo 7.

- O **Mac-Morpho**[37] (Aluísio et al., 2003) é um *corpus* em português que contém 915.367 *tokens*[38] retirados de textos jornalísticos da Folha de São Paulo, já etiquetadas com suas respectivas classes de palavras. É considerado um dos maiores *corpora* do português contemporâneo e foi anotado manualmente com PoS *tags*. Seu *tagset* original conta com 41 etiquetas de PoS, sendo 19 delas somente para marcas de pontuação. Além de ser um *corpus* anotado, ele também foi usado para treinar várias versões de *taggers* usados pela comunidade do português (Fonseca et al., 2015b).
- O **Bosque** é uma parte de um *corpus* maior, o **Floresta Sintá(c)tica**[39] (Afonso et al., 2002), mas que contém anotação de PoS, de *features* morfológicas e também de *treebanks* (anotação sintática em árvore de dependência). O Bosque possui cerca de 200.000 *tokens* e é composto por textos jornalísticos das duas variantes do português, provenientes do CETENFolha (português brasileiro) e CETENPúblico (português europeu).
- O **Porttinari**[40] (Pardo et al., 2021) é um compilado de vários *corpora* de diferentes gêneros (jornalísticos, tweets de ações e bolsas de valores, reviews de produtos de e-commerce, resenhas de livros, entre outros). A versão atualmente disponível desse *corpus* é a porção jornalística (Duran et al., 2023) dividida em três *corpora*: (i) Porttinari-base, (ii) Porttinari-check e (iii) Porttinari-automatic. O **Porttinari-base** contém 8.418 sentenças anotadas manualmente com informação morfológica e morfossintática seguindo as diretrizes do Universal Dependencies (UD) (Marneffe et al., 2021). O **Porttinari-check** é um pequeno *corpus* de 1.685 sentenças com características similares ao Porttinari-base, porém anotado automaticamente usando como conjunto de treino o Portinari-base para ilustrar o contraste com a anotação manual. O **Porttinari-automatic**, é um grande *corpus* com 3.954.189 sentenças que foi anotado automaticamente também usando o mesmo processo.
- O **MorphoBr**[41] (Alencar et al., 2018) é um léxico construído com o propósito de subsidiar de forma abrangente a análise morfológica do português. Esse léxico combina recursos livres semelhantes, corrigindo milhares de erros e lacunas dos demais recursos, e aumentando de forma significativa o número de entradas, principalmente em relação aos substantivos e adjetivos no diminutivo.

[37] http://nilc.icmc.usp.br/macmorpho/
[38] Dados da última versão (v3), disponível em http://nilc.icmc.usp.br/macmorpho/.
[39] https://www.linguateca.pt/Floresta/
[40] https://sites.google.com/icmc.usp.br/poetisa/resources-and-tools
[41] https://github.com/LR-POR/MorphoBr

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Nos repositórios do NILC[42], POeTiSA[43] e Linguateca[44] você pode ter acesso a vários outros recursos para o português, tais como *corpora*, léxicos e outros recursos lexicais com anotações morfológica e morfossintática. No Capítulo 16 também serão indicados *corpora* relevantes para o português, mas que não necessariamente contêm informação morfológica ou morfossintática. Para esclarecer de forma bem detalhada como funcionam esses recursos, a anotação de PoS e de *feats*, as etiquetas etc., selecionamos um deles, o PortiLexicon-UD, para ser apresentado em detalhes na próxima seção (Seção 4.3.3), a fim de exemplificar os conceitos abordados ao longo do capítulo.

4.3.3 PortiLexicon-UD: um recurso para processamento léxico em português

Para ilustrar o uso, na prática, de um recurso para processamento lexical, esta seção apresenta o PortiLexicon-UD (Lopes et al., 2022). O PortiLexicon-UD é um recurso que auxilia na tarefa de identificar as unidades de processamento. Ele é um léxico para o português que elenca palavras e suas anotações morfossintáticas. Especificamente, esse léxico utiliza o padrão Universal Dependencies (UD)[45] (Marneffe et al., 2021) com etiquetas PoS, lema e etiquetas de atributos morfológicos (gênero, número etc.). Na sua versão atual[46], PortiLexicon-UD possui 1.226.339 entradas.

O conjunto de etiquetas PoS da UD define **17 classes gramaticais** descritas no Quadro 4.1 com alguns exemplos de utilização no português segundo o mapeamento adotado. Todo mapeamento do português para o padrão UD foi feito durante a construção do *corpus* Porttinari-base (Duran et al., 2023). Dessa forma, o mapeamento reflete as decisões tomadas na anotação deste *corpus*.

Quadro 4.1: Etiquetas PoS da UD e sua descrição no português

PoS	Descrição e Exemplos
ADP	**adposições**, uma classe fechada que corresponde às preposições em português, como “de”, “para” e “com”.
ADJ	**adjetivos**, uma classe aberta que inclui palavras como “bonitas”, “último” e “vermelha”, mas que inclui também um subconjunto fechado, os números ordinais escritos por extenso como as palavras “primeiro”, “centésima” e “duodécimo”. No entanto, a anotação UD aceita também anotar como ADJ subconjuntos abertos como os números ordinais expressos em dígitos como “20º” ou “13ª” ou ainda todas formas de verbos no particípio como “cansado” e “cedido”.

[42] https://sites.google.com/view/nilc-usp/resources-and-tools?authuser=0
[43] https://sites.google.com/icmc.usp.br/poetisa/the-project
[44] https://www.linguateca.pt
[45] Para um estudo mais aprofundado do projeto Universal Dependencies, ver Capítulo 4.
[46] O PortiLexicon-UD está disponível no endereço: https://portilexicon.icmc.usp.br/ e lá também pode ter seus arquivos de dados baixados.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

ADV	**advérbios**, uma classe aberta com um subconjunto fechado, os advérbios primitivos (aqueles não formados com o sufixo "-mente"). Exemplos deste subconjunto fechado são as palavras "cedo", "agora" e "acima". Exemplos do subconjunto aberto são as palavras "normalmente" e "insanamente". Também são incluídas formas abreviadas destes advérbios que aparecem em expressões como "social e economicamente" onde o advérbio "social" é usado como uma forma abreviada do advérbio "socialmente".
AUX	**verbos auxiliares e de cópula**, uma classe fechada no português que engloba todas as conjugações dos verbos "ser", "estar", "haver", "ir,"ter" e "vir".
CCONJ	**conjunções coordenativas**, uma classe fechada que contém, por exemplo, "e", "mas" e "portanto".
DET	**determinantes**, uma classe fechada que inclui artigos como "um" e "o", além de palavras específicas como "cujo", mas que também tem uma grande sobreposição com pronomes como "aquele", "diversos" e "meus".
INTJ	**interjeições**, uma classe aberta que inclui, por exemplo, "tchau", "oi" e "nossa".
NOUN	**substantivos**, uma classe naturalmente aberta que inclui, por exemplo, "presidente", "quartos", "bandeirinha", "salões" e "bola".
NUM	**números cardinais**, uma classe aberta que possui um subconjunto fechado, os números cardinais escritos por extenso como "duas", "trinta" e "quinhentos", mas também um subconjunto aberto com toda sorte de números escritos com dígitos como "51", "-3.1415" e até datas como "25/12/1974".
PART	**partículas**, classe não mapeada na construção do Portinari-base.
PRON	**pronomes**, uma classe fechada que possui alguma sobreposição com determinantes como as palavras "sua", "qual" e "ambos", mas também palavras que não podem ser usadas como determinantes como "eu", "que" e "aquilo".
PROPN	**nomes próprios**, naturalmente uma classe aberta, tipicamente denominação de entidades nomeadas como "Obama" ou "Paris".
PUNCT	**todas as pontuações**, uma classe aberta, sendo as mais usuais ponto final, exclamação, interrogação, vírgula, dois pontos e aspas.
SCONJ	**conjunções subordinativas**, uma classe fechada que contém, por exemplo, "conquanto", "se" e "segundo".
SYM	**todos os símbolos simples e compostos**, uma classe aberta que inclui, por exemplo, "R$", "US$" e "%".
VERB	**verbos**, uma classe aberta que inclui todas as conjugações dos verbos plenos em português, como, por exemplo, "canta", "jogar", "chorastes" e "teríamos".
X	**tudo que não pertence ao vocabulário da língua**, naturalmente uma classe aberta, que inclui palavras estrangeiras como "détente" e "relaxation", mas ainda onomatopeias como "oinc" e "ão".

Devido à natureza dos léxicos, as palavras anotadas como pontuações (PUNCT), símbolos (SYM), além de nomes próprios (PROPN) e palavras fora do vocabulário (X) não estão presentes no PortiLexicon-UD. Levando em conta que a classe PART não é utilizada no mapeamento, o léxico possui palavras pertencentes a **12 classes da UD** que são mapeadas no português conforme sumarizado no Quadro 4.2.

Quadro 4.2: Mapeamento da classificação tradicional do português para as etiquetas PoS da UD utilizado no PortiLexicon-UD

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Português	PoS UD	Palavras
Advérbios	ADV	Todos os advérbios
Preposições	ADP	Todas as preposições
Conjunções Coordenativas	CCONJ	Todas as conjunções coordenativas
Conjunções Subordinativas	SCONJ	Todas as conjunções subordinativas
Artigos	DET	Todos os artigos
Pronomes	DET	Todos os pronomes exceto pessoais
	PRON	Todos os pronomes
Substantivos	NOUN	Todos os substantivos
Adjetivos	ADJ	Todos os adjetivos
	ADV	Todos os adjetivos que abreviam um advérbio
Interjeições	INTJ	Todas as interjeições
Numerais	NUM	Todos os cardinais exceto formas altas (“bilhão”)
	ADJ	Todos os ordinais
	NOUN	Multiplicáveis (“triplo”) e formas altas (“milhão”)
Verbos	AUX	Conjugações de “ser”, “estar”, “haver,”ir”, “ter” e “vir”
	VERB	Todos os verbos, exceto conjugações de “estar”
	ADJ	Todos os particípios

O PortiLexicon-UD foi construído a partir de um léxico pré-existente, DELAF-PB (Ranchhod et al., 1999), e suas versões UNITEX-PB (Muniz, 2004) e MorphoBr (Alencar et al., 2018). Notadamente, foi estendido o vocabulário e diversas situações foram corrigidas e adaptadas conforme descrito no lançamento do léxico (Lopes et al., 2022).

Todas as entradas no PortiLexicon-UD correspondem a uma **tupla**[47] com quatro informações:

1. Forma da palavra somente com letras minúsculas;
2. Lema da palavra somente com letras minúsculas;
3. Etiqueta PoS da UD;
4. Conjunto de pares, cada um com um atributos morfológico e seu valor associado ou, na inexistência de atributos morfológicos, utiliza-se o símbolo sublinhado (_).

Enquanto as palavras das classes ADP, ADV, CCONJ, INTJ e SCONJ usualmente contêm pouca ou nenhuma informação morfológica, outras classes possuem grande variedade de informações. Alguns exemplos de palavras destas cinco classes estão indicadas no Quadro 4.3.

Quadro 4.3: Exemplos de palavras das classes ADP, ADV, CCONJ, INTJ e SCONJ

[47] Tupla é um termo bem usado em computação, e significa uma cadeia de dois ou mais itens. Para um conjunto ordenado de elementos que são números, usamos o termo “vetor”. Mas, quando temos um conjunto ordenado no qual os elementos são outras coisas, como a cadeia de informações sobre uma dada palavra ou entrada de de um dicionário, dizemos “tupla”.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

forma	lema	PoS	atributos morfológicos
de	de	ADP	_
pra	para	ADP	Abbr=Yes
cedo	cedo	ADV	_
felizmente	felizmente	ADV	_
mas	mas	CCONJ	_
portanto	portanto	CCONJ	_
nossa	nossa	INTJ	_
oi	oi	INTJ	_
se	se	SCONJ	_
conquanto	conquanto	SCONJ	_

Já as palavras das classes ADJ, NOUN e NUM possuem uma maior variação de atributos morfológicos como pode ser visto no Quadro 4.4. Usualmente palavras destas classes possuem os atributos *Gender* (gênero) e *Number* (número), e ocasionalmente os atributos *VerbForm* (forma verbal) e *NumType* (tipo numérico) indicando a origem das palavras.

Quadro 4.4: Exemplos de palavras das classes ADJ, NOUN e NUM

forma	lema	PoS	atributos morfológicos
bonitas	bonito	ADJ	Gender=Fem|Number=Plur
primeira	primeiro	ADJ	Gender=Fem|Number=Sing| NumType=Ord
cedido	cedido	ADJ	Gender=Masc|Number=Sing| VerbForm=Part
presidente	presidente	NOUN	Number=Sing
quartos	quarto	NOUN	Gender=Masc|Number=Plur
bandeirinha	bandeirinha	NOUN	Number=Sing
salões	salão	NOUN	Gender=Masc|Number=Plur
duas	um	NUM	Gender=Fem|NumType=Card
trinta	trinta	NUM	NumType=Card
quinhentos	quinhentos	NUM	Gender=Masc|NumType=Card

As palavras etiquetadas como PRON e DET possuem uma variação ainda maior de atributos morfológicos, pois levam em consideração aspectos como o tipo de pronome e o caso. O Quadro 4.5 ilustra alguns destes casos.

Quadro 4.5: Exemplos de palavras das classes PRON e DET

forma	lema	PoS	atributos morfológicos
a	o	DET	Definite=Def|Gender=Fem|Number=Sing| PronType=Art
um	um	DET	Definite=Ind|Gender=Masc|Number=Sing| PronType=Art
cujo	cujo	DET	Gender=Masc|Number=Sing| PronType=Rel

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

esse	esse	DET	Gender=Masc|Number=Sing| PronType=Dem
minha	meu	DET	Gender=Fem|Number=Sing|Person=1| Poss=Yes|PronType=Prs
a	o	PRON	Gender=Fem|Number=Sing|Person=3| PronType=Dem
eu	eu	PRON	Case=Nom|Number=Sing|Person=1| PronType=Prs
nossas	nosso	PRON	Gender=Fem|Number=Plur|Person=1| Poss=Yes|PronType=Prs
quais	qual	PRON	Number=Plur|PronType=Rel
que	que	PRON	PronType=Int

Finalmente, as palavras que pertencem às classes AUX e VERB também possuem uma variação de atributos morfológicos relevante, porém mais padronizada, já que os atributos descrevem as conjugações dos tempos verbais. O Quadro 4.6 apresenta exemplos dos tempos verbais em português.

Quadro 4.6: Exemplos de palavras das classes AUX e VERB

forma	**lema**	**PoS**	**atributos morfológicos**
sido	ser	AUX	Gender=Masc|Number=Sing| VerbForm=Part
tava	estar	AUX	Abbr=Yes|Mood=Ind|Number=Sing| Person=3|Tense=Imp|VerbForm=Fin
vinha	vir	AUX	Mood=Ind|Number=Sing|Person=3| Tense=Imp|VerbForm=Fin
teríamos	ter	AUX	Mood=Cnd|Number=Plur|Person=1| VerbForm=Fin
cantar	cantar	VERB	VerbForm=Inf
cantarem	cantar	VERB	Number=Plur|Person=3|VerbForm=Inf
cantando	cantar	VERB	VerbForm=Ger
cantada	cantar	VERB	Gender=Fem|Number=Sing|VerbForm=Part
canto	cantar	VERB	Mood=Ind|Number=Sing|Person=1| Tense=Pres|VerbForm=Fin
cantavas	cantar	VERB	Mood=Ind|Number=Sing|Person=2| Tense=Imp|VerbForm=Fin
cantou	cantar	VERB	Mood=Ind|Number=Sing|Person=3| Tense=Past|VerbForm=Fin
cantáreis	cantar	VERB	Mood=Ind|Number=Plur|Person=2| Tense=Pqp|VerbForm=Fin
cantarão	cantar	VERB	Mood=Ind|Number=Plur|Person=3| Tense=Fut|VerbForm=Fin
cantaria	cantar	VERB	Mood=Cnd|Number=Sing|Person=1| VerbForm=Fin
cantemos	cantar	VERB	Mood=Sub|Number=Plur|Person=1| Tense=Pres|VerbForm=Fin
cantasses	cantar	VERB	Mood=Sub|Number=Sing|Person=2| Tense=Past|VerbForm=Fin
cantares	cantar	VERB	Mood=Sub|Number=Sing|Person=2| Tense=Fut|VerbForm=Fin
canta	cantar	VERB	Mood=Imp|Number=Sing|Person=2| VerbForm=Fin

Descrito o conteúdo das entradas do PortiLexicon-UD, a Tabela 4.1 apresenta a distri-

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

buição das entradas e palavras distintas por etiqueta PoS da UD. Nesta tabela também estão indicadas: (i) na coluna *palavras,* o total de palavras distintas por classe; (ii) na coluna *amb*, o número de palavras sintaticamente ambíguas, ou seja, palavras que possuem mais do que uma entrada; (iii) na coluna *non-amb*, o total de palavras não ambíguas e; finalmente, (iv) na coluna *entradas*, o número total de entradas.

Tabela 4.1: Número de palavras e entradas no PortiLexicon-UD por etiqueta PoS da UD

PoS UD	**palavras**	**amb**	**non-amb**	**entradas**
ADJ	124.332	79.165	45.167	124.449
ADP	38	26	12	38
ADV	4.938	1.838	3.100	4.944
AUX	332	277	55	435
CCONJ	25	17	8	25
DET	109	105	4	117
INTJ	41	12	29	41
NOUN	75.154	26.807	48.347	75.377
NUM	53	42	11	54
PRON	157	123	34	180
SCONJ	14	13	1	14
VERB	746.783	65.672	681.111	1.020.665
total	**862.325**	**84.446**	**777.879**	**1.226.339**

É possível perceber que, por serem classes abertas, os verbos, substantivos e adjetivos correspondem à maior parte do léxico. As palavras funcionais, como preposições, conjunções e pronomes, correspondem a uma parte bem menor do léxico, apesar de serem extremamente importantes na linguagem.

A maior parte das palavras do PortiLexicon-UD é de palavras não sintaticamente ambíguas e, portanto, possuem uma única entrada. No entanto, algumas palavras são particularmente ambíguas, como é o caso da palavra "que", que possui o maior número de etiquetas PoS associadas a uma palavra do léxico (7), podendo ser utilizada como ADP, ADV, CCONJ, DET, INTJ, PRON e SCONJ. A palavra que possui o maior número de entradas no léxico é a palavra "fora", que pode ser anotada com 5 etiquetas PoS distintas, mas possui as seguintes onze entradas no PortiLexicon-UD:

1. lema=fora, PoS=ADP, Atr. Morf.=_;
2. lema=fora, PoS=ADV, Atr. Morf.=_;
3. lema=ir, PoS=AUX,
    a. Atr. Morf.=Mood=Ind|Number=Sing|Person=1|Tense=Pqp| VerbForm=Fin;
4. lema=ir, PoS=AUX,
    a. Atr. Morf.=Mood=Ind|Number=Sing|Person=3|Tense=Pqp| VerbForm=Fin;
5. lema=ser, PoS=AUX,
    a. Atr. Morf.=Mood=Ind|Number=Sing|Person=1|Tense=Pqp| VerbForm=Fin;
6. lema=ser, PoS=AUX,
    a. Atr. Morf.=Mood=Ind|Number=Sing|Person=3|Tense=Pqp| VerbForm=Fin;

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

7. lema=fora, PoS=NOUN,
    a. Atr. Morf.=Gender=Masc|Number=Sing;
8. lema=ir, PoS=VERB,
    a. Atr. Morf.=Mood=Ind|Number=Sing|Person=1|Tense=Pqp| VerbForm=Fin;
9. lema=ir, PoS=VERB,
    a. Atr. Morf.=Mood=Ind|Number=Sing|Person=3|Tense=Pqp| VerbForm=Fin;
10. lema=ser, PoS=VERB,
    a. Atr. Morf.=Mood=Ind|Number=Sing|Person=1|Tense=Pqp| VerbForm=Fin;
11. lema=ser, PoS=VERB,
    a. Atr. Morf.=Mood=Ind|Number=Sing|Person=3|Tense=Pqp| VerbForm=Fin.

É importante salientar que o PortiLexicon-UD, por seguir o princípio de anotação de *tokens* individualmente definido pela UD, não representa palavras compostas, ênclises, mesóclises ou palavras contraídas. Dessa forma, palavras como "segunda-feira", "contá-lo", "trazê-lo-ia" ou "desta" não fazem parte do léxico. Enquanto palavras compostas e nomes próprios podem ser incluídos em novas versões, as ênclises, mesóclises e palavras contraídas devem ser objeto de um pré-processamento que transforme esse tipo de palavra em seus componentes para buscá-las no léxico, por exemplo:

- "contá-lo" deve ser desdobrado em 3 *tokens*: "contar", "-" e "lo";
- "trazê-lo-ia" deve ser desdobrado em 3 *tokens*: "traria", "-" e "lo"; e
- "desta" deve ser desdobrado em 2 *tokens*: "de" e "esta".

Dessa forma, o PortiLexicon-UD oferece um recurso que permite a identificação de unidades léxicas de processamento de maneira eficiente e precisa. Por exemplo, na anotação de *corpus* (seja manual ou automática), o léxico pode fornecer um grande suporte à atribuição de etiquetas morfológicas, mas também para a análise de características morfológicas o léxico pode funcionar como uma referência de classificação para cada palavra do português.

4.4 Considerações finais

Este capítulo abordou o processamento automático do português no nível da **palavra**, que é considerado em PLN como a menor unidade de processamento. Para definir e delimitar essa unidade de processamento (a palavra), no entanto, é necessário considerar as pequenas unidades linguísticas que a constituem, que são os **morfemas**.

Portanto, neste capítulo, trouxemos uma visão geral da linguística e os principais conceitos da morfologia (morfema, afixo, desinência, radical etc.), mas também trouxemos os conceitos da **morfossintaxe** (lexema, lexia, léxico, *token*, *type* etc.), cujo objeto principal de estudo é a classificação das **palavras** em partes do discurso (ou PoS) (Seção 4.1).

As duas áreas, **Morfologia** e **Morfossintaxe**, se imbricam e se complementam, já que, para identificar quais são os traços morfológicos de uma palavra, é necessário saber qual o seu PoS. Por outro lado, muitas vezes, para definir o PoS de uma palavra, recorre-se aos seus traços morfológicos.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Para traçar um paralelo de forma bem simples entre as áreas e seus objetos de estudo, podemos dizer que, na linguística, a menor unidade significativa da língua é o morfema, que é estudado pela Morfologia. Em PLN, a menor unidade de processamento automático é a palavra, que é estudada pela Morfossintaxe. O Quadro 4.7 apresenta um resumo dessas associações.

Quadro 4.7: Resumo dos objetos de estudo das áreas Morfologia e Morfossintaxe

área	**unidade de análise**	**tipos de etiquetas**	**unidade mínima de análise/ processamento**
Morfologia	morfema	*feats* ou *features* ou atributos morfológicos	em linguística
Morfossintaxe	palavra	*part-of-speech* ou PoS (XPoS e UPoS) ou classes gramaticais ou classes de palavras	em PLN

Após definir todos esses conceitos (Seção 4.1) relevantes para a linguística e para o Processamento de Linguagem Natural, demonstramos como se faz o processamento morfológico em PLN, indicando as principais tarefas e etapas de processamento dos textos (Seção 4.2).

Por fim, indicamos algumas ferramentas e recursos disponíveis para o português (Seção 4.3), focando mais especificamente em um deles, a fim de exemplificar a complexidade desse nível de análise e processamento linguístico-computacional. Os recursos apresentados neste capítulo são apenas exemplificativos, mas é importante lembrar que existem vários outros que podem ser mais apropriados para uma aplicação ou outra, dependendo da necessidade e dos objetivos da tarefa.

Pretendemos, em uma próxima versão deste capítulo, apresentar atividades e exercícios práticos relacionados à etiquetagem morfológica e morfossintática. Também pretendemos explicar passo a passo como utilizar cada um dos recursos e ferramentas citadas aqui, a fim de contribuir, de forma mais prática e didática, com a internalização dos conceitos e o *modus operandi* de fazer PLN.

Também está prevista para a próxima versão uma apresentação mais global dos tipos de recursos lexicais, como *thesaurus* (e.g. o TeP[48]), redes semânticas (e.g. WordNetBr[49], VerbNetBr[50], PropBankBr[51]), léxicos, dicionários comuns, dicionários especializados, ontologias, glossários e tantos outros recursos construídos para o português.

[48] http://www.nilc.icmc.usp.br/tep2/
[49] http://www.nilc.icmc.usp.br/wordnetbr/
[50] http://143.107.183.175:21380/portlex/index.php/en/projects/verbnetbringl
[51] http://143.107.183.175:21380/portlex/index.php/en/projects/propbankbringl

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Capítulo 5

Expressões multipalavras

fogo de palha ou osso duro de roer?

Renata Ramisch
Carlos Ramisch
Aline Villavicencio

Publicado em: 26/09/2023
Atualizado em: 13/03/2024

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

5.1 Introdução

Para além das palavras, as linguagens humanas em sua riqueza têm modos particulares de expressar ideias complexas de maneira convencional – e muitas vezes abreviada. Esse fenômeno, que faz parte do inventário de diversas comunidades linguísticas, é chamado de **expressões multipalavras**. Esse é geralmente um tema indigesto no universo do Processamento de Linguagem Natural (PLN), porque essas expressões estão no limite entre a sintaxe e a semântica, e sempre acabam ficando no meio do fogo cruzado. De um lado, elas apresentam idiossincrasias e especificidades que não permitem determinadas operações sintáticas e semânticas comuns a outras combinações de palavras. De outro, essas expressões se caracterizam por representarem significados complexos que, em geral, ultrapassam os limites dos sentidos das palavras individuais. Em essência, o significado do todo pode não se dar pela soma das partes. Por exemplo, uma ideia pode ser considerada "sem pé nem cabeça" quando ela não faz sentido, apesar de ideias serem abstratas e não serem dotadas de corpos, pés ou cabeças.

Assim, nosso objetivo neste capítulo é, em primeiro lugar, definir quais são os conceitos fundamentais para quem vai navegar pelas águas turbulentas do tratamento computacional de expressões multipalavras. Começaremos por discutir quais são os elementos que compõem uma expressão: seriam palavras? Seriam lexemas? Na sequência, abordaremos o conceito de expressão multipalavras (MWE) propriamente dito[1]. A literatura traz inúmeras maneiras de analisar essas expressões, e as definições são tão variadas quanto os campos de estudo que se interessam por esse tema.

Dessa forma, apresentaremos na Seção 5.2 a definição de MWEs que adotamos neste capítulo. Na sequência, vamos tentar separar o joio do trigo, uma vez que, para entender o que são MWEs, é fundamental que você saiba também o que elas não são (por exemplo, compostos, colocações e metáforas não são MWEs). Apresentamos também algumas questões relacionadas às possibilidades de classificação das expressões multipalavras, embora ainda haja um longo caminho a se percorrer para chegarmos a uma taxonomia abrangente

[1]Como boa parte da literatura sobre esse assunto está escrita em língua inglesa, decidimos adotar, neste capítulo, o acrônimo *MWE*, derivado do termo em inglês *multiword expressions*.

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

e consensual. Para completar a fundamentação teórica, trazemos uma discussão sobre três características importantes das MWEs: ambiguidade, variabilidade e arbitrariedade.

Após as conceituações, descrevemos na Seção 5.3 as principais tarefas de PLN que envolvem MWEs, as quais se dividem basicamente em dois grandes grupos: a) a descoberta; e b) a identificação de MWEs. Depois, veremos na Seção 5.4 uma breve apresentação dos recursos existentes, com foco no português brasileiro, embora eles sejam pouco numerosos ou variados em comparação com aqueles de outras línguas. Por sinal, incentivamos leitoras e leitores a colocar a mão na massa em busca de uma mudança desse cenário. Na Seção 5.5, trataremos as principais métricas de avaliação comumente utilizadas pela comunidade para avaliar o desempenho dos sistemas de processamento computacional de MWEs.

Por fim, consideramos importante olhar pelo retrovisor e entender qual foi o percurso para chegar até o ponto em que estamos, em termos científicos, na busca por resolver a complexa temática das MWEs (Seção 5.6). Qual é o cenário atual? Qual é a posição do português brasileiro em termos de recursos e pesquisas diante da comunidade internacional? Quais são os desafios que permanecem a pesquisadoras, pesquisadores e entusiastas das MWEs para descascar esse abacaxi em tempos de modelos de língua tão grandes que não cabem em si?

Quem quiser ir mais longe, poderá dar uma olhada nas referências e nos *links* listados na Seção 5.7. Sem querer prometer mundos e fundos, este capítulo tenta acrescentar mais um tijolo na construção do conhecimento linguístico em PLN para o português brasileiro. Como você pôde perceber por esta introdução, o texto contém uma série de exemplos de expressões multipalavras, tanto para ilustrar fenômenos linguísticos quanto para divertir leitoras e leitores.

5.2 Hora de dar nome aos bois

Antes de iniciarmos a nossa jornada pelo universo das MWEs, precisamos definir algumas noções importantes. Uma parte dos conceitos discutidos a seguir baseia-se no projeto internacional PARSEME, do qual fazemos parte. O PARSEME reúne especialistas em MWEs em mais de 20 línguas, incluindo o português. Seu objetivo é criar *corpora* anotados com MWEs em diversas línguas usando diretrizes de anotação unificadas. Além disso, o PARSEME organizou várias *shared tasks* (Seção 16.2) para avaliar sistemas de identificação de MWEs[2].

5.2.1 As pedras fundamentais

Como o nome indica, expressões **multipalavras** possuem, necessariamente, duas ou mais **palavras**. Logo, o primeiro conceito a definir é o próprio conceito de *palavra*. Tal definição está longe de ser um consenso (Church, 2013; Manning; Schütze, 1999; Mel'čuk et al., 1995), mas como o Capítulo 4 já abordou essa discussão de forma aprofundada, não vamos retomá-la aqui.

Na verdade, o termo *lexema* seria mais preciso do que *palavra* para definir tais expressões. **Lexemas** são unidades lexicais elementares de significado que representam blocos básicos do léxico de uma língua, como explicado na Seção 4.1.3. Porém, o termo *expressões multipalavras* já está consolidado, e não pretendemos a essa altura do campeonato erguer

[2] O termo *shared task* pode ser traduzido como avaliação conjunta, tarefa compartilhada, esforço colaborativo ou campanha de avaliação. Para evitar ambiguidade, mantemos o termo em inglês, que já está consagrado na comunidade de PLN no Brasil.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

a bandeira da mudança terminológica para *expressão multilexemas*. Portanto, vamos nos referir aos elementos que compõem uma MWE como palavras.

Em PLN, palavras (ou lexemas) são frequentemente confundidas com *tokens*. Conforme explicado na Seção 4.2.2, os **tokens** são o resultado de um processo computacional de tokenização, ou seja, a segmentação do texto em unidades menores. Aqui, adotamos a distinção entre palavra e *token* do PARSEME (Savary et al., 2018, p. 92), que, por sua vez, é baseada no projeto Universal Dependencies (Marneffe et al., 2021, p. 259), mencionado em outros capítulos do livro:

> Uma **palavra** é uma unidade (semântica) linguisticamente motivada. A identificação de palavras é, portanto, dependente da língua, e anotadores devem ter uma ideia clara de como defini-las para a sua própria língua.
>
> Um ***token*** é uma noção técnica e pragmática, definida de acordo com indícios linguísticos mais ou menos motivados e de acordo com a ferramenta de tokenização específica que se tem em mãos.

Idealmente, *tokens* e palavras teriam uma correspondência de 1 para 1. Em línguas que utilizam espaços para separar palavras (como é o caso do português), é o que geralmente acontece. Entretanto, uma tokenização perfeita é praticamente impossível, como em substantivos compostos ("girassol"), contrações ("caixa d'água") e certas convenções ortográficas ("super-herói"). Além disso, o uso do espaço como separador não é universal: o chinês e o japonês, por exemplo, não separam visualmente as palavras, enquanto o alemão não separa os substantivos compostos[3]. Como consequência, as MWEs podem ou não conter espaços, uma vez que isso depende das convenções ortográficas da língua e/ou do *software* de tokenização. Mas se a presença ou ausência de espaços não é um critério suficiente, como saber se determinado grupo de palavras é ou não é uma MWE?

5.2.2 Definições de MWE: uma pedra no nosso sapato

Infelizmente, não existe uma definição única para o termo *MWE*, pois o conceito é uma espécie de "guarda-chuva" sob o qual se agrupam diversos fenômenos linguísticos. Por um lado, poderíamos dizer informalmente que MWEs são *palavras que se dão bem*. Porém, apesar de ser intuitiva, essa definição não é suficientemente rigorosa para fins práticos. Por outro lado, muito latim já foi gasto para definir e caracterizar o fenômeno de maneira mais precisa e rigorosa. Por exemplo, as diretrizes do projeto PARSEME para MWEs verbais têm 134 páginas[4]!

Entre a definição intuitiva e lacônica (*palavras que se dão bem*) e as dezenas de páginas das diretrizes do PARSEME, várias definições alternativas foram propostas, com escopos ligeiramente diferentes. Smadja (1993) enfatiza a frequência, definindo MWEs como "combinações arbitrárias e recorrentes de palavras". Já Choueka (1988) afirma que uma MWE é "uma unidade sintática e semântica cujo significado ou conotação exata e inequívoca não pode ser derivado diretamente do significado ou conotação de seus componentes". No famoso artigo "pain-in-the-neck", Sag et al. (2002) definem as MWEs como "interpretações idiossincráticas que cruzam os limites (ou espaços) das palavras"[5].

Um dos motivos dessa diversidade de definições é o fato de que várias áreas do conhecimento, com pontos de vista diversos, têm interesse no fenômeno: linguística, PLN, ciências

[3]Veja a história da Rhababerbarbara: https://www.youtube.com/watch?v=IFoyspFAKnM
[4]Disponível em: https://parsemefr.lis-lab.fr/parseme-st-guidelines/
[5]Para saber mais sobre definições de MWEs, consulte o estudo de Seretan (2011, p. 182–184)

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

cognitivas, fraseologia, lexicografia, terminologia etc. Neste capítulo, como já anunciamos anteriormente, vamos usar a definição do PARSEME (Savary et al., 2018), cujos três aspectos principais discutidos na sequência são idiossincrasia, estrutura sintática e lexicalização.

Quadro 5.1: Expressões multipalavras

> **Expressões multipalavras** são entendidas como sequências (contínuas ou descontínuas) de palavras que: (a) contêm pelo menos duas palavras componentes que são lexicalizadas, ou seja, sempre realizadas pelos mesmos lexemas, incluindo uma palavra principal e pelo menos uma outra palavra sintaticamente relacionada; e (b) exibem algum grau de idiossincrasia lexical, morfológica, sintática e/ou semântica.

Fonte: (Savary et al., 2018)

Idiossincrasia

Uma **idiossincrasia** é um comportamento excepcional ou imprevisível. Na definição do Quadro 5.1, o termo se refere ao fato de que MWEs divergem das regras de composição padrão, resultando em combinações imprevisíveis. Assim, poderíamos definir as MWEs como "exceções que ocorrem quando as palavras são combinadas". Essa idiossincrasia é frequentemente semântica (Capítulo 8), visto que "o significado de uma MWE não deriva explicitamente das suas partes" (Baldwin; Kim, 2010). Por exemplo, o significado de "banho" e o de "maria" não resultam no significado de "**banho-maria**" (técnica de cozimento lento). Embora essa idiossincrasia semântica seja uma das características mais prototípicas das MWEs, elas também podem ter outros tipos de idiossincrasias, de natureza lexical, morfológica ou sintática.

Estrutura sintática

A definição do Quadro 5.1 estipula que as palavras que formam uma MWE devem estar "sintaticamente relacionadas", ou seja, apresentar **coesão sintática**. Existem diversos formalismos sintáticos (Capítulo 6), mas o PARSEME se baseia na sintaxe de dependências do Universal Dependencies (Seção 6.6.3). Portanto, consideramos as MWEs como subárvores de dependência formadas por lexemas que não são necessariamente adjacentes no texto. Por exemplo, a anotação de "**ter** como/por **objetivo**" exclui as preposições (variáveis), que não ligam o verbo ao objeto como na sintaxe de dependências tradicional, mas dependem do substantivo "objetivo"[6]. Assim, considerando que uma MWE "age como uma unidade atômica em algum dos níveis de análise linguística" (Calzolari et al., 2002), poderíamos atribuir etiquetas morfossintáticas (Kahane et al., 2017) e semânticas (Schneider; Smith, 2015) às expressões, da mesma maneira como atribuiríamos para palavras simples.

Lexicalização

Uma questão que surge rapidamente ao anotar as MWEs diz respeito à sua extensão: quais palavras fazem parte da expressão e quais não? Por exemplo, o determinante "as" deve ser

[6] A partir deste ponto do texto, as expressões aparecerão sempre em itálico, com seus elementos lexicalizados em negrito. Para saber mais sobre a notação, consulte Markantonatou et al. (2021).

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

anotado como parte da MWE em "**fazer** as **apresentações**"? Ou em "**dar as caras**"? No PARSEME, as diretrizes para a extensão da anotação se baseiam na noção de componentes lexicalizados. Um **componente lexicalizado** é uma palavra que, em todas as ocorrências possíveis da expressão, é sempre realizada pelo mesmo lexema. Isso quer dizer que os componentes lexicalizados não podem ser omitidos ou substituídos por sinônimos, caso contrário, a MWE se tornaria agramatical ou teria um novo sentido (não idiomático). Por exemplo, em "**tomar** um **banho**", o determinante "um" não é um componente lexicalizado, pois pode ser substituído por outro determinante ou omitido sem perder o significado idiomático ("**tomar** dois **banhos**"). Entretanto, "**fazer as pazes**" não pode significar "resolver um conflito" em "fazer a paz"; portanto, "as" é um componente lexicalizado nessa expressão[7].

5.2.3 Separando o joio do trigo

O fato de uma expressão ser composta por mais de uma palavra não a torna automaticamente uma MWE: como vimos anteriormente, expressões multipalavras precisam também apresentar alguma **idiossincrasia** com relação a expressões estruturalmente semelhantes, consideradas regulares, composicionais e produtivas. Portanto, agora que definimos o que são as MWEs, vamos explicitar brevemente o que elas **não** são.

O termo **colocação** muitas vezes é visto como sinônimo de MWE. Neste capítulo, adotamos a definição de colocações como sendo combinações de palavras que aparecem juntas com mais frequência do que o esperado por puro acaso[8]. Logo, podem ser combinações completamente regulares que apresentam apenas preferências de associação **estatística**, sem nenhuma outra idiossincrasia (por exemplo, "ler um livro"). Consideramos que a saliência estatística não é um critério suficiente para caracterizar MWEs.

Compostos são lexemas resultantes do processo de **formação de palavras** (ou seja, justaposição de lexemas, às vezes com pequenas adaptações morfológicas). Dependendo do idioma, os compostos não apresentam necessariamente um comportamento idiossincrático, de modo que nem todos eles são MWEs.

As **metáforas** até podem evoluir para MWEs ao longo do tempo, mas geralmente são mais flexíveis. Por exemplo, o "coração" está associado a emoções, e "partir o coração" pode ser parafraseado como "destruir o amor"[9]. Nas metáforas, geralmente é difícil identificar pelo menos dois componentes lexicalizados, que não podem ser parafraseados. Então, elas não são consideradas MWEs.

Por fim, as MWEs podem ser definidas como combinações que "correspondem a alguma forma convencional de dizer as coisas" (Manning; Schütze, 1999). No entanto, a **convenção** também está em jogo em **entidades nomeadas** e **termos** de domínios específicos, por exemplo, da saúde (Capítulo PLN na Saúde) ou do direito (Capítulo PLN no Direito). Entidades nomeadas se referem a entidades específicas no mundo, como pessoas ("Inês Brasil"), lugares ("Porto Alegre") e organizações ("Movimento dos Sem Terra"). Os **termos**, por sua vez, denotam conceitos especializados de um domínio técnico ou científico ("redes neurais", "ressonância magnética"). Embora seja possível ver termos e entidades nomeadas multipalavras como MWEs, isso não é conveniente. Dada a natureza complexa

[7]Essa noção está ligada a um teste formal para identificar MWEs, no qual inflexibilidade é vista como um indicador de não composicionalidade semântica (Candito et al., 2021, p. 467).

[8]Veja Evert (2009) para uma discussão detalhada.

[9]Algumas MWEs são metáforas "congeladas", embora sua etimologia não seja sempre óbvia, por exemplo, "**lágrimas de crocodilo**" (https://pt.wikipedia.org/wiki/Lágrimas_de_crocodilo).

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

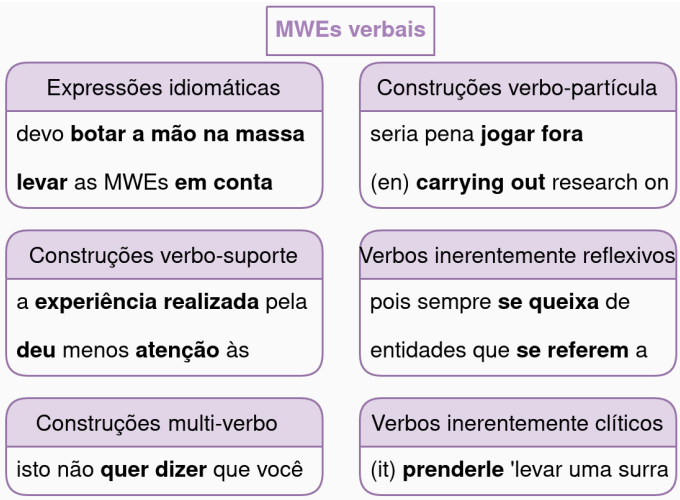
das expressões multipalavras, elas já dão bastante pano pra manga[10]. Assim, parece razoável delegar o tratamento de entidades nomeadas e termos a outras comunidades de pesquisa, por exemplo, em extração de informações (Capítulo Extração de Informação)[11]. Então, consideramos que convencionalidade não é um critério suficiente para MWEs.

5.2.4 MWEs para dar e vender

As expressões multipalavras são um fenômeno linguístico diverso, que desafia inúmeras tentativas de categorização. Porém, pode ser útil agrupar expressões semelhantes em categorias, tanto para a criação de recursos lexicais quanto para a anotação de *corpora* (Seção 5.4). Ramisch (2015) resume um conjunto de classificações de MWEs propostas na literatura, abrangendo a gramática de construções (Fillmore et al., 1988), a teoria sentido-texto (Mel'čuk; Polguère, 1987) e as classificações orientadas ao PLN (Sag et al., 2002; Smadja, 1993). O autor também propôs uma tipologia baseada em dois eixos ortogonais: a distribuição morfossintática da MWE como um todo e o nível de "dificuldade".

Parra Escartín et al. (2018) fazem uma comparação geral dessas classificações, propondo não apenas uma nova categorização, mas também critérios para categorizar as MWEs. O artigo sugere que "as tipologias de MWEs devem ser adaptadas à língua que está sendo pesquisada, e as tipologias clássicas baseadas principalmente no inglês não parecem adequadas para descrever e classificar as MWEs em outras línguas". Por exemplo, as construções verbo-partícula do inglês (por exemplo, "**take off**") são frequentemente consideradas uma categoria importante de MWEs, embora sejam irrelevantes para línguas de muitas famílias linguísticas, como as eslavas e as românicas (família da qual o português faz parte). Os verbos inerentemente reflexivos (por exemplo, "**se queixar**") são muito mais comuns nessas línguas, mas raramente são incluídos em classificações centradas no inglês.

Figura 5.1: Taxonomia das MWEs verbais proposta pelo PARSEME



Fonte: Imagem adaptada e traduzida de (Ramisch, 2023).

[10] Em português brasileiro, a troca de *pra* por *para*, ainda que possível, provavelmente geraria algum estranhamento.

[11] Para uma discussão aprofundada sobre MWEs vs. entidades nomeadas, consulte Candito et al. (2021).

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

MWEs verbais

Uma categorização de MWE única, operacional e válida em todas as línguas, abrangendo todos os fenômenos que correspondem à definição do Quadro 5.1, é uma questão de pesquisa aberta e ambiciosa. No entanto, um passo importante nessa direção é a tipologia do PARSEME para MWEs verbais, que abrange um grande número de línguas de diferentes famílias (Savary et al., 2018). A categoria das MWE verbais do PARSEME engloba seis subcategorias mais específicas, determinadas principalmente pela natureza do complemento usado pelo verbo principal: **construções multi-verbo**, se o complemento for outro verbo; **verbos inerentemente reflexivos** (IRV), se for um clítico reflexivo; **verbos inerentemente clíticos**, se for outro clítico (não reflexivo); **construções verbo-partícula**, se for uma partícula homônima a uma preposição ou um advérbio. Todos os outros tipos de MWEs verbais idiossincráticas deveriam, em teoria, pertencer à categoria **expressões idiomáticas verbais** (VID). Contudo, uma categoria especial tem precedência sobre as expressões VID: as **construções verbo-suporte** (LVC) são formadas por substantivos predicativos (que denotam eventos ou estados), acompanhados por um verbo-suporte (também chamado de verbo leve) que modifica o evento ou estado por meio de suas características morfológicas. Exemplos de cada categoria podem ser vistos na Figura 5.1[12]. Falaremos mais sobre essas categorias na Seção 5.4.3, que descreve o *corpus* PARSEME do português brasileiro.

MWEs não verbais

A generalização da categorização do PARSEME para outras categorias morfossintáticas além dos verbos, bem como a extensão das diretrizes de anotação, constitui um objetivo ousado para trabalhos futuros. Todavia, existem na literatura algumas propostas para algumas línguas específicas. Por exemplo, Schneider; Smith (2015) propõem uma categorização simples em expressões "fracas" e "fortes". As diretrizes de anotação do *corpus* STREUSLE[13] descrevem de que maneira distinguir essas duas categorias em inglês. Já Candito et al. (2021) diferenciam apenas MWEs de entidades nomeadas, anotando as MWEs em francês de acordo com seu papel morfossintático na frase (substantivo, verbo etc.). Além disso, as autoras e autores também indicam o(s) critério(s) usado(s) para considerar que determinada combinação é uma MWE.

Por fim, Ramisch (2023) propõe uma taxonomia baseada nas categorias de palavras simples do Universal Dependencies (Marneffe et al., 2021). Adotar a visão do UD pode ser interessante porque ela já foi testada para centenas de línguas e *corpora* anotados com sintaxe, como descrito no Seção 6.6.3. No UD, as unidades linguísticas são classificadas como **nominais**, que se referem a entidades (geralmente substantivos); **orações**, que se referem a eventos ou estados (geralmente verbos); **modificadores**, usados para especificar os atributos de nominais, orações ou outros modificadores (tradicionalmente adjetivos e advérbios). Além disso, um conjunto de itens **funcionais**, como determinantes e verbos auxiliares, não são independentes, mas atuam como especificadores do significado ou da função sintática das três categorias principais. A tipologia proposta por Ramisch (2023) estende essas noções às expressões multipalavras.

[12]Veja detalhes e exemplos em português em http://parsemefr.lis-lab.fr/parseme-st-guidelines/.
[13]Disponível em: https://github.com/nert-nlp/streusle

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

5.2.5 Um osso duro de roer

Depois dessa árdua missão de categorizar as MWEs, tentar persuadir a leitora ou o leitor de que essas expressões são difíceis de modelar e processar seria ensinar o padre a rezar a missa. Portanto, vamos nos concentrar em três propriedades onipresentes nas MWEs: ambiguidade, variabilidade e arbitrariedade[14]. Elas são ao mesmo tempo difíceis de lidar e interessantes de explorar para detectar a presença de MWEs em textos (Constant et al., 2017).

Ambiguidade

A ambiguidade dificulta a identificação de expressões multipalavras porque determinada combinação de lexemas pode ser ou não uma MWE, dependendo do contexto em que ela ocorre. Por exemplo, "**quebrar um galho**" é claramente uma MWE em "meu mecânico me **quebrou um galho** arrumando o meu carro no domingo", mas não em "o vento foi tão forte que quebrou um galho do jacarandá". Essa propriedade das MWEs foi amplamente estudada, sobretudo em estudos cognitivos interessados em como o significado idiomático é armazenado e acessado no cérebro humano (Geeraert et al., 2018; Popiel; McRae, 1988). Entretanto, na prática, parece que a importância do problema foi superestimada. Savary et al. (2019b) demonstram que, pelo menos para MWEs verbais em cinco idiomas diferentes (entre os quais o português), a proporção de leituras literais, em comparação com ocorrências idiomáticas, é praticamente irrelevante (2-4%). Em termos práticos, isso significa que um sistema de identificação automática baseado em regras pode ser tão eficiente quanto (ou até mais eficiente que) um sistema mais complexo, baseado em desambiguização contextual estatística usando aprendizado de máquina (Pasquer et al., 2020). No entanto, seria necessário aumentar consideravelmente a cobertura dos léxicos de MWEs (Seção 5.4) para conferir certa robustez a métodos baseados em regras (Savary et al., 2019a).

Variabilidade

Um dos desafios para se lidar com as expressões multipalavras é que, embora os exemplos prototípicos sejam completamente fixos, na prática há uma variabilidade significativa entre diferentes ocorrências de uma mesma MWE, especialmente para algumas categorias, como as verbais. A variabilidade limitada constitui uma propriedade observável que é frequentemente usada em testes linguísticos para a presença dessas expressões. Muitos testes de variabilidade foram projetados para capturar a idiomaticidade morfológica, lexical, sintática, semântica e pragmática das MWEs (Savary et al., 2018; Schneider; Smith, 2015). No nível léxico-semântico, a variabilidade limitada também foi chamada de **não substituibilidade** (Manning; Schütze, 1999). A substituição por um lexema relacionado é um teste útil para verificar se um componente é lexicalizado e se a combinação apresenta algum grau de idiossincrasia semântica, já que o resultado geralmente é inaceitável ou agramatical, ou produz uma mudança de significado inesperada (por exemplo, "**bater as botas**/?sapatilhas"). Nos níveis morfológico e sintático, a variabilidade limitada geralmente se manifesta por meio de um comportamento sintático irregular (por exemplo, "**quem me dera**") em relação a construções sintaticamente semelhantes (por exemplo, "?quem lhe deu"). Mas a variabilidade é uma faca de dois gumes: ao mesmo tempo em

[14] Detalhes e exemplos sobre por que as MWEs são um osso duro de roer para os modelos de PLN podem ser encontrados em Constant et al. (2017) e Ramisch; Villavicencio (2018).

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

que ajuda a identificar e anotar as MWEs manualmente, também dificulta sua identificação automática quando determinada forma (canônica) é conhecida, por exemplo, em um léxico, e a expressão ocorre de uma outra forma ("**throw** someone **to the lions** vs. **wolves**").

Arbitrariedade

Nas MWEs, as palavras interagem de maneiras pouco usuais umas com as outras, assumindo significados inesperados ou até mesmo perdendo por completo seus significados originais. Como os componentes lexicalizados das MWEs são arbitrários, elas são difíceis de prever e difíceis de gerar automaticamente usando mecanismos de composição. Um exemplo prototípico é a geração de MWEs em textos traduzidos por máquina. A tradução palavra a palavra desse tipo de expressão pode gerar traduções pouco naturais (e até mesmo engraçadas) ou erradas. Por exemplo, a expressão em inglês "to cost **an arm and a leg**" se tornaria "custar um braço e uma perna" se traduzida literalmente, enquanto a tradução correta e idiomática seria "custar **os olhos da cara**". Dada a arbitrariedade dessas MWEs, não parece razoável esperar que um tradutor automático seja capaz de traduzir esse exemplo para o português sem conhecimento externo sobre expressões multipalavras nas duas línguas. Assim como para a variabilidade, a arbitrariedade é tanto uma maldição quanto uma bênção: a incapacidade ou a dificuldade de tradução de uma MWE pode ser usada como um teste de identificação[15].

5.3 Tarefas: botando a mão na massa

Uma das principais confusões terminológicas da área diz respeito à nomenclatura e à definição das tarefas computacionais relativas às MWEs. O que chamamos de *processamento de MWEs* tem sido chamado na literatura de identificação (Tsvetkov; Wintner, 2011), extração (Tsvetkov; Wintner, 2012), aquisição (Ramisch, 2015), indução de dicionários (Schone; Jurafsky, 2001), aprendizado (Korkontzelos, 2011) e assim por diante. Para pôr a coisa nos eixos, o *survey* de Constant et al. (2017) propôs uma terminologia (amplamente adotada na comunidade desde então) que define as tarefas ligadas às MWEs. O quadro conceitual proposto divide o processamento de MWEs em duas subtarefas: **descoberta** e **identificação**. A **descoberta** tem por objetivo encontrar MWEs (lexemas) **novas** no texto e armazená-las para uso futuro em um léxico. A **identificação**, por sua vez, é o processo de anotar essas expressões automaticamente (*tokens*) em um texto, associando-as a MWEs (lexemas) conhecidas.

A delimitação das duas tarefas é fundamental porque, embora ambos os processos recebam texto bruto como entrada, seus resultados são diferentes, como ilustrado na Figura 5.2 e na Figura 5.3. A saída da descoberta é uma lista de unidades lexicais candidatas, enquanto a da identificação é um texto anotado. A lista de candidatas a MWE geralmente requer uma revisão manual por especialistas antes de ser adicionada a um léxico. A identificação, por outro lado, gera anotações que podem ajudar a chegar ao significado correto do texto em tarefas subsequentes de PLN.

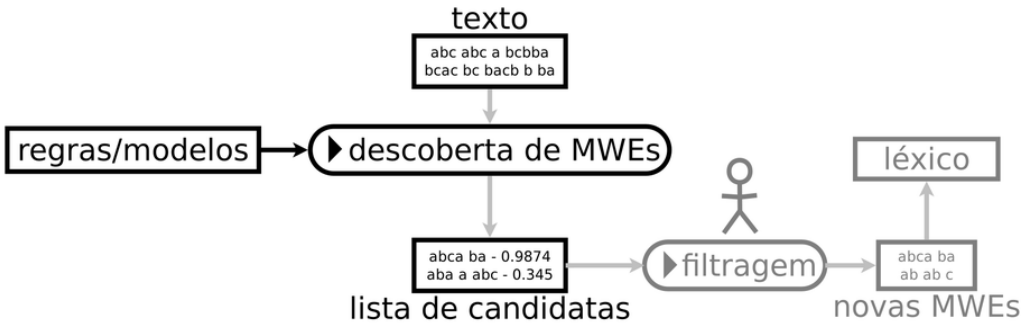
Ambas as tarefas também costumam empregar abordagens e estratégias de avaliação diferentes. Autoras e autores de métodos de descoberta tendem a aplicar técnicas não

[15] Ainda que útil, esse teste não é determinístico, já que algumas MWEs, por coincidência, possuem traduções palavra a palavra em outras línguas (por exemplo, "**febre amarela**", vira "**yellow fever**" em inglês e "**fièvre jaune**" em francês).

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

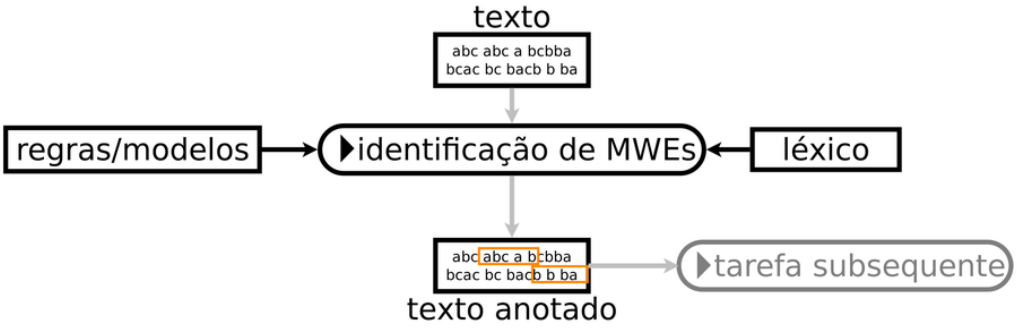
©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Figura 5.2: Descoberta de MWEs: regras ou modelos extraem do texto uma lista de candidatas. A lista é então filtrada por especialistas, que adicionam novas entradas ao léxico.



Fonte: Adaptado de (Constant et al., 2017).

Figura 5.3: Identificação de MWEs: regras, modelos ou léxicos são aplicados ao texto de entrada, gerando uma nova versão dele na qual as MWEs estão destacadas.



Fonte: Adaptado de (Constant et al., 2017).

supervisionadas, que são avaliadas em termos da qualidade das MWEs descobertas. Por outro lado, abordagens de identificação são frequentemente baseadas em modelos de aprendizado supervisionado, cujos resultados são avaliados comparando texto anotado automaticamente com anotações de referência, muitas vezes feitas por especialistas humanos (Seção 5.5).

Um aspecto importante da descoberta de MWEs é a predição de composicionalidade. O princípio da **composicionalidade** pressupõe que o significado de frases, expressões ou sentenças pode ser determinado pelos significados de suas partes e pelas regras usadas para combiná-las (Frege, 1892/1960). Dito de outro modo, o "significado de uma frase típica em uma linguagem natural é complexo, pois resulta da combinação de significados que são, de certa forma, mais simples" (Cruse, 1986, p. 24). Como consequência, somos capazes de atribuir interpretações até mesmo a novas frases, envolvendo combinações inéditas compostas por elementos conhecidos (Goldberg, 2015).

O **grau de composicionalidade** expressa, sob a forma de um valor numérico, em que proporção o significado de um grupo de palavras pode ser inferido ou adivinhado a partir dos significados das palavras que o compõem. Por exemplo, um "**processo seletivo**" é realmente um processo para seleção de pessoas (alto grau de composicionalidade), enquanto

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

um "**pé frio**" é uma pessoa com falta de sorte, ou seja, seu sentido está pouco ou nada relacionado com os sentidos das palavras "pé" e "frio" (baixo grau de composicionalidade). A predição automática do grau de composicionalidade permite decidir quais entradas devem aparecer em um léxico (pouco composicionais, cujo significado não pode ser inferido a partir das palavras) e quais não precisam (muito composicionais). Apresentaremos alguns recursos para essa tarefa na Seção 5.4.1.1.

Além dessas tarefas, existem outras aplicações de PLN que podem se beneficiar da descoberta ou identificação prévia de MWEs. Na tradução automática, essas expressões frequentemente provocam erros de tradução, e a qualidade da tradução de MWEs foi avaliada por vários trabalhos (Barreiro et al., 2013; Ramisch et al., 2013). Diversos resultados demonstraram que uma modelagem explícita das expressões multipalavras pode ajudar a gerar traduções de maior qualidade (Bouamor et al., 2012; Cap et al., 2014; Carpuat; Diab, 2010; Stymne et al., 2013; Tan; Pal, 2014; Zaninello; Birch, 2020). A identificação explícita de MWEs também pode ser útil para a análise sintática (Nivre; Nilsson, 2004) ou ser realizada em conjunto com ela (Constant; Nivre, 2016). Outras aplicações de PLN nas quais a identificação de MWEs foi avaliada incluem a recuperação de informação (Acosta et al., 2011), a desambiguação de sentido de palavras (Finlayson; Kulkarni, 2011), a etiquetagem de *supersenses* (Liu et al., 2021; Schneider; Smith, 2015), a análise de sentimentos (Hwang; Hidey, 2019), a predição de níveis de complexidade textual (Gooding et al., 2020), a identificação de metáforas (Rohanian et al., 2020) e a detecção de discurso de ódio (Zampieri et al., 2021). A próxima seção apresenta alguns dos recursos linguístico-computacionais existentes usados para treinar e avaliar sistemas que realizam as tarefas descritas nesta seção.

5.4 Recursos: uma joia rara

Frequentemente, são os recursos que conectam a descoberta e a identificação de expressões multipalavras. Recursos também fazem a ponte entre ambos os processos e outras tarefas de PLN, como as diversas aplicações listadas anteriormente. Há dois tipos principais de recursos que participam do processamento de MWEs: léxicos (Seção 5.4.1) e *corpora* (Seção 5.4.2). Na sequência, definimos e exemplificamos alguns desses recursos, cujo desenvolvimento é crucial para esse campo de pesquisa.

5.4.1 Léxicos de MWEs

Em PLN, um léxico de MWEs pode ser simplesmente uma lista contendo expressões em determinada língua. Tais listas são bastante comuns no campo da terminologia, assim como no reconhecimento de entidades nomeadas (Capítulo Extração de Informação). Formas mais sofisticadas de léxicos de MWEs podem incluir informações sobre a categoria, a função morfossintática (POS), a estrutura sintática interna da MWE, o sentido, uma ou mais definições, traduções etc. Especialmente relevantes para esses léxicos são as restrições de variabilidade aplicadas a alguns elementos da expressão, por exemplo, o plural obrigatório para "pazes" em "**fazer as pazes**" (Savary et al., 2020). A representação dessas restrições foi estudada em vários formalismos: Gross (1986) na léxico-gramática, Mel'čuk (2023) na teoria sentido-texto, Grégoire (2010) na forma de classes de equivalência e Przepiórkowski et al. (2014) em um dicionário de valência.

A maioria desses léxicos é construída manualmente, ainda que ferramentas automáticas de descoberta de MWEs (cuja saída são "pré-léxicos", como mostrado na Figura 5.2) pos-

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

sam orientar o processo. Existem léxicos de expressões multipalavras com granularidades e tamanhos variados em diversos idiomas, como o grego (Markantonatou et al., 2019), o francês (Gross, 1986; Ramisch et al., 2016a), o holandês (Grégoire, 2010), o polonês (Graliński et al., 2010; Przepiórkowski et al., 2014), e em dialetos latino-americanos do espanhol (Bogantes et al., 2016). Para a língua portuguesa, existem recursos baseados no formalismo léxico-gramática, desenvolvidos pelo projeto UNITEX-PB[16] (Muniz, 2004), ou como parte da ferramenta de análise sintática do português europeu STRING[17] (Baptista et al., 2022).

Finalmente, existe um grande número de dicionários impressos e eletrônicos para usuárias e usuários humanos (por exemplo, aprendizes de idiomas) que frequentemente contêm MWEs representadas como entradas regulares, como entradas relacionadas a uma palavra-chave simples (por exemplo, uma das palavras componentes da MWE), ou em volumes especializados em MWEs (Campress, 1997; Sinclair, 1989; Walter, 2006) [18]. O *survey* de Losnegaard et al. (2016) mostra um apanhado geral dos recursos de MWEs, com um foco especial em léxicos computacionais.

5.4.1.1 *Datasets* de composicionalidade

Tabela 5.1: Grau de composicionalidade de 0 (idiomático) a 5 (composicional) do substantivo (SUB), adjetivo (ADJ) e par substantivo-adjetivo. Média ± desvio padrão dos mais polêmicos (acima) aos mais consensuais (abaixo).

	MWE	**SUB**	**ADJ**	**SUB-ADJ**
Polêmica+	pavio curto	1.6 ± 1.8	1.1 ± 1.9	**1.9** ± 2.3
	sexto sentido	4.0 ± 1.4	2.5 ± 2.1	**2.8** ± 2.2
	gelo-seco	3.2 ± 1.6	3.2 ± 1.8	**3.0** ± 2.1
	mau-olhado	1.8 ± 1.2	4.2 ± 1.5	**2.3** ± 2.1
	câmara fria	3.6 ± 2.2	5.0 ± 0.0	**3.4** ± 2.1
	poção mágica	3.3 ± 2.3	3.8 ± 1.5	**3.4** ± 2.1
	estrela cadente	1.7 ± 2.1	4.0 ± 1.2	**1.7** ± 2.1
Consenso+	farinha integral	5.0 ± 0.0	4.5 ± 0.6	**5.0** ± 0.0
	gato-pingado	0.0 ± 0.0	0.3 ± 0.5	**0.0** ± 0.0
	núcleo atômico	5.0 ± 0.0	4.4 ± 1.8	**5.0** ± 0.0
	pão-duro	0.0 ± 0.0	1.0 ± 1.7	**0.0** ± 0.0
	sentença judicial	5.0 ± 0.0	5.0 ± 0.0	**5.0** ± 0.0
	tartaruga-marinha	5.0 ± 0.0	5.0 ± 0.0	**5.0** ± 0.0
	vôo internacional	5.0 ± 0.0	5.0 ± 0.0	**5.0** ± 0.0

Fonte: Exemplos de (Ramisch et al., 2016b).

Uma informação particularmente interessante para nós é o grau de composicionalidade das MWEs. Como dito na Seção 5.3, fazer uma predição automática do grau de composicionalidade de um grupo de palavras constitui uma subtarefa importante na descoberta

[16]Disponível em: http://www.nilc.icmc.usp.br/nilc/projects/unitex-pb/web/
[17]Disponível em: https://string.hlt.inesc-id.pt/w/index.php/Dictionaries
[18]Por exemplo, eles foram úteis ao escrever este capítulo, para encontrar exemplos de MWEs.

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

de novas expressões a serem adicionadas a um léxico. Para desenvolver e avaliar tais sistemas, precisamos de conjuntos de dados que contenham valores de referência que possam ser usados para aproximar um padrão-ouro, como exemplificado na Tabela 5.1. Esses valores podem ser fornecidos por linguistas ou via *crowdsourcing* — nesse caso, é necessário que várias pessoas anotem a mesma MWE para atenuar a subjetividade intrínseca à anotação por pessoas leigas. Vejamos a seguir alguns exemplos de conjuntos de dados de composicionalidade.

- Reddy et al. (2011) anotaram um conjunto de 90 MWEs em inglês compostas por dois substantivos — por exemplo, ***zebra crossing*** (lit. "passagem zebra"), que tem como significado idiomático "faixa de pedestres" — e por um substantivo e um adjetivo — por exemplo, ***sacred cow*** (lit. "vaca sagrada"), que tem como significado idiomático "ideia não criticável". Cada MWE possui três valores numéricos de 0 a 5: o grau de composicionalidade da MWE como um todo e a contribuição de cada uma de suas partes para o significado global da MWE (por exemplo, "**sacred**" para "**sacred cow**" e "**cow**" para "**sacred cow**"). O conjunto de dados foi construído por *crowdsourcing*, e os valores finais são a média de 30 anotações por MWE.

- Farahmand et al. (2015) coletaram anotações para 1.042 MWEs nominais em inglês. Cada MWE foi anotada por quatro especialistas quanto à sua composicionalidade e convencionalidade, usando uma escala binária. Nesse léxico, as MWEs são consideradas não composicionais se pelo menos duas pessoas assim o afirmarem (Yazdani et al., 2015), e o valor total de composicionalidade é dado pela soma das quatro anotações binárias.

- Roller et al. (2013) coletaram anotações para um conjunto de 244 MWEs compostas por dois substantivos em alemão. Cada MWE tem uma média de 30 anotações de grau de composicionalidade em uma escala de 1 a 7, obtidas através de *crowdsourcing*. O recurso foi posteriormente enriquecido com outros tipos de informação, como grau de familiaridade e abstração (Roller; Schulte im Walde, 2014). Outros trabalhos do mesmo grupo de pesquisa estenderam a cobertura e as informações coletadas para compostos nominais em alemão (Schulte im Walde et al., 2016) e criaram um conjunto de dados para construções verbo-partícula em alemão (Bott et al., 2016; Schulte im Walde et al., 2016).

- Cordeiro et al. (2019) compilaram um conjunto de dados multilíngue contendo 180 MWEs em francês e em português, e 280 MWEs em inglês (das quais 90 derivam do recurso de Reddy et al. (2011)). Cada MWE foi associada a três valores de grau de composicionalidade, de 0 (composicional) a 5 (idiomático), como exemplificado na Tabela 5.1. Além de estudar as anotações coletadas, os autores e a autora utilizaram 180 compostos em cada língua para desenvolver modelos de predição automática de composicionalidade (coluna SUB-ADJ) e para estudar parâmetros dos modelos como o pré-processamento do *corpus*, os modelos vetoriais usados ou os coeficientes de combinação de vetores. Ao final, 100 MWEs em inglês foram reservadas para teste, a fim se de avaliar a robustez dos resultados obtidos nos experimentos. Esse conjunto de dados foi estendido com paráfrases (Wilkens et al., 2017) e anotações de grau de composicionalidade em contexto de sentenças (Garcia et al., 2021; Tayyar Madabushi et al., 2022).

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Essa lista cobre apenas uma amostra dos conjuntos de dados existentes para a predição automática de composicionalidade. Ramisch (2023) apresenta um resumo de 33 conjuntos de dados em 12 línguas, entre os quais 6 possuem MWEs em português. De acordo com Savary et al. (2019a), léxicos de MWEs são essenciais para aumentar a cobertura de sistemas de identificação automática de MWEs em contexto. Se essa hipótese se confirmar, será necessário bastante trabalho para estender e melhorar a qualidade dos léxicos existentes.

5.4.2 *Corpora* de MWEs

Como ilustramos na Figura 5.3, *corpora* anotados são importantes para desenvolver e avaliar sistemas de identificação automática de MWEs em textos. Antes de 2016, havia pouquíssimos recursos desse tipo disponíveis, com destaque para o Wiki50 (Vincze et al., 2011) e o STREUSLE (Schneider; Smith, 2015), ambos em inglês. A maioria dos métodos de identificação de expressões multipalavras foi avaliada em *treebanks* (Seção 7.3.1), com anotações obtidas indiretamente a partir das árvores sintáticas (Rosén et al., 2016). Por exemplo, a identificação de MWEs foi estudada no *Talbanken* em sueco (Nivre; Nilsson, 2004), no *French Treebank* em francês (Candito; Constant, 2014), no *Arabic Treebank* em árabe (Green et al., 2013), nos *treebanks* turcos MST, IMST, IVS e IWT (Eryiğit et al., 2015), no *treebank* húngaro Szeged (Vincze et al., 2013), no *treebank* de Praga em tcheco (Bejček et al., 2013), no Penn *treebank* e nos *treebanks* UD em inglês (Cafferkey et al., 2007; Constant; Nivre, 2016; Kato et al., 2016).

Em 2016 e 2017, duas *shared tasks* deram o pontapé inicial na criação de *corpora* anotados especificamente para a identificação de MWEs. A *shared task* DiMSUM fez parte do SemEval 2016 (Schneider et al., 2016). Os dados de treinamento e teste, em inglês, derivam em grande parte do *corpus* STREUSLE (Schneider; Smith, 2015), anotado com MWEs fortes e fracas e com *supersenses*, ou seja, etiquetas semânticas genéricas. No ano seguinte, a primeira edição da *shared task* PARSEME colocou à disposição *corpora* anotados com MWEs verbais em várias línguas. Dada a importância desses *corpora* na comunidade e em função da participação das autoras e do autor do capítulo na anotação do português brasileiro, detalhamos em seguida as diferentes edições dos *corpora* do PARSEME. A Tabela 5.2 traz um resumo sobre cada uma das edições da *shared task* correspondente.

Edição 1.0

A primeira edição da *shared task* resultou em *corpora* anotados para 18 línguas usando a versão 1.0 das diretrizes de anotação[19]. Esse foi o primeiro *corpus* anotado para MWEs verbais amplamente multilíngue, seguindo um conjunto de diretrizes de anotação único, com os mesmos formatos e as mesmas ferramentas de anotação.

Edição 1.1

Na edição de 2018, houve mudanças significativas, que foram mantidas desde então: um novo formato de dados, novas diretrizes de anotação, novas línguas, um controle de qualidade mais sistemático e métricas de avaliação específicas para sistemas de identificação. Três línguas da edição 1.0 (maltês, tcheco e sueco) não foram incluídas nessa edição, mas

[19] Disponível em: https://parsemefr.lis-lab.fr/parseme-st-guidelines/1.0/

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

cinco novas aderiram ao PARSEME (árabe, basco, croata, inglês e hindi), com a versão final contendo 20 línguas.

Edição 1.2

Em 2020, foi lançada a terceira edição, sem grandes alterações nas diretrizes de anotação, mas abrangendo apenas 14 línguas, entre as quais duas novas línguas: chinês e irlandês. Essa edição exigia a preparação de um grande *corpus* bruto (não anotado com MWEs, mas automaticamente anotado com informações morfossintáticas), e como muitas equipes não conseguiram fornecer esse *corpus*, tais línguas acabaram não sendo incluídas.

Edição 1.3

Por fim, a edição mais recente no momento da escrita deste capítulo foi lançada em 2023. Essa foi a primeira versão não associada a nenhuma *shared task*. As diretrizes de anotação não tiveram modificações significativas, mas a grande contribuição dessa edição foi a compilação dos *corpora* de todas as 26 línguas que participaram das três edições precedentes.

Tabela 5.2: Dados das edições das *shared tasks* do PARSEME

Referência	línguas	sentenças	*tokens*	MWEs
v1.0 (Savary et al., 2017) http://hdl.handle.net/11372/LRT-2282	18	274,376	5.4M	62,218
v1.1 (Ramisch et al., 2018a) http://hdl.handle.net/11372/LRT-2842	20	280,838	6.1M	79,326
v1.2 (Ramisch et al., 2020) http://hdl.handle.net/11234/1-3367	14	279,785	5.5M	68,503
v1.3 (Savary et al., 2023a) http://hdl.handle.net/11372/LRT-5124	26	455,629	9.3M	127,498

Em relação ao desenvolvimento dos sistemas de identificação de MWEs, na primeira edição, os *corpora* destinados a essa tarefa foram divididos em conjunto de treino (fornecido previamente) e conjunto de teste (fornecido apenas no final da *shared task*). Em todas as demais edições, havia ainda um *corpus* intermediário de desenvolvimento (fornecido previamente), destinado ao desenvolvimento e à otimização nos sistemas após a primeira etapa de treinamento. Os sistemas foram avaliados em termos de precisão, cobertura e medida-F, conforme descrito na Seção 5.5.

5.4.3 *Corpora* de MWEs em português

Em todas as edições do PARSEME, o português se fez presente com *corpora* jornalísticos anotados por especialistas. Um deles guarda uma característica interessante em termos linguísticos, visto que deriva do jornal popular Diário Gaúcho. Sua principal característica é a linguagem mais coloquial, o que permitiu uma variedade de MWEs do tipo expressão idiomática verbal (VID)[20], que costumam se fazer mais presentes nessa modalidade linguística. Por exemplo, encontramos nesse *corpus* diversas expressões relacionadas ao futebol (como "**marcar gol**") ou à televisão (como "**ir ao ar**"). Além do interesse linguístico,

[20]As categorias foram definidas na Seção 5.2.4.

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

a linguagem mais coloquial e informal dos textos tornou a experiência de anotação mais divertida.

Na edição 1.1 da *shared task*, descrita por nós em Ramisch et al. (2018b), a categoria mais frequente de MWEs verbais é a de expressões com verbo-suporte. Essa categoria representava mais de 60% das expressões anotadas, seguida das expressões idiomáticas, representando cerca de 20% das MWEs. Em relação ao comprimento das expressões e tamanho das lacunas entre os elementos que as compõem (descontinuidade), esses valores dependem da categoria analisada.

Nos verbos inerentemente reflexivos, a maioria das expressões tem exatamente dois elementos lexicalizados, não havendo nenhum outro elemento entre eles. No que se refere aos LVCs, apesar de a maioria deles também ter exatamente dois elementos lexicalizados, há um número considerável de ocorrências que apresentam descontinuidade igual a 1, ou seja, os elementos lexicalizados estão separados por uma palavra, que geralmente corresponde a um determinante. Os VIDs tendem a ser mais longos, com média de 2,9 elementos lexicalizados — a maior MWE anotada continha 10 palavras ("**estar com a faca e o queijo na mão**"). A maioria dos VIDs são contínuos, mas eventualmente é possível encontrar advérbios ou determinantes entre os elementos lexicalizados, como em "**cair** muito **bem**".

Em Ramisch et al. (2018b), incluímos ainda uma análise linguística de alguns casos interessantes. Um dos desafios com os quais nos deparamos é a dificuldade de identificar com clareza os argumentos semânticos necessários para que um substantivo seja considerado predicativo e, com isso, permitir a anotação de um LVC. Um exemplo no contexto do futebol é a palavra "falta". Em "o jogador fez uma falta", é difícil afirmar que o substantivo de fato tem argumentos; já em "o jogador faz falta ao time", esses argumentos estão mais evidentes. Todavia, a equipe de anotação decidiu não anotar nenhum desses casos como construções verbo-suporte (LVCs).

Outro caso desafiador é o dos verbos inerentemente reflexivos (IRVs): segundo as diretrizes de anotação, só devem ser anotados com essa etiqueta os casos em que o verbo nunca ocorre sem o reflexivo (como em "**se queixar**") ou quando o verbo sem o clítico adquire um sentido completamente diferente daquele do verbo acompanhado pelo clítico (como em "**se referir**"). No entanto, em casos como "se encontrar", essa mudança de sentido não fica muito clara, por exemplo, em "A banda encontrava-se em São Paulo". Nesse caso, optamos por não anotar a expressão como MWE.

A equipe de anotação se deparou ainda com algumas dificuldades para distinguir MWEs de metáforas ou colocações. Um exemplo do primeiro caso é a expressão "pisar no freio", visto que o substantivo pode ser substituído por "acelerador", mantendo a compreensão do sentido da expressão. Um exemplo do segundo caso é a expressão "realizar um sonho". Ainda que inicialmente ela possa ser uma forte candidata a MWE, a aplicação dos testes definidos pelas diretrizes de anotação mostram que ambos os elementos podem ser substituídos, como "realizar um desejo" ou "realizar uma tentativa", ou ainda "ter um sonho".

Como podemos ver nessa análise, anotar *corpora* permite não somente criar recursos para o desenvolvimento e avaliação de sistemas de PLN, mas também para estudar e compreender com mais profundidade um fenômeno linguístico.

5.5 Avaliação: colocando os pingos nos 'i's

Como descrito na Seção 5.3, podemos dividir as tarefas de processamento de expressões multipalavras em dois grupos: descoberta de MWEs e identificação de MWEs. Cada tipo

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

de tarefa tem objetivos diferentes e, por consequência, avaliações distintas.

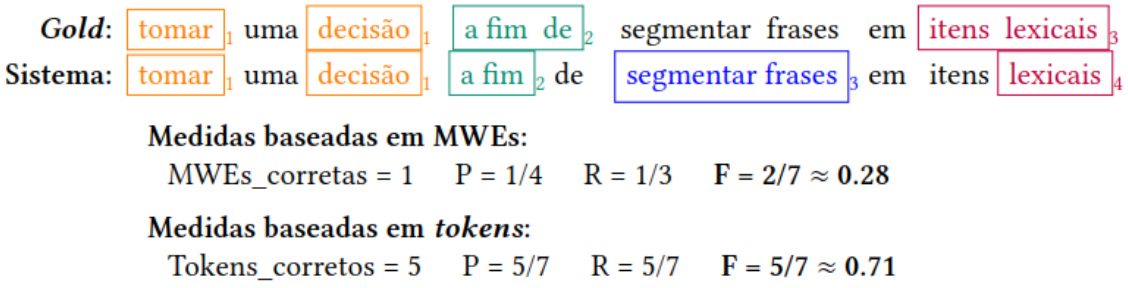
A descoberta de MWEs é uma tarefa particularmente difícil de se avaliar, pois as MWEs candidatas muitas vezes estão ausentes dos léxicos (Seção 5.4.1), exigindo avaliação por especialistas. Duas estratégias principais foram utilizadas para avaliar essa tarefa. A primeira é a comparação com listas de MWEs existentes, assumindo que as candidatas que estão incluídas no léxico são MWEs (Lin, 1999; Schone; Jurafsky, 2001). A segunda é a anotação manual de uma amostra das candidatas para a avaliação da precisão do método, sem no entanto avaliar sua cobertura (Evert; Krenn, 2001).

Nesta seção, detalharemos mais as medidas de avaliação para a tarefa de **identificação** de MWEs[21]. Para essa tarefa, a *shared task* DiMSUM foi pioneira em propor duas medidas de avaliação: a medida estrita e a baseada em *links*, para levar em conta predições parcialmente corretas (Schneider et al., 2016). Apesar dessas medidas serem interessantes, foi a *shared task* PARSEME que definiu as métricas que são hoje a referência para essa tarefa e as quais nós explicamos na sequência.

A ideia das medidas do PARSEME é comparar um conjunto de anotações de MWEs preditas por um sistema com um conjunto de anotações de referência (o *gold standard*). As medidas **baseadas em MWEs** são mais estritas, pois consideram que toda a expressão precisa ser predita corretamente, do início ao fim. As medidas **baseadas em *tokens*** são menos rigorosas, pois consideram parcialmente correto predizer parte de uma expressão (ainda que haja erros se considerarmos a MWE como um todo).

Em ambas as variantes, a qualidade das predições é medida por três valores: precisão (P), revocação (R, em inglês *recall*, também chamada de cobertura) e *F-score* (F), que é a média de P e R, como explicado na Subseção Métricas Baseadas em Conjuntos. Apenas a extensão das MWEs preditas é considerada, ou seja, quais palavras fazem ou não parte da expressão sem se avaliar o resto da sentença; e as categorias (Seção 5.2.4) preditas são ignoradas.

Figura 5.4: Exemplo: anotação de referência (acima, 3 MWEs) e predição do sistema (abaixo, 4 MWEs). O número de MWEs e *tokens* corretos é dividido pelo número total de MWEs/*tokens* preditos (P) ou de referência (R)



Fonte: Adaptado de https://gitlab.com/parseme/mwesinanutshell/

As medidas de avaliação baseadas em MWEs recompensam apenas combinações completas, considerando cada expressão como uma instância indivisível. Os valores de P e R, para esse tipo de avaliação, correspondem à proporção de MWEs completas que foram preditas corretamente (precisão) e identificadas (revocação). Por exemplo, na Figura 5.4, apenas a primeira expressão "**tomar decisão**" foi corretamente predita por inteiro. Por-

[21]Veja mais sobre Avaliação de sistemas de PLN no Capítulo 18.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

tanto, a precisão do sistema é de 1/4 (uma dentre quatro MWEs preditas estão corretas) e a revocação do sistema é 1/3 (uma dentre as três MWEs da referência foram corretamente identificadas).

As medidas de avaliação baseadas em *tokens* são similares, porém consideram os *tokens* que fazem parte das MWEs. No exemplo da Figura 5.4, dentre os sete *tokens* preditos como parte de uma MWE, cinco estão corretos (tomar, decisão, a, fim, lexicais), ou seja, a precisão é de 5/7. Como há sete *tokens* na referência anotados como parte de MWEs, a revocação também é de 5/7, por coincidência neste exemplo[22].

Além do tipo de medida (baseada em MWEs ou em *tokens*), a edição 1.1 da *shared task* PARSEME introduziu medidas especializadas, avaliando os sistemas somente em um subconjunto de MWEs que representam um **fenômeno (linguístico) específico** (Ramisch et al., 2018a). Isso significa que as medidas especializadas correspondem às medidas P, R e F baseadas em MWEs, mas calculadas apenas em um subconjunto das MWEs que apresentam determinadas características. A medida especializada em MWEs não vistas no *corpus* de treino foi o critério de avaliação principal na edição 1.2 da *shared task* PARSEME (Ramisch et al., 2020). Vários sistemas foram desenvolvidos para otimizar essas medidas, por exemplo, com relação a MWEs descontínuas (Taslimipoor et al., 2019).

5.6 Até onde chegamos e para onde vamos

Nos últimos anos, houve um interesse substancial da comunidade de PLN no tratamento de expressões multipalavras. Entre *shared tasks*, anotação de *corpora* e disponibilização dos mais diversos tipos de recursos, algoritmos e sistemas, tem-se evoluído no sentido de uma melhor compreensão sobre esse fenômeno, tanto do ponto de vista linguístico quanto do tratamento computacional dessas expressões, bem como seu impacto para diversas tarefas de PLN.

Nessa perspectiva, o PARSEME (muitas vezes citado neste capítulo) trouxe contribuições significativas em relação a *corpora* multilíngues, e é possível que, no futuro, o escopo das anotações desses *corpora* vá além das MWEs verbais[23]. No entanto, sabemos que uma andorinha só não faz verão, e há ainda muito espaço para pesquisas diversas, projetos novos e desenvolvimento amplo de trabalhos teóricos exploratórios e de tarefas aplicadas. Na sequência, apresentamos alguns dos pontos que consideramos que merecem atenção no futuro próximo.

Criação e manutenção de recursos

Apesar da falta de prestígio associada a essa tarefa, criar recursos ainda é importante na era dos modelos de linguagem em larga escala (os LLMs), como discutido na Seção 16.2. Em primeiro lugar, a supervisão humana em geral é muito mais eficaz do que aumentar o tamanho do modelo ou dos dados brutos (Ouyang et al., 2022). Em segundo lugar, a avaliação de modelos é essencial para entender e melhorar a tecnologia da linguagem, e só pode ser realizada com a ajuda de conjuntos de dados anotados (Haviv et al., 2023; Tayyar Madabushi et al., 2021). Por fim, criar recursos oferece material de análise para

[22] Para garantir que o sistema agrupe *tokens* corretamente, as MWEs são alinhadas, como indicado pelas cores na Figura 5.4, antes de se calcular as medidas baseadas em *tokens*.

[23] Se você está lendo este capítulo no referido futuro, em que já existem recursos para diversas categorias de MWEs além das verbais, saiba que chegamos aqui quando tudo isso era mato, e podemos ficar satisfeitas pelos avanços nesse quesito.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

teorias linguísticas e enriquece a descrição de fenômenos linguísticos ao fundamentá-los em dados reais. Todavia, esses recursos precisam ser robustos, homogêneos e confiáveis, o que significa que é necessário identificar e corrigir possíveis erros e inconsistências. Além disso, a disponibilidade desses recursos no longo prazo é crucial para a reprodutilibidade. Os artigos que descrevem conjuntos de dados frequentemente contêm *links* quebrados ou que direcionam a páginas pessoais de autoras e autores, onde os dados já não estão mais disponíveis. O trabalho de Ramisch (2023), por exemplo, mapeou 33 conjuntos de dados, mas conseguiu recuperar apenas 27 deles (alguns obtidos apenas via contato direto por e-mail). Logo, vale a pena considerar questões como disponibilização, manutenção e atribuição de licenças de livre acesso (sempre que isso for possível) aos recursos.

Diálogo entre projetos e equipes

Criar léxicos e *corpora* anotados é uma tarefa longa e trabalhosa, mas é ainda mais trabalhoso quebrar a cabeça tentando tornar compatíveis as anotações de diferentes níveis linguísticos que foram feitas isoladamente com base em um mesmo conjunto de textos. Uma abordagem mais interessante é que projetos grandes procurem garantir logo de início que os *corpora* sejam compatíveis nos mais diversos níveis de anotação linguística. Por exemplo, a interação entre o PARSEME e o UD, como o que foi proposto por Savary et al. (2023b), tem buscado compatibilizar tanto o nível de anotação das dependências quanto a anotação das MWEs[24]. Essa interação pode beneficiar todos os envolvidos, a partir da troca de experiências e de ferramentas de anotação e tratamento dos recursos. Além disso, a comunidade precisa se fazer presente em eventos e congressos da área, trocando experiências e desafios, especialmente em contextos multilíngues. Com isso, será possível manter alguma unidade (quando possível) em termos de decisões tomadas dentro de projetos ou mesmo entre os diversos projetos.

Diversidade linguística

A maior parte das pesquisas, comunidades acadêmicas e grandes laboratórios de PLN direcionam a sua atenção para uma única língua ou, no máximo, um conjunto limitado de línguas de prestígio. Você terá muito menos dificuldade para encontrar recursos e resultados sobre a língua inglesa, por exemplo, do que sobre o árabe, o swahili, o nheengatu, ou até mesmo o português. No contexto atual de busca pela diversidade, o desenvolvimento de pesquisas, sistemas e recursos precisa orientar seus esforços a cenários multilíngues e desafios de famílias linguísticas variadas. É evidente também a baixa representatividade de línguas minoritárias entre os recursos existentes. Projetos como o UD e o PARSEME já contam com algumas dessas línguas, mas os recursos são raros e geralmente muito pequenos para línguas menos prestigiadas cientificamente, por exemplo, dialetos locais, línguas indígenas, línguas de contato, línguas de imigração, entre outras. O céu é o limite em termos de variação linguística, então não precisamos nos limitar a poucas línguas já muito exploradas, como o inglês, o espanhol ou o francês. Mesmo nesses contextos, ainda são escassos os recursos advindos da fala, de níveis linguísticos populares ou informais, de contextos bilíngues ou mesmo da linguagem infantil. Logo, é fundamental que projetos que se ocupem das MWEs considerem a diversidade, não apenas de expressões nos *corpora*, mas também das línguas desses *corpora* (Seção 20.2).

[24] Esse é um dos objetivos do projeto UniDive: https://unidive.lisn.upsaclay.fr/

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Descrição aprimorada e ampliada de MWEs

As diretrizes multilíngues do PARSEME abrangem apenas expressões multipalavras verbais, mas outras categorias de MWEs têm sido descritas em projetos específicos de cada língua, por exemplo, para o francês (Candito et al., 2021) e para o espanhol (Parra Escartín et al., 2018). Portanto, estender as propostas de tipologia para MWEs nominais, modificadoras e funcionais a *corpora* multilíngues pode ser um próximo passo interessante para pesquisadoras e pesquisadores corajosos. Além disso, o tipo de anotação atual é bastante simples, dizendo apenas quais palavras pertencem à MWE, bem como sua categoria. No entanto, seria interessante modelizar informações mais detalhadas, como o grau de composicionalidade da expressão em contexto (Seção 5.4.1.1), ou ainda a vinculação das unidades textuais anotadas às unidades lexicais correspondentes, como entradas em uma Wordnet (Seção 9.1.1) ou *frames* semânticos (Seção 9.1.2).

Processamento cognitivo de MWEs

Expressões multipalavras são objetos linguísticos complexos, cujo estudo pode se inspirar em pesquisas de outras áreas. Na psicolinguística, expressões multipalavras e idiomaticidade são estudadas principalmente na perspectiva de abordagens baseadas no uso para aquisição de linguagem (Goldberg, 2005; Tomasello, 2015). Uma maior interação entre linguística computacional e psicolinguística cognitiva pode beneficiar ambas as áreas. Por um lado, descobertas sobre os mecanismos básicos de associação e memória que influenciam a aquisição de linguagem podem inspirar modelos de PLN, por exemplo, como viéses indutivos para arquiteturas neurais. Por outro lado, a simulação computacional pode ser usada para estudar a aquisição de linguagem, como em *corpora* de fala direcionada a crianças.

Por fim, defendemos aqui um olhar multidisciplinar e multilíngue para o fenômeno das expressões multipalavras. Reconhecemos os avanços obtidos até aqui com os LLMs, mas entendemos também que abordagens paralelas de construção de recursos e análises linguísticas podem beneficiar tanto a ciência quanto a aplicação prática desse conhecimento. Acreditamos ainda que não vale a pena olhar só para o próprio umbigo, então talvez o futuro das MWEs esteja na interação com outras comunidades de PLN e áreas de estudo, tanto da linguística quanto da computação.

5.7 Ao infinito e além

Neste capítulo, fizemos um apanhado dos conceitos e aspectos que julgamos mais interessantes no que se refere a pesquisa em MWEs, mas é evidente que há muitas mais fontes que você pode consultar, caso queira se aprofundar. Damos a seguir algumas dicas: fique à vontade para explorar esse universo por conta própria.

- O principal fórum para publicação e discussão dos avanços no tratamento computacional de MWEs é o *workshop* anual realizado em conjunto com as principais conferências em linguística computacional[25]. Esse *workshop* é organizado pela seção de MWEs do SIGLEX, e todos os anais dos eventos estão disponíveis na Antologia da Associação de Linguística Computacional (Association for Computational Linguis-

[25]Disponível em: https://multiword.org

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

tics, ou ACL)[26]. Outros *workshops* enfocam aspectos específicos do processamento de MWEs, como o MUMTTT, sobre a tradução de MWEs (Monti et al., 2017) [27].

- A coleção de livros *Phraseology and Multiword Expressions* publica livros sobre tópicos recentes na área[28]. Essa coleção é um dos resultados do projeto PARSEME, uma rede de pesquisadores na Europa que fez progressos significativos na área (Savary et al., 2015) [29]. O grupo construiu muitos recursos úteis, como uma lista de *treebanks* com identificação de MWEs (Rosén et al., 2016) [30] e uma lista de recursos lexicais de MWEs (Losnegaard et al., 2016). Além disso, a *shared task* do PARSEME sobre identificação verbal de MWEs lançou *corpora* anotados para mais de 20 idiomas[31].
- O SEMEVAL também apresenta tarefas relacionadas a MWEs, como classificação de substantivos compostos (Hendrickx et al., 2010), interpretação de substantivos compostos (Butnariu et al., 2010) e extração de *keyphrases* (Kim et al., 2010). Em 2016, a *shared task* DIMSUM do SEMEVAL enfocou a identificação de MWEs baseada em texto corrido, publicando *corpora* com anotação de MWEs para o inglês (Schneider et al., 2016) [32]. A tarefa 2 do SEMEVAL 2022 foi sobre predição e representação de idiomaticidade em contexto[33].

Agradecimentos

Agradecemos ao apoio dos projetos PARSEME (COST IC1207), UniDive (COST CA21167), PARSEME-FR (ANR-14-CERA-0001), SELEXINI (ANR-21-CE23-0033-01) e EPSRC projeto MIA: Modeling Idiomaticity in Human and Artificial Language Processing (EP/T02450X/1). Partes do capítulo são baseadas em materiais realizados com a contribuição de Agata Savary, Silvio Cordeiro, Marie Candito e Mathieu Constant, entre outras colegas, coautoras e coautores.

[26]Disponível em: https://aclanthology.org/
[27]Última edição no momento da escrita do capítulo: http://europhras.com/2022/mumttt-2022-2/
[28]Disponível em: https://langsci-press.org/catalog/series/pmwe
[29]Disponível em: http://parseme.eu
[30]Disponível em: https://clarino.uib.no/iness/page?page-id=MWEs_in_Parseme
[31]Disponível em: https://gitlab.com/parseme/corpora/-/wikis/home
[32]Disponível em: https://dimsum16.github.io
[33]Disponível em: https://sites.google.com/view/semeval2022task2-idiomaticity

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Parte IV

Estrutura

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Capítulo 6

A ordem e a função das palavras em uma sentença

Sintaxe

Adriana S Pagano
Amanda Rassi
Ana Clara S Pagano

Publicado em: 26/09/2023

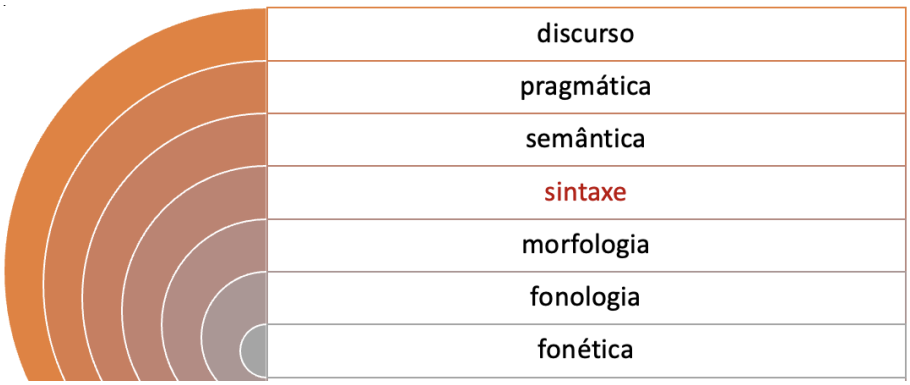
https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

6.1 Introdução

Nos capítulos anteriores, vimos algumas das unidades de análise que são examinadas nas distintas subáreas dos estudos linguísticos, cada uma das quais com desafios específicos para o PLN. Neste capítulo, nosso foco será uma subárea em particular – a sintaxe –, que estuda como as palavras se organizam nas estruturas que constroem as distintas funções gramaticais no escopo da frase ou sentença.

Retomando nossa representação do estudo da linguagem (Figura 1.2 do Capítulo 1), podemos dizer que a sintaxe é um estrato/camada central do sistema linguístico (Figura 6.1), pois organiza funções no estrato imediatamente inferior – a morfologia – e fornece estruturas de funções que serão importantes nos estratos superiores – semântica e pragmática – para, por exemplo, identificar papéis temáticos cruciais na tarefa de extração de informação, os quais respondem as perguntas "quem?" faz "o quê?", "para quem?", "como?", "quando?", "onde?" etc.

Figura 6.1: Representação das subáreas de estudo da linguagem com destaque para a sintaxe

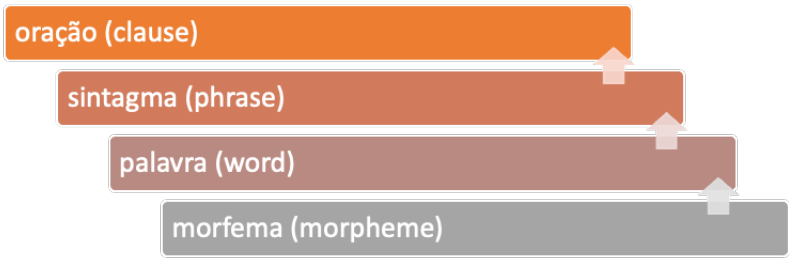


Ainda em relação ao sistema linguístico, além de sua organização em estratos, temos uma segunda forma de organização: a escala de ordens (em inglês, *rank scale*), que organiza

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

diferentes unidades de análise em níveis hierárquicos. A escala de ordens[1] está representada na Figura 6.2. Os morfemas são as menores unidades constitutivas das palavras, as quais, por sua vez, se organizam em estruturas, chamadas sintagmas, que nos permitem construir funções na oração, tais como sujeito, objetos, adjuntos e complementos.

Figura 6.2: Escala de ordens



A escala de ordens abrange as unidades utilizadas para a descrição da língua, seja esta língua falada ou escrita. A unidade superior de análise – a oração – consiste em um conjunto organizado de palavras que constroem significados sobre algum evento no mundo, que, como dissemos, podemos indagar perguntando "quem?" faz "o quê?", "para quem?", "como?", "quando?", "onde?" etc.

Na língua escrita, há uma unidade grafológica, geralmente denominada, em português, "sentença", também conhecida como "frase" ou "período"[2]. A "sentença" segue as convenções prototípicas da língua escrita, isto é, letra maiúscula inicial e sinal de pontuação que indica finalização, podendo ser este um ponto final, um sinal de interrogação ou de exclamação. A distinção entre "sentença" (em inglês, *sentence*) e "oração" (em inglês, *clause*) é muito importante em PLN, uma vez que em uma sentença (aquela que inicia com letra maiúscula e conclui com sinal de pontuação) pode haver mais de uma oração e, portanto, mais de um evento.

Neste capítulo, vamos abordar os níveis da escala de ordens e sua relevância para a análise sintática. Para compreendermos melhor como as palavras funcionam nos distintos tipos de sintagma e os sintagmas na oração, apresentaremos exemplos que esperamos que sejam esclarecedores. Após uma seção de reflexões iniciais, revisaremos conceitos básicos sobre análise sintática, passando por tipos de representação, até chegarmos nos dois tipos de análise sintática mais utilizados em PLN: a sintaxe de constituência e a sintaxe de dependência. Por último, abordaremos as intersecções entre a sintaxe e os demais estratos do sistema linguístico ao final deste capítulo, na Seção 6.8.

Antes de passarmos para nossas reflexões iniciais sobre sintaxe, cabe mencionar que o processo de analisar a estrutura de orações em PLN é denominado "*parsing*", termo tomado por empréstimo do inglês. O *parsing* sintático toma como base a classe de palavra (em inglês, *part-of-speech*) das distintas palavras que compõem os sintagmas. Como veremos no Capítulo 7, em PLN a análise sintática automática é realizada por meio de softwares denominados *parsers*. Também veremos alguns desafios que a análise em cons-

[1] Cabe aqui um esclarecimento sobre a nomenclatura. Em língua inglesa, utilizamos "*phrase*" para nomear o nível da escala de ordens acima da palavra e abaixo da oração, enquanto "*syntagma*" nomeia a organização das funções numa unidade. Assim, toda "*phrase*" possui uma organização de funções ou "*syntagma*". Na língua portuguesa, a palavra "sintagma" é utilizada para se referir ao que em inglês é denominado de "*phrase*" e de "*syntagma*".

[2] Em estudos de sintaxe em português, "sentença" nomeia a maior unidade de análise sintática, a qual também pode ser nomeada como "frase" ou "período" (Kenedy; Othero, 2018).

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

tituintes sintáticos traz para os *parsers* já existentes, sobretudo em casos nos quais a delimitação de unidades e suas relações entre si na hierarquia da oração admitem mais de uma interpretação possível.

Um outro termo que é importante introduzir neste momento é o conceito de *treebank*, que é utilizado para se nomear um conjunto de sentenças com anotação morfossintática e com representação em diagramas de árvore. Nas seções seguintes, veremos alguns exemplos de diagramas de árvore juntamente com outras formas de representação. No Capítulo 7, veremos alguns dos *treebanks* disponíveis atualmente para PLN em português.

6.2 Reflexões Iniciais

Vamos começar examinando um texto, no Exemplo 6.1. Trata-se de um texto pequeno e simples, reproduzido aqui sem qualquer letra maiúscula e sem pontuação. Até o espaçamento entre palavras foi removido. À medida que for lendo, tente identificar sentido nele:

Exemplo 6.1:

ameninapostouumafoto

Embora a forma como o texto é apresentado acima não seja a mais esperada por nós num texto hoje em dia, ao menos nas línguas de origem europeia contemporâneas, essa forma já foi utilizada na antiguidade, no latim clássico, e é conhecida como *scriptio continua*. De fato, a separação visual entre palavras e sentenças, tal como escrevemos hoje, foi um desenvolvimento histórico que levou à forma como representamos hoje a língua na sua forma escrita. Contudo, apesar do estranhamento que possa causar o texto no Exemplo 6.1, muito provavelmente você não teve muita dificuldade em reinserir o espaçamento entre palavras, chegando a uma versão como a que segue no Exemplo 6.2:

Exemplo 6.2:

a menina postou uma foto

Pensando, ainda, numa versão impressa deste texto, a letra maiúscula inicial e a pontuação final certamente não geram grandes dificuldades e você deve ter chegado à versão no Exemplo 6.3:

Exemplo 6.3:

A menina postou uma foto.

O reconhecimento de cada uma das palavras, separadamente, e da sentença toda como uma unidade, todavia, não é suficiente para podermos explicar por que ou como o texto faz sentido para nós. Entre as palavras individuais e a sentença como um todo, há agrupamentos que funcionam como unidades intermediárias e nos quais a ordem das palavras é condicionada pela estrutura da língua. Assim, na sentença acima, identificamos palavras que agrupamos como pequenos blocos de informação, pois constroem significados e contribuem com o significado do texto como um todo. Esses agrupamentos são, por exemplo, "a menina" e "uma foto".

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Em cada um desses agrupamentos de palavras que naturalmente reconhecemos, há uma ordem em que as palavras se sucedem umas às outras. Assim, uma palavra como "a" (artigo) ocorre antes de "menina" (substantivo), nunca depois. Quando a ordem das palavras dentro desses agrupamentos é trocada, temos uma forma que consideramos com problemas na sua formulação e que nos causa estranhamento. Por exemplo: "menina a" ou "foto uma".

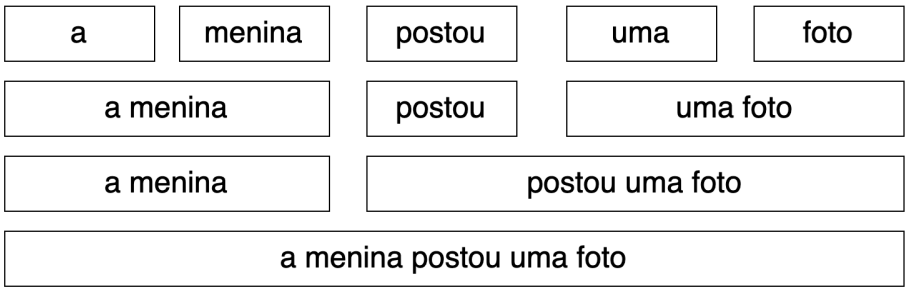
Com estas reflexões iniciais, passamos agora a revisar alguns conceitos básicos de sintaxe.

6.3 Noções Básicas de Sintaxe

O estudo de como as palavras se agrupam e se organizam em uma ordem determinada é o objeto do campo de estudos da sintaxe. É através da sintaxe que reconhecemos as regras que ditam quais agrupamentos de palavras serão aceitos por nós e quais serão considerados problemáticos.

Os agrupamentos operam como pequenos "tijolos" que vão construindo significados, ao serem agrupados em uma unidade maior até construir um significado que é relevante para a oração. Nessa analogia com tijolos que são agrupados, o Exemplo 6.3 pode ser representado como disposto na Figura 6.3:

Figura 6.3: Representação do agrupamento progressivo de unidades menores em maiores



Olhando para a Figura 6.3, temos, no topo, cada uma das palavras do nosso exemplo. À medida que vamos descendo na Figura 6.3, vemos agrupamentos entre as palavras, os quais, progressivamente, culminam na oração. Os agrupamentos mais próximos da oração são aqueles que constroem os significados mais relevantes para a compreensão de informações, como as que respondem às perguntas elencadas no Quadro 6.1:

Quadro 6.1: Perguntas utilizadas para elucidar informações sobre eventos

Pergunta	**Resposta**
Quem?	a menina
Fez o quê?	postou uma foto
Postou o quê?	uma foto

Em cada um dos agrupamentos de palavras que vamos progressivamente formando até chegar ao agrupamento maior, a oração, as palavras trabalham para realizar uma função em comum; elas são uma única unidade funcionando dentro da oração. Dito de outra forma, cada um desses agrupamentos existe pois exerce uma função dentro da oração. Por

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

exemplo, "a menina" é uma unidade constituída por um artigo, também chamado determinante ("a"), e um substantivo ("menina"). Esses dois elementos operam conjuntamente para constituir uma função e realizar um significado. Os agrupamentos acima do nível da palavra, como vimos na escala de ordens na Figura 6.2, recebem a denominação **sintagma** (em inglês, *phrase*). Cada um dos sintagmas é uma estrutura que é parte de uma unidade maior e que exerce uma função nela, sendo que esta outra estrutura pode ser um outro sintagma ou a oração.

Existem vários tipos de sintagma: sintagmas nominais, verbais, adverbiais, adjetivais e preposicionais, e cada um possui as suas próprias regras para a organização das palavras que o compõem, de acordo com a função que as palavras exercem dentro do próprio sintagma.

A cada uma das unidades que funciona dentro de um sintagma damos o nome de constituinte (em inglês, *constituent*). São constituintes as palavras individuais bem como seus agrupamentos progressivos em unidades maiores. Nossa sentença "A menina postou uma foto" está constituída por cinco palavras, as quais podem ser agrupadas em constituintes intermediários até dois grandes constituintes: [a menina] e [postou uma foto].

Alguns exemplos de constituintes no nível da palavra são: substantivos, verbos, adjetivos etc.

Retomando o Exemplo 6.3 temos, então, os seguintes sintagmas: "a menina", que é um sintagma nominal, pois o seu constituinte principal é o substantivo "menina"; e "postou uma foto", que é um sintagma verbal, pois seu constituinte principal é o verbo "postou"; dentro do sintagma verbal temos ainda um outro sintagma, "uma foto", que é considerado um sintagma nominal, pois seu constituinte principal é o substantivo "foto". É importante notar, portanto, que pode haver um ou mais sintagmas dentro de um outro sintagma. Esse constituinte principal que dita qual é o tipo de sintagma é chamado de núcleo, pois ele é o núcleo da estrutura gramatical, ou seja, do sintagma.

Para saber com que tipo de sintagma estamos lidando, precisamos, então, saber que fatores ditam qual é o núcleo desse sintagma. Isso irá depender da função que o sintagma exerce na oração. Pensemos no exemplo: o sintagma nominal "uma foto" exerce a função de objeto dentro do sintagma "postou uma foto". Objetos são realizados por sintagmas nominais; assim, o seu núcleo é um substantivo.

É importante ressaltar, também, que esses constituintes não se restringem ao nível imediatamente inferior na oração, nem a nenhum outro nível: constituintes são os componentes que integram todas as estruturas da sentença, ou seja, todos os sintagmas, e podem ser compostos por outros constituintes. Daí a relevância da escala de ordens apresentada na Figura 6.2.

No Exemplo 6.3, os sintagmas "a menina", "postou uma foto" e "uma foto" têm uma função no nível superior da escala de ordens, isto é, na oração. Na oração, essas funções podem ser associadas ao que chamamos, na Semântica, de papéis temáticos.

Papel temático é um conceito utilizado para nomear o tipo de relação que um verbo estabelece com seu sujeito e seus complementos, relação pela qual o verbo lhes atribui uma função semântica, como, por exemplo, Agente, Paciente ou Objeto de uma ação. Os papéis temáticos podem ser mapeados com as funções sintáticas na oração e podem ser indagados por meio de perguntas do tipo "quem?" faz "o quê?", "para quem?", "como?", "quando?", "onde?" etc., como as exemplificadas no Quadro 6.2. As respostas a cada pergunta nos ajudam a identificar um constituinte com uma função sintática específica na oração e um papel temático. No Exemplo 6.3, a parte da oração que responde a pergunta "quem?" correspondente ao sujeito da oração, que, neste caso, é identificado com o papel

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

temático de Agente da ação realizada. Já a parte que responde a "o quê?" corresponde ao objeto do verbo e ao papel temático Objeto Estativo.

Quadro 6.2: Exemplo de funções sintáticas e papéis temáticos

Pergunta	**Resposta**	**Função sintática**	**Papel temático**
Quem?	a menina	sujeito	Agente
Fez o quê?	postou uma foto	predicado	—
Postou o quê?	uma foto	objeto	Objeto Estativo

Como vimos, então, um sintagma pode ser construído por uma ou mais palavras, sendo uma delas a principal, o núcleo. A estrutura dos constituintes é o que conhecemos como sintagma e o tipo de sintagma é definido pela classe de palavra do seu núcleo ou componente principal. No Quadro 6.3, temos os três constituintes da oração no Exemplo 6.3 e o tipo de sintagma de cada um deles.

Quadro 6.3: Exemplo de tipos de sintagma

Sintagma	**Núcleo**	**Classe de palavra do núcleo**	**Tipo de sintagma**
a menina	menina	substantivo	sintagma nominal
postou uma foto	postou	verbo	sintagma verbal
uma foto	foto	substantivo	sintagma nominal

Palavras e sintagmas podem, assim, ser constituintes, sempre funcionando numa estrutura maior, na escala de ordens, tendo como unidade maior a oração. A análise da estrutura das orações de acordo com seus constituintes e a hierarquia estabelecida entre eles é conhecida como **sintaxe de constituência**. A análise sintática de constituência é um dos dois tipos de análise sintática mais utilizados em PLN, sendo o outro tipo a denominada **sintaxe de dependência**. Neste capítulo, vamos apresentar os dois tipos de análise e discorrer brevemente sobre suas diferenças. Mas, antes, vamos examinar os principais tipos de representação em sintaxe.

6.4 Tipos de representação

Há algumas formas convencionais de se representar as estruturas sintáticas que podem ser utilizadas de acordo com o tipo de análise sintática. Dentre elas, destacamos as seguintes: colchetes, árvores, setas, parênteses e indentação.

6.4.1 Colchetes

Colchetes (em inglês, *brackets*) são uma forma de representação hierárquica por meio da qual indicamos quais palavras estão agrupadas dentro de uma mesma unidade e quais unidades são parte de unidades maiores. Por exemplo, a sentença "A menina postou uma foto." pode ser representada, de forma progressiva, dos constituintes maiores aos menores, como ilustrado na Figura 6.4.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Figura 6.4: Representação de hierarquia de constituintes por meio de colchetes

[A menina postou uma foto.]
[A menina postou uma foto] | [.]
[A menina] | [postou uma foto] | [.]
[A] | [menina] | [postou] | [uma foto] | [.]
[A] | [menina] | [postou] | [uma] | [foto] | [.]

No topo, temos um único par de colchetes que abrange a sentença toda. Em seguida, separamos, entre colchetes, a oração e seu ponto final. Em seguida, separamos, entre colchetes, os constituintes maiores e, progressivamente, vamos separando, entre colchetes, constituintes dentro de constituintes, até chegarmos às palavras individuais, que são nossos constituintes mínimos. Barras verticais auxiliam a visualização na Figura 6.4.

Na representação por meio de colchetes, cada par de colchetes indica um **nível de continência**. O par de colchetes mais externo contém a sentença como um todo. Já o par de colchetes mais interno de todos contém as palavras. As unidades menores encontram-se contidas nas maiores, sendo as unidades mínimas cada uma das palavras individualmente. A representação completa pode ser feita numa única linha:

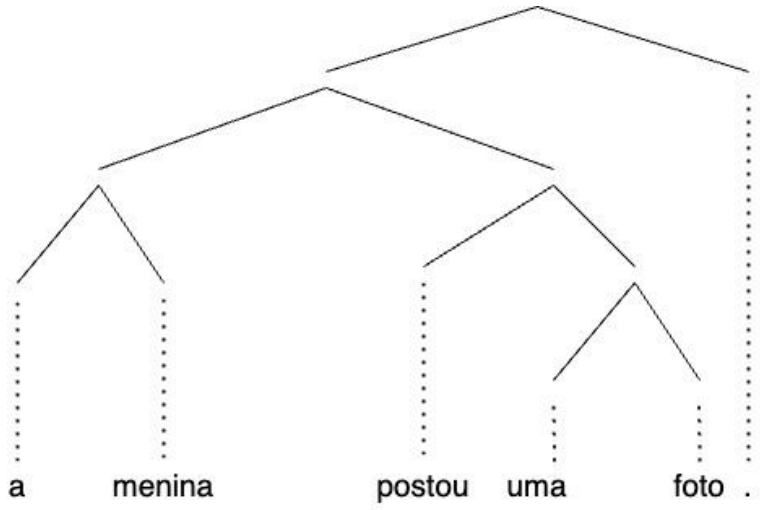
[[[[[**A**][**menina**]][[**postou**][[**uma**][**foto**]]]][**.**]]]

Este tipo de representação, como veremos neste capítulo, pode ser feito tanto para sintaxe de constituência como para sintaxe de dependência.

6.4.2 Árvores

Árvores (em inglês, *trees*) são uma forma de representação que fornece uma visualização mais clara da hierarquia dos constituintes. Assim, nosso exemplo anterior tem a seguinte representação arbórea na Figura 6.5:

Figura 6.5: Representação de hierarquia de constituintes por meio de árvore



https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

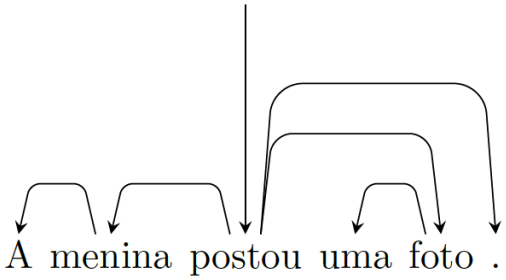
Na notação com estruturas arbóreas, um diagrama chamado de árvore representa graficamente a estrutura de uma sentença como uma hierarquia, sendo os **nós** (em inglês, *nodes*), elementos discretos (ou elementos finais) em um grafo (ou representação simbólica) e as **arestas** (em inglês, *edges*) linhas que conectam dois nós e indicam uma relação entre eles.

Este tipo de representação, como veremos neste capítulo, pode ser feito tanto para sintaxe de constituência como para sintaxe de dependência.

6.4.3 Setas

Setas (em inglês, *arced arrows*) são uma representação na qual setas unidirecionais são desenhadas de um **nó pai** (em inglês, *parent*) para um **nó filho** (em inglês, *child*), como pode ser visto na Figura 6.6. Nela, temos uma representação de relações de dependência entre nós pais e filhos. A seta parte da palavra que governa a relação e chega na palavra dependente. Essa relação será melhor explicada na Seção 6.6.

Figura 6.6: Representação de relações de dependência por meio de setas unidirecionais



A seta central, que chega no verbo "postou" é apenas uma representação conceitual de que esse é o elemento central da sentença. As setas que saem de "postou" e chegam em "menina", "foto" e "." indicam que, nessas relações, "postou" é o pai e "menina", "foto" e "." são filhos de "postou". A seta que sai de "menina" e chega em "a" indica que "a" é filho de "menina", assim como a seta de "foto" para "uma" indica que "foto" é pai de "uma".

Este tipo de representação, como veremos mais adiante neste capítulo, é prototípico da análise sintática de dependência.

6.4.4 Parênteses

Parênteses (em inglês, *parentheses*) são usados para estabelecer a relação entre pares de palavras. Nesta representação, cada um dos pares de palavras entre as quais se estabelece uma relação unidirecional de dependência é apresentado entre parênteses, sendo a primeira posição a da palavra que governa a relação e a segunda a da palavra dependente.

Na linha 1, temos a relação conceitual de `root`, que é representada pelo elemento vazio, o qual governa o verbo "postou", considerado o elemento central na sentença. Assim como na representação por setas, temos nas linhas 2, 4 e 6, as relações de dependência em que o verbo "postou" governa seus dependentes "menina", "foto" e ".", respectivamente. Na linha 3, vemos a dependência de "a" (dependente) em relação a "menina" (governante), assim como na linha 5, em que "uma" é dependente de "foto".

Este tipo de representação é utilizado pela análise sintática de dependência.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Figura 6.7: Representação de relações de dependência por meio de parênteses

```
1        (    , postou)
2        (postou, menina)
3        (menina, a)
4        (postou, foto)
5        (foto, uma)
6        (postou, .)
```

6.4.5 Indentação

Indentação (em inglês, *indentation*) também é usada para explicitar a hierarquia entre as palavras. Este tipo de representação nos permite visualizar a hierarquia nas relações de dependência.

Figura 6.8: Representação de relações de dependência por meio de indentação

```
postou
  menina
    a
  foto
    uma
```

Na Figura 6.8, os dependentes ou nós filhos, por exemplo "menina" e "foto", são indentados em relação ao nó pai "postou". Os nós "a" e "uma" são indentados em relação aos nós "menina" e "foto", respectivamente.

Este tipo de representação é utilizado pela análise sintática de dependência.

Como veremos nas seções seguintes, cada formato de representação pode ser mais ou menos adequado para cada tipo de análise sintática. Neste capítulo, vamos explicar as duas abordagens mais comuns de análise sintática que fundamentam os dois tipos de *parsing* mais utilizados em PLN: o *parsing* de constituência e o *parsing* de dependência.

6.5 Sintaxe de constituência

Na sintaxe de constituência, também chamada de sintaxe de constituintes, unidades chamadas constituintes são agrupadas em unidades maiores. Os constituintes podem ser palavras ou distintos tipos de sintagmas, podendo haver sintagmas contidos dentro de sintagmas maiores. Esse fenômeno implica uma hierarquia de unidades, que é capturada e visualizada por meio de representações como as que vimos na seção anterior.

Na sintaxe de constituência, a estrutura hierárquica pode ser representada com o uso de colchetes, que, como vimos anteriormente, indicam quais constituintes inferiores estão contidos nos constituintes hierarquicamente superiores; ou com estruturas arbóreas,

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

nas quais no topo da árvore encontra-se a sentença completa e, nos níveis inferiores, os constituintes menores. Em ambos os casos, cada constituinte está acompanhado de sua respectiva notação, de acordo com um conjunto específico de etiquetas morfossintáticas (em inglês, *tagset*).

A análise sintática tem como ponto de partida a unidade maior – a sentença – representada, prototipicamente, pela notação **S** (do inglês, *sentence*). Se uma sentença tem pontuação final, o sinal de pontuação recebe a notação **PNT** (do inglês, *punctuation*) e é anotado no mesmo nível hierárquico de **S**. De **S**, derivam constituintes cujas estruturas realizam tipicamente as funções sintáticas no nível da oração: sujeito, geralmente realizada por um constituinte com estrutura de sintagma nominal, que recebe a notação **NP** (do inglês, *noun phrase)*; e predicado, geralmente realizada por um constituinte com estrutura de sintagma verbal, com a notação **VP** (do inglês, *verb phrase*), podendo ter objetos com estrutura prototípica de sintagma nominal (**NP**). Há também frases preposicionais, anotadas como **PP** (do inglês, *prepositional phrase*). Nos níveis seguintes, cada constituinte é subdividido em subconstituintes até chegarmos às palavras individuais, as quais recebem a notação de sua classe: **V** para verbo, **N** para substantivo, **DET** para determinantes como artigos e pronomes demonstrativos, **ADJ** para adjetivo, **ADV** para advérbio, e **P** para preposição.

Vejamos a representação com colchetes e notação para a sentença: "A menina postou uma foto.".

[ROOT [S [S [NP [DET A] [N menina]] [VP [V postou] [NP [DET uma] [N foto]]]] [PNT .]]]

Como pode ser observado, a cada palavra é atribuída uma classe de palavra de acordo com um conjunto de etiquetas definido. Nesta representação, o par de colchetes mais externos contém a sentença como um todo, incluindo a pontuação, e recebe a etiqueta **ROOT** (em português, raiz) que indica o ponto inicial da análise. Já os pares de colchetes mais internos de todos contêm as palavras individuais com sua respectiva notação para classe de palavra.

Na notação com estruturas arbóreas, a sentença: "A menina postou uma foto." é representada na forma de árvore[3] na Figura 6.9.

A árvore pode ser lida, de cima para baixo, da seguinte forma: O nó denominado **ROOT** indica a raiz da árvore. O nó **S1** representa a sentença toda como um único constituinte; é o nó que domina todos os outros nós da árvore. No caso de uma sentença grafológica (unidade delimitada por um sinal de pontuação), como é o caso na Figura 6.9, temos uma representação da sentença como um todo (**S1**), abrangendo a sentença (**S2**) e o seu sinal de pontuação (**PNT**). De **S2** derivam constituintes cujas estruturas realizam tipicamente as funções sintáticas no nível da oração: sujeito realizado por um constituinte com estrutura de sintagma nominal (**NP**) e predicado, por um constituinte com estrutura de sintagma verbal (**VP**). Nos níveis seguintes, cada constituinte é subdividido em subconstituintes até chegarmos às palavras individuais.

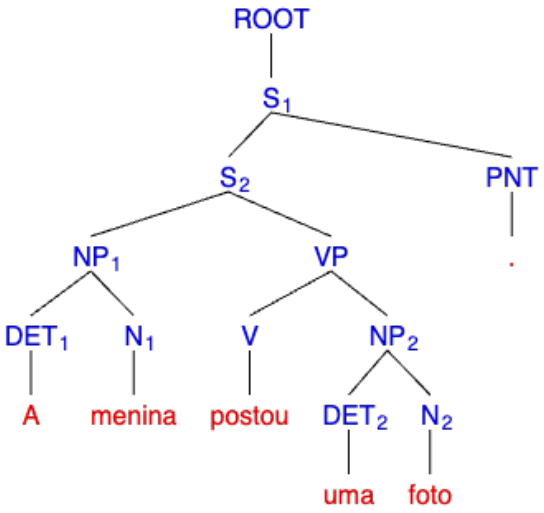
Na árvore de constituência, as arestas representam segmentos que conectam os nós filhos a um nó pai. Assim, o nó **NP** ("a menina") possui uma aresta que conecta **DET** ("a") e

[3]As árvores de constituência utilizadas neste capítulo foram elaboradas com o software jssyntaxfree, disponível em https://github.com/int2str/jssyntaxtree, com base em anotação das autoras. O conjunto de etiquetas utilizado é o adotado pelo projeto PortulanClarin, disponível em https://portulanclarin.net/.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Figura 6.9: Representação de sintaxe de constituência em diagrama de árvore



N ("menina"). As palavras individuais são as folhas da árvore (em inglês, *leaves*) e são o nível inferior na hierarquia da árvore.

Identificar os constituintes e seu lugar na hierarquia da oração é uma operação fundamental para a compreensão das funções sintáticas, que, como vimos, são a base para as funções semânticas da oração. Uma vantagem da notação com estrutura arbórea é a maior facilidade com a qual podemos visualizar as relações hierárquicas dos constituintes. Contudo, a notação com colchetes é computacionalmente mais fácil de ser processada.

6.5.1 Possibilidades de interpretação e ambiguidades sintáticas

Vejamos agora um exemplo de como a análise de constituintes requer uma análise dos papéis temáticos.

Exemplo 6.4:

> A menina postou uma foto com o celular.

No Exemplo 6.4, "com o celular" é um sintagma preposicional (**PP**), formado pela preposição "com" e o sintagma nominal "o celular". Uma característica dos sintagmas preposicionais é que eles podem estar contidos num sintagma nominal ou num sintagma verbal. No Exemplo 6.4, "com o celular" pode estar contido no sintagma nominal "uma foto" ou no sintagma verbal "postou uma foto". A decisão por considerar uma ou outra análise depende da interpretação do analista, tendo como apoio informações do contexto, ou seja, as informações que podemos depreender de outras partes do texto ou ao ter acesso a imagens, como, por exemplo, na Figura 6.10[4].

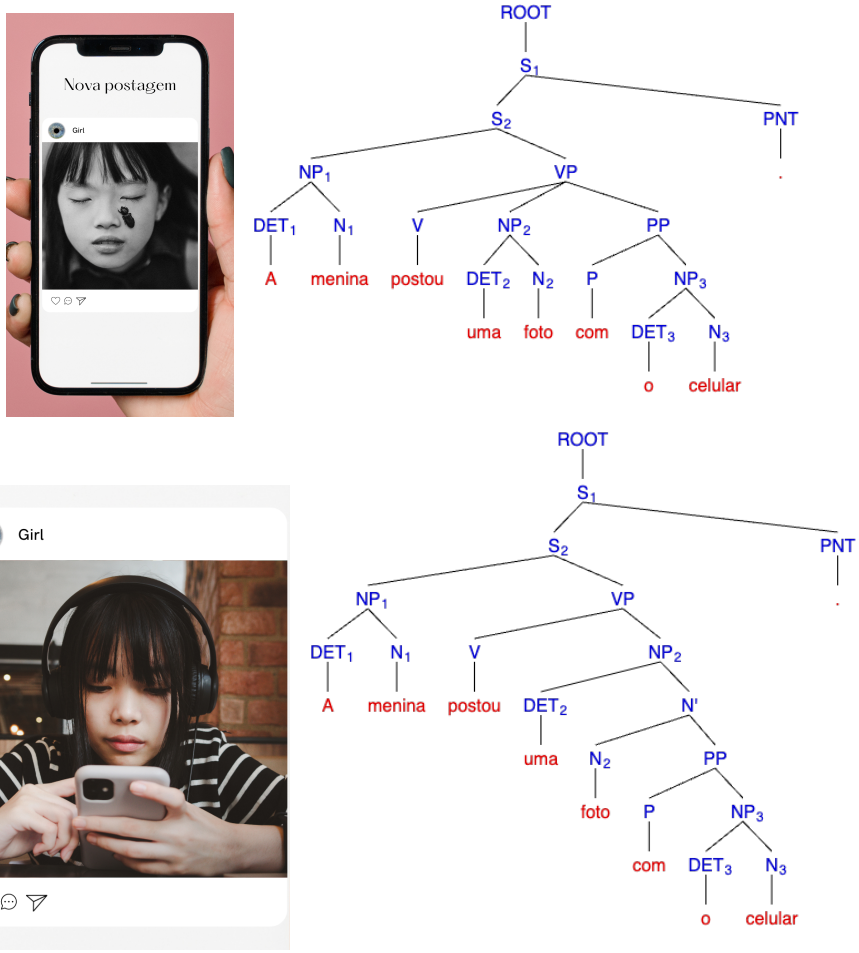
Na Figura 6.10, a imagem de cima pode ser interpretada como uma foto que foi postada com o celular, sendo o celular o instrumento utilizado para a postagem. Nesse caso, o sintagma preposicional "com o celular" estará ligado ao sintagma verbal "postou", como mostrado na representação da árvore.

[4]Imagens geradas com o aplicação https://gencraft.com/ e modificadas pelas autoras.

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

Figura 6.10: Duas imagens passíveis de serem rotuladas como “a menina postou uma foto com o celular”



Já a imagem de baixo, na Figura 6.10, pode ser interpretada como a menina tendo postado uma foto, na qual ela está segurando o celular. Nesse caso, o sintagma preposicional “com o celular” estará contido no sintagma nominal “uma foto com o celular”, como mostrado na representação da árvore.

Há alguns testes que podem ser realizados para observar se um constituinte, por exemplo, um sintagma preposicional, deve ser interpretado como vinculado a um sintagma verbal ou contido em um sintagma nominal. Se o sintagma preposicional pode ser deslocado de posição numa oração e esse deslocamento é condizente com o significado que interpretamos pelas dicas do contexto, ele estará conectado ao sintagma verbal. Assim, em “A menina postou uma foto com o celular”, podemos mudar a posição do sintagma preposicional “com o celular” e obtermos duas sentenças nas quais “com o celular” se conecta ao sintagma verbal “postou”:

Exemplo 6.5:

> Com o celular, a menina postou uma foto.

Uma outra possibilidade é fazer uma paráfrase da oração, como no Exemplo 6.6. Por outro lado, se o sintagma preposicional “com o celular” for interpretado como sendo contido

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

no sintagma nominal "uma foto com o celular", a paráfrase poderia ser a apresentada no Exemplo 6.7.

Exemplo 6.6:

> A menina postou uma foto por meio do celular.

Exemplo 6.7:

> A menina postou uma foto na qual ela está segurando o celular.

Se considerarmos as duas interpretações do Exemplo 6.4 sob a perspectiva dos papéis temáticos, podemos afirmar que o sintagma preposicional "com o celular" terá o papel temático de Instrumento quando estiver vinculado ao sintagma verbal "postou". Se o sintagma preposicional estiver contido dentro do sintagma nominal "uma foto com o celular", o sintagma como um todo terá o papel temático de Objeto Estativo.

A decisão sobre como segmentar a oração em constituintes exige do analista humano a interpretação de significados que são construídos pelas distintas funções sintáticas, associadas aos papéis temáticos e tendo como apoio o contexto, isto é, informações situacionais (por exemplo, imagens que acompanham a linguagem verbal) ou de sentenças anteriores ou posteriores no próprio texto do qual uma sentença faz parte.

Esta situação é um exemplo representativo de um dos motivos pelos quais o conhecimento linguístico e de uso da língua ainda não têm sido completamente representado/capturado por nenhum dos métodos atuais. A interpretação de significados linguísticos que dependem de informações contextuais é um dos maiores desafios para o PLN atualmente.

Além da análise sintática de constituintes da oração, há uma segunda abordagem, conhecida como sintaxe de dependência. Ela será o objeto da seção seguinte.

6.6 Sintaxe de dependência

O segundo tipo de análise sintática é chamado de sintaxe de dependência e tem suas bases na gramática de dependência de Tesnière (1959).

Diferentemente da abordagem de constituência, explicada na seção anterior, em que a estrutura de uma sentença é definida por meio de sintagmas contidos em outros sintagmas, a análise de dependência descreve as relações de **dependência** entre palavras. Nessa abordagem, uma palavra é vista como subordinada a outra ou regida por ela, de acordo com relações sintáticas tais como sujeito-verbo; sujeito-objeto; verbo-objeto; coordenação; subordinação etc.

Na sintaxe de dependência, cada palavra é um nó de uma relação com uma outra palavra. Essas relações entre palavras são estabelecidas de forma unidirecional entre uma palavra regente (*head*), que é o nó de onde a relação parte, e uma palavra regida ou dependente, que é o nó ou palavra aonde a relação chega. Essa unidirecionalidade da relação é importante para se estabelecer a hierarquia entre as palavras, pois determina quem é o regente (de onde a relação parte) e quem é o regido (aonde a relação chega), o que será explicado na Seção 6.6.1. Por fim, é importante destacar que uma palavra pode reger várias outras, mas só pode ser regida por uma única.

Se quisermos conectar dois sintagmas, precisamos, diferentemente da análise de constituintes, conectar a palavra que representa o núcleo de um sintagma à palavra que representa

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

o núcleo de outro sintagma. Além disso, também é necessário estabelecer as microrrelações dentro de um mesmo sintagma.

Na sintaxe de dependência, há basicamente dois tipos de relações:
(i) macrorrelações
(ii) microrrelações.

Como o próprio nome sugere, as **macrorrelações** estabelecem relações entre os núcleos de diferentes sintagmas e geralmente conectam palavras de classe aberta. Por exemplo, a relação que liga um verbo (núcleo do sintagma verbal) ao seu sujeito (núcleo do sintagma nominal) é chamada de macrorrelação. Já as **microrrelações** conectam elementos mais próximos, podendo ser adjacentes ou estar em uma vizinhança próxima. As microrrelações de dependência geralmente conectam uma palavra de classe aberta a uma palavra de classe fechada, como é o caso de um substantivo (palavra de classe aberta) e seu artigo (palavra de classe fechada).

6.6.1 Núcleo e dependente

Para estabelecer a relação de hierarquia entre duas palavras na sintaxe de Dependência, os termos mais usados em inglês são *head* e *dependent*. Em português, convencionou-se traduzir *dependent* por "dependente", mas, com relação a "*head*", os trabalhos de PLN usam nomenclaturas diferentes, tais como "cabeça", "núcleo", "dominante" ou o próprio termo em inglês *head*[5]. Também é comum em PLN chamar o núcleo de "pai" e o dependente de "filho", já que essa nomenclatura permite extrapolar as relações de parentesco e chamar também os nós de "avô" (quando se refere ao núcleo do núcleo), "neto" (quando se refere ao dependente do dependente) e de "irmão" (quando os dependentes possuem o mesmo núcleo)[6].

Em uma relação de dependência, o **núcleo** é o que rege, que comanda o seu **dependente**. Por exemplo, em um sintagma nominal como "a menina", o substantivo é o núcleo enquanto o artigo é o dependente, pois é o substantivo que rege o artigo, impondo-lhe os traços morfológicos de gênero (neste caso, feminino) e de número (neste caso, singular).

Definir quem é o núcleo e quem é o dependente nem sempre é uma tarefa fácil. Apesar de existirem convenções de diferentes teorias que definem essa hierarquia, nem sempre essas convenções são consensuais e, às vezes, essas definições são estabelecidas de forma arbitrária ou ad hoc. Por exemplo, preposições são consideradas núcleo em algumas abordagens de dependências (Osborne; Gerdes, 2019) e dependentes em outras (Universal Dependencies).

Por fim, vale ressaltar que as palavras podem ser núcleos ou dependentes, a depender da relação que elas estabelecem umas com as outras. Na frase "A menina postou uma foto.", por exemplo, o substantivo "menina", será núcleo na relação com seu determinante "a", mas ao mesmo tempo será dependente na relação com o verbo "postou". Algumas palavras, principalmente as de classe fechada, são consideradas sempre como dependentes, enquanto outras podem ser ora núcleo, ora dependente, considerando-se a outra palavra com a qual se relacionam.

[5]Entendemos que todos esses termos são sinônimos, mas adotaremos neste capítulo o termo "núcleo".
[6]Vale ressaltar que estes termos também são usados em inglês: "*parent*", "*child*", "*grandparent*", "*grandchild*" e "*sibling*".

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

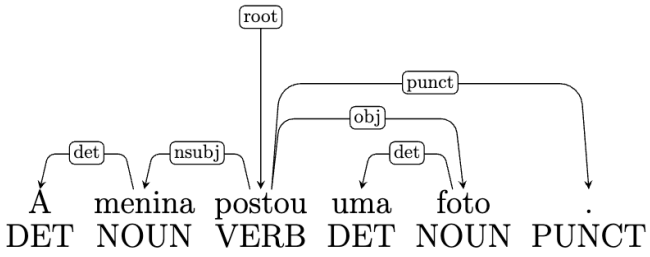
©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

6.6.2 A representação da sintaxe de dependência

Assim como na sintaxe de constituência, as relações de dependência podem ser representadas de formas distintas, sendo que todas explicitam a hierarquia e a direcionalidade da relação entre duas palavras em uma sentença. Das representações vistas neste capítulo, as duas mais utilizadas são os diagramas com setas e com representação parentética.

Como já foi dito, na representação com seta, as palavras são ligadas umas às outras por setas unidirecionais, ou seja, em um único sentido. Cada seta tem um ponto de partida e um ponto de chegada, o que significa que as relações não são recíprocas e nem uma mera concatenação entre palavras; pelo contrário, essas setas explicitam a dependência de uma palavra em relação à outra. Vale esclarecer que a representação com setas, assim como a representação de dependência com diagrama arbóreo, são ambas chamadas de "árvore de dependência". A Figura 6.11 mostra um exemplo de representação das relações de dependência com setas.

Figura 6.11: Exemplo de árvore de dependência para "A menina postou uma foto"



Como vemos na Figura 6.11, a cada palavra é atribuída uma etiqueta morfossintática, indicada em caixa alta na parte de baixo da Figura. A etiqueta que aparece acima dos arcos das setas representa o nome da relação que se estabelece entre a palavra núcleo e a palavra dependente. Essa etiqueta é escrita em letras minúsculas. As etiquetas são selecionadas dentro de um conjunto fechado que varia segundo o padrão de anotação. Na Figura 6.11, o conjunto de etiquetas é o proposto pelo projeto Universal Dependencies, sobre o qual falaremos na Seção 6.6.3.

Como vemos na Figura 6.11, toda sentença possui uma única raiz (ou `root`), que é a palavra que não depende de nenhuma outra. Em sentenças que possuem verbos, a raiz geralmente é o próprio verbo.

No exemplo da Figura 6.11, verificamos três macrorrelações entre: i) o verbo ("postou") e o núcleo do sujeito ("menina"); ii) o verbo ("postou") e o núcleo do objeto direto ("foto"); e iii) o verbo e o ponto final. Em todas elas, o verbo é o núcleo e as demais palavras são seus dependentes, conforme aponta a direção das setas.

Há também duas microrrelações entre: i) o substantivo ("menina") e o artigo ("a"); e ii) o substantivo ("foto") e o artigo ("uma").

Em uma árvore de dependência, algumas classes de palavras podem ser núcleo; outras, não. Mas cada palavra é dependente de uma outra. Isso significa que, na representação por diagrama com setas, em toda palavra deve chegar alguma (e exclusivamente uma) seta. A única exceção é para a raiz, na qual também chega uma seta, mas é a seta da relação `root`, o que significa que esta é a única palavra que não tem núcleo/pai. Todas as demais palavras, além da raiz, possuem um núcleo, de onde a seta parte.

Conforme sinalizado na Seção 6.6.1, o ponto de partida da seta representa o núcleo, e

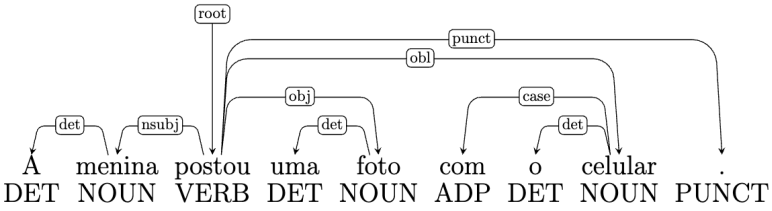
https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

o ponto de chegada representa o dependente. Algumas relações de dependência permitem qualquer direção de seta, ou seja, pode ir de uma palavra à esquerda para uma palavra à direita ou vice-versa (e.g. relações de `nsubj`, `punct`, `obj` etc.). Outras relações possuem direção obrigatória, como é o caso de `det` ou de `case`, que é sempre da direita para a esquerda em línguas românicas como o português, uma vez que artigos e preposições são antepostos aos substantivos.

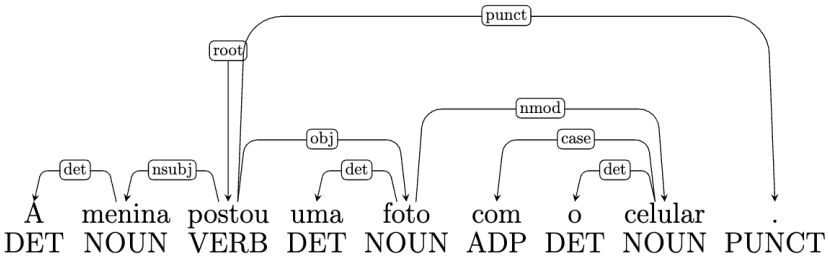
Na Figura 6.12, temos a representação da sentença "A menina postou uma foto com o celular", aqui anotada em sintaxe de dependência. Nela, "com o celular" é interpretado como Instrumento e como uma função dependente do verbo "postou".

Figura 6.12: Exemplo de uma árvore de dependência para "A menina postou uma foto com o celular"



Já na Figura 6.13, "com o celular" é interpretado como contido no sintagma nominal e com uma função dependente do substantivo "foto".

Figura 6.13: Exemplo de uma outra árvore de dependência para "A menina postou uma foto com o celular"



Na sintaxe de dependência, existe uma orientação geral para se evitar cruzamento dos arcos das setas. Na maior parte das vezes, isso é possível e recomendado. Porém, há casos pontuais em que o cruzamento de arcos é necessário para estabelecer a relação correta entre duas palavras, como ocorre no Exemplo 6.8 da Figura 6.14.

Exemplo 6.8:

> A menina postou uma foto no Instagram com filtro.

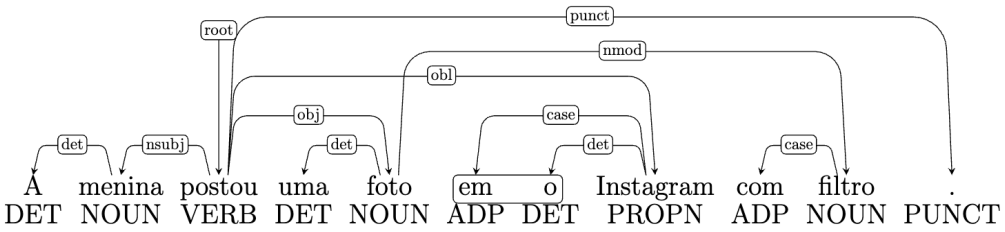
Neste exemplo, foi necessário cruzar os arcos de duas relações porque existe uma relação entre "postou no Instagram" e outra relação entre "foto com filtro". Para não haver cruzamento de arcos, a sentença teria de ter a seguinte ordem: "A menina postou uma foto com filtro no Instagram." ou ainda "A menina postou no Instagram uma foto com filtro."

O segundo tipo de representação mais comum em sintaxe de dependência é aquela feita por meio de parênteses. Na representação parentética, coloca-se o nome da relação de

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Figura 6.14: Exemplo de árvore de dependência com cruzamento de arcos



dependência, seguida de parênteses; dentro dos parênteses, coloca-se o núcleo, seguido do dependente e separados por uma vírgula.

Na Figura 6.15, temos a representação parentética da sentença "A menina postou uma foto".

Figura 6.15: Representação parentética da sentença "A menina postou uma foto".

```
1    root(ROOT, postou)
2    nsubj(postou, menina)
3    det(menina, a)
4    obj(postou, foto)
5    det(foto, uma)
6    punct(postou, .)
```

A representação na Figura 6.15 também mostra a hierarquia entre as duas palavras e a direcionalidade da relação, por meio da ordem das palavras dentro dos parênteses.

A raiz, assim como acontece na representação por diagrama arbóreo, também é representada em uma das linhas, como se fosse "dependente" de um núcleo fictício, chamado **ROOT**, representado por letras maiúsculas, em uma relação chamada `root` representada por letras minúsculas. O elemento **ROOT** representa o elemento vazio.

A vantagem da representação parentética é que ela é mais fácil de ser implementada, já que está no formato de texto e não depende de softwares de anotação. Por outro lado, a desvantagem é que a visualização da estrutura sintática da sentença, por um humano, é mais difícil do que uma representação por árvore. Na Figura 6.16 temos a representação parentética da sentença "A menina postou uma foto no Instagram com filtro", na qual há cruzamento de arcos. O cruzamento das relações, claramente visível na Figura 6.14, é mais difícil de ser identificado na representação parentética.

Destacamos, também, que, na representação parentética, a ordem das linhas não precisa seguir a ordem das palavras, podendo-se começar na primeira linha com a relação de `root` ou de `punct` ou qualquer outra macro ou microrrelação, ou colocar qualquer uma delas na sequência. O importante é que todas as relações macro e micro estejam explícitas. Nesse sentido, verificar se a anotação de todas as relações está completa é mais difícil de ser feito na representação parentética, ao passo que, na representação por diagrama arbóreo,

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Figura 6.16: Representação parentética de sentença com cruzamento de arcos.

```
1    root(ROOT, postou)    # O elemento ROOT representa o elemento vazio
2    nsubj(postou, menina)
3    det(menina, a)
4    obj(postou, foto)
5    det(foto, uma)
6    obl(postou, Instagram)
7    case(Instagram, no)
8    obl(foto, filtro)
9    case(filtro, com)
10   punct(postou, .)
```

é mais fácil visualizar caso haja palavras nas quais não esteja chegando nenhuma seta.

A sintaxe de dependência foi ganhando espaço no PLN nas últimas duas décadas e é hoje o tipo de *parsing* sintático mais utilizado, sobretudo em tarefas de extração de informação (Capítulo Extração de Informação). A seguir apresentamos um dos projetos de anotação multilíngue de sintaxe de dependência com reconhecido impacto nacional e internacional. Trata-se do projeto Universal Dependencies, cuja proposta visa maior consistência na anotação de *corpora* nas distintas línguas com base num arcabouço comum que possibilite a comparabilidade entre línguas. Apresentamos, também, um breve histórico de iniciativas de anotação que antecederam a proposta das Universal Dependencies.

6.6.3 Projetos de anotação multilingue: Universal Dependencies

Na primeira década de 2000, surgiram várias iniciativas para tentar representar as relações sintáticas entre as palavras de uma frase por meio de dependências. Uma delas é a chamada **Stanford Dependencies** (SD), que passou a fazer parte de um dos *parsers* mais utilizados, o *parser* Stanford, primeiramente para o inglês e posteriormente, para várias outras línguas[7]. Como essa iniciativa despertou interesse na comunidade linguística e de PLN, em 2006 e 2007 foram propostas *shared tasks* no CoNLL-X, nas quais os participantes treinaram e testaram sistemas nos mesmos conjuntos de dados usando anotação de dependências. Em 2007 especificamente, os participantes usaram uma abordagem multilíngue, baseada em *treebanks* de 10 línguas, e adaptada a um domínio. Nivre et al. (2007) descrevem essa tarefa, as diferentes abordagens usadas pelos participantes e os resultados dos experimentos.

Seguindo essa ideia de universalizar as representações de dependência para várias línguas, **Google** também criou seu próprio conjunto de etiquetas, baseando-se na análise de erros que foi feita para aquela *shared task* do CoNLL-X.

[7] O *parser* Stanford também dispõe de um conversor que transforma uma árvore de constituência em uma árvore de dependência, o qual funciona para várias línguas, incluindo o português. Mais informações sobre SD e Stanford parser podem ser consultadas em https://nlp.stanford.edu/software/lex-parser.shtml.

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

Outra iniciativa relevante foi o **Interset interlíngua** (Zeman, 2008), que criou uma ferramenta para a conversão de etiquetas morfossintáticas entre línguas. Essa abordagem parte do princípio de que algumas línguas são semelhantes entre si, mas nem todas elas possuem recursos suficientes para PLN. A iniciativa propõe converter a anotação morfossintática de uma língua com mais recursos para outra língua que possua menos recursos (Zeman; Resnik, 2008).

Todas essas iniciativas contribuíram para o que foi posteriormente chamado de projeto **Universal Dependency Treebank** (UDT) (McDonald et al., 2013), cuja ideia principal é universalizar as anotações das relações sintáticas e os conjuntos de etiquetas, a fim de propiciar a comparabilidade entre as línguas. Em 2013, foi lançada uma primeira versão envolvendo 6 línguas e depois, em 2014, outra versão envolvendo 11 línguas. A língua portuguesa, com *corpora* de português europeu e brasileiro, está representada nos *treebanks* com anotação em UD disponíveis (Rademaker et al., 2017). Atualmente, os *treebanks* anotados de acordo com o padrão UD abrangem mais de 100 línguas de várias famílias e troncos linguísticos diferentes.

Assim, o projeto das Dependências Universais (em inglês, *Universal Dependencies*), mais conhecido como as UD, é o resultado da combinação de todas essas iniciativas em uma abordagem única e articulada, baseada nas dependências universais de Stanford, uma versão ampliada do conjunto de etiquetas universais propostas pelo Google, um subconjunto revisado do conjunto de *features* do Interset, e uma versão revisada do formato CoNLL-X (chamado CoNLL-U)[8]. O projeto é a iniciativa mais usada e difundida hoje, motivo que justifica sua descrição em detalhes neste capítulo.

UD é um projeto que visa à anotação sintática sistemática e consistente em diversas línguas, de forma que essas línguas possam ser comparadas em relação à estrutura de suas sentenças. Apesar de buscar a maior padronização possível das anotações para garantir a comparabilidade entre as línguas, essa representação também prevê anotações particulares para línguas específicas, quando necessário. Justamente pelo fato de que algumas línguas possuem especificidades que não podem ser generalizadas para outras, o projeto UD precisa garantir que, apesar disso, as anotações sejam consistentes para que as línguas possam ser comparadas. Portanto, o projeto deve seguir seis premissas básicas[9] que reproduzimos, em nossa tradução, a seguir:

1. A proposta das UD precisa possibilitar uma análise linguística satisfatória para todas as línguas.
2. A proposta das UD precisa ser apropriada para a tipologia linguística, ou seja, fornecer uma base adequada para evidenciar o paralelismo linguístico entre línguas e famílias linguísticas.
3. A proposta das UD precisa ser adequada para uma anotação rápida e consistente pelo anotador humano.
4. A proposta das UD precisa ser compreendida e utilizada com facilidade por uma pessoa sem formação como linguista, seja um aprendiz de línguas ou um engenheiro com demandas básicas de processamento de língua.
5. A proposta das UD precisa ser adequada para o *parsing* de alto desempenho.

[8]Tradução nossa do original: "*The new Universal Dependencies is the result of merging all these initiatives into a single coherent framework, based on universal Stanford dependencies, an extended version of the Google universal tagset, a revised subset of the Intersect feature inventory, and a revised version of the CoNLL-X format (called CoNLL-U)*".

[9]As premissas podem ser consultadas em https://universaldependencies.org/introduction.html.

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

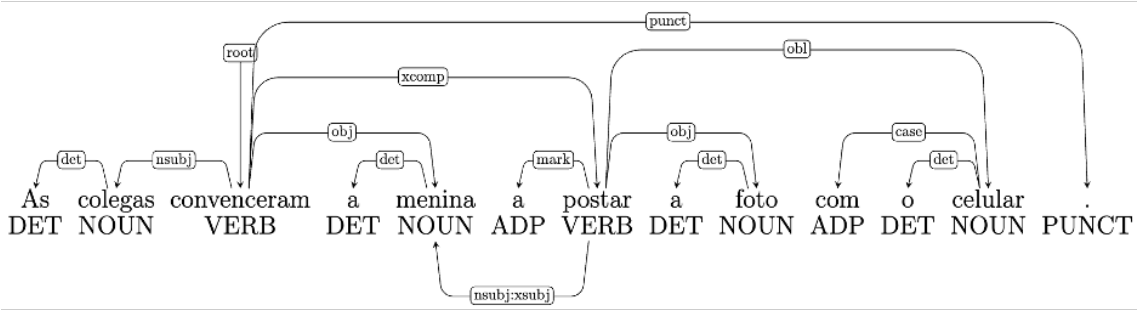
6. A proposta das UD precisa apoiar bem as tarefas de Compreensão de Linguagem Natural (Capítulo 1) subsequentes ao *parsing* sintático (extração de relações, compreensão de textos, tradução automática, dentre outras).

Além da representação básica por dependência já explicitada ao longo da Seção 6.6, que é obrigatória para todos os *treebanks*, a UD também prevê um segundo nível de representação de dependência chamada de *enhanced*, no sentido de obter uma representação mais completa e mais enriquecida. Esse segundo nível é uma camada extra de anotação das relações, realizada a fim de dar uma base mais completa para a interpretação semântica. Essa representação *enhanced* não é uma árvore tal como a árvore de dependências básicas, mas uma estrutura que agrega informação à árvore básica, conforme se apresenta para o Exemplo 6.9 na Figura 6.17.

Exemplo 6.9:

> As colegas convenceram a menina a postar a foto com o celular.

Figura 6.17: Anotação de sintaxe de dependência incluindo relações básicas e *enhanced*



A Figura 6.17 mostra, na parte de cima da sentença, as relações de dependência básicas, e na parte de baixo, as relações *enhanced*. No Exemplo 6.9, temos dois verbos, cada um dos quais tem seu sujeito. Nas relações básicas, indicamos, por meio da relação `nsubj`, o sujeito ("as colegas") do verbo principal ("convenceram"). Nas relações *enhanced*, podemos indicar, por meio da relação `nsubj:xsubj`, o sujeito ("a menina") do segundo verbo ("postar"). Ambos os sujeitos são relevantes para a interpretação de papéis temáticos e, consequentemente, para tarefas em PLN como a extração de informação.

6.6.3.1 *Tagsets*

Conforme introduzido na Seção 6.5, os chamados *tagsets* são conjuntos de etiquetas usadas para a anotação de categorias. Para garantir a comparabilidade linguística, os *treebanks* que aderem às Dependências Universais usam conjuntos de etiquetas, com as mesmas definições dadas para todas elas.

Os *tagsets* usados na UD contemplam a anotação: i) **morfológica**; e ii) **sintática**.

A anotação **morfológica** abrange três informações linguísticas: i) lema da palavra; ii) classe de palavra (em inglês, *part-of-speech* ou PoS) ; e iii) traços ou características morfológicas (em inglês, *features)*.

Utilizamos o termo **lema** para referir-nos à forma dicionarizada da palavra. Em nossa sentença, "A menina postou uma foto com o celular", o verbo "postou" tem como lema "postar", enquanto o substantivo "menina" tem o lema "menino".

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

A **classe de palavra**, como já vimos, é a categoria gramatical na qual essa palavra é classificada. Por exemplo, "postou" pertence à classe dos verbos, "menina" e "celular" pertencem à classe dos substantivos, "uma" e "o" à classe dos determinantes (ou artigos) e "com" à classe das preposições.

O *tagset* de PoS das UDs compreende 17 etiquetas, que são: **ADJ** (adjetivo), **ADP** (adposição, incluindo preposição), **ADV** (advérbio), **AUX** (verbo auxiliar), **CCONJ** (conjunção coordenativa), **DET** (determinante), **INTJ** (interjeição), **NOUN** (substantivo), **NUM** (numeral), **PART** (partícula), **PRON** (pronome), **PROPN** (nome próprio), **PUNCT** (sinal de pontuação), **SCONJ** (conjunção subordinativa), **SYM** (símbolo), **VERB** (verbo) e **X** (outros). Para uma explicação detalhada do *tagset* sugere-se consultar a Seção 4.3.3.

Já para as características morfológicas ou *features*, as UDs contemplam 24 *features*, que são divididas em lexicais e flexionais, e mais de 200 valores para essas *features*. Por exemplo, para a classe substantivo, temos em português as *features* gênero e número. Para gênero, podemos selecionar entre os valores: feminino, masculino e neutro. Para a classe verbo, temos as *features* pessoa, número, tempo verbal, modo, cada uma delas com valores específicos. Assim, o verbo "postou" está conjugado na 3a pessoa do singular do tempo pretérito perfeito do modo indicativo, e cada uma dessas informações corresponde a um valor dentro de uma *feature* morfológica.

O segundo tipo de anotação que a UD prevê é a **sintática**, por meio da qual são determinadas as relações de dependência entre as palavras de uma sentença. Hoje em dia existem 37 relações de dependência básicas. O Quadro 6.4 mostra as relações de dependência organizadas de acordo com os fundamentos das UDs.

Quadro 6.4: Relações de dependência da UD[10]

	Nominals	**Clauses**	**Modifier words**	**Function words**
Core arguments	nsubj obj iobj	csubj ccom xcomp		
Non-core dependents	obl vocative expl dislocated	advcl	admod discourse	aux cop mark
Nominal dependents	nmod appos nummod	acl	amod	det clf case
Coordination	**MWE**	**Loose**	**Special**	**Other**
conj cc	fixed flat compound	list parataxis	orphan goeswith rearandum	punct root dep

O Quadro 6.4 possui duas partes: na parte superior, estão as principais relações de dependência, distribuídas por tipo (se são relações nominais, verbais, com modificadores

[10]Referência: https://universaldependencies.org/u/dep/all.html.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

ou com palavras funcionais) e ainda classificadas por tipo de argumento. Já a parte inferior do quadro apresenta relações adicionais.

Na parte superior do Quadro 6.4, as relações estão agrupadas segundo duas dimensões:

1. no sentido horizontal, por linha, de acordo com as relações estabelecidas entre uma palavra e seu regente (*head*), e
2. no sentido vertical, por coluna, de acordo com as categorias dos dependentes.

Vemos, assim, no Quadro 6.4, que as relações que uma palavra estabelece com a palavra regente podem ser aquelas equivalentes a funções básicas da predicação em orações (*core arguments*), tais como sujeito da predição (`nsubj`, `csubj`) ou objetos (`obj`, `iobj`) ou complementos (`ccomp`, `xcomp`).

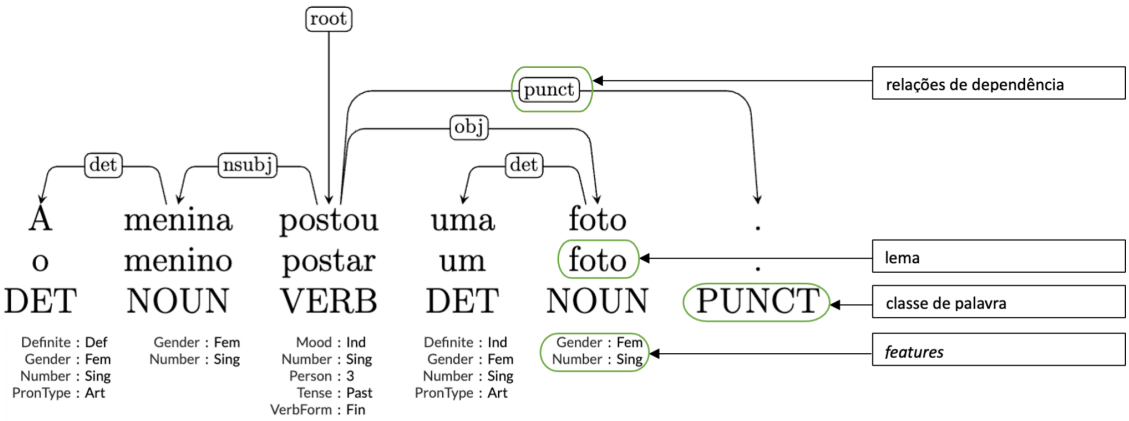
Também podem ser relações que correspondem a funções adicionais àquelas básicas (*non-core dependents of clausal predicates*), tais como adjuntos (obl), orações adverbiais (advcl), vocativos (vocative) etc. Essas relações se aplicam tanto no nível da oração como do sintagma, como é o caso de sintagmas verbais constituídos por um verbo e verbos auxiliares.

Há ainda relações de dependência dentro do sintagma nominal, como é o caso de modificadores adjetivais (amod), modificadores adnominais (nmod), determinantes (`det`), caso (`case`) etc.

Na parte inferior do Quadro 6.4, temos relações que indicam coordenação de palavras, sintagmas e orações, expressões multipalavra (MWE) (Capítulo 5), relações especiais para distintos fenômenos do discurso, tais como listas, disfluências, elipse, erros ortográficos e sinais de pontuação.

A anotação completa de uma sentença de acordo com as *guidelines* das UDs pode ser observada na Figura 6.18.

Figura 6.18: Visualização de todas as informações anotadas para uma sentença no projeto UD



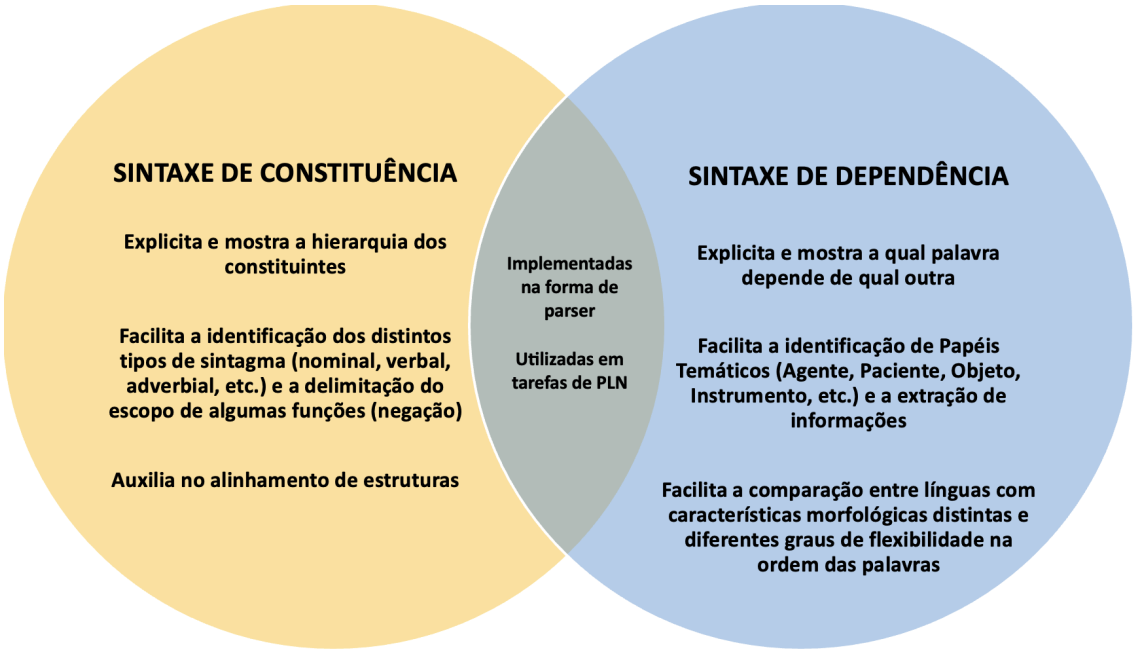
6.7 Qual é melhor: constituência ou dependência?

Uma vez apresentados os dois tipos de análise sintática mais utilizados em PLN, cabe perguntarmos quais as vantagens e desvantagens de se adotar cada um deles. A Figura 6.19 apresenta uma síntese dos principais pontos de cada tipo de análise sob a perspectiva de seu potencial em projetos de PLN.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Figura 6.19: Vantagens e desvantagens da sintaxe de constituência e de dependência



6.8 Fronteiras da sintaxe

Conforme apontado na Seção 6.1 deste capítulo, a sintaxe é um dos estratos centrais no sistema linguístico (Figura 6.1). Por estar no centro do sistema, seu estudo possui interseção com vários outros níveis, como a morfologia, a semântica, a pragmática e o discurso. Isso porque a separação em níveis ou estratos é uma forma didática de apresentar o objeto de estudo de cada área; porém, na língua em uso, esses níveis são interdependentes uns dos outros. Portanto, definir os limites entre um nível e outro é uma tarefa complexa. A seguir, exploraremos alguns conceitos e problemas linguísticos que estão na fronteira entre a sintaxe e um outro nível de análise linguística.

6.8.1 Sintaxe e Morfologia

Os traços morfológicos são gramaticalizados de diferentes formas em distintas línguas. Assim, qualquer noção que possa ser expressa morfologicamente em uma língua (por meio de traços morfológicos) pode ser expressa lexicalmente em outras línguas (por meio de perífrases ou paráfrases) (Bender, 2013).

Em português, convencionou-se que os afixos (prefixos, sufixos e infixos), as desinências (nominais e verbais) e outros componentes de palavras são unidades estudadas pela morfologia, enquanto a relação entre as palavras é estudada pela sintaxe.

Mesmo tendo-se estabelecido essa convenção, há aspectos linguísticos que estão na fronteira entre a morfologia e a sintaxe. Por exemplo:

A língua portuguesa constrói gênero gramatical (feminino ou masculino) nos substantivos por meio de uma desinência, por exemplo: "médic**o**" e "médic**a**". Há, porém, substantivos chamados "epicenos", que designam, com uma mesma forma, os dois gêneros (masculino e feminino), sendo feita a distinção de um gênero ou outro por meio da adição das palavras "macho" e "fêmea", como, por exemplo: "cobra **macho**" e "cobra **fêmea**". O

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

gênero dos substantivos em português é, portanto, estudado tanto pela morfologia (quando se utiliza uma desinência) quanto pela sintaxe (quando é necessário usar um sintagma contendo duas palavras).

Há também casos em que a mesma palavra (a mesma grafia) pode ser classificada como adjetivo ou particípio passado de um verbo (e.g. "solto", "feito", "motivada"). Isso depende da estrutura na qual a palavra opera. Assim, a distinção entre essas duas classes de palavras é de ordem sintática; contudo, tem impacto na morfologia, uma vez que, dependendo da classe de palavra atribuída, a palavra pode receber *features* de adjetivos ou *features* de verbos. Se em um texto clínico, encontramos a frase "Paciente lúcido e orientado", interpretamos "lúcido" e "orientado" como adjetivos que qualificam o substantivo "paciente" e constroem significados do estado de saúde do mesmo. Nesse caso, "orientado" é classificado como adjetivo e recebe as *features* de gênero e número. Já em "Paciente orientado a procurar fisioterapia", "orientado" é classificado como particípio passado do verbo "orientar", em uma oração passiva sem explicitação do auxiliar "ser" e sem realização de uma função que indica o Papel Temático de Agente da ação, o qual no contexto pode ser interpretado como o profissional da saúde que realizou o atendimento. Neste caso, "orientado" recebe *features* da classe verbo e também da classe adjetivo, uma vez que em português os particípios passados concordam em gênero e número com os substantivos que realizam o sujeito da oração: "O paciente foi orientado", "A paciente foi orientada".

O tempo futuro do presente, em português, pode ser construído por meio de uma desinência (por ex. "á" para a terceira pessoa do singular: "soltará", "fará", "motivará") e anotado com a *feature* morfológica de futuro. Também pode ser construído por uma perífrase verbal (e.g. "vai soltar", "vai fazer", "vai motivar"). Neste segundo caso, nenhum dos verbos recebe a *feature* morfológica de futuro, já que o significado de futuro se dá pela combinação dos dois verbos, e não por uma desinência. As formas construídas por meio de desinência são chamadas de futuro sintético e constroem a noção de futuro com recursos da morfologia. Já as formas construídas por meio de perífrase são denominadas futuro analítico e constroem a noção de futuro com recursos da sintaxe.

Formas negativas de adjetivos podem ser criadas em português por meio de morfemas (e.g. "capaz" x "**in**capaz" ou "normal" x "**a**normal" ou "moral" x "imoral") ou lexicalmente (i.e "governamental" x "**não** governamental"). O processo de formação de palavras por meio da adição de prefixos de negação **a-**, **i-**, **im-**, **in-** é estudado dentro da morfologia, ao passo que a construção da negação a partir da inserção de uma outra palavra é estudada no escopo da sintaxe.

As formas verbais que indicam imperativo em português possuem a mesma morfologia de subjuntivo. A definição entre um modo verbal e outro depende do contexto sintático em que estão inseridas. Por exemplo na sentença "Faça isso imediatamente", o verbo "faça" é classificado como modo imperativo. Já na sentença "Quero que você faça isso imediatamente", o verbo "faça" é classificado como subjuntivo. Para definir a *feature* morfológica de modo verbal a ser atribuída a formas como "faça" é preciso antes fazer a análise sintática da sentença.

6.8.2 Sintaxe e Semântica

A análise sintática pode ser complementada pela análise no nível da semântica. Essa complementação pode ser ilustrada com o seguinte exemplo.

Dada uma oração como "João deu um pulo", uma análise sintática de constituência nos permite identificar dois constituintes maiores: "João"e "deu um pulo", este último

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

passível de ser segmentado em constituintes menores contidos nele: "deu" e "um pulo". Numa análise de sintaxe de dependência, observamos as relações entre as palavras e identificamos "deu" como sendo a palavra raiz (`root`), "João" seu dependente em relação de sujeito (`nsubj`) e "pulo" em relação de objeto (`obj`). Assim, interpretamos que alguém ("João") realiza uma ação ("dar") que tem um objeto ("um pulo"). No entanto, todos nós interpretamos "dar um pulo" como expressão de uma ação equivalente a "pular". Como vimos neste capítulo, os constituintes que exercem a função de objeto geralmente correspondem a um papel temático na semântica. Em uma oração como "João deu um dinheiro", "dinheiro" é objeto do verbo "dar" e cumpre o papel temático de objeto estativo. Já em "João deu um pulo", o objeto "pulo" não possui status de papel temático; portanto, consideramos "deu um pulo" como uma única unidade do ponto de vista semântico. Assim, a análise sintática de dependência apresenta a mesma configuração para orações como "João deu um pulo" e "João deu um dinheiro". Em PLN, casos como este demandam uma análise semântica complementar à análise sintática, pois a identificação de papéis temáticos é relevante em tarefas como a extração de informação.

6.8.3 Sintaxe e Pragmática

A análise sintática pode, também, ser complementada pela análise sob a perspectiva da pragmática. Os exemplos a seguir ilustram essa complementação.

Em português, temos configurações de sentenças com uma palavra indicando negação, as quais constroem, no entanto, significados positivos. Numa sentença exclamativa como "Quantas vezes ela não ligou chorando!", temos o advérbio "não", que constrói um significado negativo. Contudo, interpretamos a exclamação como construindo um significado afirmativo: "ela ligou chorando muitas vezes". Em casos como este, de sentenças com elementos negativos que constroem um significado afirmativo, vemos que a perspectiva pragmática, isto é, a análise da linguagem em uso e dos pressupostos e inferências que fazemos como falantes, é relevante para complementar a análise sintática. Do contrário, interpretaremos a exclamação como a negação do evento: "ela não ligou chorando".

A perspectiva da pragmática também nos ajuda a interpretar sentenças nas quais não há elemento de negação, mas o significado construído é negativo. Por exemplo, em "Pessoa alguma será condenada", não temos nenhuma palavra que construa o significado de negação. No entanto, a oração é interpretada como "Nenhuma pessoa será condenada". Cabe destacar que, se trocarmos a ordem das palavras neste exemplo, teremos uma interpretação distinta, como é evidente ao contrastarmos "Pessoa alguma será condenada" e "Alguma pessoa será condenada".

6.8.4 Sintaxe e Discurso

O nível do Discurso e os Modelos Discursivos (Capítulo 11) serão apresentados mais à frente neste livro, mas convém explicitar neste momento uma confusão que se faz comumente entre os níveis sintático e discursivo.

A Gramática Tradicional (GT), nas seções referentes à sintaxe, estuda os conceitos de frase, oração e período, assim como a classificação das orações dentro de um período, por exemplo em oração principal, oração coordenada adversativa ou oração subordinada adverbial consecutiva, dentre várias outras.

Essas mesmas relações são estudadas em PLN no nível do Discurso como relações de contraste, de explicação, de equivalência, dentre várias outras. Isso porque as teorias

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

discursivas, tais como RST (*Rethorical Structure Theory*) e CST (*Cross-document Structure Theory*) não limitam sua unidade de análise linguística apenas dentro do texto, da sentença, do período ou da frase. São teorias que estudam o discurso independente do tamanho ou dos limites marcados por sinais de pontuação.

Apesar do rol de relações das teorias discursivas não representar uma equivalência exata com os tipos de orações da Gramática Tradicional, podemos identificar paralelos, por exemplo, entre uma relação de *Summary* (da RST) e uma oração coordenada conclusiva (da GT) ou entre uma relação discursiva de *Indirect Speech* (da CST) com uma oração subordinada substantiva objetiva direta).

6.9 Considerações finais

Neste capítulo, introduzimos os conceitos básicos de sintaxe, que é a área da linguística responsável por definir e descrever a ordem e a função das palavras e sintagmas na frase. Assim, foram discutidos conceitos linguísticos de **classe de palavra**, **constituinte**, **sintagma**, **frase**, **oração**, **período** (ou **sentença**), **escala de ordens**, **função sintática**, **papel temático**, assim como conceitos sintáticos voltados para o PLN, tais como *part-of-speech* (PoS), *parser*, *parsing*, *treebanks*, árvore de dependência, núcleo (*head*, *governor*, *parent*), dependente (*child*, *dependent*), entre outros.

Como o escopo deste livro é o processamento de língua natural, explicitamos as duas principais abordagens sintáticas usadas em PLN, que são: a **sintaxe de constituintes** e a **sintaxe de dependência**. Mas é importante ressaltar que existem várias correntes linguísticas teóricas que estudam e descrevem a sintaxe a partir de diferentes pontos de vista. Algumas dessas correntes são: paradigma formal ou estrutural, paradigma funcional ou funcionalista, o sistêmico-funcional, o gerativo ou gerativista, dentre outros.

Tanto a sintaxe de constituintes quanto a de dependência são abordagens amplamente usadas em PLN e, para isso, precisam de uma representação formal. Foram apresentados os tipos mais comuns de representação dessas duas abordagens, a saber: por meio de **colchetes**, de **árvores**, de **setas**, de **parênteses** e de **indentação**.

Ao final do capítulo, apresentamos alguns exemplos de questões e problemas linguísticos que podem ser estudados pelo viés da sintaxe ou de algum outro nível linguístico, já que essas fronteiras não são tão bem definidas na língua em uso. Como o foco do livro está na língua portuguesa, nos limitamos a mencionar brevemente alguns poucos exemplos. Para um estudo mais aprofundado sobre as fronteiras entre a sintaxe e os demais níveis linguísticos, ver Bender (2013).

O próximo capítulo (Capítulo 7) também é dedicado ao estudo da sintaxe dentro do PLN, porém com um olhar mais computacional voltado para os recursos e ferramentas disponíveis para fazer análise sintática automaticamente, os tipos de *parsing* e as abordagens sintáticas mais comuns em PLN.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Capítulo 7

Ferramentas e recursos para o processamento sintático

Elisa Terumi Rubel Schneider
Adriana S Pagano
Ana Clara S Pagano
Lucelene Lopes

Publicado em: 26/09/2023
Atualizado em: 20/11/2024

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

7.1 Introdução

A sintaxe é o nível de análise linguística no qual examinamos os padrões de estruturação de sentenças. Isto é, analisamos como as palavras se organizam em unidades que constroem significado dentro da sentença. Para isso, consideramos a classe de cada palavra, sua ordem na sentença e sua relação com as outras palavras. Conforme visto no Capítulo 6, em PLN, a análise computacional realizada no nível sintático é denominada *parsing*, a ferramenta que realiza essa tarefa é denominada *parser* e o recurso criado por meio da análise sintática é chamado *treebank*.

Neste Capítulo, vamos conhecer tipos de *parsing* sob a perspectiva computacional, juntamente com ferramentas e recursos disponíveis para o processamento do português brasileiro.

7.2 Tipos de *parsing*

A tarefa de *parsing* consiste em, dada uma entrada com uma sentença sem nenhuma anotação (*raw*), um modelo faz uma predição da estrutura sintática dessa sentença. Como vimos no Capítulo 6, o objetivo do processamento sintático é identificar as unidades (como palavras, sintagmas e orações) na sentença e estabelecer as relações gramaticais entre elas a fim de extrair algum tipo de informação. Essas relações podem ser analisadas em termos de:

1. **constituência**, ou seja, quais unidades são hierarquicamente inferiores às outras e estão nelas contidas; ou
2. **dependência**, isto é, quais palavras dependem de quais outras e qual é o tipo de relação entre elas, incluindo o papel de cada palavra na sentença (como sujeito, objeto, verbo, adjetivo etc.).

Assim, de acordo com o tipo de análise sintática adotada, há *parsers* de constituência e *parsers* de dependência.

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Mas há uma perspectiva adicional sob a qual podemos caracterizar tipos de *parsing* e *parsers*: trata-se do escopo ou profundidade com que a análise sintática é executada. Nesse sentido, podemos analisar as sentenças de forma exaustiva até obtermos uma análise completa de sua estrutura ou fazer uma análise mais rasa para obtermos uma análise com informações mínimas, mas relevantes para as tarefas em PLN.

O primeiro tipo é denominado *deep* (em português, *profundo)* ou *parsing* completo *e* o segundo tipo é denominado *shallow* (em português, *superficial)* ou *parsing* parcial. Contudo, cabe uma observação sobre esta terminologia. No uso geral, os termos *parsing* e *parser* acabaram sendo adotados para se referir ao *parsing* completo. Já o *parsing* parcial é conhecido como *chunking* (em português, *cortar*) e a ferramenta como *chunker,* embora *chunking* seja uma dentre várias abordagens para a implementação do *parsing* parcial (Jurafsky; Martin, 2023).

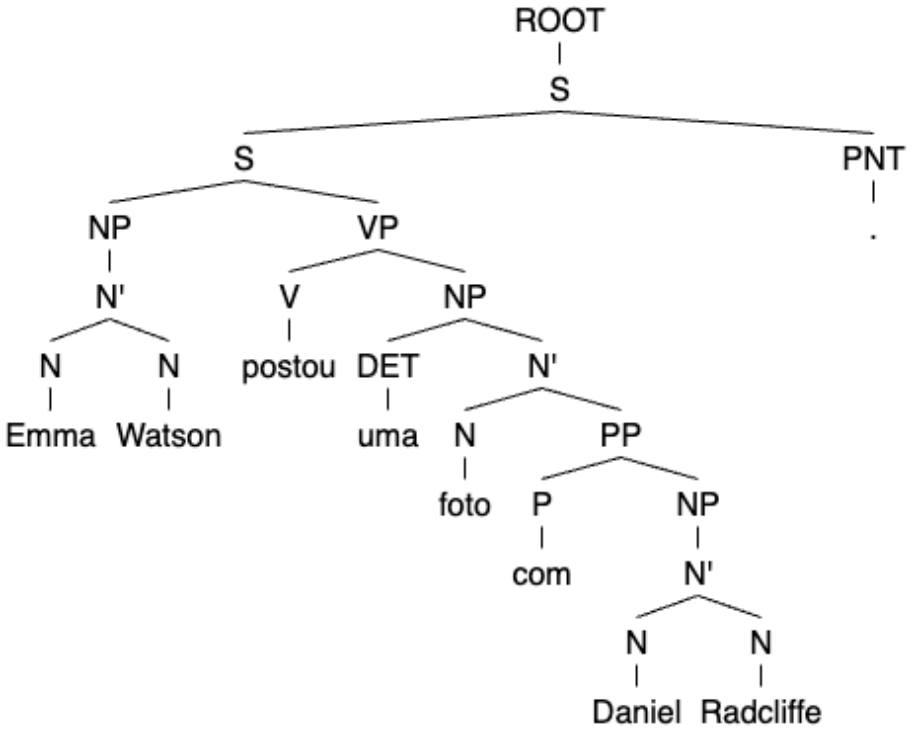
Tanto o *parsing* de constituência como o *parsing* de dependência podem ser executados de forma completa ou parcial. Tomando como exemplo o *parsing* de constituência, uma análise completa ou *deep* extrai todos os agrupamentos e as relações sintáticas em uma sentença. Por exemplo, dada a sentença:

Exemplo 7.1:

> Emma Watson postou uma foto com Daniel Radcliffe.

Temos na Figura 7.1 uma representação em diagrama de árvore que mostra a profundidade da análise.

Figura 7.1: Exemplo de saída de *parser*



Já uma análise parcial extrai constituintes delimitados, sem estabelecer a hierarquia entre eles ou de que forma uns estão contidos em outros.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Figura 7.2: Exemplo de saída de *parser* parcial

[NP Emma Watson] [VP postou] [NP uma foto] [PP com] [NP Daniel Radcliffe]

A Figura 7.2 ilustra a análise rasa, não hierárquica do *parsing* parcial para o Exemplo 7.1.

O objetivo do *parsing* parcial é gerar uma representação rasa da estrutura da sentença que possibilite um processamento mais rápido de grandes volumes de texto. É geralmente implementado por meio de tokenização de uma sentença em palavras, identificação da classe de palavra (PoS) e segmentação em pedaços ou *chunks*. O conceito de *chunk* foi proposto por Abney (1992) como uma unidade formada por uma única palavra ou por um conjunto de palavras. Em um *chunk*, há uma palavra de conteúdo circundada por palavras funcionais. A palavra de conteúdo mais explorada em *chunking* é o substantivo, dada a alta correlação de substantivos com entidades.

Assim, a tarefa de *chunking* da sentença do Exemplo 7.1, executada em Python utilizando o modelo de língua portuguesa `pt_core_news_sm`, da biblioteca spaCy, gera o resultado disposto na Figura 7.3:

Figura 7.3: Captura de tela de *notebook* em Python com código e resultado de *chunking* de uma sentença em português

```
!pip install spacy
!python -m spacy download pt_core_news_sm
import spacy
nlp = spacy.load("pt_core_news_sm")
sentence = "Emma Watson postou uma foto com Daniel Radcliffe."
doc = nlp(sentence)
for chunk in doc.noun_chunks:
    print(chunk.text)
```

```
Emma Watson
uma foto
Daniel Radcliffe
```

Como vemos na Figura 7.3, o *chunking* reconhece três unidades ou *chunks*, cada uma nucleada por um substantivo. Os três *chunks* são candidatos a entidades, sendo duas delas nomes próprios de pessoas. De fato, como veremos no Capítulo Extração de Informação, vários modelos de Extração de Informação utilizam análises rasas como a fornecida pelo *chunking*.

7.3 Recursos e ferramentas para o português

Nesta seção, vamos conhecer alguns recursos e ferramentas de PLN para análise sintática do português.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

7.3.1 *Corpora*

O primeiro recurso para o processamento linguístico é um *corpus* anotado ou *treebank*, isto é, textos enriquecidos com marcações de classe de palavras (*Part-of-Speech*) e relações sintáticas. Um exemplo de *corpus* em português anotado é o Bosque[1], amplamente utilizado para treinar modelos de análise sintática (Veja Capítulo 16).

O *corpus* Bosque é parte de um *corpus* maior, chamado Floresta Sintá(c)tica[2], que abrange, além do Bosque, outros *subcorpora*, nomeadamente: Selva, Amazônia e Floresta Virgem. O grande *corpus* foi anotado automaticamente pelo *parser* PALAVRAS (Bick, 2000). O Bosque está integrado por sentenças extraídas dos *corpora* CETENFolha (português brasileiro) e CETEMPúblico (português europeu), ambos constituídos por textos jornalísticos escritos. Uma versão do Bosque[3] foi convertida para o formato UD (Universal Dependencies), apresentado no Capítulo 6, e é hoje um dos *treebanks* mais utilizados pela comunidade de PLN no Brasil em modelos de *parsing* de dependência atuais.

Além da Floresta Sintá(c)tica, encontra-se disponível, como recurso para a língua portuguesa, o Corpus Internacional do Português – CINTIL[4], desenvolvido pela Universidade de Lisboa, que possui 1 milhão de *tokens* de texto jornalístico, com anotação de classe de palavra, lema e expressões multipalavra. Uma versão desse *corpus,* o CINTIL-UDep[5], é disponibilizada com anotações no padrão UD.

Mais recentemente, o *corpus* PetroGold[6] foi disponibilizado e, hoje, é um *corpus* passível de ser utilizado em modelos de *parsing* de dependência. PetroGold é um *corpus* de textos acadêmicos no domínio do petróleo, anotado no formato UD e revisado manualmente.

Há diversas iniciativas em andamento, no momento da escrita deste capítulo, para a criação de *corpora* anotados em português brasileiro. No escopo do projeto NLP2, desenvolvido pelo Centro de Inteligência Artificial[7] (C4A1) da Universidade de São Paulo, com o objetivo de desenvolver recursos, ferramentas e aplicações para levar o português ao estado da arte em PLN, o projeto POeTiSA[8] desenvolve o *treebank* Porttinari[9], um *corpus* multi-gênero de textos em português brasiliero anotados de acordo com o padrão UD. Inclui textos jornalísticos do *corpus* da Folha de São Paulo/Kaggle, o *corpus* MAC-MORPHO[10] de textos jornalísticos, o *corpus* DANTE[11] (*Dependency-ANalised corpora of TwEets*), integrado por tweets da Bolsa de Valores, B2W-reviews01[12], composto por resenhas e avaliações de consumidores da empresa de comércio eletrônico Americanas e um *corpus* de Resenhas online de livros. A versão Porttinari-base já se encontra disponível[13].

Uma iniciativa também em andamento é o *corpus* Veredas[14], desenvolvido na Faculdade de Letras da UFMG, que visa à construção de *treebanks* de textos anotados de acordo com o padrão das UD. Inclui amostras de uma variedade de textos em inglês, espanhol e portu-

[1] https://www.linguateca.pt/Floresta/corpus.html
[2] https://www.linguateca.pt/Floresta/principal.html
[3] https://universaldependencies.org/treebanks/pt_bosque/index.html
[4] http://cintil.ul.pt/
[5] https://universaldependencies.org/treebanks/pt_cintil/index.html
[6] https://universaldependencies.org/treebanks/pt_petrogold/index.html
[7] https://c4ai.inova.usp.br/
[8] https://sites.google.com/icmc.usp.br/poetisa/the-project
[9] https://sites.google.com/icmc.usp.br/poetisa/resources-and-tools
[10] Mac-Morpho: http://www.nilc.icmc.usp.br/macmorpho/
[11] Brazilian Stock Market Tweets with Emotions|Kaggle: https://www.kaggle.com/datasets/fernandojvdasilva/stock-tweets-ptbr-emotions
[12] https://opencor.gitlab.io/corpora/real19b2wreviews01/
[13] Corpus Porttinari: https://sites.google.com/icmc.usp.br/poetisa/porttinari
[14] http://www.letras.ufmg.br/veredas/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

guês brasileiro: colunas jornalísticas, fábulas, narrativas, receitas culinárias, questionários médicos e bulas de medicamento. Em parceria com a PUCPR, a Faculdade de Letras da UFMG desenvolveu o *treebank* DepClinBr, um *corpus* de narrativas clínicas anotadas de acordo com o padrão das UD (Oliveira et al., 2022).

7.3.2 *Parsers*

Parsers são ferramentas que podem auxiliar uma aplicação (por exemplo, tradução automática, sumarização de textos, extração de informação, *question-answering*) ou fazer parte de uma ferramenta maior ou conjunto de ferramentas (*toolkit*).

Algumas ferramentas computacionais estão disponíveis para realizar a análise sintática em português. A análise pode ser feita por meio de:

7.3.2.1 Programas e aplicativos

Existem diversos *parsers* desenvolvidos por distintos grupos de pesquisa, fornecidos através de um programa de computador ou aplicativo a ser instalado. No Brasil, podemos citar Curupira[15], Donatus[16] e PassPort[17]. Curupira é um analisador robusto de uso geral para o português brasileiro, fornecendo um conjunto das análises sintáticas possíveis para uma frase de entrada. A ferramenta analisa sentenças de cima para baixo, da esquerda para a direita, por meio de uma gramática funcional livre de contexto, restrita e relaxada, para o português brasileiro escrito padrão e um léxico extenso e de ampla cobertura. A Figura 7.4 apresenta a interface gráfica onde é possível ver a obtenção de toda informação da análise realizada pelas regras do *parser*[18].

Donatus é um projeto que consiste em ferramentas e gramáticas baseadas em Python e na biblioteca NLTK[19] para análise profunda e anotação sintática de *corpora* do português brasileiro. Inclui uma interface gráfica, conforme pode ser visto na Figura 7.5. Está disponível em repositório público, sob licença *GNU General Public License version 3.0 (GPLv3)*.

PassPort é uma ferramenta para análise de dependências de português treinado com o Stanford Parser, utilizando o *corpus* Portuguese Universal Dependency (PT-UD). Infelizmente, a página do projeto não se encontra disponível na data de escrita deste capítulo.

7.3.2.2 *Frameworks* e Bibliotecas

Devido à popularidade de Python, muitas bibliotecas de PLN foram desenvolvidas na linguagem. Entre as bibliotecas que incluem *parsing* para língua portuguesa, podemos citar spaCy[20], Stanza[21] e NLTK.

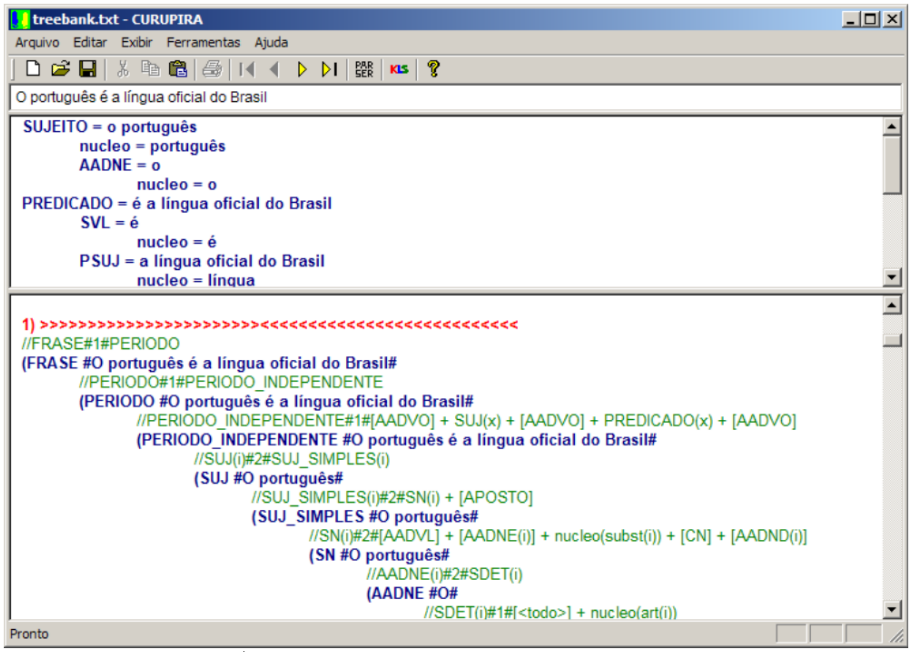
spaCy é uma biblioteca PLN que oferece análise linguística eficiente e rápida para várias línguas, incluindo o português. Inclui recursos para tokenização, marcação de parte

[15] http://www.nilc.icmc.usp.br/nilc/tools/curupira.html
[16] https://sourceforge.net/projects/donatus/
[17] PassPort (A Dependency Parsing Model for Portuguese | SpringerLink): https://link.springer.com/chapter/10.1007/978-3-319-99722-3_48
[18] No entanto, vale ressaltar que o projeto foi desenvolvido em 2002 - 2004, de acordo com o site, e pode não estar mais disponível para ser obtido.
[19] NLTK :: *Natural Language Toolkit* https://www.nltk.org
[20] https://spacy.io/
[21] https://stanfordnlp.github.io/stanza/index.html

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Figura 7.4: Captura de tela da interface gráfica da ferramenta Curupira



Fonte: (Martins et al., 2003)

do discurso (PoS *tagging*), reconhecimento de entidades nomeadas, análise sintática e outros. Através de modelos pré-treinados, o spaCy é capaz de fornecer análises detalhadas, permitindo a extração de informações semânticas de um texto em língua portuguesa.

Stanza[22] é outra biblioteca PLN que suporta vários idiomas, incluindo o português, desenvolvida pela Universidade Stanford. Fornece uma gama de recursos semelhantes ao spaCy, com suporte a análises mais profundas, como a análise de dependência neural.

NLTK (*Natural Language Toolkit*) também é uma biblioteca em Python que oferece suporte para tarefas de PLN em língua portuguesa, como a análise sintática. NLTK permite o *parse* usando expressões regulares (com *Regexp Parser*), análise de dependência com analisador de dependência probabilístico e análise de dependência com analisador de Stanford.

É importante mencionar que, além de Python, outras linguagens de programação também oferecem bibliotecas e *frameworks* para análise sintática.

7.3.2.3 Ferramentas online

A seguir apresentaremos algumas das ferramentas web disponíveis para *parsing* de texto em português, permitindo executar a análise sintática e obter como saída arquivos com distintos formatos, incluindo a visualização das árvores sintáticas. Todas as ferramentas apresentadas são de acesso livre e gratuito.

O Portparser

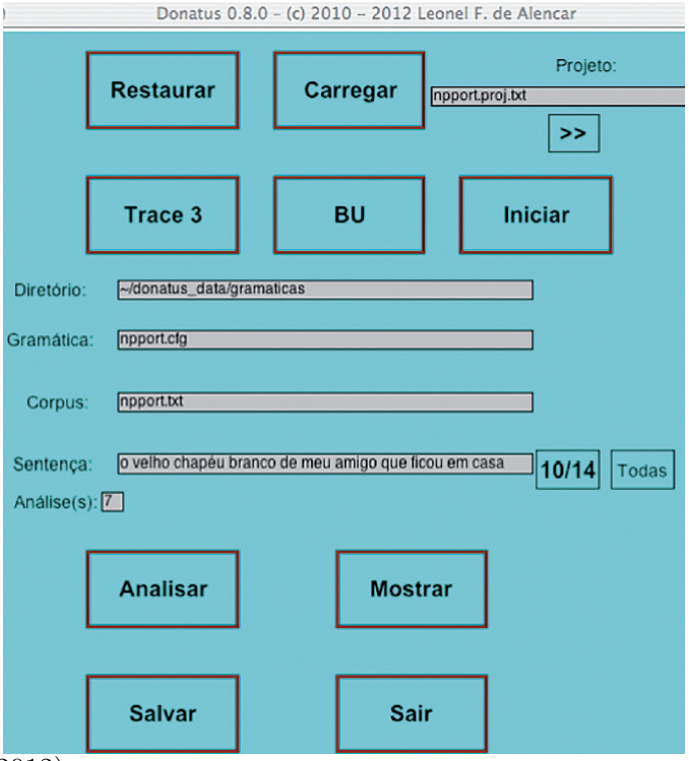
O Portparser (Lopes; Pardo, 2024) é um modelo de parsing para o português brasileiro que se alinha à iniciativa de Universal Dependencies (UD), tendo sido desenvolvido no

[22] Cabe esclarecer que embora o Stanford forneça um modelo treinado para língua portuguesa, que pode ser utilizado pelas bibliotecas Python, na interface online deste projeto não há suporte para o idioma português.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Figura 7.5: Captura de tela da interface gráfica da ferramenta Donatus



Fonte: (Alencar, 2012)

contexto do projeto POeTiSA (POrtuguese processing - Towards Syntactic Analysis and parsing) [23]. O modelo foi treinado com o subcorpus jornalístico Porttinari-base (Duran et al., 2023) do treebank Porttinari (Pardo et al., 2021). O Portparser faz anotação de Part-of-Speech (PoS), identificação de lemas, extração de características morfológicas e anotação de relações de dependência.

O desempenho para o português é de 99% de precisão para anotação de PoS, lema e atributos morfológicos, enquanto a precisão para a anotação de dependências supera 95% de LAS (Labeled Attachment Score), que é a métrica mais tradicional da tarefa de parsing (Nivre; Fang, 2017), sendo o estado da arte conhecido no momento para a língua portuguesa.

A ferramenta que permite o uso do Portparser está disponível online[24]. A Figura 7.6 apresenta a captura da tela de entrada do Portparser. O Portparser foi projetado para receber textos em português e executar segmentação sentencial, tokenização e as anotações do modelo citadas (PoS, lema, características morfológicas e relações de dependência).

O Portparser aceita entrada de sentenças isoladas (Single sentence) ou upload de arquivo com diversas sentenças (Multiple sentences). No caso de entrada de um arquivo com diversas sentenças, o Portparser oferece a possibilidade de executar a segmentação sentencial do texto ou considerar que o texto já foi segmentado e possui uma sentença por linha.

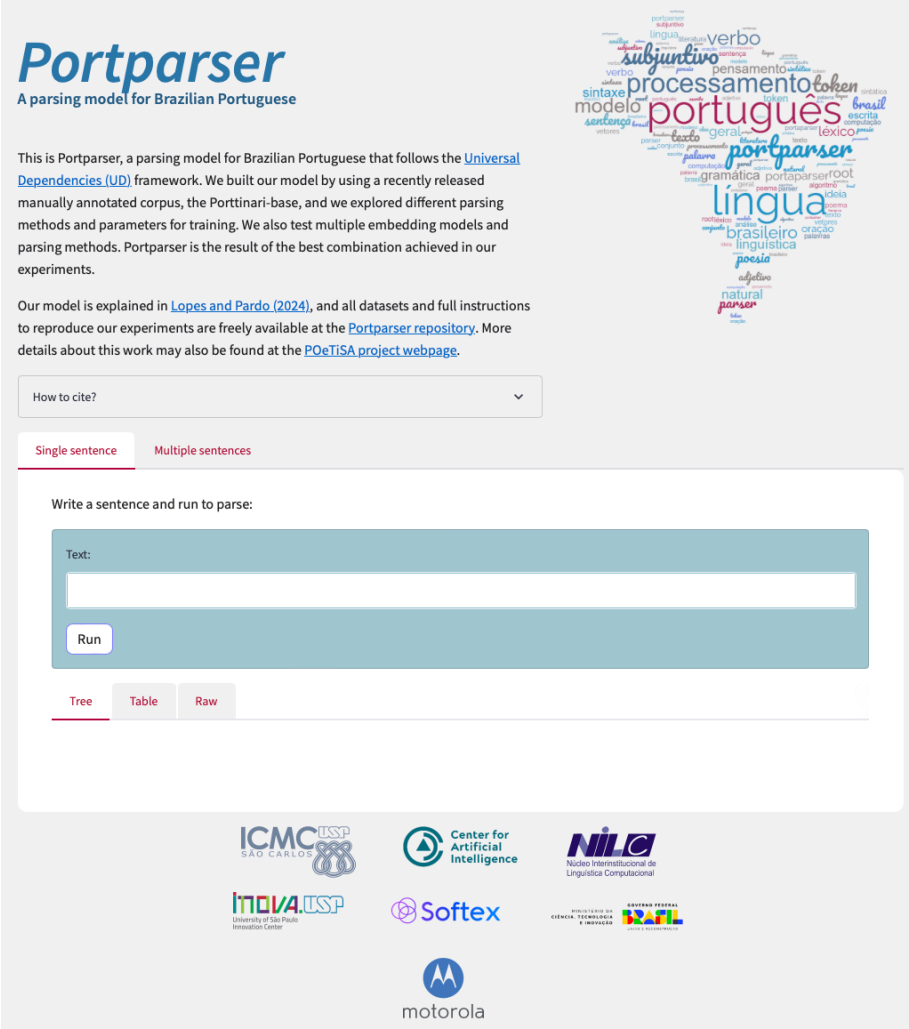
O resultado da utilização do Portparser pode ser visto na Figura 7.7, onde a sentença "Gostei da variedade e da qualidade" foi anotada. A anotação produzida pode ser visu-

[23] https://sites.google.com/icmc.usp.br/poetisa
[24] http://portparser.icmc.usp.br:8082/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

Figura 7.6: Captura da tela de entrada da ferramenta online Portparser



alizada através da árvore de dependência (Tree), como na Figura 7.7, mas também por uma tabela com o conteúdo do arquivo CoNLL-U (Table) na Figura 7.8 ou como um texto puro com o arquivo CoNLL-U (Raw) na Figura 7.9.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Figura 7.7: Execução do Portparser para a sentença "Gostei da variedade e da qualidade." visualizando a árvore de dependência gerada.

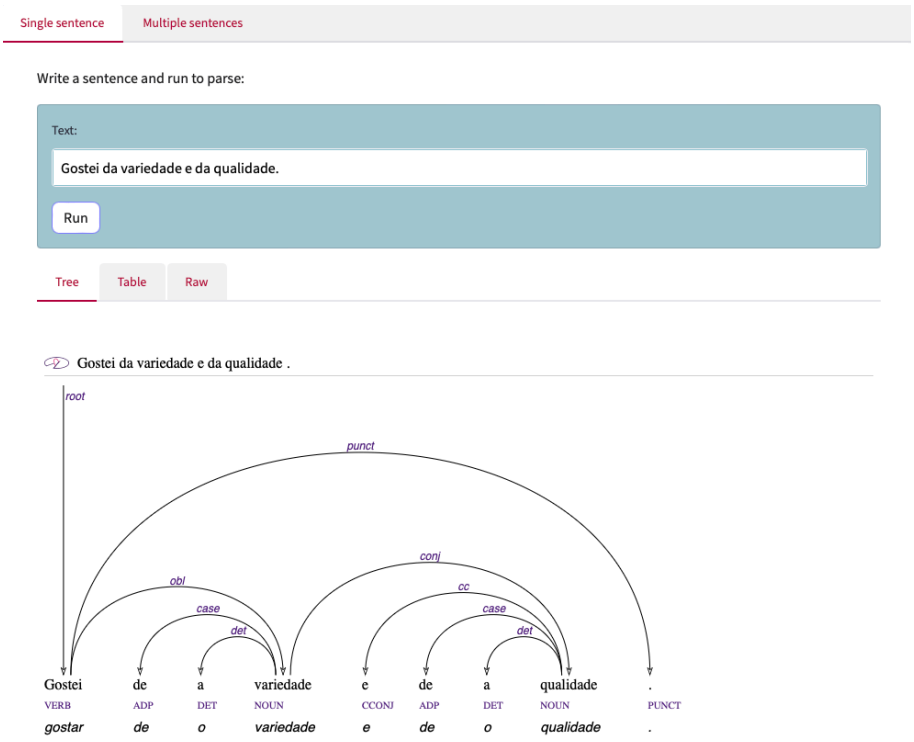
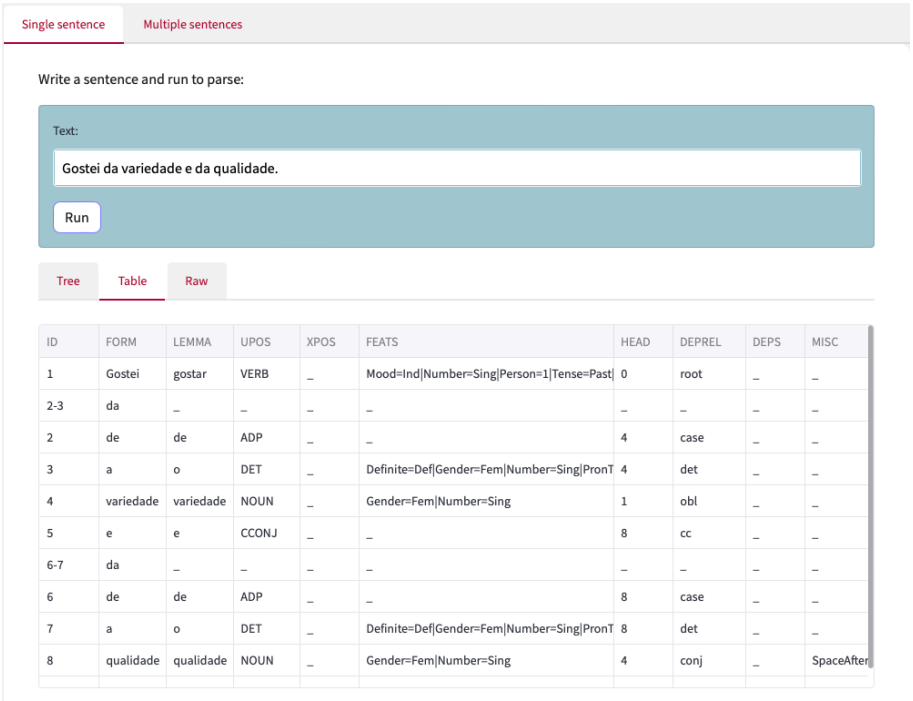


Figura 7.8: Execução do Portparser para a sentença "Gostei da variedade e da qualidade." visualizando os campos do formato CoNLL-U gerado em forma de tabela.

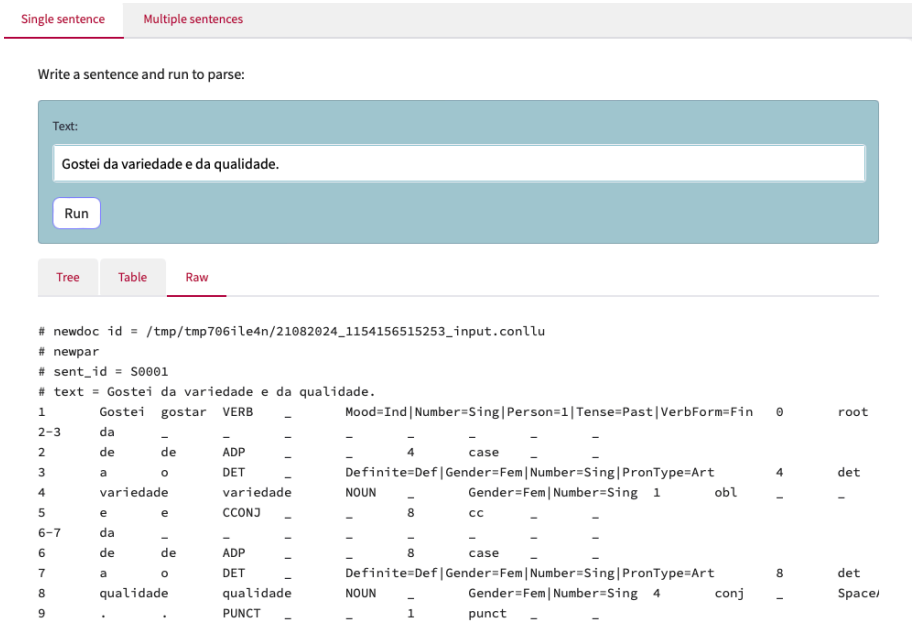


ID	FORM	LEMMA	UPOS	XPOS	FEATS	HEAD	DEPREL	DEPS	MISC
1	Gostei	gostar	VERB	_	Mood=Ind|Number=Sing|Person=1|Tense=Past|	0	root	_	_
2-3	da	_	_	_	_	_	_	_	_
2	de	de	ADP	_	_	4	case	_	_
3	a	o	DET	_	Definite=Def|Gender=Fem|Number=Sing|PronT	4	det	_	_
4	variedade	variedade	NOUN	_	Gender=Fem|Number=Sing	1	obl	_	_
5	e	e	CCONJ	_	_	8	cc	_	_
6-7	da	_	_	_	_	_	_	_	_
6	de	de	ADP	_	_	8	case	_	_
7	a	o	DET	_	Definite=Def|Gender=Fem|Number=Sing|PronT	8	det	_	_
8	qualidade	qualidade	NOUN	_	Gender=Fem|Number=Sing	4	conj	_	SpaceAfter

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Figura 7.9: Execução do Portparser para a sentença “Gostei da variedade e da qualidade.” visualizando os campos do formato CoNLL-U gerado em forma de texto puro.



https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

O Parser LX

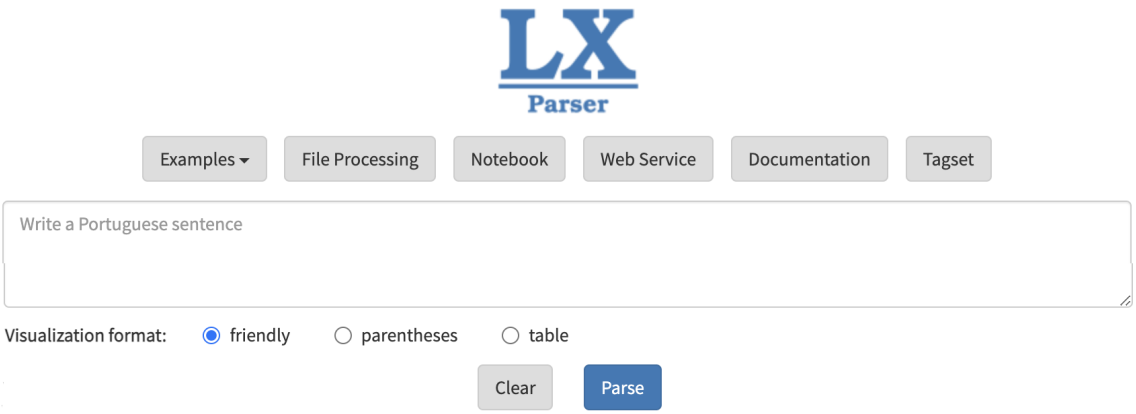
O Parser LX, ferramenta integrante do PORTULAN CLARIN[25], é parte integrante de um portal de acesso a infraestrutura de tecnologia linguística no escopo do projeto internacional CLARIN ERIC[26].

O Parser LX é disponibilizado tanto para *parsing* de constituência[27] como de dependência, este último em duas versões: LX-DepParser[28] e LX-UDParser[29]

A interface é simples e amigável, tendo como entrada uma sentença que o usuário pode digitar em campo próprio ou um arquivo que deverá ser importado.

A Figura 7.10 mostra uma captura de tela da interface do Parser LX de constituência.

Figura 7.10: Captura de tela mostrando a interface do *parser* de constituência LX.



Como vemos na Figura 7.10, o *parser* tem como saída uma visualização na forma de árvore, denominada "amigável" (em inglês, *friendly)*, ou uma representação parentética ou entre parênteses/colchetes ou ainda uma representação na forma tabular.

Para a sentença "Emma Watson postou uma foto com Daniel Radcliffe", o *parser* gera a árvore de constituência apresentada na Figura 7.11

No que diz respeito à sintaxe de dependência, LX possui duas versões, sendo uma delas adaptada ao formato Universal Dependencies, apresentado no Capítulo 6.

A Figura 7.12 mostra uma captura de tela da interface do Parser LX de dependência no formato UD.

Como a tela mostra, a saída do *parser* pode ser na forma de visualização com setas, denominada "amigável" ou no formato CoNLL-U, que é um formato próprio para *parsing* de dependência, como vimos no Capítulo 6.

Para a sentença "Emma Watson postou uma foto com Daniel Radcliffe", o *parser* gera a árvore de constituência apresentada na Figura 7.13.

[25] https://portulanclarin.net/
[26] https://www.clarin.eu/
[27] https://portulanclarin.net/workbench/lx-parser/
[28] https://portulanclarin.net/workbench/lx-depparser/
[29] https://portulanclarin.net/workbench/lx-udparser/

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Figura 7.11: Diagrama de árvore de constituência gerado pelo *parser* LX.

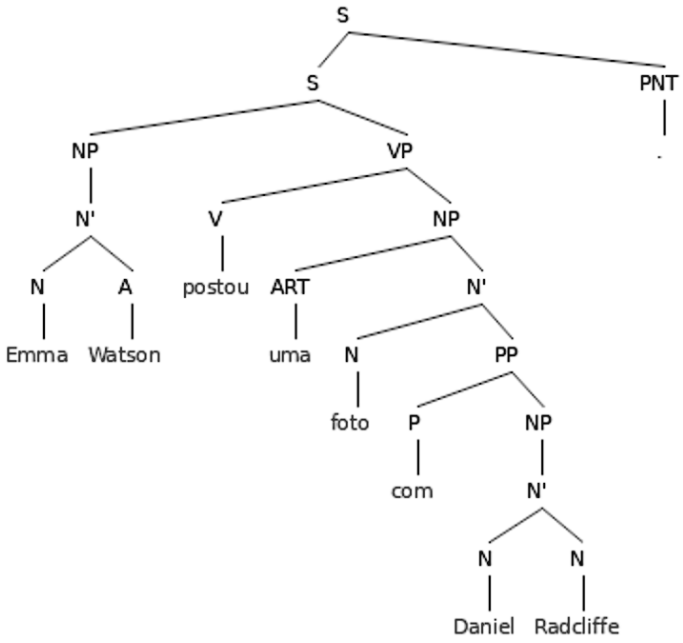
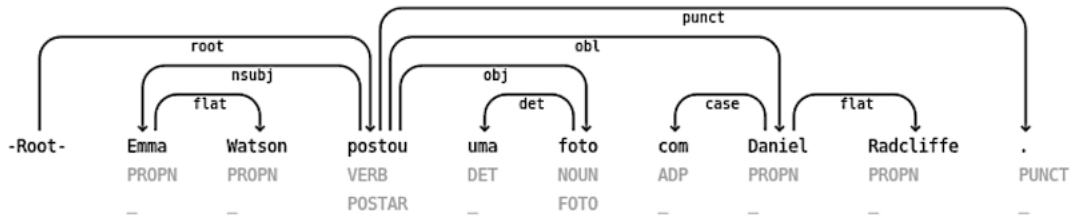


Figura 7.12: Captura de tela mostrando a interface do *parser* de dependência LX.



Figura 7.13: Diagrama com setas gerado pelo Parser LX.



©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

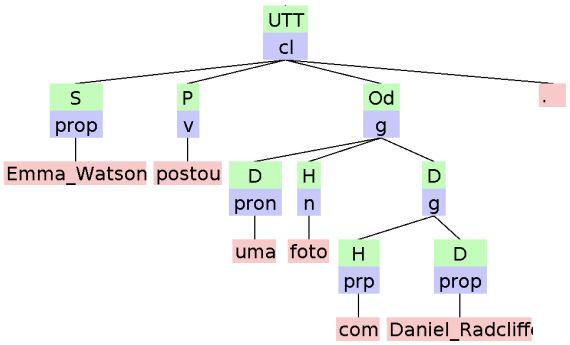
https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

O Parser VISL[30]

O projeto VISL (do inglês, *Visual Interactive Syntax Learning*) é desenvolvido pelo *Institute of Language and Communication* (ISK) da *University of Southern Denmark*. O projeto disponibiliza recursos (*corpora)* e ferramentas, tais como *parsers* de constituência e dependência. A análise se baseia no *parser* PALAVRAS (Bick, 2000) e no *corpus* Floresta Sintá(c)tica.

Para a sentença "Emma Watson postou uma foto com Daniel Radcliffe", o *parser* gera a árvore de constituência apresentada na Figura 7.14.

Figura 7.14: Diagrama de árvore de constituência gerado pelo *parser* VISL.



O UDPipe

UDPipe é um conjunto de ferramentas (*toolkit*) e um serviço web que possibilita processar texto por meio de uma *pipeline* que inclui tokenização em palavras, etiquetagem de classe de palavra (PoS), lematização, e *parsing* de dependência. É desenvolvido pelo *Institute of Formal and Applied Linguistics* da *Faculty of Mathematics and Physics da Charles University* (República Tcheca). O padrão adotado é das Universal Dependencies e conta com modelos treinados para a maioria dos *treebanks* já anotados no formato UD. Além da interface web, altamente eficiente e amigável, é possível processar texto por meio de um script em Python.

A Figura 7.15 mostra uma captura de tela da interface web do UDPipe, com a seleção do modelo Bosque para a língua portuguesa.

A entrada pode ser texto digitado em um campo próprio ou um arquivo de texto e há três formatos de saída: diagrama arbóreo, formato CoNLL-U e formato tabular.

Para a sentença "Emma Watson postou uma foto com Daniel Radcliffe", o *parser* gera a árvore de dependência apresentada na Figura 7.16.

[30] https://visl.sdu.dk/visl/about/

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Figura 7.15: Captura de tela mostrando a interface do UDPipe.

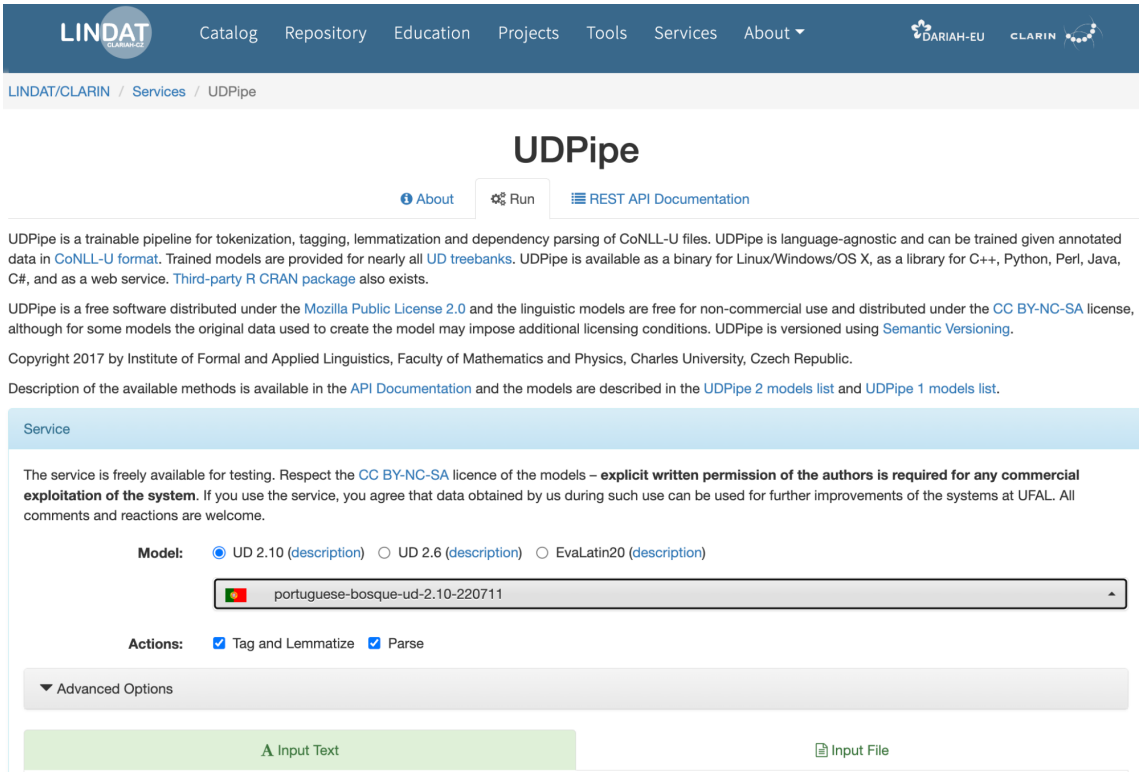
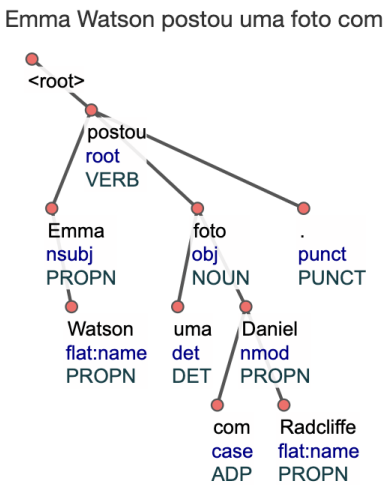


Figura 7.16: Diagrama de árvore de dependência gerado pelo UDPipe.



https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

7.4 Visualização, anotação e edição de *treebanks*

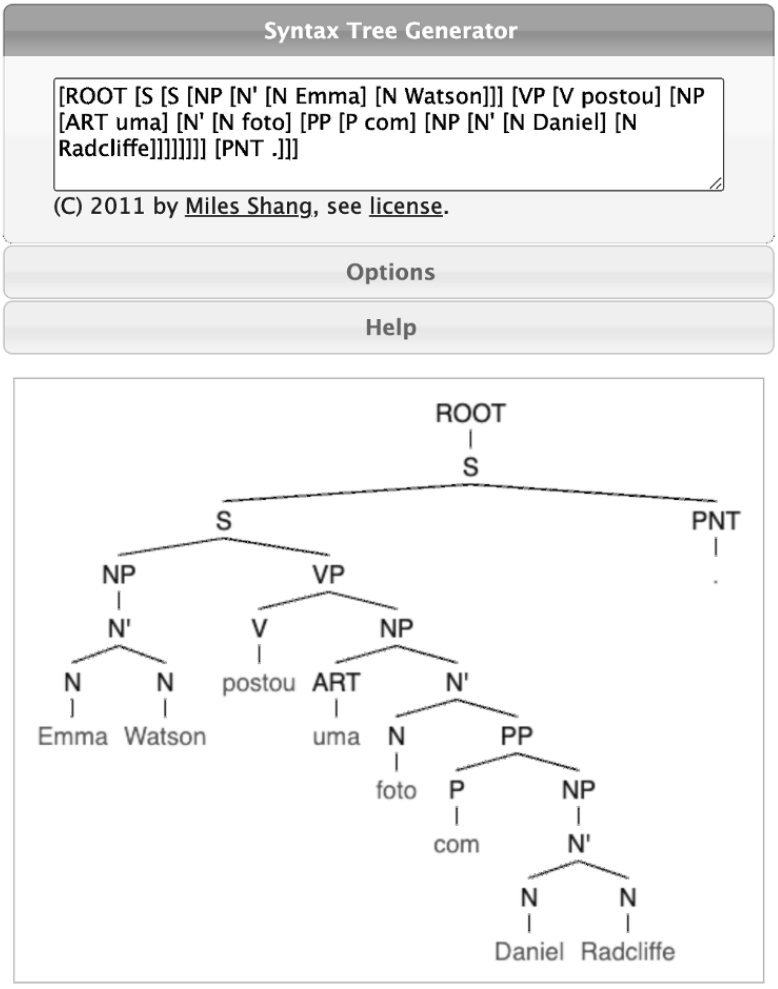
Há diversas ferramentas que permitem tanto a visualização de árvores de constituência e dependência, como a anotação de sentenças e a edição de sentenças já anotadas manualmente ou automaticamente.

7.4.1 Árvores de constituência

Há ferramentas web que oferecem uma visualização gráfica de diagrama de árvores a partir da notação entre colchetes ou parênteses dada como entrada pelo usuário.

Uma dessas ferramentas é *Syntax Tree Generator* (Syntree)[31], a qual dada uma entrada com notação em colchetes, gera um diagrama de árvore como mostrado na Figura 7.17.

Figura 7.17: Visualização de diagrama de árvore de constituência pela ferramenta Syntree.



Ferramentas semelhantes à Syntax Tree Generator são:

- Jssyntaxtree (*Dynamic JavaScript version of phpSyntaxTree*)[32]

[31] https://mshang.ca/syntree/
[32] https://ironcreek.net/syntaxtree/

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

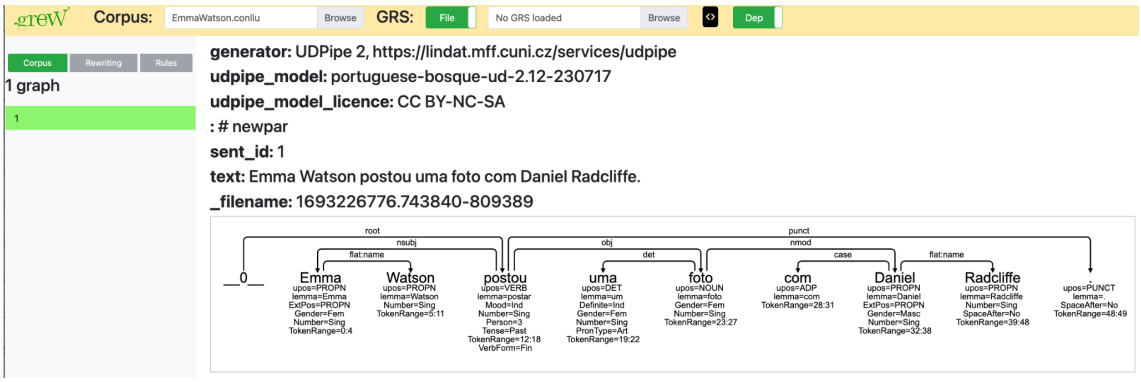
- RSyntaxTree (escrita na linguagem de programação Ruby)[33]

7.4.2 Árvores de dependência

7.4.2.1 Visualização de sentenças anotadas

Há ferramentas web que oferecem uma visualização gráfica de diagrama de árvores a partir do arquivo em formato CoNLL-U dado como entrada pelo usuário. Por exemplo, a ferramenta Grew Web possibilita carregar um arquivo CoNLL-U e visualizar o diagrama de relações de dependência, como mostrado na Figura 7.18.

Figura 7.18: Visualização de diagrama de árvore de dependência pela ferramenta Grew Web.



Uma ferramenta semelhante à Grew Web é CoNLL-U Viewer.[34]

7.4.2.2 Buscas em *treebanks*

Há ferramentas web de busca em *treebanks* anotados com relações de dependência. Esse é o caso da ferramenta Grew Match, que possibilita o acesso a 245 *treebanks* e distintos tipos de busca (por exemplo: por palavra, lema, etiqueta de PoS, etiqueta de relação de dependência, ngramas de palavras, lemas e PoS etc.). A Figura 7.19 mostra uma captura de tela com os resultados de uma busca pela palavra "foto" no *treebank* de dissertações e teses em português brasileiro no domínio do petróleo Petrogold.

Uma ferramenta semelhante à Grew Match é TüNDRA (*Tübingen aNnotated Data Retrieval Application*).[35]

7.4.2.3 Anotação e edição manual de relações de dependência

Para editar arquivos CoNLL-U previamente anotados de forma manual ou automática, uma das ferramentas mais utilizadas é ArboratorGrew[36], que possui uma versão customizada no Brasil pela equipe do ICMC da USP: Arborator-Grew-NILC[37]. Essa ferramenta permite gerenciar projetos individuais e coletivos de anotação, bem como serve de plataforma instrucional para cursos e treinamentos em anotação de sintaxe de dependência. A

[33] https://yohasebe.com/rsyntaxtree/
[34] https://universaldependencies.org/conllu_viewer.html
[35] https://weblicht.sfs.uni-tuebingen.de/Tundra/
[36] https://arboratorgrew.elizia.net/#/
[37] https://arborator.icmc.usp.br/#/

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Figura 7.19: Captura de tela da ferramenta Grew Match.

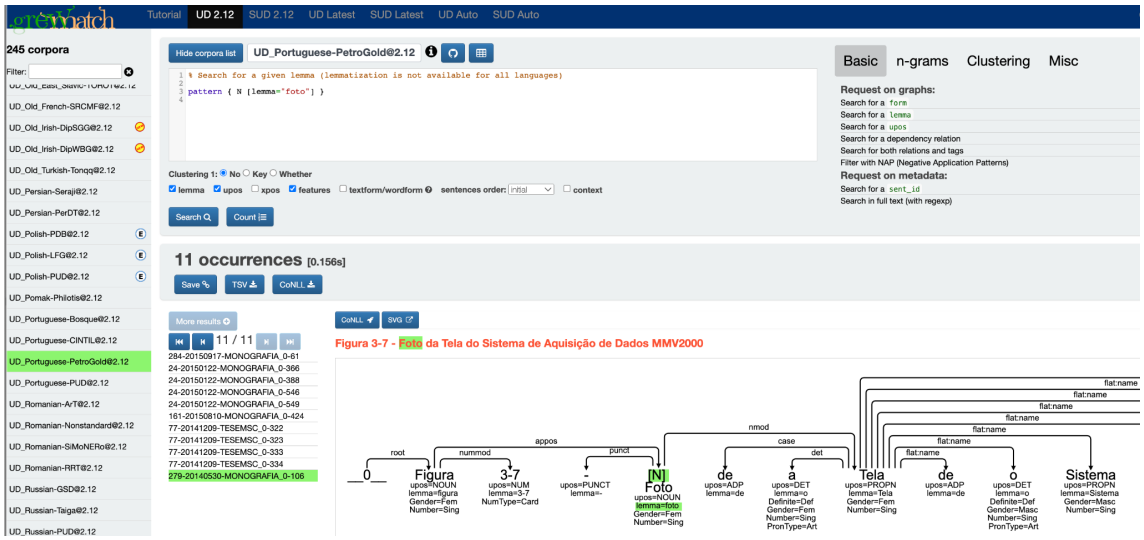


Figura 7.20 mostra a interface da ferramenta no momento de edição de uma das etiquetas de PoS.

A ferramenta Arborator Grew recebe como entrada arquivos CoNLL-U e possibilita a exportação dos arquivos CoNLL-U editados, bem como das imagens dos diagramas de dependência.

Há muitas outras ferramentas de visualização, consulta, anotação e edição de sintaxe de dependência. Nos últimos anos, o projeto Universal Dependencies vem a atualizando a lista de ferramentas disponíveis, a qual pode ser consultada no site do projeto[38].

7.4.3 Anotação de *corpus* em múltiplos níveis

Há ferramentas que permitem a anotação de sintaxe juntamente com a anotação em outros níveis, como é o caso de entidades nomeadas, relações entre entidades, correferência e outras. Uma das ferramentas mais robustas disponíveis atualmente e com interface amigável e INCEpTION[39], desenvolvida pelo *Ubiquitous Knowledge Processing (UKP) Lab* do *Department of Computer Science* da *Technische Universität Darmstadt* (Klie et al., 2018).

INCEpTION é apresentada como um ambiente computacional e plataforma de anotação semântica. É uma aplicação web que permite que vários anotadores trabalhem num mesmo projeto, sendo que a instalação é feita na máquina local do anotador, que utiliza a ferramenta por meio de um arquivo executável java e um *localhost*. Possui esquemas prontos de anotação de PoS e relações de dependência e permite anotar várias sentenças e as relações de correferência e outras relações entre elas, como ilustrado na Figura 7.21.

A Figura 7.21 mostra uma captura de tela com anotação em múltiplos níveis na ferramenta.

[38] https://universaldependencies.org/tools.html
[39] https://inception-project.github.io/

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Figura 7.20: Captura de tela no momento de edição de etiqueta de PoS na ferramenta Arborator-Grew-NILC.

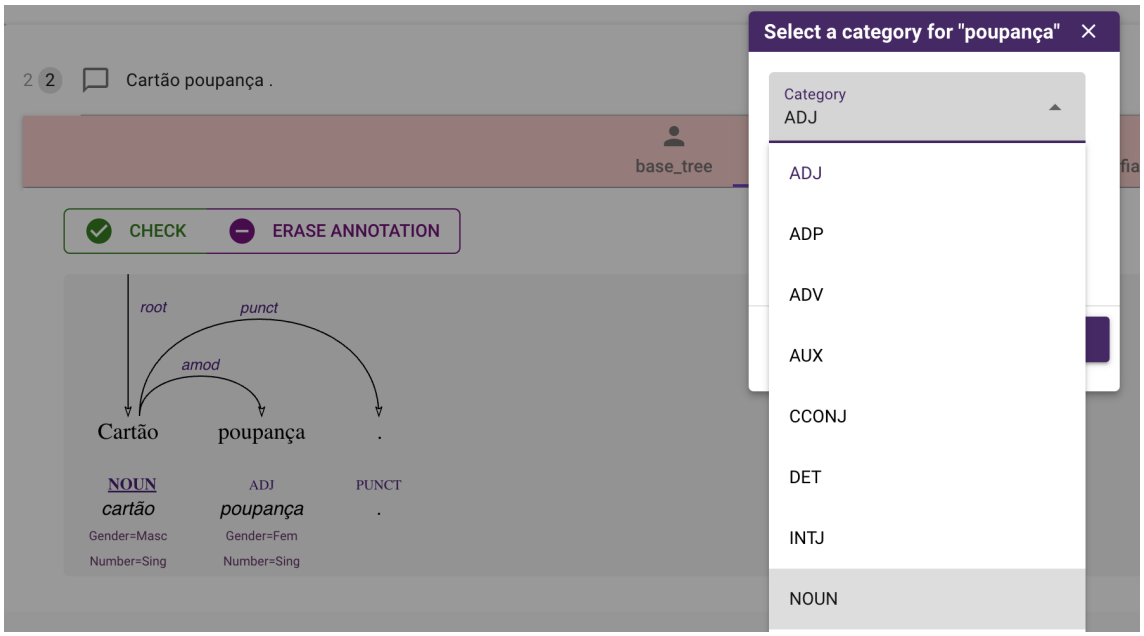
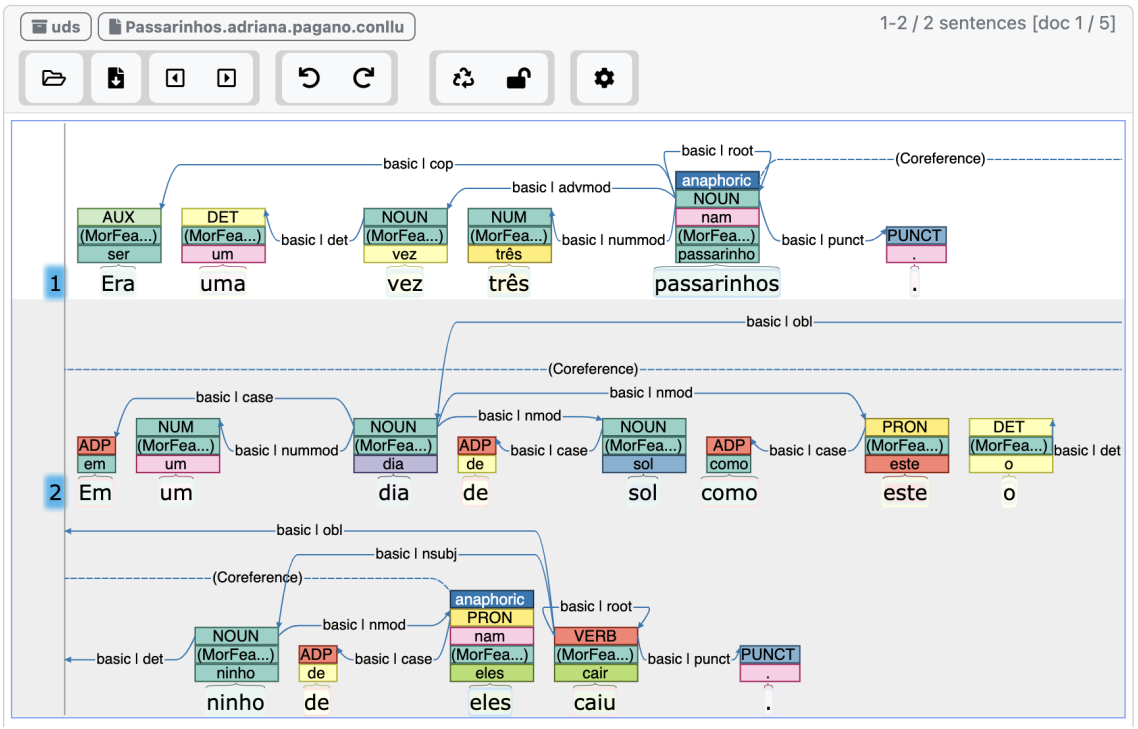


Figura 7.21: Captura de tela com anotação em múltiplos níveis na ferramenta INCEpTION.



©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

7.5 Considerações finais

No cenário de PLN, a análise sintática desempenha um papel importante na compreensão e interpretação de textos. Como vimos, várias bibliotecas e ferramentas foram propostas para trabalhar com a língua portuguesa, oferecendo soluções para a análise sintática de sentenças. A interseção entre linguística computacional e programação torna a análise sintática acessível mesmo para aqueles que não são especialistas em PLN, abrindo portas para uma compreensão mais profunda dos textos em língua portuguesa e sua estrutura intrínseca.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Parte V

Significado

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Capítulo 8

E o significado?

Cláudia Freitas

Publicado em: 26/09/2023

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

Semântica lida com o sentido do que é comunicado por meio da linguagem (em oposição ao que é comunicado por imagens ou sons não verbais, por exemplo). Assim, a **semântica estuda o significado de palavras e frases**. Mas a simplicidade relativa ao que é semântica acaba aí. Nos estudos linguísticos, a semântica é conhecida como "um domínio de investigação de limites movediços" e para o qual não há jargões bem estabelecidos (Ilari; Geraldi, 1985, p. 6). A questão "o que é o significado de uma palavra?" (e também o de uma frase) é um dos problemas nucleares da investigação semântica, e sua resposta irá depender da perspectiva teórica adotada. Essa característica é diferente de outras áreas do conhecimento, como a zoologia, por exemplo, em que não há controvérsia sobre o que é um animal.

No PLN, diferentes maneiras de conceber o significado se manifestam em diferentes abordagens para o tratamento do sentido, como veremos nos Capítulos 9 e 10.

Mas... sabemos o que é significado e os dicionários – objetos que contêm o **significado** das palavras – não só existem, como são úteis. Então, por que tanta dificuldade? Por que "limites movediços"?

Para ilustrar essa ideia, vamos fazer uma analogia entre a observação – e descrição – do significado de uma palavra e a observação de uma onda, como narrada em "Palomar na praia", de um livro de Ítalo Calvino. Trata-se de um capítulo curto, e alguns trechos são transcritos abaixo. A história gira em torno de alguém – Palomar – tentando conhecer algo – uma onda – de forma maximamente objetiva. Palomar, aliás, além de nome do protagonista, é o nome de um observatório astronômico que, durante muito tempo, ostentou o maior telescópio do mundo[1]. Vamos à história, apresentada no Quadro 8.1.

Quadro 8.1: Trechos da história "Palomar na praia" de Ítalo Calvino.

> O senhor Palomar está de pé na areia e observa uma onda. Não que esteja absorto na contemplação das ondas. Não está absorto, porque sabe bem o que faz: quer observar uma onda e a observa. (...). Em suma, não são "as ondas" que ele pretende observar, mas uma simples onda e pronto: no intuito de evitar as sensações vagas, ele predetermina para cada um de seus atos um objetivo limitado e preciso.
> O senhor Palomar vê uma onda apontar na distância, crescer, aproximar-se, mudar de forma e de cor, revolver-se sobre si mesma, quebrar-se, desfazer-se. A essa altura poderia convencer-se de ter levado a cabo a operação a que se havia proposto e ir-se embora. Contudo, isolar uma onda da que se lhe segue de imediato e que parece às vezes suplantá-la ou acrescentar-se

[1] Devo à Helena Franco Martins a apresentação deste texto como alegoria tanto para as tentativas de apreensão do significado como para a crise relativa ao conhecimento/ciência.

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

> a ela e mesmo arrastá-la é algo muito difícil, assim como separá-la da onda que a precede e que parece empurrá-la em direção à praia, quando não dá até mesmo a impressão de voltar-se contra ela como se quisesse fechá-la. (…).
> Em suma, não se pode observar uma onda sem levar em conta os aspectos complexos que concorrem para formá-la e aqueles também complexos a que essa dá ensejo. Tais aspectos variam continuamente, decorrendo daí que cada onda é diferente de outra onda; mas da mesma maneira é verdade que cada onda é igual a outra onda, mesmo quando não imediatamente contígua ou sucessiva; enfim, são formas e sequências que se repetem, ainda que distribuídas de modo irregular no espaço e no tempo. Como o que o senhor Palomar pretende fazer neste momento é simplesmente ver uma onda, ou seja, colher todos os seus componentes simultâneos sem descurar de nenhum, seu olhar se irá deter sobre o movimento da água que bate na praia a fim de poder registrar os aspectos que a princípio não havia captado (…).
> (…) Foi uma dessas línguas baixas de areia que o senhor Palomar escolheu como ponto de observação, porque as ondas nelas batem obliquamente de uma parte e de outra, e ao cavalgarem por cima da superfície semi-submersa vão encontrar-se com as que chegam da outra parte. (…).
> O senhor Palomar está procurando agora limitar seu campo de observação; se tem presente um quadrado de, digamos, dez metros de praia por dez metros de mar, pode levantar um inventário de todos os movimentos de ondas que ali se repetem com frequência variada dentro de um dado intervalo de tempo. A dificuldade está em fixar os limites desse quadrado, porque, por exemplo, se ele considera como o lado mais distante de si a linha em relevo de uma onda que avança, essa linha ao aproximar-se dele irá, erguendo-se, ocultar de sua vista tudo o que está atrás; e eis que o espaço tomado para exame se destaca e ao mesmo tempo se comprime. (…)
> Contudo, o senhor Palomar não perde o ânimo e a cada momento acredita haver conseguido observar tudo o que poderia ver de seu ponto de observação, mas sempre ocorre alguma coisa que não tinha levado em conta. Prestar atenção em um aspecto faz com que este salte para o primeiro plano, invadindo o quadro, como em certos desenhos diante dos quais basta fecharmos os olhos e ao reabri-los a perspectiva já mudou. (…).
> O vento estaria mudando? É pena que a imagem que o senhor Palomar havia conseguido organizar com tanta minúcia agora se desfigure, se fragmente e se perca. Só conseguindo manter presentes todos os aspectos juntos, ele poderia iniciar a segunda fase da operação: estender esse conhecimento a todo o universo.
> Bastaria não perder a paciência, coisa que não tarda a acontecer. O senhor Palomar afasta-se ao longo da praia, com os nervos tensos como havia chegado e ainda mais inseguro de tudo.
> CALVINO, Ítalo. *Palomar*. São Paulo: Companhia das Letras, 1994. p.7-11.

O que vemos, por trás da simplória tarefa de observação de **uma única onda**, é a dificuldade de Palomar diante de um objeto que se transforma continuamente durante a própria atividade de observação. Ainda que Palomar defina, de modo preciso, seu objetivo e seu objeto ("observar uma simples onda e pronto") e busque uma abrangência descritiva ("Colher todos os seus componentes simultâneos sem descurar de nenhum"), é difícil, na observação, isolar o objeto de suas "adjacências", reduzir as diferentes instâncias do objeto a uma essência comum ("sempre ocorre alguma coisa que não tinha levado em conta"), controlar a subjetividade, suspender as pressões externas ("bastaria não perder a paciência"), encontrar o ponto de vista superior ou ideal ("Foi uma dessas línguas baixas de areia que o sr. Palomar escolheu como ponto de observação"). Enfim, definitivamente, Palomar não é bem sucedido em sua empreitada, por mais simples que esta parecesse inicialmente.

De volta à semântica, podemos nos imaginar como Palomar na tentativa de observar o significado de uma única palavra. Podemos escolher a palavra "quente", e teremos uma "sopa quente", um "dia quente" e uma "cerveja quente". Estamos tratando da mesma

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

temperatura, do mesmo significado, o que há em comum em todos eles, e que os define? Podemos escolher outra palavra, "medo", em "medo de altura", "medo de perder o emprego", "medo de barata", "medo do mar", "medo de sofrer". Estamos falando exatamente do mesmo "medo"? Qual o significado exato, preciso, de "medo"? Se estamos diante de uma mesma palavra, e de uma palavra que sabemos usar, não seria esperado que soubéssemos definir, de forma clara e precisa, seu significado? Qual o significado (ou significados) de "tomar", tomando como exemplos combinações como "tomar um susto", "tomar um porre", "tomar cuidado", "tomar um suco", "tomar remédios", "tomar uma decisão", "tomar conta", "tomar um tombo", "tomar ciência", "tomar porrada", "tomar dois banhos", "tomar um susto" etc.

É exatamente este tipo de dificuldade que justifica a existência, nos estudos linguísticos, de duas grandes perspectivas que irão problematizar o que é o significado. São perspectivas concorrentes, e de um modo bastante simplificado podemos chamá-las de **representacionistas ou essencialistas**, por um lado, e de **pragmáticas** (ou, **pragmáticas radicais**), por outro (Martins, 2000, 2004). No PLN, estas visões se manifestam em diferentes maneiras de lidar com o significado: usando técnicas simbólicas (veja Capítulo 9) ou usando representações distribuídas (veja Capítulo 10), respectivamente.

A perspectiva **representacionista/essencialista** é a visão hegemônica, estando presente em boa parte dos estudos linguísticos e no senso comum – e, até recentemente, em boa parte do PLN também.

Nesta visão, palavras seriam como "substitutos" de **entidades extralinguísticas**, entidades externas à linguagem (entidades mentais, reais ou virtuais). As palavras, nessa perspectiva, importam pouco, importando mesmo as ideias (as entidades extralinguísticas) que elas **representam**. Significado e palavra são, assim, entidades distintas, ainda que relacionadas (falamos "do significado das palavras", por exemplo), e a relação entre elas é hierárquica, com a entidade significado se sobrepondo à entidade palavra (ou à palavra e seus sinônimos), que apenas fornece matéria/forma para "hospedar" o significado.

Ainda de acordo com esta visão, apesar da multiplicidade de usos e contextos que podem existir associados a uma mesma palavra (por exemplo "tomar", ou "quente", ou ainda "liberdade", "violência", "aprender", "significado", "compreensão") a comunicação é possível porque esses diferentes usos estão **associados** a uma **essência comum** (a entidade extralinguística), e por isso reconhecemos a palavra como sendo a mesma em diferentes situações. A associação entre a palavra (ou a palavra e seus sinônimos) e sua essência/ideia/conteúdo/conceito/significado, por sua vez, é guiada por regras. **Aprender uma língua**, aqui, é aprender a estabelecer a conexão entre a palavra e a entidade extralinguística que ela representa (e diferentes línguas irão variar quanto às palavras usadas para representar estes conceitos/ideias). Este conceito/ideia/significado, que algo é separado da palavra, é um "objeto" extralinguístico (do mundo mental, real, virtual) estável e com contornos bem definidos – mas que, por sua vez, também será descrito por meio de palavras.

Podemos agrupar sob esta visão – apresentada aqui de maneira muito simplificada – uma série de correntes teóricas que, de alguma maneira, compartilham a ideia de que a **estabilidade do significado** (e a compreensão) é o **resultado da representação de algo que lhe é exterior**.

Ainda segundo esta visão, os significados das palavras são, de certo modo, o que o dicionário diz. O fato de dicionários representarem os significados de maneira objetiva, estável e discreta (vemos isso na maneira pela qual as acepções estão claramente separadas e nu-

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

meradas), faz parecer que os significados das palavras se organizam "naturalmente" assim[2]. No PLN, reconhecemos esta maneira de lidar com o significado em recursos como *wordnets*, por exemplo, que são bases de dados lexicais que contêm "nomes, verbos, adjetivos e advérbios agrupados em conjuntos de sinônimos cognitivos, **cada um representando um conceito distinto**"[3] (grifo meu). Uma apresentação do que se pode fazer partindo desse ponto de vista, e de por que ele continua tendo espaço no PLN, está no Capítulo 9.

Já do ponto de vista pragmático (ou, mais precisamente, pragmático **radical**)[4], – e fazendo igualmente uma apresentação bastante simplificada – o significado de uma palavra é decorrência de situações concretas (e não o correspondente a uma entidade extralinguística), e situações concretas são variáveis. Nesta visão, os vários usos de uma palavra não se organizam em torno de um núcleo semântico comum (a entidade extralinguística), garantidor da estabilidade do que elas significam. A estabilidade do significado será sempre **provisória**, e o significado dependerá do uso, do contexto, do tempo, do espaço, de quem fala … A **comunicação se dá no risco** (isto é, pode dar certo ou não, podemos nos entender ou não), e os mal-entendidos existem, estão aí – não são um desvio ou uma falha, são parte do jogo. O que determina se **compreendemos o significado** de um enunciado linguístico é o fato de **a manifestação dessa compreensã**o (um comportamento) **ser considerada adequada** no contexto em que é produzida. Por exemplo, a um enunciado como "Está quente aqui", seriam manifestações legítimas ações como abrir a janela ou respostas como "Não acho" ou "Por que não tira o casaco?", entre outras. Mas dificilmente aceitaríamos como manifestações de compreensão do enunciado "Está quente aqui" dar uma cambalhota ou uma resposta como "Prefiro melão"[5]. **Aprender uma língua**, aqui, é aprender a tomar parte nas atividades humanas, um aprendizado que nunca se completa.

Assim como uma onda, os limites do significado de uma palavra não têm – aliás, podem não ter, pois não se trata de uma exigência – a precisão ou os limites definidos, necessários à formalização que sempre se buscou fazer. Segundo esta visão, o significado é flexível e maleável, não havendo uma "essência", algo que perpassa todos os usos, e sobre o qual seja possível se sustentar, se estabilizar. No PLN, esta visão se alinha às representações distribuídas (veja Capítulo 10)[6].

Sabemos que uma mesma palavra pode aparecer em contextos diferentes – desde contextos completamente distintos, como "banco" e "manga", até contextos ligeiramente diferentes, como os exemplos de "quente", "medo" ou "tomar", que já vimos[7]. Nesse caso, e considerando os modelos de representações distribuídas mais complexos e dinâmicos, cada forma "quente" ou "medo" será representada de uma maneira – e por isso nesses casos falamos de vetores contextuais (*contextual word embeddings*). Nos vetores estáticos, que irão representar de uma única maneira as várias formas "quente" ou "medo", o alinhamento à visão não-representacionista/não essencialista se mantém, uma vez que não há uma fonte

[2] Por trás de dicionários estão lexicógrafos e decisões editoriais.

[3] http://wordnet.princeton.edu/

[4] Dentre as linhas de investigação pragmáticas há as que poderiam ser também enquadradas em um paradigma representacionista. Isto porque algumas correntes da pragmática recomendam a análise das propriedades da prática da comunicação como maneira de fornecer uma explicação do que são as línguas e os significados. Por isso a especificação indicando a "radicalidade" da visão que será apresentada.

[5] Mas mesmo estas poderiam ser aceitas se assim fosse previamente estipulado.

[6] Mas não se alinha à busca do algoritmo capaz de fornecer a representação distribuída "correta", ou "verdadeira".

[7] É importante notar que, diferentemente do que supõe o senso comum, os casos de "banco" ou "manga", apesar de fartamente citados como exemplos de ambiguidade, estão longe de ser prototípicos. Pelo contrário, são raros os casos em que dois sentidos se apresentam tão claramente distintos. O mais comum são casos como "quente" ou "medo".

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

(ou entidade) externa que determina o significado da palavra. Vetores produzidos a partir de conjuntos de dados diferentes irão levar a representações diferentes[8].

Durante muito tempo, a semântica computacional esteve ancorada em visões essencialistas-representacionistas (ou simbólicas), como ilustram os capítulos *"Semantics"* de dois compêndios da área de PLN: Jurafsky; Martin (2023) e Mitkov (2003). No entanto, trabalhos de PLN que dialogam claramente com perspectivas não-essencialistas também não são novidade, como Kilgarriff (1997); Kilgarriff (2003) e Brewster; Wilks (2004), por exemplo.

Entre as técnicas simbólicas e as representações distribuídas existem ainda os *datasets* (ou *corpora*) com **anotação semântica** (veja Capítulo 16), uma terceira maneira de lidar com o significado no PLN. Se, por um lado, tais *datasets* se alinham às abordagens probabilísticas, uma vez que podem ser usados como fonte para o aprendizado de máquina (para o **aprendizado do significado**), por outro lado, a atividade de anotação de significado se alinha às abordagens representacionistas. Neste tipo de anotação (também chamada de anotação de *word senses*), cada palavra (ou segmento de texto) é anotada com informação relativa ao significado de acordo com o contexto específico em que aparece no *corpus*. A informação relativa ao significado, por sua vez, vem de **fontes externas** (como dicionários, *wordnets*, *verbnets* e *framenets*) e a tarefa de anotação pode ser descrita como um **trabalho** de desambiguação, pois consistiria em selecionar, dentre os vários sentidos possíveis de uma palavra, aquele usado no contexto da frase. O que a anotação faz, deste modo, é criar uma **representação estável** entre a palavra e o seu significado, no contexto em que está sendo usada. Cada ocorrência de uma palavra poderá estar associada a um significado diferente (e aqui vemos uma aproximação com abordagens pragmáticas), desde que este significado esteja presente no inventário de significados usado na anotação (aqui vemos uma aproximação com abordagens representacionistas). Para as pessoas responsáveis pela anotação, a principal dificuldade está na escolha do sentido adequado conforme o contexto, uma vez que os sentidos frequentemente se sobrepõem, como as ondas observadas por Palomar.

Por exemplo, tomando a palavra "trabalho" destacada no parágrafo anterior, a tarefa consiste em escolher, dentre opções listadas no quadro abaixo, retiradas do dicionário Caldas-Aulete online[9], aquela adequada ao contexto (se a anotação usasse o inventário de uma *wordnet* como fonte, o inventário de significados poderia ser diferente[10]).

Quadro 8.2: Acepções da palavra *trabalho* conforme dicionário

1. Emprego da força física ou intelectual para realizar alguma coisa
2. Aplicação dessas forças como ocupação profissional: *Seu trabalho é de gari.*
3. Local onde isso se realiza: *Mora longe do trabalho.*
4. Esmero, cuidado que se emprega na confecção ou elaboração de uma obra
5. A confecção, elaboração ou composição de uma obra
6. Obra realizada: *Essa cômoda é um belo trabalho de marcenaria.*
7. Grande esforço; TRABALHÃO; TRABALHEIRA
8. Exercício para treino: *A professora passou muito trabalho para casa.*

[8] Mas mesmo representações distribuídas podem ser associadas a visões representacionistas, quando se assume que tais representações são úteis apenas enquanto não encontramos a forma (ou a representação) correta de uma palavra.

[9] https://www.aulete.com.br/trabalho

[10] Aqui é possível consultar a OpenWordNet-PT para os significados de "trabalho": https://www.openwordnet-pt.org/search?search_field=all&term=trabalho

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

9. Ação contínua de uma força da natureza e seu efeito: *O trabalho do vento resulta na erosão eólia.*
10. Med. Fenômeno orgânico que se opera no interior dos tecidos (trabalho inflamatório; trabalho de cicatrização)
11. Resultado do funcionamento de uma máquina, um aparelho etc.: *o trabalho de uma pá mecânica.*
12. Obrigação ou responsabilidade; DEVER; ENCARGO: *Seu trabalho é protegê-lo do assédio da imprensa.*
13. Econ. Conjunto das atividades humanas empregado na produção de bens: *O capital e o trabalho são os pilares da economia.*
14. Tarefa a ser realizada: *Contratou-o para um trabalho temporário.*

Fonte: (Freitas, 2022)

Parece que as acepções 1, 2 e 14 são aceitáveis no contexto da frase, o que já é um problema se precisamos escolher apenas um sentido, e por isso não é exagero dizer que as pessoas responsáveis pela anotação se sentem como Palomar na tentativa de isolar uma onda.

Corpora anotados com este tipo de informação são escassos, e um dos motivos é justamente a dificuldade de isolar o significado/conteúdo/essência das palavras enquanto estão sendo efetivamente usadas.

O estudo de Baker et al. (2017) tentou entender por que, com este tipo de anotação, era tão difícil conseguir uma boa concordância entre anotadores (veja Capítulo 16), isto é, era tão difícil que as pessoas concordassem quanto à escolha do significado utilizado. Afinal, a tarefa é simples: associar cada palavra ao seu significado, e, se sabemos a nossa língua, sabemos o significado das palavras que usamos. No estudo, diferentes pessoas deveriam anotar as mesmas palavras, nas mesmas frases, considerando o mesmo inventário de sentidos. Os resultados indicaram uma variação bem maior que o previsto[11]. Vamos lembrar que, na **anotação**, os significados precisam ser vistos como **unidades discretas e de conteúdo estável** – uma necessidade de ordem prática que se alinha harmoniosamente com visões representacionistas, mas que não encontra respaldo em visões pragmáticas.

O fato de representações distribuídas terem levado a resultados positivos no PLN não deve ser visto como argumento contrário às técnicas simbólicas. São maneiras diferentes de lidar com o sentido das palavras. Como tirar o melhor proveito destas diferentes visões e abordagens, no PLN, é uma das questões que se coloca. O que temos visto é a limitação de cada uma delas, tomada individualmente. Se consideramos o significado como uma entidade estável, como lidar com as mudanças, que inclusive podem ser capturadas pelos dicionários (dicionários, recentemente, mudaram a definição da palavra "família"[12])? Por outro lado, se consideramos a instabilidade e a dependência dos dados, como evitar vieses indesejados, como a associação entre os sentidos, por exemplo, de "paraguaio" e "de baixa qualidade", quando dizemos "uísque paraguaio"?

Os próximos capítulos aprofundam cada uma dessas maneiras de trabalhar com o significado no PLN.

[11] A seção "Anotações Semânticas" de Freitas (2022) traz um levantamento dos principais estudos sobre anotação semântica e seus desafios, bem como uma apresentação linguística da alternativa oferecida pelas representações distribuídas para o tratamento do significado.

[12] Dicionários mudam definição de família https://www.metropoles.com/vida-e-estilo/comportamento/dicionario-houaiss-muda-significado-da-palavra-familia

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Capítulo 9

Semântica com técnicas simbólicas

Eloize Rossi Marques Seno
Valéria de Paiva
Vládia Pinheiro

Publicado em: 26/09/2023

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

Métodos simbólicos em Processamento de Linguagem Natural (PLN) envolvem a utilização de regras e representações formais explícitas para processar e entender textos em linguagem natural. Esses métodos especializam-se na manipulação de símbolos e dados estruturados, como gramáticas, ontologias e bases de conhecimento. Especificamente para o entendimento de textos em linguagem natural usando técnicas simbólicas, existem analisadores semânticos (ou *parsers* semânticos) e bases de conhecimento semântico, que visam fornecer uma representação semântica dos textos. A partir desta representação, motores de inferência são capazes de realizar raciocínio para que aplicações possam, por exemplo, extrair informações, sumarizar textos, e responder perguntas com base nos textos.

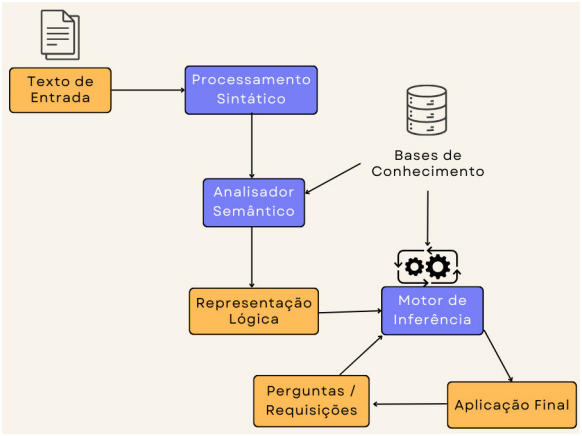
A Figura 9.1 apresenta uma arquitetura tradicional para sistemas de entendimento de textos em linguagem natural (*Natural Language Understanding* – NLU). A partir do texto de entrada, uma camada de processamento sintático realiza uma série de análises no texto, tais como detecção de língua, separação de sentenças, tokenização, análise morfológica e sintática (Capítulo 4). Na fronteira entre o processamento sintático e a análise semântica, outros processamentos linguísticos são necessários, como reconhecimento de entidades nomeadas, identificação de expressões multipalavras etc. Em seguida, o texto analisado (sintaticamente) é enviado ao analisador semântico (*parser*) que gera uma representação lógica do texto. A representação lógica e a(s) base(s) de conhecimento, no que lhes concerne, são entradas para o motor de inferência. Nesse processo, termos do texto de entrada são associados aos elementos da base de conhecimento e o motor de inferência gera respostas a perguntas (*queries*) para uma aplicação final.

É pertinente fazer uma observação neste ponto para uma definição de base de conhecimento. Uma "base de conhecimento" refere-se a um repositório centralizado, processável por máquina, que contém informações, dados, regras e procedimentos que são usados para capturar, representar e armazenar conhecimento geral ou de um domínio específico. Tais bases de conhecimento são fontes de conhecimento de mundo e suportam diversas tarefas e aplicações em PLN. Uma base de conhecimento pode ser estruturada de diversas maneiras, incluindo bancos de dados relacionais, linguagem para ontologias (e.g. a OWL[1]), formalismos para troca de dados entre sistemas (e.g. o formato JSON[2]), redes semânticas ou sistemas baseados em regras, dependendo da sua finalidade e da natureza do conhecimento armazenado.

[1] https://www.w3.org/OWL/
[2] https://www.json.org/json-pt.html

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Figura 9.1: Arquitetura de Sistemas de Entendimento de Linguagem Natural (NLU).



Fonte: Adaptada de (Ovchinnikova, 2012, p. 9)

Tradicionalmente, sistemas lógicos são usados para representação formal dos textos e seus motores de inferência servem para gerar conclusões a partir dos textos. Podemos citar os sistemas lógicos mais usados em PLN: variações de Lógica Descritiva (*Description Logic – DL*) (Baader et al., 2003), da Lógica de Primeira Ordem (Blackburn; Bos, 2005; Eijck; Unger, 2010), vários tipos de Programação em Lógica (PROLOG) (Dahl, 1994) e Lógicas Intensionais (Shapiro, 2000).

Uma característica importante dos sistemas lógicos usados para a semântica de linguagem natural é que eles dependem fortemente da forma lógica do texto ou argumento. No entanto, muitas conclusões e respostas fornecidas ao se ler um texto são justificadas pela contribuição semântica dos conceitos relacionados, não definida a priori, mas somente enquanto usados em um contexto particular. Por exemplo, considere a inferência que conclui que "Alguém foi assassinado" a partir da premissa que "Alguém foi executado". A contribuição semântica do conceito "executar" (no sentido de "assassinar") é que torna esta inferência plausível, e não a forma da sentença. Da mesma forma, a inferência "um relâmpago é visto agora" para "um trovão será ouvido em breve" é autorizada pelo conteúdo dos termos "trovão" e "relâmpago". Para se realizar inferências desta natureza, alguns filósofos como Sellars (Sellars, 1953) e Brandom (Brandom, 2001) propõem abordagens para expressão do significado que suportam análises semânticas não somente sobre a forma das sentenças, mas são capazes de, com base no domínio dos conteúdos dos termos articulados nas sentenças e textos, descobrir como estes [os conteúdos dos termos] contribuem conjuntamente para o significado das sentenças e para realização de inferências.

Este capítulo tem como objetivo examinar, além dos *frameworks* semânticos, tais como AMR (*Abstract Meaning Representation*) (Banarescu et al., 2013) e DELPH-IN (Copestake et al., 2005), os tipos de bases de conhecimento mais utilizados em PLN. Nesta versão inicial do capítulo, são apresentadas as bases (também chamadas de recursos léxico-semânticos) WordNet de Princeton (Fellbaum, 1998) e FrameNet (Baker et al., 1998), e suas versões em português: OpenWordNet-PT (De Paiva et al., 2012) e FrameNet Brasil (FN-BR) (Torrent; Ellsworth, 2013); bem como bases de conhecimento voltadas ao senso comum, tais como a ConceptNet (Speer et al., 2016) e iniciativas para o português (OMCS-BR (Anacleto et al., 2006) e a InferenceNet-BR (Pinheiro et al., 2010)). Existem outros tipos de bases de conhecimento na área do PLN, tais como dicionários e ontolo-

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

gias diversas, por exemplo, WikiData (Vrandečić; Krötzsch, 2014), YAGO (Suchanek et al., 2007) ou BabelNet (Navigli; Ponzetto, 2012). No entanto, nossa descrição aqui visa apenas a uma primeira exposição dos distintos paradigmas de expressão de conhecimento semântico. Tendo em vista cada base de conhecimento descrita, analisamos um exemplo de texto motivador. Por fim, apresentamos as considerações finais deste capítulo.

9.1 Bases de Conhecimento Semântico

Na área da Inteligência Artificial (IA), o interesse por bases de conhecimento computáveis ou processáveis por máquina surgiu na década de 60 com as primeiras redes semânticas e representações baseadas em *frames*, propostas por Minsky (Minsky, 1975) e Fillmore (Fillmore et al., 1976), respectivamente.

A comunidade de PLN foi rapidamente atraída por tais representações de conhecimento de mundo, pois pareciam prover a solução para problemas de semântica de linguagem natural. As mais antigas abordagens em PLN que utilizaram redes semânticas e *frames* remontam aos trabalhos de Bates et al. (1982) e Bobrow et al. (1977), conforme citado em (Ovchinnikova, 2012).

Neste capítulo, examinaremos dois tipos de bases de conhecimento: (1) recursos léxico-semânticos e (2) bases de conhecimento de senso comum.

1. **Recursos Léxico-Semânticos**
   Um conjunto de palavras existentes em uma determinada língua é chamado de léxico da língua e cada elemento do léxico é chamado de item lexical (Capítulo 4). Estes itens, quando organizados e agrupados de forma a facilitar o uso em processos computacionais, formam uma **base de conhecimento lexical** ou um **recurso léxico-semântico**. Uma *wordnet*[3] é um exemplo canônico desse tipo de base, onde a organização dos itens lexicais se dá através de relações semânticas, predominantemente, de hierarquia (hiperonímia/hiponímia), de inclusão (holonímia/meronímia), de equivalência (sinonímia) ou de oposição (antonímia). Essas bases incluem normalmente informação sobre os possíveis sentidos das palavras (por exemplo, um sentido de "manga" é a fruta tropical, mas "manga" também pode ter o sentido de parte de uma camisa), as relações entre sentidos (carro e pneu como merônimo-holônimo; quente e frio como antônimos), e definições e frases que exemplificam a sua utilização. Os recursos da família das *wordnets* são bases muito usadas em PLN e com uma história de sucesso dada a cobertura, variedade de relações e organização do conteúdo, além da facilidade de incorporação em aplicações e ferramentas que precisam entender textos em linguagem natural, através de *toolkits* como o NLTK[4] e spaCy[5]. Dentre as *wordnets*, temos a original e a mais proeminente – a WordNet de Princeton ou PWN[6] (Fellbaum, 1998).
2. **Bases de Conhecimento de Senso Comum**
   Para a comunidade de Inteligência Artificial, a expressão "conhecimento de senso comum" se refere aos fatos e conhecimentos informais possuídos pela maioria das pessoas, frutos da experiência da vida diária e baseados na generalização de eventos ou interpretações particulares, sem comprovação formal. Consiste em conhecimentos

[3] http://globalwordnet.org/
[4] https://www.nltk.org/
[5] https://spacy.io/universe
[6] http://wordnet.princeton.edu/

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

espaciais, físicos, sociais, temporais e psicológicos (Liu; Singh, 2004). As bases de conhecimento de senso comum expressam relações semânticas entre fragmentos de textos, tais como: relações funcionais, causais, afetivas, temporais, motivacionais, estruturais etc. Por exemplo, "bicicleta" é usada para "andar mais rápido que a pé", ou "cozinhar" é motivada por "fome". Nessa classe, se enquadram a larga base de senso comum ConceptNet[7] (Speer et al., 2016), originalmente gerada de conteúdo coletado de forma colaborativa na internet, e suas variações e congêneres.

Antes de iniciar a descrição das bases de conhecimento, introduziremos um exemplo de texto, Exemplo 9.1[8], em português brasileiro, para ser analisado conforme os insumos de cada base de conhecimento. Após o exemplo, indicamos algumas conclusões e respostas resultantes de inferências que pessoas, inseridas na cultura brasileira e proficientes no português, fariam ao ler o texto.

Exemplo 9.1:

> Um assalto do tipo saidinha bancária, ocorrido na tarde desta terça-feira, terminou com uma mulher de 42 anos baleada pelos assaltantes. O roubo ocorreu na rua Professor Costa Mendes.

Conclusões:

- O assalto teve uso de arma de fogo.
- A vítima estava em uma agência bancária.
- A motivação do crime foi conseguir dinheiro.

Nas próximas subseções, são descritas algumas das bases de conhecimento mais representativas para o PLN – Wordnet, FrameNet e a ConceptNet, e suas bases congêneres para o português. No final de cada subseção, discorremos sobre como essas bases contribuem para análise semântica do Exemplo 9.1. A escolha dessas três bases seguiu critérios de abrangência de suas entradas e representatividade para tarefas de PLN. A WordNet de Princeton é, consensualmente, o recurso léxico-semântico mais utilizado em PLN para dar suporte a tarefas como desambiguação de sentido de palavras, perguntas e respostas, e análise semântica. FrameNet é uma das bases mais relevantes para a tarefa de anotação de papéis semânticos (*Semantic Role Labeling* – SRL), pois atribui papéis semânticos não somente a verbos, mas também a termos das demais classes gramaticais. ConceptNet é a base de conhecimento de senso comum com mais entradas tanto para o inglês quanto para o português.

9.1.1 *Wordnets*

WordNet, desenvolvida por George A. Miller, Christiane Fellbaum e colaboradores, é considerada uma base de conhecimento léxico-semântica que organiza os itens lexicais (palavras ou expressões) em *synsets* (que vem de *synonym sets*, ou conjuntos de palavras sinônimas). A primeira *wordnet* foi desenvolvida para o inglês por George Miller, na Universidade de

[7] https://conceptnet.io/

[8] Esse texto foi adaptado de notícia publicada em jornal digital, disponível em https://diariodonordeste.verdesmares.com.br/seguranca/mulher-e-baleada-em-saidinha-bancaria-no-montese-1.933137?page=10.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Princeton, um projeto que se iniciou em 1985, e é ordinariamente chamada de WordNet de Princeton (ou Princeton WordNet, na sigla PWN)[9] e é descrita por Fellbaum (1998).

Wordnets são redes de palavras amplamente utilizadas em PLN para dar suporte a tarefas como desambiguação de sentido de palavras, perguntas e respostas, e análise semântica em geral. A unidade básica da WordNet são os *synsets* que representam conjuntos de palavras sinônimas. Cada *synset* expressa um conceito em particular. Os *synsets* têm uma glosa, semelhante a uma definição num dicionário e podem conter ainda frases que ilustram o emprego de alguma das suas palavras. A WordNet está dividida em quatro redes semânticas, uma para cada classe aberta de palavras: substantivo, verbo, adjetivo e advérbio.

Como exemplo, a Figura 9.2 apresenta os *synsets* da palavra *"murder"* (verbo "assassinar", em português) da PWN. Ao todo são três *synsets*, um na classe `Noun` (substantivo) e dois na classe `Verb` (verbo). O primeiro *synset* da palavra "*murder*" (na classe `Verb`) tem como tropônimos diretos os verbos "*burke*", *"execute"* e hiperônimo direto o *synset "kill"*.

Figura 9.2: Synsets da palavra *"murder"* na WordNet de Princeton.

WordNet Search - 3.1
- WordNet home page - Glossary - Help

Word to search for: murder [Search WordNet]

Display Options: (Select option to change) [Change]

Key: "S:" = Show Synset (semantic) relations, "W:" = Show Word (lexical) relations

Display options for sense: (gloss) "an example sentence"

Noun

- S: (n) **murder**, slaying, execution (unlawful premeditated killing of a human being by a human being)

Verb

- S: (v) **murder**, slay, hit, dispatch, bump off, off, polish off, remove (kill intentionally and with premeditation) *"The mafia boss ordered his enemies murdered"*
  - ***direct troponym*** / *full troponym*
    - S: (v) burke (murder without leaving a trace on the body)
    - S: (v) execute (murder in a planned fashion) *"The Mafioso who collaborated with the police was executed"*
  - *direct hypernym* / *inherited hypernym* / *sister term*
    - S: (v) kill (cause to die; put to death, usually intentionally or knowingly) *"This man killed several people when he tried to rob a bank"; "The farmer killed a pig for the holidays"*
  - *derivationally related form*
  - *sentence frame*
- S: (v) mangle, mutilate, **murder** (alter so as to make unrecognizable) *"The tourists murdered the French language"*

[9] http://wordnet.princeton.edu/

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

A PWN é a base léxico-semântica mais utilizada em PLN, com interfaces locais (APIs) em alguns dos maiores sistemas de programação (.NET/C#, dBase, Java, MySQL, OCaml, OSX, Perl, PHP, Prolog, Python, REST, SQL, Windows, XML)[10], mais de 20 mil citações no Google Scholar e dezenas de projetos que a utilizam. Apesar de ser tão utilizada, a WordNet de Princeton parou de evoluir em 2012, por falta de recursos financeiros. A última edição oficial de PWN foi a versão 3.1, lançada em 2011. Em 2019 um consórcio de pesquisadores, incluindo Christiane Fellbaum (a coordenadora da PWN), resolveu transformar a PWN em um recurso moderno, hospedado em GitHub, de tal forma que possa ser sempre atualizado (McCrae et al., 2019), mas a maior parte das aplicações continua usando PWN 3.0 ou 3.1.

Como todas as bases com conhecimento semântico, WordNet não é um projeto acabado e tampouco completo. Algumas lacunas decorrem da divisão e independência entre as redes semânticas, o que dificulta a expressão de relações estruturais (entidade-atributo), de relações semânticas que acontecem entre classes de palavras (verbos e substantivos, por exemplo) em uma particular situação ou contexto; ou informações sintagmáticas: relações que ocorrem entre os termos de um proferimento (entre verbo e substantivo, entre substantivo e adjetivo etc.). No que se refere aos tipos de relações expressas na PWN, essa dispõe de relações causais entre *synsets*, por exemplo, *"snore"* `implies` *"sleep"*, mas não numa taxa de cobertura suficiente em relação ao conjunto dos *synsets*. Outra limitação é que recursos como a PWN são mais adequados para substantivos concretos do que para conceitos abstratos como "medo", "felicidade" etc. Enquanto substantivos concretos como "gato", "felino", "mamífero", "animal" etc. são mais facilmente organizados em taxonomias, tal processo é menos consensual quando aplicado às emoções ou a verbos. Um quarto criticismo diz respeito às expressões multipalavras (MWEs – Capítulo 5). Essas existem em PWN, mas não na quantidade suficiente para a modelagem adequada da língua. De acordo com Sag et al. (2002, p. 2), o número de MWEs em PWN precisaria ser maior do que é. Um quinto criticismo diz respeito ao nível de granularidade das distinções de significado na PWN. Essas distinções são muito refinadas, o que faz com que as medidas de concordância entre anotadores sejam baixas.

A partir da WordNet de Princeton, várias *wordnets* foram propostas para diversas línguas, entre elas o português, conforme será descrito na seção a seguir.

9.1.1.1 *Wordnets* para o português

Vários recursos léxico-semânticos foram criados para o português nos últimos anos. Alguns deles são listados na página da Linguateca[11]. O NILC[12] tem uma coleção de recursos listados no portal PortLex[13], entre os quais se encontram, entre outros, VerbNet.Br (Scarton; Aluisio, 2012) e PropBank.Br (Duran; Aluísio, 2012).

Há várias versões de *wordnets* para o português, como Wordnet.BR (Dias-da-Silva, 2005), Onto.PT (Gonçalo Oliveira, 2014), PULO (Simões; Guinovart, 2014) e OpenWordNet-PT[14] (De Paiva et al., 2012). Essas *wordnets* são discutidas detalhadamente em (De Paiva et al., 2016; Gonçalo Oliveira, 2014), portanto, aqui simplesmente reiteramos a mensagem principal dessas comparações.

[10] https://wordnet.princeton.edu/related-projects
[11] https://www.linguateca.pt/
[12] https://sites.google.com/view/nilc-usp/
[13] http://143.107.183.175:21380/portlex/index.php/en/
[14] https://www.openwordnet-pt.org/

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Apesar de existirem várias alternativas de *wordnets* para o português, todas são menores e menos desenvolvidas do que a PWN. PWN é um recurso relativamente grande com 16MB, incluindo 155.327 palavras organizadas em 175.979 *synsets* num total de 207.016 pares de palavra-significado. A OpenWordNet-PT (OWN-PT) (De Paiva et al., 2012), alinhada à PWN, conta com 47.702 *synsets* (somente 27% da PWN), dos quais 32.855 correspondem a substantivos, 5.060 a verbos, 8.753 a adjetivos e 1.034 a advérbios. O número de projetos usando OWN-PT é muito limitado, possivelmente porque, construída de forma semi-automática, usando aprendizado de máquina no conjunto de wikipedias multilinguais (Melo; Weikum, 2009) e manualmente melhorando os dados obtidos.

Como exemplo, a Figura 9.3 apresenta os *synsets* da palavra "assassinar" na OWN-PT. Ao todo são quatro *synsets*, um na classe *Noun* (substantivo) e três na classe Verbo. O terceiro *synset* (02482425-v) refere-se a "matar intencionalmente e com premeditação" (glosa) e possui como hiperônimo direto o *synset* "matar" (01323958-v).

Figura 9.3: *Synsets* da palavra "assassinar" na OpenWordNet-PT (OWN-PT).

1. 02483000-v assassinate **| assassinar**
   - *(murder; especially of socially prominent persons; "Anwar Sadat was assassinated because many people did not like h*
2. 01325128-v dispatch **| assassinar**
   - *(kill without delay; "the traitor was dispatched by the conspirators")*
3. 02482425-v bump_off, slay, remove, off, polish_off, murder, hit, dispatch **| assassinar, matar, despachar**
   - *(kill intentionally and with premeditation; "The mafia boss ordered his enemies murdered")*
4. 00768701-n felony **| crime**
   - *(a serious crime (such as murder or arson))*

1

Outras *wordnets* são ainda menores (PULO (Simões; Guinovart, 2014)), ou menos acuradas, pois, construídas numa abordagem mais dinâmica (ONTO.PT (Gonçalo Oliveira, 2014)), podem mudar completamente de uma versão para a seguinte.

Algumas decisões de projeto de uma *wordnet*, assim como de outras bases de conhecimento, parecem claras e já são consenso na comunidade do PLN. *Wordnets* devem ser recursos abertos, grátis e fáceis de utilizar. Devem ter versões adequadas a usuários humanos e a agentes computacionais, isto é, devem ter interfaces de busca para usuários e interfaces ou bibliotecas para usos computacionais. Tais recursos linguísticos precisam ser mantidos e melhorados, pois nenhum é perfeito e as linguagens naturais são sistemas vivos, dinâmicos e em constante e contínua evolução.

Porém, outras decisões permanecem em aberto: uma alternativa só para o português brasileiro e outra para o português de Portugal? Ou uma alternativa para ambas variantes do português? Alternativas multilinguais tais como Open Multilingual WordNet (OMW)[15] (Bond; Foster, 2013) ou somente em português? Somente alternativas alinhadas a PWN ou o alinhamento[16] não é necessário? Somente as relações semânticas de PWN ou outras também? As entidades nomeadas devem ser incluídas no recurso ou não? Qual deve ser o registro do recurso? Deve incluir gírias e palavras de baixo-calão ou não?

[15] https://omwn.org/

[16] O alinhamento entre *wordnets* refere-se ao processo de mapeamento ou ligação de *synsets* entre diferentes *wordnets* de línguas distintas. Por exemplo, o alinhamento entre PWN e OWN-PT seria o processo de identificar que o *synset* em inglês para *"car"* é equivalente ao *synset* em português para "carro".

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

9.1.1.2 Análise do Exemplo Motivador usando a WordNet

Usaremos a OWN-PT para analisar o exemplo motivador Exemplo 9.1 apresentado no início da Seção 9.1. No Exemplo 9.2, foram sublinhadas algumas palavras que foram associadas a *synsets* na OWN-PT. Para realizar esta associação, é necessário definir o sentido ou significado da palavra usada no texto. Esta tarefa em PLN denominamos de Desambiguação do Sentido de Palavras (*Word Sense Desambiguation* – WSD). Após o exemplo, são listadas algumas afirmações (especificamente de **hiperomínia**) entre o *synset* da palavra usada no texto e outro *synset*.

Exemplo 9.2:

> Um assalto do tipo saidinha bancária, ocorrido na tarde desta terça-feira, terminou com uma mulher de 42 anos baleada pelos assaltantes. O roubo ocorreu na rua Professor Costa Mendes.

Usando a OWN-PT, definimos os seguintes *synsets* para as palavras sublinhadas em Exemplo 9.2: "assalto" (00783063-n), "terminar" (02610845-v). Não foi encontrado nenhum *synset* para o termo "balear". A seguir, algumas afirmações de hiperonímia entre esses *synsets* e outros:

- "roubo" (00781685-n) é hiperônimo de "assalto";
- "cessar" (02609764-v) é hiperônimo de "terminar";

Algumas dificuldades com a análise do exemplo, à luz da OWN-PT, foram:

- A desambiguação não é simples nem para um ser humano proficiente na linguagem natural (no caso, o português) e experiente com anotação de sentidos em *wordnets*. A diversidade e granularidade de *synsets* encontrados para uma palavra dificulta a definição do significado. Por exemplo, para a palavra "assalto" tem-se 07 (sete) *synsets* e todos parecem adequados para definir o sentido da palavra no Exemplo 9.2;
- Para algumas palavras, não foram encontrados *synsets* na OWN-PT (e.g. "baleada").

A partir da associação de uma palavra a um *synset*, um *parser* semântico pode, por exemplo, expandir o texto com tais informações semânticas, servindo como entrada para sistemas de entendimento de linguagem natural.

Como dito anteriormente, as bases não são sempre corretas e, definitivamente, não são completas. A língua muda, evolue o tempo todo e os significados das palavras seguem essa evolução. Nesse sentido, outros recursos léxico-semânticos são propostos e visam preencher lacunas na semântica das linguagens naturais. Na próxima subseção, detalharemos o recurso léxico-semântico FrameNet. Essa base se tornou relevante para a tarefa de Anotação de Papéis Semânticos (*Semantic Role Labeling* – SRL) pela abrangência e por incluir os papéis semânticos associados a substantivos e adjetivos.

9.1.2 FrameNet

FrameNet (Baker et al., 1998), da Universidade de Berkeley[17], é um recurso com conhecimento léxico e semântico baseado na semântica de *frames* (Fillmore et al., 1976) e na teoria de *frames* de (Minsky, 1975). Um *frame* é uma estrutura hierárquica conceitual

[17] https://framenet.icsi.berkeley.edu/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

que define uma situação, objeto ou evento por meio de seus participantes e relacionamentos. FrameNet faz parte da classe de recursos léxico-semânticos que suportam a tarefa de Anotação de Papéis Semânticos (*Semantic Role Labeling* - SRL), pois provê uma base de relações semânticas entre predicados e argumentos. Por exemplo, no evento de cometimento de crime, definido pelo *frame* `Commiting_crime`, são definidas as seguintes relações entre os verbos "cometer" ou "perpetrar" e os argumentos "criminoso", "crime", "explicação", "frequência", "instrumento", "maneira", dentre outros. Essas relações são denominadas de papéis semânticos, pois expressam funções que os diferentes constituintes de uma sentença desempenham em relação ao verbo ou predicado da sentença. FrameNet difere-se de outros recursos para SRL, como PropBank (Palmer et al., 2005) e VerbNet (Kipper et al., 2000), na medida em que associa papéis semânticos não somente a verbos, mas também a substantivos, a adjetivos, a advérbios, e até a proposições.

A Figura 9.4 apresenta um recorte da definição e componentes do *frame* `Commiting_crime`[18].

Figura 9.4: Descrição do *frame* `Commiting_crime` na FrameNet de Berkeley.

FrameNet Data

Committing_crime

Definition:

A Perpetrator (generally intentionally) commits a Crime, i.e. does something not permitted by the laws of society.
They PERPETRATED a felony by substituting a lie for negotiations.

The suspect had allegedly COMMITTED the crime to gain the attention of a female celebrity.

FEs:

Core:

Perpetrator [Perp] — The individual that commits a Crime.
Semantic Type: Sentient
How can he COMMIT treason against the King of England in a foreign country , if he is not English?

He PERPETRATED a crime against mother nature.

Core Unexpressed:

Crime [Cr] — An act, generally intentional, that has been formally forbidden by law.
How can he COMMIT treason against the King of England in a foreign country , if he is not English?

He PERPETRATED a crime against mother nature.

Como se pode observar na Figura 9.4, o *frame* é formado por vários componentes, descritos a seguir:

- **Elementos de *frames* (*Frame Elements* – FE)**, que definem os papéis semânticos envolvidos no *frame*. A sentença "*He committed the murder coldly and deliberately*" (em português, "Ele cometeu o assassinato fria e deliberadamente") evoca o *frame* `Commiting_crime` através do verbo *"commit"*, cujos argumentos *"He"*, *"the murder"* e *"coldly and deliberately"* expressam os seguintes papéis semânticos: *"perpetrator"*, *"crime"* e *"manner"*, conforme abaixo:

[18] A descrição completa do *frame* está disponível em https://framenet.icsi.berkeley.edu/frameIndex.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

  - *[PERPETRATOR **He**] committed [CRIME **the murder**] [MANNER **coldly and deliberately**.]*, onde:
    * *PERPETRATOR* – o indivíduo que cometeu um crime;
    * *CRIME* – um ato, geralmente intencional, que é formalmente proibido pela lei;
    * *MANNER* – uma descrição da forma e dos efeitos secundários do crime, assim como descrições gerais comparando eventos, podendo também indicar características salientes do criminoso que afetam a ação (presunçosamente, friamente, deliberadamente, ansiosamente, cuidadosamente).

- **Unidades Lexicais (*Lexical Unit* – LU)**, que são as palavras relacionadas no *frame*. Cada palavra polissêmica[19] com significados distintos pertence a um *frame* diferente. Por exemplo, a palavra *"commit"* possui quatro entradas (sentidos) no léxico, conforme os quatro *frames* dos quais participa: `Imposing_Obligation`, `Institutionalization`, `Commitment`, `Committing_crime`. O léxico da FrameNet contém, para cada unidade léxica, além do termo e de uma definição em linguagem natural, as realizações sintáticas possíveis dos elementos de *frames* relacionados à unidade léxica. Por exemplo, na sentença "*He committed the murder coldly and deliberately*" (em português, "Ele cometeu o assassinato fria e deliberadamente"), o elemento de *frame* ou papel semântico "crime" (*the murder*) possui a realização sintática de sintagma nominal (NP – *Noun Phrase*);
- **Entradas Lexicais (*Lexical Entry* – LE)**, que são unidades lexicais evocadoras de *frame*, ou seja, que chamam ou ativam *frames*. No *frame* `Commiting_crime`, as entradas lexicais são os verbos `commit.v` e `perpetrate.v` e os substantivos `commission.n` e `crime.n`. As entradas lexicais mais comuns são verbos, porém alguns *frames* são ativados por substantivos e adjetivos. Por exemplo, a sentença *"... the reduction of debt levels to $665 million from $2.6 billion."* (em português, "... a redução dos níveis de dívida para 665 milhões de dólares, de 2,6 mil milhões de dólares.") tem-se um exemplo de uso do *frame* `Cause_change_of_scalar_position`, evocado pelo substantivo *"reduction"*;
- ***Corpus* da FrameNet**, um conjunto de sentenças anotadas que exemplificam os componentes da FrameNet. O *corpus* da FrameNet é uma parte crucial, pois representa um recurso valioso para o desenvolvimento e teste de sistemas de PLN que requerem uma compreensão da semântica dos textos, especialmente, dos papéis envolvidos no evento. O conjunto total dos textos anotados contém, atualmente, 202.978 textos, divididos em:
  - Conjunto de Anotações Completas (*Full Text Annotation Sets*), que contém 28.446 anotações semânticas detalhadas para textos inteiros, e não apenas para frases isoladas. Essas anotações incluem informações sobre os elementos do *frame* (papéis semânticos da FrameNet), as unidades e entradas lexicais (léxico da FrameNet), e suas realizações sintáticas;
  - Conjunto de Anotações Lexicográficas (*Lexicographic Annotation Sets*), que contém 174.532 anotações para as palavras em uma língua, incluindo informações sobre os sentidos das palavras, os *frames* que esses sentidos evocam e os papéis semânticos (elementos de *frame*) associados a cada sentido. Em (Ruppe-

[19] Uma palavra polissêmica é uma palavra que possui vários significados ou sentidos relacionados entre si, dependendo do contexto em que é usada. Por exemplo, a palavra "banco" pode se referir a uma instituição financeira, um local para sentar, ou a uma elevação de areia no mar.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

nhofer et al., 2006, pp. 67-88), tem-se o detalhamento das camadas de anotação do *corpus* da FrameNet;

- **Tipos semânticos**, que são associados às unidades lexicais, aos elementos do *frame* ou ao *frame* como um todo. Em (Ruppenhofer et al., 2006, pp. 111-120), tem-se a definição destes marcadores semânticos. Por exemplo, o elemento de *frame* `Perpetrator` do *frame* `Committing_Crime` é marcado como sendo do tipo `sentient` (que percebe pelos sentidos, que recebe impressões).

Além da definição individual de cada *frame*, a FrameNet possui relações semânticas entre *frames*, denominadas relações *frame-to-frame*. Alguns exemplos são: `Inherits_from` (herda de), `Is_Inherited_by` (é herdado por), `Is_Used_by`*‘ (é usado por). Em (Ruppenhofer et al., 2006, pp. 104-111), tem-se a descrição das relações* frame-to-frame* suportadas pela FrameNet.

Atualmente, a FrameNet contém 1224 *frames*, 10.478 elementos de *frames* (papéis semânticos), e 13.687 unidades lexicais[20].

FrameNet fornece uma nova perspectiva para um recurso léxico-semântico. O significado de palavras ou unidades lexicais é dado no contexto das situações em que podem participar (*frames*), por meio dos papéis que podem assumir. FrameNet não poderia substituir completamente a WordNet porque falta à primeira muitas das relações semânticas úteis como meronímia e hiperonímia. Embora haja uma interseção entre essas bases, elas se distinguem em boa parte. Enquanto a WordNet foca em relações entre *synsets* organizando uma hierarquia e taxonomia do mundo, a FrameNet foca nas relações que ocorrem em eventos.

Alguns projetos visam relacionar as entradas lexicais dessas duas bases. É o caso do projeto SemLink[21], cujo objetivo é vincular diferentes recursos léxico-semânticos por meio de um conjunto de mapeamentos. Estes mapeamentos permitirão combinar as diferentes informações fornecidas por esses diferentes recursos lexicais para tarefas como inferência em linguagem natural (*Natural Language Inference* – NLI). Os recursos mapeados pelo SemLink são WordNet, FrameNet, VerbNet e PropBank.

9.1.2.1 Framenets para o português

FrameNet Brasil (FN-Br)[22] (Salomão, 2009), iniciativa de pesquisa lexicográfica, em desenvolvimento na Universidade Federal de Juiz de Fora (UFJF) desde 2008, tem o objetivo de construir e evoluir, para o português, a contraparte linguística da rede semântica original FrameNet. Atualmente, a base da FN-Br é a base mais robusta e representativa do paradigma da Semântica de *Frames* para o português. Foi construída através da tradução automática dos *frames* existentes na FrameNet original, e posterior adaptação para o português brasileiro. Este processo de adaptação envolveu traduzir e ajustar a descrição e os elementos dos *frames* para garantir que eles sejam relevantes e aplicáveis ao contexto brasileiro. Além da adaptação dos *frames* originais da FrameNet, no âmbito de alguns projetos, como o COPA 2014 (Torrent et al., 2014) e FLAME[23], relativos aos domínios de

[20] Mais detalhes sobre os números atuais da FrameNet podem ser acessados em https://framenet.icsi.berkeley.edu/current_status.

[21] https://verbs.colorado.edu/semlink/. A versão atual do SemLink é a versão 2.0 e pode ser acessada pelo GitHub https://github.com/cu-clear/semlink.

[22] https://www2.ufjf.br/framenetbr/

[23] https://www2.ufjf.br/framenetbr/projetos/

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

esporte e turismo, respectivamente, foram criados novos *frames* para representar conceitos e situações específicos da cultura e do português brasileiro. O *corpus* FN-Br é constituído pela combinação de mais de 16 *corpora*, todos caracterizados por permitir acesso público e que representam usos do português europeu e do português brasileiro. Em 2009, de acordo com (Salomão, 2009), os *corpora* totalizavam pouco mais de 280 milhões de palavras.

A Figura 9.5 apresenta um recorte da definição e componentes do *frame* `Cometer_crime` da FN-Br[24], adaptado do *frame* `Commiting_crime` da FrameNet de Berkeley (vide Figura 9.4).

Figura 9.5: Descrição do *frame* `Cometer_crime` na FrameNet Brasil (FN-Br).

Cometer_crime

Definição

Um Criminoso (geralmente intencionalmente) comete um Crime, ou seja, faz algo que não é permitido pelas leis da sociedade.

Elementos de Frame Nucleares

FE Core:

Criminoso **semantic_type:** @sentient	O indivíduo que comete o Crime.

FE Core-Unexpressed:

Crime	Um ato cometido, geralmente intencional, que é formalmente proibido por lei.

Elementos de Frame Não-Nucleares

Explicação **semantic_type:** @state_of_affairs	Um estado de coisas ao qual o Criminoso está respondendo em seu Crime.
Finalidade **semantic_type:** @state_of_affairs	A ação que o Criminoso estava tentando realizar no Crime.
Frequência	A frequência com a qual um Crime é cometido.

Outras iniciativas culminaram na geração de bases de *frames* em português, todas de menor tamanho que a FN-Br e para domínios ainda mais específicos.

A base FrameFOR (Barreira et al., 2017) é uma base com 113 *frames* em português brasileiro, adaptados da FrameNet original, contendo os papéis semânticos, unidades e entradas lexicais relacionados aos tipos de crimes mais investigados na Perícia Forense do Estado do Ceará, no Brasil (PEFOCE) – formação de quadrilha, tráfico de drogas, sequestro, corrupção, receptação, contrabando, pedofilia, estupro, agressão, tortura, falsificação, ameaça, porte ilegal de arma, estelionato, e extorsão.

O estudo de (Bertoldi, 2011) analisou os limites da criação automática de léxicos computacionais segundo o paradigma FrameNet, comparando as unidades lexicais evocadoras, os papéis semânticos e a estrutura do *frame* `Criminal_process`, em inglês e português. Esse estudo contrastivo mostrou que os *frames* do domínio jurídico são socialmente orientados e que a criação automática de léxicos em áreas cultural e socialmente orientadas tende a apresentar divergências. Em (Bick, 2009) tem-se a proposta de PFN-PT, um sistema para

[24]Ver descrição completa em https://webtool.framenetbr.ufjf.br/index.php/webtool/report/frame/main

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

a anotação semântica automática do português, consistindo numa nova framenet contendo cerca de 13.000 padrões sintáticos, cobrindo 7.300 lemas verbais com 10.700 sentidos.

Todos estes projetos, ainda que de menor porte, possuem relatos de sucesso em aplicações de PLN como extração de informação, anotação de papéis semânticos, reconhecimento de entidades nomeadas, evidenciando a importância de abordar as peculiaridades linguísticas com perspectivas contextualizadas e culturalmente relevantes.

9.1.2.2 Análise do Exemplo Motivador usando a FrameNet

Nesta seção, usaremos a FrameNet de Berkeley para analisar o exemplo motivador definido no início da Seção 9.1. No Exemplo 9.3 são destacados o *frame* associado, as unidades lexicais (elemento evocador) que evocaram o *frame* e os papéis semânticos identificados no texto.

Exemplo 9.3:

> Um assalto do tipo saidinha bancária, ocorrido na tarde desta terça-feira, terminou com uma mulher de 42 anos baleada pelos assaltantes. O roubo ocorreu na rua Professor Costa Mendes.

- *FRAME*: *Robbery* (definição do *frame*: situação em que um perpetrador prejudica uma vítima tirando algo (bens) dela … O assalto pode ser feito de uma maneira específica (por exemplo, à força) e através de um meio específico (por exemplo, ameaçando a vítima).)
- UNIDADE LÉXICA: *robbery* ("assalto" e "roubo", em português)
- PAPEL SEMÂNTICO *place*: "rua Professor Costa Mendes"
- PAPEL SEMÂNTICO *time*: "na tarde desta terça-feira"
- PAPEL SEMÂNTICO *perpetrator*: "pelos assaltantes"

Uma dificuldade com a análise desse exemplo, à luz da FrameNet, foi na identificação do papel semântico "vítima" ("uma mulher de 42 anos"), pois o complemento da sentença que a contém "…com uma mulher de 42 anos baleada …" possui a estrutura sintática Prep.Det.N (preposição + determinante + substantivo) e não é compatível com nenhuma realização sintática do elemento de *frame* `victim`.

9.1.3 ConceptNet

ConceptNet[25] (Speer et al., 2016) é uma base de conhecimento de senso comum que expressa relações rotuladas e ponderadas entre palavras ou fragmentos de textos em linguagem natural, através de um Grafo de Conhecimento (*Knowledge Graph*) contendo *edges* ou afirmações. Alguns exemplos de afirmações expressas na ConceptNet são:

- Uma rede é usada para pescar peixe (*A net is used for catching fish.*);
- "Folhas" é uma forma da palavra "folha" (*"Leaves" is a form of the word "leaf"*);
- A palavra *"cold"* em inglês é *"studyeny"* em tcheco (*The word "cold" in English is "studeny" in Czech*);
- O alimento é usado para comer (*Food is used for eating*);
- Bicicleta é usada para chegar a algum lugar rápido (*Bicycle is used for getting somewhere fast*);

[25] https://conceptnet.io/

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

- Cozinhar é motivada por você está com fome (*Cook is motivaded by being hungry*).

Sua versão original (Havasi et al., 2007; Liu; Singh, 2004) foi criada pela equipe do MediaLab do *Massachusetts Institute of Technology (MIT)* em 1999, a partir de conhecimentos extraídos do projeto de construção coletiva (*crowdsourcing*) Open Mind Common Sense (OMCS) (Singh et al., 2002). O OMCS surgiu com o objetivo de coletar, pela Internet e de colaboradores voluntários, sentenças que expressavam fatos da vida comum. Por exemplo, a sentença "*The Effect of [falling off a bike] is [you get hurt]*" foi coletada de voluntários, quando solicitados a preencher os espaços do *template* "*The Effect of [….] is [….]*". A alternativa adotada pela equipe da ConceptNet foi construir a rede semântica (nós conceituais interligados pelas relações semânticas), a partir de um processo automático sobre o *corpus* OMCS, o qual extraiu as relações semânticas e seus argumentos.

A motivação do projeto que mantém a ConceptNet é expressar os fatos que as pessoas sabem comumente sobre o mundo conhecimento de senso comum através de afirmações que relacionam conceitos. Este tipo de conhecimento é importante porque, quando as pessoas se comunicam, seus proferimentos acontecem sobre suposições implícitas e básicas, as quais suportam e explicam boa parte dos raciocínios necessários para um bom nível de entendimento e, consequentemente, uma boa comunicação. Por exemplo, quando alguém fala "Eu comprei doces", está implícito que usou dinheiro, ou quando fala "Fui a um casamento", provavelmente tinha uma noiva, um noivo, uma festa com bolo e champagne, e o interlocutor está autorizado a perguntar "A noiva estava bonita?" etc.

Atualmente, a ConceptNet[26] evoluiu como um projeto colaborativo com diversas fontes:

- *Open Mind Common Sense* (OMCS) (Singh et al., 2002) e projetos irmãos em outras línguas (Anacleto et al., 2006);
- Informações extraídas da análise do Wikcionário[27], em vários idiomas, com um analisador personalizado ("Wikiparsec");
- *"Games with a Purpose"*, que são jogos projetados para coletar conhecimento comum (Ahn et al., 2006; Kuo et al., 2009);
- *Open Multilingual WordNet* (Bond; Foster, 2013), uma representação de dados vinculados a WordNet de Princeton e seus projetos paralelos em vários idiomas;
- JMDict (Breen, 2004), um dicionário japonês multilíngue;
- OpenCyc, uma hierarquia de hiperônimos fornecida pelo Cyc (Lenat; Guha, 1989), um sistema que representa o conhecimento do senso comum na lógica de predicados;
- Um subconjunto de DBPedia (Auer et al., 2007), uma rede de fatos extraídos de *infoboxes* da Wikipédia.

A unidade de conhecimento da ConceptNet é uma afirmação ou *edge*[28] que é uma relação particular entre termos ou frases em uma linguagem natural, de uma fonte específica. Sucintamente, cada *edge* é uma tripla com um primeiro argumento (nó inicial), um rótulo da relação e um segundo argumento (nó final). Por exemplo, a afirmação *"Bicycle is used to get somewhere fast"* pode ser expressa como (*Bicycle*, *is used to*, *get somewhere fast*). Cada *edge* é representada em uma estrutura de dados com os seguintes atributos:

- *URI* – identificador único para a afirmação que está sendo expressa;

[26] Sua última versão é a ConceptNet 5.8 e a documentação completa está disponível em https://github.com/commonsense/conceptnet5/wiki

[27] Disponível em https://pt.wiktionary.org/wiki/Wikcion%C3%A1rio:P%C3%A1gina_principal

[28] https://github.com/commonsense/conceptnet5/wiki/Edges

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

- *REL* – o URI da relação expressa no *edge*. Atualmente, existem 34 relações em *edges* da ConceptNet 5 - *RelatedTo*, *IsA*, *is Used For*; *Motivated by*, *Desires* etc.[29];
- *START* – o URI do primeiro argumento da afirmação;
- *END* – o URI do segundo argumento da afirmação;
- *WEIGHT* – a força da afirmação. Um peso típico é 1, mas pode ser maior a depender do número de vezes que a afirmação foi recuperada ou generalizada a partir das fontes;
- *SOURCES* – as fontes que, quando combinadas, dizem que esta afirmação deveria ser verdadeira;
- *LICENSE* – o URI *Creative Commons* para a licença que rege esses dados;
- *DATASET* – o URI que representa o conjunto de dados de uma fonte específica que criou a afirmação;
- *SURFACE TEXT*[30] – o texto original em linguagem natural que expressou esta afirmação. Os conceitos do início e fim serão marcadas entre colchetes duplos. Um exemplo é *"[[Bicycle]] is used for [[get somewhere fast]]"*.

A ConceptNet contém mais de 21 milhões de *edges* e quase 10 milhões de nós com palavras ou fragmentos de textos. A base cobre em torno de 78 linguagens naturais[31] com pelo menos 10.000 entradas no vocabulário. As 10 (dez) principais linguagens são: inglês, francês, italiano, alemão, espanhol, russo, português, japonês, holandês e chinês. A base de afirmações na linguagem inglesa possui um vocabulário com 1,8 milhão de nós, e no português contém um vocabulário com 473 mil nós.

O *framework* computacional da ConceptNet contém uma hierarquia de URIs que identificam os principais componentes dessa base de conhecimento: afirmações (ou *edges*), termos (palavras ou frases em uma linguagem particular), relações (por exemplo, *IsA*), *datasets*, fontes de dados. Também possui uma API REST[32] pela qual você pode obter os componentes no formato JSON. Cada *edge*, termo, relação, *dataset* e fonte da ConceptNet possui uma URI que os identificam e sua definição completa em JSON pode ser acessada via API. A Figura 9.6 apresenta um recorte da definição do termo *"murder"*, contendo a lista de *edges* a partir desse termo.

Ao acessar, via API, a definição deste termo, tem-se acesso à sua definição em JSON, conforme ilustrado na Figura 9.7, com a URI deste conceito (*"/c/en/murder"*) e a lista de *edges*.

ConceptNet se tornou um *hub* de conteúdo semântico, pois provê *link* para outras bases de dados, com um *toolkit* e uma API que suportam inferências práticas de senso comum sobre textos, tais como descoberta de contexto (que habilita a extração da vizinhança contextual de um conceito, e.g., "tirar a roupa", "ir dormir", e "deitar-se" são vizinhos do conceito "ir para a cama"), cadeia de inferências (que habilita encontrar caminhos na rede semântica a partir de um conceito, e.g., "comprar comida" - "ter comida" - "comer comida" - "sentir-se cheio" - "sentir-se com sono") e analogia conceitual (que envolve encontrar conceitos que são estruturalmente similares, e.g., "funeral" e "casamento", "sofá" e "cama"). Todos esses exemplos foram extraídos de (Ovchinnikova, 2012).

[29] A lista completa das relações expressas na ConceptNet está disponível em https://github.com/commonsense/conceptnet5/wiki/Relations

[30] Pode ser nulo porque nem todas as afirmações foram derivadas de entrada em linguagem natural

[31] As bases de afirmações da ConceptNet para todas as linguagens naturais suportadas pode ser consultada em https://github.com/commonsense/conceptnet5/wiki/Languages

[32] Acessível em api.conceptnet.io. A documentação completa da API da Conceptnet pode ser acessada em https://github.com/commonsense/conceptnet5/wiki/API

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

Figura 9.6: Descrição do termo *"murder"* na ConceptNet 5.8.



9.1.3.1 Bases de conhecimento de senso comum para o português

Como visto, a Conceptnet 5.8 possui uma cobertura de 473 mil termos ou frases no português, representando assim a mais extensa base de conhecimento de senso comum para essa língua.

O projeto *Open Mind Common Sense* – Brasil (OMCS-Br) foi um projeto do Laboratório de Interação Avançada (LIA) da Universidade Federal de São Carlos – UFSCar, em colaboração com o MediaLab do MIT, para a coleta de conhecimento de senso comum em português (Anacleto et al., 2006). Este projeto em 2010 contava com 160.000 afirmações de senso comum de seus colaboradores. O projeto foi descontinuado, mas diversas aplicações e estudos foram desenvolvidos a partir desta base. Dentre eles, podemos citar, uma ferramenta que utiliza a base de conhecimento de senso comum para auxiliar a interação humana (de alunos e professores) com ferramentas educacionais (Anacleto et al., 2007).

A base InferenceNet-BR (Pinheiro et al., 2010) adapta a ConceptNet (Liu; Singh, 2004) adicionando uma camada que define o papel da afirmação em uma inferência – se como premissa (ou pré-condição) ou como conclusão (ou pós-condição). Além da tradução dos termos e suas afirmações (relações com outros conceitos), o projeto da InferenceNet-BR evoluiu a base com novo conhecimento semântico específico para o domínio de segurança pública em português.

A InferenceNet-BR compõe-se de duas bases de conhecimento:

1. **Base Conceitual** – essa base contém o conjunto de termos (palavras ou frases em linguagem natural) relacionados em uma rede semântica, representada por meio de quádruplas (ARG1, REL, ARG2, PESO, TIPO_INF) que definem as afirmações ou *edges*, onde:

    - ARG1 – identificador do termo inicial da relação;
    - ARG2 – identificador do termo final da relação;
    - REL – identificador da relação semântica de um total de 17 relações, por exemplo, "CapazDe";"PartDe"; "ÉUm"; "EfeitoDe" etc.;

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Figura 9.7: Definição em JSON do termo *"murder"* na ConceptNet 5.8.


```
"@id": "/c/en/murder",
  "edges": [
    {
      "@id": "/a/[/r/IsA/,/c/en/murder/,/c/en/crime/]",
      "@type": "Edge",
      "dataset": "/d/conceptnet/4/en",
      "end": {
        "@id": "/c/en/crime",
        "@type": "Node",
        "label": "a crime",
        "language": "en",
        "term": "/c/en/crime"
      },
      "license": "cc:by/4.0",
      "rel": {
        "@id": "/r/IsA",
        "@type": "Relation",
        "label": "IsA"
      },
      "sources": [
        {
          "@id":
"/and/[/s/activity/omcs/omcs1_possibly_free_text/,/s/contributor/omcs/highplacespam1/]",
          "@type": "Source",
          "activity": "/s/activity/omcs/omcs1_possibly_free_text",
          "contributor": "/s/contributor/omcs/highplacespam1"
        },
        ...
          ...
     ],
      "start": {
        "@id": "/c/en/murder",
        "@type": "Node",
        "label": "a murder",
        "language": "en",
        "term": "/c/en/murder"
      },
      "surfaceText": "[[a murder]] is [[a crime]]",
      "weight": 8.0
    },
```


- PESO – força da afirmação. Um peso típico é 1, mas pode ser maior a depender do número de vezes que a afirmação foi recuperada ou generalizada a partir das fontes (Conceptnet original e *corpus* de textos de domínio);
- TIPO_INF – tipo da relação inferencial - premissa (pré-condição - PRE) ou conclusão (pós-condição - POS).

2. **Base de Sentenças-Padrão** - essa base contém a estrutura sintática das sentenças-padrão e suas relações com termos da Base Conceitual. Seja a sentença-padrão $sp_1 = X$ "ser_assassinado_por" Y. Temos que X está relacionado com o termo "vítima" e Y está relacionado com o termo "criminoso", através da relação "ÉUm", e com o tipo inferencial **pós-condição**. Ou seja, na sentença "Maria foi assassinada por seu amante" podemos concluir que Maria é a vítima e seu amante é o criminoso. A rede semântica dessa base de sentenças-padrão é representada, portanto, por meio de quádruplas (SP, REL, ARG, PESO, TIPO_INF) que definem as afirmações ou *edges*, onde:

   - SP – identificador da sentença-padrão geralmente da forma "(X, sintagma verbal, Y)";
   - ARG – identificador do termo final da relação;

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

- REL – identificador da relação semântica de um total de 17 relações, por exemplo, “CapazDe”;“PartDe”; “ÉUm”; “EfeitoDe” etc.;
- PESO – força da afirmação. Um peso típico é 1, mas pode ser maior a depender do número de vezes que a afirmação foi recuperada ou generalizada a partir das fontes (Conceptnet original e *corpus* de textos de domínio);
- TIPO INF – tipo da relação inferencial - premissa (pré-condição - PRE) ou conclusão (pós-condição - POS).

A Figura 9.8 ilustra uma parte da rede de relacionamento inferencial da palavra “crime” na base InferenceNet-BR, com as seguintes relações:

- (capazDe, “crime”,“ter vítima”, “Pre”);
- (capazDeReceberAcao, “crime”,“evitar por polícia”, “Pre”);
- (primeiroSubEventoDe, “crime”,“escolher a vítima”, “Pre”);
- (usadoPara, “crime”,“vingança”, “Pre”);
- (capazDeReceberAcao, “crime”,“cometer com arma”, “Pre”);
- (capazDe, “crime”,“envolver violência”, “Pre”);
- (éUm, “crime”,“violação da lei”, “Pre”);
- (motivacaoDe, “crime”,“vingança”, “Pre”);
- (efeitoDe, “crime”,“culpa”, “Pos”);
- (efeitoDesejavelDe, “crime”,“julgamento”, “Pos”).

9.1.3.2 Análise do Exemplo Motivador usando a ConceptNet

Nesta seção, usaremos a base da ConceptNet para o português e para o inglês para analisar o Exemplo 9.1. No Exemplo 9.4 são destacadas algumas afirmações de senso comum associadas aos termos mencionados no texto (termos sublinhados).

Exemplo 9.4:

> Um assalto do tipo saidinha bancária, ocorrido na tarde desta terça-feira, terminou com uma mulher de 42 anos baleada pelos assaltantes. O roubo ocorreu na rua Professor Costa Mendes.

Afirmações (*edges*) do termo *assalto*:

- “terror” está localizada em “assalto” (da base para o português);
- “pessoas com medo” está localizada “assalto” (da base para o português);
- *“heist” is a type of “robbery”* (roubo é um tipo de assalto);
- *“revolver” is a thing used for “robbery”* (revólver é uma coisa usada para assaltar)

Afirmações (*edges*) do termo *balear* (*“shoot”*, em inglês):

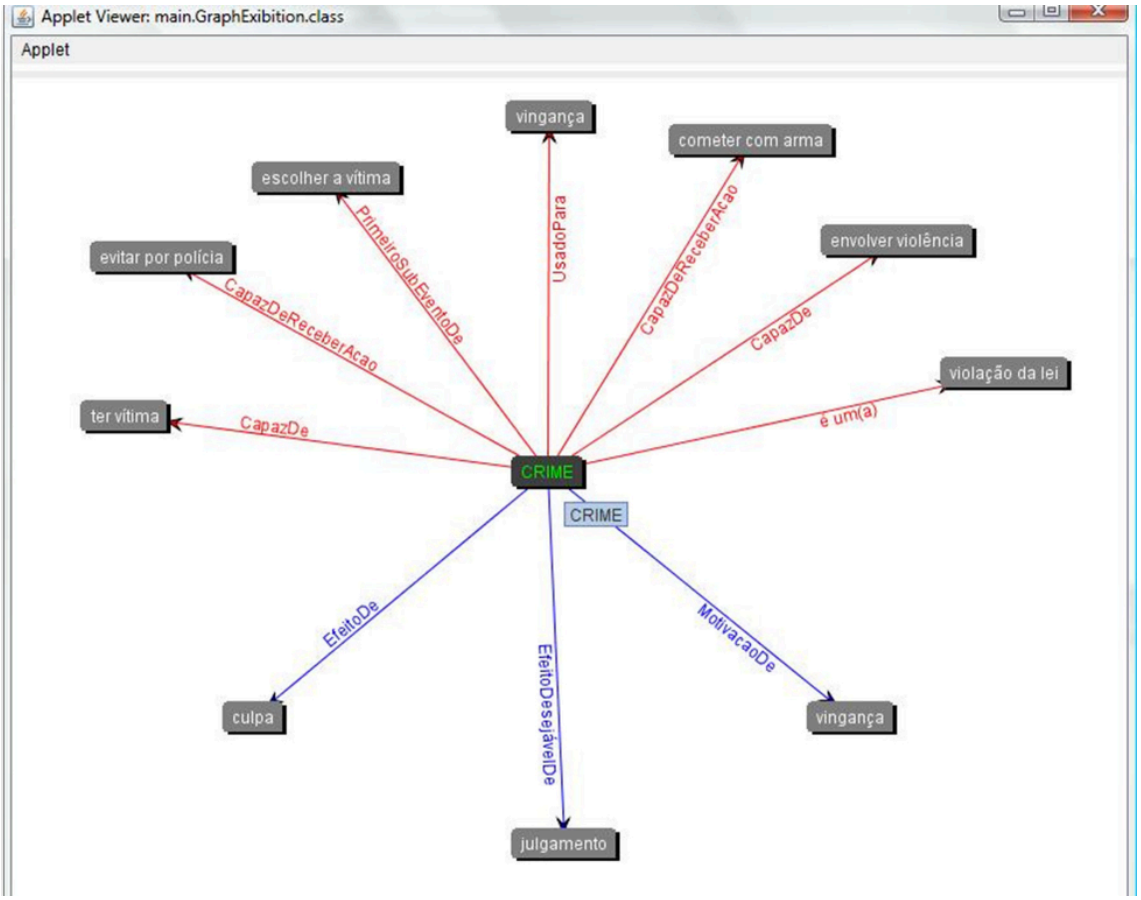
- “fight the enemy” is subevent of “shoot” (lutar contra o inimigo é subevento de balear);
- “a gun” is a thing used for “shoot” (uma arma é uma coisa usada para balear);
- *“shoot” is a way of “kill”* (balear é uma forma de matar);
- *“shoot” is a type of “sprout”* (é um tipo de broto, como broto de feijão)

Algumas dificuldades com a análise do exemplo, à luz da ConceptNet, foram:

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Figura 9.8: Rede de relações inferenciais do termo "crime" (pré-condições (PRE) são as arestas na cor vermelha, e pós-condições (POS) são as arestas na cor azul).



- A desambiguação do sentido da palavra "balear" no exemplo Exemplo 9.4 não é trivial. Na base da ConceptNet não há separação das redes semânticas por classe gramatical e de palavras polissêmicas. Por exemplo, as afirmações de "balear" no sentido de atirar e em outros sentidos estão associadas a um mesmo termo, assim como afirmações de "shoot" como verbo e como substantivo;
- As afirmações em português são incipientes na base da ConceptNet, pelo menos para os termos analisados nesse exemplo.

9.2 Considerações finais

Ao longo deste capítulo exploramos três bases de conhecimento comuns na área de PLN: WordNet, FrameNet e ConceptNet. WordNet, com sua rica estrutura hierárquica, destaca-se por mapear relações semânticas entre conjuntos de sinônimos (*synsets*), fornecendo uma compreensão sobre sinônimos, antônimos, hiperônimos e muito mais. Esse recurso se tornou uma das ferramentas mais utilizadas em aplicações de PLN, desde a análise de sentimentos até a desambiguação de sentidos.

FrameNet, por sua vez, adota uma abordagem baseada em *frames* para capturar significados em dados contextos ou situações, fazendo a ponte entre elementos lexicais e seus respectivos papéis semânticos. Através dessa base, podemos entender mais profundamente

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

como as palavras interagem dentro de estruturas semânticas e pragmáticas mais amplas. Esta abordagem, focada nos papéis semânticos, permite uma análise mais rica do texto, tornando-a particularmente útil para tarefas de anotação de papéis semânticos e análise de discurso. Em contrapartida, ela não possui um critério claro de completamento, ou seja, não sabemos quando vamos ter (ou se já temos) todos os *frames* necessários.

Por fim, a base ConceptNet destaca-se pelo seu caráter colaborativo e multidimensional. Integra conhecimentos de senso comum e conhecimento léxico-semântico de várias fontes e idiomas, oferecendo uma visão ampla das relações entre conceitos e contextos. Seu formato de rede semântica ajuda a capturar a complexidade e interconexão do conhecimento humano de uma maneira holística. ConceptNet tem sido usada em aplicações de extração de informação e reconhecimento de implicação textual. Em contrapartida, problemas de consistência da base parecem importantes. E tal como no caso de FrameNet, não temos um critério explícito de quando teremos uma cobertura suficiente.

Em resumo, cada uma destas bases de conhecimento representa um recurso para o PLN com perspectivas únicas, mas qual delas se aplica melhor e em quais casos? Recursos léxicos-semânticos parecem não serem suficientes para expressar conhecimento de mundo. De outro lado, bases de conhecimento de senso comum são mais flexíveis e expressivas, porém menos formais. Acreditamos que uma abordagem híbrida, em que tais bases de conhecimento sejam usadas de forma combinada, é mais promissora para o PLN. Nas próximas versões deste capítulo introduziremos os *parsers* semânticos que, em conjunto com as bases de conhecimento, constituem poderoso *framework* para sistemas de entendimento de linguagem natural.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Capítulo 10

Representação vetorial e semântica distribucional

Eloize Rossi Marques Seno
Daniela Claro
Laila Mota

Publicado em: 16/04/2026

Este capítulo é uma versão renovada do capítulo de Semântica Distribucional que existia nas edições anteriores deste livro.

A representação semântica pode ser ampliada com o intuito de melhor compreender o significado de uma palavra ou símbolo. No Capítulo 9, vimos a representação semântica por meio de técnicas simbólicas, unindo os elementos lexicais aos seus papéis semânticos através de WordNets, FrameNets e ConceptNets. Ao longo deste capítulo, as unidades lexicais são posicionadas em relação à sua representação numérica, transformando-as em vetores compreensíveis pelas máquinas. Essas representações podem ser obtidas por meio de duas abordagens principais: o modelo ***bag-of-words*** (BoW), que se baseia na contagem global de termos e a **semântica distribucional**, baseada em contexto.

Diferente dos humanos, que reconhecem o sentido das palavras e o significado do texto de forma intuitiva, os algoritmos computacionais requerem uma representação numérica para realizar suas operações. O BoW é um modelo de representação simplificado que transforma um texto (como uma frase ou documento) em um vetor, ignorando completamente a ordem e a estrutura gramatical das palavras, focando apenas em sua frequência ou ocorrência. O modelo inicialmente constrói um vocabulário com todas as palavras únicas do *corpus* e, em seguida, representa cada texto como um **vetor esparso** do tamanho desse vocabulário, onde cada posição no vetor indica a presença (ou a frequência) de uma palavra específica. Um exemplo clássico consiste no uso da medida estatística TF-IDF para ponderar a importância de palavras em uma coleção de documentos. Essa medida é amplamente adotada em tarefas que envolvem a comparação de similaridade entre textos, como a detecção de plágio, a inferência textual e a recuperação de informações. Com TF-IDF, o significado de uma palavra é representado por meio de uma função simples calculada com base na frequência da palavra na coleção de documentos (vide (Seção 10.2.1). Como muitas palavras nunca ocorrem em alguns documentos, frequentemente, esse modelo leva a vetores muito grandes e esparsos, ou seja, com muitos zeros.

Apesar da simplicidade, o BoW trata uma palavra como tendo um único significado, independentemente do contexto onde ela ocorre. Isso acontece porque nesse modelo as palavras (ou termos) são tratadas de forma isolada, ignorando que o significado emerge das relações entre elas. A semântica distribucional supera essa limitação ao representar palavras por seus contextos de ocorrência.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

A semântica distribucional tem sido a principal abordagem de representação do significado lexical em diversas tarefas de processamento de língua natural. Nessa abordagem, as unidades lexicais são representadas por vetores de valores reais que codificam o significado a partir de sua distribuição em textos. Ela é ancorada na **Hipótese Distribucional** (Firth, 1957; Harris, 1954) que preconiza que palavras que têm um contexto linguístico semelhante tendem a ter significado similar ou aproximado. É o caso, por exemplo, de palavras como "ensino" e "educação" que costumam aparecer no mesmo contexto de palavras como "aluno", "escola" e "professor", sugerindo que existe uma similaridade entre as duas palavras em certos contextos. Vejamos, por exemplo, a ocorrência dessas palavras nas sentenças a seguir[1]:

1. MEC deve começar a ouvir alunos sobre novo ensino médio em 8 de maio.
2. Governo prorroga inscrições para concurso público de professores na rede estadual de ensino.
3. Alunos e profissionais da educação terão aulas de comportamento seguro.
4. Aluno é colocado em ensino remoto após intimidar e tentar derrubar professor em SP.

Assim, na semântica distribucional as palavras são caracterizadas pelo contexto em que elas aparecem. Por se basearem em distribuição, os vetores podem ser aprendidos automaticamente a partir de textos, sem que haja supervisão de um humano (utilizando-se textos não rotulados, portanto). Os modelos que aprendem esse tipo de representação são denominados de **Modelos Semânticos Distribucionais - MSD** (*Distributional Semantic Models – DSM*, no inglês), cujas representações são, frequentemente, denominadas de **vetores densos**, contrapondo os **vetores esparsos** representados por modelos *bag-of-words*.

Os vetores densos são conhecidos como **vetores semânticos** ou ***embeddings*** e podem ser calculados usando modelos ou algoritmos como Word2Vec, Glove e FastText, que serão abordados na Seção 10.3. Os *embeddings* produzidos por esses modelos são ditos **estáticos**, pois cada palavra recebe um único vetor fixo, independentemente do contexto em que aparece. Por outro lado, modelos baseados em redes neurais profundas (*deep learning*, no inglês), como ELMo, BERT e GPT, geram representações vetoriais de palavras onde o vetor de uma mesma palavra muda conforme o contexto em que ela aparece. Diferentemente dos *embeddings* estáticos, nesses modelos o significado é construído dinamicamente a partir das palavras que fazem parte do contexto (ou vizinhança) de determinada palavra. Por essa razão, eles são denominados de ***embeddings* contextualizados** ou **dinâmicos**.

Este capítulo tem como objeto de estudo os modelos de representação vetorial baseados no BoW (vetores esparsos) e na semântica distribucional (vetores densos). Contudo, o foco aqui são os vetores densos estáticos ou ***embeddings* estáticos**. Os vetores densos contextualizados ou ***embeddings* contextualizados**, por sua vez, são abordados no Capítulo Modelos de linguagem.

Introduzimos o capítulo apresentando os conceitos básicos da representação vetorial (Seção 10.1) e as duas formas de representação mais comuns, a **matriz termo-documento** (Seção 10.1.1) e a **matriz termo-contexto** (Seção 10.1.2). Em seguida, introduzimos a **similaridade do cosseno** (Seção 10.1.3), uma maneira padrão de se calcular a similaridade entre dois vetores, que consiste na principal ferramenta para quantificar a

[1] Sentenças retiradas de títulos de notícias retornadas pelo Google, em 18/04/2023, a partir dos termos de busca "ensino" e "educação" combinados com "aluno", "escola" e "professor".

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

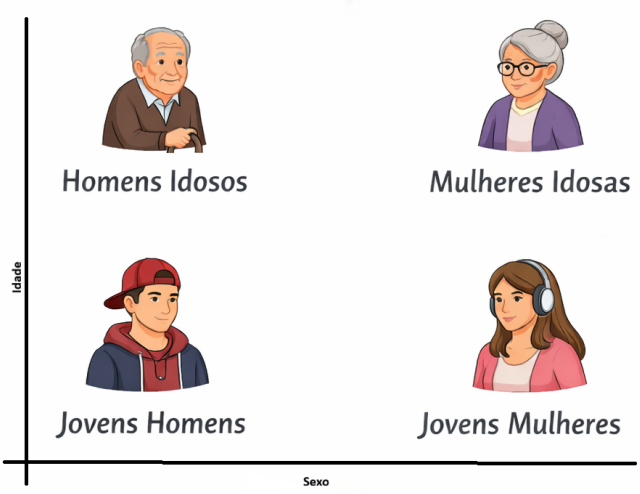
similaridade semântica entre textos e palavras. Na sequência, apresentamos os vetores esparsos, as medidas TF-IDF (Seção 10.2.1) e PMI (Seção 10.2.2), frequentemente usadas na atribuição de pesos aos termos (ou palavras) nesse tipo de representação vetorial, e a técnica LSA (Seção 10.2.3), comumente usada para reduzir a dimensionalidade de vetores esparsos. Por fim, abordamos os *embeddings* estáticos (Seção 10.3) e os modelos de geração de *embeddings* mais comuns como Word2Vec (Seção 10.3.1) , FastText (Seção 10.3.2) e Glove (Seção 10.3.3).

10.1 Representação Vetorial

As primeiras investigações no campo da representação vetorial do significado tiveram início na década de 1950, impulsionadas pela proposta de usar um ponto no espaço multidimensional para representar a conotação de uma palavra. Dentro desse campo, destacam-se os **métodos distribucionais**, fundamentados na convicção de que o significado de uma palavra pode ser definido a partir da sua distribuição nos contextos linguísticos em que ela ocorre. Essa visão é compartilhada por linguistas como (Joos, 1950), (Harris, 1954) e (Firth, 1957) e pela proposta de (Osgood et al., 1957), de usar um ponto no espaço multidimensional para representar a conotação de uma palavra.

Enquanto a representação vetorial é um termo amplo a respeito da conversão de unidades textuais em números, os métodos distribucionais focam especificamente na distribuição das palavras vizinhas que formam o contexto (Jurafsky; Martin, 2023). Por exemplo, considere um espaço semântico com duas características (*features*): sexo e idade (portanto, bidimensional), conforme a Figura 10.1. Observa-se a distinção das duas dimensões nas quais os homens estão situados nos quadrantes mais a esquerda, enquanto as mulheres estão situadas nos quadrantes mais à direita, considerando o eixo x. E considerando o eixo y, que corresponde à idade, os idosos estão mais posicionados nos quadrantes mais superiores, enquanto os jovens estão posicionados nos quadrantes mais inferiores.

Figura 10.1: Espaço semântico bidimensional (sexo e idade).



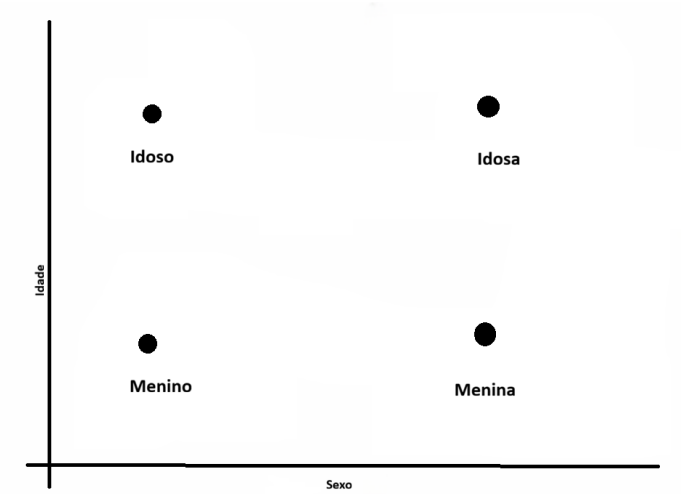
A ideia central é representar cada palavra como um ponto em um **espaço vetorial multidimensional**, construído a partir da distribuição de suas palavras vizinhas. Considerando aqui essas 2 dimensões, caso as palavras a serem representadas fossem *me-*

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

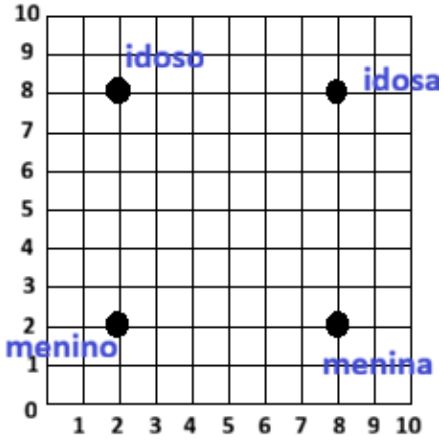
nina, menino, idoso e idosa, empiricamente poderíamos posicioná-las de acordo com a Figura 10.2.

Figura 10.2: Espaço semântico bidimensional (sexo e idade) e as representações das palavras.



Considerando que os eixos são as retas de um plano cartesiano representadas numericamente, observe na Figura 10.3 os posicionamentos das palavras em relação aos eixos x e y, dado dois segmentos de reta de 0 a 10.

Figura 10.3: Plano cartesiano com representação numérica das palavras.



De acordo com o plano cartesiano, observamos que: o ponto *menino* corresponde a (2,2); *menina* corresponde a (8,2); idoso, corresponde a (2,8) e *idosa* corresponde a (8,8). A representação numérica de cada palavra é, portanto, bidimensional dadas as dimensões sexo e idade. À medida que o número de dimensões aumenta, maior a assertividade de representar uma palavra naquele plano cartesiano. A título de exemplo, o GPT-3 foi desenvolvido com vetores de 12.288 dimensões.

Historicamente, essa área evoluiu em como representar os vetores numericamente, iniciando por representações baseadas em contagens globais (por exemplo, baseadas em fre-

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

quência ou na medida TF-IDF) até a utilização de redes neurais para desenvolver representações que produzem vetores densos (Bengio et al., 2003) ou *embeddings*.

Diferentes significados podem ser representados por meio dos *embeddings*, como a similaridade de sentido (ex. "comércio" e "negócio"); a orientação de sentimento ou polaridade (ex. "fenomenal", que conota uma avaliação positiva, e "estúpido" ', que conota uma avaliação negativa); a associação entre palavras (ex. "futebol" e "bola", que são relacionados, uma vez que futebol se joga com uma bola), entre outros aspectos.

Conforme visto na Figura 10.3, espaços vetoriais são objetos de estudo da Álgebra Linear e são bem caracterizados pela sua dimensão, que, grosseiramente falando, representa o número de direções independentes no espaço. Um espaço vetorial é formado por uma coleção de objetos chamados **vetores**. Em um Modelo Semântico Distribucional é possível representar palavras, sentenças e documentos completos.

Geralmente, os vetores são representados por meio de uma **matriz de coocorrência** (ou distribuição de coocorrência), que retrata a frequência de coocorrência das palavras. As representações matriciais mais comuns são a **matriz termo-documento**, onde cada dimensão (vetor) da matriz representa um documento, e a **matriz termo-contexto**, onde cada dimensão representa uma palavra (Jurafsky; Martin, 2023). As subseções a seguir abordam essas duas formas básicas de representação.

10.1.1 Matriz termo-documento

Na matriz **termo-documento**, o espaço vetorial é formado por uma coleção de documentos[2] que representam pontos ou **vetores** nesse espaço. Cada vetor tem dimensão $|V|$, onde $|V|$ representa o tamanho do vocabulário, que contém as palavras distintas (sem repetições) de todos os documentos da coleção. Assim, cada palavra do vocabulário é uma linha na matriz e cada coluna representa um documento da coleção. Cada célula da matriz, por sua vez, representa a frequência de uma palavra em um documento em particular, ou seja, quantas vezes aquela palavra ocorreu naquele documento. Esta abordagem é a base do modelo *bag-of-words* (BoW). A intuição por trás dessa representação é a de que o documento é tratado como fosse um saco de palavras, ou seja, um conjunto não ordenado de palavras, no qual a posição de cada palavra no documento é ignorada e apenas sua frequência é relevante.

A matriz termo-documento foi definida por (Salton; Allan, 1994) como parte do Modelo de Espaço Vetorial – MEV proposto pelos autores para a recuperação de documentos na web. No MEV, um documento é representado como um vetor de frequência de palavras (ou termos), uma coluna, como no exemplo da Tabela 10.1. Os três documentos representados na tabela, por meio de colunas, são pontos em um espaço tridimensional. A matriz foi computada a partir de três textos de divulgação científica extraídos da revista online Pesquisa FAPESP[3], da seção de Tecnologia, edições 308 e 310 de 2021. Os textos se referem aos seguintes temas: energias renováveis[4] (coluna 1), risco de escassez de energia fornecida pelas hidrelétricas brasileiras[5] (coluna 2) e veículos elétricos movidos a etanol[6] (coluna 3). A matriz da Tabela 10.1 representa apenas um subconjunto das palavras que

[2] O termo "documento" é usado nesta seção de forma genérica, podendo se referir a sentenças, parágrafos ou documentos completos.
[3] https://revistapesquisa.fapesp.br/
[4] https://revistapesquisa.fapesp.br/a-forca-das-renovaveis/
[5] https://revistapesquisa.fapesp.br/sob-o-risco-da-escassez/
[6] https://revistapesquisa.fapesp.br/eletricos-movidos-a-etanol/

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

ocorreram nesses textos. A contagem (frequência) das palavras em cada texto foi realizada considerando a sua forma lematizada.

Tabela 10.1: Matriz termo-documento para 5 palavras extraídas dos textos da revista online Pesquisa FAPESP. As células representam a frequência de uma palavra em cada texto.

	energias renováveis	**escassez de energia**	**veículos elétricos**
energia	32	36	9
eólico	15	15	1
renovável	6	8	4
etanol	1	1	22
hidrogênio	0	0	16

Em aplicações reais, os vocabulários têm milhares de palavras e o número de documentos pode ser enorme (imagine todas as páginas da web). Isso frequentemente resulta em vetores muito grandes, levando a matrizes esparsas, já que muitas palavras nunca aparecem em outros documentos. Para lidar com um grande número de dimensões, uma técnica comumente usada é a Análise Semântica Latente (em inglês, *Latent Semantic Analysis* – LSA), que reduz a dimensionalidade do espaço vetorial através da Decomposição em Valores Singulares (em inglês, *Singular Value Decomposition* – SVD), conforme veremos na Seção 10.2.3.

10.1.2 Matriz termo-contexto

Diferente da matriz termo-documento, a matriz termo-contexto é a implementação da hipótese distribucional, em que os vetores semânticos também podem ser usados para representar o significado de **palavras** com base em sua vizinhança imediata e não apenas em sua ocorrência em documentos. Para tanto, ao invés de usarmos uma matriz de **termo-documento** (Seção 10.1.1), usamos uma matriz de **termo-contexto**, também conhecida por matriz **palavra-palavra** ou matriz **termo-termo**.

Na matriz de **termo-contexto**, o espaço vetorial é formado por uma coleção de **palavras** que representam vetores nesse espaço. A matriz de coocorrência tem dimensionalidade $|V| \times |V|$ e cada célula representa o número de vezes que a palavra da linha (alvo) e a palavra da coluna (contexto) coocorrem em algum contexto em um *corpus* de treinamento. O contexto pode ser, por exemplo, um documento, no qual a célula representa o número de vezes que as duas palavras aparecem no mesmo documento. Entretanto, é mais comum usar contextos menores limitados a um número de palavras à esquerda e à direita da palavra-alvo (linha), por exemplo, três palavras à esquerda e três palavras à direita. Dessa forma, cada célula representa o número de ocorrências da palavra da coluna (contexto) em uma janela de três palavras em torno da palavra-alvo.

A Tabela 10.2 apresenta um subconjunto simplificado da matriz de coocorrência termo-contexto computada a partir dos textos da revista Pesquisa FAPESP descritos na Seção 10.1.1. As linhas da tabela representam cada palavra-alvo e cada célula indica o número de vezes que a palavra-alvo correspondente coocorreu em cada contexto (colunas). Para a criação da matriz foi considerada uma janela de contexto de tamanho 5, isto é, cinco palavras à esquerda e cinco palavras à direita da palavra-alvo. Para exemplificação, o vetor da palavra-alvo "energia" está destacado em negrito.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Tabela 10.2: Subconjunto simplificado da matriz de coocorrência termo-contexto computada a partir de três textos da revista Pesquisa FAPESP.

	elétrica	geração	combustível
etanol	0	1	10
veículo	0	0	4
energia	**20**	**5**	**2**
país	5	2	0

Os vetores que representam palavras são frequentemente denominados de ***embeddings***, embora muitas vezes esse termo seja usado de maneira mais restrita para se referir apenas aos **vetores densos** como é o caso das representações obtidas com os modelos Word2Vec, Glove e Fastext (Seção 10.3), e não aos vetores esparsos como o TF-IDF e o PMI (Seção 10.2).

10.1.3 Calculando a similaridade entre vetores

Duas tarefas bastante comuns do processamento de linguagem natural consistem em determinar a semelhança entre dois documentos e obter uma medida de equivalência entre duas palavras. Uma vez que as palavras ou documentos estão representados como vetores, é possível utilizar métricas de distância da Álgebra Linear para calcular a semelhança entre eles.

A **medida do Cosseno**, também conhecida por **distância do Cosseno**, é sem dúvida uma das mais clássicas da área de PLN (Jurafsky; Martin, 2023). Ela calcula a distância entre dois vetores no espaço vetorial a partir do valor do cosseno do ângulo compreendido entre eles. Se o ângulo compreendido for zero (ambos os vetores apontam para o mesmo lugar), a medida resultará no valor 1. Para um ângulo diferente de zero, o valor resultante será inferior a 1. Para vetores ortogonais[7], o valor será zero. Se os vetores apontarem em direções contrárias, o valor será -1. Logo, a medida do Cosseno encontra-se no intervalo fechado entre [-1, 1]. Contudo, como os valores de frequência de termos (palavras) são positivos, o cosseno desses vetores encontra-se no intervalo entre [0, 1], sendo que quanto mais próximo de 1 for o valor, maior é a similaridade entre os vetores.

A **distância do Cosseno** é dada pelo produto escalar entre dois vetores, também chamado de produto interno na Álgebra Linear. Sejam x e y dois vetores semânticos n-dimensionais, ambos representando documentos ou ambos representando palavras, o produto escalar entre x e y é definido pela Equação 10.1:

$$< x, y >= x \cdot y = \|x\|\|y\| Cos\theta$$
$$Cos\theta = \frac{x \cdot y}{\|x\|\|y\|}$$

onde <x,y> representa a somatória dos produtos das coordenadas correspondentes dos vetores x e y, calculada com base na Equação 10.1, e $\|v\|$ representa o comprimento do vetor v definido pela Equação 10.2:

$$x \cdot y = \sum_{i=1}^{n} x_i y_i = x_1 y_1 + x_2 y_2 + x_3 y_3 + \cdots + x_n y_n \tag{10.1}$$

[7] Dois vetores são ortogonais se o Produto Escalar entre eles é nulo (zero), conforme veremos na Equação 10.1.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

$$\|v\| = \sqrt{\sum_{i=1}^{n} v_i^2} \tag{10.2}$$

Em suma, a distância do Cosseno entre dois vetores x e y é definida pela Equação 10.3:

$$Cos = \frac{x \cdot y}{\|x\|\|y\|} = \frac{\sum_{i=1}^{n} x_i y_i}{\sqrt{\sum_{i=1}^{n} x_i^2}\sqrt{\sum_{i=1}^{n} y_i^2}}$$

Para fins de exemplificação, consideremos os dois textos do Exemplo 10.1:

Exemplo 10.1:

1. Eu adorei o filme. Foi incrível.
2. Eu detestei o filme. Foi horrível.

Podemos transformar esses textos em vetores que representam a frequência de ocorrência de cada palavra. Para o conjunto de palavras {eu, adorei, detestei, filme, incrível, horrível}, temos os vetores semânticos a e b, representando o texto 1 e o texto 2, respectivamente:
$a = [1, 1, 0, 1, 1, 0]$
$b = [1, 0, 1, 1, 0, 1]$
Inicialmente, vamos calcular a somatória dos produtos entre as coordenadas correspondentes de a e b, conforme a Equação 10.1:

$$a \cdot b = 1 \cdot 1 + 1 \cdot 0 + 0 \cdot 1 + 1 \cdot 1 + 1 \cdot 0 + 0 \cdot 1 = 2$$

Em seguida, calculamos o comprimento de cada vetor com base na Equação 10.2:

$$\|a\| = \sqrt{1^2 + 1^2 + 0^2 + 1^2 + 1^2 + 0^2} = \sqrt{4} = 2$$

$$\|b\| = \sqrt{1^2 + 0^2 + 1^2 + 1^2 + 0^2 + 1^2} = \sqrt{4} = 2$$

Por fim, substituímos os valores já calculados na Equação 10.3, para obtermos o cosseno do ângulo entre os vetores a e b:

$$Cos(a, b) = \frac{2}{4} = 0,5$$

A similaridade entre os vetores a e b é diretamente proporcional ao cosseno do ângulo entre eles, ou seja, 0,5.

A medida do Cosseno é menos sensível à frequência de ocorrência das palavras em um *corpus* do que outras medidas de similaridade, como a Distância Euclidiana. Isso significa que as palavras menos frequentes não terão um peso desproporcional no cálculo da similaridade entre os vetores. Essa é a principal razão que faz com que essa medida seja tão comum em PLN.

10.2 Vetores esparsos

Vimos nas Seções 10.1.1 e 10.1.2 que as matrizes termo-documento e termo-contexto associam a frequência de ocorrência de cada termo ao documento ou contexto em que ocorrem. No entanto, a frequência simples de um termo (isto é, o número de vezes que ele ocorre)

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

é pouco discriminativa, já que algumas palavras (como "porque", "durante", "após", "sobre" etc.) são bastante comuns e não caracterizam nenhum documento ou contexto em particular. Abordagens mais avançadas como é o caso das medidas TF-IDF (do inglês, *Term Frequency-Inverse Document Frequency*) e PMI (do inglês, *Pointwise Mutual Information*), costumam ser mais eficazes do que a simples frequência de um termo para discriminar o conteúdo de um documento ou um contexto. Como muitos termos nunca ocorrem em alguns documentos de uma coleção ou nunca aparecem em certos contextos, frequentemente, essas medidas levam a vetores com muitas dimensões e esparsos, ou seja, com muitos valores nulos (zeros). Por essa razão, as matrizes que se utilizam dessas medidas para atribuir valores aos termos são comumente chamadas de **vetores esparsos**. As medidas TF-IDF e PMI serão detalhadas nas Seções 10.2.1 e 10.2.2, respectivamente. Em seguida, na Seção 10.2.3, apresentamos a técnica LSA (do inglês, *Latent Semantic Analysis*), muito usada para reduzir a dimensionalidade desses espaços criados com TF-IDF ou PMI e capturar relações semânticas.

10.2.1 Atribuindo pesos aos termos da matriz termo-documento com TF-IDF

A medida TF-IDF representa uma alternativa mais eficiente do que a contagem de termos para atribuir valores aos termos de uma **matriz termo-documento**. Ela atribui um peso para cada termo de um documento multiplicando a frequência do termo no documento (TF) pelo inverso da frequência do termo em todos os documentos de um *corpus* ou coleção (IDF). Dessa maneira, um termo que ocorre muitas vezes em um documento, mas não em muitos documentos da coleção, terá um peso mais alto, enquanto um termo que ocorre em muitos documentos terá um peso mais baixo.

A frequência de um termo (TF) mede a sua importância em um documento. Ela é calculada com base no número de ocorrências de um termo t em um documento d, dividido pelo total de termos do documento d (conforme a Equação 10.3). Essa medida é importante porque, em geral, as palavras que aparecem com mais frequência em um documento são mais relevantes para discriminar o seu conteúdo. Porém, TF somente não é suficiente para identificar as palavras mais importantes de um documento, pois algumas palavras podem ser muito frequentes em muitos documentos e, portanto, não auxiliam na discriminação do seu conteúdo. A frequência inversa no documento (IDF) é, então, fundamental, para atribuir um peso maior às palavras que são frequentes mas ocorrem em poucos documentos de uma coleção.

A frequência inversa no documento (IDF) mede, portanto, a importância relativa de uma palavra em uma coleção de documentos. Ela é calculada dividindo-se o número total de documentos da coleção pelo número de documentos que contêm a palavra em questão e tomando o logaritmo desse resultado. Seja N o número de documentos de uma coleção, o IDF de um termo t é definido pela Equação 10.4. Em outras palavras, o IDF mede a raridade de uma palavra em um conjunto de documentos. Essa medida é importante porque palavras que aparecem em poucos documentos têm um maior poder de discriminação do conteúdo de um documento.

$$tf(t,d) = \frac{numero_ocorrrencias(t,d)}{total_termos(d)} \tag{10.3}$$

$$idf(t) = \log\left(\frac{N}{total_documentos(t)}\right) \tag{10.4}$$

Assim, o TF-IDF de um termo t de um documento d é dado pelo produto entre o valor

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

TF (definido pela Equação 10.3) e o valor de IDF (definido pela Equação 10.4), conforme a Equação 10.5:

$$\texttt{tf-idf}(t, d) = \texttt{tf}(t, d) * \texttt{idf}(t) \tag{10.5}$$

Quanto menor o número de documentos que contêm determinado termo, maior será o TF-IDF daquele termo. Em suma, termos que aparecem com frequência em muitos documentos recebem um peso menor do que os termos mais específicos de um determinado documento. TF-IDF é uma medida bastante versátil e amplamente utilizada em várias tarefas que envolvem o processamento de textos. Alguns exemplos de aplicação mais comuns que podem ser citados são:

- Recuperação de informação: a matriz TF-IDF é usada para classificar e recuperar documentos relevantes com base em termos de pesquisa específicos. Os termos mais importantes, com os valores mais altos de TF-IDF, são usados para classificar a relevância de cada documento.
- Análise de sentimentos: assim como na recuperação de informação, a medida TF-IDF é usada para identificar as palavras mais importantes em um texto e, posteriormente, a polaridade do sentimento associado ao texto (como positiva, negativa ou neutra) é determinada com base nessas palavras.
- Agrupamento de documentos: a matriz TF-IDF também pode ser usada para agrupar documentos que compartilham termos em comum. Os documentos que têm um alto valor de TF-IDF para os mesmos termos farão parte do mesmo grupo.
- Sumarização de textos: na sumarização automática a medida TF-IDF pode ser usada para identificar as palavras-chave ou os tópicos mais relevantes de um texto, para gerar o seu resumo.

10.2.2 Atribuindo pesos aos termos da matriz termo-contexto com PMI

Uma forma mais eficaz de pesar os termos de uma **matriz termo-contexto**, comparada à simples contagem de coocorrência de termos, é usar a medida PMI (do inglês, *Pointwise Mutual Information*). PMI é uma medida estatística que auxilia na identificação de palavras associadas. Ela mede qual é a probabilidade que dois termos ocorram juntos um do outro em relação à probabilidade de cada termo ocorrer de forma independente. Por exemplo, o termo "inteligência artificial" tem um significado específico quando as palavras "inteligência" e "artificial" aparecem juntas em um texto. Quando ocorrem isoladamente, essas duas palavras constroem outros significados.

Formalmente, o PMI entre um termo alvo x e um termo contexto y é definido pela Equação 10.6.

$$pmi(x, y) = \log_2 \frac{p(x, y)}{p(x)p(y)} = \log_2 \frac{p(x|y)}{p(x)} = \log_2 \frac{p(y|x)}{p(y)} \tag{10.6}$$

onde:

$$p(t) = \frac{frequencia(t)}{total_termos}$$

A medida PMI leva em consideração, tanto a probabilidade conjunta dos termos (numerador da Equação 10.6), quanto a distribuição geral de cada termo no *corpus* em que estão sendo analisados (denominador da equação). Vale lembrar que a probabilidade de

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

dois eventos independentes ocorrerem é dada pelo produto das probabilidades dos dois eventos[8].

A medida PMI é simétrica, ou seja, $pmi(x, y) = pmi(y, x)$. Assim, essa medida não considera a ordem de ocorrência das palavras, ou seja, "inteligência artificial" e "artificial inteligência" terão a mesma probabilidade.

Os valores de PMI podem ser negativos ou positivos. O valor de PMI será 0 quando x e y forem independentes, indicando que não há qualquer associação entre os dois termos. Embora o valor possa ser negativo ou positivo, o resultado esperado é sempre positivo. Valores negativos indicam que os termos estão ocorrendo com menos frequência do que esperaríamos que eles ocorressem ao acaso. A medida PMI é máxima quando x e y estão perfeitamente associados (isto é, $p(x|y)$ ou $p(y|x) = 1$). Ou seja, quando os termos têm alta probabilidade conjunta e cada termo tem uma baixa probabilidade de ocorrer isoladamente, como, por exemplo, o termo "fiada" que provavelmente ocorre com maior frequência com o termo "conversa" (na expressão "conversa fiada") do que isoladamente.

O fato de a medida PMI poder assumir valores positivos ou negativos e sem limites definidos torna mais difícil a interpretação dos resultados. Uma maneira de contornar esse problema é substituir os valores negativos por 0, aplicando uma variação de PMI, denominada PPMI (*Positive Pointwise Mutual Information*), definida pela Equação 10.7:

$$ppmi(x, y) = max\left(\log_2 \frac{p(x, y)}{p(x)p(y)}, 0\right) \tag{10.7}$$

A medida PPMI é motivada pela observação de que valores negativos de PMI tendem a não ser confiáveis, a menos que tenhamos um *corpus* muito grande e expressivo. Além disso, evita que tenhamos que lidar com valores negativos tendendo ao infinito ($-\infty$) para termos que nunca ocorrem juntos (isto é, $p(x, y) = 0$). Essa variante da medida PMI é mais robusta e precisa para a construção de modelos semânticos distribucionais, uma vez que ela remove os valores negativos de PMI que podem prejudicar a precisão dos modelos. Em geral, quanto maior for o *corpus*, mais confiável será a medida de associação.

Embora PPMI seja uma medida útil para quantificar a força da relação entre dois termos em um *corpus*, ela pode ser muito sensível à frequência dos termos, podendo superestimar a associação entre palavras raras e subestimar a associação entre palavras comuns. Outra limitação das medidas PMI e PPMI é que, por depender somente da coocorrência de palavras no *corpus*, não levam em consideração a posição das palavras no texto. Isso pode ser um problema em situações em que a ordem das palavras é importante como é o caso das expressões multipalavras.

Medidas de associação de termos como PMI e PPMI são úteis em uma variedade de aplicações de processamento de linguagem natural, conforme exemplificamos a seguir:

- Construção de modelos de linguagem: são úteis para identificar palavras que coocorrem com frequência em *corpora*, informação relevante que pode ser incorporada em modelos de linguagem, como os modelos Word2Vec e Glove.
- Agrupamento de tópicos: são úteis para o agrupamento de documentos por tópicos com base na identificação de palavras que são altamente correlacionadas a determinado tópico.
- Recomendação de conteúdo/produto: são importantes para recomendar produtos com base nas preferências do usuário, auxiliando na identificação de palavras frequen-

[8] Os dois últimos termos da Equação 10.6 são considerados equivalentes pelo Teorema de Bayes.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

temente associadas a conteúdos/produtos específicos. Essas associações de palavras são posteriormente usadas para fazer recomendações personalizadas.
- Tradução automática: são úteis para a identificação de palavras que ocorrem frequentemente em um mesmo contexto de tradução.
- Análise de sentimento: são úteis para analisar a opinião ou sentimento expresso em um texto, a partir da identificação de palavras-chave que são altamente correlacionadas com um determinado sentimento, como "bom" ou "ruim".

10.2.3 Reduzindo a dimensionalidade com LSA

As medidas TF-IDF e PMI/PPMI apresentadas nas Seções 10.2.1 e 10.2.2 frequentemente geram matrizes esparsas contendo muitas células com valor nulo, dado que muitos termos têm baixa frequência em muitos documentos (no caso da matriz termo-documento) ou não ocorrem em muitos contextos em um *corpus* (no caso da matriz termo-contexto). A esparsidade de uma matriz pode afetar significativamente o desempenho dos algoritmos que a processam, levando a um aumento no tempo de processamento, devido à necessidade de acessar todos valores armazenados, inclusive os valores nulos. Por conter muitos zeros, a matriz esparsa ainda demanda muita memória para poder armazenar todos os valores, mesmo com *corpora* relativamente pequenos.

Para lidar com esse problema decorrente do grande número de dimensões, é comum o uso de técnicas que permitam reduzir a dimensionalidade de uma matriz. A Análise de Semântica Latente (em inglês, *Latent Semantic Analysis* – LSA) ou, ainda, Indexação de Semântica Latente (em inglês, *Latent Semantic Indexing* – LSI), como é chamada na área de recuperação de informação, é uma técnica que pode ser usada para reduzir essa dimensionalidade, atuando como uma ponte entre as representações esparsas e a criação de espaços conceituais densos.

O objetivo da LSA é reduzir a dimensão da matriz original, diminuindo a importância de valores singulares menores. Isso ajuda a eliminar ruídos e a capturar as relações semânticas subjacentes entre as palavras. O resultado é um espaço semântico onde termos e documentos semelhantes são posicionados próximos, mesmo que não compartilhem palavras idênticas.

A LSA pode ser aplicada tanto em matrizes de termo-documento como em matrizes de termo-contexto (ou palavra-palavra). A ideia por trás dessa técnica é que existe alguma estrutura semântica latente subjacente nos dados, que é ocultada em certa medida pela incerteza na escolha das palavras. Para estimar essa estrutura latente, é usada uma técnica de análise matricial, da Álgebra Linear, chamada Decomposição em Valores Singulares (em inglês, *Singular Value Decomposition* – SVD) (Golub; Reinsch, 1970), que permite determinar o conjunto de padrões que descreve um documento ou contexto. A SVD encontra os principais eixos de variância no espaço vetorial. Assim, ao decompor uma matriz termo-documento em valores singulares, por exemplo, é possível identificar o conjunto de termos padrão que descreve qualquer documento daquela coleção. Na decomposição, somente as k dimensões mais importantes da matriz são mantidas. Como resultado, tem-se um espaço semântico no qual termos e documentos semelhantes são colocados próximos uns dos outros.

De maneira geral, decompor uma matriz A de ordem $m \times n$ (para m diferente de n) significa fatorar essa matriz em três matrizes fatores, tal que o produto dessas matrizes permite recompor a matriz A original. Formalmente, a Decomposição em Valores Singulares de uma matriz $m \times n$ de valores reais A é igual ao produto de três matrizes fatores

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

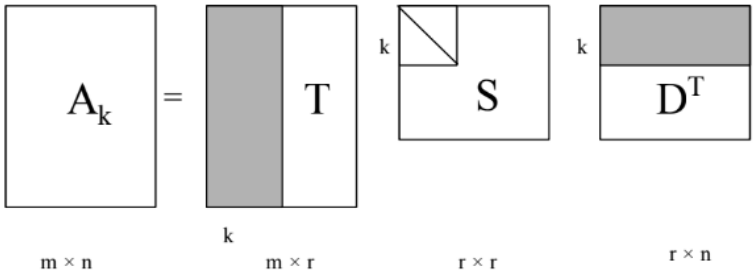
©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

T, S e D^T, conforme segue:

$$A = TSD^T$$

onde T e D são colunas (vetores) ortonormais e S é a matriz diagonal de valores singulares. As colunas de T e D representam os vetores singulares à esquerda (vetores de palavras) e à direita (vetores de documentos ou contextos) de A, respectivamente. A matriz A_k $m \times n$ é construída a partir dos k maiores valores singulares (Figura 10.4).

Figura 10.4: Decomposição em Valores Singulares



Na Figura 10.4, m representa o número de palavras (termos), n é o número de documentos (ou contextos), k é a dimensão desejada do espaço de conceitos reduzido, ou seja, representa o total de valores singulares mais significativos a serem mantidos na matriz reduzida e r é o ranqueamento da matriz A ($A(min(m, n))$.

Durante a construção da matriz termo-documento ou termo-contexto, os valores nas células da matriz representam a frequência com que uma palavra ocorre em um determinado documento (ou contexto). Essa coocorrência de palavras gera padrões de associação. Dessa forma, na matriz T resultante da decomposição SVD, as palavras que possuem semelhanças semânticas são definidas através dos padrões de coocorrência entre palavras nos documentos (ou contextos) originais. Mais especificamente, a matriz T contém vetores de palavras que representam as relações semânticas latentes entre as palavras no espaço de conceitos. Cada vetor é uma combinação linear ponderada desses conceitos latentes. As palavras que compartilham padrões semânticos semelhantes terão vetores de palavras semelhantes no espaço de conceitos. A similaridade semântica entre palavras pode ser calculada utilizando a **distância do Cosseno** (Seção 10.1.3) entre os vetores de palavras no espaço de conceitos. Palavras que são semanticamente similares terão vetores que apontam em direções semelhantes e, portanto, terão valores de similaridade próximos de 1. Essa abordagem permite capturar relações semânticas latentes entre as palavras, mesmo que elas não tenham aparecido juntas frequentemente nos documentos originais. Isso é especialmente útil para encontrar sinônimos, identificar tópicos subjacentes e lidar com a chamada "maldição da dimensionalidade" em dados textuais.

A matriz S é obtida com base no produto entre a matriz original A e a sua matriz transposta (A^T), resultando em uma matriz de dimensão $n \times n$ (se a matriz original for $m \times n$). Em seguida, calculam-se os autovalores e autovetores da matriz resultantes do produto $A * A^T$. Os autovalores representam a variabilidade dos dados em direções específicas no espaço. Os valores singulares são obtidos a partir dos autovalores. Eles são a raiz quadrada dos autovalores positivos, ordenados em ordem decrescente. A matriz S é então construída como uma matriz diagonal, onde os valores singulares são colocados na diagonal principal e os demais elementos são zeros. Quanto maior o valor singular, maior a contribuição desse

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

conceito latente na representação dos dados originais. A redução da dimensionalidade envolve manter apenas os primeiros k valores singulares mais significativos, onde k é a dimensão desejada no espaço de conceitos reduzido.

A matriz D^T (de documentos ou contextos para conceitos), por sua vez, é calculada utilizando os autovetores associados à matriz resultante do produto $A * A^T$. Os autovetores são os mesmos que compõem a matriz T. Eles são normalizados para terem comprimento igual a 1. Isso é importante para garantir que as informações de magnitude não sejam distorcidas. A matriz D^T é construída colocando os autovetores normalizados como colunas, onde cada coluna representa um vetor de documentos ou de contextos no espaço de conceitos.

Embora LSA seja uma técnica poderosa para análise de texto e redução de dimensionalidade, ela possui algumas limitações que devem ser consideradas. Por exemplo, ao tratar os termos como entidades independentes, ignorando as relações de contexto mais complexas que ocorrem em linguagem natural, nuances de significado que dependem do contexto podem não ser totalmente capturadas. Tratando as palavras de maneira independente, essa técnica também não captura a estrutura sintática da linguagem.

Outra limitação importante é que a técnica LSA tende a ter dificuldade em lidar com documentos muito curtos, uma vez que a coocorrência de termos relevantes é menor e a representação no espaço de conceitos pode ser menos robusta. A LSA é uma abordagem estática, o que significa que não é capaz de lidar bem com mudanças no significado das palavras ao longo do tempo ou em diferentes contextos.

Embora a LSA capture informações semânticas latentes, os conceitos extraídos nem sempre são facilmente interpretáveis por seres humanos. Isso dificulta a compreensão do que exatamente está sendo capturado em cada dimensão reduzida do espaço de conceitos. Além disso, ela não é capaz de capturar nuances semânticas mais complexas, como a ambiguidade.

Em resumo, LSA é uma técnica valiosa para muitas tarefas de processamento de linguagem natural, mas é importante estar ciente de suas limitações e considerar outras abordagens, como modelos baseados em redes neurais, para lidar com algumas das desvantagens mencionadas.

10.3 Vetores densos estáticos

Vimos na Seção 10.2 como representar uma palavra por meio de um vetor esparso e com muitas dimensões, correspondentes às palavras do vocabulário ou aos documentos de uma coleção. Nesta seção, introduziremos uma representação de palavras mais robusta, conhecida por ***embeddings***, de **vetores densos** e menores, com dimensões variando entre 50-1000. Essas dimensões não possuem uma interpretação clara do seu significado (Jurafsky; Martin, 2023).

Os vetores são densos, ou seja, seus valores são números reais positivos ou negativos, ao invés de contagens esparsas, na maioria das vezes zeros, como é o caso dos vetores esparsos vistos na Seção 10.2. Vetores densos (daqui em diante, *embeddings*) capturam melhor as relações semânticas e contextuais entre as palavras do que os vetores esparsos. Por exemplo, na representação vetorial esparsa, sinônimos como "alfabeto" e "abecedário" muito provavelmente têm dimensões distintas e não relacionadas, pois esse tipo de representação pode falhar ao capturar a similaridade entre palavras que estão no contexto de "alfabeto" e "abecedário". Essa é uma das razões que faz com que os *embeddings* apresentem melhor desempenho em tarefas de PLN do que os vetores esparsos.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Os *embeddings* são aprendidos a partir de *corpora* por meio de algoritmos de aprendizado de máquina supervisionado ou não supervisionado, por exemplo, usando redes neurais artificiais como é o caso do modelo Word2Vec (Seção 10.3.1), ou, ainda, usando representação estatística da matriz de coocorrência de termos, como é o caso do modelo Glove (Seção 10.3.3).

Os modelos de geração de *embeddings* podem ser **estáticos** ou **contextualizados**. Nos modelos **estáticos**, cada palavra do vocabulário é mapeada para um único vetor (uma lista de números) em um espaço multidimensional. Uma vez que o modelo é treinado, cada palavra terá a mesma representação independente do seu contexto de ocorrência. Por exemplo, o *embedding* de "manga" seria o mesmo na frase "Eu comi manga." e na frase "Minha camisa tem manga comprida.". Nos modelos de *embeddings* **contextualizados**, por outro lado, a representação numérica de palavras é gerada de forma **dinâmica**, ou seja, é definida no momento do uso e leva em consideração o **contexto da sentença** em que a palavra está inserida. Ao contrário dos modelos estáticos, os *embeddings* contextualizados permitem que uma mesma palavra tenha diferentes representações dependendo da frase em que aparece. Isso permite lidar melhor com palavras polissêmicas, tornando os *embeddings* mais adequados para tarefas que dependem do contexto. No exemplo anterior, a palavra "manga" teria *embeddings* distintos nas duas frases, ou seja, um representando o sentido de fruta (em "Eu comi manga.") e outro representando a manga da camisa (na frase "Minha camisa tem manga comprida."). A escolha entre esses diferentes modelos depende das necessidades da aplicação, do domínio e das características da tarefas em que os *embeddings* serão utilizados.

Nesta seção focamos na descrição dos modelos de geração de ***embeddings* estáticos**, como Word2Vec (Seção 10.3.1), Fasttext (Seção 10.3.2) e GloVe (Seção 10.3.3). Os modelos de geração de ***embeddings* contextualizados** como os ELMo, BERT e GPT são abordados no Capítulo Modelos de linguagem.

Os ***embeddings* estáticos** podem ser definidos para diversos tipos de unidades de representação, incluindo palavras, caracteres, subpalavras, sentenças e até mesmo textos com várias sentenças. Por exemplo, considere as sentenças do Exemplo 10.2:

Exemplo 10.2:

1. Vou ao **banco** sacar dinheiro.
2. Adoro sentar no **banco** da praça.

Considerando o contexto da sentença e o senso comum, o termo "banco" na primeira sentença corresponde a uma *instituição financeira* cujo significado é distinto do "banco" da segunda sentença que corresponde a um *assento*. Neste caso, os *embeddings* estáticos definem um mesmo vetor para representar a palavra "banco" nas duas sentenças, independente do contexto.

Formalmente, as unidades de representação e seus vetores são representados por uma matriz

$$M \in \mathbb{R}^{|V| \times d}$$

onde $|V|$ é o tamanho do vocabulário V e d é a dimensão do *embedding*, em geral, um valor entre 50-1000. Cada linha da matriz contém o vetor estático da unidade de representação $u_i \in V$.

Embora a ideia de representar elementos de um texto usando vetores no espaço multidimensional não seja tão recente (vide, por exemplo, (Joos, 1950), (Harris, 1954), (Firth,

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

1957), (Osgood et al., 1957)), somente a partir de 2013 os *embeddings* começaram a ser muito utilizados, com o desenvolvimento e a disponibilização do modelo Word2Vec (Mikolov et al., 2013b), conforme explicado a seguir.

10.3.1 Word2Vec

O Word2Vec é uma técnica de aprendizado de unidades de representações distribuídas proposta por Mikolov et al. (2013b), que tem como objetivo capturar a semântica e a relação entre unidades de representação em um *corpus*, aprendendo *embeddings* estáticos para cada palavra presente no vocabulário de treino (Jurafsky; Martin, 2023). O modelo é baseado na ideia de que palavras que ocorrem em contextos semelhantes têm significados semelhantes. Portanto, ele explora a distribuição de palavras em grandes *corpora* para aprender representações vetoriais que capturam esses padrões.

O Word2Vec possui duas arquiteturas principais: CBOW – *Continuous Bag-of-Words* e Skip-gram. Na arquitetura CBOW, o modelo tenta prever uma palavra-alvo com base em um contexto de várias palavras de entrada. O contexto é definido por um conjunto de palavras vizinhas da palavra-alvo. Ao contrário do CBOW, na arquitetura Skip-gram o modelo tenta prever o contexto (palavras vizinhas) de uma palavra-alvo. Dito de outra forma, o Skip-gram tenta encontrar as palavras que normalmente aparecem no contexto da palavra-alvo. Esse método é, em geral, mais lento de treinar, mas muitas vezes gera representações mais precisas.

As subseções 10.3.1.1 e 10.3.1.2 explicam as arquiteturas CBOW e Skip-gram, respectivamente.

10.3.1.1 CBOW

A ideia principal por trás do CBOW é prever a palavra-alvo com base no contexto. O contexto é definido por um conjunto de palavras vizinhas da palavra-alvo. Por exemplo, na frase "Vou ao **banco** sacar dinheiro.", o CBOW tentaria prever a palavra "banco" com base no contexto [Vou, ao, sacar, dinheiro], sendo [Vou, ao] o conjunto de palavras anteriores a palavra-alvo e [sacar, dinheiro] o conjunto de palavras posteriores. Dessa forma, o modelo aprende a associação entre as palavras do contexto e a palavra-alvo.

A arquitetura do CBOW é baseada em uma rede neural de uma única camada oculta (Figura 10.5). A camada de entrada possui um neurônio para cada palavra no contexto. A camada oculta (camada de projeção) tem um número fixo de neurônios, que é um hiperparâmetro definido antes do treinamento. A predição da palavra-alvo $w(t)$ é feita base nas palavras de histórico e palavras futuras e o resultado calculado na saída da projeção é a média de todos os vetores de todas as palavras presentes no contexto da camada de entrada. Este modelo pode prever tanto palavras anteriores como posteriores a uma palavra-alvo em um contexto.

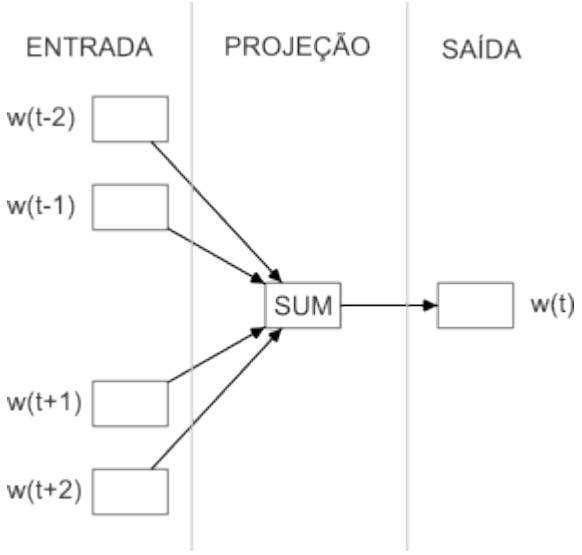
O treinamento do modelo utilizando a arquitetura CBOW é dada pela Equação 10.8, onde N é o tamanho da camada de projeção, isto é, o número de neurônios da camada oculta; D é a dimensão dos vetores; e V é o vocabulário. A primeira parte da fórmula, $N \times D$, refere-se à multiplicação de matrizes durante o treinamento. A segunda parte, $D \times \log_2(V)$, está relacionada ao *softmax*, usado para calcular a probabilidade de cada unidade de representação no vocabulário V ser a representação de destino, com base nos vetores das representações de entrada.

$$Q = N \times D + D \times \log_2(V) \tag{10.8}$$

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Figura 10.5: Arquitetura CBOW. Prediz a palavra-alvo com base no contexto.



Fonte: Adaptado de (Mikolov et al., 2013b)

O *softmax* é uma generalização da função *sigmoid* que converte um vetor numérico em um vetor de probabilidades de possíveis saídas (Jurafsky; Martin, 2023), com valores dentro do intervalo [0,1] perfazendo um somatório de 1, como representado na Equação 10.9.

$$softmax(v_i) = \frac{exp(v_i)}{\sum_{j=1}^{|V|} exp(v_j)} \tag{10.9}$$

Assim, a função *softmax* tem por objetivo converter um vetor numérico em um vetor normalizado de probabilidades.

10.3.1.2 Skip-gram

Enquanto o modelo CBOW prevê a representação de uma unidade (e.g. palavra) com base nos contextos em que ela ocorre em um *corpus* de treinamento, o modelo Skip-gram tenta prever o contexto (ou as representações vizinhas) a partir de um alvo. Considerando as representações como palavras. Nesta arquitetura, cada palavra é uma entrada para uma rede neural similar à arquitetura do CBOW, com camada de projeção para prever palavras dentro de um determinado intervalo antes e depois da palavra de entrada, conforme Figura 10.6. Utilizando o mesmo exemplo anterior, o Skip-gram recebe a palavra "banco" como entrada e tenta prever o contexto "Vou, ao, sacar, dinheiro". Essa abordagem permite que o modelo aprenda a representação de uma palavra, considerando as palavras que normalmente aparecem ao seu redor.

Nesta arquitetura do Skip-gram, aumentar o tamanho do intervalo resulta em vetores melhores; entretanto, isso também aumenta a complexidade computacional do algoritmo. A complexidade do treinamento do modelo utilizando a arquitetura Skip-gram é proporcional à Equação 10.10, onde C é a distância máxima entre as palavras.

$$Q = C \times (D + D \times \log_2(V)) \tag{10.10}$$

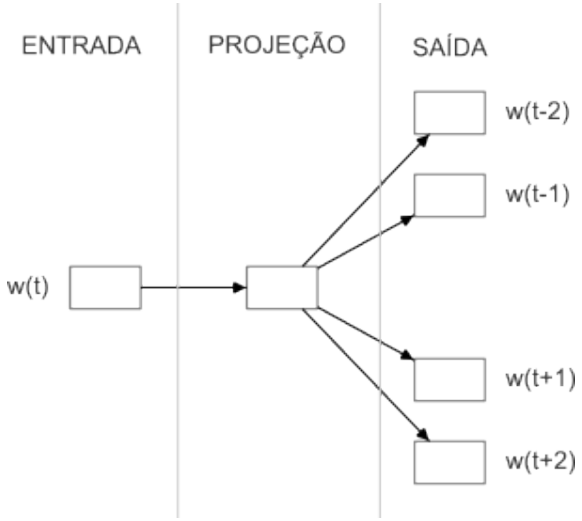
https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

O Skip-gram parte da ideia de que é mais provável que uma palavra ocorra próxima a uma palavra-alvo se seu vetor for similar ao vetor desta palavra-alvo. Esse cálculo de similaridade é baseado na similaridade do Cosseno (Seção 10.1.3).

Em suma, o Skip-gram treina um classificador que, dada uma palavra-alvo e seu contexto, calcula uma probabilidade do quão similares são o contexto e a palavra-alvo. Quanto mais distantes forem os vetores do contexto e da palavra-alvo, menos relacionados são a palavra-alvo e o contexto e, portanto, menor será o peso daquele contexto no treinamento (Mikolov et al., 2013b).

Figura 10.6: Arquitetura Skip-gram. Prediz o contexto com base na palavra-alvo.



Fonte: Adaptado de (Mikolov et al., 2013b)

10.3.1.3 Treinamento do Word2Vec

O processo de treinamento do Word2Vec envolve a criação de um vocabulário a partir do *corpus*. Durante o treinamento, o modelo ajusta os valores desses vetores para maximizar a capacidade de prever a palavra-alvo ou o contexto, dependendo da arquitetura escolhida.

Embora considerada mais complexa que a CBOW, a abordagem Skip-gram é mais comumente utilizada com a técnica de amostragem negativa (*Skip-gram with Negative Sampling* – SGNS). Na amostragem negativa apenas um subconjunto de palavras negativas (palavras não contextuais) é selecionado para a atualização dos pesos, ao invés de ajustar os pesos de todas as palavras no *corpus* a cada iteração. Isso torna o treinamento mais eficiente computacionalmente, aprendendo boas representações, especialmente para palavras mais frequentes (Mikolov et al., 2013b).

Uma vez treinado, o modelo Word2Vec é capaz de fornecer representações vetoriais para palavras (*embeddings*), nas quais palavras semanticamente similares são mapeadas para regiões próximas do espaço vetorial.

10.3.1.4 Limitações do Word2vec

Apesar das vantagens de se utilizar o modelo Word2Vec, é importante destacar que o mesmo possui algumas limitações. Embora o modelo Word2Vec capture as relações semân-

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

ticas entre palavras, as palavras com múltiplos sentidos podem ser ambíguas dependendo do contexto, dificultando a precisão da representação.

Outra limitação do Word2Vec é que ele não lida bem com palavras raras ou fora do vocabulário; consequentemente essas palavras podem gerar representações vetoriais pouco confiáveis. Além disso, a arquitetura Skip-gram não captura explicitamente relações sintáticas, tais como relações entre adjetivos, verbos, por exemplo. Sendo assim, para tarefas que exigem um entendimento mais profundo da estrutura gramatical, outros modelos ou técnicas podem ser mais adequados.

Além da questão da representação de palavras do vocabulário, o modelo não leva em consideração questões morfológicas e ignora a estrutura interna das palavras, o que é uma limitação especialmente para línguas morfologicamente ricas, ou seja, que possuem uma grande variedade de morfemas que podem para expressar diferentes funções gramaticais e que podem ser adicionados, alterados ou combinados para criar diferentes formas de palavras e estruturas gramaticais dentro da língua, como Árabe ou Finlandês.

Por fim, o Word2Vec não lida diretamente com a concatenação de palavras como uma única unidade. Por exemplo, o modelo não conseguiria reconhecer a palavra "pontapé", se ela não estivesse presente no vocabulário, mesmo se o vocabulário de treino contivesse as palavras "ponta" e "pé". No entanto, existem variações do modelo que permitem capturar informações contextuais mais ricas, como o modelo Fasttext descrito a seguir.

10.3.2 Fasttext

O Fasttext é uma extensão do modelo Word2Vec criada pelo grupo *Facebook AI Research* (FAIR), que amplia o conceito do Skip-gram, no qual cada palavra é representada como uma combinação de *n-gramas* de caracteres. O modelo leva em consideração não apenas as palavras individuais, mas também as subpalavras (morfemas) que compõem as palavras. Isso permite que o Fasttext consiga melhor representar as palavras fora do vocabulário e consequentemente, captura informações contextuais mais ricas, especialmente em idiomas com aglutinação e morfologia complexa.

A principal diferença em relação ao Word2Vec está no tratamento das palavras, ou seja, ao invés de tratá-las como uma única unidade, o Fasttext as divide em *n-gramas* menores, como *n-gramas* de caracteres, e as representa como a soma de seus *n-gramas*.

Mais especificamente, cada palavra w é representada como um saco de caracteres (*bag of characters*). Para cada palavra, são adicionados os símbolos '<' e '>' denotando o início e fim das palavras, respectivamente. Isso permite a distinção entre prefixos, sufixos e outras sequências de caracteres. Além disso, a própria palavra w é incluída no conjunto de *n-gramas*, para aprender um *embedding* para cada palavra além dos *n-gramas* dos caracteres.

A Figura 10.7 apresenta um exemplo de saco de caracteres da palavra "computador", considerando $n = 3$ para o tamanho dos *n-gramas*. É importante ressaltar que a sequência "<computador>" tem como 3-gramas de caracteres palavras do português como "com" e "dor"; entretanto, as sequências "com" e "dor" são diferentes dos trigramas "com" e "dor" da palavra "<computador>". Também é possível capturar como informação o prefixo "com" e sufixo "dor" da palavra.

Segundo Bojanowski et al. (2017), o modelo extrai os n-gramas para n maior ou igual a 3 e menor ou igual a 6. Diferentes conjuntos de n-gramas podem ser considerados, por exemplo, tomando todos os prefixos e sufixos.

Em termos arquiteturais, o FastText também possui duas arquiteturas assim como o Word2Vec: CBOW – *Continuous Bag-of-Words* e Skip-gram.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Figura 10.7: Exemplo de representação de trigramas adotados pelo modelo FastText.

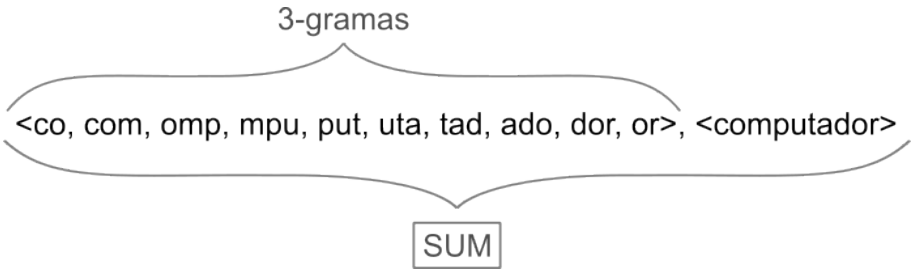
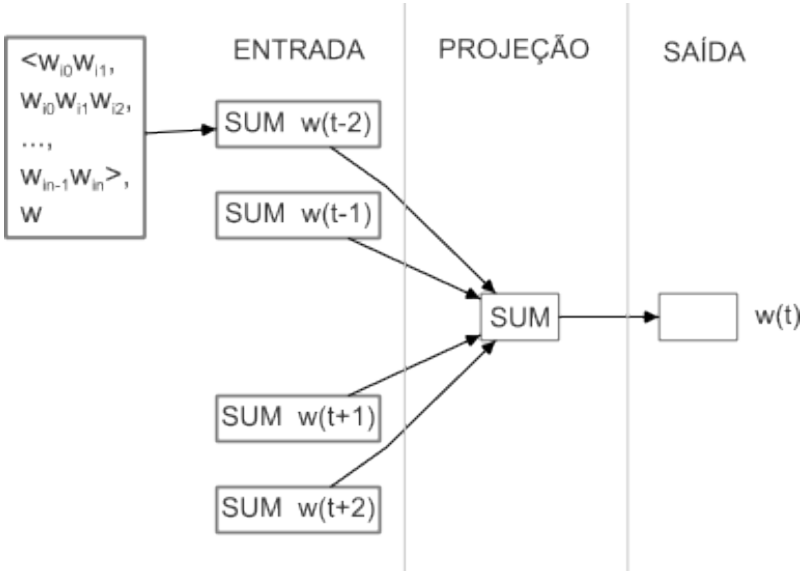


Figura 10.8: Arquitetura FastText-CBOW.



A Figura 10.8 exemplifica o funcionamento do FastText com a arquitetura CBOW. Essa arquitetura prediz a palavra-alvo com base no contexto. Cada *embedding* é gerado a partir dos vetores dos n-gramas, neste exemplo, sendo bigramas.

10.3.2.1 Limitações do FastText

Embora o FastText seja um modelo eficiente e útil para muitas tarefas de processamento de linguagem natural, ele também possui algumas limitações. A primeira é a representação limitada do contexto. Nesse modelo, os textos são representados como a soma das representações de vetor de palavras individuais. Isso significa que o Fasttext não captura informações de ordem ou estrutura sintática mais complexa nos textos.

Outra limitação importante é que o modelo é sensível ao tamanho do vocabulário. Em outras palavras, o FastText requer uma representação de vetor de palavras predefinida para cada palavra no vocabulário. Isso pode levar a problemas de dimensionamento, quando se lida com vocabulários muito grandes, pois o espaço vetorial se torna maior e o treinamento e a inferência podem se tornar mais lentos e exigentes em recursos computacionais.

Embora o FastText seja eficaz em muitos idiomas, pode não funcionar tão bem em idiomas com morfologia muito complexa, onde as palavras se desdobram em várias formas com significados diferentes. Em tais casos, modelos que incorporam informações morfológicas mais profundas podem ser mais apropriados.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

É importante notar que as limitações do FastText não o tornam inadequado para todas as tarefas de processamento de linguagem natural. Ele continua sendo uma escolha sólida para muitos cenários devido à sua eficiência e simplicidade, mas é importante considerar essas limitações ao decidir qual modelo utilizar em um projeto específico. Em tarefas mais complexas e em idiomas com características particulares, pode ser necessário explorar modelos mais avançados como os modelos contextualizados abordados no Capítulo Modelos de linguagem.

10.3.3 Glove

Ao explorar modelos de *embeddings* estáticos como Word2Vec e FastText, é importante mencionar outra abordagem: o modelo GloVe (*Global Vectors for Word Representation*) (Pennington et al., 2014). Enquanto o Word2Vec e o FastText se concentram principalmente na relação local entre as palavras, o GloVe adota uma perspectiva global, levando em consideração a contagem de coocorrência palavra-palavra em um *corpus*. Essa abordagem permite que o GloVe capture informações de relação semântica e sintática entre as palavras.

O GloVe é um modelo global de regressão log-bilinear, que relaciona as variáveis dependentes e independentes por meio de uma função logarítmica (Pennington et al., 2014). Uma função log-bilinear é uma função não linear que tem como argumentos dois vetores. Neste modelo, a regressão log-bilinear é aplicada para estimar os vetores de palavras e as matrizes de transformação necessárias para mapear as palavras em um espaço vetorial.

Primeiramente, uma matriz de coocorrência é construída a partir de um *corpus* de textos. Essa matriz registra quantas vezes duas palavras aparecem juntas em uma janela de contexto. A partir da matriz de coocorrência, é construída uma matriz de probabilidade que representa a probabilidade condicional de uma palavra ocorrer perto de outra palavra. Essa matriz tenta capturar a relação entre as palavras considerando suas frequências relativas de coocorrência. O objetivo do GloVe é encontrar representações vetoriais para palavras de forma que a relação entre os vetores corresponda à relação entre suas probabilidades de coocorrência. Isso é formulado como uma função de perda que minimiza o erro entre as relações de coocorrência reais e as estimadas. O modelo é treinado ajustando os vetores de palavras para minimizar a função de perda. Isso é feito usando um algoritmo de otimização, o Gradiente Descendente Estocástico (*Stochastic Gradient Descent – SGD*). O *SGD* é um algoritmo de otimização para ajustar os parâmetros de um modelo de acordo com uma função de custo a ser minimizada.

Ao utilizar uma combinação das vantagens dos métodos existentes (fatoração de matriz global e janela de contexto local), há uma análise das propriedades dos modelos que não eram totalmente exploradas e argumentam que o ponto de partida apropriado para o aprendizado de *embeddings* deve ser com proporções de probabilidades de coocorrência, ao invés das próprias probabilidades (Pennington et al., 2014).

Considerando a matriz de contagens de coocorrência palavra-palavra denotada por X, cuja entrada X_{ij} tabula o número de vezes que a palavra j ocorre no contexto da palavra i e seja $X_i = \sum_k X_{ik}$ o número de vezes que qualquer palavra aparece no contexto da palavra i, tem-se que $P_{ij} = P(j|i) = X_{ij}/X_i$ é a probabilidade de que a palavra j apareça no contexto da palavra i.

De maneira geral, o modelo GloVe busca converter X em matrizes de atributos W, onde as linhas de W são preenchidas por palavras do vocabulário e cada coluna corresponde a uma dimensão no espaço vetorial (Pennington et al., 2014).

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

A proporção de coocorrência (P_{ik}/P_{jk}) depende de três palavras, i, j, e k, de forma que o modelo mais geral assume a forma:

$$F(w_i, w_j, \widetilde{w}_k) = \frac{P_{ik}}{P_{jk}} \tag{10.11}$$

onde F distingue palavras relevantes de palavras irrelevantes da seguinte maneira:

- Se a palavra k estiver relacionada às palavras i e j (P_{ik} e P_{jk} são grandes), ou não relacionada às palavras i e j (P_{ik} e P_{jk} são pequenos), então o valor de F seria próximo de 1.
- Se a palavra k estiver relacionada exatamente a uma das palavras i ou j, então o valor de F estaria longe de 1.

Como os espaços vetoriais são estruturas inerentemente lineares, a maneira mais natural de codificar a informação presente em P_{ik}/P_{jk} é com diferenças vetoriais. Sendo assim é possível restringir F, modificando a Equação 10.11 para

$$F(w_i - w_j, \widetilde{w}_k) = \frac{P_{ik}}{P_{jk}}. \tag{10.12}$$

A função de custo J do modelo tem como objetivo aprender vetores de palavras otimizados para prever as probabilidades de co-ocorrência de palavras no *corpus* e é dada pela Equação 10.13:

$$J = \sum_{i,j=1}^{V} f(X_{ij})(w_i^T \widetilde{w}_j + b_i + \tilde{b}_j - \log X_{ij})^2 \tag{10.13}$$

onde V é o tamanho do vocabulário e b é bias (ou viés) (Jurafsky; Martin, 2023) que leva em conta a frequência de cada palavra. $f(X_{ij})$ é uma função de ponderação introduzida para compensar X_{ij}, uma vez que $log(X_{ij})$ é indefinido nesse ponto, e também para equilibrar a contribuição de palavras frequentes e pouco frequentes para o modelo.

Para a função de custo J, a função f deve obedecer às seguintes propriedades (Pennington et al., 2014):

1ª propriedade:

- f deve satisfazer $f(0) = 0$. Se f for contínua, deve desaparecer como $x \to 0$ rápido o suficiente dado que o $\lim_{x\to 0} f(x) \log^2 x$ é finito;

2ª propriedade:

- $f(x)$ deve ser não decrescente para que raras coocorrências (pequeno x) não tenham excesso de peso (tem f relativamente grande);

3ª propriedade:

- $f(x)$ deve ser relativamente pequeno para grandes valores de x, para que as coocorrências frequentes não sejam sobrecarregadas.

Partindo dessas propriedades, Pennington et al. (2014) propõem a função $f(x)$:

$$f(x) = \begin{cases} (x/x_{max})^\alpha) & \text{if } x < x_{max} \\ 1 & \text{c.c.} \end{cases} \tag{10.14}$$

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

onde x_{max} e α são hiper-parâmetros.

O resultado do modelo é um conjunto de representações vetoriais de palavras que capturam as relações entre as palavras em termos semânticos e sintáticos. Esses vetores podem ser usados em diversas tarefas de PLN.

O GloVe tem capacidade para capturar informações semânticas mais abrangentes em comparação aos outros métodos de representação vetorial de palavras, devido à abordagem de coocorrência global ponderada.

10.3.3.1 Limitações do Glove

Apesar das vantagens do GloVe em relação às tarefas de analogia e similaridade, cabe destacar algumas limitações do método. Em virtude de seu treinamento necessitar de uma matriz de contagem de coocorrência palavra-palavra, o modelo consome muito espaço de memória; além disso, assim como o Word2Vec, o GloVe não consegue lidar com palavras fora do vocabulário e ignora a morfologia das palavras.

Em síntese, GloVe é um modelo de representação vetorial de palavras que utiliza informações de coocorrência global entre palavras no *corpus* e estima os parâmetros do modelo por meio de regressão log-bilinear. Essa abordagem visa capturar relações mais abrangentes entre as palavras e obter vetores de palavras mais informativos e precisos.

10.4 Considerações finais

As representações vetoriais geradas pelos modelos Word2Vec, Fasttext e GloVe têm uma dimensionalidade menor em comparação com os vetores esparsos baseados em abordagens como TF-IDF e PMI/PMMI, o que ajuda a reduzir o custo computacional e a dimensionalidade dos dados. Outra vantagem dos *embeddings* é que eles podem ser transferidos e usados como recursos em tarefas de aprendizado de máquina relacionadas, melhorando o desempenho de modelos em tarefas de PLN com conjuntos de dados menores.

Embora os *embeddings* apresentam algumas vantagens em relação aos vetores esparsos, nos *embeddings* estáticos cada palavra tem uma única representação, independentemente do contexto em que a palavra ocorre. Isso limita a capacidade de compreender a ambiguidade nas palavras. Além disso, por se tratarem de representações estáticas, ou seja, que não são atualizadas durante o treinamento da tarefa na qual são empregados, eles podem não se adaptar bem a tarefas específicas, especialmente aquelas que exigem informações contextuais precisas como é o caso de tarefas de análise de sentimentos, de resolução de ambiguidades e de tradução automática. Há que se considerar, ainda, que a língua natural está em constante evolução, e o significado das palavras pode mudar ao longo do tempo. Os *embeddings* estáticos não conseguem capturar essas mudanças.

Para tentar contornar esses problemas dos *embeddings* estáticos, permitindo o ajuste do modelo durante o treinamento da tarefa específica, várias abordagens baseadas em ***embeddings* contextualizados** têm sido estudadas na literatura recentemente. Ao contrário dos *embeddings* estáticos, os *embeddings* contextualizados se adaptam ao vocabulário e às características específicas da tarefa, capturando informações contextuais, pois consideram o contexto em que cada palavra é usada em um documento ou sentença. Isso permite lidar melhor com palavras polissêmicas, tornando os *embeddings* mais adequados para tarefas que dependem do contexto.

Os modelos de *embeddings* contextualizados são o assunto do Capítulo Modelos de linguagem.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Parte VI

Discurso

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Capítulo 11

Modelos discursivos

Paula Christina Figueira Cardoso
Jackson Wilke da Cruz Souza
Roana Rodrigues

Publicado em: 26/09/2023

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

11.1 Introdução

No Dicionário Houaiss[1], discurso pode referir-se à "língua em ação, tal como é realizada pelo falante; a um segmento contínuo de fala maior do que uma sentença (Análise de discurso); a um enunciado oral ou escrito que supõe, numa situação de comunicação, um locutor e um interlocutor"; e ainda à "reprodução que alguém faz das palavras atribuídas a outra pessoa". Diante das possibilidades de definir o que é discurso, nos parece pertinente pontuar quais os limites e o objeto de estudo do nível discursivo para a Linguística e, mais especificamente, para o PLN.

Segundo Barros (2021), na Linguística há diferentes perspectivas teórico-metodológicas para o estudo do *texto* e do *discurso*, porém todas coincidem no fato de considerarem que a análise discursiva "vai além da dimensão da palavra ou da frase, e se preocupa com a organização global do texto; examina as relações entre a enunciação e o discurso enunciado e entre o discurso enunciado e os fatores sócio-históricos que o constroem". Salientamos que *texto* e *discurso* tendem a ser entendidos como elementos que se complementam. Segundo Lyons (1977), o texto se dá por meio do *discurso*, em que aquele seria qualquer passagem que apresenta a conexão do discurso, falado ou escrito, em um diálogo ou um monólogo.

Por sua vez, no PLN há uma tendência a definir *discurso* como "qualquer segmento conexo de texto ou fala, compreendendo uma ou mais frases ou segmento de frases" (Sidner, 1978). Essa parece ser uma definição bastante genérica, mas que conduz as pesquisas da área a tomarem *texto* e *discurso* como sinônimos. Diversos estudos discursivos em PLN trabalham com textos de diversos gêneros (como redações escolares, textos jornalísticos ou postagens em redes sociais) e tamanhos variados. Assim, a definição proposta por Sidner (1978) nos parece pertinente por não ter concebido *discurso* a partir de uma porção encadeada de duas ou mais sentenças, mas a partir da possibilidade de observação de questões que extrapolam os limites da materialidade e que não têm como fator limitante o tamanho. Ainda sob a perspectiva do PLN, Mitkov (2010) enfatiza que o discurso produzido não é uma mera coleção aleatória de símbolos ou palavras, mas se trata de elementos relacionados e significativos que têm um objetivo comunicativo particular.

Sendo assim, podemos afirmar que, em nível discursivo, uma preocupação comum à Linguística e, em especial, aos estudos de PLN está na relação entre os elementos de um

[1]Dicionário Houaiss, disponível em: https://houaiss.uol.com.br/.

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

texto, podendo-se, de antemão, depreender que a produção de um texto em si pressupõe um processo de interação e de intenções entre os sujeitos envolvidos em uma determinada situação comunicativa. De acordo com Oliveira (2008), podemos organizar as relações textuais em duas grandes áreas: *coesão* e *coerência*, que, para a autora, são, na verdade, faces de uma mesma moeda.

Segundo Koch (2003), é possível definir *coesão* como "o fenômeno que diz respeito ao modo como os elementos linguísticos presentes na superfície textual se encontram interligados entre si, por meio de recursos também linguísticos, formando sequências veiculadoras de sentido". A autora ainda destaca duas modalidades de coesão[2]: a remissão (reativação de referentes por anáfora, catáfora ou sinalização) e a sequenciação (elementos responsáveis pelo avanço e a continuidade dos sentidos do texto). Por sua parte, *coerência* refere-se "ao modo como os elementos subjacentes à superfície textual vêm a construir, na mente dos interlocutores, uma configuração veiculadora de sentidos" (Koch, 2003, p. 52). A coerência resulta da construção feita pelos interlocutores, por isso, embora parta do texto, envolve uma série de fatores de caráter cognitivo, interacional, situacional e sociocultural. A superfície do texto, conforme ressalta Koch (2003, p. 53), "funciona como pistas ou chaves para orientar o interlocutor na construção do sentido". Pardo (2005, p. 1) explica o fenômeno da coerência textual nos exemplos de Exemplo 11.1:

Exemplo 11.1:

a) Embora tenha chovido, as obras continuaram.
b) João não foi à aula, mas estava doente.

Segundo o autor, apenas o trecho (1a) é coerente, por apresentar um sentido global marcado por uma relação de *oposição* entre as proposições. O trecho (1b), por sua parte, é incoerente, pois "a relação de *oposição* [marcada nesse caso pela conjunção adversativa *mas*] contraria a relação de*causa* que parece mais plausível" (Pardo, 2005). Portanto, é no nível do discurso que um escritor/falante organiza e relaciona as proposições para a produção de um texto com determinados objetivos comunicativos, buscando, assim, satisfazer as suas intenções comunicativas, como persuadir, informar ou pedir algo ao seu leitor/ouvinte.

As relações estabelecidas entre os elementos no interior de um texto para a construção de sentido são bastante complexas, inclusive para a interpretação humana. Por isso, verifica-se a teorização, anotação e o processamento de dados discursivos como grandes desafios para o PLN. Com base nisso, neste capítulo, não temos a pretensão de findar as discussões sobre o nível discursivo; pelo contrário, nosso objetivo é apresentar um panorama sobre modelos discursivos que vêm sendo utilizados em pesquisas nas (sub)áreas de PLN, além de destacarmos tarefas desenvolvidas e consolidadas a partir desses modelos.

Para tanto, este capítulo se organiza da seguinte maneira: na Seção 11.2, apresentamos fundamentações teóricas gerais sobre modelos de relações discursivas, exemplificando suas preocupações e potenciais aplicações por meio das teorias GSDT, SDRT, Teoria de Centering e Teoria das Veias. Na Seção 11.2.1 e na Seção 11.2.2, em contrapartida, descrevemos com algum aprofundamento dois modelos discursivos bastante relevantes nos estudos de PLN no mundo e no Brasil: a *Rhetorical Structure Theory* (RST) e a *Cross-document Structure Theory* (CST). Na Seção 11.3, apresentamos os principais recursos disponíveis

[2]Os conceitos de coesão são definidos e discutidos no Capítulo 12.

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

e aplicações em PLN que utilizaram modelos discursivos para sua constituição e/ou realização. Em Considerações Finais (Seção 11.4), descrevemos algumas limitações, desafios e conquistas da área.

11.2 Modelos de relações discursivas

Ao longo da exposição desta seção, poderá ficar a impressão de que alguns modelos são mais detalhados que outros. Isso se deve ao fato de que muitos deles não têm sido vastamente utilizados nos últimos anos, especialmente por conta do excelente desempenho que alguns métodos estatísticos e modelagens computacionais recentes vêm apresentando na área de PLN e Inteligência Artificial. Apesar de alguns modelos apresentarem essa questão, eles estão presentes nesta seção devido à aderência a aplicações e desenvolvimento de recursos para o PLN, ou mesmo por terem servido como ponto de partida teórico para outros modelos. Há modelos clássicos que buscam tratar diversos fenômenos discursivos, também nomeados retóricos. A título de exemplo, mencionamos, inicialmente e de maneira concisa, as contribuições da GSDT, SDRT, da Teoria de Centering e da Teoria das Veias.

A teoria de Grosz; Sidner (1986), conhecida como **GSDT** (*Grosz and Sidner Discourse Theory*), visa modelar o aspecto intencional do discurso. Parte-se da ideia de que o autor de um texto possui uma ou mais intenções e estrutura seu conteúdo de forma a satisfazê-las. Identificar as intenções do autor é crucial para compreender a mensagem pretendida. Como as intenções potenciais em um discurso são praticamente ilimitadas, a GSDT organiza-o usando relações de contribuição e satisfação entre as intenções. Essas relações são em número finito e limitadas a dois tipos: a *intenção primária* do discurso e as *intenções subjacentes* aos segmentos do discurso. Define-se, nesta teoria, as seguintes relações: *Dominance*, *Satisfaction-Precedence*, *Supports* e *Generates*.

A relação *Dominance* ocorre quando a intenção subjacente a um segmento A contribui para a intenção subjacente de um segmento B, isto é, A dominates B, representado por (DOM(A,B)). A relação *Satisfaction-Precedence* ocorre quando a intenção subjacente a um segmento A deve ser satisfeita antes da intenção subjacente a um segmento B, isto é, SP(A,B). As relações *Supports* e *Generates* ocorrem entre o conteúdo dos segmentos. A primeira acontece se a aceitação de um segmento B fornece subsídios para a aceitação do segmento A, então se diz que o conteúdo de B supports A (SUP(A,B)). A segunda ocorre se a ação descrita em B contribui para a ação descrita em um segmento A (GEN(B,A)). No Exemplo 11.2, extraído de Maziero (2016, p. 14), ilustra-se tais relações.

Exemplo 11.2:

> a) A teoria XYZ é bem informativa para muitas tarefas de PLN que requerem conhecimento discursivo, e conta com diversos *parsers* disponíveis.
> b) Seu uso, portanto, é uma ótima alternativa quando se deseja automatizar totalmente uma tarefa de PLN.

Segundo Maziero (2016), no exemplo anterior, a intenção do autor do texto é persuadir o leitor que o uso da XYZ é uma ótima alternativa no campo do PLN (2b), argumentando a favor do modelo da primeira sentença. Podemos dizer, portanto, que há uma relação de DOM (2b, 2a) e SUP (2a, 2b). A teoria não visa explicitar qual a intenção do autor do texto, mas estabelece conexões entre as intenções, além de abordar questões como os focos de atenção e a estrutura linguística.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

A teoria **SDRT** (Asher; Lascarides, 2003) – Teoria da Representação do Discurso Segmentado – se interessa em identificar os segmentos discursivos e as relações retóricas entre essas unidades, que podem ser classificadas em dois tipos básicos. Uma análise SDRT abrange todas as etapas do processamento do discurso, incluindo segmentação, identificação de relacionamentos e construção de hierarquias, usando informações semânticas e pragmáticas.

O discurso é representado como um hipergrafo, no qual as arestas são as relações discursivas e os nós representam as Unidades de Discurso Elementar (EDUs) que contém apenas um elemento. O grafo pode ter ainda Unidades de Discurso Complexas (CDUs) que são nós com mais de um elemento simples. As unidades discursivas são conectadas por relações retóricas de *coordenação* ou de *subordinação*. As relações de coordenação conectam segmentos do discurso no mesmo nível hierárquico, enquanto as relações de subordinação ligam um segmento do discurso a outro segmento que está um nível hierárquico abaixo. Asher; Vieu (2005) afirmam que essa distinção (no nível do discurso) possui uma motivação intuitiva, na qual certas partes do texto desempenham um papel subordinativo (menos relevante) em relação às demais. É importante ressaltar que o conjunto de relações, sejam de coordenação ou subordinação, não é fechado, pois estudos recentes já apresentam variações do conjunto original (por exemplo, (Muller et al., 2012)).

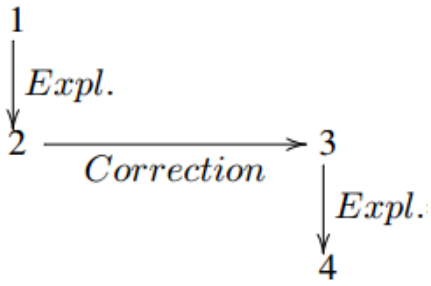
Esta teoria foi bastante explorada para modelar diálogos, pois permite representar contra-argumentação, um fenômeno pouco tratado em outros modelos discursivos (Afantenos; Asher, 2014; Asher et al., 2016; Badene et al., 2019; Li et al., 2020). Portanto, a teoria SDRT possui mecanismos que podem ser aplicados ao tratamento de diálogos, tais como *Question Elaborating*, *Correction* e *Question Answer Pair*. No Exemplo 11.3, Afantenos; Asher (2014) exemplificam um diálogo:

Exemplo 11.3:

a) [Maria irá falhar em seus exames.]^1 [Ela não estudou muito.]^2
b) [Não, ela estudou muito.]^3 [Agora ela tem até olheiras.]^4

Em (3b), o falante não questiona seu interlocutor sobre sua conclusão (EDU 1), mas expressa discordância em relação à veracidade subjacente àquela conclusão. Isso assume a forma de uma relação de *Correction* entre a EDU 2 do primeiro falante, em (3a), representando o motivo e o contra-argumento do segundo falante. O falante fornece uma razão adicional para suas crenças por meio de uma relação de *Explanation*. Na Figura 11.1, tem-se a representação na forma de gráfico para esse diálogo.

Figura 11.1: Exemplo para relação Contraste – SDRT



Outra preocupação em nível discursivo é a resolução anafórica, elemento fundamental para o estabelecimento das relações de correferência de um texto. A **Teoria de Cen-**

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

tering (Grosz et al., 1995), foca nas relações existentes entre anáforas e visa estabelecer a coerência nos segmentos discursivos adjacentes ao direcionar a atenção para a escolha de uma expressão referencial (discurso local). O principal objetivo da teoria é prever qual entidade discursiva tem maior importância em determinados segmentos, definindo um conjunto de regras e restrições que ditam as escolhas feitas pelos participantes do discurso, como demonstrado em Exemplo 11.4 e Exemplo 11.5, a seguir, em que a Teoria de Centering fornece meios para tratar essas diferenças.

Exemplo 11.4:

a) João foi a sua loja de música favorita para comprar um piano.
b) Ele havia frequentado a loja por vários anos.
c) Estava excitado porque iria finalmente poder comprar um piano.
d) Mas quando chegou, a loja estava fechada.

Exemplo 11.5:

(a) João foi a sua loja de música favorita para comprar um piano.
(b) Esta era a loja que João frequentou por vários anos.
(c) Ele estava excitado porque iria finalmente poder comprar um piano.
(d) Ela estava fechada quando João chegou.

Nos exemplos, adaptados de Grosz et al. (1995), tem-se que os dois textos expressam a mesma ideia, mas no Exemplo 11.4 "João" é a unidade central enquanto que no Exemplo 11.5, o foco é alternado entre "João" e a "loja de música". Percebe-se que as escolhas dos participantes podem variar desde a seleção da estrutura sintática (como em (4d) e (5d) que usam estruturas diferentes para tratar sobre o fato de a loja estar fechada)em até a escolha de expressões referenciais (como o uso de "a loja", em (4b) e "esta" em (5b) ao tratar do mesmo referente).

Ainda na linha de tratamento de anáforas, há a **Teoria das Veias** (*Veins Theory*), proposta por Cristea et al. (1998) , que sugere o estabelecimento de domínios referenciais de acessibilidade para cada unidade discursiva, representado pelas "veias" definidas na RST[3]. A Teoria das Veias expande as regras de coerência local da Teoria de Centering para abranger a composicionalidade das unidades do discurso (Seno, 2005). A veia de uma unidade é definida como um conjunto de unidades do discurso que podem conter o antecedente de uma anáfora. Para manter a coerência, é fundamental que o antecedente e o termo anafórico estejam presentes no mesmo veio, contribuindo para o discurso global.

No exemplo Exemplo 11.6, extraído de Seno (2005), as unidades 1 e 3 são ditas relevantes. Assim, o antecedente da anáfora "a fábrica" da unidade 4 pode estar presente em uma das unidades 1 e 3. No exemplo, seu antecedente encontra-se em 1.

Exemplo 11.6:

[1] **A empresa Produtos Pirata Indústria e Comércio Ltda., de Contagem** [2] (na região metropolitana de Belo Horizonte), [3] deverá registrar este ano um crescimento de produtividade nas suas áreas comercial e industrial de 11% e 17%, respectivamente.
[4] Os ganhos são atribuídos pela diretoria d**a fábrica** à nova filosofia.

[3] A ser detalhada na próxima seção.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Os modelos discursivos podem destacar estruturas linguísticas, intencionais, informacionais ou de foco, todos com a principal preocupação de apresentar as relações entre os elementos de um texto para depreender a sua produção e os processos de interação e intenções pertencentes a uma situação comunicativa específica. Nesta seção foram apresentadas brevemente bases teóricas dos modelos: GSDT, SDRT, da Teoria de Centering e da Teoria das Veias. Conforme já explicitado, embora existam vários modelos de análise discursiva que partem de reflexões linguísticas e possibilitam aplicações computacionais, nos deteremos, nas próximas seções, à descrição aprofundada de dois modelos discursivos: a RST e a CST, devido à sua relevância no cenário brasileiro.

11.2.1 O modelo RST (Mann e Thompson, 1988)

A **Rhetorical Structure Theory** (RST) é uma teoria linguístico-descritiva que trata da organização do texto utilizando relações retóricas (também nomeadas relações de coerência ou discurso) que existem entre os segmentos discursivos, formando uma estrutura discursiva totalmente conectada, geralmente na forma de árvore (Mann; Thompson, 1988). A RST explica a coerência postulando uma estrutura hierárquica e conectada, na qual cada parte de um texto tem uma função a cumprir, com relação às outras partes do texto (Taboada; Mann, 2006).

Cada proposição é associada a um núcleo (informação principal) ou satélite (informação adicional) de uma relação retórica. Em casos padrões, as relações se estabelecem entre duas proposições, expressas por segmentos adjacentes no texto. Quando a relação conecta um núcleo e um satélite, ela é chamada de mononuclear. Por outro lado, se a relação conectar somente núcleos, ela é chamada de multinuclear.

Mann; Thompson (1988) estabeleceram um conjunto de 23 relações retóricas que podem ser aplicadas a uma grande variedade de textos. Nesse conjunto, cada relação é classificada em semântica (*subject-matter*) ou intencional (*presentational*). As relações semânticas são aquelas que informam o leitor sobre algo, por exemplo, a relação SEQUENCE, cujo efeito pretendido é que o leitor reconheça que há uma sucessão temporal dos eventos apresentados. As relações intencionais alteram a inclinação do leitor para algo, por exemplo, a relação JUSTIFY, cujo efeito pretendido é que o leitor passe a aceitar melhor o direito do escritor de apresentar o núcleo. Vários pesquisadores modificaram e/ou complementaram esse conjunto de relações, como Marcu (1997) e Pardo (2005). No Quadro 11.1 apresenta-se o conjunto de relações de Mann; Thompson (1988) e o tipo de cada relação. Quanto à nuclearidade, as relações multinucleares estão marcadas com um asterisco.

Quadro 11.1: Relações RST

Relações semânticas	Circumstance, Solutionhood, Elaboration, Background, Volitional cause, Non-volitional cause, Volitional result, Non-volitional result, Purpose, Condition, Otherwise, Intepretation, Evaluation, Restatement, Summary, Sequence*, Contrast*
Relações intencionais	Antithesis, Concession, Enablement, Evidence, Justify, Motivation

Conforme se observa no Quadro 11.1, Mann; Thompson (1988) definiram as relações em termos de quatro campos, que devem ser observados pelo analista de um texto durante o

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

processo de construção da estrutura RST. Os campos são restrições sobre o núcleo (N), restrições sobre o satélite (S), restrições sobre a combinação de núcleo e satélite e o efeito que a relação em questão pode causar no leitor. Nos Quadros 11.2 e 11.3, apresentam-se as definições das relações *Antithesis* e *Contrast*, respectivamente.

Quadro 11.2: Definição da relação Antithesis

Anthithesis	**Restrições sobre N**: O escritor julga N válido **Restrições sobre S:** Nenhuma **Restrições sobre N+S:** N e S se contrastam e, por esse motivo, não podem ser válidos simultaneamente. Compreendendo-se S inválido levará o leitor a aceitar melhor N **Efeito no receptor:** O leitor aceita melhor N

Quadro 11.3: Definição da relação Contraste

Contrast	**Restrições sobre Ns**: Não mais do que dois Ns; as situações nos Ns são (a) compreendidas como similares em vários aspectos, (b) compreendidas como diferentes em vários aspectos e (c) comparadas em relação a uma ou mais dessas diferenças **Efeito no receptor:** O leitor reconhece as similaridades e diferenças resultantes da comparação sendo feita.

Um grande desafio encontrado na análise RST é a definição da relação retórica entre dois segmentos textuais. Se esse contexto for expandido para um texto inteiro, há diversas possíveis árvores discursivas para um mesmo texto, com segmentos, relações e nuclearidades diferentes. Por exemplo, um analista RST pode identificar que há uma oposição entre duas unidades discursivas e, assim, relações como *Antithesis* e *Concession* poderiam ser úteis na análise, gerando diferentes árvores discursivas. Para amenizar tal situação, o analista deve olhar para o campo efeito da definição das possíveis relações e identificar aquele que está mais saliente para o objetivo do autor do texto (Mann; Thompson, 1988).

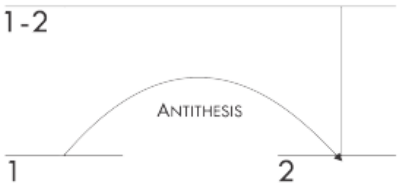
Na Figura 11.2, apresenta-se um exemplo da relação mononuclear *Antithesis*, em que os segmentos 1 e 2 não podem ser válidos ao mesmo tempo, pois, ou a "detonação" foi "acidental" ou "proposital". O segmento 2 é nuclear. Para que a crença do leitor no segmento 2 seja melhor aceita, o segmento 1 deve ser inválido. Na Figura 11.3 exemplifica-se a relação multinuclear *Contrast*.

Para fazer a análise RST de um texto, várias estratégias podem ser utilizadas. Carlson; Marcu (2001) apontam que uma estratégia bem aceita é fazer uma análise incremental, isto é, relacionar primeiro as proposições de uma sentença, o que resultará em uma subestrutura RST, a qual, por sua vez, será relacionada à outra subestrutura. Podem-se montar subestruturas de cada parágrafo do texto isoladamente e depois integrá-las, formando uma única estrutura RST. Se o analista decide por esse tipo de análise, ele pode tirar proveito da estrutura organizacional dada pelo produtor do texto. Por exemplo, se duas proposições estão diretamente relacionadas por *Condition*, é provável que elas sejam expressas em uma única sentença.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Figura 11.2: Exemplo para relação Antithesis



1. POUCO DEPOIS DA DETONAÇÃO, OCORRIDA ÀS 10H40 (3H40 EM BRASÍLIA), FONTES DA POLÍCIA MOSCOVITA ADIANTARAM QUE ELA TERIA ACONTECIDO PROVAVELMENTE POR CAUSA DA EXPLOSÃO ACIDENTAL DE UM BUJÃO DE GÁS.
2. - ISTO NÃO FOI UM ACIDENTE, FOI PROPOSITAL - DISSE RESIN, CITADO PELA AGÊNCIA OFICIAL RUSSA "ITAR-TASS".

Figura 11.3: Exemplo para relação Contraste



11.2.2 CST: Cross-document Structure Theory, 2000

A **Cross-document Structure Theory** (CST) é um modelo teórico derivado da RST, diferenciando-se acerca da quantidade de textos que podem ser analisados e, consequentemente, dos fenômenos linguísticos que ocorrem. Esta teoria foi proposta por Radev (2000), com o objetivo de realizar análises semânticas de múltiplos textos que possuem o mesmo assunto. O autor percebeu que, quando se agrupa textos que possuem o mesmo assunto, fenômenos linguísticos (como redundância, contradição e complementaridade), de estilo e de organização. Em sua proposta, o modelo CST, então, é capaz de traduzir cada fenômeno em diferentes relações, como demonstrado nos exemplos 11.7, 11.8 e 11.9, a seguir.

Exemplo 11.7:

> (S1) Todos morreram quando o avião, prejudicado pelo mau tempo, não conseguiu chegar à pista de aterrissagem e caiu numa floresta a 15 quilômetros do aeroporto de Bukavu.
> (S2) Todos morreram quando o avião, prejudicado pelo mau tempo, não conseguiu chegar à pista de aterrissagem e caiu numa floresta a 15 Km do aeroporto de Bukavu.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Exemplo 11.8:

> (S1) Ao menos 17 pessoas morreram após a queda de um avião de passageiros na República Democrática do Congo.
> (S2) O avião explodiu e se incendiou, acrescentou o porta-voz da ONU em Kinshasa, Jean Tobias Okala.

Exemplo 11.9:

> (S1) Todos morreram quando o avião, prejudicado pelo mau tempo, não conseguiu chegar à pista de aterrissagem e caiu numa floresta a 15 Km do aeroporto de Bukavu.
> (S2) A aeronave se chocou com uma montanha e caiu, em chamas, sobre uma floresta a 15 quilômetros de distância da pista do aeroporto.

As sentenças (S1 e S2) foram retiradas do *corpus* CSTNews (Cardoso et al., 2011), que contém conjuntos de textos escritos em português brasileiro e anotados com o modelo CST. Cada uma das sentenças foram extraídas de fontes jornalísticas distintas, e associadas manualmente em função dos fenômenos identificados. No Exemplo 11.7 existe uma relação de *redundância*, uma vez que o par de sentenças apresenta um conteúdo praticamente idêntico. Já no Exemplo 11.8, há uma relação de *complementaridade*, pois S2 acrescenta que o avião acidentado "explodiu e se incendiou", em relação à informação em S1. Por fim, no Exemplo 11.9 observa-se a presença de *contradição*, porque S1 informa que a causa da queda do avião foi o mau tempo, enquanto S2 destaca que a causa do acidente foi o choque contra a montanha.

Os modelos apresentados neste capítulo foram caracterizados como discursivos, pois focalizam aspectos e fenômenos de apenas um texto. Dessa maneira, é necessário destacar que o modelo CST é especialmente caracterizado por sua abordagem semântica, já que é possível identificar, como demonstrado, uma série de fenômenos linguísticos, de estilo e de estrutura. Porém, tais fenômenos se dão de maneira não intencional, diferentemente dos outros modelos em que os fenômenos ocorrem por intencionalidade de quem elabora o texto e, nesse sentido, o estrutura e o organiza de determinada forma.

A contradição no Exemplo 11.9, por exemplo, só foi possível de ser identificada porque dois textos foram agrupados e, de maneira manual e/ou automática, foi identificado o fenômeno (não intencional) em questão. Assim, o que justifica a ocorrência do modelo CST neste capítulo é o fato de ele ocorrer na relação entre textos. Nesse sentido, se dá de maneira discursiva, ainda que caminhe nas margens de uma definição clássica de discurso para o PLN.

De acordo com Radev (2000), as relações CST podem ocorrer entre diferentes unidades informativas, tais como, palavras, sintagmas, sentenças, parágrafos e documentos, formando um grafo, como ilustrado na Figura 11.4.

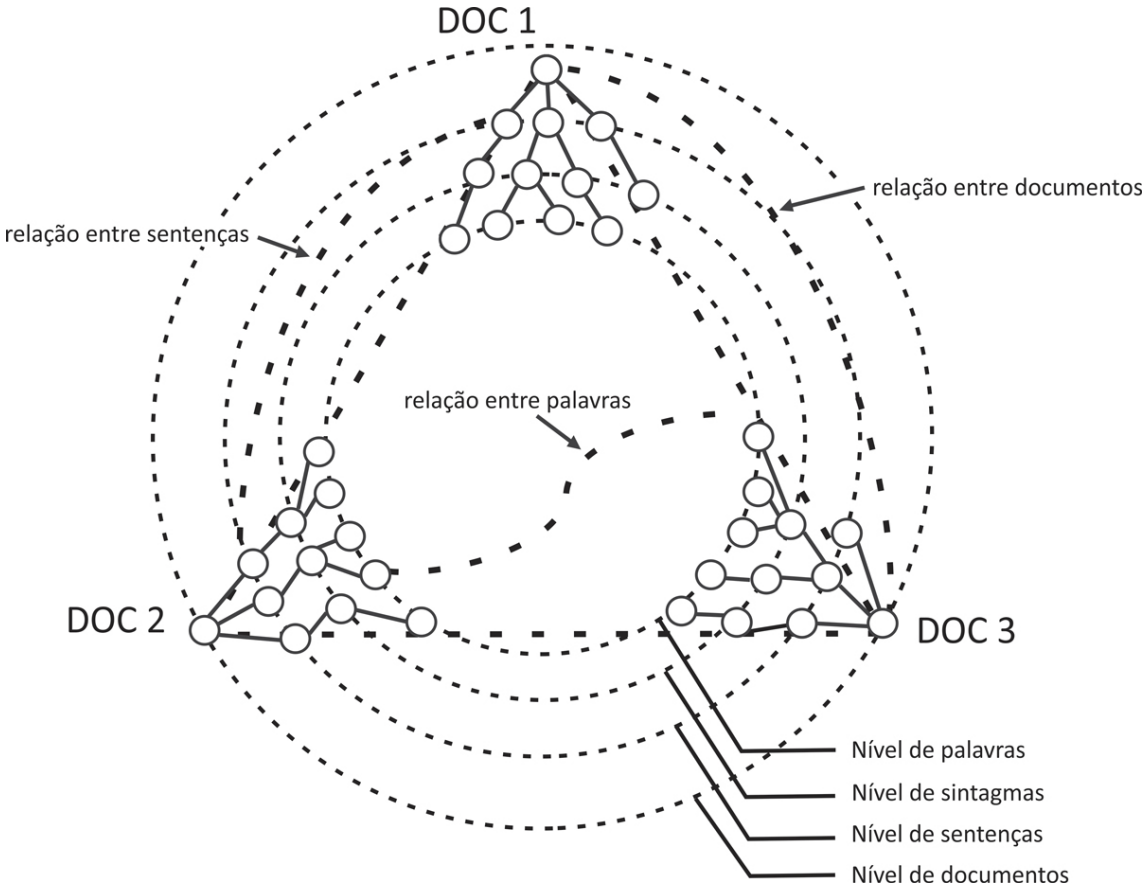
Na Figura 11.4, percebe-se que os níveis nos quais as relações CST podem ser identificadas compõem uma hierarquia (palavras → sintagma → sentença → texto), ainda que usualmente isso seja feito em nível sentencial. Cada um dos três documentos (DOC 1, DOC 2 e DOC 3) está representado por um subgrafo, que codifica relações internas aos textos. Os relacionamentos internos a cada texto podem ser caracterizados em nível sintático ou discursivo. As relações CST que podem ser estabelecidas nos diferentes níveis estão representadas por linhas pontilhadas mais grossas.

Ainda sobre a Figura 11.4, destaca-se que:

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Figura 11.4: Esquema de relacionamentos CST



- Os documentos similares são representados numa hierarquia de palavras, sintagmas, sentenças e os próprios documentos, ou seja, todos esses níveis são considerados na análise;
- Em cada nível da hierarquia podem ocorrer relações CST, apesar de sentenças serem usualmente mais utilizadas nos trabalhos da área;
- O grafo resultante da anotação é provavelmente desconectado, pois nem todos os segmentos dos textos em análise precisam estar relacionados: podem existir segmentos que não se referem diretamente ao mesmo assunto.

Para o inglês, originalmente foram propostas 24 relações CST por Radev (2000). Uma vez que o modelo CST admite haver ambiguidade entre as relações, é natural ter novas propostas de conjuntos de relações. Além disso, determinadas relações podem não ocorrer em certos *corpora* com gêneros textuais específicos. Destacam-se as relações propostas para o modelo CST aplicado ao português brasileiro. Aleixo; Pardo (2008a) chegaram a um conjunto de 14 relações multidocumento. Segundo os autores, a redução justifica-se pela não ocorrência de algumas relações em textos jornalísticos ou ainda por conta da similaridade entre algumas relações, o que resultou no agrupamento de algumas delas, como é o caso de *Equivalence* e *Paraphrase* em apenas *Equivalence*, ou *Elaboration* e *Refinement* em *Elaboration*. No Quadro 11.4, mostra-se o conjunto de relações CST utilizado no *corpus* CSTNews.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Quadro 11.4: Relações CST

Attribution	Identitty
Citation	Indirect speech
Contradiction	Modality
Elaboration	Overlap
Equivalence	Subsumption
Follow-up	Summary
Historical background	Translation

Maziero et al. (2010) propuseram uma tipologia em que as relações CST para o português brasileiro estão categorizadas entre Redundância, Complemento, Contradição, Fonte/Autoria e Estilo. É possível inferir que essa proposta seja, na verdade, uma simplificação do modelo discursivo com foco na implementação computacional, em especial, além ter contribuído com a área de PLN na compreensão de fenômenos linguísticos no contexto de Sumarização Automática Multidocumento.

Mais recentemente, alguns estudos descritivos (Souza, 2015; Souza, 2019) apontaram que a organização de relações CST entre Conteúdo e de Apresentação/Forma pode não ser suficiente para caracterizar as relações CST, em especial as relações classificadas, até então, como complementaridade. Tais estudos indicam que algumas relações de redundância (como *Subsumption* e *Overlap*) poderiam ser classificadas como relações de complementaridade, por apresentarem outras informações acerca do mesmo evento.

Vale destacar que o modelo CST contribuiu com a criação de recursos e, consequentemente, em aplicações de PLN. Na próxima seção, destacamos alguns deles, entre os outros modelos apresentados aqui.

11.3 Recursos e aplicações para o português brasileiro

A descrição dos fenômenos linguísticos em nível discursivo, a partir dos diferentes modelos de análise, como os descritos neste capítulo, contribuiu para importantes avanços de diversas aplicações de PLN. Freitas (2022) escreve que para que tais aplicações sejam bem-sucedidas, uma série de recursos e ferramentas linguístico-computacionais é acionada. Assim, destaca-se a criação de *corpus* como recurso anotado no nível discursivo, de ferramentas que facilitam a anotação automática de dados e de diversas aplicações, como de sumarização (Cardoso, 2014; Uzêda et al., 2010), tradução automática (Marcu et al., 2000) e avaliação de redações (Stab et al., 2014).

Na literatura, são encontrados pelo menos dois *corpora* padrão ouro com relações discursivas para o português brasileiro: Summ-it[4] (Collovini et al., 2007; Fonseca et al., 2016a) e CSTNews[5] (Aleixo; Pardo, 2008b; Cardoso et al., 2011). O *corpus* Summ-it reúne anotações de vários níveis linguísticos, incluindo relações retóricas da RST, correferência e entidades nomeadas. Esse recurso, concebido para promover pesquisas em discurso e sumarização automática, constitui-se de 50 textos jornalísticos do caderno de Ciências da Folha de São Paulo.

O *corpus* CSTNews, por sua vez, contém 50 grupos de textos jornalísticos de assuntos variados, coletados manualmente das fontes de notícias Folha de São Paulo, Estadão, O

[4] https://www.inf.pucrs.br/linatural/wordpress/recursos-e-ferramentas/summ-it/
[5] http://nilc.icmc.usp.br/CSTNews/login/about

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Globo, Jornal do Brasil e Gazeta do Povo. Assim como o *corpus* Summ-it, CSTNews foi orientado para a sumarização automática, sendo constituído de diversas camadas de anotação, tais como RST e CST, e sumários manuais e automáticos.

A anotação no nível discursivo de textos pode ser feita de forma manual ou automática. Para alguns modelos discursivos existem analisadores automáticos, conhecidos como *parsers* discursivos, que visam a identificação retórica do texto, gerando uma estrutura hierárquica em que as intenções do autor são explicitadas e relacionadas entre si (Maziero, 2016). Para relações RST, se tem conhecimento do *parser* DiZer[6] (Maziero et al., 2015; Maziero, 2016). Treinado com textos acadêmicos e jornalísticos, a ferramenta recebe um texto de entrada, segmenta-o, identifica a nuclearidade e monta a estrutura arbórea com as relações discursivas.

Com a finalidade de facilitar o processo de anotação de *corpus* com CST, foi desenvolvida a ferramenta semiautomática CSTTool[7] (Aleixo; Pardo, 2008a). A CSTTool possibilita os processos de segmentação dos textos-fonte em nível sentencial e a identificação, em pares, das sentenças lexicalmente relacionadas por meio de medidas de similaridade. Após a indicação dos possíveis pares relacionados, cabe ao anotador escolher uma relação CST adequada. Após a indicação dos possíveis pares relacionados, cabe ao anotador escolher uma relação CST adequada. Para uma análise totalmente automática, está disponível o CSTParser[8] (Maziero; Pardo, 2012), que recebe como entrada um conjunto de documentos relacionados e segmenta-os em sentenças. Após isso, busca os pares de sentenças mais prováveis de terem algum relacionamento multidocumento por meio de medidas de similaridade.

11.4 Considerações finais

Lidar com o nível discursivo é um desafio para os estudos em PLN, como já havia sido sinalizado por Dias-da-Silva (1996), um dos pioneiros na área no Brasil. O autor já destacava algumas questões relativas esse nível de análise linguística, como a necessidade de delimitar o objeto de estudo, determinar os limites entre análise textual e discursiva, ou ainda caracterizar o discuso como um processo. Felizmente, algumas dessas perguntas já foram respondidas, como a definição do objeto de estudo. No entanto, outras questões ainda estão sendo investigadas para encontrar possíveis respostas.

Se, por um lado, ao longo dos últimos anos, percebemos que diferentes tarefas linguístico-computacionais foram sendo demandadas e concebidas discursivamente, como resolução anafórica, por outro, há de se questionar se a análise de sentimentos e emoções, por exemplo, se enquadra no nível discursivo. Como dito anteriormente, o nível discursivo congrega outros níveis de análise linguística e, consequentemente, é esperado que determinados fenômenos sejam fronteiriços com a Morfologia, Sintaxe, Semântica e Pragmática.

Quanto ao questionamento de Dias-da-Silva (1996) sobre a possibilidade de o discurso ser um processo, é possível que as respostas residam em aprimorar modelos discursivos a partir de descrições linguísticas cada vez mais robustas. Ao longo deste capítulo ilustramos modelos discursivos que por vezes nasceram para suprir expectativas teórico-metodológicas de determinadas aplicações em PLN, mas que não se restringiram a elas. Outros modelos, no entanto, ficaram restritos a determinadas aplicações, podendo esse fato ser explicado

[6] https://github.com/egmaziero/dizer3
[7] http://www.nilc.icmc.usp.br/nilc/index.php/tools-and-resources
[8] http://www.nilc.icmc.usp.br/CSTParser/

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

por uma maior dependência de humanos para as fases de treinamento dos modelos. Assim, há ainda um vasto campo de pesquisas e descrições linguísticas a serem realizadas em todos os modelos aqui dispostos.

Sabe-se que o desenvolvimento de tecnologias sofisticadas tem substituído a reflexão e supervisão linguísticas por modelos estatísticos, com métodos não compreensíveis para os seres humanos. No entanto, conforme aponta Freitas (2022), o conhecimento linguístico para o PLN não ficará obsoleto por diversos motivos, entre os quais a autora destaca quatro:

(i) nem todo conhecimento em PLN é voltado para aplicações da indústria, portanto, há pesquisas linguísticas que dependem desse conhecimento para o desenvolvimento de aplicações linguísticas (materiais lexicográficos, didáticos, corretores gramaticais etc.);
(ii) continua sendo necessária ao menos uma amostra do conhecimento humano para as tarefas em PLN, como na construção de *datasets*, em versões iniciais de sistemas e na avaliação do desempenho da máquina);
(iii) é elevado o custo (computacional, financeiro e ambiental) das atividades desenvolvidas com base nos métodos estatísticos, por isso, informações linguísticas possibilitam a economia no processamento em comparação com o uso de dados brutos; e
(iv) desde uma perspectiva filosófica, haver apenas a eficácia – sem compreensão, nem explicação – dos sistemas não é o suficiente, pois a ciência se baseia no paradigma da verdade.

Ao passo que os modelos estatísticos/probabilísticos proporcionam avanços infindáveis a pesquisas em PLN, é necessário pontuar que o fato de não sabermos como a língua(gem) é, de fato, processada em muitos deles isso faz com que muitos desafios ainda perdurem. Muito recentemente, fomos inseridos em abordagens e aplicações que, com apenas um comando, é possível ter conteúdos readequados a exigências estruturais de determinados gêneros textuais (Capítulo ChatGPT, MariTalk e outros agentes de conversação). Para além dos especialistas, isso pode impressionar muitas pessoas, ainda que estejamos lidando substancialmente com a probabilidade de combinações de *tokens*, sem lançar mão ao sentido das unidades que estão sendo articuladas em sentenças ou em textos.

Nesse sentido, talvez o que tenhamos diante de nós sejam outras e mais complexas perguntas que se alinham às reflexões e provocações de Dias-da-Silva (1996): É possível processar o discurso sem precisar de um componente semântico? É possível compreender emoções de um texto sem o componente pragmático? É possível realizar análise discursiva sem análises linguísticas e, consequentemente, sem os modelos discursivos aqui apresentados?

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Capítulo 12

Resolução de correferência

Evandro Fonseca
Aline Aver Vanin
Renata Vieira

Publicado em: 26/09/2023

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

12.1 Introdução

No processo de construção de sentidos na língua em uso, interlocutores negociam o universo de discurso de que falam, escolhendo referir-se a algum, ou a alguns, indivíduo(s) cuja identidade estabelecem e da qual garantem a existência (Neves, 2013). Esses referentes, concretizados no texto por **expressões referenciais**, vão atravessá-lo por inteiro, garantindo unidade temática – isto é, a coerência que constitui um texto (Vieira; Faraco, 2019). Fazer referência a algo ou a alguém no mundo é uma ação intrinsecamente ligada à interação, em que se constituem os objetos de discurso, isto é, entidades que constituem termos das predicações, entidades oriundas de uma construção mental, e não de um mundo real (Neves, 2013).

A construção de referentes se dá por cadeias de texto, redes referenciais construídas pelos objetos de discurso que constituem as marcas da textualidade. Uma cadeia de referência, ou cadeia referencial, corresponde à noção de cadeia anafórica, e cadeia coesiva (Roncarati, 2010). Os elos coesivos em um texto são mecanismos semânticos e léxico-gramaticais essenciais para a tessitura textual. Os referentes se ligam por meio de relações de sentido que formam a base para retomadas em um texto (Roncarati, 2010). Nesse sentido, as cadeias coesivas ligam referentes à(s) sua(s) expressão(ões) referencial(is), em que os fios que tecem o texto são articulados por meio de procedimentos e recursos - o que chamamos de **coesão textual** (Vieira; Faraco, 2019). Trata-se de elementos cujos mecanismos gramaticais coesivos estão em consonância, sejam eles por reiteração (ou retomada), por associação (ou ligações de sentidos entre as palavras presentes), ou por conexão entre as orações (por conectores) (Antunes, 2007), os quais garantem que o texto seja coerente em sua extensão. Então, se em um dado texto temos uma dada entidade, como "Maria", nome próprio, é de se esperar que os seus referentes sejam detectados por relações léxico-semânticas do texto, como, por exemplo: por pronome, "ela", ou por sintagma nominal, "a professora", "a ativista", "a mulher" etc.

12.2 Resolução de Correferência

A Resolução de Correferência a partir de textos é uma tarefa útil e também um dos principais desafios da área de Processamento da Linguagem Natural (PLN). Isso porque essa

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

tarefa depende de diversos níveis de processamento, como análise sintática, morfológica, extração de sintagmas nominais, entre outros. Na literatura, encontramos diversas iniciativas para a língua portuguesa que abordam esse problema, geralmente separados entre a resolução de anáfora (Basso, 2009; Bick, 2010; Ferradeira, 1993; Rocha, 2000; Vieira et al., 2005) e o estudo da correferência nominal (Fonseca, 2014; Fonseca et al., 2014; Fonseca et al., 2016b; Freitas et al., 2009a). Resolução de Correferência consiste em identificar as diferentes formas que uma mesma menção pode assumir em um discurso. Em outras palavras, esse processo consiste em identificar determinados termos e expressões que remetem a uma mesma referência. Na sentença apresentada no Exemplo 12.1 podemos dizer que [*o único país de a União Europeia a não permitir patenteamento de genes*] é uma correferência de [*A França*], da mesma forma que [*A UE*] é uma correferência de [*a União Europeia*]. Agrupando esses termos formamos grupos de menções referenciais, mais conhecidos como cadeias de correferência.

Exemplo 12.1:

> A França resiste como o único país de a União Europeia a não permitir patenteamento de genes. A UE …

Na presente seção, veremos as definições de conceitos fortemente relacionados à tarefa de Resolução de Correferências, tais como os de referentes, entidades nomeadas, sintagmas nominais, entre outras definições.

12.2.1 Referentes

Referentes, ou menções, podem ser definidos como termos os quais usamos para nos referirmos a determinada entidade em um discurso. Em um texto, essas referências podem aparecer como uma entidade nomeada específica ou como parte constituinte de um sintagma nominal.

12.2.1.1 Entidades Nomeadas

Entidades nomeadas, a grosso modo, são elementos que podem ser referenciados por meio de nomes próprios (Jurafsky; Martin, 2023). Esses nomes próprios podem configurar-se em classes específicas, tais como: Pessoa (nomes de pessoas), Organização (nomes de empresas), Local (nomes de lugares), entre outras. Por meio dos exemplos abaixo, podemos identificar diversas entidades nomeadas (ENs), como Banco Nacional de Desenvolvimento Econômico e Social (a), Apple (b), nomes de bandas musicais (c)[1].

a) O Banco Nacional de Desenvolvimento Econômico e Social (BNDES), empresa pública federal, é hoje o principal instrumento de financiamento de longo prazo …
b) A Apple informou que vendeu 5 milhões de iPhone 5 só em um fim de semana …
c) Várias bandas de black metal tiveram influências do punk, tais como Venom, Celtic Frost, Bathory, Sarcófago, Darkthrone, Impaled, Nazarene, Mayhem, Hellhammer, Behemoth, entre outras …

[1] O Capítulo Reconhecimento de Entidades Nomeadas no Domínio Legal apresenta técnicas para o tratamento de entidades nomeadas no domínio do Direito.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

12.2.1.2 Sintagmas Nominais

São unidades formadas por uma ou mais palavras que, juntas, desempenham uma função sintática específica na frase (Capítulo 4). A natureza de um sintagma depende diretamente do elemento que constitui seu núcleo. Neste capítulo, damos foco a menções expressas por sintagmas nominais. Dito isso, temos então os sintagmas nominais, cujos núcleos podem configurar-se em nome comum, nome próprio ou um pronome. Os pronomes podem apresentar-se, basicamente, nas formas de pronome pessoal, demonstrativo, indefinido, possessivo ou relativo. Um sintagma nominal geralmente é composto por um determinante (artigo, pronome demonstrativo, pronome indefinido e numeral cardinal) seguido de um substantivo. Por exemplo, na sentença do Exemplo 12.2 "**O especialista**" é um sintagma nominal, e o artigo "**O**" é seu determinante. Por meio do determinante de um sintagma é possível extrair informações valiosas. Isto é, a palavra "especialista", por si só, pode assumir diferentes papéis. Contudo, "**O especialista**" qualifica uma pessoa do sexo masculino, além de informar quem é o especialista (apenas um, não dois ou mais). Notemos que o determinante carrega informações úteis para o processamento linguístico.

Exemplo 12.2:

> **O especialista** não respondeu todas as perguntas.

Contudo, sintagmas nominais podem configurar-se em apenas substantivos como no Exemplo 12.3 onde temos dois sintagmas nominais sem determinante explícito: "Rio de Janeiro" e "cidade maravilhosa". Respectivamente, um nome próprio e um substantivo comum seguido de seu adjunto adnominal. Algumas vezes, esse adjunto pode ser predicativo.

Exemplo 12.3:

> Rio de Janeiro, cidade maravilhosa

Para entendermos a diferença entre adjunto adnominal e predicativo, basta observarmos que ora um termo pode exercer a função de adjunto, ora de predicativo. Ou seja, enquanto o adjunto adnominal representa o termo acessório da oração, o predicativo se revela como um termo essencial, de modo a deixá-la compreensível, dotada de sentido. Em d), "referência em saúde e segurança" representa parte essencial à constituição do enunciado, pois sem a presença desses termos o entendimento estaria comprometido. Assim, consideramos que se trata de um predicativo, visto que atribui uma característica ao sujeito, cujo núcleo é representado por "cidade".

d) A cidade que é referência em saúde e segurança.

Em e) constatamos que o termo "limpa" pode perfeitamente ser retirado do contexto oracional sem que isso cause nenhum dano ao perfeito entendimento do discurso. Logo, trata-se de um termo acessório da oração ou adjunto adnominal.

e) A cidade limpa que é referência em saúde e segurança.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

12.2.1.3 Tipos de Referentes

Existem três tipos de referentes: referentes específicos, referentes não-específicos e referentes abstratos.

Referentes específicos: Quando a menção de uma entidade, basicamente, identifica-a por meio de um nome comum ou próprio.

f) Microsoft informou que irá resolver o bug que reinicia o Windows Phone em dezembro.
g) Luiz Inácio Lula da Silva sancionou nesta quarta-feira, 29, a lei que regulamenta as atividades de moto-taxista e motoboy de todo país...
h) Roger Waters faz seu segundo show em São Paulo.

Em (f) temos um referente específico, isto é, a menção da entidade refere-se diretamente a algo específico, à empresa Microsoft. O referente específico, nesse caso, ainda pode ser classificado como uma entidade do tipo Organização. Existem outros tipos de referentes específicos, como Pessoa (g), Local (h) , entre outros. Notemos que em (h) temos dois tipos de referentes específicos, "Roger Waters" e "São Paulo", respectivamente entidades do tipo Pessoa e Local.

Referentes não-específicos: Quando as menções referem-se a uma entidade de forma não específica (autoridades, funcionários, policiais...), como mostram os exemplos "i", "j" e "k".

i) Policiais invadiram a casa, porém os bandidos já haviam fugido.…
j) Funcionários estão descontentes. Eles afirmam ainda não terem recebido o seu décimo terceiro salário.
k) Autoridades disseram que estão cansados de fazer as mesmas declarações.

Referentes abstratos: como o próprio nome sugere, são entidades abstratas, "não físicas". Tratam de estados e qualidades, sentimentos e ações, como: medo, viagem, coragem, felicidade, esforço... Exemplos "l" e "m".

l) O medo é algo que deve ser superado. Para isso, concentre-se em seus objetivos.
m) A viagem foi ótima, porém o tempo podia estar melhor.

12.2.2 Relações Semânticas Referenciais

Nesta subseção temos como foco tornar claro os tipos de relações semânticas que podem indicar uma relação de correferência.

12.2.2.1 Hiperonímia e Hiponímia

Hiperonímia é uma relação semântica que expressa um sentido amplo entre dois termos, partindo de uma classe mais ampla para uma subclasse mais específica, por exemplo: (inseto abelha). Neste caso, o termo "inseto" é um hiperônimo de "abelha". Já Hiponímia representa uma relação contrária, ou seja, parte de uma classe mais específica para uma classe mais abrangente. Para o exemplo previamente dado temos que "abelha" é um hipônimo de "inseto". Os hiperônimos e hipônimos são importantes no campo semântico, pois

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

são muito usados na retomada de elementos em um texto, a fim de evitar repetições desnecessárias. No que diz respeito à identificação de menções referenciais em um discurso, na língua portuguesa é comum partirmos de termos específicos para termos mais abrangentes. Dessa forma, a relação de Hiponímia geralmente ocorre com maior frequência.

n) João e Maria estão muito felizes com o seu cão. O animal é fiel e companheiro.
o) Nada disso vai fazê-los mudar de carro. O pequeno veículo parece suprir todas as necessidades deles.

12.2.2.2 Sinonímia

Trata-se de uma relação entre dois termos, em que estes, mesmo sendo distintos lexicalmente, possuem significados muito próximos, por exemplo: (menino garoto). É importante referir que muitas vezes os sinônimos podem ter conotações diferentes, dependendo do contexto, como: (gato bichano) e (gato atraente). Em um texto, a utilização de sinônimos de uma palavra é importante para evitar repetições. Assim, um sinônimo é uma palavra que, apesar de ser diferente, tem o mesmo significado (ou semelhante) e, por isso, a sua inclusão não altera o sentido do texto em questão.

p) Esse carro é maravilhoso. Também, estamos falando de um automóvel de 100 mil reais.
q) Ana comprou um gato. O bichano adora dormir no sofá.

12.2.3 Correferência, Anáfora e Catáfora

Para o entendimento sobre o que é correferência, é relevante também definirmos anáfora, já que seus conceitos estão relacionados. Anáfora pode ser definida como a retomada de uma expressão apresentada anteriormente em um texto. Quando uma entidade é mencionada pela primeira vez textualmente, temos a evocação da entidade. Durante a leitura da sequência do texto, quando essa entidade é novamente mencionada, temos a realização do acesso a essa entidade. A expressão que faz o acesso é dita como anafórica, e a expressão anterior é dita como seu antecedente (Vieira et al., 2012). Há casos de anáfora em que o termo anafórico e o antecedente são correferentes, isto é, remetem a uma mesma entidade (como os Exemplos "r" e "s" ilustram), mas há também casos de anáfora sem correferência, como podemos ver em "t".

r) A Ana comprou um cão. O animal já conhece todos os cantos da casa. Nesse exemplo, o termo anafórico é o grupo nominal "o animal", que retoma o valor referencial do antecedente, "o cão". É a relação entre "cão" e "animal" que suporta a correferência.
s) Maria está com febre. Acho que ela está doente. Notemos que a interpretação referencial do sintagma nominal "ela" depende da sua relação anafórica com o sintagma nominal "Maria".
t) João faz 18 anos no dia 2 de Julho de 2001. No dia seguinte, parte para uma grande viagem pela Europa. Já nesse caso, o valor referencial da expressão sublinhada constrói-se a partir da interpretação do antecedente, a expressão adverbial temporal "no dia 2 de Julho de 2001". Assim, "No dia seguinte" designa o dia 3 de Julho de 2001.

Catáfora: semelhante à anáfora mas em ordem oposta, uma relação catafórica ocorre quando um termo se refere a outro que vem à frente e lhe dá, a partir deste, o seu sentido. Conforme podemos ver no exemplo "u":

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

u) A mãe olhou-o e disse: - Meu filho, estás com um olhar cansado.

Correferência: é um fenômeno que ocorre quando duas ou mais menções em um discurso referem-se a uma mesma entidade. O conjunto de menções a uma mesma entidade no texto é denominado de cadeia de correferência.

v) O João está doente. Vi-o na semana passada. Neste caso, o pronome "o" é uma anáfora de "João", pois, para ser compreendido, necessita resgatar a frase anterior para que seu significado seja construído.

Temos também o tipo aposto, que ocorre quando o termo da oração se relaciona a uma entidade para esclarecê-la ou explicá-la, como em "w" e "x".

w) Cubatão, a cidade mais poluída do Brasil, localiza-se na Baixada Santista.

x) Maria comprou várias frutas: mamão, melancia, abacate e uva. Normalmente, o aposto aparece isolado por sinais de pontuação, sendo mais comum aparecer entre vírgulas ou então introduzido por dois pontos. Nos exemplos acima podemos notar que "cidade" é correferente de "Cubatão", e "mamão, melancia, abacate e uva" são correferentes de "frutas".

y) (extraído do *corpus* Summ-it (Collovini et al., 2007)) "A discussão sobre a biotecnologia nacional está enviesada, pois está sendo entendida como sinônimo de transgenia. A opinião é de Miguel Guerra, da UFSC (Universidade Federal de Santa Catarina). Guerra participou do debate"Biotecnologia para uma Agricultura Sustentável"... Para o agrônomo, o Brasil deve buscar o desenvolvimento de transgenias que tentem melhorar as condições da agricultura local..."

No exemplo "y", as expressões "Guerra" e "o agrônomo" fazem referência à entidade "Miguel Guerra", já mencionada anteriormente no texto. Para não repetir a mesma expressão, faz-se uso de outra diferente, mas que retoma a mesma entidade mencionada previamente. Esse é um método muito utilizado no processo de escrita para não deixar o texto repetitivo, e está diretamente relacionado a coesão referencial e sequencial. Notemos que a coesão referencial é responsável por criar um sistema de relações entre as menções dentro de um texto, permitindo que o leitor identifique termos e expressões que remetem a uma mesma entidade. Junto a isso temos também a coesão sequencial, responsável por criar condições que auxiliam na progressão textual. De forma geral, as flexões de tempo, as de modo dos verbos e as conjunções são mecanismos responsáveis pela coesão sequencial e auxiliam na coesão referencial.

Esses fatores inferem diretamente nas dificuldades existentes na tarefa de resolução de correferência, dado que estão relacionados diretamente a questões linguísticas e a habilidades cognitivas humanas complexas, de difícil reprodução por sistemas computacionais. Em poucas palavras, o desafio é: como inferir computacionalmente que a palavra "agrônomo", que está sendo citada dois parágrafos abaixo da expressão "o agrônomo Miguel Guerra", está se referindo a esta entidade e não a uma outra?

Portanto, o conjunto dessas expressões referenciais relativas a uma mesma entidade de mundo denomina-se **cadeia de correferência**. Esse conjunto é responsável pela construção coesa de um texto e por isso sua importância, já que a coesão é responsável pela compreensão textual. No exemplo "y", podemos afirmar que "Miguel Guerra" é o antecedente e "Guerra" é a anáfora. Dessa forma, expressões correferentes fazem referência

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

à mesma entidade, enquanto expressões anafóricas e catafóricas podem retomar uma referência ou ativar um novo referente. Anáfora e catáfora pressupõem um par ordenado, enquanto que a correferência remete à ideia de conjunto (Poesio et al., 2016).

12.2.4 Referências Endofóricas e Exofóricas

Quando lidamos com relações referenciais é importante deixarmos claro que estas podem configurar-se em dois tipos: endofóricas e exofóricas. Referências endofóricas são aquelas que antecedem ou sucedem informação dentro de um texto. Essas comumente ocorrem na forma de anáforas ou catáforas, conforme visto nos exemplos anteriores. Já relações exofóricas referem-se a relações que ocorrem fora de um dado texto e necessitam de um prévio conhecimento de mundo, local ou momento para serem identificadas, como em:

z) O Bruxo do Cosme Velho foi homenageado em nossa cidade.

Notemos que no exemplo "z" não existem referências para os termos "Bruxo do Cosme Velho" e "nossa cidade". É necessário recorrermos ao conhecimento de mundo para inferirmos que "Bruxo do Cosme Velho" refere-se a Machado de Assis. Da mesma forma, a referência do termo "nossa cidade" não está no texto, mas pode estar na memória do leitor ou na memória do escritor.

12.2.5 Correferência, coerência e coesão

Quando lidamos com a resolução de correferência, existem características que geralmente ficam implícitas em textos bem escritos e estruturados, mas que merecem atenção, dado que influenciam diretamente na obtenção de bons resultados. Dentro desse contexto, temos a coerência e a coesão textual. De acordo com Koch; Travaglia (2012) a coerência textual é algo que tem a ver com a boa formação do texto, não em um sentido gramatical, mas sim em nível de interlocução. A coerência é algo que se estabelece na interação, na interlocução ou em uma situação comunicativa entre duas pessoas. Em poucas palavras, a coerência é o que faz com que o texto tenha sentido, devendo ser vista como um princípio de interpretabilidade do texto e também com a capacidade que o leitor possui para calcular seu significado. A coerência é vista também como uma continuidade de sentidos perceptíveis no texto, a qual resulta em uma conexão conceitual cognitiva entre os elementos do texto. Como podemos perceber, a coerência é, ao mesmo tempo, semântica e pragmática aplicadas, pois a forma como construímos nossas ideias pode variar, de acordo com nosso conhecimento de mundo.

Paralelamente ao conceito de coerência temos a coesão. Ao contrário da coerência, a coesão é explicitamente revelada por meio de marcas linguísticas, sendo de caráter linear, dado que manifesta-se na organização sequencial de um texto. Em poucas palavras, a coesão está muito mais ligada à sintaxe e à gramática. Notemos que esses conceitos são muito importantes para a tarefa de Resolução de Correferência, dado que a correferência de um termo e seu antecedente é guiada por essa construção de ideias.

12.3 Abordagens Computacionais para Resolução de Correferência

Na literatura, encontramos uma grande variedade de abordagens que propõem resolver correferência em diversos idiomas, como: inglês, chinês, árabe, espanhol, galego, português,

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

entre outros (Chang et al., 2012; Coreixas, 2010; Fernandes et al., 2014; Fonseca et al., 2014; Fonseca, 2014; Fonseca et al., 2015a; Lee et al., 2017b; Martschat; Strube, 2015; Ng; Cardie, 2002; Rahman; Ng, 2011a; Soon et al., 2001; Yang et al., 2008). Essas abordagens, em sua maioria, são voltadas para a língua Inglesa e baseadas em aprendizado de máquina. Contudo, é possível encontrarmos alguns modelos baseados em regras linguísticas (Garcia; Gamallo, 2014; Hou et al., 2014; Lee et al., 2013). Veremos que, diferente dos modelos baseados em regras, o aprendizado de máquina pode se ramificar em diferentes propostas, como *Mention-Pair*, *Entity-Mention*, *Mention-Ranking* e *Antecedent-Trees*.

12.3.1 Modelos Baseados em Regras

Modelos baseados em regras consistem em uma série de passos que definem se duas menções são correferentes entre si. Abordagens baseadas em regras requerem um conhecimento prévio mais aprofundado referente ao idioma e ao domínio a serem tratados. Por exemplo, ao lidarmos com aprendizado de máquina, durante a implementação e seleção de *features*, caso sejam selecionadas *features* irrelevantes, a maioria dos algoritmos de treino consegue detectar e desconsiderar tal característica. Já em abordagens voltadas a regras, não temos essa flexibilidade. Cada regra deve ser elaborada cuidadosamente, pois não temos um modelo estatístico como apoio. Outra característica forte desses modelos é a forma como as menções são agrupadas. Ou seja, em abordagens baseadas em regras não existe etapa de treinamento: definidas as regras, uma menção m_j é comparada com todas as menções que a antecedem e, caso alguma das regras seja satisfeita, essas menções são agrupadas. Esse tipo de método é o mais utilizado pelos modelos de regras atuais (Garcia; Gamallo, 2014; Lee et al., 2013).

Um dos principais modelos de regras contido na literatura foi proposto por Lee et al. (2011). Denominado Stanford Multi-Pass Sieve, é um sistema para a resolução de correferência puramente baseado em regras linguísticas. Seu modelo possui dez *Sieves*/filtros, cujo objetivo é agrupar menções correferentes, caso cada regra ou conjunto de regras sejam satisfeitos. O modelo de Lee et al. foi proposto para o inglês, durante a CoNLL[2] (*Conference on Natural Language Learning*), ficando em primeiro colocado no *ranking* de melhores modelos. Os modelos foram avaliados por meio do *corpus* Ontonotes (Pradhan et al., 2011), em conjunto do CoNLL Scorer (Pradhan et al., 2014). Alguns anos após surgiram outras abordagens semelhantes com o mesmo propósito, como o trabalho de Garcia; Gamallo (2014), voltado ao português, ao espanhol e ao galego. Embora abordagens baseadas em regras possam ser de custoso planejamento, dado que cada idioma possui suas características, estas podem provar-se eficazes e competitivas, principalmente quando há carência por *corpora* anotados.

12.3.2 Modelos Baseados em Aprendizado de Máquina

12.3.2.1 Mention Pair

A primeira delas, e a mais popular, provavelmente por sua simplicidade, é a abordagem baseada em pares de menções. Basicamente, modelos que lidam com essa abordagem optam por efetuar seu treino por meio de pares de menções, de forma a determinar se duas menções são correferentes ou não. Os modelos baseados em pares de menções têm influenciado significativamente os trabalhos que propõem a resolução de correferência utilizando técnicas de aprendizado de máquina nos últimos dezesseis anos (Soon et al., 2001). Modelos

[2] http://conll.cemantix.org/2011/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

baseados em pares de menções visam responder se devem ou não classificar como correferente uma menção m_j com um candidato antecedente m_i. Inicialmente, para treinar um modelo baseado em pares, é necessário extrair características/*features* que possibilitem obter alguma informação proveniente da comparação entre m_i e m_j. No Quadro 12.1, podemos observar as *features* mais utilizadas pela literatura.

Quadro 12.1: *Features* mais comuns na literatura.

Feature	**Descrição**
Casamento de Padrões	Se m_i e m_j são lexicalmente iguais.
Casamento de Núcleos	Se m_i e m_j possuem o mesmo núcleo.
Alias	Se m_j é sigla de m_i ou vice-versa.
I_Pronome	Se m_i é um pronome.
J_Pronome	Se m_j é um pronome.
Número	Se m_i e m_j concordam em número (singular/plural).
Gênero	Se m_i e m_j concordam em gênero (masculino/feminino).
Nome Próprio	Se m_i e m_j são nomes próprios.
Aposto	Se m_j é aposto de m_i
Distância entre Sentenças	Distância em sentenças entre m_j e m_i.
Distância entre Sintagmas	Distância em menções à entre m_j e m_i.
Classe Semântica	Se m_i e m_j possuem mesma classe semântica.
Hiponímia	Se m_i e m_j possuem uma relação de hiponímia.
Hiperonímia	Se m_i e m_j possuem uma relação de hiperonímia.
Sinonímia	Se m_i e m_j possuem uma relação de sinonímia.

Um dos grandes desafios ao utilizar uma abordagem baseada em pares de menções se dá no desbalanceamento entre as classes positiva (pares correferentes) e negativa (pares não correferentes). Ou seja, todo modelo requer, além de amostras positivas, amostras de pares negativos. Dada essa premissa, é necessário realizar a construção de pares. Nessa etapa, ao cruzarmos essas menções, consequentemente teremos muito mais amostras negativas do que positivas. Objetivando minimizar esse desbalanceamento entre as classes, alguns trabalhos propõem diferentes técnicas para geração de pares.

Soon et al. (2001) realizam um pareamento distinto para cada uma das classes: para os pares positivos, dado o conjunto de menções C={m_i,m_j,m_k, m_l} (todas correferentes entre si), apenas as menções imediatamente adjacentes formam pares (Quadro 12.2): P_p= {(m_i,m_j), (m_j,m_k), (m_k,m_l)}. Para gerar os pares negativos, considere o conjunto de menções M={m_m, m_n, m_o, m_p, m_q} em que apenas as menções m_m e m_q são correferentes. Dentro desse contexto, a última menção deste conjunto, m_q, faz par com todas as anteriores, exceto com m_m: P_n={ (m_q,m_p), (m_q,m_o) e (m_q,m_n)} (Quadro 12.3).

Notemos que no Quadro 12.2, o conjunto de menções considerado é uma cadeia de correferência. Já no Quadro 12.3, o conjunto de menções não consiste em uma cadeia. Apenas as menções m_m e m_q são correferentes. Logo, não formam par.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Quadro 12.2: Geração de pares positivos proposta por (Soon et al., 2001).

	Pareamento de amostras Positivas
Conjunto de menções	m_i, m_j, m_k, m_l
Pares	m_i, m_j
	m_j, m_k
	m_k, m_l

Quadro 12.3: Geração de pares negativos proposta por (Soon et al., 2001).

	Pareamento de amostras Negativas
Conjunto de menções	m_m, m_n, m_o, m_p, m_q
Pares	m_q, m_p
	m_q, m_o
	m_q, m_n

Martschat; Strube (2015) propõem uma mesma metodologia para geração de pares positivos e negativos: dado documento D_x, que possua um conjunto de menções M=$\{m_i$, m_j,m_k, m_l, m_m, $m_n\}$ em que apenas m_l e m_i são correferentes, o conjunto de pares (positivos e negativos) será: P= $\{(m_l,m_k)$, (m_l,m_j), $(m_l,m_i)\}$. Basicamente, para cada par correferente (m_x,m_y), a geração de amostras negativas será realizada com as menções entre (m_x,m_y). Notemos que essa construção é efetiva pelo fato de não gerar uma grande quantidade de amostras negativas. No entanto, devido a essa restrição, pode-se perder pares negativos que possuam informações relevantes.

Quadro 12.4: Geração de pares positivos e negativos proposta por (Martschat; Strube, 2015).

	Pareamento de amostras Positivas e Negativas
Conjunto de menções	m_i, m_j, m_k, m_l, m_m, m_n
Pares	m_l, m_k
	m_l, m_j
	m_l, m_i

Em (Fonseca et al., 2015a), para um dado conjunto de menções M=$\{m_i,m_j,m_k\}$, temos: P=$\{(m_i,m_j)$, (m_i,m_k), $(m_j,m_k)\}$. Basicamente, cada menção faz par com a próxima, independente de esta ser correferente ou não. Notemos que a quantidade de pares será muito maior que em (Martschat; Strube, 2015). Para minimizar o desbalanceamento entre as classes foi utilizado *random undersampling*, que consiste na seleção aleatória de n pares negativos, em que n é a quantidade de pares positivos. Por meio de experimentos, foi visto que os níveis de balanceamento "1 para 1" (um par positivo para cada par negativo) e "1 para 2" (um par positivo para cada dois pares negativos) foram os que obtiveram melhores resultados.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Quadro 12.5: Geração de pares positivos e negativos proposta por (Fonseca et al., 2015a).

	Pareamento de amostras Positivas e Negativas
Conjunto de menções	m_i, m_j, m_k
Pares	m_i, m_j
	m_i, m_k
	m_j, m_k

12.3.3 Entity-Mention

Diferente do tradicional *Mention Pair*, o *Entity-Mention* (Yang et al., 2008) explora a propriedade de representação do discurso, tendo em vista o conhecimento de quando uma entidade é nova no discurso ou anafórica (semelhante à nossa metodologia de agrupamento proposta). Para conceber os pares, assume-se que uma instância de treino positiva consiste em i{e_x,m_y}, na qual m_y é uma menção ativa e e_x é uma entidade parcial, encontrada antes de m_y. Para cada menção anafórica m_y, uma única instância de treinamento positivo é criada para a entidade parcial à qual m_y pertence. Para os pares negativos, é criado um grupo de instâncias para cada entidade cuja última menção ocorra entre m_y e o antecedente mais próximo de m_y. Por exemplo: considere o conjunto de menções M={e_i,m_j,e_k,m_l,m_m,m_n}. Assumindo que neste conjunto tenhamos duas cadeias: C1={e_i,m_j, m_m} e C2={e_k,m_n} e m_n é a menção ativa. Teremos então, como conjunto dos pares positivos, P_p={(m_n,e_k)} e como conjunto de pares negativos P_n={(m_n,m_m), (m_n,m_j), (m_n,e_i), (m_n,m_l)}. Basicamente, assumindo que m_n representa a menção ativa e e_k representa seu antecedente, notemos que temos duas menções entre elas (m_l e m_m). Nesse caso, toda menção ou cadeia pertencente a m_l e m_m forma par com a menção ativa. Nos Quadros 12.6 e 12.7 temos os pares gerados, considerando m_n como menção ativa. Notemos que a cada iteração a menção ativa será outra e com isso novos pares serão gerados, sempre utilizando o mesmo critério.

Quadro 12.6: Geração de pares positivos proposta por (Yang et al., 2008).

	Pareamento de Amostras Positivas
Conjunto de menções	e_i, m_j, e_k, m_l, m_m, m_n
Pares	m_n, e_k

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Quadro 12.7: Geração de pares negativos proposta por (Yang et al., 2008).

	Pareamento de Amostras Negativas
Conjunto de menções	e_i, m_j, e_k, m_l, m_m, m_n
Pares	m_n, m_m
	m_n, m_j
	m_n, e_i
	m_n, m_l

Outro diferencial deste modelo focado em entidades está na forma de representar suas *features*: os autores definem três tipos de instâncias, que representam como as menções se relacionam: *link*(e_x,m_y), em que m_y representa uma menção ativa e e_x representa uma entidade parcial; *has_mention(e,m)*, descrevendo todas as menções as quais determinada menção está ligada. Por exemplo, para a cadeia previamente mencionada, C1={e_i,m_j, m_m}, teremos então *has_mention*(e_i,m_j), *has_mention*(e_i,m_m); e o último denota as características de cada par de menções, seguindo a seguinte estrutura: nome_da_feature(m_x,m_y, 0), representando respectivamente: o nome da *feature*, o par de menções e um valor binário (0 para falso e 1 para verdadeiro).

12.3.4 Mention-Ranking

No *Mention-Ranking model*, assim como o *Mention pair*, cada instância de treino i(m_x,m_y) representa m_y e sua menção precedente m_x. Basicamente, as *features* que representam uma instância e um método para criar uma instância de treino são idênticas às utilizadas pelo *Mention Pair model*. A diferença reside em rotular as instâncias de treino, assumindo que I_y é um conjunto de instâncias de treino, criadas para a menção anafórica m_y, o *rank* para i(m_x,m_y) em I_y é o rank de m_y entre os candidatos antecedentes, que é 2 se m_x é o antecedente mais próximo de m_y ou 1 caso contrário. Em poucas palavras, o antecedente mais próximo de sua anáfora recebe um *ranking* maior em relação às demais menções. Considere o seguinte conjunto de menções M={m_i, m_j, m_k, m_l, m_m, m_n}, contendo as seguintes cadeias C1={m_i, m_k, m_n,} C2={m_j, m_l,}. Notemos que, para m_n, teremos as seguintes instâncias:

Quadro 12.8: Instâncias de treino geradas para m_n

$\mathbf{I_n}$		***Ranking***
Pares	m_m, m_n	1
	m_l, m_n	1
	m_k, m_n	2

Notemos que m_i não faz par com m_n, mas sim com m_k, dado que m_i é antecedente de m_k. Dado que m_k é o antecedente mais próximo de m_n, o par recebe *ranking* 2. Já os demais pares (considerando m_n) recebem valor 1. Mesmo o *Mention-Ranking* não sendo muito popular, seus resultados são superiores às abordagens baseadas em pares de menções, como podemos ver em (Rahman; Ng, 2011b) e (Martschat; Strube, 2015).

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

12.3.5 Antecedent-Tree

Na proposta de (Fernandes et al., 2014), também baseada em pares de menções, os autores propõem um conjunto de regras, as quais objetivam reduzir a quantidade de pares menos propensos a serem correferentes. Assim, para um dado par de menções, caso pelo menos uma das regras do Quadro 12.9 seja satisfeita, o par é considerado válido para utilizar em seu treinamento (seja ele um par positivo ou negativo).

Quadro 12.9: Conjunto de regras para seleção de pares proposto por (Fernandes et al., 2014)

Regra	**Descrição – Considera um par como válido se:**
Distância	a quantidade de menções entre m_i e m_j não ultrapassar um determinado *threshold*
Classe Semântica	m_i e m_j possuírem mesma classe semântica.
Combinação de Núcleos	o núcleo de m_i combinar com o núcleo de m_j.
Concordância em atributos de discurso	os atributos de discurso combinam para m_i e m_j. Esta regra consiste de um conjunto de regras proposto por Lee et al. (2013), o qual baseia-se em atributos de uma menção e seu falante.
Pronome J	m_j for um pronome e m_i concordar em gênero, número ou fala.
Pronome e Entidade Nomeada	m_j for um pronome e m_i for um pronome compatível ou uma entidade nomeada.

Referente ao motivo dos autores nomearem sua abordagem como *Antecedent-Tree model*, reside na forma de representar o agrupamento de suas menções: para representar o agrupamento de menções correferentes entre si são utilizadas estruturas chamadas de árvores. Uma árvore de correferência é uma árvore cujos nós são dirigidos às menções e os arcos representam alguma relação entre elas. Basicamente, para cada documento é gerado um conjunto de árvores e de sub-árvores, em que cada sub-árvore representa uma menção e seus referentes. Ou seja, cada anáfora pode ser considerada uma raiz ou nodo-pai e seus antecedentes podem ser considerados nodos-filhos.

Notemos que cada abordagem possui uma forma distinta para concepção de suas amostras de treino, assim como que para representar suas estruturas de agrupamento. Martschat; Strube (2015) propuseram uma forma unificada de representar tais estruturas, os autores a chamam de estrutura latente. Basicamente, uma estrutura latente é representada por um conjunto de *arrays* "V", "A" e "L". Analisando-a, podemos verificar que uma estrutura latente pode ser abstraída à forma de um grafo, o qual "V" representa um conjunto de nós/menções; "A" representa o conjunto de arestas, e "L"(label), representa um sinal, positivo ou negativo, informando se dada menção é correferente de outra.

12.3.5.1 Modelos de linguagem

Os modelos de linguagem (Capítulo Modelos de linguagem) foram também incorporados às tarefas de resolução de correferência, Joshi et al. (2019) apresentam uma análise do modelo BERT nessa tarefa.

Com os avanços recentes dos modelos de linguagem, abordagens alternativas, como a de Kirstain et al. (2021), dispensam a necessidade de representação de *spans*.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

12.3.6 Modelos Voltados à Língua Portuguesa

Para a língua portuguesa, Silva (2011) propôs um modelo para a resolução de correferência utilizando o conjunto de etiquetas semânticas providas pelo *corpus* do HAREM (Freitas et al., 2010). Para detectar tais categorias, Silva utilizou o *parser* PALAVRAS (Bick, 2000) e o reconhecedor de entidades nomeadas Rembrandt (Cardoso, 2012). Como base de conhecimento semântico, o autor utilizou o Tep2.0[3] (Maziero et al., 2008), um *thesaurus* contendo relações de sinonímia e antonímia para a língua portuguesa.

Lidando com aprendizado supervisionado, temos o trabalho de Coreixas (2010), o qual focou nas categorias "Pessoa", "Local", "Organização", "Acontecimento", "Obra", "Coisa" e "Outro". Como recursos, foram utilizados o *corpus* do HAREM (Freitas et al., 2010), o *parser* PALAVRAS (Bick, 2000) e o *corpus* Summ-it (Collovini et al., 2007). De forma a provar que o uso de categorias semânticas pode auxiliar na tarefa de resolução de correferência, o autor compara duas versões de seu sistema: a primeira, sem fazer o uso de categorias semânticas; e a segunda, fazendo uso de categorias. Como resultado, Coreixas mostrou que o uso de categorias podem prover melhorias significativas, dado que o uso de categorias pode auxiliar a determinar se dado par de menções é correferente ou não. O autor também mostrou a importância do conhecimento de mundo para esta linha de pesquisa.

Seguindo uma linha semelhante ao trabalho de Coreixas (2010), Fonseca et al. (2014) propõem um modelo baseado em aprendizado de máquina, com foco em nomes próprios e nas categorias de entidades "Pessoa", "Local" e "Organização". Para detectar as entidades, foram utilizados os recursos Repentino (Sarmento et al., 2006) e NERP-CRF (Amaral, 2013). Adicionalmente, para casos mais genéricos de entidades, os autores utilizaram listas, contendo substantivos comuns, que remetem a determinadas entidades, tais como: [advogado, agrônomo, juiz] para a categoria "Pessoa", e [avenida, rua, praça, cidade] para "Local".

Diferente dos trabalhos anteriores (Garcia; Gamallo, 2014) abordam um modelo baseado em regras (semelhante ao de Lee et al. (2013)), mas para múltiplos idiomas (português, espanhol e galego). Em seu trabalho, os autores focam apenas na categoria semântica "Pessoa". Mais recentemente Fonseca (2018) propôs um modelo baseado em regras linguísticas, similar ao modelo de Lee et al. (2013), mas totalmente voltado ao português.

Adicionalmente, Fonseca (2018) introduziu conhecimento semântico ao seu modelo, provindo do Onto.PT (Gonçalo Oliveira; Gomes, 2014) e uma nova metodologia de agrupamento de menções (Fonseca et al., 2018). Basicamente, seu método (Figura 12.1) recebe como entrada uma lista ordenada de menções "M" e devolve uma lista de Cadeias contendo essas menções devidamente agrupadas, de acordo com o critério selecionado.

O método proposto foi baseado no trabalho de Heim (2008) e consiste em explorar a representação do discurso[4]. Para isso, assume-se que qualquer menção é nova no discurso se não possuir ligação de correferência com uma ou mais menções antecedentes. Essas ligações são consideradas utilizando o conjunto de regras proposto pelo autor. Assim, sempre que uma menção não possui uma relação referencial (nenhuma regra é satisfeita), uma nova cadeia é gerada. Basicamente utilizou-se uma lista de menções M (esta lista é ordenada na ordem em que as menções ocorrem no texto), contendo todas as menções de

[3] http://www.nilc.icmc.usp.br/tep2

[4] Considera-se como representação do discurso a forma como as ideias são construídas em textos de linguagem natural, considerando sua construção linguística e seu contexto de uso, bem como suas formas de expressões comuns. Em linguística, a Pragmática é o ramo que analisa o uso concreto da linguagem pelos falantes da língua em seus variados contextos.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Figura 12.1: Algoritmo de agrupamento proposto por Fonseca (2018)

```
1:  enquanto (tamanho de M > 0) faça
2:      int j ← 0;
3:      int[ ] S;
4:      para cada i ∈ C faça
5:          se M_0 tem relação com C_i então
6:              S_j ← C_i
7:              j ← j + +
8:          fim se
9:      fim para
10:     se j > 0 então
11:         int k ← CritérioDeAgrupamento(M_0, S, C)
12:         C_k ← M_0
13:     senão
14:         C ← criaNovaCadeia(M_0)
15:     fim se
16:     M ← Remove(M, 0)
17: fim enquanto
```

um documento de entrada. Cada menção pode ter uma ligação de correferência entre uma ou mais cadeias "C". Dessa forma, os Ids dessas cadeias são armazenados em um vetor "S" (apenas se M_0 possui alguma relação de correferência com C_i (se alguma regra retorna o valor verdade). O próximo passo é responsável por agrupar uma menção atual M_0 a uma cadeia existente C_k ou criar uma nova cadeia de correferência, usando M_0[5]. Isso depende do critério de agrupamento utilizado. Dos critérios propostos pelo autor, vale mencionar o que obteve melhores resultados: o "Peso por Regra", o qual para cada menção explora o conjunto C, com o objetivo de encontrar o maior peso (em nível de cadeias). O modelo proposto possui um total de 13 regras e; para cada regra satisfeita, soma-se 1 à pontuação. Assim, se para uma dada cadeia C_x existem duas menções correferentes com uma menção M_0 (M_a e M_b) e M_a e M_b possuem respectivamente três e duas regras com valor verdade, o peso da cadeia será cinco.

12.4 Avaliação da Tarefa de Resolução de Correferência

A tarefa de resolução de correferência é complexa e envolve diferentes níveis de processamento. Logo, avaliar um modelo de correferência não é uma tarefa simples, dado que existem muitos detalhes a serem considerados, como a detecção de menções, agrupamentos realizados, agrupamentos não realizados[6]. Na literatura encontramos cinco métricas propostas para avaliar esses modelos: MUC (Vilain et al., 1995), B-CUBED (Bagga; Baldwin, 1998), $Ceaf_e$, $Ceaf_m$ (Luo, 2005) e BLANC (Recasens; Hovy, 2011). Cada uma dessas métricas visa avaliar uma característica específica de cada modelo. Anualmente, competições como a CoNLL (Pradhan et al., 2012) são realizadas, visando motivar o desenvolvimento

[5]Notemos que, para cada iteração, M_0 muda.
[6]Veja mais sobre Avaliação de sistemas de PLN no Capítulo 18.

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

de sistemas. Nos anos de 2011 e 2012 essas competições foram voltadas à tarefa de Resolução de Correferência. Com o objetivo de avaliar os modelos participantes por meio de uma pontuação única, a conferência propôs uma nova métrica, chamada CoNLL (Pradhan et al., 2014). A métrica CoNLL consiste na média da medida-F de três outras métricas da literatura, como veremos nessa seção.

12.4.1 Métricas de Avaliação

12.4.1.1 MUC:

Baseada em cadeias, mede quantos agrupamentos de menções são necessários para cobrir as cadeias padrão. Por exemplo, considere que o conjunto K (cadeia de referência) seja composto pelas seguintes ligações(*links*) de correferência {AB, BE, CD} e que o conjunto R (cadeia predita pelo modelo) seja composto por {AB, CD}. Para este caso podemos ver que falta uma ligação no conjunto R. teremos então Abrangência= $\frac{2}{3} = 0,67$ (67%) e Precisão = $\frac{2}{2} = 1$ (100%). De forma mais geral, o cálculo da métrica MUC pode ser obtido por meio das seguintes fórmulas:

$$Abrangência = \frac{\sum_{i=1}^{N_k}(\|K_i\| - \|p(K_i)\|)}{\sum_{i=1}^{N_k}(\|K_i\| - 1}$$

$$Precisão = \frac{\sum_{i=1}^{N_r}(\|R_i\| - \|p'(R_i)\|)}{\sum_{i=1}^{N_r}(\|R_i\| - 1)}$$

Em que: K_i é i-ésima entidade padrão (*key entity* referência) e p(K_i) é o grupo de partições criado por meio da intersecção de K_i e os *links* preditos pelo modelo; R_i é a i-ésima entidade predita pelo modelo (*Response entity*) e p'(R_i) é o conjunto de partições criadas por meio da intersecção de R_i e K_i. N_k e N_r representam a quantidade de menções padrão e resposta, respectivamente.

12.4.1.2 B-CUBED:

Baseada em menções, gera resultados considerando as menções presentes e ausentes de cada entidade em dada cadeia. Basicamente, a métrica B-Cubed atribui um peso para as menções, baseando-se na quantidade total de menções existentes. Sua abrangência e precisão são obtidas por:

$$Abrangência = \frac{\sum_{i=1}^{N_k}\sum_{j=1}^{N_k}\frac{\|K_i \cap R_j\|^2}{K_i}}{\sum_{i=1}^{N_k} K_i}$$

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

$$Precisão = \frac{\sum_{i=1}^{N_k}\sum_{j=1}^{N_k}\frac{\|K_i \cap R_j\|^2}{R_j}}{\sum_{i=1}^{N_k} R_j}$$

Em que: K representa o conjunto das *key entities* (menções padrão) e R o conjunto de menções preditas pelo modelo. Por exemplo, dadas as cadeias de referência :

- C_{K1} = {A, B, C, D, E};
- C_{K2} = {F, G};
- C_{K3} = {H, I, J, K, L}.

E as cadeias preditas pelo modelo:

- C_{R1} = {A, B, C, D, E};
- C_{R2} = {F, G, H, I, J, K, L}.

Cada menção possuirá o peso de $\frac{1}{12}$, dado que o total de menções existente é 12. Dito isso temos então:
Abrangência $= \frac{1}{12} * [\frac{5}{5} + \frac{5}{5} + \frac{5}{5} + \frac{5}{5} + \frac{5}{5} + \frac{2}{2} + \frac{2}{2} + \frac{5}{7} + \frac{5}{7} + \frac{5}{7} + \frac{5}{7} + \frac{5}{7}] = 1$ (100%)
Precisão $= \frac{1}{12} * [\frac{5}{5} + \frac{5}{5} + \frac{5}{5} + \frac{5}{5} + \frac{5}{5} + \frac{2}{7} + \frac{2}{7} + \frac{5}{7} + \frac{5}{7} + \frac{5}{7} + \frac{5}{7} + \frac{5}{7}] = \frac{16}{21} = 0,76$ (76%)

12.4.1.3 CEAF:

Baseada no alinhamento de menções e entidades, possui duas variações: CEAF_m (Φ_3) e CEAF_e (Φ_4).

$$\Phi_3(K, R) = \|K \cap R\|$$

$$\Phi_4(K, R) = \frac{2\|K \cap R\|}{\|K\| + \|R\|}$$

$$Abrangência = \frac{\Phi_x}{\sum_{i=1} \|K_i\|}$$

$$Precisão = \frac{\Phi_x}{\sum_{i=1} \|R_i\|}$$

Por exemplo, dadas as cadeias de referência:

- C_{K1} = {A, B, C, D, E};
- C_{K2} = {F, G}.

E as cadeias preditas pelo modelo:

- C_{R1} = {A, B, C, D, E}.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

As métricas CEAF utilizam o alinhamento entre as entidades ou menções para calcular seus resultados, dessa forma C_{K1} será alinhado com C_{R1} e C_{K2} não possuirá um alinhamento, dado que o modelo não obteve tal cadeia. Notemos que o número de menções alinhadas é 5. Portanto $\Phi_3 = 5$. Dito isso, temos:

$CEAF_m$: Abrangência $= \frac{5}{7} = 0,71$ (71%) e Precisão $= \frac{5}{5} = 1$ (100%)

Para $CEAF_e$, dado que $\Phi_4 = \frac{2*5}{5+5} = 1$, temos:

Abrangência $= \frac{1}{2} = 0,5$ (50%) e Precisão $= \frac{1}{1} = 1$ (100%)

Notemos que para a métrica $CEAF_m$ o cálculo de precisão e abrangência é realizado considerando a quantidade de menções, para a métrica $CEAF_e$ esse valor é considerado em Φ_4. Contudo, para obtenção dos valores de precisão e abrangência, são usados os valores referentes a quantidade de entidades/cadeias.

12.4.1.4 BLANC:

BiLateral Assessment of NounPhrase Coreference avalia tanto *links* de correferência quanto os de não correferência. Basicamente, um *link* de não correferência é formado por duas menções que não são correferentes entre si. A métrica BLANC tem como objetivo recompensar as cadeias de correferência corretas, de forma proporcional ao seu tamanho. Temos, então, C_K e C_R respectivamente como: *links* de correferência padrão e preditos automaticamente e; N_K e N_R como grupo dos *links* de não correferência padrão e preditos automaticamente; $Abrangência_C$ e $Precisão_C$ remetem ao cálculo de abrangência e precisão dos *links* de correferência, e $Abrangência_N$ e $Precisão_N$, aos *links* de não correferência.

$$Abrangência_C = \frac{\|C_k \cap C_r\|}{C_k}$$

$$Precisão_C = \frac{\|C_k \cap C_r\|}{C_r}$$

$$Abrangência_N = \frac{\|N_k \cap N_r\|}{N_k}$$

$$Precisão_N = \frac{\|N_k \cap N_r\|}{N_r}$$

Por fim, a precisão e a abrangência da métrica BLANC são calculadas, respectivamente, por meio das médias de Precisão e de abrangência, obtidas entre os *links* de correferência e de não correferência:

$$BLANC_{Precisao} = \frac{Precisão_C + Precisão_N}{2}$$

$$BLANC_{Abrangencia} = \frac{Abrangência_C + Abrangência_N}{2}$$

Por exemplo: dados os seguintes links de correferência:

- C_{K1} = {A–B, B–C, C–D, D–E};
- C_{K2} = {F–G}.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

E os seguintes links preditos pelo modelo:

- C_{R1} = {A–B, B–C, C–D, D–E};
- C_{R2} = {F–G, F–I}.

Temos, então: $Abrangência_C = \frac{5}{5} = 1\ (100\%)$, $Precisão_C = \frac{5}{6} = 0,83\ (83\%)$
Considerando que os links de não correferência representam ligações entre todas as menções que não são referenciais, teremos então:

- N_K = {F–A, F–B, F–C, F–D, F–E, G–A, G–B, G–C, G–D, G–E};
- N_R = {F–A, F–B, F–C, F–D, F–E, G–A, G–B, G–C, G–D, G–E, I–A, I–B, I–C, I–D, I–E}.

$$Abrangência_N = \frac{10}{10} = 1\ (100\%)\ ,\ Precisão_C = \frac{10}{15} = 0,67\ (67\%)$$
$$BLANC_{Precisao} = \frac{0,83 + 0,67}{2} = 0,75\ (75\%)$$
$$BLANC_{Abrangencia} = \frac{1+1}{2} = 1\ (100\%)$$

12.4.1.5 CoNLL:

Amplamente utilizada para avaliar modelos de resolução de correferência, a métrica CoNLL calcula um score único, baseando-se no cálculo da medida-f das métricas MUC, B^3 e $CEAF_e$:

$$CoNLL = \frac{(F(MUC) + F(B^3) + F(CEAF_e))}{3}$$

12.5 Aplicações

Os ganhos da tarefa de Resolução de Correferência podem ser significativos, principalmente se considerarmos abordagens que utilizam apoio semântico (Fonseca, 2018; Rahman; Ng, 2011a). Em poucas palavras, existem muitas utilidades para a tarefa, e muitas outras tarefas de PLN podem se beneficiar de tal processamento. Na literatura, encontramos alguns trabalhos que fazem uso de tais modelos, como o de Vargas; Pardo (2018). Na presente abordagem, os autores fazem uso da ferramenta de prateleira chamada CORP (Fonseca et al., 2016c), até o momento a única ferramenta disponível para a língua portuguesa. Em sua produção os autores mostraram que, por meio da resolução de correferências, foi possível obter ganhos significativos na tarefa de Agrupamento de Aspectos para Análise de Sentimentos.

Muitas outras tarefas de PLN podem se beneficiar de seus resultados; como o Reconhecimento de Entidades Nomeadas (REN) (Amaral, 2013), Extração de Relação entre Entidades Nomeadas (Collovini et al., 2014) (ER), entre outras. Na Figura 12.2, no que diz respeito a tarefa de Reconhecimento de Entidades Nomeadas, considerando a cadeia [o agrônomo Miguel Guerra, de a UFSC, Guerra, Guerra, o agrônomo], podemos dizer que o sintagma nominal "Guerra" pode ser ambíguo e existe a possibilidade de que modelos de REN (Capítulo 12) não o classifiquem corretamente. Por meio da tarefa de Resolução de Correferências podemos identificar que a menção "Guerra" corresponde ao agrônomo Miguel Guerra e, portanto, inferir uma mesma categoria de entidade nomeada (Pessoa).

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

No contexto de extração de relação entre entidades nomeadas, considerando o sintagma nominal [o agrônomo Miguel Guerra, de a UFSC] é possível identificarmos a seguinte relação (Miguel Guerra, de, UFSC). E, identificando que "Guerra" faz referência a "Miguel Guerra" é possível inferirmos uma relação direta entre "Guerra" e "UFSC".

Figura 12.2: Resolução de correferência e sua aplicabilidade (imagem extraída utilizando a ferramenta CORP (Fonseca et al., 2016c))

A discussão sobre a biotecnologia nacional está enviesada , pois está sendo entendida como sinônimo de transgenia . A opinião é de **[o agrônomo Miguel Guerra, de a UFSC [5]]** (Universidade Federal de Santa Catarina) . **[Guerra [5]]** participou de o debate "Biotecnologia para uma Agricultura Sustentável" , realizado ontem durante a 52ª Reunião_Anual de a SBPC (Sociedade Brasileira para o Progresso da Ciência) , sobre as biotecnologias apropriadas a o desenvolvimento de o país. **[Guerra [5]]** citou a micropropagação de vegetais (produção de mudas em laboratório , feita para evitar doenças e selecionar vegetaissaudáveis) como exemplo de biotecnologia de baixo custo . Com ela , aumentou -se a produção de moranguinho , em o sul de o país , de 3,2 kg para 60 kg por hectare . Para **[o agrônomo [5]]**, o Brasil deve...

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Parte VII

Geração e Interação

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Capítulo 13

Geração de linguagem natural

Thiago Castro Ferreira
Ivandré Paraboni

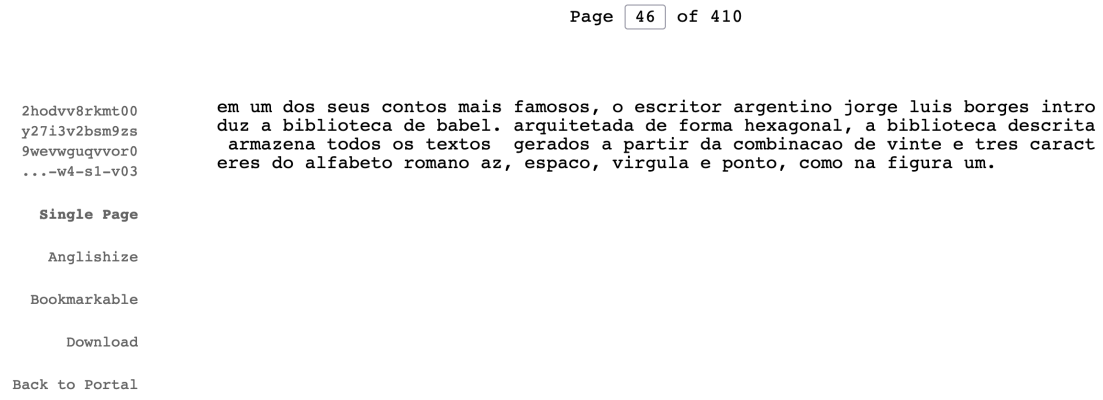
Publicado em: 20/11/2024

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

13.1 Introdução

Em um dos seus contos mais famosos, o escritor Argentino Jorge Luis Borges introduz a Biblioteca de Babel. Arquitetada de forma hexagonal[1], a biblioteca descrita armazena todos os textos gerados a partir da combinação de 23 caracteres do alfabeto romano: A-Z, espaço, vírgula e ponto, como na Figura 13.1.

Figura 13.1: Versão do primeiro parágrafo deste capítulo, encontrado no projeto "The Library of Babel".



O conto sugere o cenário (distópico?) onde todos os textos já escritos, assim como todos que ainda serão escritos, como jornais, livros, este próprio texto e o conto do autor, estão presentes na biblioteca universal. Apesar de alertar que esta pode ser a questão errada a ser feita, Gatt; Krahmer (2018) questionam quem poderia escrever todos estes documentos. Poderia um computador gerar esta biblioteca?

Em teoria, tal questionamento pode ser facilmente respondido através de um algoritmo simples (mas pouco elegante e eficiente) como o representado no algoritmo da Figura 13.2[2], capaz de gerar (em um tempo infinito) tal biblioteca com livros de até 50 mil caracteres.

[1] A figura geométrica sem curvas mais próxima de um círculo.
[2] Pseudocódigo gerado por GPT-4 Turbo e traduzido e revisado pelos autores.

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Um exemplo mais elaborado é o gerado pelo projeto *The Library of Babel*[3] que, de fato, disponibiliza uma variação da biblioteca descrita pelo autor argentino. De maneira digital, o projeto disponibiliza uma biblioteca de câmaras hexagonais, cada uma com 4 paredes de estantes, cinco estantes por parede e 32 volumes por estante. Cada volume consiste em um texto de 410 páginas representando combinações dos 23 caracteres. O projeto, que talvez seja a melhor concretização da ideia de Borges, nos permite buscar por trechos de até 3200 caracteres, como, por exemplo, uma versão representada em 23 caracteres do parágrafo anterior, presente na página 329 do volume 26 da quinta estante da quarta parede de uma das câmaras hexagonais da biblioteca (Figura 13.1). Páginas aleatórias também podem ser acessadas, como a disponibilizada na Figura 13.3, representando uma combinação completamente aleatória e sem sentido. Tal exemplo nos permite apontar uma das deficiências desta biblioteca: apesar de armazenar todo o conhecimento presente em todos os livros que já ou ainda serão escritos, estes se encontram entre uma vasta gama de conteúdo sem sentido. Neste contexto, como encontrar e catalogar os textos que, de fato, contêm um significado?

Figura 13.2: Gerando a Biblioteca de Babel

Algoritmo 1 Gerando a Biblioteca de Babel

```
1: procedure GenerateBabel(n)
2:     caracteres ← ['a', 'b', ..., 'z', ',', '.', ' ']        ▷ lista com os 23 caracteres
3:     total_livros ← ∞                          ▷ Definir número de livros como infinito
4:     livro_tamanho ← 50,000                    ▷ Definir tamanho de livros como 50,000
   caracteres
5:     for livro ← 1 to total_livros do
6:         texto ← string vazia
7:         for i ← 1 to livro_tamanho do
8:             char_indice ← Random(1, |caracteres|)  ▷ Definindo um caractere
   aleatório
9:             texto ← texto + caracteres[char_indice]
10:        end for
11:        Print(texto)                              ▷ Imprimindo o livro gerado
12:    end for
13: end procedure
```

Voltando ao conto de Borges, talvez mais interessante que a biblioteca em si seja a figura do bibliotecário. Na vasta quantidade de textos sem significado ou sem ao menos seguir o padrão gramatical de uma língua como o da Figura 13.3, cabe a este o trabalho de catalogar os volumes que de fato estejam escritos em uma linguagem natural e contenham um significado. Neste cenário, seria onisciente o bibliotecário que conseguisse catalogar todo o conhecimento contido na biblioteca. Como mostramos, até certo ponto uma máquina poderia gerar a biblioteca de Babel com suas deficiências, mas seria esta capaz de se tornar a figura sacra do bibliotecário onisciente capaz de adquirir e gerar todos textos relevantes já escritos ou que serão escritos? Talvez seja a geração de linguagem natural que busque

[3] https://libraryofbabel.info

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Figura 13.3: Primeira página de um texto buscado aleatoriamente no projeto "The Library of Babel".



responder esta pergunta.

13.1.1 Geração de Linguagem Natural

Geração de linguagem natural é o campo de estudo que visa desenvolver soluções computacionais para gerar automaticamente linguagem natural, no formato de texto ou de voz, a partir de dados linguísticos ou não linguísticos como imagens, vídeos e representações de significado diversas (Ferreira, 2018; Gatt; Belz, 2010; Reiter; Dale, 2000). Esta ciência, também conhecida pelo acrônimo GLN (ou NLG, em inglês) remonta o início da Ciência da Computação, como data o artigo (Yngve, 1961) escrito no século 20 por Victor H. Yngve[4], o qual apresenta o primeiro programa de computador para produção de sentenças gramaticais em inglês[5]. O objetivo do autor era avaliar a adequação da gramática, gerando sentenças sem necessariamente ter um objetivo comunicativo ou significado.

[4]Professor de Linguística da Universidade de Chicago e do Instituto de Tecnologia de Massachusetts (MIT).

[5]As sentenças foram automaticamente geradas a partir do livro infantil chamado Engineer Small and the Little Train.

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

Como expresso no próprio artigo, as sentenças geradas "eram em sua maior parte bastante gramaticais, embora, é claro, sem sentido". Em uma relação com a Biblioteca de Borges, o programa apresentado em (Yngve, 1961) se enquadraria em um bibliotecário capaz de gerar textos que seguem a gramática da língua inglesa.

Mais tarde, surgem os primeiros sistemas de GLN capazes de gerar texto a partir de um objetivo comunicativo ou semântico fornecido como entrada. Este é o caso de (Simmons; Slocum, 1972), por exemplo, que propõe um sistema de geração de sentenças em inglês a partir de redes semânticas de entrada. Cada vez mais convergindo com o objetivo de um bibliotecário do conto de Borges, tais sistemas passaram a ser capazes de catalogar e expressar um significado ou objetivo de comunicação dado como entrada em um texto gramaticalmente correto expresso em uma ou mais línguas.

13.1.2 Seleção de Conteúdo e Realização Textual

Em GLN, a tarefa de converter a representação semântica em um texto é chamada de "Realização Textual". Em outras palavras, sistemas como os propostos por (Yngve, 1961) e (Simmons; Slocum, 1972) são Realizadores Textuais focados em "como" gerar um texto que verbalize o conteúdo. Contudo, prévio à etapa de Realização Textual, também cabe a um sistema de geração de linguagem natural a tarefa de selecionar "qual" conteúdo deve ser expresso. Esta tarefa é popularmente conhecida como "Seleção de Conteúdo". Salvo algumas exceções (Appelt, 1980), diferentes arquiteturas de GLN têm em comum a organização nessas duas etapas: primeiro a Seleção de Conteúdo, responsável por selecionar o conteúdo a ser expresso (e.g. uma imagem, um conjunto de textos de contexto a uma pergunta, um livro a ser sumarizado etc.) seguido da fase de Realização Textual, que converte o conteúdo selecionado em texto (e.g. descrição de uma imagem, resposta a uma pergunta, o sumário de um livro etc.).

13.1.3 Avaliação de sistemas de GLN

Um aspecto crítico do projeto de sistemas de GLN é como avaliar a sua qualidade. Ao contrário de aplicações de interpretação de linguagem natural (e.g. análise de sentimentos, classificação de textos etc.), cujo objetivo é extrair algum tipo de significado do texto, e que podem assim ser facilmente avaliadas pela contagem de seus erros e acertos, sistemas de GLN não podem ser avaliados desta forma. Via de regra, não existe uma forma única e correta de expressar um significado em forma textual, mas sim diversas alternativas possíveis com maior ou menor grau de gramaticalidade, fluência e outras características, o que torna a avaliação de sistemas deste tipo uma questão de pesquisa tão antiga quanto a própria história destes sistemas.

Assim como o processo computacional de GLN em duas etapas, sua avaliação também é normalmente feita em duas fases. Os textos gerados por sistemas de GLN são normalmente julgados pela sua **adequação**, i.e. se o texto gerado atende ao propósito ou tarefa específica para o qual foi criado, e pela sua **fluência**, i.e. se o texto gerado tem qualidade linguística, incluindo sua gramática, sintaxe e estilo. Por exemplo, se dado um estímulo como "Quem é Elis Regina?" um sistema de GLN gerasse o texto "Elza Soares foi uma cantora, compositora musical e intérprete de samba-enredo brasileira", podemos considerar o resultado como fluente mas não adequado. Por outro lado, se a resposta fosse "Regina cantora brasileira", este texto seria avaliado como adequado mas sem fluência.

A avaliação de um sistema de GLN pode ser conduzida de maneira automática ou manual, esta última também chamada de avaliação humana. De maneira automática,

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

destacam-se métricas tradicionais como BLEU (Papineni et al., 2002), METEOR (Lavie; Agarwal, 2007), chrF++ (Popović, 2017), dentre outras. Estas métricas determinam a qualidade de um texto gerado automaticamente comparando-o com um texto de referência, geralmente conhecido como "padrão ouro". Esta comparação é realizada no nível de *tokens*, e quanto mais alto o nível de sobreposição entre os *tokens* do texto gerado e os *tokens* do padrão ouro, maior será o valor destas métricas e, por extensão, da qualidade do texto gerado. Estudos como (Callison-Burch et al., 2010) mostram que métricas como BLEU (Subseção *Bilingual Evaluation Under-study* (BLEU)), a métrica automática mais popular, tende a explicar melhor a fluência de um texto do que a adequação. Outros estudos, porém, demonstram que tais métricas tradicionais não correlacionam com avaliações humanas, questionando se as mesmas, de fato, explicam de maneira fidedigna a qualidade de um texto (Novikova et al., 2017; Reiter, 2018).

Visando resolver os problemas de métricas tradicionais por sobreposição de *tokens*, métricas de qualidade estado-da-arte funcionam a partir da comparação dos *embeddings* extraídos dos textos gerados e o padrão ouro. BLEURT (Sellam et al., 2020), BERTScore (Zhang et al., 2020) e COMET (Rei et al., 2020) são alguns exemplos de métricas deste tipo. Estudos demonstram que estas métricas têm maior correlação com avaliações humanas do que as métricas tradicionais (Ferreira et al., 2020), apesar de serem mais custosas do ponto de vista computacional.

Para a avaliação precisa de um sistema de geração de linguagem natural, as avaliações humanas são geralmente consideradas a melhor opção. Estas são normalmente conduzidas na forma de testes cegos onde avaliadores humanos são apresentados a textos a serem avaliados sem revelar sua autoria. Os avaliadores são geralmente solicitados a atribuírem uma nota em uma escala Likert (e.g. de 1 a 5, onde 1 é muito ruim e 5 muito bom) em dimensões como fluência e adequação. Para um guia de melhores práticas sobre a avaliação humana em GLN, (Lee et al., 2021) pode ser consultado.

13.1.4 Arquiteturas de GLN

Apesar de terem pontos em comum, arquiteturas de sistemas de GLN vêm evoluindo ao longo do tempo, passando de sistemas tradicionais baseado em regras para sistemas que representam o atual estado-da-arte, conhecidos por serem orientados a dados e desenvolvidos a partir de técnicas de Inteligência Artificial. Neste capítulo, buscaremos explicar esta evolução a partir de quatro arquiteturas distintas de GLN. Duas destas arquiteturas, mais tradicionais, são a arquitetura baseada em *templates* e a arquitetura *pipeline*. As outras duas, mais recentes, são as arquiteturas baseadas em Redes Neurais e as baseadas em modelos pré-treinados, também conhecidas como Inteligências Artificiais Generativas.

13.2 Arquitetura Baseada em Templates

A arquitetura de GLN a partir de *templates* é a mais simples entre as apresentadas neste capítulo. Seguindo a mesma divisão em fases de Seleção de Conteúdo e Realização Textual de outras arquiteturas, neste caso são utilizados *templates* pré-determinados para verbalizar cada tipo de conteúdo selecionado. Os *templates* geralmente contam com marcações específicas - ou *tags* - que no texto final são substituídas por valores selecionados na primeira fase da geração.

Reiter; Dale (2000) definem os *templates* como "textos enlatados" (*canned texts*) que, de acordo com a definição, são especificações de conteúdo onde a sequência de caracteres

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

a ser usada já foi determinada, embora algum processamento ortográfico ainda possa ser necessário. Estes processamentos podem incluir, por exemplo, apresentar a primeira letra da frase como maiúscula se o texto em questão for uma frase completa, colocar (em certos contextos) uma frase utilizada como título em letra maiúscula, apresentar elementos específicos do texto em negrito ou itálico para indicar ênfase ou algum outro propósito, ou ainda determinar a pontuação final da frase dependendo se esta for uma afirmação, uma pergunta, um imperativo, ou se serve como introdução a uma lista detalhada.

13.2.1 Aplicações

A arquitetura baseada em *templates* é adotada, por exemplo, em diversos robôs jornalistas brasileiros na antiga plataforma Twitter. Alguns exemplos são sumarizados a seguir.

- Robotox[6] é um robô jornalista criado no projeto "Por trás do Alimento"[7] que cria um tweet na antiga plataforma Twitter a toda nova liberação de um agrotóxico pelo Governo Federal.
- colaborabot[8] é um sistema de linguagem natural criado pelo grupo Colaboradados[9] para monitorar o acesso aos portais de transparência pública governamentais. Na fase de seleção de conteúdo, o sistema identifica os portais web governamentais que se encontram fora do ar e, na etapa de Realização Textual, cria um texto informando sobre o problema, para que então o texto seja publicado na antiga plataforma Twitter.
- Rui[10] é um robô desenvolvido pelo portal JOTA, e visa identificar e reportar os processos parados no Supremo Tribunal Federal brasileiro.
- "Elas no Congresso"[11] é um robô jornalista de iniciativa da revista AzMina que acompanha e publica sobre a tramitação de proposições que tratam dos direitos das mulheres no Congresso Brasileiro.
- Rosie[12], do grupo "Operação Serenata de Amor", analisa e identifica suspeitas em gastos de deputados federais em exercício de sua função no congresso Brasileiro.

13.2.2 Exemplo – Sistema Rosie

A Figura 13.4[13] mostra exemplos de publicações feitas pelo sistema Rosie, robô que analisa e identifica suspeitas em gastos de deputados federais em exercício de sua função no congresso brasileiro.

Como pode ser deduzido a partir deste exemplos, os textos produzidos seguem o mesmo *template* geral:

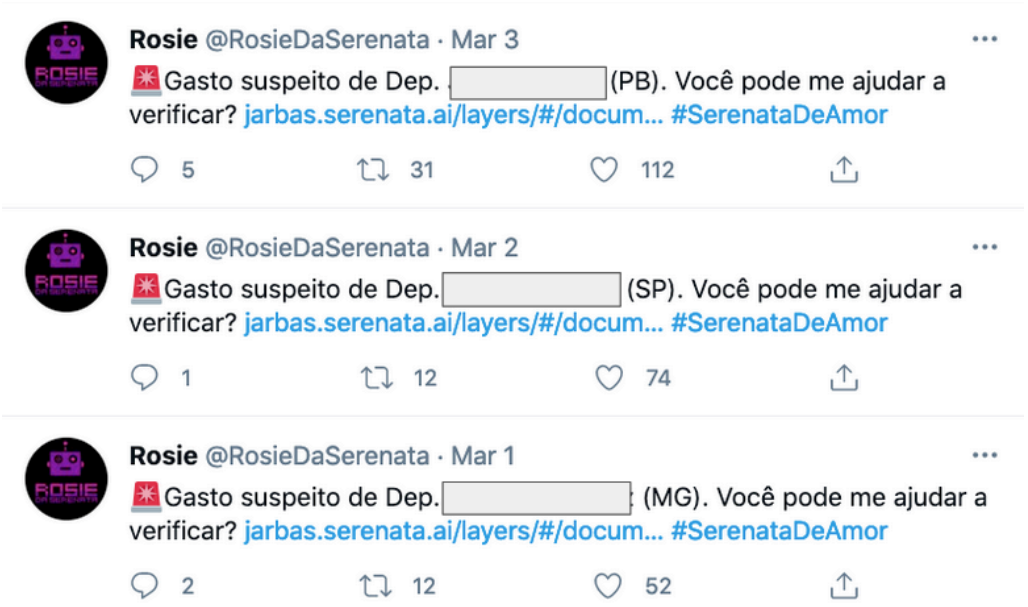
```
Gasto suspeito de {TITULO} {NOME} ({PARTIDO}).
Você pode me ajudar a verificar?
```

[6] https://x.com/orobotox
[7] Parceria entre a Agência Pública e *Repórter Brasil*.
[8] https://x.com/colabora_bot
[9] https://colaboradados.github.io
[10] https://x.com/ruibarbot
[11] https://x.com/elasnocongresso
[12] https://x.com/RosieDaSerenata
[13] Extraída de https://x.com/RosieDaSerenata.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Figura 13.4: Publicações geradas pelo sistema Rosie.



O *template* consiste de três *tags*, `TITULO`, `NOME` e `PARTIDO`, as quais são substituídas pelos valores selecionados durante a fase de seleção de conteúdo. Por exemplo, consideremos a lista fictícia apresentada na Tabela 13.1 com nomes de deputados com seus respectivos partidos e gastos.

Tabela 13.1: Demonstração fictícia da etapa de Seleção de Conteúdo

Título	Nome	Partido	Gasto (R$)	Alertas
Dep.	Antonio	PF	1500	1
Dep.	João	PJ	500	0
Dep.	Maria	PK	450	2

Neste exemplo, um gasto poderia ser classificado como suspeito se excedesse a média de gastos de todos os deputados. Caso esta tabela fosse dada como entrada para a etapa de Seleção de Conteúdo do sistema, tendo em vista que a média de gastos é de R$ 800, o gasto do deputado Antonio seria identificado como suspeito. Para publicação de um texto sobre este gasto, bastaria assim substituir as informações da tabela no *template* pré-estabelecido: "Gasto suspeito de **Dep. Antonio** (**PF**). Você pode me ajudar a verificar?"

13.2.3 Limitações

Arquiteturas baseadas em *templates* são possivelmente ideais para verbalização de mensagens simples, mas exibem problemas a medida que o volume de geração e a complexidade dos textos gerados aumentam. Como pode ser visto na Figura 13.4, os três textos da imagem seguem o mesmo formato geral. Uma estrutura engessada deste tipo pode incomodar o leitor, como argumentado em (Clerwall, 2014), onde notícias geradas automaticamente foram avaliadas como "chatas" apesar de também serem avaliadas como mais informativas, precisas, confiáveis e objetivas que notícias escritas por humanos. Além disso, a abordagem

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

baseada em *templates* apresenta problemas para verbalizar estruturas de linguagem mais complexas como, por exemplo, as que exigem flexão de tempo, número e gênero. Para o mesmo exemplo anterior, consideremos que a abordagem Serenata de Amor selecionasse e gerasse um texto a partir do seguinte *template*:

Gasto suspeito de TITULO NOME (PARTIDO). Você pode me ajudar a verificar? O parlamentar possui ALERTAS alertas no mês.

De acordo com ambas mensagens, o gasto suspeito do deputado Antônio, selecionado a partir dos dados na Tabela 13.1, seria verbalizado da seguinte maneira:

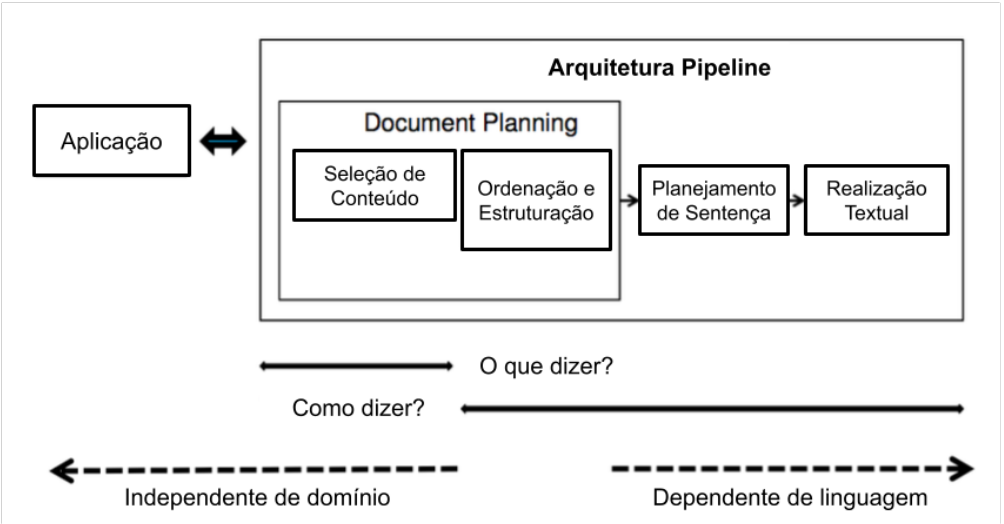
Gasto suspeito de **Dep. Antonio** (**PF**). Você pode me ajudar a verificar? O parlamentar possui **1** alertas no mês.

Percebe-se então que a flexão de número está incorreta, já que o deputado possui somente um alerta no mês, enquanto o objeto direto da segunda mensagem encontra-se no plural. O mesmo aconteceria para verbalizar gastos suspeitos da deputada Maria, já que o sujeito da segunda sentença, "O parlamentar", estaria com a flexão de gênero incorreta. Questões deste tipo são naturalmente difíceis de lidar em um esquema baseado em *templates* fixos, e acabam limitando seu uso em aplicações mais sofisticadas.

13.3 Arquitetura *pipeline*

A arquitetura *pipeline* visa gerar linguagem natural por meio de várias transformações da representação de entrada. Essa entrada é composta de vários módulos especialistas em diferentes funções do processo. Cada módulo gera uma representação intermediária para o próximo módulo, até a obtenção do texto final. A arquitetura *pipeline* é a mais popular entre as arquiteturas clássicas de geração de linguagem natural, popularizada por Reiter; Dale (2000) com a proposta de geração automática de linguagem natural em três fases sequenciais: (1) Planejamento de Documento, ou Macro-planejamento; (2) Planejamento de Sentença, ou Micro-planejamento; e (3) Realização Textual. A Figura 13.5 demonstra esta arquitetura.

Figura 13.5: Arquitetura *pipeline* de Reiter; Dale (2000).



Fonte: (Santhanam; Shaikh, 2019)

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

O Planejamento do Documento contempla a seleção do conteúdo e sua estruturação na forma de de capítulos, seções, parágrafos, sentenças etc. O Macro-planejamento enfoca o planejamento de cada sentença do texto, incluindo as escolhas lexicais e das formas referenciais a serem usadas para mencionar as entidades-alvo do discurso. Por último, a fase de Realização Textual é responsável pelos ajustes finais do texto, como a conjugação dos verbos no formato correto de tempo, aspecto, modo e voz, bem como a concordância entre substantivos e verbos. Muitas vezes, os realizadores também precisam inserir palavras funcionais como verbos auxiliares e preposições e sinais de pontuação (Gatt; Krahmer, 2018).

Figura 13.6: Arquitetura *pipeline* de Castro Ferreira et al. (2019) aplicada como o sistema "DaMata", um Robô-Jornalista Cobrindo o Desmatamento na Amazônia Brasileira.



Fonte: (Teixeira et al., 2020)

Mais recentemente, Castro Ferreira et al. (2019) propuseram uma versão adaptada da arquitetura *pipeline* de Reiter; Dale (2000) onde cada módulo é implementado com o uso de redes neurais. A proposta converte uma representação não-linguística de entrada em texto a partir de seis etapas sequenciais: Seleção de Conteúdo, Ordenação de Discurso, Estruturação de Texto, Lexicalização, Geração de Expressões de Referência e Realização Textual. A Figura 13.6 mostra o robô-jornalista "DaMata", desenvolvido a partir da arquitetura *pipeline* de Castro Ferreira et al. (2019) para cobrir o desmatamento na Amazônia Brasileira.

Na proposta de Castro Ferreira et al. (2019), as três etapas iniciais - Seleção de Conteúdo, Ordenação de Discuso e Estruturação de Texto - são parte do Planejamento de Documento, e visam selecionar, ordenar e estruturar o conteúdo a ser verbalizado em seções. Já no Planejamento de Sentença, o módulo de Lexicalização visa encontrar as frases e palavras adequadas para expressar o conteúdo a ser incluído em cada frase (Reiter et al., 2005; Smiley et al., 2016), enquanto a fase de Geração de Expressões de Referência envolve referenciar as entidades a serem mencionadas do texto (Krahmer; Deemter, 2012). Por fim, a Realização Textual é responsável pelos últimos ajustes para obter o texto final no idioma alvo de interesse.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

13.3.1 Aplicações

Até recentemente, a arquitetura *pipeline* era a mais popular na geração de linguagem natural, fazendo com que esta abordagem tenha sido aplicada na geração automática de textos de diversos domínios, como nos esportes (Lee et al., 2017a; Theune et al., 2001), previsão do tempo (Goldberg et al., 1994; Sripada et al., 2004), previsão de polenização (Turner et al., 2006), resumos de mergulhos autônomos (Sripada; Gao, 2007), descrições de eventos em turbinas a gás (Yu et al., 2007), informações do mercado de ações (Kukich, 1983), auxílio no tratamento do tabagismo (Reiter et al., 2003), relatórios de cuidados intensivos neonatais (Portet et al., 2009; Reiter, 2007), geração de dados enciclopédicos (Androutsopoulos et al., 2013; Duma; Klein, 2013) e de notícias (Campos et al., 2020; Teixeira et al., 2020).

Muitos estudos também focaram em aspectos específicos do processo como Seleção de Conteúdo (Barzilay; Lapata, 2005; Bouayad-Agha et al., 2012; Perez-Beltrachini et al., 2016), Estruturação de Texto/Agregação (Barzilay; Lapata, 2006; Bayyarapu, 2011), Lexicalização (Reiter et al., 2005; Smiley et al., 2016) e Geração de Expressões de Referência (Castro Ferreira et al., 2018; Castro Ferreira; Paraboni, 2014a; Dale; Haddock, 1991; Deemter, 2016a; Krahmer; Theune, 2002; Van Deemter et al., 2012).

13.3.2 Exemplo

Para exemplificar a arquitetura *pipeline*, utilizaremos o exemplo do CoronaReporter (Campos et al., 2020), robô jornalista desenvolvido com base na arquitetura de Castro Ferreira et al. (2019) para reportar os casos de COVID-19 no Brasil.

13.3.2.1 Seleção de Conteúdo

A fase de Seleção de Conteúdo do CoronaReporter visa selecionar um subconjunto das oito mensagens seguintes:

- `ACTIVE_CASES`: número de casos ativos de COVID-19 no Brasil no dia atual;
- `TOTAL_CASES`: número total de casos de COVID-19 no Brasil;
- `NEW_CASES`: número de novos casos de COVID-19 no Brasil no dia atual;
- `ACTIVE_CASES_VARIATION_LAST_DAY`: variação de casos ativos de COVID-19 no Brasil em comparação ao dia anterior;
- `CASES_VARIATION_LAST_DAY`: variação de casos ativos e inativos de COVID-19 no Brasil em comparação ao dia anterior;
- `NEW_DEATHS`: número de mortes por COVID-19 no Brasil no dia atual;
- `TOTAL_DEATHS`: número total de mortes por COVID-19 no Brasil;
- `DEATHS_VARIATION_LAST_DAY`: variação no número de mortes por COVID-19 no Brasil em comparação ao dia anterior;

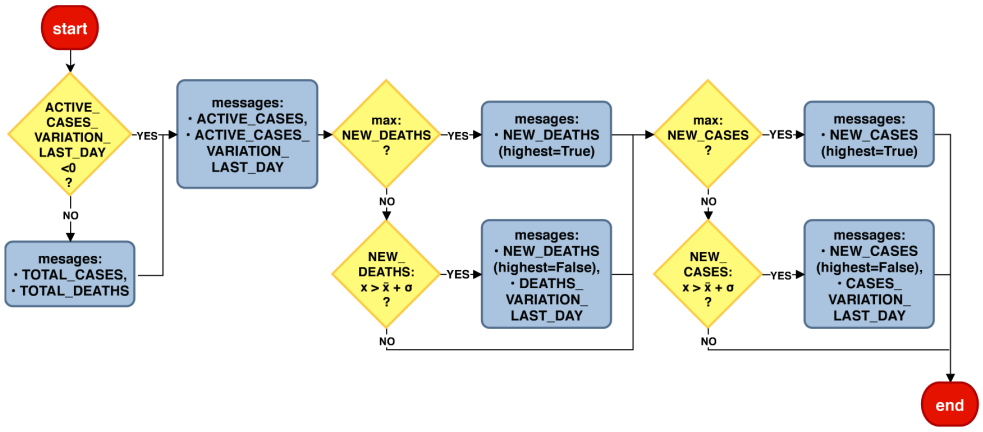
A Figura 13.7, extraída de Campos et al. (2020), mostra o processo baseado em regras de Seleção de Conteúdo do robô-jornalista. Caso o número de casos ativos tenha diminuído em comparação ao dia anterior, este número (`ACTIVE_CASES(cases)`) e sua variação

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

(`ACTIVE_CASES_VARIATION_LAST_DAY(variation, trend=low)`) são selecionados. Caso contrário, a variação de casos tenha se mantido ou aumentado, os números do total de casos (`TOTAL_CASES(cases)`) e de mortes (`TOTAL_DEATHS(deaths)`) são selecionados.

Figura 13.7: Módulo de Seleção de Conteúdo do robô-jornalista "CoronaReporter".



Fonte: (Campos et al., 2020)

Na sequência, o módulo de Seleção de Conteúdo analisa o número de mortes pela doença. Caso o número de mortes seja o maior da série histórica, esta mensagem é selecionada com um atributo que aponta para a ordem de grandeza: `NEW_DEATHS(deaths, highest=True)`. Caso contrário, o módulo checa se o número de mortes é relevante para ser informado ao comparar se este é maior que a média de mortes mais o desvio padrão da série histórica. Em caso positivo, as mensagens correspondentes ao número de mortes no dia ((`NEW_DEATHS(deaths)`)) e sua variação (`DEATHS_VARIATION_LAST_DAY(variation, trend=high)`) são selecionadas.

Por fim, o número de novos casos é analisado da mesma maneira que o número de novas mortes. Caso o número de casos ativos seja o maior da série histórica, esta mensagem é selecionada com um atributo que aponta para a ordem de grandeza: `NEW_CASES(cases, highest=True)`. Caso contrário, o módulo checa se o número de casos ativos é maior que a média de casos mais o desvio padrão da série histórica. Em caso positivo, as mensagens correspondentes ao número de casos ativos no dia ((`NEW_CASES(cases)`)) e a variação (`CASES_VARIATION_LAST_DAY(variation, trend=high)`) são selecionadas.

Para exemplificar as próximas etapas, consideremos as seguintes mensagens selecionadas:

```
DEATHS_VARIATION_LAST_DAY(trend="high", variation=0.06)
NEW_CASES(cases=15305, highest=True)
NEW_DEATHS(deaths=824, highest=False)
TOTAL_CASES(cases=218223)
TOTAL_DEATHS(deaths=14817)
```

13.3.2.2 Ordenação do Discurso

Uma vez feita a seleção de conteúdo, a etapa de Ordenação de Discurso visa disponibilizar as mensagens selecionadas na ordem que devem aparecer no texto. A ordenação pode ser

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

feita por meio de regras ou de técnicas de aprendizagem de máquina, como em Castro Ferreira et al. (2019).

No nosso exemplo, as mensagens selecionadas poderiam ser ordenadas da seguinte maneira:

```
TOTAL_CASES(cases=218223)
NEW_CASES(cases=15305, highest=True)
NEW_DEATHS(deaths=824, highest=False)
TOTAL_DEATHS(deaths=14817)
DEATHS_VARIATION_LAST_DAY(trend="high", variation=0.06)
```

13.3.2.3 Estruturação do Texto

Última fase do módulo de Planejamento de Documento, a estruturação do texto visa agrupar as mensagens selecionadas em capítulos, seções, parágrafos, sentenças ou outras unidades textuais de interesse. No nosso exemplo, as mensagens selecionadas e ordenadas podem ser organizadas em um único parágrafo de duas sentenças da seguinte maneira:

```
<PARAGRAPH>
    <SENTENCE>
        TOTAL_CASES(cases=218223)
        NEW_CASES(cases=15305, highest=True)
    </SENTENCE>
    <SENTENCE>
        NEW_DEATHS(deaths=824, highest=False)
        TOTAL_DEATHS(deaths=14817)
        DEATHS_VARIATION_LAST_DAY(trend="high", variation=0.06)
    </SENTENCE>
</PARAGRAPH>
```

13.3.2.4 Lexicalização

A fase de Lexicalização, que é a primeira no Planejamento de Sentença, visa encontrar as frases e palavras adequadas para expressar o conteúdo a ser verbalizado. No exemplo do sistema CoronaReporter (Campos et al., 2020), os desenvolvedores definiram *templates* de lexicalização para cada estrutura de sentença. Para a primeira sentença do nosso exemplo, composta pelas mensagens de total (`TOTAL_CASES`) e novos casos com máxima histórica de COVID ((`NEW_CASES(highest=True)`)), um exemplo de *template* seria:

```
COUNTRY_1 VP[aspect=simple,tense=present,voice=active,
person=COUNTRY_1,number=COUNTRY_1] totalizar CASES_1
NN[number=CASES_1,gender=male] caso de #COVID-19 , CASES_2
NN[number=CASES_2,gender=male] caso a mais de o que
em o dia anterior .
```

Assim como na arquitetura baseada em *templates*, *tags* são usadas para que as referências às entidades possam ser inseridas pelo módulo subsequente de Geração de Expressões de Referência. Contudo, os *templates* de lexicalização da arquitetura *pipeline* são mais flexíveis que os da arquitetura baseada em *templates* ao representar parte das palavras

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

abertas em sua forma base, antecedidas pelas informações de como devem ser flexionadas nas etapas posteriores. Por exemplo, o verbo "totalizar" seria representado no *template* da seguinte forma:

```
VP[aspect=simple,tense=present,voice=active,
person=COUNTRY_1,number=COUNTRY_1] totalizar
```

Veja que o verbo é definido no aspecto simples, no tempo presente e na voz ativa. Já a pessoa e o número são definidos de acordo com a entidade a ser referenciada na *tag* `COUNTRY_1`.

No exemplo anterior, todas as sentenças poderiam ser lexicalizadas da seguinte maneira:

```
<PARAGRAPH>
    <SENTENCE>
        COUNTRY_1 VP[aspect=simple,tense=present,voice=active,
        person=COUNTRY_1,number=COUNTRY_1] totalizar CASES_1
        NN[number=CASES_1,gender=male] caso de #COVID-19 ,
        CASES_2 NN[number=CASES_2,gender=male] caso a mais
        de o que em o dia anterior .
    </SENTENCE>
    <SENTENCE>
        Desde ontem , VP[aspect=simple,tense=past,voice=passive,
        person=DEATHS_1,number=DEATHS_1,gender=female] registrar
        em COUNTRY_1 DEATHS_1 ADJ[number=DEATHS_1,gender=female]
        novo NN[number=DEATHS_1,gender=female] morte , que agora
        VP[aspect=simple,tense=present,voice=active,
        person=DEATHS_2,number=DEATHS_2] somar DEATHS_2 ,
        representando DT[number=singular,gender=TREND_1] um TREND_1
        ADJ[number=singular,gender=TREND_1] diário de VARIATION_1 .
    </SENTENCE>
</PARAGRAPH>
```

13.3.2.5 Geração de Expressões de Referência

Após a etapa de Lexicalização, o módulo de Geração de Expressões de Referência do sistema CoronaReporter tem primeiramente a tarefa de mapear as *tags* dos *templates* para as entidades a serem referenciadas. Em seguida, deve gerar as expressões de referência para essas entidades, levando em consideração o contexto de cada uma, e atualizar as informações de pessoa e gênero dos termos relevantes.

Para a tarefa de mapeamento, o resultado seria algo como:

```
COUNTRY_1 -> Brasil
CASES_1 -> 218223
CASES_2 -> 15305
DEATHS_1 -> 824
DEATHS_2 -> 14817
TREND_1 -> high
VARIATION_1 -> 0.06
```

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Uma vez feito o mapeamento entre as *tags* e entidades, as expressões de referência a cada entidade devem ser geradas e colocadas no lugar das respectivas *tags* com base no contexto onde se encontram. Por exemplo, a primeira referência ao país "Brasil" na primeira sentença pode ser feita usando-se o seu nome próprio, ou seja, "O Brasil". Já a segunda referência na segunda sentença pode ser feita através de uma descrição como ("o país") para manter a fluência do texto evitando que se torne repetitivo. Juntamente com a geração das expressões de referência a uma entidade, as informações de gênero e número das palavras abertas afetadas podem ser atualizadas. Por exemplo, como a entidade "Brasil" se encontra na terceira pessoa do singular, o verbo "totalizar" na primeira sentença pode ser atualizado de acordo:

```
VP[aspect=simple,tense=present,voice=active,
person=3rd,number=singular] totalizar
```

Por fim, a representação intermediária após o módulo de Geração de Expressão de Referência para o nosso exemplo poderia ser feita da seguinte forma:

```
<PARAGRAPH>
    <SENTENCE>
        O Brasil VP[aspect=simple,tense=present,voice=active,
        person=3rd,number=singular] totalizar 218,223
        NN[number=plural,gender=male] caso de #COVID-19 ,
        15,305 NN[number=plural,gender=male] caso a mais
        de o que em o dia anterior .
    </SENTENCE>
    <SENTENCE>
        Desde ontem , VP[aspect=simple,tense=past,voice=passive,
        person=3rd,number=plural,gender=female] registrar
        em o país 824 ADJ[number=plural,gender=female]
        novo NN[number=plural,gender=female] morte , que agora
        VP[aspect=simple,tense=present,voice=active,
        person=3rd,number=plural] somar 14,817 ,
        representando DT[number=singular,gender=female] um alta
        ADJ[number=singular,gender=female] diário de 6% .
    </SENTENCE>
</PARAGRAPH>
```

A Geração de Expressões de Referência é uma tarefa complexa, envolvendo não apenas a seleção de conteúdo adequado a um determinado contexto como também medidas para evitar, por exemplo, a geração de expressões ambíguas, entre outros objetivos muitas vezes conflitantes. Assim, por exemplo, além de selecionar o conteúdo adequado para referenciar uma entidade (e.g. "o deputado maranhense"), o algoritmo precisa garantir que este conteúdo especifica a entidade-alvo (o deputado em questão) de forma única naquele contexto, ou seja, sem risco de ser confundido com outro alvo (e.g. outros deputados maranhenses) em discussão. O fenômeno de referência tem sido um tópico de interesse recorrente na GLN baseada na arquitetura *pipeline* e áreas correlatas, tanto no que diz respeito à proposta de abordagens computacionais (Castro Ferreira; Paraboni, 2014a, 2014b; Deemter, 2016b; Krahmer et al., 2003; Santos Silva; Paraboni, 2015) como conjuntos de dados anotados com informação semântica (Gargett et al., 2010; Gatt et al., 2007; Paraboni et al., 2017; Viethen; Dale, 2011).

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

13.3.2.6 Realização Textual

A fase de Realização Textual é responsável pelos últimos ajustes para obtenção do texto final, como a flexão das palavras no formato correto de acordo com o seu contexto. Por exemplo, o verbo "totalizar" da primeira sentença seria flexionado como:

```
VP[aspect=simple,tense=present,voice=active,
person=3rd,number=singular] totalizar -> totaliza
```

O substantivo "caso" seria flexionado no plural como indicado:

```
NN[number=plural,gender=male] caso -> casos
```

O artigo indefinido "um" na última sentença seria verbalizado na forma singular e feminina:

```
DT[number=singular,gender=female] um -> uma
```

Além da correta flexão das palavras abertas, as contrações do português seriam resolvidas ("em o" `->` "no") e o espaçamento seria solucionado em um processo chamado de *detokenização* (e.g. "Desde ontem ," `->` "Desde ontem,").

Por fim, o texto do nosso exemplo seria realizado como:

```
O Brasil totaliza 218.223 casos de #COVID-19, 15.305 casos a mais do
que no dia anterior. Desde ontem, foram registradas no país 824 novas
mortes, que agora somam 14.817, representando uma alta diária de 6%.
```

13.3.3 Limitações

A arquitetura *pipeline* possui algumas limitações bastante relevantes para o desenvolvimento de aplicações de GLN de propósito mais geral. Para o desenvolvimento de uma abordagem em regras de um módulo como o de Estruturação de Texto, por exemplo, seria necessário construir manualmente um grande número de regras para agrupar em sentenças a combinação de mensagens ordenadas vindas do módulo de Ordenação do Discurso. No caso de uma abordagem baseada em aprendizagem de máquina, seria necessária uma grande quantidade de dados paralelos anotados entre mensagens ordenadas e seus agrupamentos em sentenças. O mesmo se repete para qualquer módulo da arquitetura *pipeline* após a Seleção de Conteúdo.

Em conclusão, seja em uma abordagem baseada em regras ou de aprendizagem de máquina, a arquitetura *pipeline* demanda um grande esforço para implementação das regras e anotação das representações intermediárias de cada módulo, respectivamente. Tal esforço pode tornar inviável a construção de sistemas de GLN que demandam a verbalização de uma grande quantidade de mensagens e sua combinação.

13.4 Arquitetura Ponta-a-Ponta

Diferentemente da arquitetura *pipeline* discutida na seção anterior, a arquitetura ponta-a-ponta (em inglês, *end-to-end*) visa gerar um texto a partir de uma representação (ou outro texto) de entrada sem representações intermediárias explícitas. Esta arquitetura

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

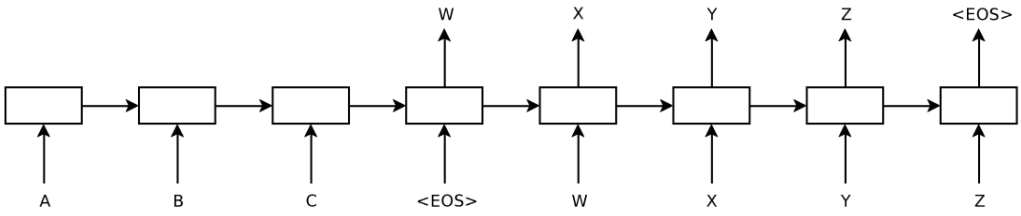
©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

popularizou-se a partir do avanço dos modelos de aprendizado profundo, principal ferramenta utilizada na geração de linguagem natural ponta-a-ponta. Estas arquiteturas são discutidas a seguir, fazendo uso de alguns conceitos que também são discutidos na Subseção Modelos de Linguagem Neurais Modernos, como o uso de *encoders*, mecanismos de atenção e as arquiteturas do tipo Transformer.

Treinamento e Inferência Modelos do tipo ponta-a-ponta são treinados em uma grande quantidade de dados paralelos que fazem um mapeamento das representações de entrada para seus respectivos textos de saída. Uma vez adequadamente treinados, estes modelos são aplicados à geração de textos.

Encoder-Decoder Sutskever et al. (2014) foi um dos primeiros estudos a aplicar uma arquitetura ponta-a-ponta na tarefa de geração de texto, no caso usando uma máquina de tradução. Nesta abordagem, dado um texto de entrada em um determinado idioma, um modelo neural é usado para codificar esta sequência textual, *token* por *token*, em representações vetoriais do tipo *embeddings* (Mikolov et al., 2013b). A representação vetorial de entrada obtida após a codificação do último *token* de entrada é então decodificada, também *token* por *token* em uma sentença no idioma alvo até a obtenção de um *token* sinalizador de parada. Esta arquitetura, conhecida como *encoder-decoder* é ilustrada na Figura 13.8.

Figura 13.8: Arquitetura Ponta-a-Ponta Encoder-Decoder.



Fonte: (Sutskever et al., 2014)

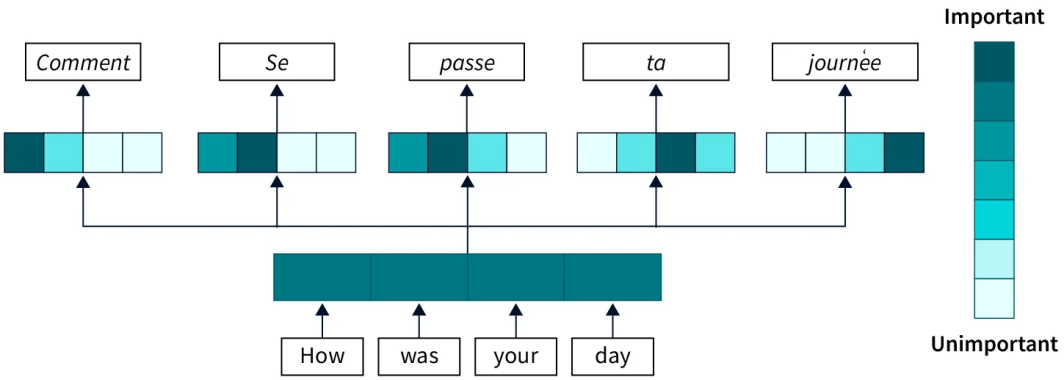
Mecanismo de Atenção Em (Bahdanau et al., 2015), a arquitetura *encoder-decoder* foi acrescida de um mecanismo de atenção do tipo comumente aplicado a máquinas de tradução, sendo distinta da arquitetura *encoder-decoder* simples no que diz respeito ao processo de decodificação. De forma mais específica, para cada *token* decodificado, as representações vetoriais codificadas são combinadas em uma única representação a partir do peso de importância de cada um dos *tokens* de entrada. Assim, o processo de decodificação aprende em qual *token* de entrada prestar atenção quando gerar cada *token* no idioma alvo. Este mecanismo de atenção é ilustrado na Figura 13.9.

A Figura 13.10 mostra o mapa de calor dos pesos de atenção ao traduzir uma sentença do inglês para o francês. Observa-se neste exemplo como um maior peso de atenção ao *token* de entrada *agreement* é dado ao gerar o *token accord* em francês. Nas avaliações conduzidas, Bahdanau et al. (2015) mostram como o mecanismo de atenção aumenta significativamente a adequação e fluência das traduções.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Figura 13.9: Arquitetura Ponta-a-Ponta Encoder-Decoder com Mecanismo de Atenção.



Fonte: (Malingan, 2024)

Transformer Vaswani et al. (2017) introduziram a arquitetura neural conhecida por Transformer, que é considerada o estado-da-arte até o momento da escrita deste capítulo. Esta arquitetura é ilustrada na Figura 13.11.

Como pode ser visto na Figura 13.11, a arquitetura também se divide entre um módulo codificador e outro decodificador. Uma das principais diferenças da arquitetura *encoder-decoder* com mecanismo de atenção (Bahdanau et al., 2015) é o fato de que Transformers também aplicam os pesos de atenção no processo de codificação, ou seja, ao gerar a representação vetorial de um *token* de entrada, leva-se em conta os pesos de atenção de seus *tokens* antecessores e sucessores de entrada. Além disso, esta arquitetura é otimizada para o processo de treinamento, que tende a processar mais *tokens* em um menor espaço de tempo e ser mais eficiente do que sua antecessora[14].

13.4.1 Aplicações

Arquiteturas ponta-a-ponta são aplicadas a uma ampla gama de sistemas de GLN recentes. O estudo em (Wen et al., 2015) propõe um modelo generativo ponta-a-ponta *encoder-decoder* com camadas de *Long Short-Term Memory* (LSTM) semanticamente condicionadas para realizar textualmente informações sobre locais de hotéis e restaurantes; em (Wen et al., 2016), utiliza-se um modelo semelhante para verbalizar representações semânticas sobre avaliações de *laptops* e televisores; em (Lebret et al., 2016) e (Chisholm et al., 2017), são propostos modelos neurais para gerar frases biográficas a partir de tabelas de fatos de biografias da Wikipédia; em (Dong et al., 2017), é proposto um modelo neural para realizar textualmente avaliações de produtos; finalmente, em (Mei et al., 2016) é proposto um modelo de GLN ponta-a-ponta para verbalização de dados do tempo e competições robóticas.

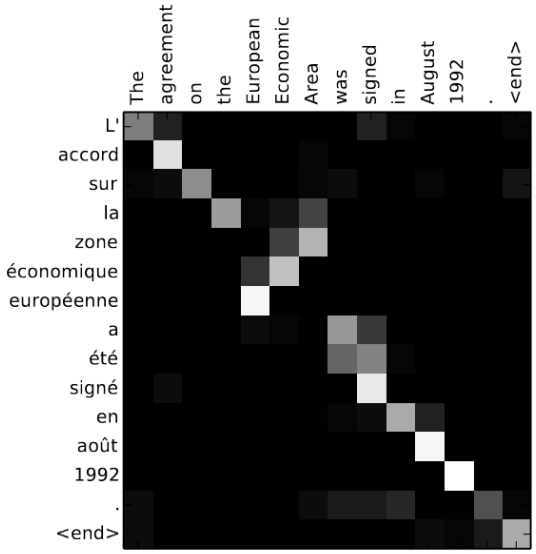
No campo dos desafios computacionais do tipo *shared-task*, muitos dos sistemas competidores seguem a arquitetura ponta-a-ponta e tornaram-se referências na área. Dentre os mais famosos destaca-se o desafio WebNLG (Ferreira et al., 2020; Gardent et al., 2017),

[14]Apesar do processo de treinamento da arquitetura Transformer ser mais rápido do que a arquitetura proposta em (Bahdanau et al., 2015), o processo de inferência tende a ser mais lento.

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

Figura 13.10: Mapa de Calor demonstrando os Pesos de Atenção na tradução inglês-francês.



Fonte: (Bahdanau et al., 2015)

cujo o objetivo é a verbalização de representações semânticas da base de web semântica DBPedia (Bizer et al., 2009); o E2E Challenge (Dušek et al., 2018), que visa a verbalização de informações sobre restaurantes; e o desafio GEM (Gehrmann et al., 2021) cujo o objetivo é criar um *benchmark* dinâmico para GLN, sua avaliação e métricas, enfrentando as dificuldades na medição do progresso em GLN, que dependem de métricas automatizadas, conjuntos de dados e padrões de avaliação humana em constante evolução. Os artigos que descrevem cada um destes desafios apresentam informações sobre os sistemas participantes.

13.4.2 Exemplo – sistema CoronaReporter

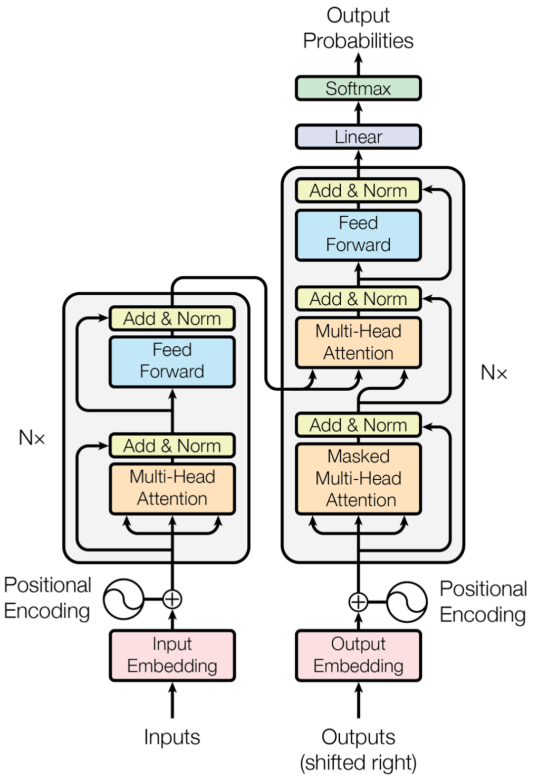
Como exemplificado na Seção 13.3.2, o sistema CoronaReporter (Campos et al., 2020) utiliza uma arquitetura *pipeline* com múltiplas representações intermediárias entre os números de COVID-19 no Brasil e o relatório textual final sobre o assunto. Se este mesmo robô-jornalista fosse construído de acordo com a arquitetura ponta-a-ponta, observa-se que, após a seleção de conteúdo, uma rede neural *encoder-decoder* ou Transformer poderia ser treinada em uma grande quantidade de dados paralelos para realizar textualmente o conteúdo selecionado sem representações intermediárias explícitas. Neste caso, o conteúdo selecionado de exemplo na Seção 13.3.2 seria diretamente convertido no texto final correspondente como exemplificado a seguir.

```
Conteúdo Selecionado:
DEATHS_VARIATION_LAST_DAY(trend="high", variation=0.06)
NEW_CASES(cases=15305, highest=True)
NEW_DEATHS(deaths=824, highest=False)
TOTAL_CASES(cases=218223)
TOTAL_DEATHS(deaths=14817)
```

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

Figura 13.11: Arquitetura Transformer.



Fonte: (Vaswani et al., 2017)

```
Texto Realizado:
O Brasil totaliza 218.223 casos de #COVID-19, 15.305 casos a mais do
que no dia anterior. Desde ontem, foram registradas no país 824 novas
mortes, que agora somam 14.817, representando uma alta diária de 6%.
```

Outros exemplos de arquiteturas neurais para o caso específico da geração de texto em português incluem as tarefas de sumarização neural (Costa; Paraboni, 2019), e a transferência de estilo para geração de posicionamentos (Costa; Paraboni, 2024) e discursos políticos (Costa; Paraboni, 2023).

13.4.3 Limitações

Conforme discutido nesta seção, arquiteturas de GLN ponta-a-ponta são baseadas em modelos neurais como *encoder-decoder* e Transformer. Tais abordagens demandam uma ampla quantidade de dados para treinamento, o que implica na necessidade de um grande volume de dados paralelos entre representações semânticas e seus textos correspondentes. Isso representa um alto custo de treinamento, sendo uma das limitações dessas arquiteturas.

Além disso, apesar de produzirem textos mais fluentes que as arquiteturas mais tradicionais, a primeira geração de sistemas ponta-a-ponta tem sérios problemas de adequação e, em especial, o problema conhecido como alucinação (Ji et al., 2023) em que o modelo

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

pode gerar informações que parecem plausíveis, mas que não correspondem à realidade ou ao conhecimento factual correto. Algumas destas limitações têm sido mitigadas com o uso de técnicas como pré-treinamento e refinamento com base em *prompts*, como veremos na próxima seção.

13.5 Arquiteturas Pré-treinadas

Inicialmente, arquiteturas neurais ponta-a-ponta eram treinadas uma única vez para um domínio específico de GLN, como a verbalização de casos de COVID-19 no Brasil exemplificado na Seção 13.4.2. No entanto, como neste caso os pesos das redes neurais são inicializados de forma aleatória durante o treinamento, abordagens deste tipo exigem uma grande quantidade de dados anotados no domínio específico.

Como forma de otimizar este processo, atualmente as aplicações ponta-a-ponta de GLN têm seu treinamento iniciado a partir de modelos neurais pré-treinados, que são construídos com base em tarefas genéricas com uso de grandes volume de dados. Segundo (Raffel et al., 2020), esse pré-treinamento faz com que o modelo desenvolva habilidades e conhecimentos de uso geral que podem então ser transferidos para tarefas posteriores, como a GLN. A seguir discutimos brevemente alguns exemplos populares de modelos neurais pré-treinados refinados para aplicações deste tipo.

BERT BERT (*Bidirectional Encoder Representations from Transformers*) utiliza a arquitetura Transformer para capturar contextos bidirecionais em textos (Devlin et al., 2019). Neste modelo, discutido em mais detalhes na Subseção Codificador: BERT e seus amigos, o processo de pré-treinamento envolve duas tarefas principais: a *Masked Language Modeling* (MLM), onde palavras aleatórias em uma frase são mascaradas e o modelo é treinado para prever essas palavras, e *Next Sentence Prediction* (NSP), onde o modelo aprende a entender a relação entre pares de frases. Essa fase resulta em representações ricas e contextualmente robustas. Apesar de ter sido refinada para tarefas de geração, este modelo é comumente aplicado em tarefas de classificação de texto.

T5 T5 é um modelo de linguagem *encoder-decoder* conhecido como *Text-To-Text Transfer Transformer* e desenvolvido pelo Google Research (Raffel et al., 2020), e discutido em mais detalhes na Subseção Instâncias de um Transformer inteiro: T5 e seus aliados. Este modelo adota uma abordagem unificada para lidar com diversas tarefas de PLN ao transformá-las em problemas de entrada-saída de texto para texto. Durante o pré-treinamento, o modelo T5 é alimentado com uma vasta quantidade de dados textuais em diversas tarefas estruturadas como `*Instrução $+$ texto de entrada` → `texto de saída` do *corpus* C4, publicado juntamente com o modelo de linguagem. Por exemplo, para uma tarefa de tradução, a entrada pode ser "*translate English to French: How are you?*" e a saída esperada seria "*Comment ça va?*". Essa formulação unificada permite que o mesmo modelo trate diversas tarefas como tradução, análise de sentimento, sumarização de texto e respostas a perguntas de maneira consistente. Este estilo de treinamento é muito popular na atualidade, e ficou conhecida como aplicações baseadas em *prompts* (i.e. instruções) ou Inteligência Artificial Generativa.

GPT GPT (*Generative Pre-trained Transformer*) é uma família de modelos neurais de linguagem baseado na parte decodificadora da arquitetura Transformer (Achiam et

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

al., 2023; Brown et al., 2020). A abordagem de geração é representada na forma de um processo auto-regressivo, em que o modelo prevê a próxima palavra em uma sequência com base nas palavras anteriores, e gera texto de forma sequencial. No processo de pré-treinamento, o modelo é treinado em grandes *corpora* de textos para prever a próxima palavra em uma sequência dada, utilizando apenas o contexto das palavras anteriores. Esse treinamento em grande escala permite ao GPT aprender representações linguísticas e semânticas muito ricas, assim como ser capaz de resolver diversas tarefas formatadas de maneira textual. Por esta razão, este modelo é comumente utilizado como uma IA generativa.

BART BART é um modelo neural encoder-decoder que combina as vantagens dos Transformers bidirecionais, como o BERT, com propriedades auto-regressivas, como o GPT (Lewis et al., 2020). Semelhante aos modelos anteriores, no processo de pré-treinamento as entradas textuais são corrompidas por meio de diversas técnicas, como *Masked Language Modeling* (MLM) do BERT e permutação de palavras, e o modelo é treinado para reconstruir o texto original. Esse processo permite que o modelo BART aprenda a compreender e corrigir distorções em textos, resultando em representações robustas e flexíveis.

13.5.1 Aplicações

Modelos pré-treinados são amplamente utilizados em sistemas de GLN da atualidade. O modelo T5 foi treinado com instruções relacionadas a tarefas de GLN como máquinas de tradução, sumarização de texto e perguntas e respostas (Raffel et al., 2020); em (Guo et al., 2020) é abordado o refinamento da arquitetura T5 no *corpus* WebNLG para realização textual de representações da web semântica; em (Rothe et al., 2020) é investigado o refinamento de modelos pré-treinados como GPT, BERT e suas variantes nas tarefas de máquinas de tradução, sumarização e *split-and-rephrase*, a qual busca reescrever uma frase longa em duas ou mais frases curtas e coerentes; em (Almasi; Schiønning, 2023), é proposta uma versão refinada do modelo GPT-3 para a geração de notícias em dinamarquês; em (Koppatz et al., 2022), é proposta uma aplicação semelhante para geração de títulos de notícias em finlandês; (Xie et al., 2023) faz uma investigação profunda do uso de IAs generativas na geração de histórias (e.g. *storytelling*); em (Wang et al., 2023b) é abordada uma aplicação de IA generativa no contexto de sistemas de diálogo; finalmente, em (Hicke et al., 2023) foi avaliado o uso de modelos de linguagem como GPT-2, DialogGPT e GPT-4 na geração de respostas acuradas no domínio da educação.

13.5.2 Exemplo

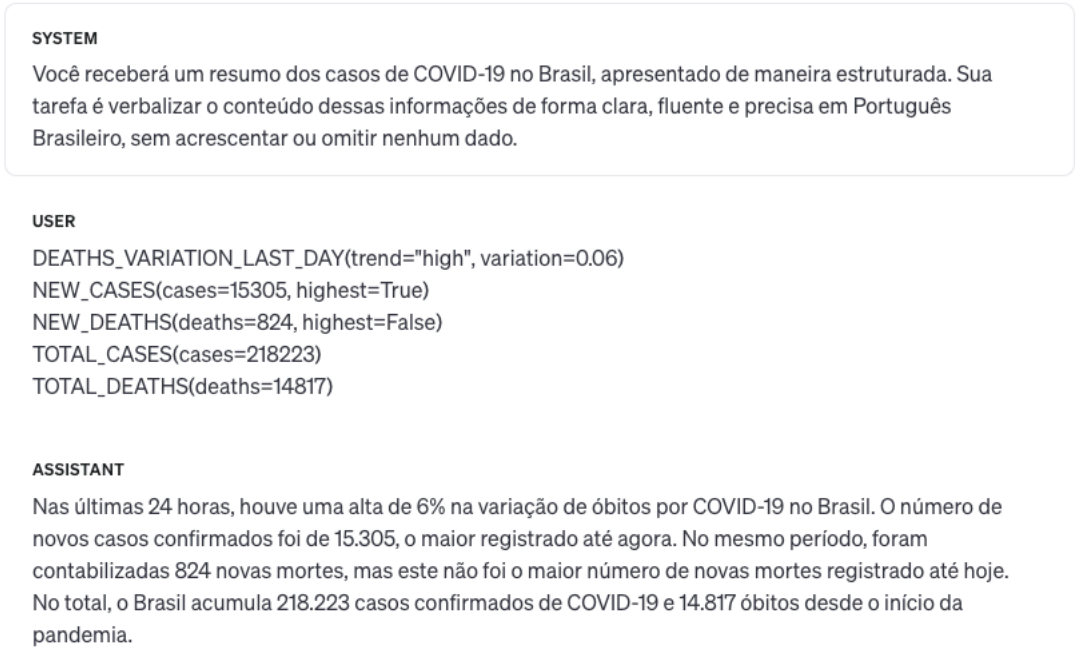
A Figura 13.12 ilustra a implementação do módulo de Realização Textual do robô-jornalista CoronaReporter utilizando a IA Generativa GPT-4o que, no momento da escrita deste capítulo, é uma das mais recentes versões da arquitetura GPT.

Como pode ser visto na Figura 13.12, neste caso o conteúdo selecionado é dado como entrada para o modelo, antecedido por uma mensagem de sistema instruindo-o como processá-la. Esta instrução explica o domínio e contexto da entrada fornecida (e.g. resumo dos casos de COVID-19 no Brasil) e como a mesma deve ser processada (e.g. converter o conteúdo em um texto em português claro, fluente e preciso). A representação estruturada de entrada consiste nas mensagens selecionadas a serem verbalizadas, cada uma formatada

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Figura 13.12: Aplicação da IA Generativa GPT-4o para verbalização de casos de COVID-19 no Brasil. Instrução executada no dia 25 de Agosto de 2024.



com um predicado identificador do tipo da mensagem (pe. `NEW_CASES`, `NEW_DEATHS`) juntamente com seus atributos e respectivos valores (e.g. `cases=15035` etc.). A saída gerada pelo modelo é apresentada no parágrafo *Assistant* na mesma figura. Seguindo a instrução de sistema, o modelo de linguagem é capaz de verbalizar o texto resultante em português brasileiro de maneira fluente e adequada, mesmo com a representação semântica de entrada tender à língua inglesa.

13.5.3 Limitações

Uma das principais limitações das IAs Generativas é o fato de não terem informações atualizadas. Como exemplificado na Figura 13.13, ao perguntar sobre o número de medalhas que o Brasil ganhou nas Olimpíadas de Paris em 2024, o modelo GPT-4o informaria que não possui esta informação, já que seu último treinamento data de Outubro de 2023.

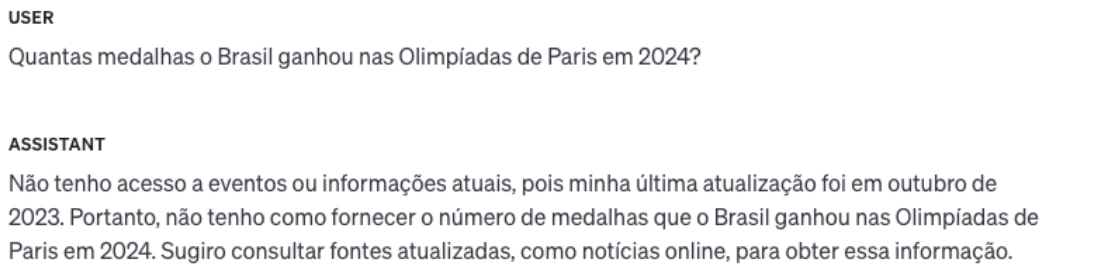
Técnicas utilizadas para sanar este tipo de problema incluem refinar (*fine-tuning*) o modelo com informações atualizadas, ou aplicar a técnica de engenharia de *prompt* conhecida como *Retrieval Augmented Generation* (Gao et al., 2024), que consiste de um sistema de busca seguido da IA generativa em que, dada uma instrução de entrada, o sistema de busca recupera os documentos mais semelhantes à instrução, e os fornece como contexto para a IA generativa prover uma resposta adequada.

Modelos de linguagem neurais também possuem limitações na resolução de alguns tipos de tarefas, como as que demandam operações aritméticas. Pesquisadores têm buscado resolver este problema a partir do uso de IAs Generativas como Agentes (Karpas et al., 2022; Yao et al., 2023). Nesta abordagem, os agentes têm disponíveis uma série de ferramentas, como web serviços, calculadoras, outras soluções de Inteligência Artificial e, dada uma instrução, cabe ao agente interpretar o problema, segmentá-lo em problemas menores, e

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Figura 13.13: Instrução dada ao modelo GPT-4o no dia 24 de Agosto de 2024. O exemplo demonstra como o modelo não possui informações atualizadas após seu treinamento.



delegar suas soluções às ferramentas mais adequadas para resolvê-los.

13.6 Conclusão

Um dos principais objetivos deste capítulo foi o de mostrar a linha do tempo dos modelos de geração de linguagem natural e sua nítida evolução ao longo dos anos. Se nos primórdios estes sistemas visavam somente avaliar a adequação de gramáticas de língua inglesa sem necessariamente considerar um significado real (Yngve, 1961), hoje vemos modelos de linguagem neurais poderosos, capazes de gerar textos variados e de alta adequação para verbalização de informações em diferentes domínios e linguagens. Usando a analogia introduzida no começo do capítulo, em termos de qualidade, os novos textos gerados automaticamente se distanciam dos ruídos da biblioteca de Borges, estando muito mais próximos daqueles indexados pelos bibliotecários da instituição fictícia.

Do modelo de Yngve (1961) aos modelos de linguagem como ChatGPT, a evolução da área de GLN passa por arquiteturas tradicionais como a baseada em *templates* e a arquitetura *pipeline*, capazes de gerar textos adequados e fiéis à informação de entrada, contudo repetitivos e engessados em termos de fluência. Além disso, estes sistemas demandam um grande esforço manual para implementação de regras e anotação de suas representações, tornando-os inviáveis para uma aplicação que demande a verbalização de uma grande variedade de tipos de mensagens. Com base nestas limitações, podemos assim concluir que a aplicação de arquiteturas de GLN tradicionais é atualmente recomendada apenas para domínios restritos e bem específicos.

Mais recentemente, arquiteturas de GLN ponta-a-ponta mostraram-se capazes de gerar textos mais fluentes que suas arquiteturas antecessoras, e sem a necessidade do desenvolvimento de regras ou a anotação manual de representações intermediárias. Contudo, o treinamento de aplicações ponta-a-ponta demanda uma grande quantidade de dados de treinamento e um alto poder computacional, tornando seu desenvolvimento mais custoso que no caso das arquiteturas antecessoras. Estas arquiteturas também sofrem com problemas de adequação, em especial o fenômeno da alucinação, podendo omitir ou adicionar informações que deveriam ou não estar no texto gerado.

Por fim, arquiteturas ponta-a-ponta pré-treinadas, também conhecidas como IAs generativas, representam o atual estado da arte em muitas tarefas desafiadoras para seres humanos, incluindo diversas aplicações de GLN. Embora possam ser refinados para um

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

domínio específico (processo conhecido como *fine-tuning*), esses modelos já são capazes de gerar texto em sua versão pré-treinada, mitigando a necessidade de grandes volumes de dados de treinamento típicos das arquiteturas ponta-a-ponta. No entanto, um ponto negativo deste tipo de abordagem é o crescimento exponencial do tamanho desses modelos. Por exemplo, enquanto o modelo GPT-2 possui 1,5 bilhões de parâmetros, estima-se que o GPT-4 tenha quase 2 trilhões. A demanda por dados e capacidade computacional para treinar e executar esses modelos está diretamente relacionada ao seu tamanho e, consequentemente, o custo de pré-treinamento e uso de um modelo como o GPT-4 é tão elevado que seu acesso fica restrito a grandes empresas. Cidadãos comuns, pequenas empresas e laboratórios de pesquisa geralmente só conseguem utilizá-los por meio dessas grandes corporações.

Em conclusão, o futuro da área de GLN passa por melhor entender as falhas de fluência e adequação das IAs Generativas, atuais modelos estado-da-arte. Em específico, necessita-se entender melhor o que estes modelos de linguagem neurais **não** podem ou têm dificuldades em fazer. Tal entendimento passa pela pesquisa e desenvolvimento de melhores *corpora* e métricas de avaliação. Uma vez que o conhecimento das limitações destes modelos ficar mais clara, o andamento do campo de pesquisa em GLN indica para a aplicação de Agentes (Karpas et al., 2022; Yao et al., 2023), que são modelos de linguagem instruídos em gerenciar um grande número de ferramentas para resolução de problemas. Além disso, espera-se uma democratização do acesso destes modelos, seja através da disponibilização de modelos públicos, pesquisa de técnicas para barateamento de custos de execução, e avaliação de seu uso numa maior variedade de línguas.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Capítulo 14

Perguntas e Respostas

Eduardo G. Cortes
Renata Vieira
Dante A. C. Barone

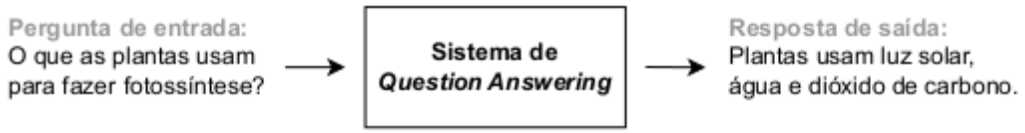
Publicado em: 13/03/2024

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

14.1 Introdução

A área de Resposta Automática a Perguntas, em inglês *Question Answering* (QA), estuda como criar sistemas capazes de responder de forma automática a perguntas em linguagem natural. Esses sistemas buscam a capacidade de compreender a pergunta, recuperar informações relevantes e fornecer respostas precisas e úteis. Se compararmos com um sistema convencional de Recuperação de Informação (RI), que tem o objetivo de fornecer documentos relevantes a partir de uma consulta de entrada (Capítulo Recuperação de Informação), os sistemas de PR diferem-se principalmente pela sua precisão em fornecer apenas aquilo que lhe foi solicitado. Enquanto os sistemas de RI retornam listas ranqueadas de documentos, cabendo ao usuário/a explorar estes documentos em busca de informações mais específicas, os sistemas de PR irão fornecer apenas a resposta ao usuário/a, conforme o exemplo da Figura 14.1.

Figura 14.1: Exemplo de entrada e saída de um sistema de PR.



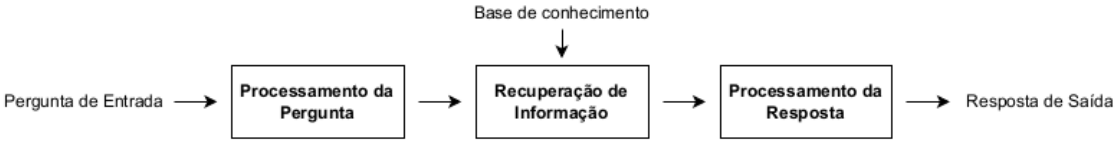
Neste capítulo, abordamos a área de PR no contexto de sistemas especialistas focados em entregar informações diretas sobre o que foi perguntado. Por outro lado, existem os sistemas *chatbots*, conhecidos também como agentes de conversação, que visam manter um diálogo contínuo e engajante com o usuário/a. A principal diferença desses sistemas é a sua capacidade de adaptar suas respostas considerando o contexto mais amplo da conversa. Estes sistemas de conversação não serão o foco deste capítulo (veja Capítulo ChatGPT, MariTalk e outros agentes de conversação); ao invés disso, abordaremos os sistemas de PR projetados para responder a perguntas de forma isolada, ignorando o contexto mais amplo da conversa, mas focados na eficiência e precisão em fornecer respostas informativas.

A área de PR se relaciona diretamente com duas grandes subáreas do Processamento de Linguagem Natural (PLN), que são a compreensão da linguagem natural e a geração de linguagem natural, uma vez que sistemas de PR são desenvolvidos para processar a

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

pergunta de entrada e, muitas vezes, a geração de linguagem natural para a resposta de saída. Além disso, PR também se relaciona com a área de RI, uma vez que estes sistemas utilizam algum tipo de base de conhecimento; é necessário recuperar as informações relevantes para o desenvolvimento da resposta. A Figura 14.2 apresenta a arquitetura geral de um sistema de PR, contendo três passos convencionais, que consistem em: 1) processar a pergunta de entrada buscando sua compreensão, 2) buscar por informações relevantes em alguma base de conhecimento; e 3) definir a resposta de saída.

Figura 14.2: Arquitetura geral de um sistema de QA.



Os sistemas de PR utilizam diferentes etapas de processamento que envolvem diversas tarefas. Por exemplo, a compreensão da pergunta pode envolver diversas etapas, desde classificação de texto, extração de tópicos e Reconhecimento de Entidades Nomeadas (REN). Cada uma destas etapas podem ser aprofundadas e contém seus próprios desafios e problemas de pesquisa. Por outro lado, existem também soluções com menos etapas explícitas, como os modelos *end-to-end*, nos quais o modelo de PR pode ser entendido como uma caixa preta, onde a entrada é uma pergunta (podendo conter informações de contexto) e a saída é uma resposta. Estes modelos são redes neurais profundas, treinadas especificamente para a tarefa de PR. Assim, podemos considerar que estas etapas estariam de alguma forma implícita absorvida nas suas diversas camadas de neurônios artificias (apesar de não se poder assumir de fato uma hierarquia de representações desse tipo). Atualmente, estes modelos estão atingindo quantidades consideráveis de parâmetros neurais, que estão saindo de milhões para bilhões e até trilhões de parâmetros (o que pode exigir capacidades disponíveis para muitos poucos). O aumento na quantidade de parâmetros destes modelos está relacionado ao aumento da sua capacidade de entendimento da linguagem natural, armazenamento de informação e geração de linguagem natural.

Existem diversas maneiras de classificar um sistema de PR, com base em características que influenciam significativamente seu funcionamento e complexidade. Estas classificações incluem:

- **Tipo de pergunta:** Os sistemas podem ser especializados em responder a perguntas factuais (ex.: "Quem escreveu Dom Casmurro?") ou perguntas que requerem instruções detalhadas (ex.: "Como fazer um bolo de chocolate?").
- **Fonte de conhecimento:** Os sistemas podem utilizar documentos (com etapas de processamento de IR) ou dados estruturados (como grafos de conhecimento), onde as perguntas são transformadas em consultas (*queries*) que buscam informações relevantes.
- **Domínio de conhecimento:** Existem sistemas de amplo domínio, que cobrem uma variedade de assuntos, e sistemas de domínio restrito, focados em áreas específicas como saúde (Oliveira et al., 2021), mudanças climáticas (Cação et al., 2021), ou direito (Quaresma; Rodrigues, 2005).

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Cada um desses eixos representa uma dimensão independente, permitindo que os sistemas de PR sejam classificados e desenvolvidos combinando essas características de qualquer maneira, para atender a necessidades específicas de informação.

Além das tarefas convencionais de PR, existem diferentes tarefas com peculiaridades e desafios únicos, como ranquear uma lista de respostas fornecida para uma pergunta, correlacionar informações de múltiplas fontes de dados para fornecer uma resposta correta que requer múltiplos passos de raciocínio, e responder a perguntas com base em informações visuais (sistemas multi modais), o que requer que o sistema entenda a pergunta textual, correlacionando-a com a informação visual para gerar uma resposta apropriada.

Este capítulo explora os conceitos previamente introduzidos, categorizando os diferentes tipos de sistemas e apresentando as abordagens e técnicas comumente empregadas em cada um. São discutidos os diferentes tipos de tarefas de PR, destacando também os métodos de avaliação e outras categorias de tarefas relacionadas à PR. O capítulo é concluído com considerações finais que sintetizam as principais ideias e direcionamentos futuros.

14.2 Classificação de Sistemas de Perguntas e Respostas

Existem diferentes tipos de sistemas de PR que se diferenciam por diferentes critérios. Estas diferenças estão principalmente relacionadas à riqueza de possibilidades da linguagem natural, à grande amplitude de áreas de conhecimento e profundidade de detalhes que cada uma apresenta, como também às diferentes formas de recursos de informação. Portanto, o desenvolvimento de um sistema que domine toda a complexidade da linguagem natural, capaz de lidar com todos os tipos de estruturas de dados, e com conhecimento aprofundado em todas as áreas, é desafiador. Assim, é comum que existam diferentes tipos de sistemas de PR adequados para aspectos particulares. Esta seção busca organizar as diferentes categorias de sistemas de PR através de três formas de categorização baseadas no tipo da pergunta de entrada do sistema, no tipo de fonte de conhecimento e no domínio de conhecimento dos sistemas.

14.2.1 Tipo de pergunta

As perguntas em linguagem natural são uma forma de discurso que busca obter informações. Cada tipo de pergunta serve a um propósito específico e requer um tipo de resposta diferente. Existem diversos tipos de perguntas, cada uma servindo a um propósito diferente. Podem existir perguntas que requerem uma informação factual e curta, como por exemplo “Qual é a capital do Brasil?”, que requer apenas o nome de uma cidade como resposta. Outras perguntas podem requerer um texto mais longo, como por exemplo “Por que o céu é azul?”, que requer uma explicação. O tipo de pergunta pode variar não apenas no tamanho da resposta, mas também na necessidade de etapas de raciocínio para fornecer a resposta. Por exemplo, a pergunta “Quantos gols marcou o time campeão da *Champions League* de futebol de 2020?”, requer primeiro identificar o time campeão da *Champions League* de 2020 e depois contabilizar quantos gols este time fez durante o campeonato.

Dependendo do tipo de pergunta, a funcionalidade e complexidade do sistema de PR pode mudar consideravelmente. Por exemplo, perguntas que requerem informações factuais e curtas podem apenas extrair a informação solicitada de um documento de texto, enquanto uma pergunta que busca a comparação de duas obras literárias requer que o sistema analise dois livros diferentes para gerar uma resposta. Sendo assim, o tipo de pergunta que o sistema busca responder tem um papel significativo em seu desenvolvimento.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Diferentes estudos apresentam taxonomias que tentam organizar perguntas em diferentes categorias. Uma das principais é o trabalho de (Li; Roth, 2002) que apresenta uma taxonomia em dois níveis de granularidade com várias categorias em cada nível, conforme a Figura 14.3. O primeiro nível apresenta uma categorização mais abstrata, como "ENTIDADE", enquanto o segundo nível apresenta suas subcategorias mais detalhadas, como "Animal". Por exemplo, uma pergunta como "Quem é o atual presidente de Portugal?" pode ser classificada como "HUMANO" e "Indivíduo", já que a resposta esperada é o nome de uma pessoa, enquanto que a pergunta "O que é um prisma?", pode ser classificada como "DESCRIÇÃO" e "Definição", já que a resposta esperada é um texto definindo o que é um prisma.

Figura 14.3: Taxonomia de tipo de perguntas de (Li; Roth, 2002).

ENTIDADE	NUMÉRICO	LOCALIZAÇÃO
Animal	Código	País
Corpo	Contagem	Cidade
Cor	Data	Estado
Criativo	Distância	Montanha
Moeda	Dinheiro	Outro
Doença	Ordem	
Evento	Período	
Comida	Porcentagem	**DESCRIÇÃO**
Instrumento	Velocidade	Definição
Linguagem	Temperatura	Descrição
Planta	Tamanho	Maneira
Produto	Peso	Motivo
Religião		
Esporte		
Substância	**HUMANO**	**ABREVIAÇÃO**
Símbolo	Grupo	Exp
Técnica	Individuo	Abb
Termo	Título	
Veículo		
Palavra		
Outro		

Uma maneira mais geral de dividir as perguntas é em dois grandes grupos: perguntas factuais e não factuais.

- **Factuais**: São perguntas que exigem essencialmente que um único fato ou um pequeno trecho de texto seja retornado como resposta. Por exemplo, "Quando começou a Segunda Guerra Mundial?" ou "Qual é a capital da Espanha?". A maioria das categorias da taxonomia de (Li; Roth, 2002) podem ser consideradas tipos de perguntas factuais, exceto a categoria "DESCRIÇÃO";
- **Não factuais**: São perguntas que normalmente requerem respostas longas, como uma descrição, opinião ou explicação. Por exemplo, "Por que o céu é azul?" ou "Quais os passos para obter um título de mestre?".

O tipo de pergunta tem um papel significativo no funcionamento de um sistema de PR. Perguntas factuais normalmente são mais simples e requerem uma abordagem extrativa, onde a resposta final é diretamente extraída da base de conhecimento e retornada. Já o tipo não factual tipicamente contém perguntas mais complexas e desafiadoras para os sistemas, que podem requerer etapas de processamento especializadas em gerar linguagem

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

natural, em caso de respostas longas. A Tabela 14.1 apresenta diferentes tipos de perguntas não factuais. Percebe-se que algumas dessas categorias têm desafios distintos para o sistema de PR, como por exemplo, perguntas de confirmação podem requerer etapas de checagem de fatos, enquanto que perguntas comparativas podem requerer a análise de múltiplas informações. Além destas categorias, podem existir diversos outros tipos de perguntas nos grupos de factuais e não factuais, como perguntas que requerem listas ("Quais são as maiores cidades do mundo?"), outras hipotéticas ("O que aconteceria se a lua desaparecesse?"), ou perguntas que precisam de exemplos como resposta ("De que maneira a arte influencia a sociedade?").

A definição de categorias de perguntas pode ser ajustada e ampliada conforme as necessidades e capacidades evoluem nos sistemas de PR. É fundamental considerar a flexibilidade contextual das perguntas, onde até questões aparentemente diretas podem variar em sua classificação conforme o contexto. Por exemplo, a pergunta "Quem é o líder?" pode ser classificada de forma diferente dependendo do contexto em que é feita. Se estiver relacionada a uma discussão sobre uma empresa, pode referir-se a uma "Organização"; se o contexto for um evento esportivo individual, pode ser classificado como "Indivíduo". À medida que os sistemas de PR se tornam mais sofisticados, eles também devem ser adaptáveis e flexíveis para acomodar novos tipos de perguntas e mudanças nas expectativas dos usuários, enfatizando a importância de interpretar corretamente o contexto para fornecer respostas precisas.

Tabela 14.1: Alguns tipos de perguntas não factuais com exemplos.

Tipo	**Descrição**	**Exemplo**
Definição	Perguntas que requerem a definição de alguma coisa. Normalmente iniciam com "O que é ..."	"O que é um prisma?"
Explicativas	Perguntas que requerem uma explicação ou contextualização. Normalmente iniciam com "Por que ..."	"Por que o céu é azul?"
Procedimentais	Perguntas que requerem um conjunto de passos ou instruções para realizar alguma coisa. Normalmente iniciam com "Como ..."	"Como fazer um bolo de chocolate?"
Comparativa	Perguntas que requerem uma comparação entre dois ou mais assuntos.	"Quais são as diferenças entre SSD e HDD?"
Opinião	Perguntas que requerem uma perspectiva pessoal ou avaliação.	"O que você acha de arte moderna?"
Confirmação	Perguntas que requerem "Sim" ou "Não" como resposta.	"Atenas é a capital da Grécia?"

14.2.2 Tipo de Fonte de Conhecimento

Um dos principais componentes de um sistema de PR é a fonte de conhecimento, que é utilizada para extrair as informações necessárias para responder à pergunta de entrada.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Existem diferentes tipos de componentes que representam a fonte de conhecimento, os quais se diferenciam na maneira como o conhecimento é criado, armazenado, modificado e consultado.

14.2.2.1 Dados Não Estruturados

Os dados não estruturados não seguem uma estrutura padrão e têm como vantagem a facilidade na criação do conjunto de dados que será a fonte de conhecimento do sistema. Deste modo, é possível utilizar qualquer conteúdo textual, sem a necessidade de estruturar o conteúdo em algum modelo pré-definido (Capítulo 16). Porém, a desvantagem desta abordagem é o maior desafio de processar e analisar estes dados pelos sistemas, que terão que extrair as informações relevantes da coleção que são necessárias para a resposta. Isto requerer uma capacidade maior de compreensão da linguagem natural. Nesta categoria se encaixam principalmente os documentos brutos de texto, que normalmente são uma coleção de documentos textuais sem uma estrutura que represente a informação em detalhes.

Para entendermos um sistema de PR com documentos textuais, podemos imaginar um sistema que utilize uma coleção de artigos jornalísticos como fonte de conhecimento e que responda perguntas sobre estes artigos. Este sistema pode ser classificado como um sistema de PR com dados não estruturados, uma vez que os dados estão em arquivos de texto sem nenhuma estrutura aprofundada para representar as informações. Além disso, o sistema utiliza alguma solução de RI para encontrar artigos relevantes em sua coleção para a pergunta de entrada. Depois de encontrar os documentos relevantes, é necessário extrair a informação específica solicitada destes documentos através de algum mecanismo que consiga associar a informação com a pergunta de entrada, como técnicas envolvendo REN ou modelos de compreensão de leitura.

Da mesma forma, um sistema de PR pode utilizar a *web* como fonte de informação, extraindo páginas HTML com algum mecanismo de busca, como o Google. Embora as páginas **web** contenham o seu conteúdo estruturado com tags HTML, estas tags não seguem um mesmo padrão entre páginas de diferentes sites. Além disso, a principal dificuldade é que a estruturação do texto com **tags** não é profunda o suficiente para representar cada informação da página de forma individual. Normalmente, o conteúdo textual é dividido em parágrafos com uma **tag** "<p>", como é mostrado na Figura 14.4. Porém, outras *tags* não são utilizadas para informações mais detalhadas, como por exemplo, que os componentes absorvidos pelas plantas são o carbono e a água. Esta informação está entre outras diversas informações de um único parágrafo textual. Assim, podemos classificar o conteúdo de páginas *web* como dados não estruturados.

14.2.2.2 Dados estruturados

Diferentemente dos dados não estruturados, fontes de informações com dados estruturados apresentam uma estrutura padrão que representa e organiza a informação, facilitando o acesso e a interpretação dos dados pelo sistema. A informação estruturada segue um esquema rígido, como tabelas em um banco de dados relacional, onde cada tipo de dado é armazenado em uma coluna específica e cada registro ou linha representa uma entidade completa. Assim, todos os dados seguem um mesmo formato, reduzindo a necessidade do sistema em interpretar textos brutos para extrair as informações.

Como exemplo, podemos imaginar um sistema de PR que fornece informações sobre médicos de um hospital. As informações destes profissionais estão armazenadas em tabelas

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Figura 14.4: Exemplo de dados não estruturados com uma página HTML.

```
<html lang="pt">
<head>
<title>Fotossíntese</title>
</head>
<body>
<h1>Entendendo a Fotossíntese</h1>
<p>
A fotossíntese é um processo biológico complexo realizado por plantas, algas e algumas bactérias,
pelo qual a luz solar é convertida em energia química. Esse fenômeno incrível é a base para a vida na
Terra, pois fornece a energia necessária para o crescimento das plantas, além de ser a fonte primária
de energia para quase todas as formas de vida. Durante a fotossíntese, as plantas absorvem dióxido
de carbono (CO2) e água (H2O) e, na presença de luz solar, transformam-nas em glicose e oxigênio.
O processo pode ser simplificado pela equação química 6CO2 + 6H2O + luz solar → C6H12O6 + 6O2.
</p>
...
```

de um banco de dados relacional, onde cada uma contém colunas para detalhes específicos como nome, área de especialização e horário de trabalho. Assim, para a pergunta "Qual a área de especialização da Dra. Maria da Silva?", o sistema poderia simplesmente retornar o valor da coluna "especialização" da linha do registro de Dra. Maria da Silva. Para encontrar este valor, o sistema precisaria transformar a pergunta de entrada em uma consulta SQL. SQL, que significa Linguagem de Consulta Estruturada (do inglês, "*Structured Query Language*"), é uma linguagem de programação usada para gerenciar e manipular bancos de dados relacionais. A consulta SQL equivalente à pergunta poderia ser algo como:

```
SELECT especializacao FROM medicos WHERE nome = 'Maria da Silva';
```

Esta consulta instrui o banco de dados a selecionar o valor da coluna "especialização" na tabela "medicos" onde a coluna "nome" é igual a "Maria da Silva". A execução dessa consulta resultaria na resposta desejada pelo sistema.

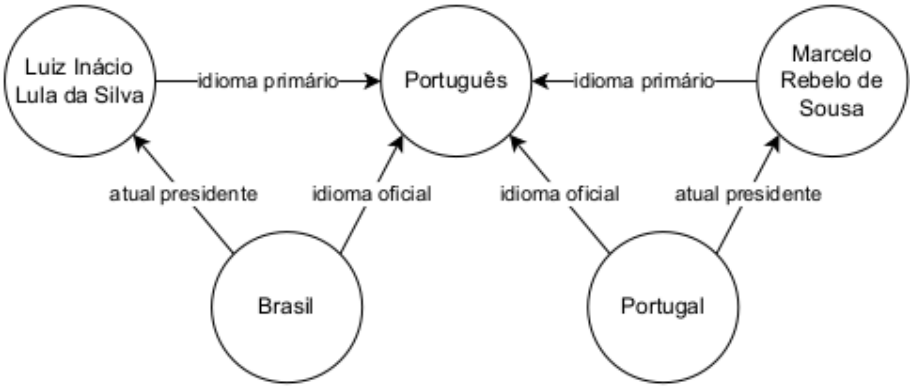
No grupo de dados estruturados, além de tabelas e bancos de dados relacionais, se encaixam principalmente os dados linkados, como ontologias e grafos de conhecimento, que são comuns de serem utilizados como fonte de conhecimento de sistemas de PR. Estas estruturas oferecem uma maneira rica e flexível de representar e organizar informações através de entidades e relacioná-las. Elas ajudam na compreensão semântica, que é crucial para responder perguntas de maneira eficaz e natural. Como mostrado na Figura 14.5, a entidade "Marcelo Rebelo de Sousa" contém diferentes relações, e entre estas entidades existe a relação de "atual presidente" com "Portugal". Este tipo de estrutura facilita a busca por informações relevantes. Por exemplo, dada a pergunta "Quem é o presidente de Portugal?", com o grafo de conhecimento é possível verificar todas as relações da entidade "Portugal" e verificar qual está mais próxima, que neste caso é "atual presidente", e retornar a entidade relacionada.

Podem existir sistemas que utilizem ambas as formas de fonte de conhecimento. Por exemplo, dada a pergunta "Como o atual presidente de Portugal entrou na política?", o sistema de PR pode utilizar o grafo de conhecimento para extrair a entidade "Marcelo Rebelo de Sousa" como atual presidente de "Portugal" e depois extrair as informações

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Figura 14.5: Exemplo de dados estruturados com grafo de conhecimento.



de como ele entrou na política, consultando documentos não estruturados que ajudarão a formular a resposta de saída.

14.2.2.3 Memória paramétrica

A transição dos sistemas de PR baseados em dados estruturados e não estruturados para os que utilizam memória paramétrica marca uma evolução significativa. Enquanto as abordagens tradicionais dependem de fontes de conhecimento externas, claramente definidas e acessíveis, a memória paramétrica introduz um paradigma onde o conhecimento é embutido diretamente na arquitetura e nos parâmetros de modelos de aprendizado profundo.

Com o advento das redes neurais profundas, os modelos de linguagem foram equipados com a capacidade de armazenar informações em seu elevado número de parâmetros ajustáveis (Capítulo Modelos de linguagem). Estes parâmetros podem ser entendidos como pesos de conexões entre neurônios artificiais que são ajustados durante a fase de treinamento. Esses modelos são treinados em grandes conjuntos de dados para aprender uma variedade de padrões e fatos, e são capazes de aplicar esse aprendizado para responder perguntas e realizar outras tarefas de PLN.

No entanto, com redes neurais, ainda não temos garantias de um processo perfeito de armazenamento, e não sabemos se as informações estão realmente sendo representadas de forma fiel aos dados originais. Além disso, apesar de serem capazes de aplicar esse aprendizado para responder perguntas e realizar outras tarefas de PLN, as informações contidas nesses modelos não são facilmente acessadas da mesma forma que um grafo de conhecimento ou documentos textuais, e também não são interpretáveis para os humanos até o momento. A informação está contida em diversos valores numéricos e que são utilizados exclusivamente pelos modelos de redes neurais profundas quando uma determinada entrada de dados é passada ao modelo.

Os modelos de aprendizado profundo, particularmente aqueles baseados em *transformers* (Vaswani et al., 2017), como o BERT (Devlin et al., 2019) e GPT (Achiam et al., 2023; Brown et al., 2020), contêm um mecanismo chamado de atenção, que permite ao modelo dar pesos diferentes para partes de uma entrada de texto de maneira mais dinâmica e contextual. A memória paramétrica é profundamente influenciada pelo processo de treinamento. O pré-treinamento em grandes corpora de texto é uma técnica comum para estabelecer uma base sólida de conhecimento geral, além de aprender a linguagem, estrutura, e relações contextuais. O *fine-tuning* é o próximo passo, onde o modelo pré-treinado é

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

ajustado para tarefas específicas de PR. Nesta etapa, o modelo aprende especificidades do domínio em questão, adaptando sua memória paramétrica para ser mais eficaz na resposta a perguntas relevantes.

Um sistema de PR que utiliza um grande modelo de *transformers* como memória paramétrica é um exemplo clássico de uma abordagem *end-to-end*. Nesse sistema, um único modelo normalmente executa todas as etapas desde o processamento da pergunta textual até a resposta de saída. As etapas de processamento são realizadas de forma implícita nas inúmeras camadas do modelo. Durante a fase de treinamento, os parâmetros neurais do modelo são ajustados para mapear perguntas a respostas corretas, com base em um grande volume de pares de pergunta-resposta. Este processo, após extensivo desenvolvimento, ajustes e comparações entre diversas versões, possibilita que o modelo não apenas memorize respostas específicas, mas também aprenda padrões e relações que o capacitam a generalizar e responder a perguntas novas e não vistas durante o treinamento, embora esse resultado seja o fruto de um esforço significativo de otimização e refinamento.

É importante notar que a geração de respostas por modelos *transformers end-to-end* é resultado do entendimento contextual e da aplicação de padrões aprendidos, e não do acesso direto a respostas predefinidas. Além disso, a natureza "caixa-preta" desses modelos pode tornar desafiador entender exatamente como uma resposta é formada, levantando questões de interpretabilidade e confiabilidade. Portanto, enquanto sistemas de PR baseados em *transformers end-to-end* oferecem eficiência para processar e responder perguntas, eles também vêm com desafios importantes que precisam ser considerados.

Modelos *end-to-end* podem ser efetivamente combinados com outras abordagens para enriquecer suas capacidades. Especificamente, é possível fornecer não apenas a pergunta, mas também um contexto relevante como entrada para esses modelos. Para isso, o modelo precisa ser treinado com pares de pergunta+contexto e resposta, permitindo que ele aprenda a extrair e utilizar informações pertinentes do texto fornecido. Essa abordagem pode ser complementada com etapas de RI para identificar e fornecer o contexto mais relevante para cada pergunta. Ao integrar fontes de informação confiáveis e específicas, essa estratégia aumenta a confiabilidade e precisão do modelo, ajudando a reduzir o problema de delírio ou alucinação comum em modelos de linguagem gerativos, onde podem gerar respostas plausíveis, porém incorretas ou sem fundamento.

14.2.3 Domínio de conhecimento

Sistemas de PR podem ser categorizados com base no domínio de conhecimento ao qual se dedicam. Alguns são desenvolvidos para responder perguntas específicas de campos como biomedicina ou mercado financeiro, enquanto outros são projetados para abranger uma gama mais ampla de assuntos. Essencialmente, podemos dividir os sistemas em dois grupos principais: de amplo domínio e de domínio restrito.

14.2.3.1 Domínio Amplo

Os sistemas de amplo domínio visam responder perguntas sobre uma vasta gama de tópicos. Dado que o escopo do conhecimento é mais amplo, as fontes de informação utilizadas precisam ser extensivas e diversificadas para cobrir o máximo possível de áreas. Comumente, esses sistemas optam por fontes de conhecimento não estruturadas devido à sua ampla disponibilidade e variedade. Documentos de texto brutos e mecanismos de busca de páginas web são frequentemente empregados como fontes de informação, pois oferecem uma quantidade massiva de dados. No entanto, a vastidão de informações em domínios

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

amplos implica em desafios relacionados à precisão e relevância das respostas geradas, demandando etapas de processamento específicas para recuperação e interpretação de dados.

14.2.3.2 Domínio Restrito

Os sistemas de domínio restrito concentram-se em responder perguntas dentro de um tópico específico ou conjunto limitado de tópicos. Devido à característica focada desses sistemas, a amplitude de informações necessárias é relativamente menor, permitindo uma exploração mais profunda e detalhada do assunto em questão. É comum a utilização de fontes de dados estruturados, pois estes podem oferecer informações específicas e detalhadas, essenciais para fornecer respostas precisas. Os sistemas de domínio restrito geralmente se beneficiam de um entendimento mais profundo e especializado do tópico, resultando em respostas mais confiáveis e acuradas. No entanto, eles enfrentam o desafio de manter-se atualizados e abrangentes dentro do seu campo específico, o que pode exigir atualizações regulares e monitoramento contínuo. Tornar-se assim uma tarefa custosa, já que os dados precisam ser constantemente estruturados e atualizados na fonte de informação, o que demanda esforços consideráveis em termos de tempo e recursos, tanto para a coleta quanto para a manutenção desses dados, garantindo que o sistema permaneça relevante e preciso ao longo do tempo.

14.3 Abordagens

Uma vez que a tarefa de PR busca compreender a pergunta de entrada, recuperar informações relevantes em sua base de conhecimento, e muitas vezes, gerar linguagem natural para a resposta de saída, um sistema de PR pode conter diversas etapas de processamento. Além disso, essas etapas não são necessariamente as mesmas entre diferentes sistemas. As etapas podem depender principalmente das diferentes categorias de sistemas de PR apresentados na Seção 14.2.

Considerando as diferentes arquiteturas e abordagens de sistemas de PR, podemos separar as abordagens apresentadas nesta seção em duas categorias: as de sistemas que utilizam etapas explícitas de processamento, que chamamos de **abordagem modular**; e as de sistemas que abstraem todas as etapas da abordagem modular em um modelo de redes neurais profundas, que chamamos de **abordagem *end-to-end***. Assim, nesta seção serão vistas três possibilidades de arquiteturas de sistemas de PR, variando as abordagens e principalmente a fonte de conhecimento do sistema, que impacta as demais etapas de processamento. A primeira é uma abordagem modular que utiliza dados não estruturados como fonte de conhecimento. A segunda é novamente uma abordagem modular mas que utiliza grafos de conhecimento como fonte de conhecimento. Para finalizar, a terceira é uma abordagem *end-to-end* utilizando documentos não estruturados e memória paramétrica. Para cada possibilidade, são descritas a lógica da arquitetura do sistema e suas etapas.

É importante mencionar que as abordagens discutidas são baseadas em estudos da literatura que apresentam arquiteturas similares entre sistemas do mesmo tipo. Portanto, um sistema de PR não precisa necessariamente seguir exatamente estas etapas e pode utilizar diferentes etapas. Em resumo, a escolha da abordagem para o sistema de PR deve ser guiada por uma análise do contexto no qual será aplicado, dos recursos atuais e da viabilidade de desenvolver novos recursos, como conjuntos de dados para treinamento e bases de conhecimento do sistema. Além disso, é importante considerar as vantagens e desvan-

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

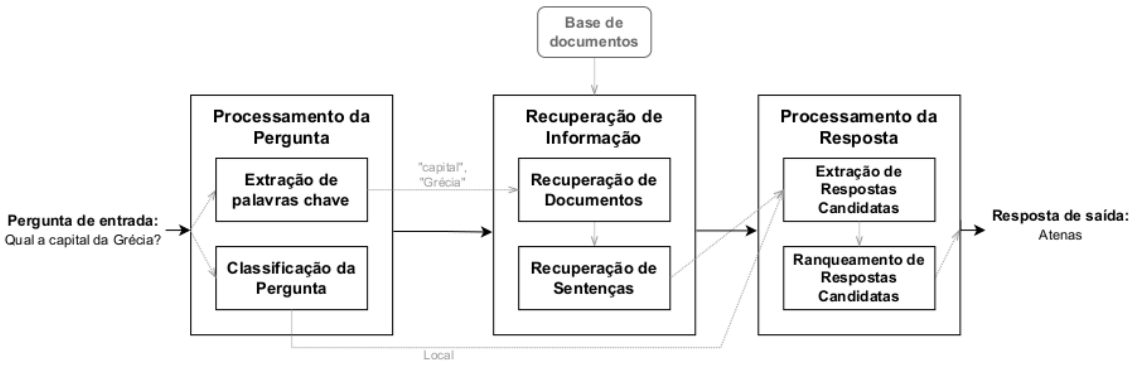
tagens das abordagens modular e *end-to-end*, considerando suas implicações em termos de flexibilidade, complexidade de implementação e capacidade de atender às necessidades específicas do projeto.

14.3.1 Abordagem modular com documentos

Nesta abordagem, vamos considerar um sistema de PR que busca responder a perguntas factuais de domínio amplo. A fonte de conhecimento será composta por documentos de texto bruto. A Figura 14.6 apresenta a arquitetura deste sistema e seus diferentes passos. Basicamente, o sistema é dividido em três grandes etapas, conhecidas como:

1. Processamento da Pergunta, responsável por realizar processamentos na pergunta de entrada, buscando "compreender" o que está sendo solicitado;
2. Com as informações do primeiro passo, é realizada a Recuperação de Informações relevantes da base de documentos do sistema para serem utilizadas na criação da resposta final;
3. Por fim, utilizando as informações dos dois primeiros passos, a última etapa é responsável por determinar a resposta final do sistema. Cada uma destas etapas contém fases específicas de processamento que serão discutidas nas próximas subseções.

Figura 14.6: Exemplo de uma arquitetura de um sistema de PR para perguntas factuais utilizando documentos como base de conhecimento.



14.3.1.1 Processamento da Pergunta

Neste primeiro passo, busca-se determinar o que está sendo solicitado na pergunta de entrada. Nesta abordagem, vamos utilizar duas etapas: uma para extrair palavras-chave que serão utilizadas pela etapa de recuperação de documentos e outra para classificar o tipo da pergunta, que será utilizada para extrair respostas candidatas no passo de Processamento da Resposta.

Extração de palavras-chave: Sistemas de RI normalmente utilizam palavras-chave para buscar documentos relevantes. Assim, são extraídas palavras-chave contidas na pergunta de entrada. Nesta etapa, pode-se utilizar técnicas baseadas em REN, como no estudo do sistema RAPPORT (Rodrigues; Gomes, 2015), ou também como a remoção de *stop-words*, que são palavras menos significativas para a busca, mantendo somente as

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

palavras mais significativas. Por exemplo, na pergunta "Qual a capital da Grécia?", podemos dividí-la em palavras, através de alguma ferramenta de tokenização, e desconsiderar as *stop-words* "Qual", "a" e "da". Para isso, é possível usar alguma lista de *stop-words* pré-definida, como a utilizada no código abaixo que emprega a lista de *stop-words* em português da biblioteca NLTK para Python.

```
import nltk
nltk.download('stopwords')
from nltk.corpus import stopwords
pt_stop_words = set(stopwords.words('portuguese'))
```

Classificação da pergunta: Uma das primeiras etapas de processamento de um sistema de PR é a classificação da pergunta, que também pode ser chamada de classificação do tipo de resposta. Esta é uma tarefa de classificação de texto, onde um modelo recebe o texto da pergunta de entrada e deve determinar a sua classe. Por exemplo, se a pergunta for "Quando Pedro Álvares Cabral descobriu o Brasil?", a classe será "Data", já que a resposta é uma data. Já a pergunta "Onde nasceu Pedro Álvares Cabral?" é classificada como "Local", já que a resposta será um local. Pode-se utilizar diferentes taxonomias de classes dependendo do tipo de perguntas com que o sistema irá trabalhar. Normalmente, esta classe será cruzada com as classes de entidades dos modelos de REN.

Uma abordagem comum para a classificação da pergunta é o treinamento de modelos supervisionados. Estudos da literatura mostram que os modelos propostos para a tarefa de classificação de perguntas para o português superam os 90% de F1-score (Cortes et al., 2020; Cortes et al., 2018). Além disso, abordagens atuais com *transformers* vêm apresentando modelos cada vez mais eficientes para tarefas de classificação de texto (Zhou et al., 2024). Outra abordagem possível é através de regras manuais, que no caso da classificação da pergunta, podem considerar palavras-chave como "Quem", "Quando" e "Onde" para determinar a classe, que poderiam ser respectivamente "Pessoa", "Data" e "Local". Porém, este tipo de abordagem tem problemas com entradas não previstas e pode requerer grandes esforços na criação de regras para cobrir o máximo possível de possibilidades. Além disso, a linguagem natural apresenta desafios de ambiguidade, que podem ser desafiadores, principalmente para abordagens manuais. Assim, é mais comum a utilização de modelos supervisionados quando um conjunto de dados anotados para o treinamento está disponível.

A classe predita pelo modelo de classificação da pergunta pode ajudar nas demais etapas do sistema. Uma das principais utilizações desta classe é ajudar a filtrar documentos, sentenças e respostas candidatas. Por exemplo, se a classe da pergunta é "Pessoa", as etapas dos passos de Recuperação de Informações podem considerar apenas informações que contenham entidades do tipo "Pessoa", já que a resposta deve ser uma pessoa que será extraída dessas informações. Como no exemplo da Figura 14.7, o sistema pode focar apenas em sentenças que contenham este tipo de entidade, reduzindo consideravelmente o escopo de busca. No caso do nosso sistema de exemplo, a classe da pergunta será utilizada na etapa de extração de respostas candidatas, considerando apenas as sentenças que contêm uma entidade da mesma classe.

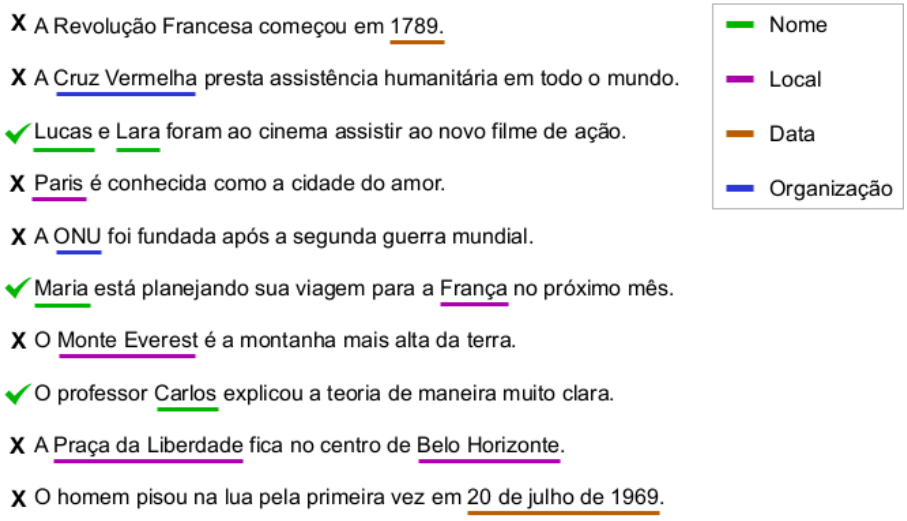
14.3.1.2 Recuperação de Informação

Este passo é responsável por buscar as informações relevantes da base de conhecimento que são determinantes para a resposta final do sistema. Podem haver diferentes etapas

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Figura 14.7: Exemplo ilustrativo de sentenças com suas entidades identificadas, onde somente as sentenças com entidades do tipo "Pessoa" são selecionadas.



que filtram cada vez mais a informação em unidades cada vez menores. Por exemplo, pode haver uma etapa que começa filtrando quais documentos textuais são relevantes para a pergunta, em seguida outra etapa que extrai parágrafos relevantes destes documentos, e por fim, uma etapa que extrai sentenças relevantes destes parágrafos, como é feito no sistema IdSay (Carvalho et al., 2009). No caso do sistema de exemplo desta seção, será utilizada uma etapa para recuperação de documentos e outra na sequência para recuperar sentenças.

Recuperação de documentos: Normalmente, esta etapa utiliza abordagens de RI (Capítulo Recuperação de Informação) para encontrar documentos relevantes do conjunto da base de conhecimento do sistema utilizando os termos da pergunta de entrada para a consulta (Gonçalo Oliveira et al., 2019). É possível também utilizar outras etapas do passo de Processamento da Pergunta que otimizem os termos da pergunta de entrada, como extração de palavras-chave. Em (Costa; Cabral, 2008), é utilizada uma etapa de reformulação da pergunta para padrões de perguntas do português já definidos, facilitando a busca por informações para a resposta.

Recuperação de sentenças: Mesmo que poucos documentos sejam selecionados como relevantes, estes normalmente apresentam diversas informações textuais, onde muitas podem ser irrelevantes para a pergunta. Assim, novas etapas que buscam filtrar ainda mais as informações relevantes trazem mais precisão ao modelo, uma vez que o conjunto de respostas candidatas derivadas das informações relevantes selecionadas deve ser reduzido. Uma possibilidade é buscar apenas as sentenças relevantes destes documentos, descartando as demais. Para isso, é necessário primeiro dividir o documento textual em sentenças.

A divisão do documento em sentenças não é uma tarefa trivial através da divisão pelos caracteres de pontuação, que normalmente dividem o texto em sentenças, pois estes caracteres podem ser ambíguos, e dependendo do contexto, não significam uma divisão por sentenças. Por exemplo, o ponto final '.', pode ser utilizado dentro de números, como neste exemplo "Foram gastos R$ 3.500.000,00 no investimento". De qualquer forma, existem bibliotecas especializadas na divisão de textos em sentenças para o português, como o *SpaCy*, conforme o exemplo de código abaixo.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

```
# Baixar modelo: python -m spacy download pt_core_news_sm
import spacy
nlp = spacy.load("pt_core_news_sm")
texto = "Hoje, o preço do petróleo subiu 5.7%. Amanhã haverá uma reunião importante sobre economia global."
doc = nlp(texto)
sentencas = [sent.text for sent in doc.sents]
```

Após a divisão das sentenças, é preciso determinar quais delas são relevantes ou não para a pergunta. Para isso, existem diferentes estratégias, como, por exemplo, verificar se existem termos da pergunta presentes na sentença. Outra forma é verificar se a sentença contém alguma entidade do mesmo tipo da pergunta. Neste caso, é necessário uma etapa de classificação da pergunta e também a utilização de modelos de REN para identificar as entidades nas sentenças (Capítulo Extração de Informação). Por fim, é possível também calcular um valor de similaridade entre a pergunta de entrada e a sentença através de métodos que verifiquem a similaridade entre textos. Existem diferentes métricas e modelos que buscam um valor que represente esta similaridade. Esta abordagem tem a vantagem de considerar aspectos semânticos, como sinônimos.

14.3.1.3 Processamento da Resposta

O último passo do sistema de PR é o Processamento da Resposta, que realiza as etapas de processamento para determinar qual será a resposta de saída. Neste passo, são utilizadas as informações dos passos anteriores, principalmente as informações do passo de Recuperação de Informação. As etapas deste passo podem mudar significativamente com o tipo de pergunta que o sistema está trabalhando. Sistemas para perguntas que requerem respostas longas podem utilizar abordagens de geração de texto para a resposta de saída. No caso do nosso sistema de exemplo, estamos trabalhando com perguntas factuais que requerem uma entidade como resposta. Assim, optamos por utilizar etapas de extração de respostas candidatas e ranqueamento de respostas candidatas.

Extração de respostas candidatas: Uma vez que o passo de Recuperação de Informação encontrou as informações relevantes, é comum que a próxima etapa seja a definição de respostas candidatas. A abordagem desta etapa deve mudar conforme o tipo de pergunta. Podem existir abordagens que extraem respostas baseadas em padrões, como a do sistema Esfinge (Costa, 2009). No caso de respostas longas, podem ser utilizados: modelos de sumarização capazes de sumarizar as informações dos documentos relevantes em respostas; modelos de geração de texto, que receberiam a pergunta e informações de contexto para gerar o texto da resposta; existe também a possibilidade de extrair diretamente pedaços de texto, como um parágrafo ou conjunto de frases, como a resposta candidata do sistema; por fim, o uso de templates pode ser uma possibilidade, onde um esqueleto de resposta pré-definido é preenchido com informações extraídas dos documentos, permitindo a geração de respostas mais estruturadas e controladas.

No caso do nosso sistema de exemplo, focado em perguntas factuais, a etapa de extração de respostas candidatas pode utilizar um modelo de REN para extrair as entidades de cada sentença relevante e considerar apenas as entidades do mesmo tipo da classe como resposta candidata. Podemos considerar o exemplo da Figura 14.7, onde a classe da pergunta é "Pessoa". Logo, podemos utilizar apenas os nomes de pessoas como resposta candidata.

Outra maneira de extração de respostas candidatas é através de modelos para a tarefa de compreensão de leitura (*reading comprehension*). Esses modelos são treinados em grandes conjuntos de dados que contêm perguntas com trechos de texto correspondentes e as respostas corretas. Eles aprendem a processar o contexto e a extrair a informação relevante

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

que responde à pergunta. Para perguntas factuais, esses modelos podem ser especialmente eficazes, pois são capazes de identificar e extrair a entidade exata que responde à pergunta a partir do texto fornecido.

Ranqueamento de respostas candidatas: Após criar uma lista das respostas candidatas, a próxima etapa é ranquear essa lista, onde as respostas mais prováveis ficarão no topo deste ranque. Para criar este ranque, é necessário atribuir um valor de pontuação para cada resposta candidata. Existem diferentes estratégias para determinar este valor de pontuação. Uma abordagem possível é verificar quais as respostas mais comuns na lista de candidatas. Por exemplo, se a lista de respostas candidatas permite repetições, uma pontuação possível seria determinar quantas vezes a resposta ocorre nesta lista. Assim, as respostas mais repetidas ficariam no topo do ranque. Outra possibilidade é verificar a semelhança da resposta candidata com a pergunta de entrada. Neste caso, pode-se utilizar técnicas para determinar semelhança de texto, como a similaridade por cosseno (Si et al., 2019), onde é necessário mapear tanto a resposta como a pergunta em vetores utilizando um mesmo espaço semântico multidimensional. Depois, verificar a proximidade entre eles comparando o cosseno do ângulo entre os dois vetores. Quanto menor o ângulo, maior a similaridade, com um ângulo de 0 graus indicando máxima similaridade ou vetores idênticos.

Uma possibilidade de método para ranquear respostas candidatas é usar modelos de aprendizado de máquina que foram treinados para avaliar a relevância de uma resposta candidata dada a pergunta e o contexto. Esses modelos podem levar em consideração diversos fatores, como a similaridade semântica, a presença de palavras-chave, a confiabilidade da fonte de onde a resposta foi extraída, e até feedback de usuário/as anteriores. Para perguntas factuais, o ranqueamento também pode envolver a verificação da precisão factual das respostas candidatas. Isso pode ser feito através de consulta a bases de dados confiáveis ou utilizando modelos que foram treinados para validar a veracidade das informações.

Ao final deste processo de ranqueamento, a resposta candidata que recebe a pontuação mais alta é selecionada como a resposta final do sistema. Esta resposta é então entregue ao usuário/a, completando o ciclo do sistema de PR. É importante notar que os usuários podem ter a opção de visualizarem várias respostas top ranqueadas, permitindo-lhes escolher a que acham mais satisfatória ou explorar diferentes perspectivas sobre o tema questionado.

14.3.2 Abordagem modular com grafo de conhecimento

Nesta abordagem, consideramos um sistema de PR que utiliza grafos de conhecimento como fonte principal para recuperar respostas de perguntas factuais sobre um domínio restrito. Grafos de conhecimento são estruturas que armazenam informações de forma semântica, representando entidades como nós e relações entre elas como arestas, como mostrado no exemplo da Figura 14.5, apresentada na Seção 14.2.2.2. Esta abordagem permite a recuperação precisa de informações, pois o sistema pode navegar pelo grafo para encontrar respostas específicas. Além disso, esta abordagem é particularmente útil para perguntas que exigem compreensão e inferência complexas, além de oferecer uma maneira mais estruturada de representar e acessar dados em comparação com os dados não estruturados. Como exemplo, temos o sistema apresentado em (Sousa et al., 2020) para o português, que propôs uma abordagem baseada em ontologias sobre fatos.

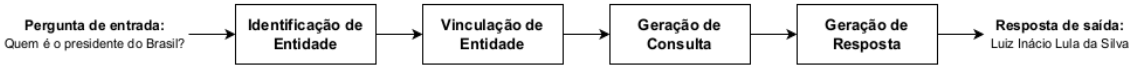
A Figura 14.8 apresenta a arquitetura geral deste tipo de sistema, destacando suas

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

principais etapas e fluxo de processamento. As etapas escolhidas foram a identificação de entidades, vinculação de entidades, geração de consultas e geração de respostas. Existem diferentes abordagens de arquitetura e processamento com grafos de conhecimento para sistemas de PR. Neste caso, foram escolhidas etapas baseadas em *parsing* semântico. Porém, existem outras abordagens como baseadas em *template* e modelos *end-to-end*, conforme o estudo de revisão de (Pereira et al., 2022).

Figura 14.8: Exemplo de uma arquitetura de um sistema de PR utilizando um grafo de conhecimento como base de conhecimento.



Identificação de Entidade: Assim como na abordagem modular com documentos, a primeira etapa envolve o processamento da pergunta do usuário/a. No entanto, nesta abordagem, o foco está em entender quais entidades estão sendo referenciadas na pergunta. Isso pode envolver modelos de REN para identificar entidades e desambiguação. Esta etapa é crucial, pois a identificação correta de entidades influenciará todas as etapas subsequentes. Por exemplo, na pergunta "Quem é o presidente do Brasil?", a entidade de interesse é "Brasil".

Vinculação de Entidade: Após identificar as entidades, o sistema tenta mapear a entidade identificada a um nó correspondente no grafo de conhecimento. Este processo é desafiador, pois uma única entidade pode ter múltiplas representações. Isso é similar à etapa de Recuperação de Informação na abordagem não estruturada, mas em vez de procurar documentos ou trechos, o sistema busca nós específicos dentro do grafo. No exemplo anterior, o sistema vincularia "Brasil" ao nó correspondente no grafo de conhecimento que representa o país.

Geração de Consulta: Uma vez que as entidades são vinculadas corretamente aos seus nós correspondentes, a próxima etapa é a geração de consultas. Esta etapa envolve a construção de uma consulta estruturada (geralmente uma consulta SPARQL se estiver usando um grafo como Freebase ou DBpedia) que será usada para extrair informações do grafo de conhecimento. A geração de consultas depende do entendimento do sistema sobre a pergunta do usuário/a e das entidades vinculadas. Seguindo nosso exemplo, uma consulta SPARQL poderia ser gerada para buscar a pessoa que tem a relação "presidente de" com o nó "Brasil".

Da mesma forma que existem abordagens que utilizam modelos de *transformers* para transformar uma entrada textual em uma consulta SQL, como a utilizada no sistema de (José et al., 2022), é possível também treinar um modelo para transformar a pergunta de entrada em linguagem natural em uma consulta SPARQL para um grafo de conhecimento.

Geração de Resposta: Finalmente, uma vez que a consulta é executada e os dados relevantes são recuperados do grafo, o sistema precisa gerar uma resposta compreensível para o usuário/a. A complexidade desta etapa pode variar dependendo da natureza da pergunta e da estrutura do grafo de conhecimento. Isso pode envolver simplesmente retornar o nome de uma entidade ou uma lista de entidades, ou pode envolver mais processamento para gerar uma resposta longa em linguagem natural, como utilizar modelos de geração de linguagem natural.

A abordagem com grafos de conhecimento tem a vantagem de utilizar uma base de conhecimento estruturada e semântica, o que pode melhorar a precisão e a relevância das respostas, especialmente para perguntas que requerem compreensão e inferência comple-

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

xas. No entanto, também apresenta desafios, como a necessidade de manter e atualizar constantemente o grafo de conhecimento para refletir informações precisas e atuais. Enquanto a abordagem modular com documentos pode ser mais flexível e capaz de lidar com uma gama mais ampla de perguntas, a abordagem modular com grafo de conhecimento oferece maior precisão e eficiência para perguntas específicas onde a vinculação direta a entidades conhecidas é possível. A escolha entre as duas abordagens dependerá das necessidades específicas do sistema de PR, como tipo de perguntas que ele visa responder e o tipo de informação disponível para consulta.

14.3.3 Abordagem *end-to-end*

A abordagem *end-to-end* representa um design de sistemas de PR onde o objetivo é criar um modelo que possa lidar com todas as etapas do processo de PR, desde a compreensão da pergunta até a geração da resposta, sem intervenção ou etapas de processamento intermediárias. Ao invés de separar as tarefas em diferentes passos (como Processamento da Pergunta, Recuperação de Informações e Processamento da Resposta), a abordagem *end-to-end* busca unificar todas essas operações em um único modelo.

Neste caso, estamos considerando como exemplo um sistema de PR de amplo domínio para perguntas factuais e não factuais. Ao invés de utilizar somente a pergunta como entrada do modelo, incluímos também um contexto, que é um texto com informações relevantes e confiáveis, que normalmente tem tamanho de um grande parágrafo de texto. Assim, o modelo utiliza memória paramétrica junto com o contexto textual extraído de alguma fonte de conhecimento externa confiável. Essa estratégia traz mais confiabilidade ao sistema, já que confiar unicamente na memória paramétrica do modelo pode trazer problemas de imprecisão e delírio ao gerar a resposta.

Diferente das figuras de arquiteturas dos modelos modulares, a Figura 14.9 ilustra as etapas de desenvolvimento deste modelo, destacando que, no contexto de um modelo *end-to-end*, as etapas de processamento convencionais não são explicitamente definidas, mas sim integradas de forma implícita. Porém, é importante ressaltar que não podemos afirmar com certeza que o modelo adota uma hierarquia ou sequência de etapas conhecidas de forma modular. Na verdade, o modelo gera uma série de representações que, combinadas, podem incorporar certas subtarefas de maneira codificada. Por fim, antes de encaminhar uma pergunta ao modelo *end-to-end*, torna-se necessário contextualizá-la com informações de alguma fonte de conhecimento externa. Portanto, adicionamos essa etapa única de processamento, análoga ao passo de Recuperação de Informação descrito na Seção 14.3.1.

Figura 14.9: Exemplo de passos para criação de um modelo *end-to-end* para PR.



Ao invés de focarmos unicamente na descrição, funcionalidades e aplicabilidades das etapas do sistema, nosso foco aqui se voltará, principalmente, para as etapas empregadas no desenvolvimento do modelo.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

14.3.3.1 Pré-processamento dos Dados

Antes de qualquer treinamento, é fundamental preparar um conjunto de dados para treinamento dos modelos de linguagem. O primeiro treinamento destes modelos (pré-treinamento) utilizam grandes corpora de texto que não precisam ser anotados e servem para o modelo aprender os padrões fundamentais da linguagem natural, que formarão a base para tarefas mais específicas de PLN. O segundo treinamento deve utilizar dados específicos para a tarefa de PR. Assim, é necessário um conjunto de pares de perguntas e resposta junto com o contexto relevante. Este conjunto pode ser consideravelmente menor do que o conjunto de dados utilizado na fase de pré-treinamento. Além das instâncias para treinamento, podem ser usadas instâncias para testar o modelo.

14.3.3.2 Pré-treinamento

Esta etapa envolve o treinamento do modelo em uma grande quantidade de dados de texto não anotados. Aqui, modelos de *transformers* que já foram pré-treinados em tarefas gerais de processamento de linguagem natural podem ser utilizados como ponto de partida, o que pode economizar tempo e recursos. Modelos como BART (Lewis et al., 2020), GPTs (Achiam et al., 2023; Brown et al., 2020), ou T5 (Roberts et al., 2019) são exemplos comuns que podem ser adaptados para a tarefa de PR. O pré-treinamento serve para que o modelo adquira um entendimento básico da linguagem. Esta etapa normalmente requer elevado recurso de tempo e computacional para o treinamento utilizando o grande volume de dados de corpus.

É possível aproveitar modelos pré-treinados, como Sabiá (Pires et al., 2023b) e BERTIMBAU (Souza et al., 2020), que são específicos para o português e já foram treinados em extensos conjuntos de dados. Esses modelos podem ser refinados para tarefas específicas de PLN, como PR (Oliveira et al., 2021). Para a geração de respostas longas, especialmente para perguntas não factuais, recomenda-se o uso de arquiteturas *full-transformers*, que são compostas por *encoders* e *decoders*, ou arquiteturas *decoder only*. Essas arquiteturas são apropriadas para gerar texto e com capacidade de gerar saídas textuais mais extensas. Por outro lado, arquiteturas *encoder only*, como a do modelo BERT, consistem apenas de módulos de codificação da linguagem e são mais adequadas para tarefas como classificação de texto.

14.3.3.3 Treinamento específico

Após o pré-treinamento, a próxima etapa é o ajuste fino, onde o modelo é treinado novamente em um conjunto de dados específico para a tarefa de PR. Este conjunto de dados consiste em pares de perguntas e respostas, junto com o contexto relevante. É nesta etapa que o modelo realmente aprende a função de PR, ajustando-se para compreender como as perguntas se relacionam com os contextos e quais respostas são as mais adequadas.

14.3.3.4 Teste e validação

Uma vez treinado, é recomendado que o modelo seja rigorosamente testado e validado para garantir que está produzindo respostas corretas e relevantes. Isto é feito usando um conjunto de dados separado que não foi visto pelo modelo durante o treinamento. As métricas de desempenho, como precisão, revocação, BLEU (ver Subseção *Bilingual Evaluation Under-study* (BLEU)) e BERTScore, são comumente utilizadas para avaliar a

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

qualidade das respostas do modelo. Os métodos de avaliação serão vistos em mais detalhes na Seção 14.4.

Em comparação com as abordagens modulares, o modelo *end-to-end* abstrai muitos dos processos intermediários. No entanto, a etapa de Recuperação de Informação para a obtenção do contexto é um ponto em comum entre as abordagens. A principal vantagem de um sistema *end-to-end* é sua capacidade de aprender representações internas e relacionamentos complexos nos dados, o que pode levar a um melhor desempenho, maior generalização, e a respostas mais precisas. No entanto, sistemas *end-to-end* também podem ser mais difíceis de interpretar e requerem conjuntos de dados substanciais e poder computacional significativo para o treinamento. Além disso, a abordagem *end-to-end* representa o estado da arte em sistemas de PR, aproveitando as recentes inovações em modelos de aprendizado profundo, grandes conjuntos de dados e poder computacional.

Esta abordagem oferece uma integração mais fluída e direta entre as etapas do processo de PR, potencialmente reduzindo os erros que podem ocorrer em sistemas com múltiplas etapas independentes. Além disso, ao aprender diretamente de exemplos de pares de pergunta-resposta com contexto, os modelos *end-to-end* podem desenvolver uma compreensão mais refinada das sutilezas e variações da linguagem natural, o que é crucial para responder perguntas não factuais e complexas.

Contudo, é importante ressaltar que, apesar da eficiência e sofisticação, a abordagem *end-to-end* ainda enfrenta desafios, especialmente relacionados à qualidade e diversidade do conjunto de dados utilizado (Gururangan et al., 2020), bem como à necessidade de interpretabilidade e explicabilidade dos resultados produzidos pelo modelo (Linardatos et al., 2021; Tjoa; Guan, 2021). A constante evolução dos modelos de linguagem e o avanço das técnicas de aprendizado de máquina continuam a impulsionar o desenvolvimento de sistemas de PR mais robustos e precisos.

14.4 Métodos de avaliação

O Capítulo 18 apresenta de forma geral as questões de avaliação de tecnologias de linguagem, aqui vamos situar o uso de algumas métricas para o contexto de PR. Existem diferentes métodos de avaliação de sistemas de PR que podem ser mais adequados para as diferentes tarefas computacionais envolvidas nas etapas de processamento do sistema. Por exemplo, para a etapa de classificação de perguntas, é adequado empregar métodos de avaliação utilizados em problemas de classificação. Já para a etapa de recuperação de documentos, é adequado métodos de avaliação utilizados para RI. Em relação à avaliação direta das respostas de saída do sistema, existem métodos adequados para respostas curtas, como as perguntas factuais, e outros mais adequados para respostas longas.

Um dos principais recursos de avaliação de sistema de PR é conjunto de dados (*dataset*), que normalmente contém perguntas de entrada e respostas consideradas corretas para estas perguntas. Além disso, os conjuntos de dados podem oferecer outros recursos adicionais, como a base de conhecimento que o sistema deve utilizar para consulta, informações adicionais sobre a pergunta, como o seu tipo, entidades no texto, entre outros. Assim, os sistemas testados recebem como entradas as mesmas perguntas do conjunto de dados, podendo assim comparar as respostas geradas pelos diferentes sistemas de PR, como também as saídas das etapas de processamento de cada sistema, e compará-los a fim de verificar qual apresenta o melhor desempenho.

Para a comparação entre sistemas, são utilizadas métricas de avaliação que fornecem uma base quantitativa para medir o desempenho, através da comparação da resposta de

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

saída do sistema com respostas de referencias encontradas do conjunto de dados. Estas métricas variam de acordo com o tipo de resposta que o sistema de PR está projetado para gerar. Para respostas curtas e precisas, como um nome, uma data ou um fato específico, as métricas são normalmente usadas para avaliar a capacidade do sistema de identificar corretamente a resposta exata dentro de um conjunto de respostas candidatas. As principais métricas são:

- **Precisão**: mede a proporção de respostas corretas entre todas as respostas fornecidas. Por exemplo, um sistema que responde a 100 perguntas, mas apenas 80 respostas são corretas, a precisão será de 80%.

$$Precisão = \frac{\text{Número de Respostas Corretas}}{\text{Número Total de Respostas Fornecidas}} \tag{14.1}$$

- **Revocação**: avalia a proporção de respostas corretas identificadas pelo sistema em relação ao total de respostas corretas possíveis. Por exemplo, se para 100 perguntas, o sistema identifica corretamente 80 respostas de um total de 120 respostas corretas possíveis, a revocação é de 66.67%.

$$Revocação = \frac{\text{Número de Respostas Corretas}}{\text{Número Total de Respostas Corretas Possíveis}} \tag{14.2}$$

- **F1-Score**: considera a precisão e a revocação juntas, através de sua média harmônica.

$$F1\text{-}Score = 2 \times \frac{\text{Precisão} \times \text{Revocação}}{\text{Precisão} + \text{Revocação}} \tag{14.3}$$

- **Acurácia**: simplesmente mede a porcentagem de respostas corretas fornecidas pelo sistema. Por exemplo, se o sistema responde corretamente a 85 de 100 perguntas, a acurácia é de 85%.

$$Acurácia = \frac{\text{Número de Respostas Corretas}}{\text{Número Total de Respostas}} \tag{14.4}$$

Para respostas longas, podem ser utilizadas métricas que determinam a similaridade entre textos, que incluem:

- **BLEU (*Bilingual Evaluation Understudy*)**: Originalmente desenvolvida para avaliação de tradução automática, compara a resposta do sistema com uma ou mais respostas de referência. Essa comparação é feita com base na sobreposição de n-gramas (sequências de palavras) entre a resposta gerada e as respostas de referência (Papineni et al., 2002).

- **ROUGE (*Recall-Oriented Understudy for Gisting Evaluation*)**: Foca mais na capacidade do sistema de reproduzir o conteúdo das respostas de referência. Utiliza sobreposição de n-gramas, subsequências mais longas e co-ocorrências de palavras para avaliar resumos automáticos, por exemplo (Lin, 2004).

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

- **METEOR (*Metric for Evaluation of Translation with Explicit ORdering*)**: É similar ao BLEU, mas mais sofisticado, pois também considera sinônimos e a estrutura gramatical das respostas. METEOR tenta alinhar melhor com a avaliação humana do que o BLEU (Banerjee; Lavie, 2005).

As respostas longas são mais desafiadoras de serem comparadas e avaliadas, principalmente para as métricas baseadas em sobreposição de palavras. Uma vez que a linguagem natural permite diversas maneiras de expressar uma informação, muitas vezes respostas que representam a mesma ideia mas com palavras diferentes podem receber uma pontuação menor. Existem métricas que utilizam modelos de redes neurais profundas para predizer o valor da métrica, buscando superar este problema de sobreposição através da análise semântica e da contextualização das palavras dentro das respostas. Neste caso, temos como exemplo o modelo BERTScore.

BERTScore: alinha as palavras entre as respostas do sistema e de referência com base em seus *embeddings*, levando em conta o contexto em que as palavras são usadas. Isso permite que a métrica avalie não apenas a sobreposição exata de palavras, mas também a semelhança semântica e contextual. Assim, mesmo que as palavras usadas sejam diferentes, se elas compartilharem significados semelhantes no contexto dado, a resposta pode ser avaliada positivamente (Zhang et al., 2020).

Um dos maiores desafios na avaliação de sistemas de PR é lidar com a subjetividade, especialmente em respostas longas, onde a "correção" pode ser aberta a interpretação. Assim, muitas vezes métricas automáticas não conseguem lidar com tais sutilezas. Neste caso, podem ser empregadas métricas qualitativas, que envolvem a avaliação manual humana das respostas fornecida pelo sistema. Isso pode incluir:

- **Avaliação por Juízes Humanos**: Avaliadores humanos analisam as respostas geradas pelo sistema para julgar critérios específicos, como relevância, precisão e naturalidade.
- **Testes de Usabilidade**: Observação de como os usuário/as finais interagem com o sistema e coletar seus feedbacks sobre a eficácia e utilidade das respostas.

Finalmente, a avaliação de sistemas de PR contém múltiplas abordagens e pode levar em conta tanto métricas quantitativas quanto qualitativas para obter uma compreensão completa do desempenho do sistema. É importante escolher as métricas adequadas com base no tipo de pergunta, na natureza das respostas esperadas e nas etapas específicas do processo de PR envolvidas.

14.5 Outras tarefas relacionadas a PR

Esta seção aborda tarefas adicionais e específicas no campo de PR, que se destacam por suas peculiaridades e desafios únicos. Essas tarefas ampliam o escopo convencional de PR, integrando aspectos como comunidades online, processamento de múltiplos passos e dados visuais.

14.5.1 Comunidades de Perguntas e Respostas - *Community* QA

Community QA refere-se a sistemas onde as perguntas e respostas são geradas pela comunidade, como em fóruns online ou plataformas de perguntas e respostas. Um desafio chave

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

aqui é o ranqueamento das respostas. Em um *Community QA*, uma pergunta pode receber várias respostas de diferentes usuário/as. O objetivo é então classificar estas respostas com base na sua relevância e utilidade para a pergunta feita.

Por exemplo, em um fórum sobre programação, uma pergunta sobre um problema específico em Python pode receber diversas soluções propostas por outros usuário/as. O sistema de *Community QA* analisará estas respostas levando em consideração fatores como clareza, exatidão técnica, votos dos usuários e talvez até a reputação do respondente no fórum. Em seguida, ele classificará as respostas para que a mais útil apareça no topo.

O estudo de (Castro Ferreira et al., 2021) buscou desenvolver um *benchmark* para o português de *Community* QA no domínio da diabetes. Neste caso, a tarefa consiste em: dada uma pergunta de entrada, é necessário recuperar perguntas armazenadas em uma base de perguntas e respostas, que sejam semanticamente semelhantes e já respondidas. Assim, um sistema com estas características não precisa necessariamente gerar um resposta para uma pergunta, já que a resposta estaria pronta e armazenada em uma base de perguntas e respostas.

14.5.2 PR de múltiplas etapas *Multihopping* QA

A tarefa de *Multihopping* QA aborda perguntas que requerem múltiplos passos de raciocínio para encontrar a resposta. Diferente de perguntas diretas, onde a resposta pode ser encontrada em um único local, o *Multihopping* QA exige que o sistema combine informações de múltiplas fontes de dados.

Por exemplo, considere a pergunta: "Qual o autor do livro favorito do presidente dos Estados Unidos". Para responder, o sistema primeiro precisa identificar quem é o presidente atual, depois encontrar qual é o seu livro favorito, e finalmente, descobrir o autor desse livro. Cada um desses passos é um "salto" no processo de raciocínio, tornando a tarefa complexa e desafiadora.

14.5.3 PR com dados visuais

PR com dados visuais combina elementos de visão computacional com PR, onde as perguntas estão relacionadas a conteúdos visuais, como imagens ou vídeos. O desafio aqui é entender a pergunta e correlacioná-la com a informação visual para gerar uma resposta apropriada.

Um exemplo clássico é o de perguntas sobre o conteúdo de uma imagem. Suponha que temos uma foto de uma rua movimentada, e a pergunta é: "Quantos carros azuis estão na imagem". O sistema de PR precisa processar a imagem para identificar e contar os carros azuis. Este tipo de tarefa requer habilidades de processamento de imagem além do entendimento da linguagem natural.

14.6 Sistemas de PR para o Português

Os sistemas de PR para o idioma português têm registrado um avanço notável, com progressos tanto na análise linguística quanto na aplicação de tecnologias de inteligência artificial. A revisão de trabalhos na área mostra uma evolução que abrange diversas metodologias e áreas de aplicação, como sistemas focados em textos jurídicos, que se baseiam em bases de dados estruturados, ou com corpora especializados para treinamento de modelos de aprendizado de máquina.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Inicialmente, sistemas como o do estudo de (Quaresma; Rodrigues, 2005) evidenciaram os desafios no desenvolvimento de PR em domínio específico, como o jurídico, onde se buscam respostas exatas diante da dificuldade de compreender a linguagem natural. Também, notou-se desafios similares com os sistemas de domínio amplo, como o sistema Esfinge (Costa, 2009), onde se encontram desafios relacionados com a análise semântica. Esses trabalhos, junto a outras iniciativas que buscavam respostas na web e em fontes externas, marcaram esforços para aprimorar etapas de processamento do sistema de PR para o português em busca da acurácia das respostas.

Outros sistemas se destacaram, como o IdSay (Carvalho et al., 2009) e Priberam (Amaral et al., 2008), que representaram um avanço nas abordagens de PR, aplicando técnicas mais avançadas de PLN e RI. Tais sistemas demonstraram melhorias significativas na precisão das respostas e na capacidade de processar um espectro mais amplo de tipos de perguntas. Além destes, houveram trabalhos focados em avaliação de sistemas, como Págico, destacado no LREC 2012 por (Mota et al., 2012). Este trabalho focou na melhoria de PR ao utilizar a Wikipedia em português, criando dados para avaliações futuras.

Com a maior disponibilidade de grandes volumes de dados e o avanço das técnicas de aprendizado, trabalhos mais recentes como o sistema de (Gonçalo Oliveira et al., 2019), DEEPAGÉ (Cação et al., 2021), e ClinicalQA (Oliveira et al., 2021) exploraram o uso de modelos pré-treinados e algoritmos avançados para melhorar ainda mais a eficácia dos sistemas de PR. O uso do modelo BERTimbau (Souza et al., 2020) e a aplicação de fine-tuning específico para domínios restritos, como visto no estudo sobre *Blue Amazon* (Spindola et al., 2021) e no trabalho de (Silva et al., 2022a), destacam a importância da personalização dos modelos para contextos específicos e a capacidade de superar resultados.

Adicionalmente, a utilização de novas estratégias, como a integração de PR com a geração de consultas SQL a partir de perguntas em linguagem natural (José et al., 2022), mostra uma diversificação de técnicas de PR. Essas inovações mostram uma tendência de sistemas de PR cada vez mais versáteis, capazes de oferecer respostas precisas e relevantes em variados domínios e condições.

Em resumo, a evolução dos sistemas de PR para o português é caracterizada por um desenvolvimento em direção à melhorias de precisão, abrangência e flexibilidade. Através da aplicação de métodos de PLN, RI e aprendizado de máquina cada vez mais sofisticados, esses sistemas estão se consolidando como ferramentas fundamentais para acesso à informação e apoio à decisão baseada em dados em diversas áreas de conhecimento e atividades.

14.7 Considerações finais

Neste capítulo, exploramos a área de PR, oferecendo uma visão abrangente de suas diversas características, desde a classificação de sistemas até as metodologias e desafios associados. A área de PR, como visto, lida com vários desafios e envolve a interseção de várias subáreas do PLN, bem como desafios únicos em termos de compreensão e geração de linguagem natural.

Abordamos sistemas de PR sob diferentes perspectivas, incluindo a natureza das perguntas (factuais vs. não factuais), o tipo de fonte de conhecimento (documentos não estruturados, dados estruturados, e memória paramétrica), e o domínio de conhecimento (amplo vs. restrito). Cada uma dessas classificações apresenta desafios e abordagens distintas, demonstrando a necessidade de estratégias adaptativas e soluções no design de sistemas de PR.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

As metodologias discutidas refletem a evolução contínua da área, desde abordagens modulares que separam explicitamente as etapas de processamento, até modelos *end-to-end* que integram todas as operações em um único sistema. O desenvolvimento de modelos *end-to-end*, em particular, destaca-se como um avanço significativo, trazendo eficiência ao processamento de PR, embora ainda apresente desafios em termos de interpretabilidade e necessidade de grandes conjuntos de dados para treinamento.

A avaliação de sistemas de PR foi discutida, mostrando abordagens de avaliação para diferentes características. Também, foram exploradas tarefas adicionais e específicas no campo de PR, como *Community QA*, *Multihopping* QA e PR com dados visuais. Essas tarefas expandem o escopo de PR, integrando desafios de comunidades online, raciocínio de múltiplos passos e processamento de dados visuais.

É importante destacar os avanços com sistemas baseados em *Large Language Models* (LLMs), como o ChatGPT, que representam uma inovação significativa na capacidade de sistemas de PR de responder perguntas de maneira eficaz. Esses modelos têm demonstrado habilidades impressionantes em compreender e gerar linguagem natural, oferecendo respostas contextualizadas e conversacionais que podem abranger uma vasta gama de tópicos e domínios. Porém, os LLMs ainda enfrentam limitações significativas, como a ocorrência de alucinações. Isso destaca a necessidade de mecanismos de verificação de fatos e integração com fontes de conhecimento confiáveis para garantir a precisão das respostas fornecidas.

Além das questões de confiabilidade, os sistemas baseados em LLMs apresentam desafios significativos relacionados ao custo e aos requisitos de recursos computacionais para sua utilização. O treinamento e a execução destes modelos exigem uma quantidade substancial de poder computacional. Esses requisitos podem resultar em custos proibitivos para pesquisa e desenvolvimento, especialmente para organizações menores ou pesquisadores independentes. Neste contexto, os sistemas de PR convencionais continuam a desempenhar um papel importante. Estes sistemas, que muitas vezes utilizam abordagens mais tradicionais como abordagem modular, podem ser significativamente menos custosos em termos de recursos computacionais e financeiros.

Olhando para o futuro, espera-se que os sistemas de PR continuem a evoluir, integrando avanços em áreas como aprendizado profundo, compreensão de linguagem natural e processamento de dados multimídia. Além disso, a crescente importância da ética e da privacidade na IA sugere um futuro onde os sistemas de PR devem ser cada vez mais transparentes, justos e seguros. Em termos de aplicação prática, espera-se uma expansão contínua dos sistemas de PR em vários setores, como saúde, educação, assistência jurídica e atendimento ao cliente. A integração de sistemas de PR em interfaces conversacionais e assistentes virtuais promete uma interação mais natural e intuitiva entre humanos e máquinas.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Capítulo 15

Diálogo e Interatividade

Brielen Madureira

Publicado em: 13/03/2024

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

15.1 Introdução: Dialogando

"O diálogo é uma exigência existencial". Assim se referiu Freire (1989) a uma atividade que, de modo geral, é feita com naturalidade e desenvoltura pelos seres humanos: conversar. De fato, **boa parte das relações humanas são intermediadas por alguma forma de diálogo**: aprendizagem escolar, entrevista de emprego, negociações, cultivo de relacionamentos afetivos, psicoterapia, debates acadêmicos e políticos etc. É algo tão habitual que, enquanto mantemos o propósito principal da interação em foco, os fenômenos dialógicos em si passam até despercebidos. Todavia, basta haver algum impedimento à fluência típica de uma conversa (patologias que afetam a fala, interlocutores desatentos, um ruído alto, uma frase mal formulada) para nos atentarmos mais aos detalhes de como se dá essa forma tão fundamental de comunicação humana.

Conversar é pôr em uso a linguagem natural em um contexto, para algum fim, em uma atividade que exige muita coordenação, agilidade e planejamento. Diálogos podem ocorrer entre dois ou mais participantes e, embora a forma oral seja mais autêntica e convencional na existência humana, o diálogo através de mensagens de texto instantâneas permeia também o cotidiano em diversos grupos sociais da atualidade. Muitas dimensões podem variar em uma conversa, como espontaneidade, formalidade, sinceridade e naturalidade, além de ela estar sujeita a influências sociolinguísticas, de estruturas de poder, de adequação ao ambiente e de pressuposições (Abbott, 2008; Boxer, 2002; Danescu-Niculescu-Mizil et al., 2012; Prabhakaran; Rambow, 2013). Todos nossos microatos durante uma conversa podem carregar significado, inclusive os silêncios, o olhar, as expressões faciais, os gestos, as hesitações e o tom da fala.

Vamos inspecionar brevemente o exemplo na Figura 15.1, o qual contém uma transcrição de duas pessoas dialogando[1]. Podemos observar algumas peculiaridades: há diversas pausas, frases que não seguem formas idealizadas de sintaxe (é os avós), ou que são interrompidas ("achei uma" dá lugar a "adorei"), sinais de compreensão (hum hum) e sobreposição (a palavra "nunca" de MRN ocorre junto com a de HLS no áudio) e uma clarificação

[1] Este exemplo e todos os demais ao longo deste capítulo são de conversas genuínas, extraídas do *Corpus* de Referência do Português Contemporâneo (CRPC) do Centro de Linguística da Universidade de Lisboa – CLUL (versão 3.0 2012, através da plataforma CQPWeb no período de novembro de 2023 e fevereiro de 2024). Em seu *site* é possível ouvir os áudios correspondentes. Nas transcrições, / significa uma pausa curta, // significa uma pausa longa, + se refere a interrupções ou abandonos e trechos riscados são repetições, reformulações ou cortes. Disponível em https://clul.ulisboa.pt/projeto/crpc-corpus-de-referencia-do-portugues-contemporaneo.

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

("meus pais não / mas meus avós"). Já podemos perceber que a **fala-em-interação** (e sua transcrição) diverge bastante da forma canônica dos textos em grandes *corpora*, extraídos de livros ou de matérias de jornais, os quais geralmente foram bem lapidados por processo de edição.

Figura 15.1: Transcrição de um trecho de diálogo entre MRN e HLS em português brasileiro falado em conversa face a face, onde observamos diversos fenômenos do diálogo como pausas curtas / e longas //, frases inacabadas, sobreposições, sinais de compreensão e clarificação.

HLS - ~~achei~~ ~~achei uma~~ adorei //

HLS - agora / eu sou louca para conhecer Olinda //

MRN - ah é ?

HLS - é que eu não conheço o norte / eu tenho maior vontade //

MRN - os pais pernambucanos / mas você

MRN - nunca foi ?

HLS - nunca

HLS - fui a Pernambuco //

HLS - ~~não~~ ~~não s~~ meus pais não / mas meus avós //

MRN - é os avós //

HLS - a minha família

HLS - toda é pernambucana / é //

MRN - hum hum //

HLS - e eu nunca fui //

HLS - mas eu tenho uma vontade / muito grande / de conhecer //

HLS - Olinda / ir a Maranhão / a São Luís / isso tudo //

Fonte: CRPC, arquivo Arte_Urbana

Antes de prosseguir com a leitura, sugerimos um outro pequeno exercício. Procure um vídeo em que haja pessoas em uma conversa espontânea, como amigas e amigos em uma festa ou um debate livre (isto é, evite trechos de seriados, filmes ou gravações com falas planejadas). Se lhe for possível, assista primeiro sem som. Preste atenção nos olhares, gestos e expressões faciais. Depois, assista-o novamente, agora focando na *forma* com que as pessoas conversam, sem se ater tanto ao conteúdo. Tome nota do que você percebe,

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

por exemplo: Como elas decidem quem vai falar? Como elas corrigem seus próprios erros ou os erros da interlocutora ou do interlocutor? Quando ocorrem pausas? As frases são todas bem formuladas? Que estratégias os participantes usam para demonstrar e resolver pontos que não ficaram claros? Que efeitos parecem ter os olhares? Que tipos de gestos e expressões são feitos? Como elas se referem ao ambiente em que estão ou a pessoas, lugares ou objetos distantes?[2]

Sendo um fenômeno tão multifacetado, imbuído de tantas características de vasta complexidade, é evidente que criar modelos computacionais que capturem todas as nuances do diálogo é extremamente desafiador (e, ousamos dizer, impossível com a tecnologia atual). Mesmo os agentes conversacionais comerciais que impressionam pelas aparentes habilidades de manipulação da linguagem ainda estão longe de modelar diálogo por completo. Por exemplo, na forma usual de conversação de *chatbots*, há uma alternância forçada de turnos (eu escrevo, você escreve, eu escrevo, você escreve e assim sucessivamente) que não se assemelha à conversa espontânea entre humanos. Embora seja possível uma conversa dessa forma, ela não é tão natural e rica em fenômenos (compare-a com interações em grupos agitados do WhatsApp, por exemplo: como um *chatbot* se sairia nesse contexto?). É também comum assistentes conversacionais se confundirem em várias situações, como quando há ambiguidade ou uma correção (uma rápida busca por vídeos na internet traz diversos exemplos reais), e há pesquisa investigando limitações de seu uso no contexto brasileiro (Guerino; Valentim, 2020).

Neste capítulo, apresentaremos um panorama geral de conceitos e fenômenos do diálogo (Seções 15.2 e 15.3), bem como de tipos de modelos e tarefas (Seção 15.4), tanto os mais clássicos quanto os que se enquadram no atual paradigma de técnicas de aprendizado de máquina. Discutiremos os desafios de se avaliar modelos de diálogo, ressaltando que não basta otimizar uma métrica de performance para se ter um resultado satisfatório (Seção 15.5). Trataremos, ainda, de recursos, incluindo *corpora* existentes para o português e, mais amplamente, das diversas formas de coleta de dados interativos, que difere da coleta de textos em monólogo comumente usados para treinar modelos de PLN baseados em dados (Seção 15.6). Para concluir, discorreremos sobre algumas **questões imperativas quanto ao desenvolvimento e ao uso responsável desses modelos**, frisando a importância da preservação, em vez da banalização ou corrosão, dessa atividade humana construída socialmente ao longo da nossa evolução e que é tão valiosa para nossa existência.

15.2 Conceitos: O que forma um diálogo?

O que é diálogo? Pense, por alguns instantes, em uma definição. Nesta seção, buscaremos delinear esse fenômeno, trazendo à tona algumas de suas atribuições e alguns de seus ingredientes. Compreender as particularidades e circunstâncias do que é diálogo é essencial para definirmos e avaliarmos modelos capazes de tomar parte em conversas com humanos. Isso nos possibilita **dar o devido valor ao objeto de estudo**, indo além da técnica em si.

Iniciemos pela palavra "diálogo", que vem do grego, sendo composta pelo prefixo "dia", que significa "através", e "logos" que quer dizer "palavra" ou "discurso"[3]. Há muitas atividades que podemos realizar **através do discurso**. Intuitivamente, sabemos que, em

[2] Em tempo, prestar mais atenção em **como** as pessoas conversam (em vez de focar apenas **no quê** elas falam) pode se tornar um passatempo ou ato contemplativo interessante.

[3] Fonte: Dicionário de Oxford. Disponível em https://www.oed.com/dictionary/dialogue_n?tab=etymology#6924392.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

um diálogo, há dois ou mais participantes que se revezam em proferir falas (ou sinais)[4], e, em geral, enquanto uma pessoa fala a outra está em silêncio, supostamente prestando atenção no que está sendo dito (ou gesticulado)[5].

Vamos caracterizar melhor como esse uso **colaborativo da linguagem** ocorre. Diversas fontes consideram a conversação face a face como uso primário da linguagem humana, do qual outras formas derivam (ver Tabela 1.1 em (Bavelas, 2022)). É uma premissa razoável, visto que, como aponta Clark (1996a), a escrita e os aparatos tecnológicos (telefone, rádio etc.) não estão disponíveis em todas as sociedades humanas, enquanto a conversa é universal, além de ser a forma básica de aquisição de linguagem (Clark, 2020; Clark, 2014). Comecemos, portanto, com um formato elementar: duas pessoas estão conversando, através da voz, no mesmo ambiente que ambas podem ver. Em (Clark, 1996a), encontramos dez características dessa forma de interação:

- **co-presença**: ambos os participantes estão inseridos no mesmo ambiente físico no momento da conversa;
- **visibilidade**: ambos os participantes podem se ver;
- **audibilidade**: ambos os participantes podem se ouvir;
- **instantaneidade**: a percepção das ações do outro ocorre sem um atraso perceptível;
- **evanescência**: o meio (*e.g.* as ondas sonoras) se dissipa rapidamente;
- **ausência de registro**: as ações dos participantes não deixam rastro;
- **simultaneidade**: os participantes podem produzir e receber as falas de imediato e simultaneamente;
- **extemporaneidade**: as ações são formuladas e executadas em tempo real;
- **autodeterminação**: os participantes decidem por si próprios que ações fazer e quando fazê-las;
- **autoexpressão**: os participantes agem conforme si mesmos (isto é, não representam um papel).

A partir dessas características, conseguimos classificar melhor as várias formas de diálogo. Todavia, a proposta não é atribuir um valor de superioridade sobre essa forma em relação às demais. Interações que não atendem a todos os requisitos também podem ser diálogos, e são igualmente válidas. Por exemplo, conversas telefônicas não exibem as características de co-presença e visibilidade, e, se são gravadas, elas deixam registro. Em conversas por língua de sinais, audibilidade não precisa estar presente, enquanto pessoas com limitações de visão podem igualmente conversar. Mas a ausência de certas características pode vir a ser um critério excludente na hora de definir diálogo. Uma conversa por cartas em que as pessoas estão simulando ser celebridades não atende a nenhuma dessas características. Isso seria um diálogo? E uma peça de teatro, em que as falas não são

[4]Usaremos "fala" de modo genérico para a produção linguística, englobando também língua de sinais.

[5]Não vamos tratar aqui de outros tipos de participação em conversas, mas, para registrar, também pode haver ouvintes passivos (mediadores que apenas observam), espectadores (como uma plateia) e bisbilhoteiros ouvindo atrás da porta, que acabam tendo um papel na conversa ou a influenciando indiretamente (Clark, 1996a).

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

espontâneas? Do ponto de vista teórico, provavelmente não. Porém é difícil chegar a uma definição categórica e definitiva, de modo que, na prática, **a delimitação vai depender diretamente do uso específico da tecnologia**.

Podemos, ainda, pensar em gradações para o nível de interatividade de uma conversa (Schlangen, 2023c). Uma conversa entre familiares próximos pode ser bem espontânea, com cada interlocutor se expressando livremente, como, quando e o quanto quer. Já em uma entrevista, há uma assimetria evidente de papéis, onde uma pessoa majoritariamente faz perguntas e a outra responde. Ambientes institucionais, como um julgamento, têm regras pré-estabelecidas de quem pode tomar a palavra e em quais momentos. O número de participantes também afeta algumas características. Quando há mais participantes, tomar a palavra pode se tornar mais competitivo, além de não ser possível focar visualmente em todos ao mesmo tempo.

O que mais é relevante quando pessoas estão conversando? Clark (1996a) salienta que um diálogo é uma **ação conjunta** entre os participantes, que precisam coordenar suas ações em busca de um propósito. A ação mais evidente é a fala (o que falar, quando falar e quando não falar) junto com seu processamento (entender o que está sendo falado e que rumo a conversa vai tomando a cada passo). Na fala, todas as dimensões linguísticas discutidas nos demais capítulos deste livro têm um papel, desde fonética, prosódia e entonação, até sintaxe, semântica, análise do discurso e pragmática. Em especial, a pragmática é extremamente relevante no diálogo, pois a conversa está ocorrendo em um contexto, que inclui a situação como um todo, o discurso, o ambiente físico e a bagagem de conhecimento de cada um.

Há muitos motivos que cotidianamente exigem essa ação conjunta como narrar um fato, explicar um assunto, tomar decisões e confidenciar emoções. Mesmo em interações em que os papéis são assimétricos (por exemplo, médica e paciente), a cooperação existe: enquanto está ouvindo a contribuição narrativa das queixas do paciente, a médica não interfere muito mas pode demonstrar que algo não ficou claro com sua expressão facial, interromper para pedir esclarecimentos, ou ajudar o paciente a se lembrar de uma palavra que lhe foge.

No contexto de PLN, também devemos levar em conta o que acontece quando pelo menos um dos participantes da conversa é um agente artificial, seja incorporado em um robô físico, um avatar virtual ou apenas uma voz em um aparato imóvel. **A atitude das pessoas muda conforme o interlocutor e o ambiente** (Giles, 2016); por exemplo, não falamos da mesma maneira com uma criança em casa e com uma juíza em tribunal. Diferentes expectativas e adaptações também podem ser desencadeadas quando o interlocutor é um programa de computador (Bernsen et al., 1996; Mol et al., 2009; Vinciarelli et al., 2015). Estudar os fenômenos do diálogo entre humanos é uma fonte valiosa de informações, mas não é suficiente para desenvolver aplicações. É preciso **buscar entender como a interação humano-máquina de fato ocorre**. Hayes (1980) argumenta, inclusive, que a meta de criar agentes que busquem **imitar** o comportamento humano sequer é um ideal, dada a capacidade de adaptação dos humanos ao interlocutor. Em vez disso, ele propõe que os modelos exibam habilidades que permitam uma "interação graciosa" com os humanos, que seja robusta a problemas de comunicação (Hayes; Reddy, 1983).

Munidas e munidos, agora, de uma compreensão mais abrangente do que é (ou pode ser) um diálogo, temos de nos debruçar sobre alguns dos fenômenos linguísticos que ocorrem quando pessoas conversam, já que a (im)possibilidade de processar esses fenômenos e de usar estratégias similares ou equivalentes pode afetar a performance de agentes conversacionais e nossa percepção deles.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

15.3 Fenômenos: O que ocorre em uma conversa?

Vamos agora examinar uma série de conceitos relativos a alguns dos principais fenômenos que ocorrem em diálogos. São manifestações e estratégias que muitas pessoas conseguem usar e compreender intuitivamente com maestria sem nunca terem pensado sistematicamente sobre como se dá esse processo. Prestar mais atenção a esses pormenores nos possibilita vislumbrar a riqueza linguística e cognitiva de uma conversa. Além disso, levar em conta a ampla pesquisa teórica e empírica sobre esses fenômenos, oriunda de campos como teoria do diálogo, ciência cognitiva e análise de conversação, faz toda diferença na hora de propor e avaliar modelos de conversação bem fundamentados.

Para entendermos melhor a função desses fenômenos, vamos emprestar a metáfora dos dois trilhos usada em (Clark, 1996a). Uma conversa consiste em dois trilhos paralelos. No primeiro deles, ocorrem os **atos comunicativos**, ou seja, os enunciados sobre os "negócios oficiais" da interação. No segundo, chamado de trilho colateral, ocorrem os **atos metacomunicativos**, em que se lida com os atos comunicativos do outro trilho de modo a manter a comunicação funcionando. Vamos voltar ao exemplo da Figura 15.1. Quando HLS diz que é louca para conhecer Olinda, que não conhece o norte (na verdade, nordeste) e tem vontade, ela está no primeiro trilho, contribuindo com o tema substancial da conversa. Já quando ela diz "meus pais não / mas meus avós", ela está, principalmente, corrigindo o mal entendido de MRN ter dito que seus pais são pernambucanos. Da mesma forma, quando MRN pronuncia "hum hum", faz um ato metacomunicativo para mostrar compreensão.

Antes de prosseguirmos, precisamos de mais uma definição. Em um diálogo, cada participante participa, em seu turno, com uma "**unidade de contribuição**", que pode ser desde apenas um fonema ou uma palavra até múltiplas frases em sequência, ou ainda fragmentos. Essa unidade é comumente chamada de **enunciado** (*utterance*), termo que usaremos para incluir também unidades equivalentes em língua de sinais ou em mensagem de texto.

15.3.1 Incrementalidade

Há vastas evidências científicas de que humanos processam linguagem de forma **incremental** (ver, por exemplo, (Altmann; Kamide, 1999; Altmann; Mirković, 2009; Crocker, 2010; Ferreira; Swets, 2002; Levelt, 1993; Marslen-Wilson, 1973), *inter alia*), ou seja, a compreensão da estrutura e do significado se constrói conforme as palavras vão sendo recebidas, e a produção também se dá passo-a-passo ao longo do tempo. Mais concretamente, nós não precisamos esperar chegar ao final de uma frase que estamos lendo ou ouvindo para só então começar a compreendê-la, tampouco já temos sempre uma frase completa formulada em nossa mente quando começamos a pronunciar suas primeiras palavras. Produção e compreensão da linguagem são processos que ocorrem **em tempo real**.

Dessa forma, em cenários interativos, o **tempo** se torna um componente central, de forma que a linguagem deixa de ser vista como um **produto** e passa a ser considerada uma **ação** ou um **processo** (Clark, 1992; Trueswell; Tanenhaus, 2005). Esse paradigma ocasiona diversos fenômenos metacomunicativos que não são bem capturados em *corpora* de textos escritos em monólogo, os quais veremos a seguir. Um exemplo é a possibilidade de terminar as frases da outra pessoa (Gregoromichelaki et al., 2011; Sidnell, 2012), como vemos na Figura 15.2. Aqui, CAU hesita ao pronunciar "são super", então CAV continua com a sugestão "receptivas", opção aceita por CAU no turno seguinte. Para pre-

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

dizer ou propor uma continuação antes do enunciado acabar, a ouvinte precisou já estar processando o enunciando ainda incompleto.

Figura 15.2: Transcrição de um trecho de diálogo em português brasileiro falado em conversa face a face, no qual CAV sugere uma continuação para a frase de CAU, evidenciando o processamento incremental.

CAU - todo o mundo te trata / muito bem //

CAU - assim / ~~voeê~~ ah / ~~voeê quer sa voeê quer saber o que esta aco~~ ah / as pessoas querem saber de você e ~~querem s~~ assim / querem saber ~~o~~ quem você é / o que é que você está acostumado a fazer lá / e tal / elas são super / eh /

CAV - receptivas //

CAU - receptivas / e / eu assim / achei superlegal //

CAU - então acho que eu aprendi muitas coisas por esse lado / não é //

Fonte: CRPC, arquivo Surpresas_da_Fotografia

Grande parte dos sistemas atuais em diversas tarefas de PLN é treinada em textos já existentes, e, portanto, podem acessar frases ou trechos por completo na hora de fazer predições. Em particular, dois dos melhores modelos atuais dependem, por definição, da disponibilidade do texto completo: as BiLSTMs (Schuster; Paliwal, 1997) constroem representações *forward* e *backward* da frase e os Transformers (Vaswani et al., 2017) são otimizados para processamento em paralelo, usando o mecanismo de atenção que acessa todo e qualquer *token* de entrada de uma vez (Madureira; Schlangen, 2020). Já no processamento em tempo real, os modelos precisam fazer essas predições com base em **prefixos** dos textos, conforme eles vão sendo construídos, sem saber o que virá em seguida (Köhn, 2018; Schlangen; Skantze, 2011). Quando o próximo trecho fica disponível, o modelo pode ter inclusive de revisar previsões passadas, propondo uma nova predição que leve em conta a última porção da entrada recebida, como se vê na Figura 15.3.

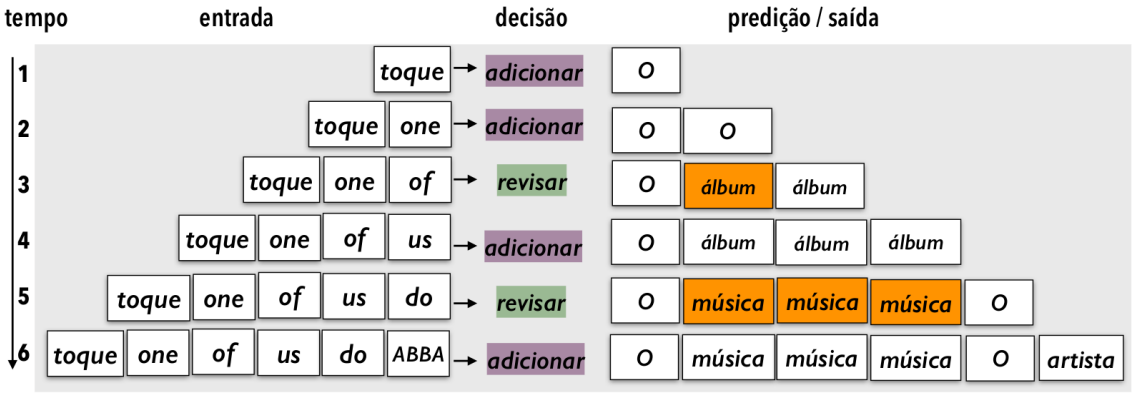
Vemos como é desafiador o PLN feito em tempo real. Então, por que tentar modelar o processamento incremental nos modelos de diálogo? Não é mais simples esperar o final da contribuição do interlocutor e trabalhar com material completo? De fato, tudo depende da aplicação. Como discutido por Schlangen; Skantze (2011), há vantagens em dois principais aspectos: **reatividade e naturalidade**. O quesito reatividade se manifesta em um ganho do ponto de vista de engenharia, já que, em uma conversa em tempo real, um agente que espera o interlocutor terminar de falar para só então começar a processar o conteúdo tem reações mais lentas, o que é perceptível e acaba causando silêncios longos e incômodos na fluência da conversa. Por mais que o processamento interno ocorra rapidamente, variações mínimas de silêncio já são passíveis de interpretações adicionais pelos humanos (Bruneau, 1973; Roberts et al., 2006; Wilson; Zimmerman, 1986). Se o processamento começa de imediato e vai sendo atualizado a cada nova palavra, quando o interlocutor chega ao final do enunciado o modelo já tem uma construção parcial de como vai agir em seguida, e a reação é, portanto, mais rápida. Já em questão de naturalidade, o modelo consegue lidar de forma mais genuína com os fenômenos que trataremos a seguir, como produzir e compreender sinais de feedback, construir declarações colaborativamente (como vimos no exemplo) e lidar melhor com interrupções e troca de turnos.

Em (Schlangen; Skantze, 2011), encontramos uma abrangente formalização de possíveis implementações do processamento incremental para modelos de diálogo, levando em conta modularidade, granularidade, relações entre entrada e saída e possibilidade de revisões.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Figura 15.3: Processamento incremental ocorrendo para a tarefa de reconhecimento de entidades nomeadas. Aqui um assistente virtual recebe um pedido de uma usuária, palavra a palavra, e precisa ir identificando que deve tocar uma determinada música. O modelo integra cada nova palavra recebida em sua hipótese, e a corrige quando necessário. Quando o processamento não é incremental, apenas o passo 6 ocorre, isto é, o modelo espera a conclusão da entrada para gerar a saída completa.



Fonte: Exemplo fictício, adaptado de (Madureira et al., 2023a)

15.3.2 Troca de Turnos e Estrutura Sequencial

Como vimos, conversas são compostas de enunciados de dois ou mais participantes. Mas **como esses enunciados se organizam sequencialmente**? Como cada participante decide quando falar e quando estar em silêncio? Basicamente, um diálogo é uma sequência de enunciados e de silêncios; os enunciados podem ser de uma só pessoa ou sobreposições de mais de uma pessoa ao mesmo tempo, enquanto os silêncios podem ocorrer durante o enunciado de uma pessoa (pausas) ou nos momentos entre troca de turnos dos falantes (vãos) (Heldner; Edlund, 2010).

Dadas as boas maneiras, intuitivamente podemos crer que, enquanto uma pessoa fala, a outra escuta, mas não é sempre isso que acontece: sobreposições são relativamente frequentes (Liesenfeld et al., 2023). Há diversas tentativas de formalização das dinâmicas do processo denominado **troca de turnos**, isto é, como os participantes da conversa decidem como manter seu turno e continuar falando, passá-lo deliberadamente para outra participante (por exemplo, ao fazer uma pergunta) ou tomar o turno por conta própria (Ruiter, 2019; Sacks et al., 1978). Há também a possibilidade de tomar o turno mesmo enquanto uma outra pessoa ainda está falando, por meio de interrupções (Bennett, 1978; Yang et al., 2011a).

As decisões de troca de turno se baseiam em **deixas** (*cues*) de várias fontes: verbais, prosódicas, de intonação, gestuais, de olhares e de respiração (Duncan, 1972; Skantze, 2021). Especialmente nos momentos em que falas parecem se aproximar de um fim, ou quando há momentos de silêncio, é comum haver sobreposições. Há também tentativas de formalizar mecanismos de resolução das sobreposições, isto é, como decidir quem continua falando quando elas ocorrem (Schegloff, 2000). Mostramos um exemplo de trocas de turno na Figura 15.4, na qual há pausas, sobreposição e interrupção de turno.

A sequência de turnos não ocorre de forma aleatória em uma conversa coerente. Schegloff; Sacks (1973) propuseram o conceito de **pares de adjacência** para classificar os

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Figura 15.4: Transcrição de um trecho de diálogo em português brasileiro falado em conversa face a face, no qual observamos diversas pausas e vãos. Além disso, CAV toma o turno antes de CAU ter acabado o primeiro enunciado, e o turno de DAC se sobrepõe ao de CAU no áudio.

CAU - e foi assim / uma viagem que eu aproveitei / muita coisa / porque +

CAV - para qual cidade você foi ?

CAU - eu fui para Porto Alegre / mas / eh / eu fiquei ~~em cida~~ hospedada / numa outra cidade vizinha / se não me engano / aí de Esteio / é perto de Canoas / é uma coisa assim //

CAU - e mas depois ~~eu peguei~~ junto com o pessoal ~~da~~ da Federal de São Carlos / nós fomos fazer os circuitos ~~das~~ eh / das vinícolas / não é / então / Bento Gonçalves / toda aquela região ~~de~~ eh / de vinícolas e tal //

CAU - ah / ~~a go~~

DAC - foi tomar vinho //

CAU - essa gozação / eu aguentei depois / porque quando eu cheguei aqui / eu tirei foto / para caramba / porque eu / tirava foto de tudo

Fonte: CRPC, arquivo Surpresas_da_Fotografia

tipos típicos de pares de turnos em um diálogo, por exemplo: pergunta e resposta, saudações, oferta e aceita ou recusa, etc. Cada uma dessas contribuições se relaciona com o que foi dito antes e influencia o que será dito depois, podendo ser expandida com adendos antecipatórios, intermediários ou posteriores (Stivers, 2013). Uma conversa não é feita apenas de pares: outras categorias de sequências existem, como narrativas (Stivers, 2013) e os chamados sub-diálogos (Larsson, 2017; Litman; Allen, 1987), isto é, subsequências coerentes de turnos inseridas dentro de sequências mais longas (como na Figura 15.5). Os atos de fala (Levinson, 2017; Sadock, 2006) e os mecanismos de atenção e intenção na estrutura do discurso (Grosz; Sidner, 1986) também estão presentes na organização da estrutura do diálogo.

Evidentemente, é desafiador para modelos reconhecerem, integrarem e produzirem tantos sinais e subjetividades. Os modelos atuais não dão conta de tudo, mas há algumas estratégias possíveis para tornar esse processo mais controlado, como imposição de turnos alternados, com limites claros de que o enunciado acabou (como na comunicação por rádio ou com os grandes modelos de linguagem atuais) e duração da pausa como sinal mais notório. Para mais detalhes e uma recente revisão de literatura sobre o estado da arte, ver (Skantze, 2021).

15.3.3 Disfluências e Reparos

Como vimos no Capítulo 2, a fala tem disfluências, sendo composta por diversos **fragmentos** que nem sempre formam frases em uma gramática idealizada, e sua transcrição difere bastante de textos editados. Como descrito por Shriberg (1994, 2001), há várias formas de disfluências e descontinuidades. Resumidamente, temos as seguintes, com alguns exemplos exibidos na Figura 15.6:

- **pausa preenchida**: sons como “hmm”, “éé́é”, por exemplo quando a pessoa está pensando, hesitando ou tentando ganhar tempo (ver (Clark; Tree, 2002) para alguns usos);
- **repetição**: uma palavra ou trecho repetido logo em seguida;

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Figura 15.5: Transcrição de um trecho de diálogo em português brasileiro falado em conversa face a face, no qual há um sub-diálogo. BAR decide, no meio da narrativa, confirmar se ela deve prosseguir, e o faz após obter permissão de BAS.

BAR - eu mesma levei um susto agora / esse fim de +

BAS -

BAR -

BAS - esse ~~fi~~ +

BAR - estou fugindo do meu problema de

BAS -

BAR - casa / não é ?

BAS - não / não /

BAS - não está não / pode falar à vontade //

BAR - nós levámos um susto / tremendo ~~essa~~ agora em Julho //

Fonte: CRPC, arquivo A_Fazenda

- **supressão**: uma palavra ou trecho que é proferido, mas abandonado;
- **substituição**: uma palavra ou trecho que é substituído por outros;
- **inserção**: uma palavra ou trecho adicionado a trecho anterior;
- **erro de articulação**: fonemas pronunciados incorretamente, geralmente por acidente.

Há mecanismos linguísticos de **reparo** para se lidar com as disfluências e com outros problemas de comunicação. As **correções** podem vir da própria pessoa que fala (*self-correction)*, imediatamente ou nos turnos subsequentes, ou como intervenção da pessoa que ouve (*other-correction*) e decide reparar um enunciado (Kitzinger, 2012; Schegloff et al., 1977). Um outro mecanismo essencial para ajustar mal entendidos, dúvidas ou erros (desde problemas acústicos a questões pragmáticas) é o **pedido de clarificação**, no qual o interlocutor demonstra que algo não ficou claro e pede esclarecimento (Purver, 2004). Na Figura 15.7, vemos um exemplo de pedido de clarificação, possivelmente por motivo acústico, seguido de um reparo em forma de repetição do que já foi dito, de forma mais bem articulada sonoramente.

Não devemos, todavia, considerar que esses fenômenos são ruídos que devem ser evitados no *design* de um modelo ou eliminados dos dados. São fenômenos naturais da produção e

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Figura 15.6: Transcrição de um trecho de diálogo em português brasileiro falado em conversa face a face, no qual observamos diversas disfluências. Há diversas pausas preenchidas (eh). No primeiro turno, há repetição de "fala sobre". Há também diversas substituições ao longo desse enunciado: "que tem" por "que cabe", "fazer" por "dar" e "ao" por "o". Há, ainda, um início de palavra "reso" que é abandonada e substituída por "saber resolver".

CAR - tem o costume / eh / que eles falam / a analogia / que / por exemplo / se tem uma lei que / eh / ~~fala sobre a / f~~ fala sobre uma coisa / mas não é bem isso ~~que tem~~ que cabe / ~~ao~~ o facto que ocorreu / certo / aí ele pode / fazer analogia dessa lei //

CAR - quer dizer / ~~fazer~~ dar um jeito nessa lei para ela / caber dentro / do que está acontecendo / sabe ?

CAR - então isso / quer dizer / Direito é difícil / não é / porque além de você ter que saber as leis / você vai ter que ~~reso~~ / saber resolver problema que / nem previsto é / nem nada //

MRL - tem que saber o número da lei e tudo ?

CAR - olha / os professores / eles sabem //

CAR - eles falam assim / " a lei seis mil e quinze " / que é a que mais usa / que é enorme / que tem quase tudo / é na lei seis mil e quinze //

CAR - mas / eh / eu acho que isso você acaba / eh / pegando com o tempo / não é //

Fonte: CRPC, arquivo O_Mundo_do_Direito

Figura 15.7: Transcrição de um trecho de diálogo em português brasileiro falado em conversa face a face, no qual há um pedido de clarificação feito por CAC e atendido no último turno.

GRS - então / que é que o senhor acha dos jovens / por exemplo /

CAC - acha de quê ?

GRS - de hoje em dia ?

GRS - dos jovens / de hoje em dia //

Fonte: CRPC, arquivo Criar_Filhos

compreensão de linguagem em tempo real pelos humanos, que vão fazer parte do diálogo e, inclusive, carregar significado. Como discutido por Ginzburg et al. (2014), disfluências exercem efeitos no discurso e se relacionam aos componentes da estrutura do diálogo. Embora em diálogos via texto algumas disfluências sejam corrigidas antes de a mensagem ser enviada (como erros de digitação que são percebidos na hora), ainda assim reparos são necessários (por exemplo, quando o corretor automático substitui uma palavra que não queremos dizer, ou quando não formulamos bem o conteúdo da mensagem e precisamos nos explicar).

Implementar formas de lidar com todos os tipos de disfluências e reparos é complexo, mas há algumas estratégias que podem ser utilizadas perante incertezas do ponto de vista do sistema, como pedir para o interlocutor repetir o que disse, confirmar se uma hipótese está correta ou propor alternativas e deixara usuária ou o usuário decidir qual é a desejada (ver, por exemplo, (Skantze, 2007)).

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

15.3.4 Construção de Base Comum e de Referências

Para uma conversa fazer sentido, deve haver coerência entre as contribuições de cada participante, que cooperam para atingir **compressão mútua** em busca de algum objetivo (Cervone et al., 2018; Perrault; Allen, 1978). Essa cooperação se manifesta em um processo denominado **embasamento** (*grounding)* com os participantes negociando significados e trabalhando juntos para dar e obter evidência positiva ou negativa de que enunciados foram compreendidos de forma satisfatória para o propósito da conversa (Benotti; Blackburn, 2021; Clark; Brennan, 1991). Mostramos um exemplo do embasamento acontecendo no diálogo da Figura 15.8.

Figura 15.8: Transcrição de um trecho de diálogo em português brasileiro falado em conversa face a face, no qual vemos o processo de embasamento. Antes de responder, TIN verifica se entendeu corretamente a dimensão temporal da pergunta ("hoje em dia"), o que é confirmado por CAL, para que ela possa prosseguir.

TIN - quando a gente começou a crescer um pouquinho / a gente foi / mudando //

CAL - mas / uma usa roupa da outra / ou não ?

CAL - a roupa dela é dela / a sua é sua ?

TIN - você está falando hoje em dia /

CAL - é hoje / é //

TIN - como está ?

TIN - hum / a minha irmã é o seguinte / não é / quando interessa para ela / ela troca alguma coisinha / ela empresta / tudo //

Fonte: CRPC, arquivo Muito_Iguais_e_Muito_Diferentes

A construção a cada turno da conversa edifica passo a passo a **base comum** (*common ground*) entre os participantes (Clark; Brennan, 1991). O conhecimento pode ser privado ou compartilhado entre os participantes (Ginzburg, 2012), e os significados vão sendo construídos **em conjunto**, coordenada e colaborativamente, ao longo da conversa (Clark, 1996a). Esse processo demanda que cada um "tome nota" mentalmente do que já foi compartilhado e do que ainda é confidencial (Lewis, 1979). Falando mais concretamente, se em um determinado ponto da conversa eu lhe conto uma novidade que você não sabia, ela passa, daí em diante, a fazer parte de um saber compartilhado entre nós. Se voltarmos a conversar daqui alguns dias, eu posso pressupor que você já sabe desse fato sem repeti-lo. Claro que, às vezes, algum lapso de memória ocorre, e os participantes precisam reconstruir a base juntos.

A base comum contém tanto conhecimento da experiência pessoal partilhada quanto conhecimento comunal, ou seja, normas sociais e aspectos culturais (Clark, 1996b). Sendo assim, torna-se relevante o conceito de teoria da mente (Brennan et al., 2010), pois cada participante precisa criar e levar em conta um modelo mental do outro (o que eu creio que minha interlocutora sabe ou não sabe) para ajustar o que vai dizer.

Dois fenômenos que emergem nesse processo são os sinais de feedback ou *backchannels*, termo cunhado por Yngve (1970), e a construção colaborativa de enunciados (Clark; Schaefer, 1987; Purver et al., 2009). Os sinais de feedback são emitidos por quem ouve, para demonstrar em tempo real que está compreendendo ou acompanhando o que está sendo dito, por exemplo, dizendo "aham" ou "é" durante um enunciado da outra pessoa. Vemos

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

alguns exemplos na Figura 15.9. Já a construção colaborativa ocorre quando um enunciado é construído conjuntamente pelos participantes, tentando chegar em um acordo quanto ao significado antes de prosseguir. Em particular, o que pode acontecer em parceria é a construção de referências, quando um participante inicia uma menção a alguma coisa e, a partir daí, com mecanismos de reparo, expansão e substituição, chegam juntos a uma referência de mútua compreensão (Clark; Wilkes-Gibbs, 1986; Heeman; Hirst, 1995).

Figura 15.9: Transcrição de um trecho de diálogo em português brasileiro falado em conversa face a face, no qual há diversos sinais de feedback emitidos por DAL enquanto CAN formula o contexto da pergunta.

CAN - dona Dália / mas nós estávamos falando de religião / não é ?

DAL - certo //

CAN - o brasileiro sempre diz que é católico / mas ele nunca pratica //

DAL - é /

CAN - é assim que você

DAL - exacto //

CAN - sente / também ?

DAL - é / eu sinto assim / não é //

Fonte: CRPC, arquivo O_Acidente

15.3.5 Sinais Multimodais

Durante uma conversa face a face, podemos usar gestos para apontar objetos, sinalizar direções ou demonstrar tamanhos. Mexemos a cabeça afirmativa ou negativamente. Nossa expressão facial pode exteriorizar emoções que se tornam perceptíveis ao interlocutor, como espanto, alegria ou irritação. Além disso, olhar para alguém também pode ser um meio de se comunicar, e às vezes é natural desviarmos um pouco o olhar enquanto estamos no meio de um enunciado. Ou seja, além dos enunciados linguísticos, muitas outras ações ocorrem durante um diálogo: **há dinâmicas dos olhares** (Bavelas et al., 2002; Beattie, 1978; Rossano, 2012), **dos gestos** (Bavelas et al., 1992; Wagner et al., 2014) **e das expressões faciais** (Bavelas; Gerwing, 2007; Chovil, 1991) em curso, em paralelo com a fala, e que também carregam ou adicionam significado. Como são dimensões além do processamento de linguagem natural, não trataremos de cada uma em detalhes aqui, mas queremos frisar que elas são também muito relevantes em interações em que os participantes podem se ver.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

15.3.6 Análises de diálogo em português

Ao longo desta seção, apresentamos definições de fenômenos que ocorrem em diálogos de maneira geral. Há trabalhos publicados no contexto da língua portuguesa, trazendo análises e também contribuições empíricas. Para facilitar a identificação, reunimos alguns trabalhos por tópico na Tabela 15.1. É impraticável listar todos os trabalhos existentes, de modo que esta seleção serve como ponto inicial para busca de outras contribuições.

Tabela 15.1: Seleção de trabalhos acadêmicos acerca de fenômenos do diálogo no contexto da língua portuguesa.

Tópico	**Trabalhos**
análise de conversação e fala-em-interação	Martins (1987), Barros (1999), Bulhões et al. (2006), Garcez (2006), Alencar (2008), Cunha Recuero (2008)
atos de fala e outras questões pragmáticas	Korsko (2004), Osborne (2010)
base comum e ação conjunta	Oushiro (2011), Hilgert (2012), Oushiro; Mendes (2012), Hilgert (2014), Kanitz; Frank (2014), Souza (2021)
continuações e feedback	Freschi (2017), Sousa et al. (2022)
diálogos em língua de sinais	Gomes et al. (2020)
inserções, repetições, reformulações	Fávero et al. (1998), Oliveira et al. (1998), Essenfelder; Rodrigues (2005)
interfaces conversacionais	Jacintho; Penha (2016)
jogos de linguagem	Nunes (2016)
marcadores conversacionais	Nunes (2017)
múltiplos participantes	Rocha et al. (2014)
narrativas e digressões	Koch (2012), Tesch (2015), Ferla (2020)
organização de turnos	Antonio (2003), Bernardo (2005), Reis; Silva (2013), Carvalho; Acioli (2017), Carapinha; Plag (2018)
perguntas e respostas, informações	Fávero et al. (1996), Konrad (2018)
referenciação	Marcuschi (2001)
reparo	Toscano (2001), Loder et al. (2002), Garcez; Loder (2005), Oliveira; Dias (2018)
sinais multimodais	Rodrigues (2003), Schröder (2015), Ostermann et al. (2016), Avelar; Ferrari (2017), Vogel (2018), Kanitz; Luz (2019)
sobreposições	Stein (2010), Marega; Jung (2011), Moraes Garcez; Stein (2015)
tópico, foco e ponto de vista	Bernardo (2001), Bernardo (2002), Bernardo (2003), Botelho (2011), Bernardo et al. (2021)
transcrição	Gago (2002), Pimentel (2016)

15.4 Modelos e Tarefas: Implementando um processo complexo

Fundamentamos bem até aqui quão **multifacetado e rebuscado** é o processo de um diálogo. Ao longo do texto, já fomos destacando alguns desafios de se criar sistemas que consigam "jogar o jogo" do diálogo de acordo com todas as suas estratégias. Vamos, agora, analisar esse tema mais tecnicamente, abordando dois aspectos: modelos e tarefas. Trataremos de modelos como tentativas de realizar computacionalmente o diálogo (ou algum de seus fenômenos), e tarefas, como formas de simplificar a representação desse processo de modo a aplicar, por exemplo, métodos de aprendizado de máquina, como se

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

tornou usual no atual paradigma de PLN.

15.4.1 Modelos

Podemos conceber dois principais objetivos de nos aventurarmos a modelar diálogo: definir modelos computacionais para fins de pesquisa, em busca de expandir a compreensão científica de como o uso interativo da linguagem funciona, ou como uma pura aplicação que funcione para determinado fim, na qual as técnicas de engenharia podem acabar se sobrepondo às teorias linguística e cognitiva. No segundo propósito, já existem diversos exemplos em funcionamento, como os *chatbots* em uso na indústria, que atendem a demandas concretas. Já quanto ao primeiro propósito, há muito a se desvendar.

Como vimos, o uso interativo da linguagem ocorre de forma situada, em um contexto. Schlangen (2023a) propôs alguns requisitos teóricos para agentes conversacionais, divididos entre demandas representacionais e em processos de suporte. Ou seja, por um lado, é necessário que o agente ideal tenha um modelo de linguagem, um modelo de representação do mundo, um modelo da situação em que está inserido, um modelo do discurso e um modelo de si próprio e dos outros. Por outro lado, esse conhecimento é aplicado e atualizado através de processos, tendo como pilares o processamento e o aprendizado incrementais, e o embasamento conversacional e multimodal.

Ele constata que, em PLN, há tentativas de implementar partes desse conjunto de forma isolada, mas ainda não um modelo completo; além disso, várias dessas tentativas são simplesmente focadas mais no método de aprendizagem de máquina que está em voga do que no fenômeno interativo em si. Em outras palavras, uma vez que temos, por exemplo, métodos de sequência-a-sequência ou de rotulagem de sequência funcionando bem em diversas áreas, acaba-se por tentar moldar o diálogo de forma a se adequar às representações exigidas por tais métodos, em vez de partir da compreensão do fenômeno para o método adequado. Isso é uma forma de viés cognitivo conhecida como lei do instrumento[6], e suas implicações são discutidas na literatura (Wagstaff, 2012).

Tendo o ideal teórico em mente, vamos apresentar algumas abordagens principais na construção de modelos de diálogo. Para sermos sucintos, vamos focar em sistemas mais voltados a aplicações práticas, tratando de seus objetivos, componentes e paradigmas. Dada a natureza introdutória deste capítulo, não introduziremos detalhes técnicos, os quais podem ser encontrados em fontes dedicadas à implementação de sistemas de diálogo. Para uma exposição mais detalhada sobre modelos de bate-papo, veja o Capítulo ChatGPT, MariTalk e outros agentes de conversação; todavia, note que produtos como o chatGPT não são modelos de diálogo propriamente ditos. Eles são apenas modelos de linguagem preditores de próximas palavras, que foram encaixados no uso para bate-papo, no que Skantze; Doğruöz (2023) descrevem como uma "solução" que achou seu problema (em vez do contrário).

Primeiramente, **para quê se utiliza um sistema de diálogo**? Há muitas possibilidades: atendimento ao consumidor, suporte técnico, busca de informações em base de dados, interface de usuário, assistentes virtuais em automóveis ou casas, entre outros. O objetivo em si, somado ao contexto onde vai ser utilizado, determina algumas dimensões principais (Jurafsky; Martin, 2023), a saber:

- **propósito**: o sistema pode ser orientado a tarefas, no qual o diálogo ocorre com um propósito bem definido (reservar um restaurante ou preencher um formulário),

[6] Também conhecida pela máxima "se você tem um martelo, tudo se parece um prego". Fonte: https://en.wikipedia.org/wiki/Law_of_the_instrument.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

ou buscar manter diálogos "abertos", ou seja, bate-papo, com conversas sobre temas genéricos mais voltadas a entretenimento;

- **dispositivo**: o sistema pode estar incorporado em um robô ou em um aparelho específico, ter uma representação virtual em forma de avatar ou ícone, ou, ainda, ser uma voz acessória saindo de um dispositivo que realiza outras atividades;
- **domínio**: o sistema pode ser tanto restrito a um único tópico (reserva de viagens) quanto multitópico (reserva de viagem e reserva de restaurante) ou genérico (*i.e.* qualquer tópico);
- **modalidade**: o sistema pode ter de lidar com texto e/ou com som e/ou com multimodalidade, que inclui imagens e vídeos;
- **iniciativa**: o sistema pode permitir interações nas quais apenas ele tem iniciativa (por exemplo, o sistema faz perguntas e interlocutores apenas respondem), ou apenas a usuária ou o usuário tem iniciativa (o sistema apenas responde a comandos) ou uma mistura dos dois.

Modelos orientados a tarefas só precisam atingir um objetivo, o que limita bastante o tipo de interação, facilitando tanto a modelagem quanto a avaliação. Já os modelos de bate-papo exigem maior versatilidade quanto aos rumos da conversa, embora ainda exista a possibilidade de restringi-los quanto a tópicos (*e.g.* um *chatbot* que saiba bater papo apenas sobre cinema).

Agora, **o que um sistema de diálogo precisa ter**? Vamos retomar os elementos básicos do formato elementar de uma conversa, expostos na Seção 15.2. Momento a momento, um participante precisa (i) captar o que está sendo dito, (ii) compreendê-lo, (iii) atualizar o estado da situação em sua mente, (iv) decidir o que fazer a seguir, se necessário buscando informações na memória, e, então, (v) gerar e (vi) proferir o próximo enunciado.

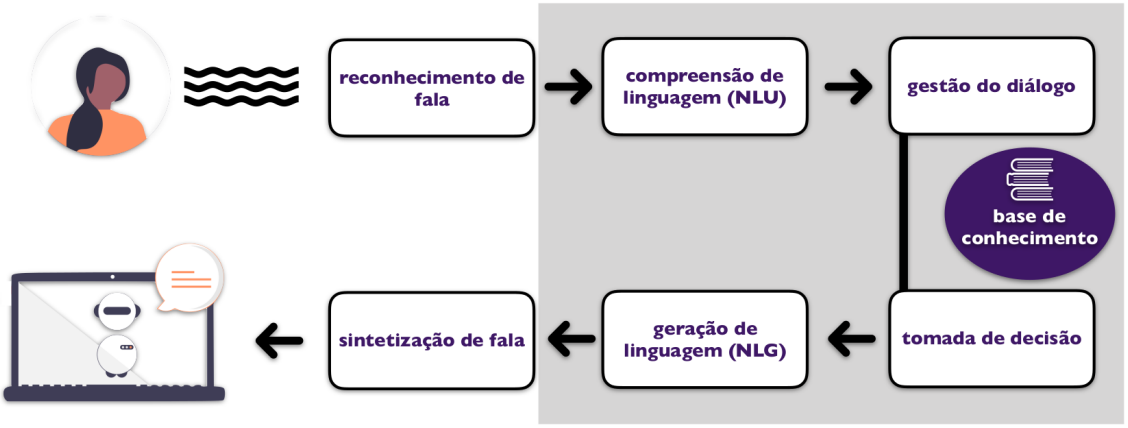
Os sistemas mais clássicos realizam esse processo de forma modular (e possivelmente seriada), em que cada componente se especializa em uma dessas tarefas (Williams et al., 2016). Primeiro, (i) há um sistema de reconhecimento de fala que capta o som e o converte para texto. Então, (ii) há um componente que executa a compreensão da linguagem e transforma a entrada em algum tipo de representação linguística formal. Depois, (iii) há um "gestor do diálogo" (*dialogue manager)* coordenando como ajustar seu estado interno com base nas informações recebidas, mantendo uma representação do histórico da conversa. Em seguida, (iv) há um módulo "tomador de decisão" que gere uma política do que fazer a seguir, possivelmente com acesso a uma base de dados. A representação dessa próxima ação é (v) convertida em texto por um componente de geração de linguagem natural, que então é (vi) transformado em som por um sintetizador. Na Figura 15.10, vemos uma ilustração dessa arquitetura. Em sistemas só de texto, o primeiro e o último componentes não são necessários, e em sistemas por língua de sinais ou multimodais, a captura da entrada e a sintetização da saída se dá incluindo imagens.

Atualmente, na era do aprendizado de máquina, a modularização tem sido substituída por tentativas de propôr arquiteturas de ponta a ponta (*end-to-end*) (Serban et al., 2016; Vinyals; Le, 2015), ou seja, modelos que recebem uma entrada em forma de texto, criam uma representação interna abstrata em forma de vetores numéricos, e geram uma saída em forma de texto. Isso reduz todo o processo do diálogo a uma única tarefa: dado um contexto, produzir o próximo enunciado, sem um componente explícito de gestão

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Figura 15.10: Principais componentes de um sistema de diálogo. Quando o sistema é baseado em texto, apenas a parte cinza é relevante.



Fonte: Adaptado de (Williams et al., 2016), com ilustrações de unDraw

do diálogo. Há inclusive tentativas de ir direto de som a som, sem usar o texto como mediador (Lakhotia et al., 2021). Arquiteturas de ponta a ponta têm limitações, de modo que há também alternativas híbridas, que incluem alguns componentes de representação linguística no processo.

Finalmente, **como operacionalizar o processo**? Temos dois principais paradigmas: **sistemas baseados em regras** (*rule-based)* e **sistemas baseados em dados** (*data-driven)* (Jurafsky; Martin, 2023). Nos sistemas baseados em regras, todo o fluxo da interação é definido por (muitas) diretrizes no estilo "se isso, então aquilo". Já nos sistemas baseados em dados, busca-se extrair padrões e generalizações de grandes *datasets* contendo uma vasta quantidade de exemplares de interações já realizadas no passado. Enquanto o primeiro parte do princípio de que seria possível formalizar os processos que governam o diálogo em forma de uma lista finita de regras, o segundo pressupõe que todos os fenômenos que discutimos poderiam ser implicitamente extraídos diretamente de grandes *datasets* por métodos estatísticos de aprendizado de máquina. Na prática, ambos têm limitações.

A geração dos enunciados também pode ocorrer com base em regras, ou seja, serem formados a partir de *templates* pré-definidos, bem como ser treinada a partir de dados por aprendizado de máquina ou, ainda, depender de técnicas de recuperação de informação, como as discutidas no Capítulo Recuperação de Informação e (Jurafsky; Martin, 2023).

Especialmente em sistemas orientados a tarefas, é comum definirem-se estruturas de representação (*frames*) com base na *expertise* de domínio, para orientar o **design** do modelo (Jurafsky; Martin, 2023). Tomemos como exemplo um agente que recebe e direciona pedidos de pizza. Aqui, o domínio é bem restrito (pedido de pizza), de forma que podemos definir as seguintes lacunas a serem preenchidas: `QUANTIDADE`, `SABORES`, `ENDEREÇO`, `CLIENTE`. Essas lacunas estão presentes em qualquer pedido, e a tarefa do agente é instanciar os valores de cada um para cada pedido, extraindo as informações da conversa com o cliente. Essa representação estruturada do pedido pode, então, ser passada para um dispositivo que o processe automaticamente.

Uma vantagem de sistemas baseados em regras é a interpretabilidade, pois sabe-se exatamente como cada passo da interação é processado pelo sistema. Já sistemas baseados em dados podem dar mais flexibilidade ao reduzir a necessidade de o processamento ter de passar por uma representação formal "imposta" por quem desenvolve o sistema, além

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

de conseguir extrair padrões que nem sempre estão patentes aos humanos; mas isso reduz o controle e a interpretação, o que pode ser arriscado.

Os métodos de aprendizado de máquina supervisionados costumam ser ancorados em *datasets* estáticos. Esse paradigma é limitante para diálogos: enquanto há muitas trajetórias válidas para uma conversa, só um exemplar é observado nos dados para cada caso. Por isso, há muitos sistemas de diálogo que se baseiam em processos de decisão de Markov, o que permite o uso de métodos de **aprendizado por reforço**, nos quais a interatividade tem papel central. Para mais detalhes, ver (Rieser; Lemon, 2011).

15.4.2 Tarefas

Para fazer tudo que um humano faz, modelos de diálogo podem precisar integrar várias áreas de PLN de uma vez. Minimamente, é preciso reconhecimento de fala, compreensão de linguagem natural (que em si já envolve diversas "tarefas"), produção de linguagem natural (mais muitas "tarefas") e síntese de fala. Cada um desses componentes é uma área de pesquisa em si, e integrá-los em um único sistema é, portanto, uma árdua empreitada.

Na verdade, a versatilidade dos humanos ainda é impossível de modelar por completo de uma só vez. O que ocorre, então, é a definição de sub-tarefas que tentam capturar um ou alguns aspectos do diálogo de cada vez. Algumas tarefas são bem motivadas pela teoria do diálogo, como predizer quando um pedido de clarificação é necessário, identificar o que é uma disfluência, atualizar informações que foram corrigidas, ou decidir quando é um momento oportuno para emitir um sinal de feedback. Já outras são concebidas de forma a fazer o fenômeno se subjugar às ferramentas de aprendizado de máquina disponíveis (usando por exemplo um paradigma de classificação), o que nem sempre é ideal, por ofuscar a motivação científica e a justificação teórica para o funcionamento do sistema.

Vamos listar algumas das tarefas mais conhecidas. A lista não é exaustiva, já que há muitas possibilidades e novas tarefas podem ser definidas a depender dos dados disponíveis. Incluímos os nomes em inglês para facilitar uma eventual busca por mais conteúdo. Na Figura 15.11, mostramos um exemplo ilustrativo de como o gerenciamento do diálogo pode ser feito com base em lacunas e intentos.

- **Gerenciamento de diálogo** (*dialogue management)*: monitoramento, controle e tomada de decisões sobre todo o fluxo da conversa, com coordenação de todos os componentes do sistema.
- **Monitoramento do estado do diálogo** (*dialogue state tracking)*: controle do atual estado da conversa através de uma representação interna, por exemplo, quais informações já estão disponíveis ou não para um determinado fim; comum em cenários em que está pré-estabelecido quais requisitos devem ser atendidos ao longo da conversa, como em uma reserva de restaurante.
- **Classificação de atos do diálogo** (*dialogue act classification*): enquadramento de um enunciado em uma taxonomia de unidades da conversa, isto é, identificar o que é, por exemplo, pergunta, afirmação, pedido ou promessa.
- **Monitoramento de conjecturas** (*belief state tracking*): estimação da certeza de hipóteses construídas pelo sistema ao longo da conversa.
- **Detecção de final de turno** (*end-of-turn detection*): identificação de quando um enunciado está prestes a acabar para o sistema tentar tomar o turno, se for o caso.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

- **Detecção de disfluências** (*disfluency detection*): identificação das partes do enunciado que são disfluências e, mais especificamente, como atuar perante elas (ou seja, ignorar repetições mas acrescentar ou atualizar o que foi inserido ou substituído).
- **Segmentação de enunciados** (*utterance segmentation*): decisão de como quebrar um enunciado, possivelmente longo, em sub-componentes. Por exemplo, quando uma pessoa dá uma série de instruções de uma vez, o sistema precisa definir cada passo.
- **Diarização de locutor** (*speaker diarization)*: identificação de qual participante está falando a cada momento.
- **Detecção de intenção** (*intent detection*): identificação do propósito de um enunciado; por exemplo, um assistente virtual que identifica que a usuária deseja ouvir música.
- **Preenchimento de lacunas** (*slot filling*): identificação de valores que se aplicam a *slots* pré-definidos em um contexto. Novamente, um assistente virtual que detectou que a usuária deseja ouvir uma música deve extrair o nome da música e/ou da artista no enunciado.
- **Geração do próximo enunciado** (*next turn generation*): produção de uma fala ou mensagem a ser proferida a seguir com base na atual representação interna do sistema.
- **Resposta a perguntas** (*question answering)*: como responder apropriadamente a uma pergunta. Esta é uma tarefa de PLN em si (Capítulo 14), que pode ocorrer em sistemas de diálogo. Embora algumas fontes considerem que uma sequência fixa de perguntas e respostas forma um diálogo, isso é controverso.

15.5 Avaliação: Verificando a qualidade de um sistema

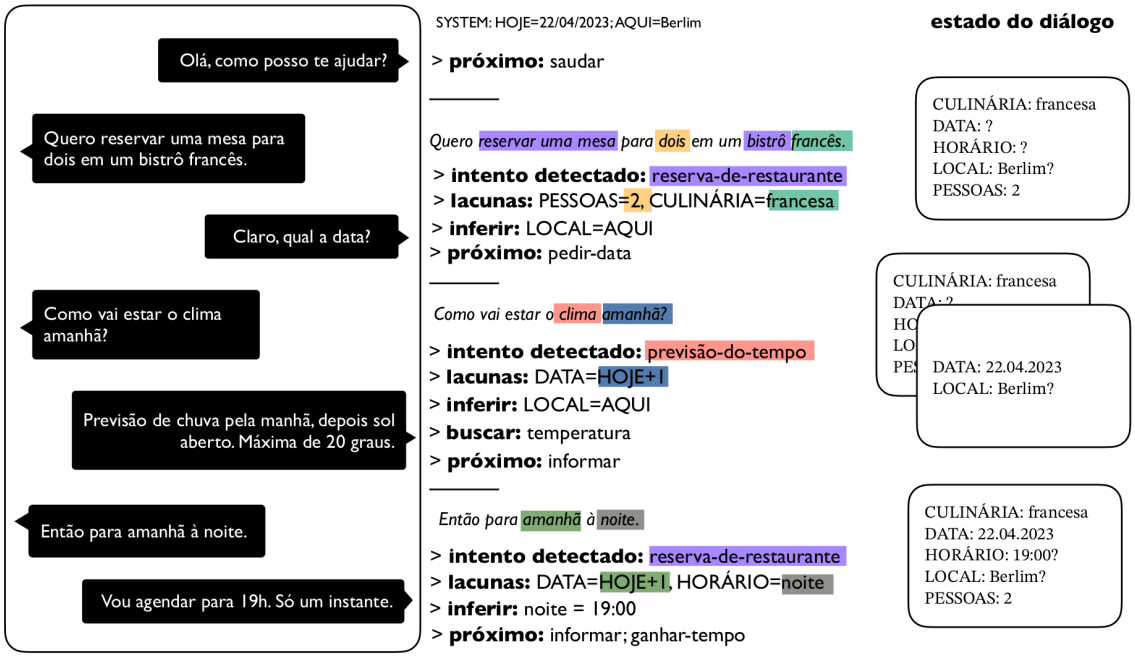
Chegou a hora de nos ocuparmos de uma incumbência essencial: avaliar os modelos de diálogo. Como podemos mensurar quão bem um modelo participou de uma conversa? Quais aspectos devem ser levados em conta? Como medir características subjetivas como qualidade, efetividade e engajamento? Que impactos e influência esse sistema pode ter sobre usuários e usuárias e como ele pode afetar a realidade?

A avaliação de sistemas interativos exige uma análise intrincada. A cada turno, temos, por um lado, o tópico, o propósito e o contexto limitando as possibilidades de pertinência da próxima contribuição, mas, por outro lado, temos também uma infinidade de enunciados que são equivalentemente apropriados para dar continuidade. Por exemplo, se eu lhe pergunto que livro você está lendo, não seria muito conexo responder que ontem você viu um raio de sol pela janela ou que o preço do pacote de feijão subiu; há uma expectativa de que o próximo turno seja ou uma resposta direta à pergunta ou algo relacionado a ela. Mesmo assim, há diversas continuações válidas: “um livro muito interessante sobre PLN”, “um livro que minha professora me indicou”, “vários”, “você já esqueceu?”, “é segredo” etc. Algumas podem ser menos apropriadas que outras, mas não seriam necessariamente “erradas”.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Figura 15.11: Exemplo ilustrativo de como se dá o gerenciamento do diálogo em um sistema orientado a tarefas. A cada entrada vinda da usuária ou do usuário, o sistema detecta qual o intento principal. Com isso, ativa uma representação que contém as lacunas necessárias para realizar esse intento. Os valores para preencher as lacunas são extraídos da mensagem ou inferidos com base na situação. Com base nas lacunas faltantes e consulta à base de dados pertinente, o sistema decide o próximo passo e gera a realização linguística da próxima mensagem. A cada passo, as informações mais recentes são acrescentadas ao estado do diálogo.



Fonte: Ilustração de elaboração própria, inspirada em (Coucke et al., 2018; Williams et al., 2016)

A avaliação de modelos orientados a tarefas é um pouco mais fácil, pois tem-se um objetivo claro a ser atingido, de modo que podemos definir métricas de sucesso (isto é, quanto se atingiu desse objetivo) como *proxy* (isto é, um "representante") para a efetividade da conversa. Já em modelos de bate-papo, não há algo tão palpável. Devido à multiplicidade de "respostas certas", não basta ter um único *gold standard* para se comparar. Qualquer *corpus* vai conter apenas uma entre as infinitas amostras do que a conversa poderia ter sido, ou seja, aquela que foi realizada pelos participantes no momento da coleta. E se o participante tivesse mudado uma frase? É possível que a conversa teria tomado um rumo totalmente distinto, mas nós nunca saberemos.

A despeito das dificuldades, já há bastante literatura a esse repeito para nos direcionar. Dois principais marcos históricos foram as iniciativas PARADISE (Walker et al., 1997) e Trindi (Bos et al., 1999). A primeira é uma proposta de sistematização da avaliação de agentes conversacionais, definindo uma hierarquia que leva em consideração a satisfação dos usuários e usuárias, o sucesso da tarefa e a minimização de custos, com medidas de eficiência (tempo da conversa, número de enunciados) e de qualidade (demora na reação

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

do sistema, taxa de reparos, etc.). Já na segunda, elabora-se uma *checklist* com critérios de comportamento do sistema que uma avaliadora ou um avaliador deve verificar; por exemplo, se o sistema lida com excesso ou falta de informação em um enunciado, se ele se adapta a falhas de comunicação, se a interpretação de um enunciado leva o contexto em conta, etc.

Como tratado em detalhe no Capítulo 18, pode-se optar por avaliação automática, através de métricas, ou por humanos. Idealmente, deve-se fazer ambas em conjunto, mas, como a avaliação por humanos é custosa e demorada, é importante termos também boas métricas para serem usadas de forma automática, deixando a avaliação por humanos para etapas mais cruciais. Notemos, todavia, que a avaliação automática é limitada, pois nem todos os aspectos de interesse podem ser bem formulados matematicamente. Por ser um tópico muito abrangente, há várias iniciativas recentes de revisão de literatura na área de diálogo que concentram o entendimento atual sobre o tema ou tentam propor uma unificação (ver, por exemplo (Braggaar et al., 2023; Deriu et al., 2021; Finch; Choi, 2020; Liu et al., 2016; Yeh et al., 2021)).

Em protocolos de avaliação automática, busca-se uma avaliação bem sistemática e objetiva, através do uso de (se necessário, múltiplas) métricas que operacionalizem as dimensões desejadas (Finch; Choi, 2020). Nesse caso, o processo deve ser repetível, focado e explicável, e as métricas devem preferencialmente ser bem correlacionadas com os julgamentos humanos (Deriu et al., 2021). Uma longa lista de métricas em uso foi compilada por Yeh et al. (2021), e novas métricas são propostas constantemente pela comunidade, pois cada aplicação pode precisar mensurar diferentes características do diálogo e da tarefa em questão. Algumas mais convencionais capturam similaridade entre o enunciado produzido pelo modelo e o que está realizado nos dados, coerência com o contexto, diversidade (entropia e inércia) e perplexidade do modelo de linguagem (Finch; Choi, 2020; Liu et al., 2016).

Na avaliação por humanos, quem deve avaliar o modelo? A responsabilidade de garantir uma boa avaliação é, inicialmente, da pessoa física ou jurídica que o desenvolve, todavia quem desenvolve o modelo não consegue ter uma abordagem neutra, devido a possíveis conflitos de interesse. Claro que durante o desenvolvimento avaliações parciais são feitas pela equipe desenvolvedora, mas no final e em alguns pontos cruciais é preciso uma avaliação imparcial. Algumas formas de fazer isso é através de experimentos, ou seja, deixando usuárias e usuários interagirem com o sistema e avaliá-lo, o que pode ocorrer trazendo pessoas ao laboratório, definindo tarefas de *crowdsourcing* ou fazendo experimentos de campo, em que o sistema é inserido em uma situação real; pode-se, ainda, simular o comportamento de um humano (deixando um modelo interagir com outro ou usando técnicas de *self-play)* (Deriu et al., 2021), mas isso se torna um outro problema[7].

Há duas formas de se efetuar avaliação humana: interativa, na qual a pessoa avalia uma conversa que ela mesma teve com um sistema (o que é mais genuíno), ou estática, quando ela lê ou ouve uma interação do modelo com outra pessoa e julga sua qualidade (Finch; Choi, 2020). Pode-se, ainda, avaliar apenas um sistema de forma isolada ou capturar a preferência da avaliadora ou do avaliador ao julgá-lo em comparação a um outro sistema. Nesses cenários, é comum o uso de formulários de satisfação que são apresentados aos indivíduos após a interação com o modelo, para dar notas ou feedback sobre a experiência. Algumas das dimensões possíveis estão listadas na Figura 15.12.

[7] Se já é difícil avaliar um modelo, como modelar e avaliar uma simulação de um humano interagindo com esse modelo?

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Figura 15.12: Algumas dimensões subjetivas de um diálogo que podem ser integradas na avaliação humana

acurácia	emoção	legibilidade
adequação	engajamento	lógica
coerência	especificidade	proatividade
coerência com o contexto	fluência	qualidade
consistência	gramaticalidade	relevância
diversidade	humanidade	sensatez
empatia	informatividade	uso do conhecimento

Fonte: (Finch; Choi, 2020)

Considerações éticas

Embora não seja uma tarefa simples, avaliar bem e de forma responsável é essencial. As aplicações do diálogo, como os *chatbots* e os assistentes virtuais, já foram inseridas em nossa realidade (muitas vezes de forma frenética e precipitada) e, portanto, atuam sobre ela e a transformam como qualquer agente. Por isso, é imprescindível ter meios de verificar esses efeitos, conhecer as limitações e os possíveis riscos do sistema, e ser transparente quanto a isso. Dessa forma, avaliação é muito mais do que obter um número que indica uma performance de sucesso. Ela deve ocorrer de forma holística ao longo de todo o processo de criação e inserção de um modelo, envolvendo escuta plena dos grupos sociais afetados por ele. Principalmente, deve estar pautada na ética, considerando todos os aspectos mais amplos de custo energético, fator humano, grupos vulneráveis, impacto ambiental e social, entre outros (Bender et al., 2021). Como qualquer tecnologia, aplicações desta área podem ser usadas para fins escusos e as considerações éticas do capítulo Capítulo 20 se aplicam aqui. Além disso, há reflexões éticas específicas de agentes conversacionais, tratando de vieses implícitos adquiridos através dos dados, uso de exemplos adversariais, violações de privacidade, dilemas de segurança, barreiras à reprodutibilidade dos resultados, perpetuação de injustiças, inclusão, transparência e antropomorfismo que têm de ser levadas em conta (Henderson et al., 2018; Murtarelli et al., 2021; Ruane et al., 2019).

15.6 Recursos: Trabalhando com dados e sistemas interativos

Modelar ou avaliar agentes que dialogam é tarefa que não precisa nem deve começar do zero. Já existem abundantes recursos disponíveis, tanto literatura, documentação e tutoriais quanto *corpora*, modelos pré-treinados, ferramentas e plataformas de desenvolvimento. Infelizmente, o problema da hegemonia da língua inglesa também afeta a área de diálogo, particularmente na disponibilidade de dados. Por outro lado, podemos ver isso como oportunidade de valorizar mais a língua portuguesa, buscando preencher esta lacuna de forma consciente e responsável.

15.6.1 Atuando de Forma Responsável

Uma ressalva importante antes de prosseguirmos: a facilidade com que se criam e acessam tantos recursos e ferramentas não deve ofuscar o alto risco envolvido em usá-los na realidade. O diálogo é uma **atividade social**, onde ética e confiança são valores

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

centrais, e seus participantes têm direitos e obrigações (Allwood et al., 2000). Além disso, o uso da linguagem segue processos de natureza normativa ancorados na realidade, sendo ativamente construído para ser **coletivamente útil** (Schlangen, 2022). **Se um agente artificial vai tomar parte nessa atividade, isso não deve ser feito com leviandade**.

Há áreas extremamente delicadas (saúde mental, aconselhamento médico, populações vulneráveis) nas quais só se deve atuar se houver uma equipe multidisciplinar envolvida, com pessoas treinadas, capacitadas e bem informadas para tomar decisões baseadas em códigos de ética e valores morais, e com base em estudos meticulosos da real necessidade e do impacto de uma aplicação de diálogo em tal contexto. Há, ainda, formas perigosas de usar tecnologia (uso militar, ameaças, golpes, vigilância, perseguição) que lamentavelmente continuam a ser propostas, de modo que precisamos estar atentos para saber identificá-las. Muitas fontes propõem, com naturalidade ou suposta "neutralidade científica", técnicas que podem ser facilmente prejudiciais e trazer alto potencial nocivo (como simulacros de "empatia" em máquinas que não têm sentimentos, técnicas de convencimento e dissimulação, discriminação, geração de conteúdo linguístico tóxico na internet, etc.). O uso enganoso, que induz usuários e usuárias a acreditar que o agente é uma pessoa, também é negativo e contraproducente (McGuire et al., 2023).

Trataremos mais dessas questões na conclusão, mas enfatizamos o alerta: **tecnologias de linguagem têm impacto em nossa existência**. Ao usar ou construir qualquer recurso, **busquemos sempre nos formar e informar além da técnica**, e ter garantias suficientes de que um sistema não será maléfico **antes** de desenvolvê-lo ou usá-lo (e não depois de já estar disponível no mercado).

15.6.2 *Corpora*

Na introdução, aludimos brevemente à inviabilidade de se modelar diálogo usando textos em monólogo, devido à natureza estática e confeccionada desse gênero. Mesmo *corpora* oriundos de redes sociais, que de fato contêm formas mais orgânicas da linguagem em uso, ainda coletam amostras de mensagens cujos usuários ou usuárias tiveram oportunidade de formular e editar o que queriam expressar antes de postar. Precisamos de dados que tenham sido gerados de forma interativa, com sua estrutura bem preservada durante a coleta. Quanto mais próximas forem essas interações do uso-alvo de uma tecnologia, melhor, pois os fenômenos podem variar em distribuição e frequência conforme o contexto.

A primeira dimensão importante a se considerar é se desejamos **diálogos entre humanos ou entre humano e máquina**, pois a forma de interação pode divergir em cada caso, como mencionamos na Seção 15.2. Como exposto no Capítulo 16, **tamanho e a qualidade** do *corpus* também são relevantes. Para pesquisas de fins mais qualitativos, pode ser suficiente um número mais baixo de exemplares. Muitos *corpora* foram coletados para fins de estudos linguísticos ou de documentação, sendo muito bem curados e anotados, mas relativamente pequenos. Já os métodos de aprendizado de máquina, famintos por dados, exigem quantidades copiosas de exemplares. Isso nem sempre está disponível e, quanto está, é raro a qualidade ser mantida, pois, devido ao tamanho, fica até difícil verificar o que o *corpus* contém (Paullada et al., 2021). Isso exige atenção redobrada ao utilizá-los, mesmo que sejam *datasets* populares e vastamente empregados.

Na área de diálogo, há ainda uma dimensão muito relevante: **autenticidade da interação**. Em alguns casos, os dados dos quais precisamos podem ser coletados em situações reais ou em circunstâncias que já ocorreram e ficaram registradas. Por exemplo, há diá-

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

logos em vídeos da internet, históricos de mensagens de texto, ou gravados por um robô em um evento. Uma empresa de celular possivelmente já terá um grande *dataset* de gravações de ligações de clientes, de forma que poderá explorar a estrutura desses diálogos já existentes para construir um assistente virtual. Há também muitos *corpora* de uso aberto ou de licença permissiva que podemos e devemos examinar para o nosso uso desejado, de modo a extrair o máximo de valor de um recurso já constituído.

Já quando precisamos de um *corpus* de diálogos em uma situação bem específica ou controlada, geralmente é preciso gerá-los, pois é algo que nem sempre está já disponível na realidade. Por exemplo, para modelar como as pessoas usam pedidos de clarificação, podemos definir um experimento no qual elas vão montar um quebra-cabeça juntas, para limitar um pouco a variabilidade das expressões e expor todos participantes ao mesmo tipo de tarefa.

Vamos começar com duas formas de criar dados conversacionais: através de experimentos, trazendo participantes a um laboratório, ou usando plataformas de *crowdworking*, o que se tornou bem comum na atualidade (mas não deixa de envolver questões éticas, ver o Capítulo 18). A vantagem desses métodos é ter controle sobre o tipo de interação, os equipamentos de captura de texto, som e imagem e a tarefa que os participantes realizam. Além disso, embora possa haver variações que se distanciem da realidade (como em qualquer experimento controlado), até certo ponto o diálogo é genuíno, ou seja, é gerado por humanos, de forma espontânea, mantendo suas estratégias internas autênticas. Mas há um entrave: para coletar dados de conversas face a face, é preciso sincronizar participantes para estarem no mesmo local ao mesmo tempo, ou, pelo menos, conectados simultaneamente. Ainda que estejamos interessados em interações via texto, é preciso que elas ocorram em tempo real. Isso implica a necessidade de coordenar horários, o que se torna uma tarefa adicional. Coletar dados de diálogo humano-máquina é mais fácil, pois um dos participantes é um programa que pode ser rodado quantas vezes for preciso e até mesmo simultaneamente em conversas diferentes, além de trazer garantias de que os participantes sempre serão expostos ao mesmo estímulo.

Há alguns procedimentos para contornar essa questão, que usam de subterfúgios para emular diálogos ou abrem mão de algumas das características discutidas na Seção 15.2. Há recursos compostos de dados assincrônicos, como conversas em fóruns da internet ou redes sociais. Há iniciativas que usam mão de obra humana para construir conversas artificialmente, pedindo a anotadores que imaginem o que falariam ou fariam em um determinado ponto de uma conversa da qual eles não participaram. Pode-se também simular diálogos de forma sintética, com *templates* e estruturas pré-definidas ou usando modelos de linguagem. Existem, ainda, *corpora* de conversas em que os participantes encenam um papel, como em jogos, ou que seguem um *script*, como legendas de filmes. A criatividade vai longe nessa área. Devemos estar cientes de que tais dados não são diálogos propriamente ditos, embora possam ser úteis para algumas aplicações específicas.

Há fenômenos que são raros em um diálogo típico, mas às vezes é exatamente eles que queremos estudar ou modelar. Por exemplo, para modelar uma estratégia para se lidar com reparos por inserção, precisamos de muitas observações que contenham reparos por inserção. Devido à sua relativa raridade, precisaríamos de imensos *datasets*, que fatalmente conteriam muitos diálogos sem qualquer reparo por inserção[8]. Tentar produzi-los artificialmente de forma explícita acaba resultando em dados menos autênticos. Há formas mais sutis de induzi-los implicitamente a acontecer.

[8]Esse dilema é conhecido como esparsidade de dados (*data sparcity*), usual em fenômenos linguísticos.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Primeiramente, temos os **jogos de diálogo**, que são atividades curtas e bem delineadas, com um objetivo claro a se atingir em conjunto (Schlangen, 2023b). O jogo é definido de forma a fazer emergir certo fenômeno sem que os participantes se deem conta disso, pois seu foco se mantém em realizar a tarefa. Além disso, há experimentos que usam a técnica *man-in-the-middle*, que manipula propositalmente algumas mensagens da conversa (Healey; Mills, 2009), ou o *Wizard of Oz*[9], no qual o participante pensa estar falando com um programa de computador mas, na verdade, ele é controlado em tempo real por um humano[10].

Quando os diálogos se dão por voz ou sinais, é preciso primeiramente transcrevê-los se quisermos aplicar os métodos usuais de PLN. Transcrever dados de diálogo é difícil, pois é preciso representar pausas, sobreposições, pausas preenchidas, e todos os demais fenômenos da oralidade (Heeman, 1995; Kowal; O'Connell, 2014). Além disso, se quisermos uma representação mais rica, é preciso representar também gestos, olhares, emoções e entonação, tudo sincronizado com a transcrição das falas. Fora a transcrição, às vezes nos interessamos por algum fenômeno específico que precisa ser identificado nos dados através de camadas de anotações linguísticas. Esses procedimentos, além de levar tempo, envolvem quase inevitavelmente um grau de interpretação dos anotadores e dependem dos esquemas de anotação e transcrição que por vezes não dão conta de todos os fenômenos ou de casos ambíguos (Basile et al., 2021). Não é razão para não fazê-lo, devemos apenas estar bem informados sobre as consequências daquilo que anotamos ou deixamos de anotar.

Há numerosos *corpora* disponíveis hoje em dia. Listá-los seria difícil e a lista ficaria defasada em pouco tempo. Há alguns trabalhos de revisão de literatura recentes que trazem compilações (ver (Gonçalo Oliveira et al., 2022; Mahajan; Shaikh, 2021; Serban et al., 2018; Sundar; Heck, 2022; Thakkar; Pise, 2019)), e a plataforma ParlAI[11] também disponibiliza diversos deles com uma interface comum, servindo como ponto inicial. Porém, uma boa estratégia é sempre usar ferramentas de busca com referências a trabalhos acadêmicos para encontrar os *datasets* existentes que sirvam ao uso em questão, lembrando de checar se a licença permite o uso desejado.

15.6.2.1 *Corpora* de diálogos em português

Vamos examinar em mais detalhes algumas fontes de dados de diálogo que estão disponíveis para a língua portuguesa. A Tabela 15.2 apresenta uma seleção de *corpora*, dando preferência aos que estão acessíveis (há mais trabalhos, mas nem todos disponibilizam os dados de forma aberta), com uma breve descrição. Há também uma lista (um pouco antiga) disponibilizada pela Linguateca[12].

[9] https://en.wikipedia.org/wiki/Wizard_of_Oz_experiment
[10] Obviamente, nesses casos os participantes deve ser informados sobre isso após o experimento.
[11] https://parl.ai/about/
[12] https://www.linguateca.pt/corpora_info.html

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

Tabela 15.2: Seleção de *corpora* com diálogos (ou aproximações) de diversos tipos no contexto da língua portuguesa.

	Trabalho	**Descrição**	**Tamanho**
CORAA	Candido Junior et al. (2022)	*corpus* para reconhecimento de fala, contém porções de diálogo em português brasileiro falado e transcrito	290,77 h (total)
CORAL	Trancoso et al. (1998)	diálogos orientados a tarefas, falados e etiquetados, em português europeu	64 diálogos
C-ORAL-ROM	Cresti et al. (2004)	documentação da forma oral de diversas línguas românicas, contém diálogos em português	300 mil palavras
C-ORAL-BRASIL	Raso et al. (2015)	coletânea de diversos *corpora* de fala espontânea em português brasileiro e indígena, contém diálogos	139 itens
CORP-ORAL	Santos; Freitas (2008)	conversas face a face em português europeu	50 horas
CRPC	Nascimento; Gonçalves (1996)	contém áudio e transcrição de diálogos em várias variedades do português	milhões de *tokens*
FakeWhatsApp.Br	Cabral et al. (2021)	*dataset* de mensagens de texto instantâneas para fins de detecção de desinformação	5.284 mensagens
Iboruna (ALIP)	Gonçalves (2019)	entrevistas sociolinguísticas e interações dialógicas em português brasileiro do interior paulista	163 diálogos
NURC	Oliviera Jr. (2016)	contém entrevistas e diálogos, com áudio e transcrição, em português brasileiro	279 h (total)
PALMA	Hagemeijer et al. (2022)	entrevistas semi-estruturadas, português em variações africanas	108 h
Pirá	Paschoal et al. (2021)	*dataset* de perguntas e respostas sobre tema ambiental	2.258 pares
SP2010	Mendes; Oushiro (2012)	conversas em português brasileiro, variação paulistana	60 gravações
–	Forte Martins et al. (2021)	*dataset* de mensagens de texto instantâneas sobre COVID-19	1.390
–	Sanches et al. (2022)	traduções de *dataset* com diálogos orientados a tarefas e coletânea de *posts* em fóruns da internet	milhões

15.6.3 Ferramentas

Assim como os recursos de dados, é infrutífero tentar listar as muitas ferramentas disponíveis. Há diversas novas ferramentas sempre aparecendo, mas muitas delas são efêmeras, pois deixam de ser atualizadas e se tornam defasadas. Além disso, muitas são soluções comerciais, e não temos intento de promover nenhuma em particular. Desta forma, tam-

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

bém optamos por não tentar indicá-las. Em vez disso, vamos mencionar que tipos de ferramentas podemos procurar para atuar com modelos de diálogo:

- **coleta de dados**: plataformas que permitam reunir dois um mais participantes com alguma interface, possibilitando (se for o caso) lhes apresentar conteúdo como imagens, e mantenha registros bem detalhados dos dados e metadados durante sua interação;
- **representação de dados**: ferramentas que facilitem a criação de estrutura pertinente para os dados, de modo a representá-los de forma fidedigna à interação original mas em formato que facilite o processamento computacional;
- **transcrição de dados**: sistemas que permitam transformar voz em texto, mantendo uma representação dos fenômenos do diálogo e da voz;
- **anotação de dados**: programas que viabilizem identificar, anotar e armazenar metadados linguísticos de um diálogo;
- **implementação e autoria**: *software* que propicie o desenvolvimento de agentes artificiais (*authoring tools*);
- **avaliação**: soluções que favoreçam a análise quantitativa e qualitativa do comportamento e performance de um modelo.

Modelos de diálogo para o português

Além de soluções comerciais (Capítulo ChatGPT, MariTalk e outros agentes de conversação), há uma série de trabalhos acadêmicos que já se ocuparam de esquematizar sistemas que interajam em língua portuguesa ou com alguma tarefa de diálogo em português. Agrupamos uma amostra deles na Tabela 15.3.

Tabela 15.3: Seleção de trabalhos sobre modelos de diálogo e *chatbots* no contexto da língua portuguesa.

	Trabalho
agente conversacional pedagógico da área de programação	Mattos et al. (2022)
agente conversacional universitário	Cruz et al. (2020)
agente recuperador de informações jurídicas	Quaresma; Rodrigues (2003)
Ambrósio: assistente de casa	Neto et al. (2006)
ANA: *chatbot* para auxiliar o combate à COVID-19	Fernandes et al. (2021)
AstroBot: *chatbot* para ensino e aprendizagem de física	Dantas et al. (2019)
BLAB: *chatbot* para disseminação de conhecimento sobre o território marítimo brasileiro	Matos et al. (2023)
CAERS: agente de conversação para intervenção pedagógica	Rossi et al. (2021)
chatbot para alocação de leitos em hospitais	Engelmann et al. (2021)
chatbot para universitários com possível perfil depressivo	Pires et al. (2023a)
Cobaia: modelo de respostas a perguntas econômicas	Santos et al. (2020)
EDUARDO: modelo semântico para integração de conteúdo	Santos et al. (2016)
FLOSS FAQ: *chatbot* baseado em *software* livre	Lacerda; Aguiar (2019)
Guardião: *chatbot* para idosos	Ferreira et al. (2019)
Mordomo virtual	Coelho et al. (2008)

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

	Trabalho
NL-SIIUE: agente recuperador de informações acadêmicas	Quintano; Rodrigues (2003)
Plantão Coronavirus: *chatbot* para detecção de sintomas da COVID	Coelho Da Silva et al. (2023)
Renan: *chatbot* para criação semi-automática de ontologias	Azevedo (2015)
TOB-SST: *chatbot* para suporte de educação sobre testes de *software*	Paschoal et al. (2019)

15.6.4 Para Saber Mais

Evidentemente, este capítulo introdutório não abarca todo o conhecimento no universo das tecnologias de diálogo atuais. Há célebres livros e capítulos sobre várias perspectivas de diálogo, diversos dos quais foram citados ao longo deste capítulo e podem ser encontramos nas referências. Para encontrar mais artigos científicos, recomendamos as seguintes fontes como ponto de partida[13]:

- o periódico *Dialogue & Discourse*: http://dialogue-and-discourse.org/
- as conferências da Associação de Linguística Computacional, mais especificamente os *tracks* de diálogo nas grandes conferências: https://aclanthology.org/
- o *workshop* do grupo de interesse dedicado a diálogo (SIGdial): https://aclanthology.org/sigs/sigdial/
- o *Workshop* de Semântica e Pragmática do Diálogo (SemDial): https://www.semdial.org/
- o *Workshop* Internacional de Sistemas de Diálogo Falado (IWSDS): https://dblp.org/db/conf/iwsds/index.html

Para a língua portuguesa, não parece haver uma fonte centralizadora. Há artigos sobre diálogo publicados em conferências de PLN (PROPOR, BRACIS, STIL) e também diversas teses que podem ser consultadas nos repositórios das universidades.

15.7 Conclusão: Mantendo o valor humano em foco

Antes de encerrar, vamos refletir um pouco sobre o que exploramos neste capítulo. Constatamos que o diálogo é parte vital de nossa humanidade, examinamos como ele ocorre como construção linguística, apresentamos formas de modelá-lo, bem como alguns recursos disponíveis em língua portuguesa, e consideramos a indispensabilidade de uma avaliação sólida e ampla. Vimos que é uma área de certa forma **unificadora**, na qual muitas tarefas de PLN podem ser relevantes e têm de trabalhar de forma orquestrada.

Ainda cabe uma pergunta, e quiçá a principal: **por que estamos fazendo isso?** Por que tentar criar sistemas que conversem conosco como humanos? Queremos, realmente, substituir um interlocutor humano por um programa de computador? O fato de **conseguirmos** criar aplicações (por exemplo, *chatbots*) não significa que **devamos** necessariamente fazer isso, nem que tal tecnologia seja sempre uma solução segura e desejada. Temos, outrossim, imensa responsabilidade acerca de nossos atos e de nosso conhecimento.

[13] Devemos, claro, sempre manter nosso olhar crítico e questionador: não é porque um artigo foi publicado que tudo nele está necessariamente "certo", validado ou justificado.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Ora, claro que há utilidade e benefícios: pode-se, por exemplo, automatizar certas tarefas monótonas, criar interfaces que facilitem a comunicação com a usuária ou o usuário de um dispositivo ou contribuir com o processo criativo de artistas. Principalmente, modelos de diálogo podem nos ajudar a progredir no entendimento científico do que é e de como funciona nossa habilidade cognitiva de uso interativo de linguagem.

Mas também **há riscos e perigos reais e concretos** em questão de vigilância, uso bélico, disseminação de informações falsas, segurança, privacidade e grupos vulneráveis, como ocorre com diversas tecnologias. Mais especificamente, é comum ver tentativas (na academia ou na indústria) de criar *chatbots* otimizados para técnicas de persuasão, de tentar simular emoções (obviamente, não genuínas), de incorporar uma "personalidade" ao sistema, de fingir se passar por um humano ou de substituição de profissionais especializados (como médicas e psicoterapeutas). Há, ainda, os riscos da antropomorfização (Abercrombie et al., 2023). São caminhos que evidentemente apresentam conflitos éticos e morais e que devem envolver pessoas capacitadas para pensar, identificar e, se necessário, regular essas questões, e não apenas implementar os detalhes técnicos. Infelizmente, há frequentemente uma supremacia de interesses comerciais ou políticos em vigor que, somados ao ritmo vertiginoso de criação de tecnologia nos últimos anos, torna muito difícil o controle e a remediação de efeitos colaterais nocivos.

É comum uma retórica em que a justificação para uma tecnologia sempre soe nobre. "O objetivo desse *chatbot* é fazer companhia para idosos sofrendo de solidão" ou "esse assistente virtual vai ajudar as pessoas cegas com a tecnologia" ou ainda "é para ajudar as tribos indígenas a preservar sua língua". Nessas horas, usualmente os grupos vulneráveis são lembrados. Mas quanto das tecnologias são propostas de forma colonizadora, sem qualquer conexão com a escuta real dos membros dessas comunidades, para saber quais suas reais necessidades e desejos (Bird, 2020)? Em que contextos realmente um *chatbot* treinado em grandes dados é necessário? Às vezes, outras soluções mais simples e controláveis são plenamente suficientes ou interessantes, mas há uma pressão pelo uso, a qualquer custo, de algo que está na moda (e que, lembremos bem, geralmente cria uma dependência comercial com um fornecedor). Outras vezes, há atividades cujo fator humano deveríamos tentar preservar. Queremos um interlocutor artificial (e pouco controlável) fazendo companhia a idosos? Interagindo com indígenas? Dando conselhos a pessoas com risco de suicídio? É essa sociedade que buscamos?

E quanto aos *chatbots* "de propósito geral", ou seja, que não são orientados a uma tarefa específica? Precisamos de um único agente artificial centralizador que saiba ensinar a fazer um guacamole, indicar filmes, dicas de maquiagem e ainda responder a questões existenciais? E que preço ético paga-se por práticas comerciais por trás desses modelos?[14] Nesse âmbito, Skantze; Doğruöz (2023) identificam um paradoxo muito pertinente envolvendo os atuais modelos de linguagem otimizados para *chat*: eles são modelos genéricos que não constroem base comum (além do contexto dado no *prompt*), fomentando apenas o uso para "jogar conversa fora", o que reduz a interação a um *chat* com fim em si mesmo. Além disso, eles sequer têm **intento comunicativo**: a atribuição de significado aos textos gerados é dada por quem os usa (Bender; Koller, 2020). Se ao mesmo tempo permite-se ser atribuída certa autoridade de conhecimento e "inteligência" a um modelo enquanto não se pode garantir ou verificar que seus enunciados têm o intento desejado e que as informações estão corretas (e, ainda, às vezes com tudo ocorrendo por trás de uma API

[14] Por exemplo, a mão-de-obra precarizada de seres humanos expostos por longas horas a conteúdo violento e perturbador usada para tornar um modelo de linguagem comercial "menos tóxico". Fonte: https://time.com/6247678/openai-chatgpt-kenya-workers/

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

intransponível e paga), arriscamos ir corroendo o valor social do diálogo,[15] um altíssimo e irreversível custo a se pagar.

Para concluir, voltemos a Freire (1989): um diálogo é, ainda, um "**ato de criação**", um caminho pelo qual ganhamos **significação** enquanto pessoas, encontrando-nos para pronunciar o mundo e conquistá-lo para nossa libertação. Tenhamos a coragem de nos opor a práticas destrutivas e adversas que visem tornar modelos de diálogo uma ferramenta de dominação, manipulação ou desinformação. Que saibamos construir aplicações que façam emergir e florescer nosso melhor como indivíduos e sociedade. Que aprendamos a usá-las com **sabedoria, solidariedade e justiça**, identificando e modulando de forma consciente e responsável seus impactos na realidade e priorizando sempre o **profundo valor da linguagem e da existência humana**.

15.8 Exercícios

1. Escolha um dos diálogos dos exemplos na Seção 15.3 e clique no *link* correspondente para acessar o exemplar completo. Anote todos os fenômenos discutidos na Seção 15.3 que você conseguir identificar. Quão fácil é fazer interpretações *post factum* com base apenas no áudio e/ou na transcrição?

2. Assista ou ouça um diálogo de um filme, série ou peça de teatro de sua preferência. Quais fenômenos discutidos na Seção 15.3 não estão presentes? Quais parecem ocorrer mais ou menos raramente do que no exercício anterior?

3. Acesse o *site* do *corpus* NURC[16], clique no ícone "Acessar página com o material" e depois em um título de sua escolha. Inspecione o conteúdo dessa página. Tente entender os formatos de arquivos de meta-dados. Por que todas essas informações são relevantes? Explore as anotações no arquivo de extensão *.TextGrid*. Você consegue entender a estrutura de dados deste arquivo?

4. Na página inicial do site to NURC, clique agora em "Acessar página de busca". Construa uma pesquisa, usando o filtro para buscar apenas por diálogos. Escolha um ID entre os resultados e clique nele para acessar a conversa completa. Verifique a transcrição e, se lhe for possível, ouça o áudio acompanhado o texto. Que aspectos a transcrição não consegue capturar muito bem?

5. Considere os seguintes tipos de interação social: (a) duas amigas jogam um jogo online, simultaneamente, e se comunicam por uma ferramenta de *chat* na própria interface do jogo, sabendo que a comunicação está interceptada por um membro da equipe concorrente (b) um casal mantém um relacionamento a distância, se comunicando ora por chamadas de vídeo, ora por mensagens gravadas de áudio, (c) um grupo de quatro pessoas não-ouvintes conversa por língua de sinais em um palco de auditório, (d) uma cientista se corresponde por cartas com um editor de um periódico para publicar um artigo (e) dois atores atuam em uma cena de filme, um deles tem uma deficiência visual, perante uma equipe de cineastas que intervêm se necessário. Classifique-as conforme as dez características do diálogo discutidas na Seção 15.2 e considere também quais aspectos de multimodalidade estão envolvidos.

[15] Visão discutida por Kevin Munger em sua palestra "*Chatbots* para o bem e o mal" na conferência EACL de 2023, disponível em https://doi.org/10.48448/5tew-e744.

[16] https://fale.ufal.br/projeto/nurcdigital/

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

6. Tente formalizar as situações do exercício anterior de forma esquemática, por exemplo através de um fluxograma que descreva todos os componentes e como eles interagem (isto é, quem participa da conversa, como se dá a geração de enunciados e a troca de turnos, que aspectos da situação são compartilhados, que meio usam para se comunicar, etc).

7. Escolha três aplicações de sistemas de diálogo que você acharia útil em seu dia a dia. Defina seu propósito, dispositivo, domínio, modalidade, iniciativa, conforme o que foi exposto na Seção 15.4.

8. Imagine que você precisa construir (separadamente) *chatbots* (por fala e áudio) orientados a tarefa para administrar as seguintes atividades por conversa telefônica: (a) receber pedidos de compra de material escolar infantil, (b) agendar consultas médicas em hospital com várias especialidades e (c) registrar reclamações sobre um serviço de internet em diferentes localidades. Defina uma lista de possíveis intentos que virão de das pessoas que buscam atendimento. Com base nisso, defina também as lacunas básicas dessa tarefa e os possíveis tipos de valores previsíveis e válidos para preencher cada uma.

9. Como você avaliaria os sistemas de diálogo do exercício anterior? Pense em quais dimensões, qualidades e características são importantes nessas tarefas e que tipos de métricas poderiam capturar esses aspectos.

10. Repita os dois exercícios anteriores, agora considerando que esses agentes se comunicam por mensagens de texto instantâneas em uma interface de *chat*.

11. O que ocorre se um agente conversacional não mantiver uma representação interna do estado do diálogo (isto é, quais informações já foram recebidas, quais ainda não, e as implicações disso)?

12. Quais as possíveis consequências práticas de um modelo que se baseia na duração de silêncios como único sinal para troca de turno?

13. O que ocorre quando um agente conversacional não é programado para lidar com pedidos de clarificação? Por um lado, se ele não souber realizá-los: o que acontece quando há ambiguidade e ou subespecificação de uma instrução dada? Por outro, se ele não souber interpretá-los, o que acontece com a comunicação dali em diante?

14. Relembrando, jogos de linguagem são interações curtas, com um propósito simples e bem definido. Defina jogos de linguagem que viabilizem a ocorrência dos seguintes fenômenos: construção de base comum, referências a objetos, comandos de direção. As conversas no jogo devem ocorrer de forma natural, ou seja, sem que os participantes sejam explicitamente instruídos para produzir esses fenômenos.

15. Imagine que você desenvolveu a lógica de um *chatbot* para lembrar uma pessoa de tomar medicamentos e se exercitar, e agora precisa definir que enunciados ele vai proferir. Você quer enunciados que tenham variação linguística, para não se tornar algo repetitivo, mas que possam ser bem controlados. Defina 5 *templates* (se necessário, com lacunas a serem preenchidos pelo sistema) para serem usados em cada um dos seguintes casos: (a) dar uma saudação de bom dia que informe o horário e a temperatura, (b) orientar a pessoa que está na hora de se alongar, (c)

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

lembrar a pessoa de qual medicamento tomar e em qual quantidade, e (d) perguntar como a pessoa está se sentindo após diferentes atividades cotidianas.

16. Se você interage (ou já interagiu) com algum *chatbot* ou assistente virtual, tente verificar sob a ótica do que aprendeu neste capítulo quais problemas de comunicação ocorrem e como eles são resolvidos (ou não). O que você acha que é preciso melhorar?

17. Sentenças do tipo *garden path*[17] contêm algum tipo de ambiguidade sintática temporária. Há também ambiguidades locais de significado que são resolvidas quanto há contexto suficiente. O efeito disso é que, geralmente, somos levados a crer que o enunciado terá um sentido e, no meio da frase, ele "muda de rumo" e temos de reajustar nossa interpretação. Por exemplo, ao ler "o corredor da maratona é muito estreito", possivelmente você primeiro pensou que se tratava de uma pessoa que corre a maratona, e só quando leu "estreito" pôde interpretar que se tratava de um corredor por onde a maratona passaria. Que consequências esse tipo de fenômeno pode ter em um sistema de processamento incremental? Qual mecanismo é necessário para que o sistema se recupere da interpretação errada?

18. Quais considerações éticas devem ser levadas em conta no uso de grandes modelos de linguagem como componente de um *chatbot*? Pense nas consequências tanto do uso intencional da ferramenta quanto de possíveis usos escusos.

Agradecimentos

Agradeço às colegas Renata Vieira e Amanda Rassi, bem como às editoras do livro, pela valiosa revisão deste capítulo. *Disclaimers*: os recursos disponíveis foram elencados para fins informativos e didáticos; não nos responsabilizamos pelo conteúdo e pelo uso de cada um. Os exemplos de diálogos reais foram selecionados com base nos fenômenos que precisamos exemplificar, o que não significa que endossamos o conteúdo das conversas completas. A estrutura e a bibliografia desse capítulo seguem parcialmente os materiais de aula sobre modelos de diálogo do Prof. David Schlangen. Em termos de conteúdo, esse capítulo se assemelha também ao célebre capítulo didático escrito por (Jurafsky; Martin, 2023), disponível em https://web.stanford.edu/~jurafsky/slp3/ (capítulo 15), cuja leitura sugerimos para aspectos mais técnicos.

[17] https://pt.wikipedia.org/wiki/Teoria_do_Garden_Path

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Parte VIII

Dados

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Capítulo 16

Conjunto de dados, *dataset* e *corpus*

Cláudia Freitas

Publicado em: 26/09/2023
Atualizado em: 16/04/2026

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

16.1 Introdução

Dados linguísticos são centrais no PLN, sobretudo nos paradigmas estatístico, neural e híbrido (Seção 1.4). A preparação de conjuntos de dados – *datasets* – de qualidade para o PLN é um empreendimento que costuma envolver conhecimentos variados – de computação e linguística, no mínimo. Neste capítulo, fazemos uma apresentação de conceitos básicos e metodologias relacionados à criação de *datasets* que contêm algum tipo de intervenção ou curadoria humana nos dados[1].

Apesar da relativa popularização das abordagens neurais e dos LLMs Capítulo Modelos de linguagem, que têm como insumo dados brutos/textos crus, sem intervenção humana, a preparação cuidadosa de conjuntos de dados se justifica por, pelo menos, dois fatores: o primeiro é a já reconhecida melhoria no desempenho de modelos quando são ajustados para tarefas ou domínios específicos (o *ajuste fino supervisionado*); a segunda é a constatação de que os altos custos financeiros envolvidos no treinamento dos grandes modelos de língua tornam abordagens estatísticas tradicionais, computacionalmente menos custosas, alternativas viáveis para a representação de informação linguística no PLN.

Um *dataset*, literalmente, é um conjunto (*set*) de dados (*data*). Dados são elementos que, organizados (ou distribuídos) de uma(s) certa(s) maneira(s), isto é, tratados, produzem informação. Praticamente qualquer coisa pode ser um dado. No PLN, os dados que usamos são dados linguísticos; nossa matéria prima é a linguagem humana, e cada língua individualmente.

Os dados podem ser

- palavras, que podem ser classificadas como substantivos, advérbios, verbos etc;
- postagens em rede social, que podem ser classificadas como ofensivas ou não;
- palavras ou unidades maiores, que podem ser classificadas como pessoa, lugar etc;
- pronomes, que podem ser classificados como masculinos, femininos ou neutros;
- documentos ou frases, que podem ser classificados como simples ou complexos;
- pares de frases, que podem ser classificadas como sinônimas (mais precisamente, classificadas quanto ao seu grau de similaridade) ou não;

[1] Este capítulo contém adaptações de (Freitas, 2022), onde apresento os bastidores do processo de anotação -- planejamento e esquema de anotação, documentação e concordância entre anotadores - com mais algum detalhe, e de um ponto de vista linguístico. E, apesar de ter sido publicado primeiro, o capítulo (Caseli et al., 2022) aproveitou muito do que já havia sido escrito para este capítulo.

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

- segmentos de textos, que podem ser classificados e relacionados como uma pergunta e a resposta a ela associada;
- sequência de caracteres que aparecem entre espaços em branco ou espaços em branco e pontuação, que podem ser classificados como palavras;
- posição que cada palavra ocupa em uma frase ou em um texto inteiro;
- frequência de cada palavra ao longo de um texto;
- sons, que podem ser classificados como fala humana, uma tosse ou uma risada;
- sons de fala, que podem ser classificados como uma palavra ("agente") ou mais de uma ("a" "gente").

Partindo dos exemplos acima, o elemento "palavra" pode virar um dado quando atribuímos a ele algum valor, como a sua classe gramatical (substantivo, verbo etc.), classe semântica (pessoa, lugar etc.), sua posição no texto ou a sua frequência.

No PLN, estes valores podem ser atribuídos aos dados de duas maneiras. A primeira delas é de maneira explícita – por exemplo, com cada palavra associada a uma informação do tipo PoS (classe de palavra, do inglês, *part-of-speech*), sendo essa informação do tipo *substantivo*, *verbo*, *pronome*, *advérbio* etc. Ou cada frase (ou palavra) associada a uma informação do tipo *polaridade de opinião*, sendo essa informação do tipo *positiva*, *negativa*, *neutra*. No primeiro caso podemos dizer que organizamos (ou distribuímos, ou classificamos, ou rotulamos) as palavras do texto conforme sua classe morfossintática, no segundo, podemos dizer que organizamos (ou distribuímos, ou classificamos, ou rotulamos) as frases (ou palavras) do texto conforme sua polaridade. O que há em comum em ambos os casos é a **organização (ou classificação) dos dados** conforme classes pré-estabelecidas que nos parecem relevantes para explorar o conteúdo linguístico, e a partir delas produzimos informação: por exemplo, se há mais opiniões positivas ou negativas (ou neutras) com relação a um determinado objeto.

Mas nem todo *dataset* linguístico precisa ter seus dados organizados de acordo com atributos "externos" ao texto. Grandes modelos de língua – modelos de previsão de palavras – têm como entrada imensos volumes de texto, sem informação linguística explícita associada (Capítulo Modelos de linguagem). A informação capturada é a posição da palavra no texto e a frequência com que é usada, e isso já permite saber muito sobre as palavras, desde que tenhamos muitas delas. Mas não é deste tipo de *dataset* que nos ocuparemos aqui (veja o [sec-cap-modelos-metodos-dados]), e nem dos procedimentos que transformam posição e frequência em informação linguística, tema do Capítulo 10. Nosso foco está nos dados linguísticos que possuem alguma organização explícita humana, feita conforme classes pré-definidas por nós – *dados anotados*, que são *elementos* linguísticos que possuem classificações, anotações ou rótulos linguísticos que codificam alguma dimensão do nosso entendimento sobre as palavras, frases ou textos. Porque contêm classificações (ou análises) atribuídas aos elementos linguísticos, estes conjuntos de dados também podem ser considerados *corpora anotados*.

16.1.1 *Dataset* ou *corpus* anotado?

Um *corpus* é um conjunto de dados linguísticos. A utilização de *corpus* – palavra latina que significa *corpo* e que tem como plural a palavra *corpora* – vem de longa data nos estudos linguísticos e lexicográficos[2], e ganha força nos anos 1980 com a popularização dos computadores. A partir daí, e cada vez mais, *corpus* diz respeito a uma coleção de

[2]Lexicografia é a área que se dedica a fazer dicionários.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

textos que pode ser processada por computadores. O material que compõe um *corpus* (os textos) é coletado com algum propósito (investigar ou explorar algum aspecto da linguagem, teórico ou aplicado) e foi produzido "naturalmente", isto é, não estamos diante de frases artificialmente inventadas com o objetivo de construir um *corpus*.

Como inicialmente o uso de *corpus* estava vinculado aos estudos linguísticos, o propósito mais comum é (ou era) o de estudar a/uma língua. Mas, à medida que a utilização de *corpus* vai sendo ampliada para áreas além da linguística, como no PLN, os critérios que caracterizam um *corpus* vão também se ampliando. A *naturalidade* – enunciados naturalmente construídos em situações reais de trocas linguísticas, aspecto fundamental se o interesse é estudar uma língua – é deixada de lado quando pensamos em *corpora* criados artificialmente (automaticamente) a partir de dados "naturais". Este é o caso de materiais criados por meio de tradução automática, ou por meio do aumento artificial de dados, que pode ser feito utilizando (i) regras ou (ii) LLMs e *prompts* capazes de induzir certas construções linguísticas. Um exemplo do primeiro caso é o WinoBias[3] (Zhao et al., 2018), criado para estudar viés de gênero e no qual, por meio de regras, foi criado um material balanceado quanto ao gênero com o mesmo número de entidades (pessoas) do gênero masculino e do gênero feminino. Um exemplo do segundo caso é o ToxSyn-PT, criado para permitir a detecção de discurso de ódio[4] (Brito et al., 2025).

Corpus ou *dataset*? Podemos usar três critérios para diferenciar *corpus* e *dataset*: utilidade, tipo de anotação, e tamanho.

Um *corpus*, para ser considerado um *dataset* linguístico no PLN, precisa ter algum **tamanho**. No mínimo, tamanho suficiente para permitir avaliar de maneira confiável o resultado de uma análise automática e, idealmente, tamanho suficiente para treinar um modelo de aprendizado de máquina. Nos estudos linguísticos, conforme o fenômeno pesquisado, quantidade pode não ser uma exigência. Em ambos os casos (nos estudos linguísticos e no PLN), podemos dispor de material já classificado – ou anotado, ou rotulado, ou etiquetado. Esta anotação ou classificação pode ser de diferentes naturezas e pode estar associada a diferentes segmentos de texto (palavras, expressões, frases, parágrafos ou o texto inteiro). Por exemplo:

- Indicar a classe gramatical de uma palavra;
- Indicar se duas frases são sinônimas ou não;
- Indicar se um tweet expressa discurso de ódio;
- Indicar se uma palavra se refere a uma PESSOA;
- Relacionar uma palavra a pronomes e demais outras maneiras pelas quais esta palavra é referida ao longo de um texto;
- Indicar se uma palavra, expressão, ou tweet é favorável ou desfavorável com relação a alguém ou algo;
- Indicar se uma frase é uma boa legenda para uma imagem;
- Indicar, para um trecho de áudio de um enunciado linguístico, a sua transcrição textual;
- Indicar qual sentimento uma palavra ou segmento de texto evoca.

Todos esses rótulos ou anotações têm em comum o fato de codificarem a interpretação humana, assim outro critério relevante para a distinção entre *corpus* e *dataset* é o **tipo**

[3] https://github.com/uclanlp/corefBias#readme

[4] Como os próprios autores reconhecem, o fato de usarem um conjunto fixo de **prompts** para a geração limita o surgimento de novos padrões discursivos, além de incluir os vieses do LLM gerador.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

de anotação. Nos estudos linguísticos é possível (e comum) a utilização de *corpora* que foram anotados de maneira completamente automática, desde que a anotação automática seja de qualidade. Já quando pensamos em *datasets* para o PLN, como sua utilidade está na avaliação e treinamento (ou ajuste) de ferramentas ou modelos, as anotações são feitas por pessoas, ou feitas por máquinas e revistas por pessoas. Neste caso, *datasets* são equivalentes a *corpora padrão ouro* (*gold standard*).

É possível que a diferença quanto à nomeação – *corpus* anotado ou *dataset* – também se deva à **utilidade** em foco: quando aquilo que a Linguística chama de *corpus* anotado passa a ser usado como material para treinar modelos de linguagem, ele também pode ser visto como um *dataset*. Vejamos a apresentação do já referido WinoBias, criado para a verificação de viés de gênero na resolução de correferência (Capítulo 12). Retomaremos o WinoBias mais à frente, mas interessam aqui as duas primeiras frases do resumo do artigo que apresenta o material (grifo meu)[5]:

> Apresentamos um novo benchmark, WinoBias, para resolução de correferência com foco no viés de gênero. Nosso ***corpus*** contém sentenças no estilo do esquema de Winograd com entidades correspondentes às pessoas às quais se faz referência por meio de sua profissão .

Já no repositório do projeto onde está o WinoBias, encontramos a seguinte apresentação (grifo meu)[6]:

> Analisamos diferentes sistemas de resolução de anáfora para entender as questões de viés de gênero presentes em tais sistemas. Dando uma mesma frase como entrada para o sistema, mas apenas mudando o gênero do pronome na frase, há variação no desempenho dos sistemas. Para demonstrar a questão do viés de gênero, criamos o ***dataset*** WinoBias.

Nem todo *dataset* tem dados linguísticos, e nem todo conjunto de dados linguístico é um *corpus* anotado. Nem todo *corpus* anotado é considerado, no PLN, um *dataset* linguístico. Neste capítulo, trataremos de datatsets – *corpora* padrão ouro – que são conjuntos de dados linguísticos classificados conforme orientação humana. (Figura 16.1).

16.1.2 Anotação linguística

A anotação é uma **atividade de classificação**: temos um conjunto de classes (as etiquetas) – também chamado de *tagset* – previamente definidas e critérios que guiam a classificação. O que torna a anotação interessante – ou desafiadora – é que a classificação é feita levando em conta o contexto do enunciado linguístico. Nos exemplos do Quadro 16.1, temos anotações de classes de palavras (anotação de PoS) e anotação de polaridade, utilizada na tarefa de análise de sentimentos. A anotação de polaridade, nos exemplos, está presente em dois níveis: no nível da palavra propriamente – cada palavra recebe uma etiqueta indicando se a polaridade é positiva [+], negativa [-] ou neutra [0] – e no nível da frase – cada frase recebe uma etiqueta indicando se a polaridade é positiva,

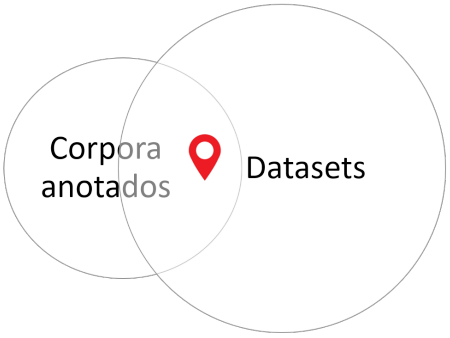
[5] "Our corpus contains Winograd-schema style sentences with entities corresponding to people referred by their occupation (e.g. the nurse, the doctor, the carpenter)." https://aclanthology.org/N18-2003.pdf

[6] "We analyze different resolution systems to understand the gender bias issues lying in such systems. Providing the same sentence to the system but only changing the gender of the pronoun in the sentence, the performance of the systems varies. To demonstrate the gender bias issue, we created a WinoBias dataset." https://github.com/uclanlp/corefBias/tree/master

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Figura 16.1: Nosso ponto de partida



negativa ou neutra com relação ao objeto sendo analisado. A anotação pode ser codificada (ou formalizada) de diferentes maneiras, e a codificação do Quadro 16.1 é apenas ilustrativa. Na Seção 16.4.4 retomaremos aspectos de codificação e formalização.

Quadro 16.1: Exemplos de frases anotadas com PoS e polaridade de sentimento

1. TRISTE$_{_SUBST_[0]}$ é$_{_V_[0]}$ uma$_{_ART_[0]}$ palavra$_{_SUBST_[0]}$ de$_{_PREP_[0]}$ 6$_{_NUM_[0]}$ letras$_{_SUBST_[0]}$ [frase neutra]
2. Um$_{_NUM_[0]}$ dos$_{_PREP+ART[0]}$ livros$_{_SUBST_[0]}$ mais$_{_ADV_[0]}$ tristes$_{_ADJ_[-]}$ que$_{_PRON-Rel_[0]}$ já$_{_ADV_[0]}$ li$_{_V_[0]}$ [frase neutra?]
3. Sofri$_{_V_[-]}$ com$_{_PREP_[0]}$ a$_{_ART_[0]}$ protagonista$_{_SUBST_[0]}$ a$_{_PREP_[0]}$ cada$_{_PRON_[0]}$ nova$_{_ADJ_[0]}$ página$_{_SUBST_[0]}$; sofri$_{_V_[-]}$ quando$_{_ADV_[0]}$ o$_{_ART_[0]}$ livro$_{_SUBST_[0]}$ acabou$_{_V_[0]}$ [frase positiva]
4. Nunca$_{_ADV_[0]}$ sofri$_{_V_[-]}$ tanto$_{_ADV_[0]}$ para$_{_PREP_[0]}$ ler$_{_V_[0]}$ um$_{_ART_[0]}$ livro$_{_SUBST_[0]}$ [frase negativa]

Anotação é uma **atividade interpretativa**. Como em boa parte das vezes o que está sendo anotado é o resultado de um amplo consenso, ou do "senso comum", ficamos com a impressão (errada) de que estamos diante de uma classificação objetiva. Decidir, por exemplo, se a palavra "TRISTE", na frase 1, deve ser classificada como *substantivo* ou *adjetivo* é uma decisão que irá depender da teoria adotada na anotação. Na frase 2, analisar a palavra "um" como *artigo* ou *numeral* também é fruto de uma escolha (teórica), e não a codificação de um dado objetivo. E, embora aqui o aspecto interpretativo da anotação pareça um detalhe teórico que só interessa a linguistas, veremos na Seção 16.4.5 consequências nem um pouco irrelevantes desta crença na objetividade da classificação, que aparece disfarçada de senso comum.

Quanto à **importância do contexto**, vemos pelos exemplos que ainda que as frases 1 a 3 contenham palavras de polaridade de sentimento considerada negativa ("triste" e "sofri"), no exemplo (1) a frase não tem polaridade (ou tem polaridade neutra), no exemplo (2) é difícil decidir se estamos diante de um julgamento de valor (negativo) sobre o livro ou se diante de uma constatação, por isso a polaridade sugerida é seguida com um "?", e no exemplo (3) temos uma frase que indica uma opinião positiva sobre um livro, apesar da menção ao sofrimento.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

16.2 *Datasets* pra quê?

A existência de *datasets* linguísticos, ou *corpora* padrão ouro, é fundamental para uma série de tarefas e aplicações de PLN. E por que fundamental? São três os motivos, todos igualmente importantes.

O primeiro deles é consequência da popularização, no PLN e na IA, de métodos baseados em aprendizado de máquina (AM): precisamos de exemplos do que se precisa aprender. E mesmo com os avanços da área, **bons exemplos**[7] **continuam necessários**, com a vantagem de agora os algoritmos necessitarem de uma quantidade menor deles, graças ao ajuste fino (Capítulo Modelos de linguagem). Já se sabe que quanto mais cuidado na preparação dos dados, melhor o desempenho dos modelos – melhor a qualidade das predições[8]. Neste contexto, um bom *dataset*, fruto de uma preparação cuidadosa, pode ser visto como um atalho para o aprendizado, como um empurrãozinho que damos nos modelos para que atinjam logo o melhor resultado possível.

Então um motivo para investirmos na criação de *datasets* linguísticos é fornecer exemplos para que certos tipos de aprendizado possam acontecer de maneira eficiente.

O segundo motivo que torna *datasets* fundamentais em diversas tarefas de PLN é que eles **facilitam o processo de avaliação** de um sistema, ferramenta ou modelo, e de **comparação** entre eles. Isto porque se a anotação codifica, no *corpus*, a compreensão humana sobre algo, e o que queremos das máquinas em certas tarefas é que elas reproduzam esta compreensão humana, a melhor maneira de saber em que medida um resultado é bom é comparando-o com o entendimento humano. Nesse contexto, poder dispor de um *corpus* padrão ouro facilita muito as coisas. Sem ele, a alternativa para a avaliação é selecionar uma amostra do material analisado automaticamente e avaliar. Embora esta seja a única opção disponível em certos casos, não é ideal porque dificulta comparações com o desempenho de outros modelos/sistemas/ferramentas. Isto é, se cada ferramenta for avaliada de maneira "independente" a partir de uma amostra do seu resultado, será difícil uma comparação com os resultados de outras ferramentas. Outra desvantagem da avaliação por meio da análise de uma amostra, ainda que mais facilmente contornável, é que quando separamos uma amostra para fazer uma análise de erros, é mais fácil perceber aquilo que foi analisado de maneira errada (falsos positivos) do que aquilo que não foi analisado, mas deveria ter sido (falsos negativos)[9]. Quando se trata de LLMs, a relação entre avaliação e *datasets* é diferente. Tarefas de **geração de texto**, como sumarização, simplificação textual, tradução, respostas a perguntas, não dispõem de um gabarito fechado e, nestes casos, o *dataset* contém apenas um gabarito de referência, já que a partir de um mesmo texto podemos produzir diferentes resumos, simplificações, traduções, ou respostas, que podem estar corretos mesmo compartilhando poucas palavras a referência. Como textos gerados por modelos neurais apresentam um amplo domínio de sinonímia e paráfrase, uma avaliação baseada apenas na **forma** das palavras será pouco informativa. Por fim, é preciso

[7] Bons exemplos são exemplos representativos, variados e analisados de maneira consistente, como veremos na Seção 16.3.

[8] Os trabalhos (Souza et al., 2020) e (Pires et al., 2023b) trazem dados da língua portuguesa a respeito disso, e a palestra de Rodrigo Nogueira sobre a adaptação e modelos de linguagem para o Português também: Adaptando Modelos de Linguagem para o Português: Passado, Presente e Futuro - com Rodrigo Nogueira

[9] Quando estamos tratando de *datasets* para avaliação no contexto de aprendizado de máquina, é importante que o material usado na avaliação não tenha sido utilizado nem nas etapas de treinamento e nem de validação do modelo. Esta separação do *dataset* em partes diferentes é importante para que a avaliação seja de fato um teste, e não uma avaliação em que o modelo está "roubando", uma vez que seu desempenho está sendo testado nos mesmos dados em que foi treinado.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

ter cuidados adicionais na utilização de *datasets* para avaliar LLMs devido ao fenômeno da contaminação dos dados (Balloccu et al., 2024): uma vez que o *dataset* esteja público na internet, é preciso garantir que ele não tenha sido usado no treinamento do modelo, o que levaria a um desempenho artificialmente bom na tarefa avaliada (o capítulo Capítulo Avaliação de Grandes Modelos de Linguagem aborda a avaliação de LLMs).

O último motivo é um desdobramento dos anteriores: a partir do momento em que temos condições de treinar, avaliar e comparar resultados, temos condições de **avançar no PLN**, uma vez que o avanço na área pode ser medido pelo desempenho em tarefas. Ou seja, no PLN, um *dataset* é criado para ajudar a resolver algum problema ou tarefa[10]. Assim, para que o *dataset* exista, houve uma pergunta/problema/tarefa anterior que motivou a sua existência – alguma dimensão do PLN foi percebida como sensível e foi considerada digna de uma medição: o quanto o PLN é bom em responder perguntas, encontrar palavras de um certo tipo semântico, relacionar palavras que se referem à mesma entidade ao longo de um texto, traduzir, resumir ou simplificar um texto, dentre tantas outras.

E aqui aproximamos *datasets* e **avaliações conjuntas** (*shared tasks*). Uma avaliação conjunta tem como principal objetivo incentivar a pesquisa e desenvolvimento de uma área, uma vez que fornece uma estrutura experimental comum (os mesmos conjuntos de dados e as mesmas medidas de avaliação). Além disso, avaliações conjuntas também são maneiras de divulgar uma tarefa. Se achamos que um determinado aspecto do PLN precisa ser avaliado, ou precisa de atenção, a criação de uma avaliação conjunta que a tematize é um caminho. E, como já sinalizado, avaliações conjuntas só existem se existem *datasets* associados (e quanto melhor o *dataset*, mais bem-sucedida a avaliação) (Veja o capítulo Capítulo 19).

Voltando um pouco no tempo para exemplificar, foi a criação de um *corpus* padrão ouro (chamado "Coleção Dourada") no âmbito da avaliação conjunta HAREM[11], em 2007-2008, que permitiu o avanço na tarefa de identificação e classificação de entidades mencionadas em português. Foi a criação do *corpus* ASSIN que permitiu a realização da avaliação conjunta ASSIN (Avaliação de Similaridade Semântica e Inferência Textual)[12], levando ao avanço em tarefas de similaridade semântica e inferência em português.

No entanto, em ambos os casos estamos diante de tarefas (identificação de entidades mencionadas, de similaridade semântica) que já existiam para outras línguas e que careciam de recursos para que pudessem ser abordadas também em língua portuguesa. Que outras tarefas poderíamos ter? Que desafios ou tarefas o PLN tem e ainda não foram abordados por falta de recursos? Por isso, além de avaliar e treinar, um *dataset* permite, ainda que indiretamente, **pautar os rumos do PLN**.

É definindo tarefas que vamos abrindo caminhos no PLN – tanto caminhos previamente explorados para outras línguas, mas ainda não pavimentados para o português, quanto caminhos realmente novos, ainda não explorados em nenhuma língua. Um problema, ou tarefa, é enfrentado quando temos os meios para fazê-lo – e a construção de conjuntos de dados é um dos meios de que precisamos, considerando as abordagens atuais.

[10] Já um *corpus* anotado padrão ouro pode ter sido criado com a intenção de estudar um determinado fenômeno linguístico.

[11] https://www.linguateca.pt/HAREM/.

[12] http://propor2016.di.fc.ul.pt/?page_id=381 e https://sites.google.com/view/assin2/.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

16.2.1 Sobre a importância dos dados

Quando indicamos que, para o AM, *datasets* são relevantes porque fornecem exemplos e permitem criar modelos, nem sempre nos damos conta da importância dos dados de um ponto de vista qualitativo. Isto é, pensamos nos dados como recursos de que precisamos dispor em abundância para produzir bons modelos de linguagem (Capítulo Modelos de linguagem), mas não como fonte de "visões de mundo". Afinal, se uma língua constrói e representa as visões de mundo e crenças de seus falantes, um modelo de língua (ou de linguagem) irá, mesmo que indiretamente, codificar visões de mundo e crenças subjacentes aos dados linguísticos com que foi treinado. De forma mais direta: se, no paradigma do AM com representações distribuídas, todo o conhecimento vem dos dados, precisamos entender bem que dados são esses.

Um exemplo famoso da importância dos dados vem da pesquisadora Robyn Speer, que criou um algoritmo de Análise de Sentimento baseado em representações distribuídas (Capítulo 10), usando como dados textos publicados da internet. Ela percebeu que o algoritmo estava classificando restaurantes mexicanos de maneira negativa, o que não encontrava respaldo na pontuação dada pelos usuários para avaliar os restaurantes e nem nos próprios textos das resenhas dos restaurantes. O motivo da avaliação ruim dos restaurantes mexicanos era que a palavra *mexicano* sempre aparecia associada a *ilegal*: *imigrantes mexicanos* estavam associados a *imigrantes ilegais,* levando à classificação de restaurantes mexicanos como algo negativo. A experiência e suas lições estão relatadas no post "*How to make a racist AI without really trying*" ("Como fazer uma IA racista sem fazer muito esforço").

E, apesar das tentativas de mitigar viés indesejado utilizando pessoas comuns para classificar dados (Seção 16.4.5), ainda há muito por fazer. Um estudo de 2021, por exemplo, mostrou que textos ficcionais criados pelo modelo de linguagem GPT-3 reproduzem estereótipos de gênero: se há um personagem feminino, ele tem muitas chances de estar associado a palavras que remetem a ambiente doméstico/familiar e à aparência/corpo; se o personagem é masculino, há muitas chances de estar associado a palavras que remetem à política, guerra e tecnologia (Lucy; Bamman, 2021). E por que isso? Justamente porque é o padrão (e viés) encontrado nos dados. Em uma exploração da caracterização de personagens literários em obras de língua portuguesa que já estão em domínio público – obras brasileiras e portuguesas –, encontramos um padrão bastante parecido quando analisamos as caracterizações atribuídas preferencialmente a personagens masculinos e personagens femininos: personagens femininos são caracterizados sobretudo quanto à aparência, com destaque para a beleza (ou ausência dela); personagens masculinos caracterizados sobretudo quanto a traços de caráter como excelência, coragem, liberdade e sabedoria (Freitas; Santos, 2023).

Ou seja, os estereótipos produzidos pelo modelo de linguagem nada mais são do que a reprodução dos padrões que já estão nos dados. E os padrões que estão nos dados nada mais são do que comportamentos linguísticos que estão na nossa sociedade. Modelos de IA não são neutros nem isentos de viés, e talvez seja impossível fazer diferente[13]. Os dados são sempre um registro do que foi dito em algum tempo/espaço, são sempre dados do passado, e por isso, invariavelmente, modelos de previsão estão fadados a refletir o passado. Isso nos leva a uma reflexão: podemos, por meio de um passado alterado, prever um futuro que queremos? Que futuro queremos? Será que *datasets* construídos artificialmente com o objetivo de eliminar vieses indesejados, ao estilo do WinoBias, que artificialmente (e

[13] https://theconversation.com/ai-free-from-bias-and-ideology-is-a-fantasy-humans-cant-organise-data-without-distorting-reality-262523

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

indiretamente) constroem relações igualitárias de gênero, raça ou etnia, seriam capazes de produzir linguagem e, consequentemente, de criar mundos, mais igualitários? Quais as implicações éticas disso? Em 2025, o presidente estadunidense Donald Trump anunciou a exigência de que LLMs usados pelo governo fossem "ideologicamente neutros", acusando modelos associados à promoção de diversidade, igualdade e inclusão de serem "IA woke". "Quem controla o passado, controla o futuro.", dizia o slogan do estado totalitário do romance *1984*, de George Orwell.

16.3 Características de um bom *dataset* linguístico

Quando falamos de um bom *dataset* linguístico, ou de um bom *corpus* anotado, são cinco as características desejáveis: consistência, variedade, representatividade, documentação detalhada e tamanho.

- **Consistência** – por consistência, entenda-se que fenômenos semelhantes devem ser anotados (isto é, analisados) da mesma maneira. A consistência permite que algoritmos de aprendizado de máquina generalizem a partir dos dados e que as avaliações sejam confiáveis. Em outras palavras: se para aprender são necessários exemplos, mas os exemplos são inconsistentes – às vezes *Brasil* em construções do tipo *morar no Brasil* está anotado como LOCAL, às vezes está anotado como ORGANIZAÇÃO –, não será possível uma boa generalização, pois os dados ruidosos trarão dificuldade para esse processo.
- **Variedade** – na medida do possível, um bom *dataset* deve ser variado (ou balanceado) com relação aos fenômenos para os quais foi construído. Por exemplo, se desejamos um *dataset* para a tarefa de análise de sentimento/opinião com resenhas de produtos, um compilado de avaliações de páginas do tipo "Reclame Aqui" não é recomendado, pois conterá, invariavelmente, muito mais avaliações negativas do que positivas. Se desejamos um *dataset* para reconhecimento de fala (Capítulo 3), é importante que ele contemple a diversidade dos sotaques brasileiros.
- **Representatividade** – esperamos que os textos que compõem o *corpus*/*dataset* sejam representativos do tipo de texto que será alvo da aplicação, isto é, ao qual o modelo baseado nesse *dataset* será aplicado. Se a ideia é criar um modelo/ferramenta que irá procurar informações em relatórios técnicos, não é indicado que o modelo seja gerado a partir de um *dataset* que contém apenas textos jornalísticos, por exemplo. Por outro lado, considerando o que foi dito sobre a presença de viés indesejado nos dados, a representatividade pode ser uma armadilha. Dados representativos poderão conter também estereótipos, então talvez seja mais prudente pensar na representatividade como uma característica que não é absoluta. Podemos desejar um material que seja representativo da "vida real" em vários aspectos, mas talvez não em todos[14].
- **Documentação** – um bom *dataset* é um *dataset* bem documentado, que informa a origem do conteúdo textual, as classes de anotação e as diretrizes usadas para

[14]De um ponto de vista linguístico, a representatividade é uma característica bastante discutível. Se pensamos em *corpora* "gerais", de que representatividade estamos falando? Ser representativo de algo significa ser uma parte (uma amostra) que contém as principais propriedades e características do todo que ela representa. Quando nosso objeto é uma língua, sabemos qual é o todo? Até onde vai uma língua? No mundo do PLN, a representatividade está associada sobretudo às tarefas – que por sua vez estarão associadas a um ou mais tipos de texto – , o que torna a busca pela representatividade mais factível.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

anotar, as características e/ou formação dos anotadores. É por meio da documentação, por exemplo, que poderemos saber que o conteúdo de um *dataset* de análise de sentimento foi extraído do site "Reclame Aqui", e que talvez seja pouco indicado para aprendizado equilibrado de avaliações positivas e negativas. Como já mencionado, outro ponto cada vez mais importante diz respeito aos cuidados relativos à presença de viés indesejado nos dados. Uma vez que não é possível obter dados sem viés nenhum– línguas não são neutras, e uma língua constrói e dissemina visões de mundo e valores de uma comunidade linguística –, o material com que o PLN trabalha (enunciados reais proferidos por pessoas reais, e anotados ou revisados por pessoas reais) pode acabar incluindo e reproduzindo preconceitos. Com o objetivo de minimizar limitações científicas e éticas decorrentes de conjuntos de dados enviesados, tem sido proposto que se inclua na documentação de *datasets* informação detalhada a respeito dos falantes que produziram tais dados ou dos anotadores que analisaram e classificaram os dados em termos de *gênero*, *classe social*, *idade*, *etnia* e o que mais for possível obter (sempre respeitando questões de privacidade). Esta preocupação não é nova (Couillault et al., 2014), mas tem importância crescente no PLN (veja-se, mais recentemente, (Bender; Friedman, 2018)).

- **Tamanho** – por fim, o tamanho é um aspecto que precisa ser levado em conta. E por tamanho não me refiro apenas ao tamanho do *corpus* em *tokens* ou *palavras*, mas também à quantidade de fenômenos rotulados. Na anotação de PoS e de sintaxe, por exemplo, todas as palavras recebem alguma etiqueta; na anotação de entidades, polaridade ou correferência, por outro lado, apenas algumas palavras são classificadas, isto é, recebem uma etiqueta. Consequentemente, um *dataset* de anotação morfossintática poderá ser menor, em número de palavras, do que um de correferência ou de entidades. Ou seja, uma coisa é o número de palavras (ou de *tokens*), e outra o número de palavras (ou *tokens*) que recebem alguma etiqueta. Uma das características dos LLMs (grandes modelos de linguagem) é terem sido treinados com *datasets* gigantescos. Para conseguirem produzir material com o volume desejado e em quantidade de tempo razoável, grandes empresas fazem uso da anotação colaborativa (ou anotação *crowdsourcing*), que cada vez mais é alvo de críticas e preocupações, como veremos na Seção 16.4.5.

Por fim, como nem sempre teremos à disposição *datasets* com todas as características desejáveis, sobretudo no que se refere à variedade e tamanho, podem ser utilizadas diferentes técnicas para criar ou aumentar artificialmente o conjunto de dados linguísticos, como **aprendizado ativo**, no qual amostras de elementos críticos para o aprendizado são selecionadas para a anotação, guiando o aprendizado de pontos sensíveis e diminuindo a necessidade de exemplos e, consequentemente, o custo da anotação manual (Zhang et al., 2022) e (Settles, 2010). Outra possibilidade é a utilização de linguagem artificialmente gerada por LLMs para alimentar modelos. No entanto, esta é uma abordagem delicada, uma vez que a utilização de dados criados por LLMs para o treino e ajuste de outros modelos de IA, de forma recursiva, pode levar ao colapso do modelo (Shumailov et al., 2024). Neste sentido, dados e interações genuinamente humanos são recursos de grande valor. Um bom *dataset* é aquele que contém enunciados linguísticos proferidos por pessoas.

16.4 Por onde começar?

Para criar um *dataset*, precisaremos de

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

- um problema ou tarefa: por exemplo, classificar enunciados como ofensivos ou não; classificar palavras ou grupos de palavras como remetendo a PESSOAS etc.
- etiquetas de anotação (*tagset*), com as quais iremos classificar os dados;
- instruções sobre como classificar os dados, conhecidas como esquema ou diretrizes (*guidelines*) de anotação;
- compilação ou criação de um *corpus* adequado para esta tarefa;
- uma maneira de codificar a anotação;
- pessoas responsáveis pela anotação;
- uma ferramenta de anotação (se for o caso);
- estratégias para otimizar o processo de anotação (se for o caso);
- maneiras de avaliar o resultado da anotação, isto é de avaliar a qualidade do *dataset* produzido (embora não tenha a ver com "por onde começar?", avaliar a qualidade do material produzido é parte importante da preparação de *datasets*).

16.4.1 Definição do problema ou tarefa

O primeiro passo é ter clara a motivação para a criação do *dataset* – que problema ele se propõe a ajudar a resolver? Assim, é fundamental ter clareza sobre a motivação para a classificação dos dados: ao organizar os livros em uma estante, diferentes motivações (encontrar rapidamente os livros de que preciso ou tirar fotos de uma estante de livros para decoração) selecionam diferentes critérios para classificação (assunto, por um lado, e tamanho do dos livros, por outro), e levam a diferentes resultados – diferentes maneiras de dispor os livros na estante. Do mesmo modo, uma palavra como "Sofri" em "Sofri para terminar o livro" pode ser classificada como "negativa" se a motivação para a classificação é encontrar opiniões, e também pode ser classificada como um "verbo" se a motivação para a classificação é o comportamento morfológico. A sequência "Abri mão" em "Abri mão do prêmio" pode ser classificada como uma unidade se a motivação para a classificação é encontrar unidades de sentido (equivalente a "abdicar", por exemplo), mas considerada duas unidades se a motivação para a classificação é construir um corretor gramatical, já que será necessário verificar, na correção, se as várias formas do verbo "abrir" estão corretas ("eles abriram mão do prêmio", "abrirei mão do prêmio").

16.4.2 Conjunto de etiquetas e instruções: o esquema de anotação

A tarefa/problema bem definidos também são cruciais para o desenvolvimento das classes de anotação (as etiquetas, e também chamamos de *tagset* o conjunto de etiquetas), que podem ser poucas como "positivo" "negativo" ou "neutro", no caso de anotação de polaridades, ou mais de dez, como na anotação morfossintática.

Definir um conjunto de etiquetas e sua utilização refletem uma maneira de ver a tarefa. Por isso, quanto mais bem definido o problema (a tarefa), mais chances de sucesso. Caso seja necessário criar um esquema de anotação, devemos logo responder às seguintes perguntas: Qual o objetivo da anotação? A que ela serve?

Além disso, um esquema de anotação deve favorecer a generalização, mas sem perder informatividade. Esta generalização é o que fazemos quando, na frase (a), anotamos *nunca* e *tanto* como advérbios; *sofri* e *ler* como verbos; quando anotamos as frases (a) e (b) como frases positivas com relação a um alvo; ou quando anotamos, na frase (c) *sarampo*, *caxumba* e *rubéola* como DOENÇAS e *tríplice viral* como VACINA. Em todos os casos, embora cada uma das palavras ou frases comentadas (e anotadas) seja uma palavra/frase diferente, dizemos que elas são de um mesmo tipo – algumas são do tipo *advérbio*, outras do tipo *verbo*;

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

algumas do tipo *positivo*; outras do tipo *negativo*; algumas do tipo DOENÇA, outras do tipo VACINA. E, apesar de as igualarmos, mantemos informação relevante (para as tarefas). Poderíamos classificar, por exemplo, *tanto* e *livro* como palavras de um mesmo tipo, porque ambas terminam em *o*, mas esta classificação não carrega informação relevante; ou podíamos classificar *sarampo*, *caxumba*, *rubéola* e *tríplice viral* com palavras do campo da SAÚDE ou MEDICINA. Em ambos os casos, embora até estejamos favorecendo alguma generalização, propondo classes mais amplas, estamos perdendo informação.

a. Nunca$_{_ADV_[0]}$ sofri$_{_V_[-]}$ tanto$_{_ADV_[0]}$ para$_{_PREP_[0]}$ ler$_{_V_[0]}$ um$_{_ART_[0]}$ livro$_{_SUBST_[0]}$ [frase negativa]
b. Sério, sofri pra terminar o livro, o livro ainda consegue ser pior que os filmes. [frase negativa]
c. Sarampo[DOENÇA], caxumba[DOENÇA] e rubéola[DOENÇA] são combatidas pela tríplice=viral[VACINA].

Na definição de um esquema de anotação é fundamental uma rodada inicial de anotação para as primeiras observações e ajustes. Porque uma coisa é a teoria e, outra, o contato com os dados. Nessa primeira rodada de anotação, pode ser que classes que julgávamos claras não estejam tão claras assim quando as palavras estão em contexto. Ou pode ser que as classes sugeridas sejam insuficientes para dar conta do que o *corpus* apresenta, e então é necessário criar novas classes. Assim, ao longo de um processo de anotação, as etiquetas iniciais (provisórias) podem ser confirmadas, ou os dados podem levar à reformulação das categorias iniciais, e o processo de anotação recomeça.

O processo de anotação com refinamento do esquema de anotação segue as seguintes etapas:

- Levantamento bibliográfico sobre o que já existe relacionado à questão, em termos teóricos e aplicados: *Existem anotações do mesmo tipo, ou diretamente relacionadas? Existem datasets para tarefas similares?*
- Elaboração de um esquema de anotação que contenha as primeiras generalizações acerca do fenômeno observado, isto é, a primeira proposta de etiquetas (classes);
- Aplicação dessas etiquetas a uma amostra mais ampla;
- Refinamento progressivo do esquema de anotação.

As instruções podem ser complexas como um manual ou uma gramática (como a documentação do *treebank* Bosque[15], e da coleção dourada do HAREM[16], que codificam morfossintaxe, e entidades genéricas, respectivamente), ou podem conter apenas alguns parágrafos de explicação, como ilustrado no Quadro 16.2, que traz (uma versão traduzida) das instruções para a anotação do *corpus* SNLI (*Stanford Natural Language Inference*), que contém 560 mil pares de frases. Nesse caso, o problema (ou tarefa) associado ao *corpus* é a identificação de certos tipos de relação semântica entre duas frases. As relações de interesse são *acarretamento*, *contradição*, *neutra*. Essas, portanto, são as etiquetas atribuídas aos pares de frases. As instruções são interessantes porque não pedem que os anotadores classifiquem pares de frases com as referidas etiquetas, pede "apenas" que produzam frases a partir de certas instruções, e voltaremos a esse *corpus* (e essas instruções) na próxima seção.

[15] https://www.linguateca.pt/Floresta/BibliaFlorestal/.
[16] https://www.linguateca.pt/aval_conjunta/HAREM/directivas.html.

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

Quadro 16.2: Instruções de anotação do *corpus* SNLI (Bowman et al., 2015)

Vamos te mostrar a legenda de uma foto. Não vamos te mostrar a foto. Usando apenas a legenda e o seu conhecimento de mundo:

• Escreva uma legenda alternativa que seja definitivamente uma descrição verdadeira da foto. Exemplo: para a legenda "Dois cães correndo em um campo." você pode escrever "Existem animais ao ar livre".

• Escreva uma legenda alternativa que possa ser uma descrição verdadeira da foto. Exemplo: para a legenda "Dois cães correndo em um campo." você pode escrever "Alguns cachorros estão correndo para pegar um graveto."

• Escreva uma legenda alternativa seja definitivamente uma descrição falsa da foto. Exemplo: para a legenda "Dois cães correndo em um campo." você poderia escrever "Os animais de estimação estão sentados em um sofá". Isso é diferente da categoria talvez correta porque é impossivel para os cães correr e sentar.

16.4.3 Escolha do *corpus*

A escolha do *corpus* também está diretamente associada à tarefa – se o interesse está na detecção de opinião, é pouco indicado um *corpus* da Wikipédia, por exemplo, a não ser que o problema seja justamente a busca por opinião em textos que supostamente não deveriam conter opinião. Além disso, outras preocupações devem estar associadas à seleção do material:

- Tem direitos autorais ou pode ser usado (e disponibilizado) livremente?
- Leva em conta ou viola a privacidade de quem escreve (Capítulo PLN em Redes Sociais)?
- Como lidar com conteúdo duplicado ou repostado, comum na internet?
- Como é a qualidade do texto?
  - Produzido originalmente em formato eletrônico ou resultado de processamento por OCR?
  - Possui muitas imagens, gráficos, tabelas? Neste caso, como lidar com este material?
  - O texto será normalizado (tudo em minúsculas, por exemplo), ou será mantida a grafia original?

Também é possível que o material textual que compõe um *corpus* tenha sido especificamente produzido para o próprio *corpus*, como o exemplo do SNLI mostrou[17]. Outro exemplo de *dataset* em que o material textual é produzido pelas pessoas com a finalidade de criar o *dataset* é o SQuAD (*Stanford Question Answering Dataset*), criado para a tarefa de perguntas e respostas (para a língua inglesa). Na tarefa de criação do *corpus*, as pessoas receberam parágrafos de documentos da Wikipédia, e sobre este material deveriam formular 5 perguntas sobre o conteúdo, e respondê-las. Além disso, não era possível usar o recurso de copiar & colar, o que forçou as pessoas a usarem suas próprias palavras na

[17] Embora, no PLN, não haja qualquer problema em considerar o SNLI um *corpus*, é possível que, de um ponto de vista dos estudos linguísticos, este material seja discutível no que se refere à dimensão *naturalidade*, isto é, "enunciados naturalmente produzidos".

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

formulação das perguntas e das respostas. Nestes casos, a etapa de seleção do *corpus* deixa de existir, e é substituída pela etapa "Crie uma maneira engenhosa de produzir certos fenômenos linguísticos em grande escala". Do mesmo modo, não há anotação ou classificação propriamente, uma vez que os enunciados criados já "nascem" organizados conforme certas classes (do tipo *pergunta* e do tipo *resposta*, ou do tipo *contradição*).

16.4.4 Codificação

Na preparação de um *dataset* é preciso decidir o formato dos dados (txt, csv, tsv, JSON, XML ou outros formatos) e o formato da anotação propriamente. Muitas vezes, o formato é determinado pela ferramenta que se usa para anotar; outras vezes, a ferramenta é escolhida em função (também) dos arquivos que suporta.

Por exemplo, um formato de anotação bastante comum (para a anotação de entidades mencionadas, mas não só) é o formato IOB (Inside-Outside-Beginning), ou BIO, em que o B significa o início (*begin*) de uma entidade, o I (*in*) a continuação dela, e o O (*out*), indica que a palavra em questão não pertence à entidade. Por exemplo, em "Ana conheceu a Serra da Mantiqueira ontem", ou em "Sarampo e rubéola são combatidas pela tríplice viral" teríamos o seguinte, usando um *token* por linha:

Figura 16.2: Anotação no formato IOB para entidades mencionadas.

```
Ana             B-PESSOA
conheceu        O
a               O
Serra           B-LOCAL
da              I-LOCAL
Mantiqueira     I-LOCAL
ontem           B-TEMPO


Sarampo         B-DOENÇA
e               O
rubéola         B-DOENÇA
são             O
combatidas      O
pela            O
tríplice        B-VACINA
viral           I-VACINA
```

Já a anotação de dependências sintáticas, quando feita pela abordagem Universal Dependencies (Capítulo 4), segue o formato CoNLL-U, que por sua vez é um formato TSV (*tab-separated values*, isto é, valores separados por tabulação). São características deste formato:

- Cada frase tem um identificador;
- Na linha seguinte temos o texto da frase;
- Na linha seguinte começa a representação das palavras (*tokens*) da frase, com um *token* por linha;

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

- A separação entre uma frase e outra é feita por uma linha em branco;
- Cada *token* contém anotação em 10 campos separados por uma tabulação. Cada campo codifica diferentes informações morfossintáticas, e informações não preenchidas são marcadas com o caractere especial "_".

A Figura 16.3 mostra uma frase no formato CoNLLL-U[18].

Figura 16.3: Anotação no formato CoNLLL-U para dependências sintáticas.

```
# sent_id = 1
# text = Ana conheceu a Serra da Mantiqueira ontem.
1       Ana     Ana     PROPN   _       Gender=Fem|Number=Sing  2       nsubj   _       _
2       conheceu        conhecer        VERB    _       Mood=Ind|Number=Sing|Person=3|Tense=Past|VerbForm=Fin   0       root    _
3       a       o       DET     _       Definite=Def|Gender=Fem|Number=Sing|PronType=Art        4       det     _       _
4       Serra   Serra   PROPN   _       Gender=Fem|Number=Sing  2       obj     _       _
5-6     da      _       _       _       _       _       _       _       _
5       de      de      ADP     _       _       7       case    _       _
6       a       o       DET     _       Definite=Def|Gender=Fem|Number=Sing|PronType=Art        7       det     _       _
7       Mantiqueira     Mantiqueira     PROPN   _       Number=Sing     4       nmod    _       _
8       ontem   ontem   ADV     _       _       2       advmod  _       _
9       .       .       PUNCT   _       _       2       punct   _       _
```

Também é preciso decidir qual segmento de texto será anotado. Na preparação de *datasets* para detecção de desinformação e discurso de ódio, ou para análise de sentimento, por exemplo, é possível tanto uma anotação localizada, que classifica palavras e expressões ou, por outro lado, uma anotação que classifica segmentos maiores como frases ou documentos inteiros. Na Figura 16.4, temos um exemplo adaptado do *dataset* HateXplain[19], utilizado na tarefa de detecção de discurso de ódio (e que foi utilizado para avaliar a capacidade do modelo de linguagem GPT de identificar este tipo de discurso). A apresentação do material explica que cada postagem no *dataset* é anotada de três perspectivas diferentes: (i) a classificação em 3 classes "frequentemente usadas para este tipo de anotação" ("ódio", "ofensivo" ou "normal"), (ii) a comunidade-alvo (ou seja, a comunidade que foi vítima de discurso de ódio/discurso ofensivo na postagem) e (iii) as justificativas para a classificação, ou seja, as partes do texto que justificam a decisão de anotar algo como *odioso*, *ofensivo* ou *normal*. A Figura 16.4 traz um exemplo (traduzido e adaptado) do material, que está no formato JSON. Cada campo codifica o seguinte:

- Id: identificador único de cada postagem
- Anotadores: traz as anotações produzidas por cada anotador
- Label: indica a etiqueta (ou classe) atribuída a cada postagem. Os valores possíveis ódio (0), normal (1) ou ofensivo (2)
- [anotador_id]: o identificador exclusivo atribuído a cada anotador
- [target]: O alvo da postagem (no caso, adaptamos para os "extraterrestres", dentre os quais estão os marcianos)
- justificativas: os elementos do texto (os *tokens*) selecionados como justificativa para a etiqueta atribuída pelos anotadores. Cada justificativa representa uma lista com valores 0 ou 1. Um valor de 1 significa que o *token* faz parte da justificativa selecionada pelo anotador. Para obter o *token* específico, podemos usar a mesma posição de índice em "post_tokens"
- post_tokens : A lista de *tokens* representando a postagem que foi anotada.

[18] Mais informações sobre o formato em https://universaldependencies.org/format.html.

[19] https://huggingface.co/datasets/hatexplain#dataset-structure e https://github.com/hate-alert/HateXplain.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Figura 16.4: Anotação de discurso de ódio no formato JSON.

```
{
    "id": "24198545_gab",
    "anotadores": [
      {
        "label": 0,  # ódio
        "anotador_id": 4,
        "target": ["extraterrestres"]
      },
      {
        "label": 0,  # ódio
        "anotador_id": 3,
        "target": ["extraterrestres"]
      },
      {
        "label": 2,  # ofensivo
        "anotador_id": 5,
        "target": ["extraterrestres"]
      }
    ],

    "justificativas":[
    [0,0,0,1,1,1],
    [0,0,0,1,0,0],
    [0,0,0,1,0,0]

    ],

    "post_tokens":
    ["marcianos","são","o","câncer","dessa","nação"]
  }
```

Como podemos ver pela Figura 16.4, diferentes porções do texto (diferentes *tokens*) foram selecionadas como justificativas para a classificação geral (ódio, ofensivo ou normal). Além disso, a mesma postagem foi considerada "discurso de ódio" para duas pessoas (anotadores 4 e 3), mas apenas "discurso ofensivo" para uma delas. E com isso passamos a uma dimensão fundamental da anotação: as pessoas que anotam.

16.4.5 Anotação: sabedoria de especialistas ou sabedoria da multidão?

A anotação é valiosa porque codifica o entendimento humano sobre alguma coisa, mas esta maneira de apresentar a anotação dá a entender que o "entendimento humano" sobre mundo é homogêneo, o que sabemos não ser verdade (e a divergência das interpretações sobre a classificação da frase dos marcianos é um pequeno exemplo). Além disso, nem todas as pessoas têm o conhecimento necessário para fazer todos os tipos de anotação.

Quando mencionamos anotadores **especialistas**, a referida especialidade pode ser de diferentes naturezas: conhecimento dos conceitos linguísticos (por exemplo, de classes de palavras, anáfora, acarretamento) e/ou do domínio (por exemplo, medicina em um *corpus* de Medicina) e/ou da tarefa em questão (anotar certos elementos usando certas ferramentas e seguindo certas instruções de anotação).

Inicialmente, boa parte das anotações utilizadas no PLN tinha como fonte ou inspiração

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

os níveis de análise linguística ou alguma teoria linguística: morfologia, sintaxe, semântica, pragmática, papéis semânticos. E, por demandarem este tipo de conhecimento especializado, linguistas sempre foram responsáveis pela anotação. Mas isto não quer dizer que linguistas estão aptos a fazer todo e qualquer tipo de anotação, justamente porque anotações (ou boa parte delas) precisam ser feitas por especialistas, e linguistas são especialistas em linguagem. Há anotações que, idealmente, precisarão do apoio ou supervisão de outros profissionais, como em projetos de anotação e/ou de criação de *datasets* de áreas como medicina, direito, genética, geologia, dentre tantas outras.

Ao lado de tarefas que demandam conhecimento especializado, há tarefas que demandam "apenas" capacidade de leitura e fluência de escrita, como atribuição de orientação semântica (positiva ou negativa) a enunciados, identificação de entidades genéricas como PESSOA, ORGANIZAÇÃO, ou a criação de frases de determinados tipos, como no *corpus* SNLI, apresentado na Seção 16.4.2. E tarefas de PLN que no fim dos anos 2000 seriam vistas com descrédito por serem complexas e precisarem de uma quantidade brutal de dados são realidade 20 anos depois. A criação dos já referidos *datasets* SNLI e SQuaD são exemplos dessa tendência, apoiada por grandes empresas que pagam (mal) para que pessoas do mundo todo produzam dados de treino para as máquinas. Trata-se da *anotação colaborativa*, ou anotação feita por "trabalhadores de multidão" ou "microtrabalhadores" (*crowdworkers*). Anotações colaborativas envolvem um grande número de anotadores e possibilitam produzir *datasets* enormes em uma quantidade de tempo relativamente baixa.

Apesar de possibilitar a geração de dados em larga escala, esta forma de anotar é alvo de muitas críticas, que vão desde a falta de comprometimento dos anotadores com a tarefa, remuneração baixa e exposição de quem trabalha a material com conteúdo repulsivo (por exemplo, classificar postagens em tarefa de detecção de discurso de ódio) (veja também Capítulo 20) até a crítica de que o que se chama de inteligência artificial é o fruto de milhares de trabalhadores invisíveis e precarizados[20].

A falta de comprometimento com a tarefa pode levar a características indesejadas nos *datasets*, que por sua vez irão impactar tanto o que é aprendido quanto a avaliação da tarefa. O estudo de Gururangan et al. (2018), por exemplo, mostrou que os anotadores do *corpus* SNLI usaram estratégias bastante previsíveis, como o emprego de (i) advérbios de negação para criar frases com *contradição*, ou de (ii) hiperônimos (relação entre *animal* e *gato*, por exemplo) para frases com *acarretamento* (as instruções para a tarefa estão no Quadro 16.2, na Seção 16.4.2). Em consequência disso, a previsibilidade tornou a tarefa artificialmente fácil para as máquinas, levando a números de desempenho enganosamente altos.

A Amazon Mechanical Turk (AMT)[21] é a mais antiga plataforma para este tipo de anotação, e ficou mais conhecida pelo público pelas reportagens sobre quem são e as condições desumanas a que estão submetidos os "anotadores", chamados *turkeys*. Admitindo que muito do que tem sido gerado pelas IAs tem como fonte os dados produzidos por estas pessoas, não chega a ser um grande exagero pensar que elas são as representantes do nosso senso comum.

Assim, outra crítica a este tipo de trabalho é que, se na imensa maioria das vezes o

[20] A intensa divulgação e popularização de grandes modelos de linguagem tem chamado a atenção para a maneira pela qual estes *datasets* são construídos, veja-se http://www.uol.com.br/tilt/reportagens-especiais/a-vida-dura-de-quem-treina-inteligencias-artificiais/, https://www.bbc.com/portuguese/geral-49234093 e http://time.com/6247678/openai-chatgpt-kenya-workers/ .

[21] O nome inspirado no "Turco Mecânico", um robô jogador de xadrez do século 18 que os adversários enfrentavam pensando estar competindo contra uma máquina, quando na verdade havia um mestre de xadrez escondido lá dentro https://www.bbc.com/portuguese/geral-49234093.

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

trabalho envolve anotações que codificam um certo senso comum, não sabemos exatamente de quem é este senso comum. A decisão sobre classificar um determinado enunciado como discurso de "ódio"*,* "ofensivo" ou "normal" (tarefa codificada no *dataset* HateXplain e em muitos outros do mesmo tipo), por exemplo, é feita com base em alguns exemplos fornecidos pela empresa responsável pela anotação. Mas será que, em temas como esses, o que desejamos é a codificação do "senso comum"? Como garantir que a anotação não vai justamente reforçar e amplificar, uma vez que provavelmente irá alimentar modelos de linguagem, o comportamento que deveria detectar e suprimir? Vejamos a apresentação do HateXplain, traduzida abaixo:

> Antes de iniciar a anotação, os anotadores são explicitamente avisados de que a tarefa exibe algum conteúdo de ódio ou ofensivo. Preparamos instruções que explicam claramente o objetivo da tarefa de anotação, como anotar os segmentos de texto e também incluímos uma definição para cada categoria. Fornecemos vários exemplos com classificação, comunidade-alvo e segmentação das anotações para ajudar os anotadores a entender a tarefa.

A apresentação faz parecer que a identificação de algo como "discurso de ódio" (em oposição a "discurso ofensivo", ou "normal") é trivial. A manifestação de discurso de ódio, no Brasil, é crime previsto por lei[22], mas os limites entre discurso de ódio e liberdade de expressão são alvo de discussão inclusive no meio jurídico. É razoável confiar no senso comum, ou na sabedoria da multidão, para indicar às máquinas aquilo que especialistas estão debatendo?

E voltamos à relevância da participação de especialistas nos processos de anotação. Se queremos codificar conhecimento, precisamos **qualificar** este conhecimento. Se estamos tentando fazer as máquinas nos ajudarem a classificar um determinado conteúdo, e esta classificação tem implicações jurídicas e limites pouco definidos, é importante que especialistas (do direito e dos direitos humanos, por exemplo) tomem parte no processo. E, neste caso específico, é uma maneira de restringir apenas a especialistas o contato com conteúdo sensível. Sem contar o óbvio: quanto melhor a curadoria, isto é, quanto melhor a classificação dos dados, melhor a qualidade das previsões que serão feitas e menos dados são necessários para um bom desempenho.

Por fim, quando menciono especialistas não-linguistas não quero dizer que o conhecimento linguístico formalizado só é útil em projetos de anotação linguística explícita, isto é, em projetos que envolvem conhecimento de teorias linguísticas. Diferentes tipos de anotação podem se beneficiar se o material já contém alguma anotação linguística anterior, e o ideal é que profissionais de diferentes especialidades colaborem. A anotação linguística, mesmo a mais simples como PoS, já é uma primeira organização dos dados textuais brutos, já codifica informação. O que difere a anotação linguística – de PoS, sintaxe, semântica etc. – das demais é que, por serem genéricas e fornecerem um primeiro nível de organização nos dados, facilitam e impulsionam outros tipos de anotação, como veremos na Seção 16.5.

16.4.6 Ferramentas de anotação

É importante contar com ferramentas que auxiliem o processo de anotação e revisão. Devido à expansão da atividade de anotação para além de tarefas estritamente linguísticas,

[22]Veja-se esta página do Ministério Público Federal dedicada ao tema: https://respeiteadiferenca.mpf.mp.br/www/discurso-odio.html.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

há um grande número de ferramentas, algumas gratuitas, criadas para este fim. Em geral, a escolha da ferramenta é influenciada pela natureza da anotação ou da tarefa, e pontos para se levar em conta são:

- o formato dos arquivos de entrada e de saída;
- se a ferramenta suporta receber textos com outros tipos de anotação, ou apenas o texto cru que será anotado;
- se é possível corrigir (ou editar) anotações usando regras/padrões linguísticos, ou a revisão/edição só pode ser feita caso a caso;
- se há diferentes maneiras de visualizar os textos que serão anotados;
- se há opção de atalhos usando teclado (precisar clicar para anotar, apesar de aparentemente mais fácil, é pouco produtivo para a maioria das pessoas que anotam);
- se é possível trabalhar online ou apenas localmente;
- se a ferramenta suporta diferentes anotadores trabalhando simultaneamente;
- se a ferramenta é customizável;
- o tempo de familiarização necessário para começar a usar.

Em geral, teremos sempre uma tensão entre uma ferramenta com várias funcionalidades, mas cuja utilização é mais difícil, e uma ferramenta mais fácil de usar, mas que oferece menos funcionalidades. Alguns exemplos de ferramentas são

- BRAT[23]: para anotação de entidades mencionadas, correferência, dependências sintáticas, entre outras.
- Label Studio[24]: para anotação de entidades mencionadas, perguntas e respostas, análise de sentimento, entre outras.
- Inception[25]: para diversos tipos de anotação semântica e discursiva.
- WebAnno[26]: para anotação morfológica, sintática e semântica.
- Arborator[27]: para anotar e buscar dependências sintáticas em formato UD.
- ET: dois ambientes integrados para buscar, revisar e avaliar dependências sintáticas em formato UD: Interrogatório[28], para buscar e revisar anotações e Julgamento[29], para avaliar .

16.4.7 Estratégias de anotação

Um *dataset*/*corpus* padrão ouro pode ser construído (i) de maneira totalmente manual por uma única pessoa, por grupo pequeno ou por centenas ou milhares de pessoas, ou (ii) de maneira híbrida, quando é feita uma primeira rodada de anotação automática que depois é revista por pessoas[30]. Este será o tema da Seção 16.5.

[23] https://brat.nlplab.org/
[24] https://labelstud.io/
[25] https://inception-project.github.io/
[26] https://webanno.github.io/webanno/
[27] https://arborator.icmc.usp.br
[28] https://github.com/alvelvis/Interrogat-rio
[29] https://github.com/alvelvis/Julgamento
[30] Já quando falamos de um *corpus* anotado, além dessas duas maneiras, existe a possibilidade de uma anotação 100% automática.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

16.4.8 Formas de avaliação

Quem garante que um determinado *dataset* é bom, isto é, que os dados que contém são confiáveis, diversificados, representativos, codificados de maneira consistente e adequados à tarefa? Cada tarefa tem suas especificidades, e é avaliada de uma maneira. Este será o tema da Seção 16.6.

16.5 Procedimentos e estratégias de anotação e revisão

Depois de definido o problema ou a tarefa, escolhido o *corpus*, o esquema de anotação, a codificação, a ferramenta e as pessoas responsáveis pela anotação – e nada impede que a anotação seja feita por uma única pessoa, ainda que esta não seja a situação ideal – é hora de anotar propriamente.

Como vimos, diferentes práticas podem ser consideradas anotação:

- Ler (ou ouvir) um enunciado e atribuir uma etiqueta ao enunciado todo ou a partes dele;
- Produzir uma frase (enunciados em geral) a partir de determinadas instruções;
- Produzir perguntas e respostas a partir de um texto.

O foco desta seção está nas anotações do primeiro tipo, e em enunciados escritos. Quando a anotação consiste em atribuir uma etiqueta a um elemento do texto, é possível começar usando uma lista de palavras vindas de um léxico ou recursos lexicais (Capítulo 4 e (Freitas, 2022)). A língua portuguesa dispõe de alguns, para diferentes tarefas[31].

16.5.1 Palavras, regras e padrões linguísticos na construção de um *corpus* padrão ouro

Quando a anotação parte de uma lista de palavras (um léxico), a cada vez que uma palavra de interesse é encontrada, ela recebe uma etiqueta (ou mais de uma, caso o *esquema de anotação* preveja isso).

Por exemplo, o Emocionário[32] contém uma lista de palavras que, em algum contexto, descrevem algum tipo de emoção e/ou sentimento da língua portuguesa[33]. O léxico tem o formato de uma lista de lemas associados a uma classe de emoção/sentimento.

Mas, como as palavras podem assumir sentidos diferentes conforme o contexto em que estão sendo usadas, listas de palavras, sozinhas, são insuficientes. Será necessária uma **revisão** da classificação inicial das palavras (da anotação) para corrigir os erros. Léxicos são sempre ótimos pontos de partida, mas não são eficientes para uma boa anotação. Por exemplo, quando usamos o léxico do Emocionário para anotar um texto, o verbo *chocar* é anotado com a emoção SURPRESA nas duas frases abaixo, mas apenas a ocorrência 2 remete à SURPRESA, e deveria ter sido anotada.

1. Após a queda, sua moto ***chocou***-se com o muro e pegou fogo.
2. O mundo ***chocou***-se com a notícia da morte do navegador.

[31] Ver GitHub do Brasileiras em PLN: https://github.com/brasileiras-pln
[32] https://www.linguateca.pt/acesso/corpos/dicemocoes.iso.txt
[33] Para mais informações, veja https://www.linguateca.pt/Gramateca/Emocionario.html.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Por isso, após a aplicação do léxico, ficamos diante da necessidade de revisão. Tanto a revisão quanto a anotação podem se beneficiar se o *corpus* já contiver uma análise linguística prévia: quanto mais informação (anotação) disponível no *corpus*, mais otimizada pode ser a revisão. Para a língua portuguesa, podemos contar com modelos de anotação morfossintática com desempenho bastante satisfatório, e esta anotação prévia pode ajudar o trabalho de revisão semântica se estamos diante de anotadores humanos especializados em conhecimentos linguísticos/morfossintáticos. Voltando aos exemplos 1 e 2, contar com uma **anotação prévia** e com uma *ferramenta adequada* permite listar os sujeitos do verbo "chocar", facilitando a análise dos casos corretos e dos que precisam ser corrigidos. Ou seja, ao ler uma lista com palavras como "mundo", "moto", "caminhão", "país", "padre", avião", "asteróide" e "Maria" e já sabendo que estas são palavras que estão sendo usadas como *sujeito* do verbo "chocar" ("mundo se chocou", "moto se chocou", "país se chocou" etc), podemos dividir as palavras em três grupos, como no Quadro 16.3.

Quadro 16.3: Distribuição de palavras que exercem a função de *sujeito* do verbo "chocar"

Grupo 1 = surpresa	**Grupo 2 = colisão**	**Grupo 3 = dúvida**
mundo	moto	Maria
país	caminhão	
	avião	
	asteróide	

Das 7 ocorrências listadas, precisaremos ler o contexto apenas daquelas listadas no grupo 3, já que pessoas podem ficar *surpresas* (grupo1) mas também podem *colidir* com outras pessoas ou objetos (grupo 2). Após a análise do contexto e decisão sobre os casos de dúvida, precisaremos eliminar a etiqueta SURPRESA quando os sujeitos forem aqueles listados nas palavras do Grupo 2. Com uma ferramenta adequada, podemos fazer todas essas alterações de uma única vez. Será preciso escrever uma regra, verificar se ela faz o que era esperado, e aplicá-la ao *corpus*. A regra poderia ser algo do tipo:

```
token.lemma = "palavras_do_grupo2" and token.deprel == "nsubj"
and token.head_token.lemma == "chocar" >> token.emo == "0"
```

que utiliza informação morfossintática (codificada nos atributos lemma e deprel, que indica a função sintática). Ou seja, a regra diz[34]:

Se

> O lema do *token* é alguma das palavras do grupo 2 (informação do atributo *lemma*) E
>
> A função sintática do *token* é sujeito (informação do atributo *deprel*, e *nsubj* é o código para a função de *sujeito*) E
>
> O "pai" do sujeito é o lema "chocar" (informação do atributo *token.head_token.lemma*)

Então

[34] O exemplo é baseado na utilização da ferramenta Interrogatório https://github.com/alvelvis/Interrogatorio.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

A anotação de emoção do *token* será zerada (informação do atributo *emo*)

E por que não usar esta forma de trabalhar, baseada em regras, para todo o PLN? Isto é, se o interesse está na anotação de emoção, por que precisamos de um *dataset* para treinar e criar um modelo, por que não podemos fazer tudo por meio de regras? Acontece que esta forma de trabalhar é bem pouco prática/eficaz para textos diferentes daqueles que originaram as regras de revisão, e também para *corpora* realmente grandes. Para o "mundo controlado"[35] do *dataset*, abordagens baseadas em léxicos e regras são uma boa estratégia, mas nem tudo poderá ser (bem) resolvido por regras, e isto se deve a uma propriedade das línguas: independentemente do tipo de texto e do assunto tratado, sempre haverá um número imenso de fenômenos que não se repetem, e para os quais não teremos regras[36].

Quando observamos a distribuição das palavras em um *corpus*, sempre veremos muitas palavras com baixa ocorrência – muitas palavras que ocorrem 5 vezes, ainda mais palavras que ocorrem 4 vezes, ainda mais palavras que ocorrem 3 vezes, ainda mais palavras com ocorrência 2, e ainda mais palavras com apenas uma única ocorrência. Temos uma imensa proporção de palavras do *corpus* que são ocorrências singulares, de casos que ocorrem apenas uma vez. Este fenômeno tem um nome: *hapax legomenon* (*hapax legomena*, no plural), termo que vem do grego e que significa "sendo dito uma vez".

Esta distribuição segue um padrão: poucos casos com muitas ocorrências, um número intermediário de casos com frequência média, e um número enorme de casos de frequência baixa. Além disso, quanto menor a frequência, mais palavras compartilham essa mesma frequência. Por isso há mais palavras de frequência 1 do que palavras de frequência 2, mais palavras de frequência 2 do que palavras de frequência 3 etc. Em *corpora* grandes, 40% a 60% das palavras são *hapax legomena* e outros 10% a 15% são *dis legomena* (ocorrem 2 vezes). E esta distribuição não se aplica apenas a palavras isoladas, mas a qualquer fenômeno, como a distribuição das entidades em um *corpus*, ou a estrutura dos sintagmas nominais. Este tipo de fenômeno é capturado pela lei de Zipf, uma lei empírica formulada no contexto da Linguística, e assim nomeada devido ao trabalho do linguista George Kingsley Zipf (1902–1950).

A Figura 16.5 mostra a distribuição das palavras do livro Ulisses, de James Joyce, e vemos que a frequência de uma palavra é inversamente proporcional à sua posição no *ranking* de frequências (isto é, a palavra na posição 1 tem frequência muito maior que a palavra na posição 10 mil). E vemos que o padrão se repete na Figura 16.6, que mostra a distribuição das 10 milhões de palavras mais frequentes em 30 wikipédias, isto é, em 30 línguas diferentes.

Quando ficamos cientes desta propriedade distribucional, conseguimos entender melhor por que as regras linguísticas são importantes, mas até um certo ponto: as regras capturam regularidades, e regularidades só existem se existe repetição. Por outro lado, se descartamos os casos com frequência 2 ou 1 (que não se repetem, portanto), estaremos olhando para uma língua mutilada, da qual uma imensa parte foi dispensada. Como uma boa parcela da língua não se repete, é difícil, apenas com regras, conseguir generalizar fenômenos. Assim, não importa a quantidade colossal de dados que tenhamos à disposição, sempre teremos uma proporção enorme de casos raros e não previstos. Por isso, abordagens baseadas em regras serão limitadas, e por isso também o aprendizado de máquina

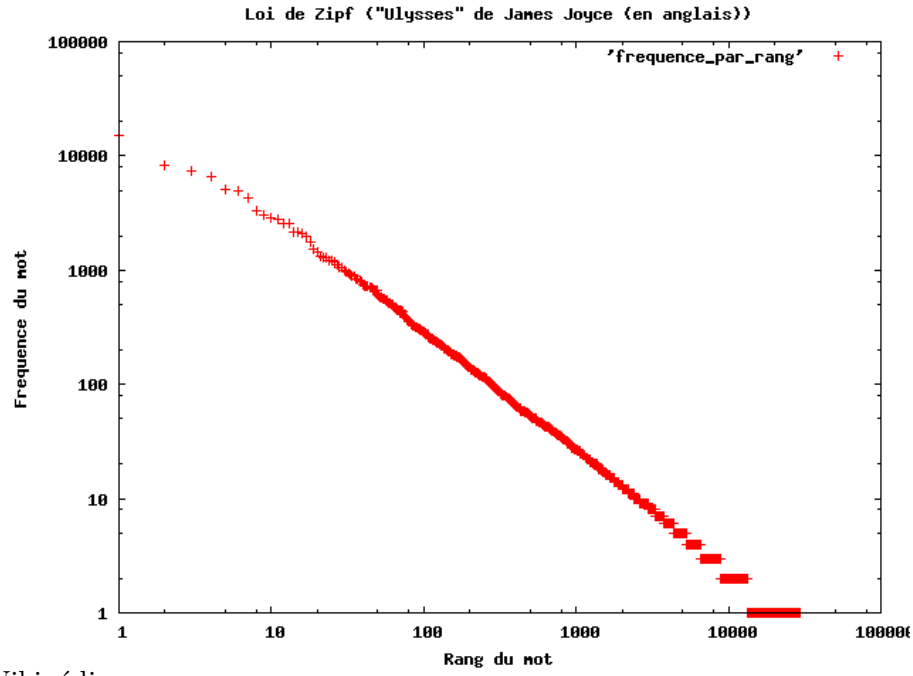
[35] "Controlado" porque sabemos onde começa e onde termina, isto é, é um conjunto finito de enunciados linguísticos.

[36] E não custa lembrar que os anos iniciais do PLN foram dominados pela crença de que o processamento linguístico poderia ser inteiramente resolvido por meio de regras linguísticas.

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

Figura 16.5: Distribuição das palavras do livro Ulisses, de James Joyce.



Fonte: Wikipédia

apresenta bons resultados: desde que haja dados, algoritmos estão cada vez melhores em prever eventos raros. Apesar da ressalva, o combinado léxico & regras continua sendo uma ótima maneira de construir *datasets* e *corpora* padrão ouro, considerando a alternativa de ler cada frase isoladamente, e anotar cada fenômeno de uma vez. A anotação de entidades no DANTEStocks (Piai, 2025), de papéis semânticos no Portinari-Base PropBank (Freitas; Pardo, 2025) e de entidades de geologia e petróleo no PetroNER (Freitas et al., 2023) seguiu esta maneira de trabalhar, que tem a vantagem de não precisar de uma equipe grande de pessoas anotando.

16.5.2 Revisão de anotação

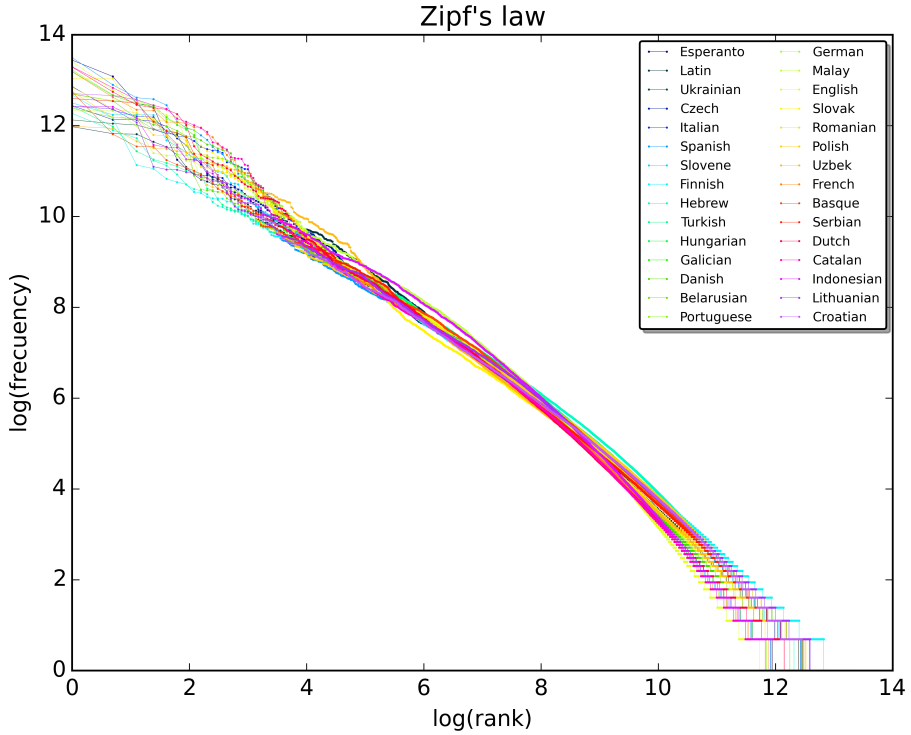
A revisão de uma anotação pode ser motivada pelo reconhecimento de que aquilo que se considerava padrão ouro, no fim das contas, não era tão ouro assim, ou porque estamos diante de um *corpus* novo, de um gênero ou domínio novos, o que acaba trazendo novos desafios para a anotação ou, como vimos na seção anterior, ou como parte de uma estratégia de anotação.

O que estas situações têm em comum é já contarem com uma anotação, ainda que de má qualidade. Por isso, o que precisamos neste momento é de **estratégias** para encontrar erros ou inconsistências na anotação. Um caminho óbvio é ler cada frase no *corpus*, e esse pode ser o primeiro passo se este é o primeiro contato com o *corpus* e conforme o tipo de anotação. Mas, como única abordagem para a detecção de erros, é uma estratégia limitada, por dois motivos: (i) uma mesma frase pode conter erros de diferentes naturezas, o que tem como consequência dificuldade em manter o foco e a consistência da revisão, deixando o processo de revisão mais suscetível a erros e mais demorado; (ii) conforme o tipo de anotação que está em jogo, há fenômenos que não costumam trazer erros – o

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Figura 16.6: Distribuição das 10 milhões de palavras mais frequentes em 30 wikipédias.



Fonte: Wikipédia

reconhecimento de certas formas como artigos, verbos, advérbios, sujeitos, por exemplo – e rever cada unidade da frase envolve rever também estes casos mais fáceis. Ou seja, analisar *token* a *token* (ou palavra por palavra) também pode ser um desperdício de energia, em *certos casos.* Wallis (2003), por exemplo, recomenda trocar uma revisão linear, *token* a *token*, por uma **revisão transversal**, que permitiria ver os fenômenos em questão de forma ampliada e garantiria uma revisão consistente, como ilustramos na seção anterior com o exemplo de *chocar* e SURPRESA. Por isso – porque analisar linearmente, frase a frase, é pouco eficaz – uma etapa importante do processo de revisão consiste em encontrar uma maneira de detectar erros e corrigi-los.

O trabalho de revisão de anotação pode se tornar um imenso labirinto, e para chegar até o (padrão) ouro precisaremos de um mapa e de um guia que nos ajudem a encontrar os erros ou inconsistências: um caminho que seja, de preferência, rápido e seguro, isto é, que nos permita encontrar erros no menor tempo possível sem abrir mão da qualidade. A preocupação é dupla: do ponto de vista do resultado, buscamos análises consistentes e adequadas – uma anotação padrão ouro –; do ponto de vista do processo de revisão, buscamos eficiência – chegar aos melhores resultados consumindo pouca energia, isto é, em pouco tempo. O estudo de (Freitas; Souza, 2024) avalia diferentes abordagens para a revisão de *corpus*, e mostra que utilizar divergências entre anotações automáticas como ponto de partida para a revisão (assumindo que as divergências são casos suspeitos) economiza tempo e garante uma revisão de alto nível.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

16.5.3 Até onde precisamos rever?

Quem já trabalhou com revisão de *corpus* sabe que o trabalho parece não ter fim – e por isso também é importante termos estratégias para conduzir e finalizar a anotação.

Do ponto de vista do aprendizado de máquina, a presença de alguns erros aleatórios não chega a prejudicar justamente porque não há o que ser aprendido, já que não será possível generalizar a partir de erros aleatórios. Mas o fato de um *dataset* ruidoso não influenciar a generalização não significa, necessariamente, que não seja necessário torná-lo o melhor possível. Em primeiro lugar, porque nem todas as ferramentas ou tarefas precisam estar associadas ao paradigma de aprendizado automático. Em segundo lugar, e mais importante, porque se quisermos que o *dataset* produzido também sirva para avaliação, a presença de erros será sempre um problema, uma vez que pode penalizar sistemas injustamente caso a análise automática esteja correta, mas a anotação padrão ouro esteja errada.

De todo modo, no ambiente de aprendizado automático, temos visto que é possível aprender com dados um pouco ruidosos, e que uma anotação de alta qualidade não se reflete, necessariamente, em facilidade de generalização.

O trabalho de Souza; Freitas (2023) traz alguns dados para informar a discussão no que se refere à língua portuguesa. Na preparação do *treebank* PetroGold v3[37], cada rodada de revisões foi acompanhada de uma rodada de "avaliação de aprendizagem" (uma *avaliação intrínseca*, como veremos na Seção 16.6) com o objetivo de verificar o impacto das melhorias linguísticas na aprendizagem automática. Os resultados estão na Tabela 16.1[38]:

Tabela 16.1: Evolução de um *corpus* comparando quantidade de revisões e melhoria no aprendizado de máquina.

	***Tokens* revistos**	**F1**
V1	–	88,53
V2	8.802	88,82
V3	9.314	90,22

Pela Tabela 16.1, vemos que apesar do esforço de revisão, o impacto na aprendizagem é limitado, sobretudo entre as versões 2 e 3. No fim das contas, um *corpus* maravilhosamente anotado leva a um desempenho muito semelhante àquele treinado em um *corpus* "apenas" bem anotado. No entanto, levar um melhor desempenho não é garantia absoluta de um *dataset* melhor, como veremos na próxima seção.

16.6 Como avaliar a qualidade do *dataset*?

Na Seção 16.2, vimos o que são e qual a relevância das avaliações conjuntas, que avaliam modelos e ferramentas a partir de um mesmo conjunto de dados/*dataset*. Nesta seção, o foco está em avaliar a qualidade dos *datasets*, já que um *dataset* de baixa qualidade não será capaz de dar suporte a uma avaliação confiável, seja ou não uma avaliação conjunta, e irá gerar um modelo de linguagem (ou previsões) de baixa qualidade[39].

Cada tarefa de PLN tem seus próprios métodos de avaliação. No entanto, quando tratamos de *datasets* anotados, alguns elementos da avaliação são comuns às diferentes

[37] https://petroles.puc-rio.ai/files/Corpora/petrogold-v3.zip
[38] O trabalho está detalhadamente descrito em (Souza, 2023) e (Souza; Freitas, 2023).
[39] Veja mais sobre Avaliação de sistemas de PLN no Capítulo 18.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

tarefas. Uma vez que a anotação pode ser considerada uma tarefa de classificação, a avaliação também é feita nesses moldes.

Para avaliar um modelo ou ferramenta de classificação é comum utilizar as medidas de *precisão* e abrangência. A **precisão** mede (avalia) se a classificação que foi feita está correta (se a palavra analisada como *verbo* é realmente um *verbo*, ou se um comentário classificado como *ofensivo* é realmente *ofensivo*). A **abrangência** mede (avalia) se tudo o que deveria ter sido encontrado (e classificado) foi encontrado e classificado corretamente (se todos os verbos ou comentários ofensivos foram encontrados). A precisão mede a *qualidade* das classificações realizadas; a abrangência mede a qualidade da quantidade de elementos classificados, isto é, indica se tudo aquilo que deveria ter sido encontrado foi, de fato, encontrado.

Para calcular precisão, abrangência e a medida F (que é uma média harmônica entre ambas), classificamos os resultados da seguinte maneira:

- verdadeiro positivo (VP): o elemento foi detectado pela análise automática e foi classificado de forma correta.
- verdadeiro negativo (VN): o elemento foi detectado pela análise automática, mas foi classificado de forma errada.
- falso positivo (FP): o elemento foi detectado pela análise automática, mas não deveria.
- falso negativo (FN): o elemento não foi detectado pela análise automática, mas deveria.

Para calcular a precisão fazemos:

$$Precisão = \frac{VP}{VP + FP}$$

Para calcular a abrangência fazemos:

$$Abrangência = \frac{VP}{VP + FN}$$

Para calcular a medida F fazemos

$$F = 2 * \frac{Precisão * Abrangência}{Precisão + Abrangência}$$

Uma ferramenta pode ser muito precisa – todas as classificações que ela faz são corretas – e pode, igualmente, ter uma baixa abrangência – apesar de acertar bastante, há muitos casos que ficam de fora. Em geral, há uma tensão entre essas duas medidas: se afrouxamos a abrangência para encontrar mais casos, podemos diminuir a precisão, trazendo muitos casos errados. E, tentando melhorar a precisão, corremos o risco de perder em abrangência. Por isso, um bom desempenho se reflete em um equilíbrio entre essas medidas, e esta é a proposta da **medida F**: indicar em um único número uma combinação entre precisão e abrangência que reflita o desempenho geral.

Usamos **medidas de F1, precisão e abrangência para avaliar modelos e ferramentas** quanto à capacidade de generalizar a partir dos dados a que foram expostos no treinamento. Mas o quanto os dados permitem essa generalização? A capacidade de generalizar a partir dos dados está associada aos algoritmos utilizados, mas *datasets* também têm um papel nessa história – para o bem e para o mal –, pois algoritmos não fazem mágica.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

16.6.1 Concordância entre-anotadores

No que se refere a *datasets* com dados anotados, além de bem documentado, de tamanho suficiente, variado e adequado à tarefa, um bom *dataset* é aquele no qual as anotações foram feitas de maneira consistente.

Nos *datasets* criados por meio de anotação *crowdsourcing*, algumas estratégias usadas para garantir a consistência são a revisão das análises por outros anotadores (e apenas aquelas anotações em que não foi necessária correção são consideradas) e o descarte de respostas ou análises desviantes (estratégia que vem sendo cada vez mais criticada, como veremos na (Seção 16.6.2)).

Na anotação feita por equipes menores, em geral compostas por especialistas, a consistência de uma anotação é avaliada por meio da concordância entre anotadores (ou **concordância inter-anotadores**), que nada mais é do que a comparação entre duas ou mais anotações (análises) humanas. Quanto maior o índice de concordância, isto é, quanto mais convergência entre as análises, mais confiáveis elas são. Na concordância entre anotadores, a ideia de uma anotação *correta* é substituída pela ideia de uma anotação *consistente* (todos os anotadores analisaram os fenômenos da mesma maneira). O raciocínio subjacente é este: se diferentes pessoas, seguindo as mesmas instruções (esquema e documentação de anotação), analisaram um fenômeno da mesma maneira, esta análise é confiável.

Levando em conta o trabalho envolvido na anotação, é comum que a verificação da concordância seja feita utilizando apenas uma amostra do *corpus*, isto é, temos as mesmas pessoas anotando os mesmos textos apenas em um subconjunto do *corpus*. O resultado da comparação destas anotações é um número que nos diz o grau de confiança que podemos ter nas análises/anotações; que nos diz o quanto as análises/anotações convergiram (ou divergiram), e a partir dele temos uma estimativa de o quão consistente está a anotação no *corpus* todo. Uma alta concordância entre anotadores é indicativa do potencial de reprodutibilidade, isto é, da possibilidade de reprodução das análises por outras pessoas (e espera-se que as máquinas sejam capazes de reproduzir estas mesmas análises).

Quando os resultados da concordância entre anotadores indicam uma baixa consistência entre as análises, é possível explorar algumas alternativas:

- Melhoria das instruções de anotação, como a inclusão de exemplos, tanto positivos ("faça assim nesses casos") quanto negativos ("não faça assim nesses casos");
- Aumento do tempo de familiarização com a tarefa e com o fenômeno sendo analisado, para que as pessoas responsáveis pela anotação estejam seguras do que estão fazendo;
- Reformulação das classes de anotação (do *tagset*). Os resultados da concordância entre anotadores podem ser baixos porque o conjunto de classes não está bem desenhado/modelado, mesmo que ele corresponda às classes de uma teoria.

A prática de medir a concordância entre anotadores é comum não apenas para avaliar a qualidade e consistência da anotação de um *dataset*, mas também em outras tarefas que visam avaliar a qualidade (e confiança) da análise humana. Este é o caso, por exemplo, da **anotação de erros de tradução automática**, uma das maneiras de avaliar a qualidade de uma tradução automática (Subseção Métricas Humanas).

Algumas medidas utilizadas para o cálculo da concordância entre anotadores foram propostas por Cohen (1960), Fleiss (1971) e Krippendorff (1970). O Cohen's Kappa foi originalmente projetado para medir a concordância entre apenas dois anotadores. Se a intenção é calcular a concordância entre múltiplos anotadores, é possível então levar em conta a média de cada par possível de anotadores. Já o Fleiss' Kappa pode ser utilizado

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

com mais de dois anotadores, e o Alfa de Krippendorff é flexível para ser aplicado em cenários com múltiplos anotadores, permitindo considerar diferentes níveis de desacordo.

A interpretação dos valores de Kappa irá depender da tarefa avaliada. Em termos gerais, Cohen sugere que valores 0 indicam nenhuma concordância, 0.01–0.20 indicam concordância fraca a nula, 0.21–0.40 indicam concordância razoável, 0.41–0.60 indicam concordância moderada, 0.61–0.80 indicam concordância substancial e 0.81–1.00 indicam uma concordância quase perfeita.

Por exemplo, na anotação semântica que visa a desambiguação de sentidos (Capítulo 8), as medidas de concordância costumam ser baixas, em torno de 0.7. Já na anotação sintática, que aparentemente seria mais complexa, é comum números superiores a 0.9.

Por fim, existem diferentes maneiras de lidar com as discordâncias, partindo da ideia de que alguns tipos de divergência entre análises humanas são mais graves (ou mais inofensivos) do que outros. O coeficiente Kappa de Cohen, por exemplo, pode ser usado na versão ponderada ou não ponderada. A diferença entre **Kappa ponderado** (*weighted Kappa*) e **Kappa não ponderado** (*non-weighted Kappa*) está na maneira como eles lidam com as discordâncias: o Kappa não ponderado trata todos os pares de categorias/etiquetas da mesma forma, ou seja, todas as discrepâncias são tratadas igualmente, independentemente de sua natureza ou importância. Por outro lado, o Kappa ponderado atribui diferentes pesos às discordâncias entre as categorias/etiquetas, baseando-se em uma escala que reflete a importância atribuída a essas categorias. Por exemplo, em uma escala Likert de 1 a 5, o Kappa não ponderado considera que, no cálculo uma concordância (mais precisamente, de uma discordância), os valores 1, 2, 3, 4 e 5 são equivalentes, isto é, são, todos eles, indicativos de discordância. Já o Kappa ponderado reconhece que apesar dos valores 1, 2, 3, 4 e 5 indicarem a presença de uma discordância, o valor 1 está mais próximo de 2 do que 5, indicando que a discordância entre 1 e 5 é mais grave ou relevante que uma discordância entre 1 e 2.

16.6.2 Quando a discordância não é erro: divergências legítimas

Como vimos, o coeficiente Kappa é uma das principais maneiras de medir a concordância entre anotadores. Uma alta concordância – isto é, várias interpretações humanas convergentes – é certamente uma maneira válida de avaliar a confiança que podemos ter nestas interpretações. No entanto, embora esta seja uma prática comum, não há necessidade de restringir a concordância a uma única interpretação – podemos concordar, simultaneamente, sobre mais de uma interpretação. Vejamos os exemplos do Quadro 16.4.

Quadro 16.4: Exemplos de frases com trechos que levaram à discordância entre anotadores

1. A equipe gaúcha **perdeu** a oportunidade de manter a distância para o líder ao ser derrotada pela Chapecoense por 1 a 0, em casa.
2. Relata sua amizade com Janot e afirma que o ex-procurador-geral **chamava** Dodge de "bruxa" em conversas reservadas

Na frase 1, podemos interpretar o trecho sublinhado como CAUSA (a equipe perdeu a oportunidade *porque* foi derrotada), TEMPO (a equipe perdeu a oportunidade *quando* foi derrotada) e MANEIRA (*Como* a equipe perdeu a oportunidade?), sem que as leituras sejam excludentes. Do mesmo modo, na frase 2, podemos interpretar o trecho sublinhado como EVENTO ou MANEIRA (chamava, *reservadamente*, Dodge de "bruxa")

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Um estudo com anotadores experientes mostrou exatamente esta variação na interpretação (Freitas; Pardo, 2025). O reconhecimento desta "divergência legítima" no processo de anotação não é novidade Basile et al. (2021) e trabalhos mais recentes questionam o "mito" do *ground truth* (a "verdade fundamental") na anotação, reconhecendo que incorporar uma ampla gama de perspectivas no processo de anotação, ao invés de considerar a variação um ruído, é fundamental não apenas para capturar a variabilidade nos julgamentos humanos (Meer et al., 2024), mas também para minimizar a exclusão de vozes minoritárias (Leonardelli et al., 2021). Esta é uma tendência que ganha força no momento em que estão em foco a anotação de elementos discursivos e pragmáticos, na qual a multiplicidade de interpretações é mais bem aceita que na anotação de elementos morfossintáticos (veja-se também o movimento *NLP Perspectives*[40]). Segundo Basile et al. (2021), *datasets* com anotações múltiplas estão se tornando cada vez mais comuns, assim como métodos capazes de integrar a discordância na modelagem.

No que se refere ao PLN de língua portuguesa, um dos primeiros esquemas de anotação que admitem anotação múltipla como alternativa para lidar com as diferentes interpretações/classificações que uma mesma palavra/entidade, em um mesmo contexto, pode assumir, é o utilizado nas duas edições do HAREM, avaliação conjunta de entidades mencionadas.

Ao admitir a classificação/anotação múltipla em um esquema de anotação, teremos que enfrentar dois desafios relacionados: (i) como distinguir anotações/interpretações divergentes, mas legítimas (como aquelas atribuídas aos exemplos 1 e 2 do Quadro 16.4), daquelas que são fruto de lapso/falta de atenção, diretivas mal feitas ou esquemas de anotação mal desenhados, e (ii) como calcular a concordância entre anotadores, levando em conta (i). Marchal et al. (2022), por exemplo, propõem uma adaptação da medida k no contexto de anotações discursivas.

De qualquer maneira, em diversos domínios e aplicações, é necessário que as análises (respostas) sejam precisas, sem qualquer traço de vagueza ou polissemia.

16.6.3 Avaliação intrínseca

Por meio de uma avaliação da concordância entre anotadores ficamos sabendo que o material está consistente do ponto de vista da análise humana. Porém, não temos muitas pistas sobre o quanto esta análise humana – ou, o quanto o *dataset* (a combinação de análise humana + dados) – permite generalizar. Além disso, consistência na amostra utilizada no cálculo da concordância entre anotadores não significa, necessariamente, que não haja lapsos na anotação ao longo do material.

No PLN, **avaliação intrínseca** é, tradicionalmente, uma avaliação dos modelos e ferramentas (não do *dataset*), e é calculada utilizando medidas de precisão e abrangência apresentadas no início desta seção. Como o nome indica, este tipo de avaliação é intrínseca à tarefa que está sendo avaliada. Ou seja, se temos uma ferramenta/modelo que faz anotação de sintaxe, ou de entidades, a ferramenta/modelo será avaliada nesta tarefa (parece óbvio, mas pode não ser assim, como veremos na **avaliação extrínseca**). Considerando a geração de um modelo de aprendizado de máquina, para que seja possível realizar uma avaliação intrínseca é importante que o material tenha um *tamanho* que permita todo o ciclo do aprendizado: treino, validação e teste. A partição de *teste* é uma parte do *dataset*, e apenas esta parte será alvo da avaliação (Capítulo Modelos de linguagem).

[40] https://nlperspectives.di.unito.it/perspectivist-approaches-to-nlp/

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Quando pensamos na avaliação intrínseca de *datasets* (e não dos modelos/ferramentas), estamos fazendo uma *inversão*, ou uma mudança de perspectiva. Se na avaliação intrínseca "original" verificamos a capacidade do modelo ou ferramenta de generalizar a partir dos dados a que foram expostos, na avaliação intrínseca de *datasets* verificamos (indiretamente) o quanto o *dataset* (mais especificamente, a natureza dos textos + a classificação dos dados) permitiu esta generalização, levando em conta as características do modelo/ferramenta. A partir dessa mudança de perspectiva, quando olhamos para o desempenho de um modelo e vemos números de precisão, abrangência e medida F, não vemos apenas uma avaliação do modelo ou ferramenta, ou o quão bom o modelo/ferramenta é, mas também até onde os dados permitiram ir, considerando os limites do modelo/ferramenta. Podemos entender da seguinte maneira: avaliação intrínseca é uma **análise dos erros** de um modelo/ferramenta que pressupõe que o material que serviu de treino está bem anotado e que o modelo/algoritmo tem um desempenho que não é aleatório. Na Tabela 16.2, por exemplo, vemos o resultado de uma avaliação intrínseca de um modelo/ferramenta que tem um desempenho geral de 89,09%, um desempenho muito bom, portanto. O que vemos na tabela, no entanto, é o desempenho de cada classe individualmente (considerando um *tagset* com 30 classes/etiquetas).

Tabela 16.2: Desempenho de um modelo (desempenho global de 89,09%) para cada classe aprendida

Etiqueta	**Qtde**	**F1**	**Etiqueta**	**Qtde**	**F1**	**Etiqueta**	**Qtde**	**F1**
classe 1	1786	99,27%	classe 11	314	89,81%	classe 21	564	74,47%
classe 2	101	99,01%	classe 12	346	89,02%	classe 22	185	71,89%
classe 3	1920	98,80%	classe 13	1269	88,02%	classe 23	78	62,82%
classe 4	178	96,07%	classe 14	140	86,43%	classe 24	29	62,07%
classe 5	447	94,85%	classe 15	166	86,14%	classe 25	92	61,96%
classe 6	319	93,73%	classe 16	1375	84,73%	classe 26	435	61,84%
classe 7	129	93,02%	classe 17	154	83,77%	classe 27	115	60,00%
classe 8	332	92,77%	classe 18	308	80,19%	classe 28	81	55,56%
classe 9	248	92,34%	classe 19	213	77,93%	classe 29	36	41,67%
classe 10	666	90,09%	classe 20	110	77,27%	classe 30	14	78,57%

Na tabela, vemos que as classes 14 a 30 foram as mais difíceis, e que as classes/etiquetas 23 a 29 foram especialmente difíceis. Vemos também que, exceto pela classe 26, as classes 23 a 30 são menos frequentes (na partição teste) que as demais. Assim, quando trocamos de perspectiva e olhamos para o desempenho de cada classe individualmente, e não para o desempenho global do modelo, pensamos: como fazer para melhorar a classificação prevista nesses casos críticos, sem piorar os demais? O que esses casos têm de especial? Acrescentar mais exemplos para cada classe resolve? Ou o problema está nas instruções de anotação, que não deixam claro como exatamente classificar em certos casos? Ou não há o que fazer, e o problema de generalização é da ferramenta/modelo?

A limitação desta maneira de avaliação é, justamente, a impossibilidade de diferenciar uma dificuldade de generalização oriunda do algoritmo de aprendizado de uma dificuldade de generalização oriunda de análises inconsistentes ou da falta de exemplos suficientes no material padrão ouro. Por isso, é importante garantir que há **consistência** na análise humana, o que vemos na concordância entre anotadores. Outra limitação desta forma de avaliar é que ela se restringe à análise da partição 'teste' de um *dataset*, uma partição pequena.

Por outro lado, uma das vantagens deste tipo de avaliação – que joga luz sobre os *datasets*, e não sobre os modelos – é a possibilidade de perceber, de forma localizada, os

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

obstáculos para generalização, e então atuar de forma direcionada para construir *datasets* "otimizados" para tarefas. Trata-se de uma forma de avaliar a qualidade da anotação que sinaliza onde há espaço para melhoria no conjunto de dados, dialogando com estratégias de aprendizado ativo.

Esta é uma abordagem que nos permite uma visão simultaneamente quantitativa e qualitativa dos resultados de uma análise automática, capaz de guiar intervenções e melhorias onde de fato elas são necessárias. Além disso, repetindo o ciclo de aprendizagem após as intervenções é possível verificar se a solução proposta (inclusão de mais exemplos, redefinição das classes, melhoria das instruções de anotação) realmente facilitou a generalização. Ou seja, esta é uma abordagem não só para avaliar, mas que permite também medir o quanto as revisões melhoram um *corpus*, e nos ajuda a decidir um ponto de corte na revisão, respondendo à pergunta "Até onde precisamos rever?", da Seção 16.5.3.

A segunda vantagem desta abordagem já foi esboçada nos parágrafos anteriores: uma avaliação intrínseca de *datasets* permite verificar o resultado de experimentações na classificação dos dados; ela ajuda e informa a tomada de decisões no processo de anotação. Se há duas ou mais maneiras legítimas e adequadas de analisar um mesmo fenômeno e o esquema de anotação só permite usar uma classe (anotação *single label*), qual escolher? Facilitar a generalização deve ser um critério relevante na escolha, no contexto do PLN. Por exemplo, na frase exemplo (1) do início deste capítulo ("TRISTE é uma palavra de 6 letras"), a palavra TRISTE foi analisada como um *substantivo* (e não como *adjetivo*), mas essa é uma escolha de anotação: decidir se as classes de palavras são propriedades das palavras (e, portanto, estáticas, e nesse caso TRISTE seria um *adjetivo* em qualquer contexto) ou se as classes são atribuídas em função do papel que exercem na frase (e, portanto, dinâmicas, e nesse caso TRISTE seria um *substantivo* no contexto da frase) deveria ser fruto de uma escolha que leva em conta as tarefas e demandas do PLN (Capítulo 4). Qual maneira de classificar leva à maior generalização sobre a anotação de PoS? E qual maneira de classificar leva ao melhor desempenho em uma outra tarefa – na análise sintática, por exemplo? Esta segunda pergunta se relaciona à *avaliação extrínseca*, tema da próxima seção.

16.6.4 Avaliação extrínseca

A avaliação extrínseca verifica se a informação codificada no *dataset* melhorou o desempenho em tarefas consideradas mais complexas. Por isso, está associada à **adequação**, e não à consistência. Assim como na avaliação intrínseca, esta é uma forma de avaliar aplicada tradicionalmente a modelos ou ferramentas, mas *datasets* também podem ser avaliados dessa maneira (novamente fazendo uma mudança de perspectiva e tomando o ponto de vista dos dados/*dataset*, e do que permitem generalizar). A avaliação é extrínseca porque ela não avalia aquilo que, diretamente, o modelo/ferramenta faz, ou aquilo que o *dataset* codifica. Ela avalia indiretamente, verificando o quanto outras tarefas são beneficiadas quando aquilo que o *dataset* codifica é levado em conta. Ou: uma avaliação extrínseca verifica o quanto a informação codificada em um *dataset* é adequada para as tarefas mais complexas que o *dataset* pretende auxiliar. Por trás dessa ideia, o reconhecimento de que, no PLN, a codificação de certos tipos de informação (principalmente informação linguística especializada, como sintaxe, semântica etc.) é um meio, e não um fim, para resolver as tarefas de PLN. Ou seja, um *dataset* sintático, para tarefas de PLN, interessa não porque codifica informação sintática, mas porque, dispondo de informação sintática, é possível

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

melhorar o desempenho em outras tarefas, como extração de informação[41] (Capítulo Extração de Informação).

O estudo de Nooralahzadeh; Øvrelid (2018), por exemplo, estava interessado em verificar a contribuição de três tipos de gramáticas diferentes (ou, de três esquemas de anotação) na tarefa de extração de relações semânticas entre entidades, com o objetivo de saber qual o tipo de anotação gramatical (ou, de representação gramatical) mais adequado à tarefa. Ou seja, o estudo queria saber qual gramática ajudaria mais o modelo/ferramenta a chegar aos melhores resultados na extração de relações entre entidades. Para tanto, dispunham de um modelo que utilizava, como entrada, informação de dependências sintáticas[42] (Capítulo 4), e prepararam *datasets* anotados conforme as três abordagens gramaticais: as chamadas dependências sintáticas de Stanford, as dependências sintáticas do projeto Universal Dependencies e dependências sintáticas usadas nas tarefas CONLL[43] (ou seja, exatamente os mesmos textos anotados, a única diferença era o esquema de anotação sintática utilizado em cada um).

Voltando ao exemplo da palavra TRISTE, podemos imaginar dois *datasets* de PoS, que contém exatamente os mesmos documentos, mas em um deles as classes são atribuídas independentemente do contexto (e "triste" será sempre um adjetivo); e no outro as classes são atribuídas conforme o papel que exercem em cada frase. Não é difícil prever que a opção pelas classes estáticas ("triste" é sempre um adjetivo) será de mais fácil generalização, e levará a um desempenho melhor em uma avaliação intrínseca. Mas será que essa é a melhor opção se desejamos "aprender" informação sintática?

Em resumo, a avaliação intrínseca verifica a consistência interna da anotação (o que não necessariamente é sinônimo de qualidade, embora, em geral, seja), mas nada nos diz sobre a adequação do *dataset* para uma tarefa específica, e aqui a avaliação extrínseca pode ser útil. E nem sempre um *dataset* que obtêm melhor desempenho na avaliação intrínseca leva aos melhores resultados quando em uma avaliação extrínseca.

No entanto, a avaliação extrínseca só pode ser realizada se a) dispomos do *dataset* para a "tarefa subsequente" – no estudo de Nooralahzadeh; Øvrelid (2018) foi utilizado o *dataset* de uma *avaliação conjunta* de extração de relações entre entidades; b) dispomos de diferentes versões do mesmo *corpus*, que variam apenas quanto ao esquema de anotação (as etiquetas e maneira de utilizá-las): diferentes tagsets de PoS; diferentes representações sintáticas, diferentes representações de papéis semânticos etc.

Para a língua portuguesa, dispomos – ainda que de maneira tímida – de alguns *corpora* com diferentes anotações de um mesmo tipo de atributo, como vemos na Tabela 16.3, que lista *corpora* que são fruto de projetos de pesquisa acadêmica, e estão disponíveis para uso. MacMorpho[44] e Bosque, os mais antigos, são os que contém mais variações, e para o Bosque-UD[45] também há uma versão em que os casos de omissão do sujeito (sujeito oculto, nos termos da Gramática Tradicional) foram explicitados (o material foi criado com o objetivo de verificar o quanto a explicitação de sujeitos nas frases é capaz de

[41] Para os estudos linguísticos a anotação sintática pode ser um fim, quando o interesse está em estudar aspectos sintáticos da língua.

[42] Importante notar que a tarefa não pressupõe a existência de uma anotação sintática anterior.

[43] Considerando o foco deste capítulo, basta sabermos que são três maneiras diferentes de codificar a informação sintática, e para detalhes é possível consultar o artigo original.

[44] O *corpus* MacMorpho e a documentação estão disponíveis em http://nilc.icmc.usp.br/macmorpho/ e o MacMorpho-UD em https://github.com/own-pt/macmorpho-UD.

[45] O Bosque anotado com a abordagem UD está disponível em https://universaldependencies.org/treebanks/pt_bosque/index.html. Demais versões estão disponíveis em https://www.linguateca.pt/Floresta/levantamento.html

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

facilitar o processamento linguístico ou o quanto a omissão do sujeito – característica da língua portuguesa, mas não da língua inglesa – dificulta o processamento automático do português)[46].

Tabela 16.3: Variações de atributos linguísticos em *corpora* do português

ATRIBUTO	VARIAÇÃO	*CORPUS*	REFERÊNCIA
PoS	*Tagset* LácioWeb clássico	MacMorpho	(Aluísio et al., 2003)
	Tagset LácioWeb modificado	MacMorpho	(Fonseca; Rosa, 2013)
	UD	MacMorpho	(Freitas et al., 2018)
		PetroGold	(Souza; Freitas, 2023)
		Bosque-UD	(Rademaker et al., 2017)
	PALAVRAS/Floresta Sintá(c)tica	Bosque	(Freitas et al., 2008b)
Sintaxe	UD	PetroGold	(Souza; Freitas, 2023)
		Bosque-UD	(Rademaker et al., 2017)
		Bosque-UD com sujeitos ocultos explicitados	(Freitas; Souza, 2021)
	UD - Petrolês	PetroGold	(Souza; Freitas, 2023)

Qual dessas maneiras de representar o conhecimento linguístico é melhor? Aliás, melhor para que?

Em termos de avaliação intrínseca, precisamos garantir condições semelhantes de avaliação: mesmo algoritmo/ferramenta e mesmo conjunto de treino e teste. Todos os *corpora*/*datasets* citados estão públicos.

Quanto à adequação à tarefa, a avaliação extrínseca é uma forma interessante de avaliar a qualidade de um *dataset*, mas para que seja possível utilizá-la é fundamental a existência de um segundo *dataset*, que codifica a tarefa mais complexa. E voltamos ao que foi mencionado no início deste capítulo: a relevância de *datasets* para o avanço do PLN.

16.7 Em resumo...

- *Datasets* são importantes para o PLN porque permitem avaliar, treinar e pautar os rumos do PLN. (Seção 16.2)
- *Datasets* bons são consistentes (boa avaliação intrínseca), bem documentados, de tamanho adequado e com dados adequados. (Seção 16.3)
- A anotação de um *dataset* não precisa ser 100% perfeita para funcionar no aprendizado de máquina. (Seção 16.5.3)
- A preparação de um *dataset* se beneficia se a revisão/anotação é feita de maneira sistemática. (Seção 16.5)
- Avaliação intrínseca não significa, necessariamente, melhor adequação à tarefa, que é medida pela avaliação extrínseca. (Seção 16.6)

Um bom projeto cuidadoso de anotação, por sua vez, deverá levar em conta (Seção 16.4):

- Clareza quanto ao fenômeno que será anotado (que se reflete em um bom esquema de anotação);

[46] A preparação do *corpus* está descrita em (Freitas; Souza, 2021) e o *corpus* está disponível em https://github.com/alvelvis/desocultando-sujeitos

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

- Escolha do *corpus* adequado (um *corpus* composto por relatórios de pesquisa é pouco adequado para a anotação de ironia, por exemplo);
- Conhecimento linguístico – para identificação e descrição do fenômeno anotado;
- Conhecimento do problema – para um melhor recorte das classes;
- Uma boa dose de inspiração ou criatividade – para chegar ao equilíbrio em termos de granularidade e generalização;
- Um outro tanto de experimentação – para validação e reformulação das classes, se for o caso;
- Verificação quanto à eficiência da anotação (por exemplo, o tempo levado e o nível de treinamento/conhecimento necessário por parte de quem vai anotar);
- Infraestrutura adequada – diretivas, documentação e ferramenta;
- Avaliação (verificação da concordância entre os anotadores).

A avaliação de um *dataset* pode levar em conta diferentes perspectivas (Seção 16.6):

- A concordância inter-anotadores avalia a consistência das anotações humanas;
- A avaliação intrínseca também avalia a consistência, mas privilegia a capacidade de generalizar a partir dos dados e a possibilidades de experimentações na classificação dos dados, tendo em vista o aprendizado de máquina;
- A avaliação extrínseca leva em conta a adequação de uma anotação para uma determinada tarefa.

Agradecimentos

Agradeço muitíssimo à Helena Caseli e à Graça Nunes pelas sugestões e provocações que deixaram o capítulo bem melhor (espero!).

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Capítulo 17

Aprendizado Transdutivo em PLN

Nataly Leopoldina Patti da Silva
Norton Trevisan Roman

Publicado em: 16/04/2026

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

17.1 Introdução

Suponha que você deseja avaliar como um determinado comportamento linguístico influencia alguma tarefa. Tal avaliação pode ser, por exemplo, como o conteúdo de notícias influencia o resultado de determinada empresa, como comentários em uma rede social influenciam a popularidade de uma pessoa, ou mesmo como resenhas em um sítio de comércio eletrônico influenciam vendas futuras de um determinado produto.

Nesse sentido, uma das primeiras decisões a serem tomadas é o que corresponde ao seu conjunto de dados, seu *corpus* (Capítulo 16). Serão notícias? Textos de plataformas de *microblogging*? Textos científicos? Comentários feitos por usuários? Enfim, há uma infinidade de opções. Uma vez definido com o que se deseja trabalhar, há que se explicitar o que se deseja identificar nesses dados. A polaridade do texto como um todo? A estrutura sintática de cada sentença que compõe o texto (como exemplificado no Capítulo 7)? Veja que, a depender do que se deseja identificar, a unidade básica de anotação, ou seja, a menor porção do texto à qual será atribuído um rótulo, é também definida.

Sabendo já que fenômenos e quais tipos de dados são de interesse, resta definir sua fonte, ou seja, iniciar a busca de tais dados. Se você conseguir uma base de dados que seja grande o suficiente para seus propósitos, além de já estar anotada com o fenômeno de interesse, considere-se uma pessoa de extrema sorte. Infelizmente, como toda sorte extrema, ela também é rara. Então o mais comum é você se deparar com nada ou, na melhor das hipóteses, com um *corpus* já coletado, mas não anotado com o que interessa a você.

Que fazer então? Se o *corpus* existir e for de uso público, simples... obtenha-o. Mas, e se não existir? Então você deve criá-lo, seja via *crawlers* de internet, seja por geração automática ou *crowdsourcing*. Veja que você já não tirou a sorte grande, ou seja, seu *corpus* não está perfeitamente ajustado aos seus interesses. Nesse caso, você se encontra em uma das seguintes situações, ilustradas na Figura 17.1:

1. Você conseguiu um *corpus* anotado, mas de tamanho insuficiente ou parcialmente anotado;
2. Você conseguiu um *corpus* de tamanho suficiente, mas não anotado; ou
3. Você ficou sem nada nas mãos (seu pior pesadelo!).

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Figura 17.1: Cenários de Disponibilidade de *Corpora*. Os desafios na busca por um conjunto de dados incluem: encontrar um conjunto de dados anotado, porém insuficiente (esquerda); encontrar um conjunto de dados não anotado (centro); e não encontrar nenhum conjunto de dados disponível (direita).



Como exemplo, imagine que você se interessa por uma tarefa de classificação de *tweets*. Nesse caso, todas as situações acima são possíveis de acontecer, vista a popularidade desse tipo de dado. No primeiro caso, você pode ter tido sorte e encontrado uma base de dados anotada com os rótulos do seu interesse (por exemplo, a base de dados **DANTEStocks**, descrita em (Di Felippo et al., 2024)). No segundo caso, você pode ter encontrado a base de dados ou implementado um *crawler* pra coletar os *tweets* da sua preferência, mas os rótulos não existem. No pior cenário, você não possui nem mesmo os *tweets* de interesse.

No caso de haver um *corpus* anotado disponível, mas pequeno, uma possibilidade é ampliá-lo, coletando mais dados de modo a formar assim um *corpus* grande o suficiente, mas parcialmente anotado (a parte correspondente ao *corpus* originalmente anotado). O segundo caso já exige que você anote os dados como um todo, o que pode ser uma tarefa extremamente custosa e demorada (Alzubaidi et al., 2023; Zhu et al., 2009). O terceiro caso une o pior dos mundos: você terá que construir o *corpus* do zero e então anotá-lo.

Que fazer? Se sua sorte não estiver tão ruim, é possível que existam ferramentas para anotação automática do *corpus*, como as descritas no Capítulo 7. Restaria então analisar uma amostra dessa anotação para se obter uma estimativa do desempenho da ferramenta em seu *corpus*. Agora, se você não tiver acesso a uma ferramenta, é possível anotar parte do *corpus*, de tamanho tal que a anotação possa ser feita dentro de um prazo e custo aceitáveis, e então apelar a métodos de aprendizado semi-supervisionado e, mais especificamente, ao aprendizado transdutivo, usando-os para incorporar os dados não rotulados ao processo de treinamento dos modelos supervisionados usados em seu projeto.

17.2 E o que seria esse tal de aprendizado transdutivo?

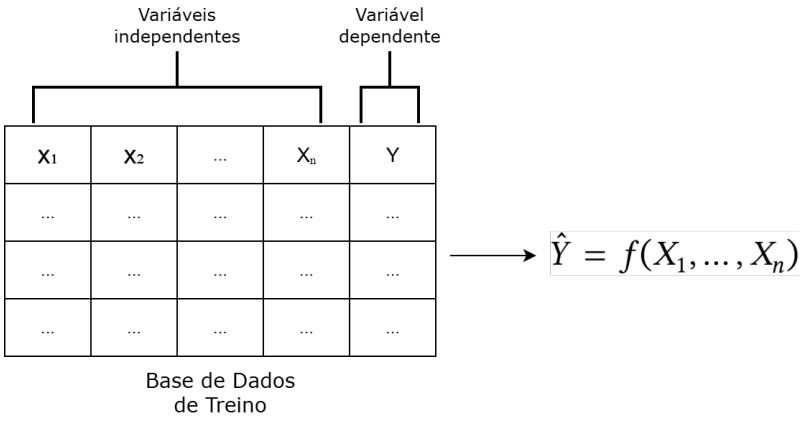
Para entender o aprendizado transdutivo, é necessário antes entender a diferença entre dedução e indução como formas de inferência, e como essas são usadas dentro do aprendizado supervisionado (ver (Russell; Norvig, 2009)). Na **indução**, uma relação (em geral, uma função) entre variáveis é derivada a partir dos próprios dados (Vapnik, 2000). Assim, ao vermos, por exemplo, determinados valores para um conjunto de variáveis $\{X_1, \ldots, X_n\}$ em conjunção com determinados valores para uma variável Y, induzimos uma relação $\hat{Y} = f(X_1, \ldots, X_n)$ entre elas[1], conforme ilustra a Figura 17.2.

[1]Usamos a notação $\hat{Y}$ para denotar que $\hat{Y}$ é uma estimativa do valor de Y.

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

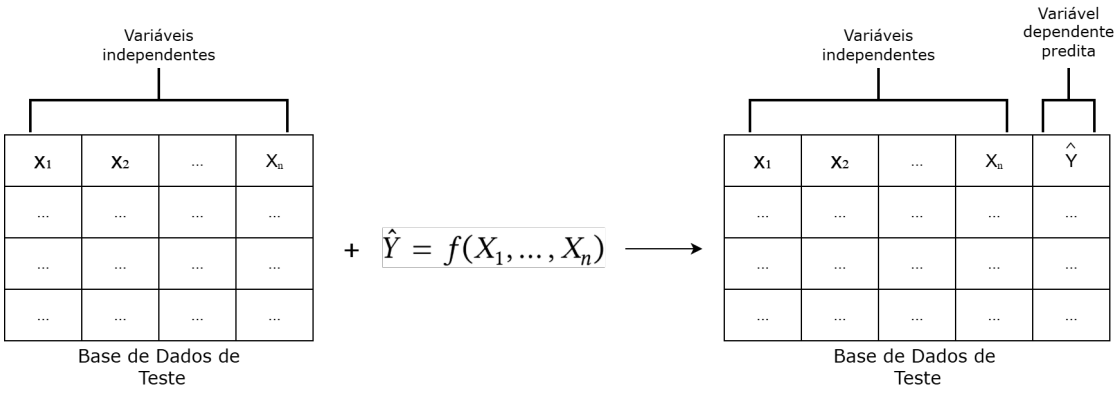
https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

Figura 17.2: Processo de Indução. A partir de um conjunto de observações para um grupo de variáveis preditoras $(X_1, \ldots, X_n)$ em conjunção com a variável alvo (Y), o modelo infere uma relação matemática $(\hat{Y})$ entre elas.



Na **dedução**, por sua vez, partimos de uma relação pré-definida entre as variáveis para, a partir do valor de uma ou mais variáveis, inferir o valor de alguma outra variável de interesse para um determinado ponto ou conjunto de pontos (Vapnik, 2000). Ou seja, a partir da relação $\hat{Y} = f(X_1, \ldots, X_n)$, e do conhecimento do valor de $\{X_1, \ldots, X_n\}$, deduzimos o valor de Y ou, no caso, apresentamos uma estimativa $\hat{Y}$ para Y (Figura 17.3).

Figura 17.3: Processo de Dedução. A partir de um conjunto de observações para um grupo de variáveis preditoras $(X_1, \ldots, X_n)$, correspondentes a exemplos de dados não rotulados, em conjunção com uma relação matemática entre elas e a variável alvo desejada $(\hat{Y} = f(X_1, \ldots, X_n))$, o modelo gera estimativas $\hat{Y}$ para os rótulos reais (desconhecidos) Y de cada exemplo não rotulado.



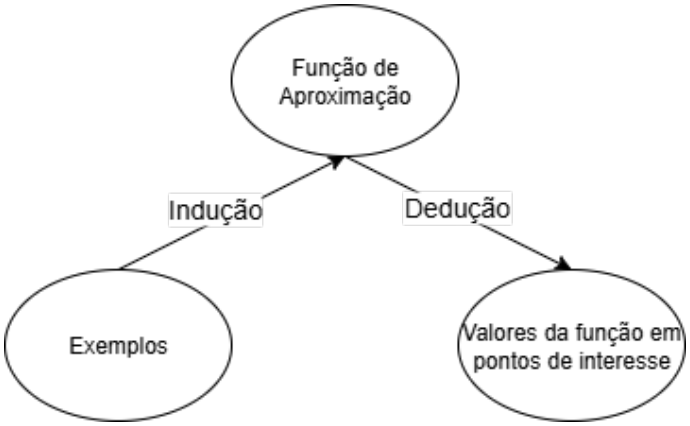
Em conjunto, indução e dedução formam a base do treinamento de modelos supervisionados. Ao treinarmos, olhamos os dados e induzimos uma relação $\hat{Y} = f(X_1, \ldots, X_n)$ entre eles. Para verificar a plausibilidade de nossa indução, aplicamos $f()$ aos valores de $\{X_1, \ldots, X_n\}$ em pontos não vistos – o conjunto de teste (Figura 17.4). Se a função induzida for uma estimativa boa, então seu resultado deve ficar próximo, dentro de uma segurança estatística, do valor de Y nesses novos dados. Essa é a parte da dedução.

Certo, mas e a transdução, onde ela entra nisso? A **transdução** é uma forma de inferência que, assim como a dedução, infere o valor de uma variável de interesse em um

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Figura 17.4: Indução e dedução no treinamento de modelos. A partir de um conjunto de exemplos rotulados (treino), uma relação entre as variáveis independentes e a variável dependente é induzida. A partir dessa função, valores da variável dependente são estimados para pontos de interesse não vistos pelo modelo (teste). Se os valores estimados ficarem próximos dos reais, então podemos deduzir que a função induzida é plausível diante desses dados.



Fonte: Adaptada de (Vapnik, 2000, p. 293).

determinado conjunto de pontos. Diferentemente da dedução, contudo, ela não parte de uma relação entre variáveis, fazendo essa inferência diretamente a partir dos dados iniciais (Vapnik, 2000). Assim, nenhuma relação geral é buscada, como na indução, mas apenas uma estimativa $\hat{Y}$ do valor da variável alvo Y para os pontos específicos de interesse (Chapelle et al., 2006), conforme ilustra a Figura 17.5.

17.2.1 Transdução como forma de aprendizado

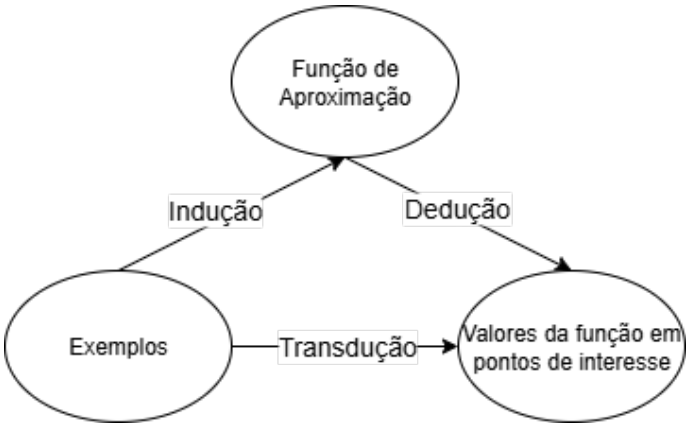
No contexto apresentado, o **aprendizado transdutivo**, conforme proposto por Vapnik[2] (Gammerman et al., 1998; Vapnik, 1998), surge como uma alternativa ao aprendizado supervisionado, ao permitir que o modelo treinado utilize diretamente os dados não rotulados do conjunto de teste durante o processo de treinamento, com o objetivo de melhorar seu desempenho especificamente nesse conjunto (Zhu et al., 2009). Dessa forma, enquanto o paradigma indutivo busca aprender uma função generalizável a partir dos dados de treinamento para ser aplicada a quaisquer novos exemplos, o aprendizado transdutivo se concentra em prever somente os rótulos dos exemplos específicos não rotulados disponíveis no momento do treinamento (Chapelle et al., 2006; Zhu et al., 2009), construindo uma espécie de "atalho" entre os passos do aprendizado indutivo clássico, conforme ilustrado na Figura 17.6.

[2]Também conhecido por seu papel fundamental no desenvolvimento das Máquinas de Vetores de Suporte (SVMs).

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

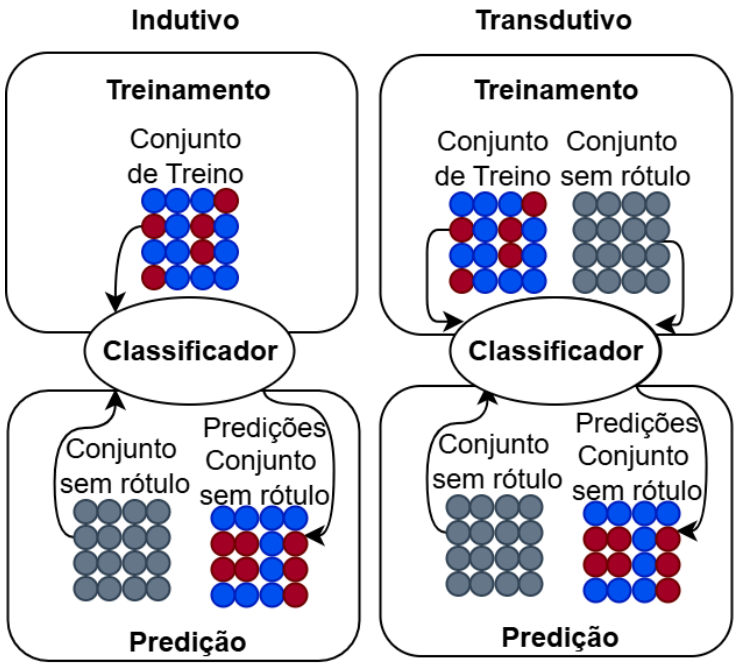
https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

Figura 17.5: A transdução trata de, a partir de uma amostra de dados rotulados e não rotulados, estimar diretamente os valores da variável dependente para os pontos não rotulados, sem a necessidade de indução de uma relação geral entre as variáveis independentes e a variável dependente a partir dos dados rotulados.



Fonte: Adaptada de (Vapnik, 2000, p. 293).

Figura 17.6: No aprendizado indutivo, uma relação entre as variáveis preditoras e a variável alvo é inferida. Essa, então, é aplicada a dados não rotulados, gerando estimativas para seus rótulos. Em contraste, no aprendizado transdutivo, dados rotulados e não rotulados são apresentados ao modelo, desta vez com a intenção de produzir inferências apenas para este conjunto de dados não rotulados em específico.



https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Formalmente[3], o aprendizado transdutivo parte de um conjunto de m instâncias:

$$S = \{s_1, s_2, \ldots, s_m\}$$

e de uma coleção de m vetores de características X, em que cada $x_i \in X, 1 \leq i \leq m$ é associado à sua instância correspondente $s_i \in S$:

$$X = (\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_m)$$

A partir de X, definem-se dois subconjuntos, $X_{\text{treino}} = (\mathbf{x}_{tr_1}, \mathbf{x}_{tr_2}, \ldots, \mathbf{x}_{tr_r})$ e $X_{\text{teste}} = (\mathbf{x}_{ts_1}, \mathbf{x}_{ts_2}, \ldots, \mathbf{x}_{ts_n})$, em que X_{treino} corresponde ao conjunto de r instâncias cujos rótulos são conhecidos, e X_{teste} corresponde ao conjunto de $n = m - r$ instâncias não rotuladas na base, de modo que $X = X_{\text{treino}} \bigcup X_{\text{teste}}$.

Ao conjunto de treino X_{treino} associamos seus rótulos $Y_{\text{treino}} = (y_{tr_1}, y_{tr_2}, \ldots, y_{tr_r})$ correspondentes, tais que cada rótulo $y_{tr_i} \in Y_{\text{treino}}$, onde $1 \leq i \leq r$ está associado a seu respectivo $\mathbf{x}_{tr_i} \in X_{\text{treino}}, 1 \leq i \leq r$. Durante o processo de aprendizado, o método transdutivo utiliza esses conjuntos para inferir os rótulos $\hat{Y}_{\text{teste}} = (\hat{y}_{ts1}, \hat{y}_{ts2}, \ldots, \hat{y}_{ts_n})$ dos exemplos de teste.

O aprendizado transdutivo é, de fato, um caso particular de aprendizado semi-supervisionado (Chapelle et al., 2006; Zhu et al., 2009) em que o foco é exclusivamente rotular corretamente os exemplos não rotulados conhecidos durante o treinamento, sem a preocupação de generalizar para dados futuros desconhecidos. Essa característica o torna especialmente atrativo em contextos de PLN nos quais temos uma base com poucos dados rotulados e muitos dados não rotulados, como em tarefas de anotação e construção de *corpora*. Essa abordagem pode então ser utilizada tanto para gerar uma base inteiramente rotulada quanto para incorporar exemplos não rotulados ao processo de treinamento supervisionado.

Do ponto de vista de aplicação prática, o aprendizado transdutivo pode ser entendido como uma estratégia viável para situações em que:

- O conjunto de dados de teste é conhecido no momento do treinamento;
- Não há interesse em inferência fora desse conjunto específico;
- A rotulação manual do conjunto completo seria inviável ou muito custosa.

17.3 Aplicações do Aprendizado Transdutivo em PLN

A aplicação do aprendizado transdutivo em tarefas de PLN tem se mostrado promissora em diversos contextos. Entre as abordagens mais comuns estão o uso de modelos com a propagação de rótulos e a adaptação de modelos supervisionados tradicionais para permitir o aprendizado transdutivo.

17.3.1 Métodos baseados em propagação de rótulos

Métodos baseados em propagação de rótulos têm como objetivo utilizar os rótulos de exemplos previamente anotados para inferir pseudo-rótulos em exemplos não rotulados (a transdução). Essa inferência possibilita que os dados originalmente não rotulados possam ser incorporados ao treinamento de modelos supervisionados, tornando-se especialmente

[3]Ver (Chapelle et al., 2006) e (Zhu et al., 2009).

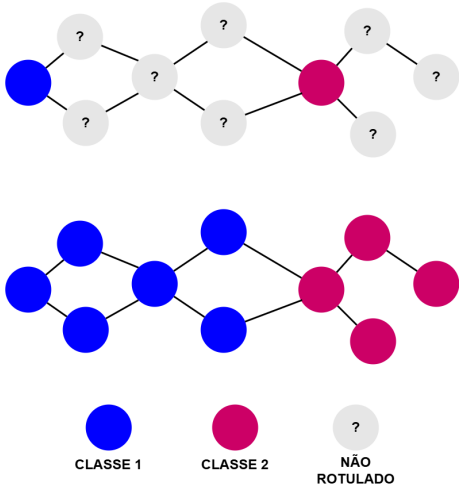
©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

útil em cenários com baixa disponibilidade de dados anotados manualmente (Iscen et al., 2019).

Uma classe importante de métodos para propagação de rótulos são os baseados em grafos, como representado na Figura 17.7. Nessa abordagem, os dados (instâncias rotuladas e não rotuladas) são representados como nós de um grafo, enquanto as arestas indicam a similaridade entre pares de exemplos. Essa similaridade pode ser definida com base tanto em *embeddings*, quanto em medidas léxicas ou representações contextuais. A partir dessa estrutura, os pseudo-rótulos podem ser atribuídos aos nós não rotulados por meio de algoritmos que propagam a informação dos rótulos ao longo do grafo, respeitando a estrutura local de vizinhança.

Figura 17.7: Processo de propagação de rótulos em um grafo. Partindo de poucos exemplos rotulados (topo), o algoritmo propaga os rótulos aos nós vizinhos não rotulados, mudando esses rótulos até que um equilíbrio seja atingido (base).



Entre os algoritmos mais utilizados para esse tipo de abordagem, destacam-se a propagação de rótulos (*label propagation*) propriamente dita e o espalhamento de rótulos (*label spreading*). A **propagação de rótulos** baseia-se em um processo iterativo de disseminação de rótulos ao longo da estrutura do grafo, até que a convergência seja atingida. Esse processo consiste dos seguintes passos (Needham; Hodler, 2019), detalhados no Algoritmo 17.1:

1. **Inicialização**: cada nó da rede recebe inicialmente um rótulo único. Alternativamente, podem ser usados rótulos iniciais (*seeds*) previamente definidos.

2. **Propagação**: os rótulos se propagam através das conexões da rede, passando de um nó para seus vizinhos.

3. **Atualização**: a cada iteração, cada nó atualiza seu rótulo com base nos rótulos dos vizinhos. O novo rótulo é então definido como sendo o rótulo apresentado pela maioria dos nós vizinhos (um voto de maioria). Alternativamente, pode-se dar pesos diferentes aos vizinhos, conforme sua similaridade no grafo.

4. **Convergência**: o processo se repete até que todos os nós adotem o rótulo mais comum entre seus vizinhos. Com isso, grupos de nós densamente conectados tendem

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

a atingir rapidamente um consenso sobre um mesmo rótulo, formando comunidades distintas na rede.

Algoritmo 17.1: Técnica de Propagação de Rótulos (Chapelle et al., 2006)

Entrada: Grafo $g = (V, E)$, com nós $V = 1, ..., n$ representando os dados de treino e as arestas E representando a similaridade entre os nós. Essa similaridade é dada por uma matriz $W : W_{i,j}$ onde x_i e x_j são vizinhos no grafo e a aresta (i, j) está em E com peso $W_{i,j}$.
Calcular a matriz de afinidade $\mathbf{W}$ por meio da função de Kernel Gaussiana com largura σ, em que $\mathbf{W}_{ij} = e^{-\frac{\|x_i - x_j\|^2}{2\sigma^2}}$
Calcular a matriz diagonal de graus $\mathbf{D}$, em que $\mathbf{D}_{ii} \leftarrow \sum_j W_{ij}$
Inicializar $\hat{Y}^{(0)} \leftarrow (y_1, ..., y_l, 0, 0, ..., 0)$
enquanto não convergir para $\hat{Y}^{(\infty)}$ **faça**
  $\hat{Y}^{(t+1)} \leftarrow \mathbf{D}^{-1}\mathbf{W}\hat{Y}^{(t)}$
  $\hat{Y}_l^{(t+1)} \leftarrow Y_l$
fim enquanto
Rotular o ponto x_i pelo sinal de $\hat{y}_i^{(\infty)}$

O **espalhamento de rótulos**, por sua vez, consiste de uma variante suavizada da propagação de rótulos, em que são introduzidas regularização e suavização nos rótulos atribuídos, permitindo uma propagação mais controlada e robusta, conforme detalhado no Algoritmo 17.2. Embora os algoritmos de propagação de rótulos e espalhamento de rótulos sejam similares na forma como propagam rótulos pelo grafo, a principal diferença está na regularização aplicada. No modelo do Algoritmo 17.2, é utilizada uma matriz Laplaciana normalizada e um parâmetro de controle α para suavizar a propagação, garantindo maior estabilidade e robustez contra rótulos ruidosos, enquanto que no Algoritmo 17.1 é realizada uma atualização direta usando a matriz de graus sem esse mecanismo de suavização.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Algoritmo 17.2: Técnica de Espalhamento de Rótulos (Chapelle et al., 2006)

Entrada: Grafo $g = (V, E)$ com nós $V = (1, ..., n)$ representando os dados de treino e as arestas E representando a similaridade entre os nós. Essa similaridade é dada por uma matriz $W : W_{i,j}$ onde x_i e x_j são vizinhos e a aresta (i, j) está em E com peso $W_{i,j}$.
Calcular a matriz de afinidade $\mathbf{W}$ por meio da função de Kernel Gaussiana com largura σ, em que $\mathbf{W}_{ij} = e^{-\frac{\|x_i - x_j\|^2}{2\sigma^2}}$, para $i \neq j$ (sendo que $\mathbf{W}_{ii} \leftarrow 0$)
Calcular a matriz diagonal de graus $\mathbf{D}$ por $\mathbf{D}_{ii} \leftarrow \sum_j W_{ij}$
Calcular o Laplaciano normalizado do grafo $\mathbf{L} \leftarrow \mathbf{D}^{-1/2}\mathbf{W}\mathbf{D}^{-1/2}$
Inicializar $\hat{Y}^{(0)} \leftarrow (y_1, ..., y_l, 0, 0, ..., 0)$
Escolher um parâmetro $\alpha \in [0, 1]$
repita
  $\hat{Y}^{(t+1)} \leftarrow \alpha\mathbf{L}\hat{Y}^{(t)} + (1 - \alpha)\hat{Y}^{(0)}$
até que convirja para $\hat{Y}^{(\infty)}$
Rotular o ponto x_i pelo sinal de $\hat{y}_i^{(\infty)}$

Essas técnicas transdutivas baseadas em grafos são especialmente relevantes em tarefas como classificação de documentos, análise de sentimentos, categorização de tópicos ou outras em que há grande volume de dados não anotados e recursos limitados para anotação manual (Jafarlou; Kubek, 2025). Um exemplo ilustrativo da utilidade do aprendizado transdutivo em PLN pode ser observado na tarefa de análise de sentimentos em resenhas de usuários (Marcacini et al., 2018). Suponha que temos um pequeno conjunto de resenhas rotuladas e um grande volume de resenhas não rotuladas. Nesse cenário, é possível construir um grafo ou uma rede onde cada nó (rotulado ou não) representa uma resenha.

Utilizando então algoritmos como Propagação ou Espalhamento de Rótulos, os rótulos das resenhas anotadas são propagados pelo grafo para os nós não rotulados. O resultado é a atribuição de pseudo-rótulos às resenhas originalmente sem anotação. Dessa forma, esses métodos demonstram como o aprendizado transdutivo pode ser uma solução eficaz e escalável para problemas do mundo real, onde a rotulação de dados em larga escala é um desafio constante.

17.3.2 Adaptação de modelos supervisionados

Além das abordagens baseadas em grafos, outra estratégia para aplicar aprendizado transdutivo em PLN envolve a modificação do processo de treinamento de modelos supervisionados. Nessa seção, focaremos em uma modificação específica, feita na função de custo dos modelos, com base na proposta do método TransBoost (Belhasin et al., 2022). Embora originalmente desenvolvido para classificação de imagens, esse mesmo princípio pode ser aplicado ao ajuste fino de modelos de linguagem (Capítulo Modelos de linguagem), permitindo uma forma eficiente de aprendizado transdutivo com dados textuais.

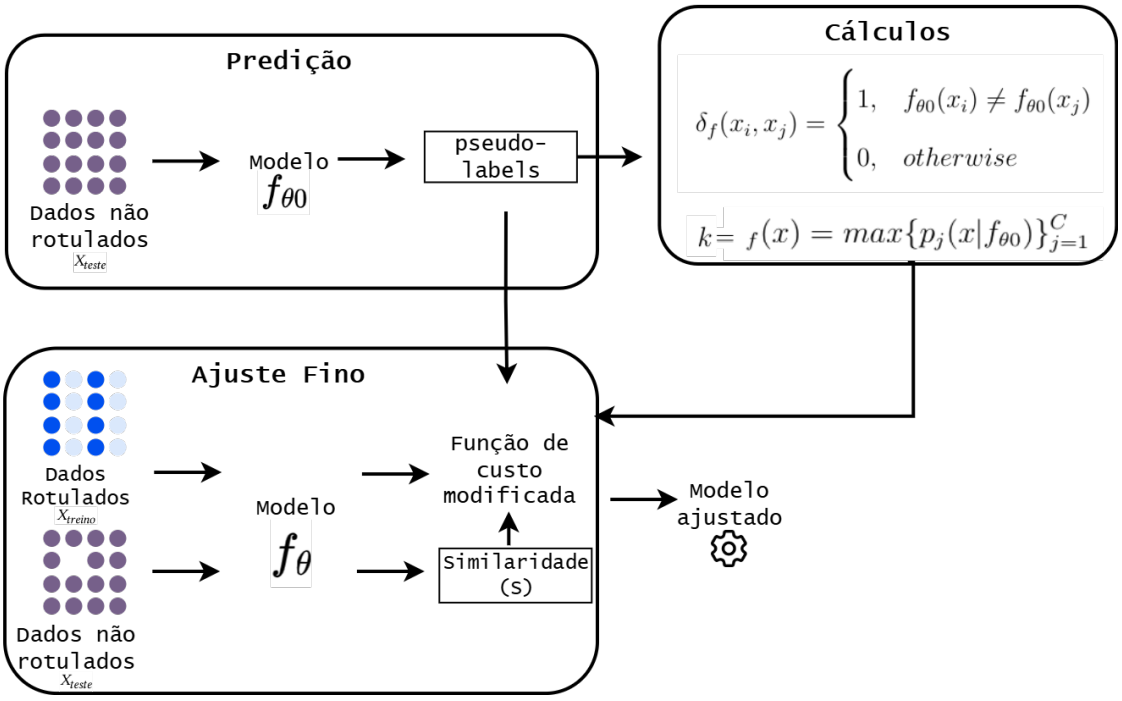
Imagine então um cenário como o descrito ao longo deste capítulo, em que você possui uma base de dados volumosa, mas cuja maior parte não está rotulada. Seu objetivo é treinar um modelo de aprendizado profundo para uma tarefa de classificação, porém o custo de rotular manualmente os dados inviabiliza a criação de um conjunto supervisionado de tamanho adequado.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Inspirado no método TransBoost (Belhasin et al., 2022), é possível modificar a função de custo do processo de ajuste fino de um modelo pré-treinado para que ela leve em consideração não apenas os exemplos rotulados, mas também os não rotulados. Essa abordagem permite então ao modelo explorar a estrutura presente nos dados não rotulados durante o ajuste fino, fazendo com que ele aprenda também a partir de sua estrutura, conforme representado na Figura 17.8.

Figura 17.8: Transboot (Belhasin et al., 2022) com um LLM. O Aprendizado Transdutivo neste caso é dividido em duas etapas: predição com o modelo pré-treinado $f_{\theta 0}$ e o ajuste fino transdutivo. O objetivo da primeira etapa é fornecer informações relevantes para o próximo passo do Aprendizado: os pseudo-rótulos preditos por $f_{\theta 0}$ para cada exemplo de X_{teste}. Esses pseudo-rótulos serão utilizados pela função δ para definir a probabilidade de dois exemplos serem ou não da mesma classe, e por k para definir a confiabilidade de $f_{\theta 0}$ nessa probabilidade dada. Essas informações serão então utilizadas durante o ajuste fino transdutivo para adicionar maior penalidade ou não à função de custo.



Dentro dessa abordagem, faz-se necessário inicialmente atribuir rótulos provisórios aos exemplos não rotulados. Isso pode ser feito por meio de um modelo pré-treinado (com ou sem ajuste fino) ou utilizando LLMs em configurações *few-shot* ou *zero-shot*. O modelo utilizado para esse fim será chamado de $f_{\theta 0}$, e o objetivo dessa etapa é gerar uma primeira estimativa razoável de rótulos, ainda que imperfeita.

O passo seguinte consiste na modificação da função custo usada no ajuste fino do modelo f_θ, ou seja, durante o ajuste fino, a função de custo é adaptada para incluir um termo adicional correspondente aos exemplos não rotulados. Esse termo transdutivo leva em conta a similaridade S entre exemplos, a confiança k do modelo nas predições e a consistência δ entre os pseudo-rótulos gerados. Em geral, esse termo tem um peso menor em relação ao erro supervisionado, controlado por um hiperparâmetro de regularização λ. Dessa forma,

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

três pontos principais são considerados para o cálculo do termo transdutivo na função de custo:

- A probabilidade δ de dois exemplos não rotulados pertencerem a classes diferentes;
- A confiança k do modelo na predição gerada; e
- A similaridade S entre os vetores preditos para os exemplos não rotulados.

Com isso, espera-se que uma penalidade maior, ponderada pela confiança do modelo na predição, seja dada ao aprendizado quando exemplos que provavelmente pertencem a classes diferentes possuírem probabilidades de classes empíricas similares. O Algoritmo 17.3 detalha a aplicação desse processo, conforme descrito em (Belhasin et al., 2022).

Algoritmo 17.3: Aplicação do Método Transdutivo por meio da alteração da função de custo (adaptado de (Belhasin et al., 2022))

Entrada: Conjunto rotulado $S_r \triangleq (X_{treino}, Y_{treino})$; conjunto de teste não rotulado X_{teste}; modelo f_θ pré-treinado; função de perda supervisionada ℓ; implementações das funções do TransBoost: $\mathcal{S}, \delta$ e κ. Hiperparâmetros: número de épocas E, tamanho do lote rotulado R', tamanho do lote não rotulado N', parâmetro de regularização λ.

para *epoch* $\leftarrow 1$ **até** E **faça**
  // Amostrar lotes rotulados e não rotulados de forma cíclica
  para $S'_r \triangleq (X'_{treino}, Y'_{treino}) \subset S_r,\ X'_{teste} \subset X_{teste}$ **até** $\max\left\{\left\lceil \frac{R}{R'} \right\rceil, \left\lceil \frac{N}{N'} \right\rceil\right\}$ **faça**
    // Preparar pares aleatórios de teste
    $X'^{\pi}_{teste} \leftarrow$ gerar uma permutação aleatória de X'_{teste}
    // Aproximar a perda do TransBoost
    $L_{\text{TransBoost}} \leftarrow 0$
    para $i \leftarrow 0$ **até** N' **faça**
      $L_{\text{TransBoost}} \leftarrow L_{\text{TransBoost}} + \kappa(x_i)\kappa(x_i^{\pi})\delta(x_i, x_i^{\pi})\mathcal{S}(x_i, x_i^{\pi})$
    fim para
    $L_{\text{TransBoost}} \leftarrow \begin{cases} \frac{L_{\text{TransBoost}}}{\sum_{i=1}^{N'} \delta(x_i, x_i^{\pi})} & \text{se } \sum_{i=1}^{N'} \delta(x_i, x_i^{\pi}) \neq 0 \\ 0 & \text{caso contrário} \end{cases}$
    // Aplicar uma etapa de otimização
    $\theta \leftarrow$ otimizar $\left[\frac{1}{R'} \sum_{i=1}^{R'} \ell(f_\theta(x_i), y_i) + \lambda \cdot L_{\text{TransBoost}}\right]$
  fim para
fim para

Para tal aplicação, o objetivo do aprendizado é minimizar a função de custo L durante o ajuste fino do modelo f_θ, a partir um conjunto de dados rotulados $X_{treino} = \{(x_1, y_1), ..., (x_r, y_r)\}$ e de um conjunto finito de dados não rotulados $X_{teste} = \{x_{r+1}, .., x_{r+n}\}$. Este objetivo é dado por:

$$L(X_{treino}, Y_{treino}, X_{teste} | f_\theta, S, \delta, k) = \frac{1}{R} \sum_{n=1}^{R} \underbrace{l(f_\theta(x_i), y_i)}_{\text{perda indutiva}} + \lambda \cdot \underbrace{l_{Transdutivo}(X_{teste} | f_\theta, S, \delta, k)}_{\text{perda_transdutiva}}$$

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

onde o primeiro termo representa uma função de perda indutiva padrão, enquanto no segundo é introduzido o termo transdutivo junto com λ, representando um hiper-parâmetro de regularização.

Para ajudar no entendimento dessa adaptação transdutiva, vamos analisar um exemplo simples, em que um conjunto de dados contendo dois pontos, x_1 e x_2, possui rótulos y_1 e y_2, respectivamente, pertencentes a duas classes $C = \{0, 1\}$. A esse conjunto rotulado é adicionado um conjunto não rotulado, contendo também dois exemplos. Um modelo pré-treinado é então ajustado nesses dados, e sua predição durante o ajuste fino, para cada um deles é:

- x_1, com $y_1 = 1$ e predição do modelo $f_\theta(x_1) = [0{,}2; 0{,}8]$
- x_2, com $y_2 = 0$ e predição do modelo $f_\theta(x_2) = [0{,}7; 0{,}3]$

Note que a predição indica que o modelo está relativamente confiante nos dois casos. Para x_1, a confiança de que o primeiro exemplo pertence à classe correta $y_1 = 1$ é 0,8, enquanto que a confiança de que x_2 pertence à classe 0 é 0,7.

O conjunto de dados não rotulados, por sua vez, teve os seguintes pseudo-rótulos gerados por $f_{\theta 0}$:

- x_3, com pseudo-rótulo $f_{\theta 0}(x_3) = [0{,}6; 0{,}4]$, ou seja, a classe 0 (com predição durante o ajuste fino $f_\theta(x_3) = [0{,}6; 0{,}4]$);
- x_4, com pseudo-rótulo $f_{\theta 0}(x_4) = [0{,}3; 0{,}7]$, ou seja a classe 1 (com predição durante o ajuste fino $f_\theta(x_4) = [0{,}7; 0{,}3]$)

Note que os exemplos x_3 e x_4 têm pseudo-rótulos distintos, 0 e 1 respectivamente, conforme gerados por $f_{\theta 0}$, mas predições de f_θ similares, conforme apresentadas pelo modelo. Segundo o modelo, ambos deveriam ser associados ao rótulo 0, portanto. Essa discordância entre $f_{\theta 0}$ e f_θ será importante nos próximos passos.

Pensando no nosso exemplo ilustrativo, o termo indutivo da função de perda seria calculado da seguinte maneira, utilizando a entropia cruzada para os exemplos rotulados:

$$L_{\text{indutiva}} = \frac{1}{2}\left[-\log(0{,}8) - \log(0{,}7)\right] \approx \frac{1}{2}(0{,}223 + 0{,}357) = 0{,}29$$

Esse valor representa o erro do modelo com base apenas nos dados rotulados, x_1 e x_2, tratando o problema como um aprendizado puramente supervisionado. Já o termo transdutivo apresentado é responsável por calcular a perda do modelo considerando X_{teste}. Seu funcionamento é dado por:

$$l_{Transdutivo}(X_{teste}|f_\theta, S, \delta, k) = \frac{1}{N_\delta} \sum_{1 \le i \le N} k(x_i)k(x_j)\delta(x_i, x_j)S(x_i, x_j)$$

Notam-se aqui três termos principais, k, δ, S, onde S mede a similaridade de duas instâncias de X_{teste}, por meio do cálculo de uma função de similaridade nos vetores de predição gerados por f_θ. Esse cálculo é feito apenas para instâncias de X_{teste} que possuam probabilidade de pertencer a classes diferentes. Essa probabilidade é dada por $\delta : X \times X \to \{0, 1\}$, obtida da seguinte maneira:

$$\delta_f(x_i, x_j) = \begin{cases} 1, & f_{\theta 0}(x_i) \neq f_{\theta 0}(x_j) \\ 0, & \textit{caso contrário} \end{cases}$$

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

onde $f_{\theta 0}$ representa o modelo responsável por gerar os pseudo-rótulos, e $f_{\theta 0}(x_i)$ a predição de um pseudo-rótulo gerado para x_i.

Voltando para o exemplo, como $f_{\theta 0}(x_3) \neq f_{\theta 0}(x_4)$, temos:

$$\delta(x_3, x_4) = 1$$

Como o cálculo anterior indicou uma maior confiança do modelo de que (x_3, x_4) tenham rótulos distintos, vamos considerar esse par de exemplos não rotulados para medirmos a similaridade $S(x_3, x_4)$ entre os vetores de predição gerados por f_θ usando a similaridade do cosseno (esse passo busca penalizar predições similares para exemplos que provavelmente pertencem a classes diferentes):

$$\begin{aligned} S(x_3, x_4) = \cos(\theta) &= \frac{f_\theta(x_3) \cdot f(x_4)}{\|f_\theta(x_3)\| \cdot \|f_\theta(x_4)\|} \\ &= \frac{0{,}6 \cdot 0{,}7 + 0{,}4 \cdot 0{,}3}{\sqrt{0{,}7^2 + 0{,}3^2} \cdot \sqrt{0{,}6^2 + 0{,}4^2}} \\ &= \frac{0{,}54}{0{,}7616 \cdot 0{,}7211} \\ &\approx 0{,}9827 \end{aligned}$$

Por fim, k é dado por:

$$k_f(x) = max\{p_j(x|f_{\theta 0})\}_{j=1}^{C}$$

utilizando uma função *softmax* padrão, representando a confiança da predição de $f_{\theta 0}$ para o pseudo-rótulo gerado com a intenção de ponderar a penalidade baseando-se na confiança do modelo na predição gerada.

Voltando para o nosso exemplo, k é a predição feita pelo modelo auxiliar $f_{\theta 0}$ de maior confiança, ou seja:

$$k(x_3) = 0{,}6; \quad k(x_4) = 0{,}7$$

Agora já temos todas as informações necessárias para o cálculo do termo transdutivo. A perda transdutiva será, portanto:

$$l_{\text{Transdutivo}} = k(x_3) \cdot k(x_4) \cdot \delta(x_3, x_4) \cdot S(x_3, x_4) = 0{,}6 \cdot 0{,}7 \cdot 1 \cdot 0{,}9827 = 0{,}4127$$

Esse valor representa uma penalidade adicional para o modelo, considerando o desacordo entre exemplos não rotulados que são similares (do ponto de vista das predições) mas foram pseudo-rotulados de forma distinta pelo modelo auxiliar. Por fim, considerando o hiperparâmetro $\lambda = 1$, a função de custo total será:

$$L = L_{\text{indutiva}} + \lambda \cdot l_{\text{Transdutivo}} = 0{,}29 + 0{,}4127 = 0{,}7027$$

Como podemos observar na Tabela 17.1, o custo total aumenta com a adição do termo transdutivo. Essa penalidade extra força o modelo a ajustar melhor suas predições para evitar que exemplos não rotulados sejam rotulados de forma incoerente, especialmente quando o modelo auxiliar está confiante nas predições.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Tabela 17.1: Comparação entre a função de custo padrão e a função com termo transdutivo.

Componente	Valor Aproximado
Perda indutiva	0,290
Perda transdutiva	0,4127
Total (com termo transdutivo)	**0,7027**

Esse exemplo mostra na prática como o aprendizado transdutivo pode ajudar a refinar o processo de ajuste de modelos mesmo quando os dados rotulados são escassos. Ao incorporar a estrutura dos dados não rotulados e considerar a confiança e coerência entre pares, é possível alcançar um desempenho mais robusto e coerente, o que é especialmente útil em tarefas de PLN com limitação de anotação supervisionada.

17.3.3 Avaliação dos dados anotados

A confiança nos rótulos artificiais gerados pelo método transdutivo precisa ser avaliada para evitar a propagação de erros e garantir a consistência dos dados anotados. Nesse sentido, uma forma inicial de avaliação é baseada na **Confiança nas Predições**, ou seja, o valor da confiança do modelo nas predições realizadas para os exemplos não rotulados. Isso é normalmente extraído da distribuição de confiança da camada de saída do modelo, podendo ser feito considerando como predições corretas aquelas a partir de um limiar ou também considerando uma margem mínima entre as duas classes com maiores confianças (Sohn et al., 2020).

A qualidade dos pseudo-rótulos também pode ser verificada por meio de sua consistência com instâncias similares, chamada de **Consistência Semântica Local**. Isso pode ser feito verificando se os vizinhos mais próximos de um exemplo (no espaço vetorial do modelo) possuem os mesmos pseudo-rótulos ou agrupar os *embeddings* e medir se os rótulos atribuídos coincidem com os agrupamentos naturais dos dados (Botzer et al., 2023).

Alternativamente, é possível realizar uma **avaliação manual amostral**. Embora mais custosa, a avaliação humana de uma amostra dos pseudo-rótulos gerados é uma forma direta e confiável de verificar sua qualidade. Esta pode ser feita em duas etapas principais:

- **Seleção de uma amostra de exemplos representativos:** selecionar exemplos de diferentes níveis de confiança, classes e regiões do espaço vetorial para revisão manual.
- **Taxa de concordância com anotadores humanos:** calcular métricas de concordância entre rótulos automáticos e os atribuídos por especialistas.

Por fim, uma forma pragmática de avaliar os rótulos gerados transdutivamente é observar seu **impacto no treinamento posterior**, ou seja, seu impacto quando utilizados em etapas posteriores de treinamento, comparando a acurácia sobre um conjunto de validação real ao incorporar os pseudo-rótulos ao treinamento supervisionado, e medir o desempenho final com e sem os dados rotulados artificialmente para avaliar sua contribuição. Essa abordagem permite verificar se os rótulos gerados estão, de fato, contribuindo positivamente para a tarefa-alvo (Oliver et al., 2018).

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

17.4 Benefícios e Limitações

O aprendizado transdutivo tem ganhado destaque como uma abordagem promissora em cenários onde o custo de anotação de dados é elevado ou a disponibilidade de dados rotulados é limitada, uma realidade comum em diversas aplicações de PLN. Ao contrário do aprendizado indutivo, cujo objetivo é generalizar para instâncias futuras e desconhecidas, o aprendizado transdutivo busca otimizar a predição para um conjunto fixo de exemplos de teste conhecidos previamente. Essa mudança de foco permite que a informação presente nos próprios dados de teste seja explorada diretamente durante o treinamento e consequentemente rotulada.

Entre os principais benefícios do aprendizado transdutivo, destacam-se:

- **Redução da dependência de bases de dados rotuladas:** ao incorporar dados não rotulados no processo de aprendizagem.
- **Melhor aproveitamento de recursos humanos e financeiros:** a diminuição da necessidade de rotulagem torna o processo de desenvolvimento mais econômico, especialmente em domínios de difícil anotação ou com poucos recursos computacionais disponíveis.
- **Predição especializada:** como o modelo conhece previamente os exemplos de teste, ele pode adaptar melhor suas representações e decisões a esses dados específicos, sem a preocupação de vazamento de dados.

No entanto, essa abordagem também apresenta desafios e limitações importantes:

- **Maior complexidade computacional:** alguns métodos transdutivos exigem a construção de grafos, inferência sobre todas as instâncias ou múltiplas etapas de ajuste fino, o que pode tornar sua aplicação mais cara e demorada computacionalmente.
- **Baixa generalização fora do conjunto de teste:** por estar centrado em um conjunto específico de exemplos, o modelo transdutivo pode não generalizar bem para novos dados, o que limita seu uso em cenários dinâmicos ou contínuos, apesar de não ser o objetivo desse tipo de aprendizado.
- **Dependência de uma boa inicialização:** muitos métodos transdutivos se beneficiam de modelos pré-treinados ou de boas predições iniciais (por meio de pseudo-rótulos), o que introduz um grau de sensibilidade ao desempenho do modelo base.

Visto isso, o aprendizado transdutivo representa uma alternativa aos métodos supervisionados tradicionais, especialmente quando incorporado a arquiteturas modernas de modelos. Sua utilidade se destaca em tarefas onde os dados de teste estão disponíveis antecipadamente, como em classificações offline, revisões, ou anotação de *corpora*, e quando há interesse em maximizar o desempenho sobre um conjunto específico de instâncias.

Espera-se um avanço na integração entre modelos pré-treinados e técnicas transdutivas, potencializando ainda mais o uso de dados não rotulados com maior eficiência. O aprendizado transdutivo é uma ferramenta complementar importante no repertório de técnicas de aprendizado de máquina aplicadas ao PLN, com potencial para democratizar o acesso a soluções de alta qualidade mesmo em cenários com recursos limitados.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Parte IX

Avaliação

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Capítulo 18

Avaliação de tecnologias de linguagem

Brielen Madureira

Publicado em: 13/03/2024
Atualizado em: 20/11/2024

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

18.1 Introdução: Avaliando

Imagine que você quer comprar um novo celular. Faz de conta que uma empresa está anunciando um novo modelo chamado weTalk. Sua campanha de *marketing* garante que o weTalk traz tecnologia de ponta, sendo melhor que todos os concorrentes. Isso lhe convence a já sair comprando? Geralmente, não acreditamos em tudo que a própria empresa que comercializa o aparelho fala, porque o interesse principal dela é vender. Nós buscamos informações de terceiros. Ao ler os depoimentos de clientes, a cena começa a mudar: há muitas pessoas insatisfeitas reclamando que o celular parou de funcionar em poucos dias ou que é muito lento. Uma luz amarela acende, mas como o preço está acessível, você fica em dúvida: arriscar ou optar por um modelo mais caro? Você, então, encontra um relatório de controle de qualidade feito por uma agência de proteção ao cliente, com uma análise bem detalhada, contendo resultados de testes feitos em diversos celulares e uma comparação minuciosa de funcionamento em vários quesitos. Esse modelo recebeu uma baixa classificação. Está decidido, melhor não adquirir o weTalk.

Esse exemplo ilustra que, para tomadas de decisão sobre um produto ou um sistema, precisamos ter acesso a uma avaliação baseada em fontes de informação confiáveis. Em PLN, não basta apenas construirmos modelos, é preciso entender quando e por que eles acertam ou erram para decidirmos se eles estão prontos para serem usados e também para podermos aperfeiçoá-los. Sendo assim, **uma avaliação adequada, justa, abrangente, detalhada, sistemática e transparente é um passo essencial ao se desenvolver, construir, analisar, comparar e usar tecnologias de linguagem**.

Em PLN, esforços consideráveis são dedicados a conceber, definir e implementar modelos e sistemas que funcionem para determinadas tarefas. Há uma variedade de possibilidades (Sparck Jones, 1994): temos sistemas de PLN avulsos (por exemplo, um algoritmo de segmentação de palavras), sistemas de PLN que dependem de outros sistemas de PLN (um sistema de sumarização que faz uso de um componente de reconhecimento de entidades nomeadas), ou, ainda, sistemas para tarefas que não são majoritariamente linguísticas mas que contêm algum componente de PLN embutido (por exemplo, um sistema de monitoramento de catástrofes que analisa mensagens de texto em redes sociais, entre outros sinais).

Mas como podemos nos certificar de que um sistema está funcionando bem? Ou, antes, o que significa "funcionar bem" em PLN? Qualquer sistema tem uma vasta gama de dimensões que devemos considerar: o que ele faz, qual seu propósito, qual é a manutenção

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

necessária, que custos gera e que benefícios traz, qual é sua performance em variados contextos, quem o utiliza, quando é utilizado, qual a infraestrutura computacional que ele exige, o que pode dar errado, quais as considerações éticas, quem o disponibiliza, como foi desenvolvido, quão fácil é de usar, sob qual licença está etc. Avaliar um sistema abarca tudo isso. Note que são análises que demandam não apenas conhecimento de PLN em si, mas também do cenário econômico, do modelo de negócios, da disponibilidade de recursos, de aspectos sociológicos e morais, além de análise de risco, controle financeiro, pesquisas de satisfação etc. Fugiria da competência deste livro tratar de todos esses temas em pormenores, e evidentemente precisa-se de uma equipe multidisciplinar para avaliar tantos quesitos.

Por isso, neste capítulo vamos restringir um pouco o escopo dessa missão. Trataremos mais precisamente de como medir, analisar e comparar a performance de um sistema, e de como fazê-lo com responsabilidade e transparência. Ao longo dos capítulos desse livro, métodos de avaliação específicos para cada tarefa já foram expostos. A ideia agora é tomarmos uma visão mais panorâmica quanto à avaliação de tecnologias de linguagem como um todo. Mais especificamente, vamos mostrar que procedimentos e ferramentas temos para responder perguntas do tipo:

- como medir a qualidade dos *outputs* produzidos pelo sistema?
- o sistema está fazendo o que deveria?
- em que casos o sistema está falhando?
- o que ocorre quando mudamos algum componente do modelo?
- o sistema A tem vantagens sobre o sistema B?
- as usuárias e os usuários estão satisfeitos?
- como o sistema pode ser melhorado?

Em PLN, tanto a perspectiva computacional quanto a linguística são fundamentais. Apesar de haver intersecção com procedimentos de avaliação em aprendizado de máquina e desenvolvimento de *software*, apenas importá-los não é suficiente. As línguas humanas em seus diversos usos tem especificidades cruciais que devemos abordar com respeito ao trabalharmos com elas. O conhecimento dos fenômenos da linguagem permite à comunidade de PLN adaptar ou criar procedimentos de avaliação customizados para suas necessidades específicas.

A quem interessa a avaliação? A que público estamos nos dirigindo? Basicamente, a todas as pessoas. Há alguns agentes que ela afeta diretamente: quem pesquisa, desenvolve ou usa um sistema, empresas e seus clientes, autoridades e reguladores (Hirschman; Thompson, 1997; King, 1996). Mas tecnologias de linguagem estão inseridas em um ecossistema, de modo que elas também têm impactos em comunidades como um todo. Inclusive pessoas que não usam um sistema podem acabar sendo indiretamente impactadas por seus efeitos (Friedman et al., 2013). Por exemplo, se um aplicativo de tradução automática falha, haverá erros no texto que se propagam quando for lido por uma pessoa que não teve contato com o sistema de tradução em si.

É fato que toda avaliação está situada em um contexto e deve ser feita de acordo com ele (Belz, 2009; Sparck Jones, 1994). Embora o processo venha se aprimorando ao longo dos anos, com boas (e não tão boas) práticas se consolidando, é uma área de

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

bastante versatilidade em questão de escolha de métricas e procedimentos. Não podemos, portanto, dar uma receita de bolo definitiva sobre como avaliar, apenas indicar perguntas que devem ser feitas e apontar possíveis formas de respondê-las. Vamos nos familiarizar com os principais métodos, técnicas e métricas, além de entender por que a avaliação é um componente crucial em qualquer tarefa, envolvendo muito mais do que a otimização de uma métrica. É algo que exige pensamento crítico e ética, além de uma profunda compressão do contexto no qual o sistema é usado e de uma série de boas práticas.

18.2 Contexto: Por onde começar?

Para começar, vamos contextualizar o tópico deste capítulo fundamentando-o em três eixos: (i) um pouco da trajetória histórica de avaliação em PLN; (ii) a formulação teórica de tarefas de PLN e (iii) uma categorização abstrata dos tipos básicos de tarefas que ocorrem em concepções computacionais envolvendo linguagem humana.

18.2.1 Um pouco de história

Embora seja questionável chamar de histórico algo que ocorreu há menos de 50 anos, vamos voltar algumas décadas para trazer uma perspectiva histórica da consolidação da avaliação em tecnologias de linguagem.

Cohen; Howe (1988) propuseram uma sistematização inicial de avaliação em pesquisa de inteligência artificial, argumentando que, enquanto outras áreas tinham métodos experimentais e técnicas analíticas já bem estabelecidas, a metodologia em inteligência artificial ainda era vaga. Sua proposta divide o ciclo de desenvolvimento em cinco estágios (definição do problema, escolha do método, implementação do método, *design* de experimentos e apreciação dos resultados), cada um com perguntas e reflexões pertinentes à avaliação, que reproduzimos no apêndice B. Já estava claro que avaliar vai muito além de métricas de performance: os autores salientaram que avaliar envolve também identificação de deficiências, comunicação, convencimento, responsabilidade e replicabilidade.

Vindo para as tecnologias de linguagem, Paroubek et al. (2007) trazem um panorama dos principais marcos históricos desde 1960. Diversas dessas iniciativas ocorreram em forma de avaliações conjuntas para aplicações específicas, como tradução automática (Capítulo Tradução Automática) e recuperação de informação (Capítulo Recuperação de Informação), pelas quais foram-se construindo propostas e coletando *expertise* na sistematização da comparação entre sistemas.

Quanto à busca por se pensar na avaliação dirigida ao PLN como um todo, houve em 1980 uma tese de doutorado (Tennant, 1980) notando que muitas iniciativas não eram sistemáticas ou eram incompletas, o que levava a uma certa confusão acerca das conquistas do PLN. O autor menciona que a maioria dos artigos acadêmicos que ele verificou não incluía sequer uma tentativa de se avaliar o sistema, o que chega a ser inconcebível nos dias atuais (ainda bem!). Essa tese propôs uma das tentativas pioneiras de caracterizar a avaliação de tecnologias de linguagem, situando-as nas dimensões de habitalibidade, completude e conceitos abstratos.[1] Um pouco depois, em 1988, ocorreu um *workshop* dedicado à avaliação de sistemas de PLN (Palmer et al., 1988), no qual se percebia que

[1] O autor caracteriza habitabilidade como "quão bem o sistema realiza as atividades para o qual foi designado", isto é, de seu próprio domínio. Completude, para ele, diz respeito ao sistema realmente realizar o que usuário/as esperam que ele faça. Análise abstrata engloba outros aspectos como não-determinismo, portabilidade, suposições sobre usuário/as e lidar com compreensão parcial de enunciados.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

os sistemas já estavam entrando no mercado e era preciso pensar em como avaliá-los, buscando uma perspectiva que funcionasse para além de tarefas específicas. O foco esteve em debater respostas para as seguintes perguntas:

1. O que são métricas válidas para a performance *black box* (i.e., quando não se tem acesso aos mecanismos do sistema)?
2. Quais teorias linguísticas são relevantes para o desenvolvimento de conjuntos de teste?
3. Como podemos caracterizar eficiência?
4. Qual é uma expectativa razoável de robustez?
5. O que constitui conjuntos de dados de treino e de validação válidos?
6. Como tudo isso se relaciona com medir o progresso da área?

Anos mais tarde, surgem alguns trabalhos proeminentes e concomitantes, com propostas já mais amadurecidas e bem fundamentadas sobre o tema. Sparck Jones; Galliers (1995) publicaram um livro que trata só de avaliação de sistemas de PLN, consolidando-a como uma área suprajacente às tarefas específicas. King (1996) relata que, por muito tempo, avaliações eram, em sua maioria, confidenciais, feitas por agências de consultoria ou revisão de pares. A mentalidade começou a mudar quando a comunidade passou a compartilhar recursos, como materiais de teste, técnicas e resultados. Passou-se a criar **conjuntos de teste**, ou seja, dados com *inputs* já mapeados aos *outputs* desejados, divididos em porcões para treino e teste. Também houve uma transição de mensurações feitas por humanos para tentativas de automatização. Em especial, nesta época podemos citar as iniciativas EAGLES I e II (Group et al., 1996), que buscaram ajustar critérios de ISO 9126 para qualidade de *software*[2] ao contexto de tecnologias de linguagem. Esse projeto formulou uma receita de sete passos, disponível no apêndice C, além de um extenso relatório com diretrizes primordiais para a área.

De lá pra cá, a comunidade tem dado bastante valor à avaliação. Belz (2009) traz um panorama da situação em 2009, que é aproximadamente o que ainda vigora. Nessa década, já havia técnicas de avaliação bem mais estabelecidas, com foco em comparações e competições. A autora identifica uma supremacia de avaliações voltadas ao desenvolvimento, feitas de forma intrínseca (ou seja, avalia-se o sistema em si) e focando na similaridade entre representações do *output* do sistema e um *gold standard*, além de dependência do nem sempre confiável julgamento humano. Veremos mais sobre esses termos na Seção 18.3.

Mais recentemente, podemos citar a série SemEval[3] que a cada ano organiza campanhas de avaliação para determinadas tarefas, ajudando a criar recursos e fomentar a comparação de diferentes sistemas. Para saber mais sobre campanhas de avaliação conjunta em português, veja o Capítulo 19. Há também *workshops* internacionais dedicados exclusivamente a esse tema, como o Eval4NLP[4], sobre avaliação em geral, e o HumEval[5], sobre avaliação por humanos. A conferência LREC[6] também tem um forte foco nesse tópico, e

[2] https://pt.wikipedia.org/wiki/ISO/IEC_9126
[3] https://semeval.github.io/
[4] https://eval4nlp.github.io/2023/index.html
[5] https://humeval.github.io/
[6] http://lrec-conf.org/

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

as grandes conferências costumam ter uma sub-área de avaliação. Até mesmo o prêmio de melhor artigo na conferência da ACL de 2020 foi dado a ao trabalho de Ribeiro et al. (2020) sobre esse tema.

Em suma, avaliação de tecnologias de linguagem é um tópico que está em alta, apesar de já vir sendo discutido há décadas. Sabemos avaliar melhor, e como os sistemas passaram realmente a fazer parte do dia a dia das pessoas, há também interesses comerciais ou regulatórios para compreender as consequências de seu uso.

18.2.2 Formalização de problemas e de modelos

Um requisito vital para uma boa avaliação é entender bem o problema que o sistema busca resolver, tanto em sua dimensão linguística como na forma em que é representado matemática ou computacionalmente. Para tanto, é útil **formalizar** o problema de forma abstrata, para permitir o raciocínio teórico e poder enquadrá-lo em um determinado método ou aproximá-lo de outras tarefas de uma mesma classe. Além disso, quando a tarefa é representada por uma amostra de dados, devemos tentar entender e esquematizar o processo que modela como os dados são gerados.

Para construir uma formalização, temos alguns "tijolos" que a matemática nos oferece: definições formais, suposições, símbolos, conjuntos, tuplas, sequências, operadores, funções, fórmulas, índices, variáveis, relações, constantes, vetores, matrizes, grafos etc. Com esses objetos, podemos seguir os seguintes passos:

- pensar sobre o problema de forma sistemática
- generalizá-lo a um nível abstrato
- listar todas as suposições e condições necessárias, refletir sobre as consequências de relaxar cada uma delas
- definir precisamente todos os seus componentes
- mencionar suas propriedades relevantes (discreto ou contínuo, categórico ou binário, ordenado ou não, finito ou infinito, domínio e co-domínio, tamanho, cardinalidade, máximo, mínimo, limites, dimensões, condições etc.)
- definir símbolos para representar e se referir a cada componente
- definir variáveis e parâmetros
- explicar como os componentes se relacionam
- descrever como os componentes são gerados e manipulados
- esquematizar as dinâmicas da tarefa
- definir matematicamente como valores e resultados são computados

Ao fazermos uma formalização, é importante sermos consistentes com a notação: cada símbolo deve se referir a um único objeto e vice-versa. Definimos o problema de forma mais geral possível, para depois instanciarmos valores específicos do nosso contexto ou *dataset* (exemplo, usar $|V|$ para representar o tamanho do vocabulário, e no problema específico informar que $|V| = 10.000$). É interessante tentar achar um equilíbrio entre

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

explicações em linguagem humana e notação matemática, retendo o rigor da formalização mas mantendo a descrição compreensível.

Vamos examinar um exemplo extraído da literatura, formulado por Madureira et al. (2023b):

Quadro 18.1: Formalização de rotulagem de sequência

> *Seja $L = \{L_1, \ldots, L_M\}$ um conjunto de etiquetas. A rotulagem de sequência é a tarefa de mapear uma sequência de n tokens $(w_i)_{i=1}^n$ para uma sequência de n etiquetas $(l_i)_{i=1}^n$, $l_i \in L$. Cada etiqueta l_i classifica o token w_i correspondente. A tarefa é mais complicada do que uma classificação de cada token, porque a natureza sequencial do input e do output devem ser levadas em conta na predição de cada etiqueta. Caso esteja disponível, uma sequência de referência $(g_i)_{i=1}^n$, com $g_i \in L$, é usada para avaliar a exatidão da sequência de etiquetas gerada.*

Esta é uma descrição abstrata de qualquer tarefa que consista em designar uma etiqueta para cada item do *input*, quando as decisões não são feitas de forma isolada para cada um, mas levam em conta o contexto sequencial. É a tarefa que está por trás de, por exemplo, anotação de PoS (Capítulo 4) e reconhecimento de entidades nomeadas (Capítulo Extração de Informação). Primeiro foi usada a letra L para representar um conjunto, com a notação convencional na matemática usando {}, e M para representar o número de elementos no conjunto. Isso é para que a formulação de adéque posteriormente a qualquer número, pois cada tarefa terá uma quantidade diferente de etiquetas. A letra n foi usada para representar quantidade de *tokens* na sequência de *input* e, consequentemente, na de *output*, já que se trata de uma relação um pra um. A notação compacta e indexada $(w_i)_{i=1}^n$ foi usada para representar as três sequências (*input*, *output* e referência), cada uma com uma letra. Com isso podemos referenciar cada elemento da sequência quando precisarmos, por exemplo, w_4 para o quarto *token e* l_2 para a segunda etiqueta prevista pelo modelo.

Note, todavia, que os autores não definiram o que é w_i. Fica subentendido que são palavras, mas seria ideal terem também definido um outro conjunto para representar o vocabulário, por exemplo, $V = \{w_1, \ldots, w_K\}$ e dizer que $w_i \in V$. Mas deve haver um equilíbrio na formalização, pois seria impraticável definir todos os objetos por completo. Por exemplo, nessa tarefa não importa muito o fato de cada w_i ser composto de uma sequência de caracteres pertencentes a um alfabeto, por isso esse aspecto foi deixado de fora.

A formalização permite a definição clara e explícita de objetos com os quais vamos trabalhar e de suas relações, servindo para ajudar a exposição ao longo de um documento e também a compreensão de quem o lê. Isso facilita também a definição de métricas e a demonstração de como são computadas, a classificação de erros, e a assimilação do que o modelo faz. Esse passo é particularmente útil para dar mais transparência e rigor para a avaliação.

18.2.3 Caracterização de tarefas de PLN

Palmer et al. (1988) propõem um passo-a-passo para o desenvolvimento de um sistema de PLN:

1. Escolher a aplicação
2. Caracterizar os fenômenos necessários

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

3. Selecionar as teorias relevantes, se disponíveis
4. Desenvolver e testar algoritmos que implementem essas teorias
5. Implementar a primeira versão do sistema
6. Caracterizar novos fenômenos que aparecem, especialmente os que têm a ver com interações
7. Refinar os algoritmos para melhorar eficiência, ou substitui-los conforme a caracterização do fenômeno muda
8. Implementar a segunda versão do sistema
9. Implementar a terceira versão do sistema, focando em questões de extensibilidade
10. Trabalhar na quarta e última versão da implementação na qual ele avança para um ambiente de produção; este estágio presta atenção especial às questões de robustez

Pois bem, vamos focar por ora no item 2: caracterizar o problema. Devemos refletir um pouco sobre algumas propriedades de dados em língua humana (veja o Capítulo 6 para mais detalhes). As línguas humanas são **sequenciais** e **temporalmente ordenadas**, ou seja, cada letra ou palavra ou sentença vem após a outra em uma sequência estabelecida por quem as profere ou escreve. Se embaralharmos as as letras de uma palavra, ou as palavras de uma sentença, ou as sentenças de um texto, eles podem deixar de fazer sentido ou mudar de significado.[7] Além disso, sendo uma sequência, ela tem uma representação **linearizada** ao longo do tempo, mas é também **estruturada**, com hierarquias sintáticas mais complexas que não são diretamente observáveis na escrita. Além disso, material linguístico escrito é composto de **unidades discretas** (letras, fonemas, palavras, frases, sentenças, enunciados, parágrafos, textos, conjuntos de textos etc.), que podem ser enumeradas. Já em sua forma oral, há um **sinal contínuo**, como vemos em um espectrograma. Compreender essas propriedades é necessário para julgarmos que tipos de métodos podem ser aplicados ao nosso objeto de estudo.[8]

A metodologia que vigora atualmente em PLN se baseia na definição de **tarefas de linguagem**, isto é, em mapear um conjunto de *inputs* a um conjunto de *outputs*, com pelo menos um deles sendo ou contendo linguagem natural, por meio da definição ou da aproximação de uma função (Schlangen, 2019). Há tarefas de compreensão, interpretação, geração, referência e inferência (Schlangen, 2019), embora algumas tarefas acabem não tendo uma aplicação muito garantida (Belz, 2009). Por exemplo, na área de tradução automática, modelam-se funções que mapeiam textos de uma língua (*input*) para textos em outra língua (*output*), buscando preservar seu significado. Na geração de legendas, o *input* é uma imagem e o *output* é uma frase ou parágrafo que descreve o que está na imagem. É comum que essas funções sejam definidas através de métodos que otimizam parâmetros extraindo padrões e generalizações de uma amostra de dados de *input* já mapeados a seu respectivo *output*. Até os modelos de redes neurais artificiais que se tornaram tão populares nada mais são do que uma função (bem complicada) que foi otimizada para mapear *inputs* a *outputs* com base em uma amostra de dados.

[7] Não queremos dizer que há uma ordem fixa para cada elemento, apenas que há uma ordem definida no momento em que as palavras são proferidas e que alterações podem vir a mudar o significado. Além disso, a rigidez da ordem dos elementos na frase varia conforme o idioma.

[8] Conceitos baseados em (Köhn, 2018).

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Há muitas tarefas de linguagem já célebres em PLN (como tradução automática, sumarização de textos e resolução de correferência), mas novas tarefas também podem ser definidas conforme a necessidade e a aplicação. Embora a lista seja longa, há subgrupos de tarefas que se assemelham em termos abstratos, de forma que podemos agrupá-las em algumas categorias principais usadas em métodos matemáticos e de aprendizado de máquina. Resnik; Lin (2010) definem alguns deles: para cada *input* podemos ter um único *output* ou múltiplos *outputs*, sendo que o *output* pode ser texto, objetos estruturados ou valores em uma escala. No Quadro 18.2, apresentamos uma taxonomia um pouco mais detalhada. Todavia, note que não há delimitações muito rígidas entre esses grupos. É comum haver a estimação de uma pontuação por trás de tarefas de classificação e ranqueamento, por exemplo, e uma sequência é também uma forma de estrutura. Todavia, esses são agrupamentos que nos ajudam a perceber que a tarefa subjacente pode ser a mesma em diversos tipos de problemas de PLN, podendo haver intercâmbio de métodos de modelagem e de avaliação entre eles.

Quadro 18.2: Exemplo de categorias de tarefas de PLN. Em cada uma delas, ou *input* e/ou *output* contêm ou são de teor linguístico.

tarefa	***input***	***output***	**exemplos**
estimar uma pontuação ou probabilidade	qualquer	um número	similaridade de textos, análise de sentimento
classificação	qualquer	uma categoria	desambiguação de significado, identificação de atos de fala ou atos de diálogo
etiquetagem	uma sequência com n elementos	uma sequência com n elementos, em relação um pra um com o *input*	anotação de *PoS*, reconhecimento de entidades nomeadas
sequence-to-sequence	uma sequência com n elementos	uma sequência com m elementos	tradução automática, sumarização
predição estruturada	qualquer	uma estrutura	extração de relações, *parsing*
agrupamento (*clustering*)	um conjunto de objetos	mesmos objetos organizados em subconjuntos	modelos de tópicos
ranqueamento	uma lista de itens	uma ordenação dessa	recuperação de informação

18.3 Paradigmas: Tipos de avaliação

Há várias abordagens e perspectivas possíveis para se avaliar um sistema. Como procedimentos de avaliação variam em propósito, escopo e natureza do objeto avaliado, não conseguimos simplesmente construir uma única ferramenta, ou um único procedimento ou um conjunto de dados de teste padrão para todos os modelos (King, 1996). A avalia-

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

ção precisa ser moldada e adequada conforme os requisitos de sua circunstância, devendo ser abrangente e sistemática e levar em conta o contexto no qual o sistema está inserido (Sparck Jones, 1994). Ainda assim, há algumas formas já bem estabelecidas que podem nos nortear. Em diversas fontes, encontramos uma divisão em três principais enfoques (Hirschman; Thompson, 1997; King, 1996; Paroubek et al., 2007):

- **adequação**: determinar se um sistema é adequado a um propósito, se atende a necessidades específicas, se ele faz o que deve, como se deve, a que custos etc.
- **diagnóstico**: descrever o perfil de um sistema com base em uma taxonomia do espaço de possíveis *inputs*, para se compreender onde e por que motivo ele está falhando
- **performance** ou **progresso**: medir resultados do sistema em uma ou mais áreas, possivelmente comparando-o com outros sistemas, para verificar se ele atingiu ou está em vias de atingir determinadas metas

Sparck Jones (1994) definiu quatro grupos de conceitos que devem ser aplicados em uma avaliação. O primeiro envolve **o sistema avaliado e fatores de performance**. Aqui, devemos considerar o que é o sistema, em que ambiente ele opera, quais fatores afetam sua performance, quais seus parâmetros e configurações, quais valores foram designados a suas variáveis, quais experimentos são rodados e quais suas funções e objetivos. O segundo grupo diz respeito a **tipos e níveis de definição da análise de performance**, ou seja, que critérios serão usados (por exemplo, eficiência, efetividade ou aceitação por humanos), como eles serão traduzidos em métricas, quais métodos serão utilizados, e a que o sistema será comparado. No terceiro grupo se encontram as **formas dos dados para teste e avaliação**. Nesse item devemos considerar o tipo dos dados e quão realistas eles são. Por um lado, precisamos de dados representativos e legítimos para a avaliação, de modo que a distribuição dos fenômenos nos dados seja condizente com a realidade em que o sistema vai funcionar. Por outro lado, devemos considerar a cobertura dos fenômenos linguísticos, podendo-se criar *test suites*, isto é, exemplos selecionados manualmente que cubram casos específicos que queremos avaliar (ver mais em Seção 16.2). Finalmente, o quarto grupo versa sobre **estratégias para o *design* e a condução da avaliação**. Devemos ter em mente o objetivo, a alçada e o *design* da avaliação em si, conforme a proposta de decomposição em critérios ilustrados na Figura 18.1.

Boa parte das avaliações em PLN pressupõe a existência de uma referência "padrão ouro" (*gold standard*), ou seja, a resposta considerada correta. Um procedimento costumeiro é comparar a resposta do sistema com a referência e julgar, com base em uma métrica, quão próximas ou parecidas elas são. Nem sempre é trivial ter uma resposta correta quando lidamos com linguagem natural, devido tanto à multiplicidade de maneiras de se expressar um significado quanto à interpretação humana, que nem sempre está em concordância entre as pessoas (Capítulo 8).

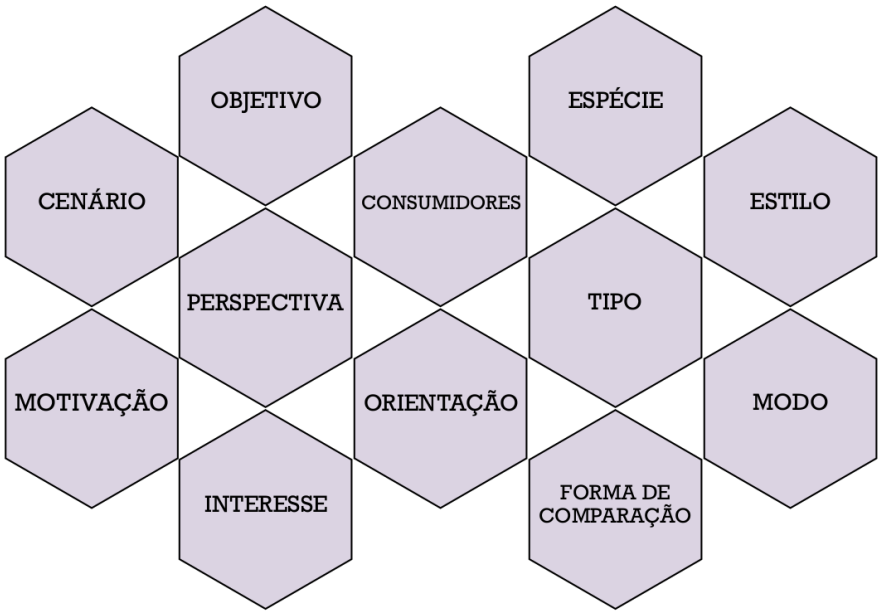
Podemos listar dez pares de características de uma avaliação (Paroubek et al., 2007; Resnik; Lin, 2010; Sparck Jones; Galliers, 1995). Vamos chamá-las dicotomias para frisar o contraste entre os elementos de cada par, mas nada impede que elas sejam usadas em conjunto ou em paralelo, a saber:

1. **manual ou automática**: Na avaliação manual, recrutam-se participantes ou avaliadores humanos para que julguem o *output* de um sistema de acordo com critérios

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Figura 18.1: Estratégia de decomposição do *design* de uma avaliação em critérios. Cada critério remete a perguntas que devem ser feitas para determinar a circunstância em que ocorre a avaliação. Espécie, tipo e estilo são algumas das dicotomias listadas a seguir.



Fonte: Adaptado de (Sparck Jones, 1994) e (Sparck Jones; Galliers, 1995).

pré-determinados. Nesse caso, a performance do humano acaba sendo avaliada indiretamente junto com o sistema (King, 1996). Na avaliação automática, usam-se algoritmos que computem os critérios de forma automatizada, idealmente aproximando o comportamento ou o julgamento de humanos.

2. **intrínseca e extrínseca**: Na avaliação intrínseca, o sistema é diretamente avaliado quanto à sua funcionalidade, com fim em si mesmo. Na avaliação extrínseca, avalia-se o impacto do sistema em uma tarefa externa, da qual ele é um componente.

3. **formativa ou sumativa**: A avaliação formativa informa sobre o progresso que está sendo feito, durante o desenvolvimento do sistema, rumo a um certo objetivo, tendo caráter temporário. A avaliação sumativa é aquela conduzida quando se atingiu um marco intermediário ou o final, com caráter mais definitivo.

4. **por usuários ou por especialistas**: A avaliação pode ser feita por usuárias e usuários do sistema, que não necessariamente têm conhecimento técnico da tarefa em si ou dos mecanismos, ou pode ser feita por pessoas que têm *expertise* no domínio ou na implementação.

5. **componente ou ponta-a-ponta**: Pode-se avaliar cada componente de um sistema individualmente ou múltiplos componentes funcionando juntos como um todo.

6. **transparente ou às escuras**: A avaliação transparente (*glass box*) tem acesso aos mecanismos que conectam *input* ao *output*, e os leva em conta. Já a avaliação às escuras (*black box*) considera apenas as relações entre *input* e *output*, sem conhecer os mecanismos pelos quais elas foram obtidas.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

7. **investigativa ou experimental**: A avaliação investigativa preocupa-se com um sistema operacional ou um *setup* e visa estabelecer características de performance, enquanto a experimental busca responder não "o que está acontecendo" mas sim "o que aconteceria se fizéssemos isso ou aquilo".

8. **qualitativa ou quantitativa**: Em avaliações qualitativas, o resultado é um perfil descritivo do comportamento do sistema, com base em aspectos não numéricos. Em avaliações quantitativas, o resultado é dado por valores numéricos através da mensuração de variáveis.

9. **objetiva ou subjetiva**: Na avaliação objetiva, as mensurações são feitas diretamente nos dados produzidos pelo processo que está sendo testado. Já a avaliação subjetiva capta mensurações baseadas na percepção que humanos têm de tal processo.

10. **supervisionada ou não-supervisionada**: Avaliações de forma supervisionada são feitas com base em um *gold standard* que traz as respostas ou resultados considerados corretos ou ideais. Avaliações não-supervisionadas não têm um *gold standard* disponível, de modo que é preciso estabelecer outras formas de julgar a qualidade do sistema.

Há, ainda, diversos atributos de qualidade de um sistema que podem ser levados em conta. Uma seleção deles é mostrada na Figura 18.2.

Figura 18.2: Atributos de qualidade de um sistema.

abrangência	confiabilidade	fidelidade	pertinência	representatividade
acurácia	conformidade	funcionalidade	portabilidade	rigor
adequação	custo	informatividade	precisão	validade
classificação de	diversidade	inteligibilidade	qualidade	velocidade
comparação	eficiência	manutenibilidade	relevância	usabilidade

Fonte: Adaptado de (King, 1996) e do ISO 9126.

Além disso, uma tecnologia é tão boa quanto o uso que se faça dela, e esse uso varia de forma individual (Sparck Jones, 1994). Sendo assim, também é relevante fazer avaliações orientadas aos usuários e usuárias do sistema, incorporando características comportamentais no protocolo de avaliação, tanto em busca de melhorar a performance quanto de compreender como o sistema é usado, em que circunstância e para quais fins (Paroubek et al., 2007).

Avaliação Conjunta e *Leaderboards*

Há bastante tempo já foi identificado um foco excessivo na busca por atingir boas pontuações ou aumentos incrementais de uma métrica, em detrimento de se fazer uma boa análise (Hirschman; Thompson, 1997).

Hoje em dia, muitas avaliações se resumem a tentar melhorar uma métrica em um *leaderboard*, ou seja, um tipo de "placar" em que se comparam diferentes sistemas em uma mesma tarefa, ou conjunto de tarefas, chamado *benchmark*. A Figura 18.3 traz um exemplo de um leaderboard para LLMs. Há uma grande variedade de benchmarks,

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

baseados em uma ou mais tarefas e datasets. Alguns deles se tornaram bem populares, por exemplo MMLU (Hendrycks et al., 2021a, 2021b), GLUE (Wang et al., 2018), SuperGLUE (Wang et al., 2019) e SQuAD (Rajpurkar et al., 2016). Embora esse paradigma ajude a fomentar um progresso mensurável, ele vem acompanhado de muitas críticas. Primeiro, ele tem a competição como cerne, o que pode se tornar algo predatório. Além disso, tentar cegamente atingir a primeira posição em um placar facilmente leva pessoas e equipes a se concentrarem em uma única pontuação, deixando de lado as diversas nuances do fenômeno com que trabalham e as limitações inerentes a qualquer *benchmark*, por exemplo, sua descontextualização (Raji et al., 2021). Ethayarajh; Jurafsky (2020) argumentam que a utilidade de um sistema é dependente das condições de quem o usa. Ou seja, às vezes o sistema que atinge pontuação máxima é tão computacionalmente custoso que é melhor abrir mão de um pouco de performance mas ter um sistema mais leve. Além disso, nem sempre melhorar a pontuação em um placar equivale a melhorar de fato as habilidades linguísticas de um sistema (Dunietz, 2020; Schlangen, 2021). Dunietz (2020) argumenta, ainda, que a busca incessante por melhorias com base em *benchmarks* faz perdermos de vista o objetivo real, que é o uso efetivo de uma tecnologia.

Figura 18.3: Um exemplo de *leaderboard* extraído do *benchmark* `clembench` (Chalamalasetti et al., 2023). Aqui, modelos de linguagem são comparados em um *benchmark* de jogos de diálogo através de uma métrica de qualidade e de percentagem de interações bem sucedidas, combinadas em uma pontuação chamada *clemscore*

Model ▲	Clemscore ▲	% Played ▲	Quality Score ▲
claude-2.1	36.38	83.08	43.79
claude-2	33.71	82.12	41.05
gpt-3.5-turbo-0613	32.53	91.96	35.37
gpt-3.5-turbo-1106	30.45	77.12	39.49
openchat_3.5	19.72	57.57	34.26
sheep-duck-llama-2-70b-v1.1	17.12	40.82	41.93
Yi-34B-Chat	16.77	63.76	26.3
WizardLM-70b-v1.0	16.7	51.65	32.34
Mixtral-8x7B-Instruct-v0.1	16.53	57.68	28.66
tulu-2-dpo-70b	15.9	54.49	29.18
claude-instant-1.2	15.44	59.61	25.91
CodeLlama-34b-Instruct-hf	10.34	23.96	43.15

As avaliações conjuntas (*shared tasks*) são eventos organizados pela comunidade nos quais uma tarefa é definida, com dados e métricas padronizados, para fomentar que diferentes soluções sejam propostas e comparadas para o mesmo problema. Elas têm vantagens

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

e desvantagens. Por um lado, ajudam a promover o progresso em subáreas, melhorar o estado da arte, discutir e comparar modelos, e criar recursos; por outro lado, elas se baseiam em competição e confidencialidade, podem ter conflitos de interesse e tirar o foco de questões éticas e de cunho mais amplo sobre um problema (Parra Escartín et al., 2017). Para saber mais sobre campanhas de avaliação conjunta em português, veja o Capítulo 19.

18.4 Procedimentos: Como avaliar?

Como vimos, há várias formas de avaliação. Podemos, por exemplo, inspecionar diretamente o funcionamento do sistema, tentar refiná-lo, avaliar suas limitações, compará-lo com outros sistemas, estudar a influência de certos componentes ou parâmetros e verificar quão bem ele funciona em áreas afins (Cohen; Howe, 1988). Vamos agora tratar de alguns procedimentos usuais para se avaliar um sistema, supondo que a avaliação vai ser concretizada em um relatório ou documento expondo uma análise dos resultados.[9]

18.4.1 Hipóteses e Experimentos

Durante o desenvolvimento de um sistema, usualmente ocorre um ciclo que alterna entre implementação de melhorias ou variações no sistema e experimentos com essas diferentes possibilidades para verificar quais são vantajosas ou atingem performance superior em algum quesito. Para a avaliação final, também é comum testarmos diversas versões do sistema, variando componentes como dados de treino, parâmetros, hiper-parâmetros, versões de código e de bibliotecas, representações dos dados, arquiteturas, métodos de otimização, etc. Cada teste pode ser considerado um experimento para mostrar o efeito de alguma escolha ou algum componente.

Há uma infinidade de possibilidades e é impossível testar exaustivamente todas elas. Dessa forma, devemos tentar definir de forma sistemática quais experimentos são úteis, necessários e suficientes para nossos propósitos, e encontrar uma forma de organizá-los para facilitar comparações e interpretações dos resultados. Não é recomendável simplesmente sair rodando experimentos a esmo. Devemos nos guiar por hipóteses sobre o que seria esperado em cada experimento. Sem uma hipótese ou uma expectativa, fica difícil julgar se o experimento foi realmente informativo ou bem sucedido.

Por exemplo, se queremos desenvolver um sistema que classifica textos como tóxicos ou não. Podemos implementar uma rede neural artificial simples, com uma camada. Verificamos o resultado. Em seguida, melhoramos sua arquitetura usando duas camadas. Qual é nossa expectativa? Esperamos que aumentar a complexidade da arquitetura leve a um resultado melhor. Se isso ocorre, ótimo, prosseguimos com as melhorias. Se isso não ocorre, vamos investigar. Há um erro no código? Há um erro na definição do modelo? Ou será que aumentar a complexidade do modelo é realmente prejudicial nessa tarefa?

Já ha hora de avaliar, podemos querer mostrar o efeito de usar *word embeddings* pré-treinados no nosso sistema. Qual é a hipótese? Esperamos que eles melhorem a performance em relação a um modelo que usa *word embeddings* próprios. Rodamos dois experimentos, um com cada tipo, e verificamos os resultados. Se ocorre o que esperamos, demonstramos a vantagem daquele componente para nosso sistema. Se não ocorre, argumentamos sobre os possíveis motivos para nossa hipótese estar errada. Talvez estejamos

[9] Para alguns exemplos mais práticos de procedimentos e métricas que trataremos aqui, recomendamos os *notebooks* disponibilizados por Chris Potts: https://github.com/cgpotts/cs224u/blob/main/evaluation_methods.ipynb e https://github.com/cgpotts/cs224u/blob/main/evaluation_metrics.ipynb.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

em um domínio tão específico que *word embeddings* genéricos não sejam tão úteis. Como podemos verificar isso? Há outros *word embeddings* que podemos testar? E se fizermos uma análise de dados para verificar a abrangência desses vetores em nosso domínio?

Todos esses raciocínios constroem análises que podem fazer parte do relatório de avaliação. Uma característica essencial de um experimento, e também do procedimento de avaliação como um todo, é que eles sejam reprodutíveis e replicáveis. Se alguém atinge uma certa métrica de performance, anota o resultado, e joga fora todos os dados e recursos que o levaram àquele número, de que ele nos serve, se não é mais possível verificarmos como se chegou a ele? Vamos simplesmente acreditar no que estão nos falando? Toda conclusão deve estar baseada em uma forma de demonstrá-la que seja acessível a quem for se respaldar nessa avaliação para fazer uso do sistema. **Avaliação não é um exercício de fé, é algo que precisa de embasamento passível de verificação**.

18.4.2 Referências

Mencionamos brevemente, na seção anterior, que é comum o uso de referências "padrão ouro" (*gold standards*) para se julgar a performance de um sistema. Vamos refletir um pouco sobre por que criar referências é desafiador em PLN através de alguns exemplos.

Há alguns problemas para os quais é relativamente possível definir uma resposta correta. Em um sistema de resolução de referências visuais, queremos que o sistema detecte um único elemento de uma imagem com base em uma referência linguística. Há uma resposta considerada correta, e podemos mais facilmente verificar se o objeto que o sistema detectou é o que desejamos.[10] Em sistemas de pergunta e resposta (Capítulo 14), há também informações factuais para as quais há um certo consenso: Quanto é 30 dividido por 3? Qual a atual capital do Equador? Quantas estrelas estão representadas na atual bandeira do Brasil? Nesses casos, há uma resposta que podemos considerar correta, e ela costuma estar bem definida (uma vez que consideremos, por exemplo, parâmetros temporais e geográficos).[11]

Mas a linguagem humana tem ambiguidades e nuances cujas interpretações variam até entre os próprios humanos, além de haver diversas formas de expressar um mesmo conceito ou ideia.

Na tarefa de tradução automática, por exemplo. Vamos imaginar que queremos traduzir a frase *the birds will migrate soon*; poderíamos traduzi-la, por exemplo, como "os pássaros vão migrar em breve", ou "as aves logo vão migrar", ou ainda "logo mais os passarinhos migrarão". Apesar de terem nuances diferentes, nenhuma dela está errada. Além disso, como geralmente não temos acesso ao que o autor ou autora da frase original realmente queria dizer, não temos como saber qual é, realmente e definitivamente, o significado original. Por isso, a tarefa de tradução automática geralmente trabalha com um referência composta de múltiplos exemplos. Mas note que, ainda assim, eles continuam sendo apenas uma amostra: é bem possível que o sistema gere "em breve haverá migração de aves", que não está errado, apesar de ser uma construção diferente que não estava entre nossas referências.

[10] Na verdade, não é assim tão simples, pois o sistema pode ter de detectar partes de objetos e fazê-lo de forma apenas parcialmente correta. Mas para efeitos de explicação vamos considerar que pode haver uma referência bem definida nesse caso.

[11] Novamente, é difícil chegar a consensos absolutos: até informações que parecem ser factuais podem estar em disputa por diferentes grupos, mudar ao longo do tempo, ou depender de contexto. Para uma discussão mais detalhada, ver (Freitas et al., 2012).

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Já em sistemas de diálogo, a criação de um *gold standard* é quase impraticável. A cada momento de uma conversa, há inúmeras formas de continuá-la de forma coerente e válida. Querer que o sistema gere exatamente o que um único humano disse naquele momento, com base em uma única amostra nos dados disponíveis, é bastante limitante.

Geralmente, a criação de uma referência ou de anotação exige uma interpretação *post factum*, feita por uma pessoa que não é a que originalmente proferiu a frase. Por exemplo, se coletamos *posts* na internet e queremos julgar se eles contêm ironia: é muito difícil ter certeza, pois não fomos nós que escrevemos o *post*. Se tivermos acesso ao contexto, podemos ter um pouco mais de segurança na resposta mas, ainda assim, nunca teremos 100% de certeza. Toda anotação por humanos de informações subjetivas vai sofrer desse dilema.

Um ponto importante para a avaliação é que não devemos confiar cegamente no *gold standard*, considerando-o uma verdade incontestável. Há erros de anotação, bem como divergências legítimas que precisamos considerar (Basile et al., 2020, 2021). Há inclusive um manifesto para que as "divergências" sejam levadas em conta ao se modelar e avaliar sistemas de PLN, em vez de tentarmos eliminá-las.[12] Além disso, as próprias instruções dadas aos anotadores podem enviesar os dados que essas pessoas geram (Parmar et al., 2023).

Ter isso em mente não deve nos impedir de usar padrões para avaliação Devemos apenas fazê-lo tendo consciência de que o próprio instrumento que usamos para avaliar pode conter falhas e limitações. Alguns *gold standard* não serão adequados o suficiente para nosso uso ou, ainda, serão de baixa qualidade. Faz parte da avaliação avaliar não só o sistema como também os materiais que temos para analisá-lo (ver mais na Seção 16.6).

18.4.3 Mensurações

Um componente primordial em avaliações são mensurações. Definimos formas de medir as propriedades e os aspectos relevantes de um sistema como desempenho, qualidade e eficiência, traduzindo-os em métricas, pontuações, estatísticas e medidas. Embora não devamos confiar irrefletidamente nelas, são formas úteis de verificar quão bem (ou mal) um sistema está funcionando e compará-lo com outros sistemas, ou com versões prévias dele mesmo.

Métricas só fazem sentido em seu contexto. Se alguém nos disser que implementou um sistema de detecção de textos tóxicos e atingiu uma métrica de qualidade de 90, o que isso quer dizer? Isoladamente, não quer dizer nada. Não sabemos se o máximo dessa métrica é 100 ou 1000. Se for 100, pode ser uma boa performance, se for 1000 a performance está calamitosa. E se o máximo for 100, mas o valor desejado seja seu mínimo?

Vamos supor que seja 100 e que queremos maximizá-la: isso não garante que atingir 90 faz dele um bom sistema. E se há outros sistemas que, nos mesmos dados, atingem 99,5 e ainda por cima fazem isso de forma mais rápida e com menos memória? Nesse caso, o sistema está bem aquém do estado da arte. Temos também de pensar nos dados. Se o sistema foi avaliado em um conjunto de dados que é balanceado (ou seja, contém por volta de 50% de textos tóxicos e 50% de textos neutros), 90% de acurácia pode ser um número bom. Mas se os dados contiverem 90% de textos neutros e apenas 10% de textos tóxicos, um "sistema" trivial que prediz que todo e qualquer texto é neutro vai atingir 90% de acurácia nesses dados, mesmo não fazendo nada! Quando a classe de interesse é rara, ou há outras formas de desbalanço nos dados da realidade, precisamos nos atentar ainda

[12] http://pdai.info/

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

mais às métricas que usamos, pois algumas delas são enviesadas. Trataremos de métricas em mais detalhes na Seção 18.5.

18.4.4 Partição de dados

Em sistemas treinados em dados, tornou-se prática comum particionar os dados que temos disponíveis em dois principais conjuntos: um de treino e um de teste (Resnik; Lin, 2010). Fazemos dessa forma porque os modelos costumam capturar muito bem os dados em que são treinados, mas o que nos interessa realmente é usá-lo na realidade, em dados que só vão existir no futuro. Ou seja, buscamos modelos capazes de generalizar. Se o modelo captar os dados de treino bem demais, ele "decora" as relações entre *input* e *output* apenas nesse conjunto e funciona mal para novos dados. Isso é conhecido como *overfitting*.

Por isso, separamos uma porção dos dados para serem usados apenas na avaliação final do sistema. Não devemos deixar que esses dados tenham qualquer influência no treino, pois eles devem ser uma *proxy*[13] dos dados da realidade que o sistema vai encontrar quando for usado para algum fim de verdade. É ideal nem sequer olhar para eles, pois mesmo que o modelo não os acesse, poderíamos influenciar seu desenvolvimento por saber o que eles contêm.

Mas como fazer, então, para checar resultados durante o desenvolvimento, sem usar os dados de teste? A solução comum é particionar os dados de treino em dois conjuntos: um para treino e um para validação. Os dados de validação são considerados dados de teste para fins de desenvolvimento (ou seja, busca de hiper-parâmetros, escolha de parâmetros, refinamento, verificação de progresso), e também para fazer análises de erro.

Não há uma regra de como fazer essa partição. Devemos ter dados suficientes para o treino, que exige bastante observações de modo a extrair padrões estatísticos, mas também deixando o suficiente para que os testes tenham validade. Uma possibilidade é dividir em 70% para treino, 10% para validação e 20% para teste (mas isso pode variar). É recomendável fazer a partição de forma aleatória, para que a distribuição dos exemplos em cada partição fique relativamente igual.

Entender a distribuição dos dados de treino é importante para escolha dos métodos que vamos implementar. Por exemplo, se estamos trabalhando com classificação e temos dados onde as classes não se distribuem de forma uniforme, pode ser necessário usarmos técnicas específicas de balanceamento dos dados de treino ou da função de otimização. Por isso, embora os dados de teste devam ser deixados fora de acesso, podemos e devemos inspecionar os dados de treino, para entender melhor com que tipo de dados estamos lidando.

18.4.5 Validação Cruzada

A partição de dados em conjuntos fixos de treino, validação e teste é um caso particular de uma técnica mais geral chamada validação cruzada (*cross validation*) (Resnik; Lin, 2010).[14] Uma desvantagem de fixar três conjuntos é que a partição é feita de forma arbitrária. E se, por acaso, todos os casos complicados acabaram indo parar nos dados de teste? O modelo não vai ter acesso a exemplos durante o treino e provavelmente não vai dar boas predições para eles. Mas e se todos os casos complicados ficarem nos dados de treino? Nesse caso, seria fácil atingir uma boa performance no conjunto de validação ou de teste, mas ela não

[13] Ou seja, um "representante substituto".
[14] https://pt.wikipedia.org/wiki/Validação_cruzada

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

refletiria o que vai ocorrer quando o sistema for usado na vida real. Para um estudo dos efeitos da arbitrariedade da partição nos resultados, ver (Gorman; Bedrick, 2019).

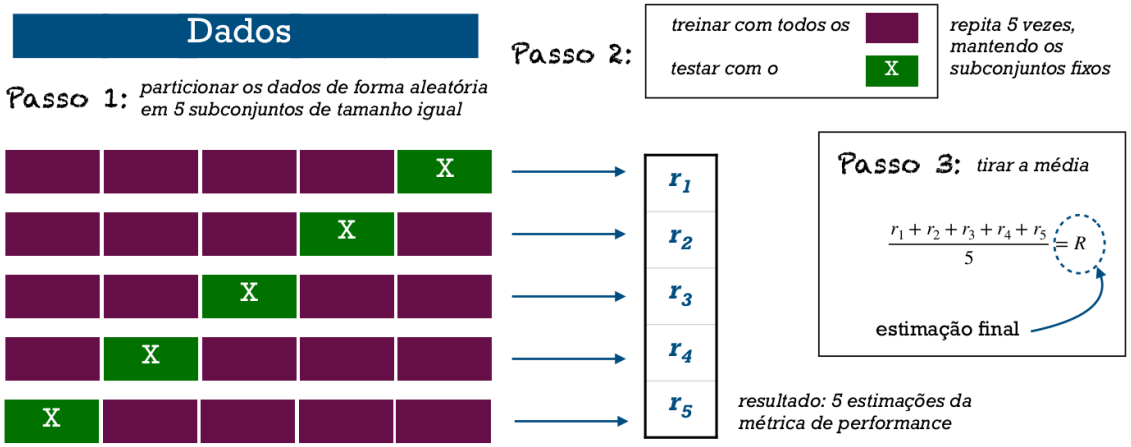
Toda partição vai sofrer dessa questão de arbitrariedade. Uma solução é usar, então, a chamada validação cruzada, que também é útil quando temos poucos dados para fazer uma partição com dados suficientes em cada porção. Nesse procedimento, os resultados são avaliados usando-se n diferentes partições. Em vez de ter um só resultado, temos n resultados. Isso nos dá uma distribuição de resultados, para os quais podemos usar análise estatística de dispersão ou simplesmente usar a média e o desvio padrão deles como uma estimação.

Há algumas formas diferentes de se realizar esse procedimento:

- ***holdout***: a partição fixa discutida no item anterior, trata-te do caso específico em que uma porção fixa é usada para teste
- ***k-fold***: divisão dos dados de treino em k partições; a cada iteração, uma delas é usada como conjunto de validação e as $k-1$ demais são usadas para treino
- ***leave-p-out***: usar p itens para validação e todo o resto para treino, alternadamente, de forma a criar todas as possíveis combinações de p itens para validação
- ***leave-one-out***: um caso particular do anterior, em que cada item é usado para validação sozinho

Os dois últimos itens são computacionalmente muito custosos para *datasets* grandes, pois seria preciso treinar uma grande quantidade de versões do sistema. A Figura 18.4 traz um exemplo da validação *k-fold* para $k = 5$. Note que, quando um *dataset* é disponibilizado já com as três partições tradicionais definidas, é interessante segui-las para facilitar a comparação. Todavia, quando uma comunidade inteira está usando os mesmos dados e a mesma partição de teste, pode ocorrer *overfitting* "comunitário", otimizando soluções para funcionarem bem apenas nesses dados, um trabalho após o outro, mesmo que os dados não estejam sendo usados diretamente no treino por ninguém (Gorman; Bedrick, 2019).

Figura 18.4: Uma ilustração de como pode ser feita a validação cruzada usando o método *k-fold*. Neste exemplo, $k = 5$ e os r são os valores numéricos de uma métrica de performance.



https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

18.4.6 Comparação de Performance

Assim como o valor numérico de uma métrica sozinho não é muito informativo, o desempenho de um sistema também faz mais sentido se ocorrer em relação a outros sistemas ou às referências chamadas *baselines*. Há quatro principais *baselines* com as quais podemos comparar um sistema. A primeira é saber qual o máximo e mínimo teóricos de uma métrica, como discutimos anteriormente, para saber quão perto ou longe o sistema está deles.

A segunda é a performance de predições aleatórias ou feitas de forma muito simples (as chamadas *baselines* triviais, das quais trataremos a seguir). Elas servem como base de performance mínima: se um sistema tem um desempenho pior do que o aleatório, há algo de muito errado com ele. Um sistema que capturou alguma coisa de significativo do processo precisa ter desempenho melhor do que tomar decisões randômicas.

Em seguida, temos a referência do estado da arte atual, que pode mudar a cada dia. Para que um novo sistema seja considerado melhor, ele deve ter uma performance superior ao atual melhor sistema (em algum aspecto de interesse). Caso contrário, ele pode até ser um sistema bom, mas não há por que usá-lo se já temos um outro que funciona melhor que ele. Claro que há muitas formas de um sistema ser superior a outro: ele pode ter uma performance inferior mas ser mais rápido, e preferirmos usá-lo em um contexto que exija agilidade.

Finalmente, há a referência de performance humana na tarefa, que é considerada um máximo desejável para um sistema, podendo ser estimada por experimentos específicos com humanos ou com métricas de concordância da anotação feita também por vários humanos (Resnik; Lin, 2010) (ver detalhes na Seção 16.6.1). Há sistemas que a "ultrapassam", mas isso não significa que o sistema é superinteligente: pode ser apenas um caso de a performance humana não levar em conta os desacordos entre anotadores ou ter sido computada em pessoas que não entenderam bem o que tinham de fazer, ou, ainda, haver erros conceituais nas tarefas que compõem o *benchmark* (Tedeschi et al., 2023). Como esperar que um modelo acerte em um caso em que nem os humanos estão de acordo? Só por escolhas arbitrárias. Humanos cometem erros, se distraem, esquecem instruções ou as interpretam de formas muito variadas. Performance humana é apenas uma referência de quão bem um grupo selecionado de pessoas resolveu aquela tarefa, o que é interpretado como uma meta a se atingir em um sistema para ser considerado competente.

Para facilitar a comparação entre diferentes sistemas, podem-se usar gráficos, tabelas e placares. Voltaremos a esse ponto na Seção 18.6.

18.4.7 Testes de Significância Estatística

Uma questão crucial ao se comparar modelos é ter cuidado na hora de generalizar. Se testarmos o modelo A e o modelo B no conjunto de dados D usando a métrica M, e o valor da métrica de A foi melhor do que o da métrica de B, isso não quer dizer que o sistema A é melhor que o sistema B. Só podemos concluir que A foi melhor que B especificamente nos dados D e pela métrica M. Talvez se tivéssemos usado as muitas outras métricas M_2, M_3, M_4 ... a conclusão teria sido diferente. Um valor melhor em uma métrica não é garantia de superioridade, por isso devemos usar várias métricas para se comparar sistemas.

Ainda há algo um pouco menos evidente porém igualmente crítico: e se tivéssemos feito a análise usando outros dados? E se tivéssemos conjuntos de dados D_2, D_3, D_4 para testar? É possível que B fosse melhor que A em todos eles, e que só foi pior em D por puro azar. Talvez D contivesse exemplos muito fora do domínio de B, enquanto D_2, D_3,

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

D_4 são mais representativos da realidade. Esse dilema é conhecido, mas um pouco mais difícil de resolver, porque muitas vezes nós só temos um conjunto de dados D, e produzir outros é muito custoso, demorado ou até inviável.

Nesses casos, há testes de significância estatística que podem nos ajudar quando precisamos argumentar que um sistema se mostra superior a outro conforme uma métrica. Eles tampouco nos dão garantias inquestionáveis, apenas fornecem evidências adicionais de que nossa conclusão pode (ou não) estar correta, e devem ser usados e interpretados com cautela. Há diversos testes, cada um com suas suposições que devem ser levadas em conta. Dois testes úteis em PLN são os testes não-paramétricos de *bootstrapping* e o de randomização em suas formas pareadas. Ambos se baseiam em simulações para computar as mesmas métricas em diversas amostrar dos dados e estimam um p-valor que indica a probabilidade de observamos a diferença entre A e B, supondo que eles são sistemas equivalentes (a hipótese nula). Quando essa probabilidade é baixa o suficiente, podemos rejeitar essa hipótese, em favor de haver uma diferença. Mas lembre-se: nunca "aceitamos" uma hipótese, apenas rejeitamos ou não rejeitamos a hipótese nula. Dessa forma, a palavra significante só deve ser usada na descrição se resultados se algum teste estatístico foi feito. Para uma exposição mais detalhada dos usos e limitações de testes de significância estatística em PLN, consultar (Berg-Kirkpatrick et al., 2012; Dror et al., 2018; Søgaard et al., 2014) e o subsequente livro (Dror et al., 2020).

18.4.8 *Baselines*

O termo *baseline* é comumente usado para dois tipos distintos de referência: a performance trivial e a performance do estado da arte atual. Sparck Jones (1994) usa *baseline* apenas para o primeiro caso, e *benchmark* para o segundo. Independente de como os chamemos, o importante é aprendermos os conceitos e suas funções.[15]

Uma *baseline* trivial serve para verificar se a performance de um sistema é pelo menos superior ao resultado que teríamos se fizéssemos predições aleatórias (por exemplo, jogando um dado ou uma moeda). Como a performance aleatória costuma ser baixa, há também a possibilidade de definir *baselines* triviais usando definindo um algoritmo que usa informações muito básicas sobre o problema ou que faz uma predição muito "crua".

Por exemplo, digamos que tenhamos um sistema que classifica textos de forma binária como sendo ou não sobre o tema de mudanças climáticas. Algumas *baselines* triviais seriam:

1. predizer que todos os textos são sobre mudanças climáticas
2. predizer que nenhum texto é sobre mudanças climáticas
3. jogar uma moeda: se der cara, classificar o texto como sendo sobre mudanças climáticas
4. se o texto contiver a palavra "clima", classificá-lo como sendo sobre mudanças climáticas

Suponhamos que a avaliação esteja sendo feita em um conjunto de dados contendo exatamente 50% de exemplos de cada classe. Nesse contexto, as *baselines* (a) e (b) atingiriam

[15] Esta seção se baseia em https://blog.ml.cmu.edu/2020/08/31/3-baselines/ e https://ehudreiter.com/2018/08/30/use-proper-baselines/.

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

performance de 50% de acurácia; a *baseline* (c) atingiria um número bem próximo a esse. Já a opção (d) é um pouco mais informativa, pois ela ao menos leva em conta uma informação (ainda que muito trivial) sobre o problema. Nesse caso, a performance dependeria da natureza dos dados: se houvesse muitos textos sobre previsão do clima, isso iria confundir as predições; mas se os textos que não tratam de mudanças climáticas fossem todos sobre programação, provavelmente essa *baseline* iria ter uma performance muito boa, pois "clima" seria um termo muito improvável nesse grupo. Nesse contexto, se implementamos um sistema que atinge acurácia de 40%, isso é pior do que simplesmente jogar uma moeda! Ou seja, um sistema, ainda que sofisticado, que não se sobrepõe sequer à performance de uma *baseline* trivial não capturou nada sobre o processo de classificação (ou há algo de errado em sua implementação).

Mas é preciso cuidado para interpretar *baselines*. Imagine que nosso conjunto de dados seja desbalanceado (o que é muito comum em fenômenos naturais), contendo apenas 15% de textos sobre mudanças climáticas. Implementamos um classificamos e mensuramos sua performance, observando que atingiu 85% de acurácia. É um valor bem alto, devemos comemorar? Quando inspecionamos os resultados do sistema, temos uma surpresa: todas as predições foram da classe negativa, ou seja, de que o texto não é sobre mudança climática. Como esses textos compõem 85% dos dados, nosso sistema apenas achou um "atalho" para atingir uma acurácia razoavelmente alta sem fazer qualquer classificação. Se tivéssemos computado a *baseline* da classe mais frequente, teríamos tido ferramentas para imediatamente suspeitar de que algo estava errado.

Baselines triviais servem para capturar esses casos, nos dando um "teste de sanidade" do sistema. Mas toda *baseline* deve ser interpretada em seu contexto, pois seus valores podem mudar conforme os dados. Elas são especialmente úteis quando estamos implementando um sistema para uma tarefa nova, para a qual não existem ainda resultados referenciais. Quando outros sistemas já existem, e nós queremos propor melhorias, eles se tornam nossas *baselines*. Nesse caso, são usados para checar se uma nova abordagem produz resultados melhores do que os já atingidos anteriormente. O uso de sistemas que performem pior que o estado da arte só se justifica se o sistema apresentar algum outro tipo de vantagem (por exemplo, uso mais eficiente de memória).

18.4.9 Ablação e Substituição

Por vezes, um sistema depende de diversas fontes de informação, ou de diversos subcomponentes ou de diversos tipos de *input*. É possível que ele fique tão complexo que não saibamos mais exatamente como os componentes interagem ou se tal nível de complexidade é realmente necessário para manter a performance. Nesse caso, podemos fazer estudos de ablação e de substituição (Cohen; Howe, 1988).

Em estudos de ablação, queremos entender qual é a contribuição das partes para o todo. Nesse caso, removemos um ou mais componentes, mantendo o resto constante, e verificamos como a performance é impactada ou quão bem o sistema "mutilado" consegue sobreviver sem ele(s) (Newell, 1975). Se a performance se mantém equivalente, pode ser interessante não usar esse componente para não termos uma complexidade adicional desnecessária. De forma similar, podemos também examinar o que acontece com a performance se, em vez de removermos um componente, o substituímos por outro.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

18.4.10 Análise de Dados

Os modelos de PLN treinados em dados podem captar padrões espúrios, vieses sociais e artefatos da anotação, de modo que as escolhas implícitas e explícitas que fazemos em relação os dados vão afetar a realidade quando esse sistema for posto para atuar (Rogers, 2021). Para avaliar um sistema que foi treinado a partir de dados e de anotações, precisamos entender bem a estrutura, as propriedades e a qualidade desses dados e que influência elas podem ter na performance do sistema. Para tanto, sugerimos a consulta à literatura sobre análise de dados e, se for possível, o diálogo com profissionais especializados nesse tema que é tão vasto. É ideal também a interlocução com profissionais capacitados para análise de dados linguísticos.

Mesmo que seja uma tarefa que vai além da competência desse capítulo, há algumas análises básicas que já são muito informativas para quem desenvolve o sistema. Por exemplo:

- Em quais idiomas estão os dados? Qual o gênero, o tema, o dialeto? São textos formais ou informais, em monólogo ou em conversas, escritos ou fala transcrita?
- Qual o tamanho e a diversidade do vocabulário? Quão representativo ele é do vocabulário usado nessa tarefa ou nesse gênero?
- Quão longos ou curtos são os textos ou as frases?
- Quais são os termos mais frequentes? Qual a proporção de *tokens* que ocorrem apenas uma vez (*hapax legomena*)?
- Que tipos de fenômenos parecem ocorrer com frequência? Por exemplo, referências, negações, perguntas, correferências, erros de digitação, ironias, humor, metáforas?
- Os dados contém algum tipo de conteúdo ofensivo ou sensível? É possível filtrá-los ou eles são parte essencial do modelo? Por exemplo, um modelo de detecção de conteúdo tóxico obviamente precisa ser treinado em um *dataset* com textos tóxicos (mas isso deve ser feito com cautela, ver (Kirk et al., 2022) e Seção 16.4.5). Já um sistema que gera textos precisa, de alguma forma, filtrar isso.
- Qual a distribuição das categorias anotadas? Qual classes são mais ou menos frequentes?

Geralmente, temos uma concepção muito idealizada dos dados em nossa mente, com várias ideias prescritivas de como os dados devem ser. Basta inspecionar qualquer *corpus* mais de perto para ver que a realidade difere bastante do formato platônico que imaginamos. Para um panorama mais abrangente quanto ao uso de dados em PLN, veja o Capítulo 16, (Bender; Friedman, 2018) e (Rogers, 2021).

18.4.11 Análise de Erro

Não devemos considerar apenas o valor nominal de uma métrica para avaliar um sistema. Uma boa avaliação deve, sim, usar métricas, mas em conjunto com uma boa análise de erro. Como o nome diz, o objetivo é analisar os erros, ou seja, buscar entender que tipos de erros (ou de acertos) estão ocorrendo nos *outputs* do sistema, quando, como, qual a relação entre eles e qual a relação deles com os *inputs* e com os mecanismos internos do sistema.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Para analisar os erros, devemos entender bem a estrutura dos dados e, se tivermos acesso, também a estrutura e os mecanismos do sistema (ver (Belinkov; Glass, 2019) para um resumo de métodos de análise de modelos). Analisar os erros requer olhar para os *outputs*. Em PLN, a análise de erro tem uma característica essencial: ou o *input* ou o *output* são ou contêm amostras de linguagem natural. O conhecimento linguístico deve, portanto, ser levada em conta na hora de se analisar os problemas do sistema.[16]

É muito arriscado confiar apenas nas métricas que o sistema computa: temos de pôr a mão na massa e abrir os arquivos contendo o que o sistema efetivamente produziu. Devemos tentar encontrar padrões e conexões entre os erros (por exemplo, pode ser que o sistema sempre se confunda quando há uma negação ou uma palavra que não está no vocabulário), agrupá-los e classificá-los, e buscar entender que componentes do sistema podem estar causando esse comportamento incorreto.[17]

Erros se propagam de um componente para o outro e de um sistema para o outro, por isso é essencial tentar mitigá-los. Entender por que eles ocorrem serve tanto para melhorar o sistema durante seu desenvolvimento quanto para informar usuários sobre suas limitações. Para saber mais, ver o capítulo 8 em (Freitas, 2022).

18.4.12 Avaliação Humana

Há aspectos da qualidade de um sistema muito difíceis de se expressar por uma métrica, principalmente aspectos linguísticos. Como mensurar fluência, naturalidade, gramaticalidade ou coerência? Muitas dessas propriedades não são binárias e envolvem subjetividade. Métricas automatizadas não conseguem capturar todas as nuances do julgamento e da interpretação humana.

Por isso, uma forma muito comum de avaliação, porém mais demorada e custosa, é usar seres humanos para julgarem a qualidade de um *output* específico ou quão bom é um sistema em geral. A avaliação humana pode ser feita tanto por *experts* quanto por leigos e pode avaliar tanto um *output* isolado quanto comparar *outputs* de sistemas diferentes. Esse instrumento permite medir aspectos subjetivos, preferências, compreensão e interpretação, e está mais bem alinhada com o propósito final de várias tecnologias de linguagem: utilidade para humanos que vão usá-la para algum fim. As medidas podem ser feitas por escalas Likert,[18] observação de comportamentos, questionários, interação com um sistema ou tarefas específicas (Shimorina; Belz, 2022).

Todavia, todo julgamento humano vai ter um viés, e as pessoas nem sempre estão de acordo. Para diminuir o risco de uma medida enviesada ou para capturar de fato a pluralidade inerente ao problema, podemos coletar a opinião de diversas pessoas acerca dos mesmos exemplos, de forma a ter uma distribuição de julgamentos humanos para cada instância que queremos avaliar. Veja o Capítulo 16 para algumas métricas de concordância.

Devido ao custo e à demora da avaliação humana, o desenvolvimento de uma tecnologia de linguagem pode combiná-la com avaliação automática, que pode ser repetida diversas vezes, deixando a humana para o final de etapas estratégicas durante o desenvolvimento e uso.

A avaliação humana deve ser feita de forma sistemática e prezar pela reprodutibilidade. Se resultados de avaliação humana são reportados para demonstrar que um sistema fun-

[16] Ver, por exemplo, as recomendações em https://naacl2018.wordpress.com/2017/12/19/putting-the-linguistics-in-computational-linguistics/.

[17] Para detalhes da taxonomia em tradução automática, ver a Subseção Métricas Humanas.

[18] https://pt.wikipedia.org/wiki/Escala_Likert

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

ciona bem, mas nós não sabemos como essa coleta foi feita e não conseguimos repeti-la, qual o valor dessa informação? De fato, há uma crise na área de PLN, reportada por Belz et al. (2023b) e Belz et al. (2023a), tanto acerca da dificuldade de repetir experimentos de avaliação quanto em campanhas de avaliação que contêm erros em sua implementação, pondo em cheque a validade dos resultados. Belz et al. (2023a) propõem alguns pontos que devem ser bem documentados. Shimorina; Belz (2022) também formulam uma *checklist* de aspectos (qualidade, propriedades, provocação de respostas, *design* do experimento, bibliografia e recursos) que devem ser bem pensados e documentados para realização de um experimento de avaliação humana. Para saber mais, recomendamos ter contato com a vasta literatura acerca de avaliação humana, por exemplo, nas edições do *workshop* HumEval.[19]

18.4.12.0.1 *Crowdworking*

Hoje em dia, se tornou usual usar mão de obra humana, através de micro-tarefas em plataformas de colaboração coletiva (*crowdsourcing*), tanto para gerar e anotar dados quanto para se avaliar sistema de PLN. De fato, é um recurso conveniente e de custo acessível, mas carrega o preço de diversas questões éticas. A pressão por criação de dados em escala massiva fomenta condições precarizadas de trabalho (Paullada et al., 2021). Além disso, há uma despersonalização e um desbalanço de poder na relação entre quem oferta trabalho e quem trabalha, o que gera situações de abuso (Leidner; Plachouras, 2017). Os problemas quanto ao uso de plataformas como a Amazon Mechanical Turk já vêm sendo expostos há mais de dez anos (Fort et al., 2011). Os impasses vão além de garantir um pagamento justo: há discussões sobre tratar os trabalhadores como sujeitos de experimentos, levando em conta direitos de privacidade, riscos psicológicos e grupos vulneráveis (Shmueli et al., 2021). Caso esse recurso seja utilizado, é também preciso lembrar que os trabalhadores podem estar em situação precarizada, com pressa por realizar o máximo de tarefas possíveis para garantir seu ganha pão. Por isso, a qualidade do serviço pode estar em risco, mas não devemos culpabilizar os indivíduos que ofertam trabalho. É preciso realizar testes de compreensão da tarefa e qualidade do trabalho, para garantir que a avaliação seja feita como se deve.

18.5 Métricas: Medindo a performance

Um dos principais procedimentos na avaliação é a medição de variáveis que capturem quão bom é o desempenho de um sistema. Métricas de avaliação são indicadores de performance que expressam o comportamento do sistema ou a qualidade de seus *outputs* de forma numérica, para tornar possível determinar quanto falta para ele atingir um nível máximo ou um nível desejado de desempenho. Além disso, servem para detectar problemas, quantificar sua acurácia e facilitar sua comparação com outros sistemas. Mas mensurações precisam ter validade interna, externa, estatística e conceitual no que se propõem a medir (Flake; Fried, 2020).

18.5.1 Usando métricas de forma responsável

Há métricas que podem ser implementadas como algoritmos, de forma a permitir, em parte, a automatização da avaliação. Todavia, a automatização só faz sentido se as métricas

[19] https://humeval.github.io/

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

tiverem correlação com os julgamentos humanos acerca das propriedades desejáveis de um sistema. Ou seja, uma métrica que melhora em termos numéricos sem ocorrer uma melhora correspondente no funcionamento do sistema na vida real não tem muita utilidade.

Lembremos também que métricas são apenas um substituto (*proxy*) do que realmente queremos medir (e nem sempre é possível), e devem sempre ser usadas com cautela (Belz, 2009; Thomas; Uminsky, 2022). Por exemplo, se uma professora aplica uma prova e computa uma nota, essa nota é usada para medir o desempenho da aluna naquela matéria, naquele momento. Mas ela não captura o que a aluna *realmente* sabe sobre a matéria, e muito menos deve ser usada para julgar o "nível de inteligência" dela, porque há muitas variáveis em jogo: a aluna poderia estar com dor de cabeça ou com uma preocupação pessoal no momento da prova, ter tido uma crise de ansiedade, saber muito mais sobre vários tópicos que não caíram na prova, mas não saber um tópico específico que compôs a maior parte das questões.

Toda métrica tem limitações, capturando apenas uma faceta de um problema. Métricas tampouco devem ser consideradas de forma isolada: elas sempre estão inseridas em um contexto. É preciso compreensão tanto do fenômeno e do domínio da tarefa quanto da definição da métrica em si para saber interpretá-la. Cada métrica tem seus próprios valores máximo e mínimo, intervalos, escopos e usos apropriados. Boas métricas devem fazer sentido no mundo real, com variações e valores sendo ancorados em processos da realidade.

Finalmente, métricas não são decretos. Em geral, especialistas propõem uma métrica para um determinado problema, ela passa a ser adotada pela comunidade e algumas se tornam bem notórias e populares, ao ponto de parecer que sempre estiveram lá e que são a única forma de se avaliar o sistema. Mas nem sempre devemos apenas seguir tradições. Por um lado, métricas tradicionais são úteis para se comparar um sistema novo com sistemas que já existem. Mas também é possível definir novas métricas ou adaptar métricas existentes para que se adéquem às necessidades do sistema, contanto que isso seja feito para melhoria da avaliação e não para mascarar deficiências do sistema ou distorcer resultados. Eventualmente, é também preciso abandonar métricas tradicionais e adotar novas medidas. De fato, novas métricas nascem o tempo todo.

Por exemplo, tomemos a famosa métrica BLEU (ver Subseção *Bilingual Evaluation Understudy* (BLEU)) para avaliação de geração de linguagem natural em tradução automática (Capítulo Tradução Automática): ela foi proposta quando o paradigma era usar modelos n-gram, de modo que ela se baseia diretamente em comparações entre n-grams. Hoje em dia, a geração de texto já não costuma se basear mais diretamente neles. A pertinência do BLEU já foi criticada, mostrando que pequenas variações nessa pontuação não têm significado na realidade, e que há métricas mais bem correlacionadas com avaliação humana. Ainda assim, muitos trabalhos continuam a reportá-la, e continua havendo uma busca por aumentos incrementais. Se um sistema alcança um BLEU de 60 e o outro de 63, há alguma melhoria real na qualidade? Essa métrica pode até ser útil para comparações com estudos mais antigos, mas a avaliação não deveria mais focar apenas em aumentar alguns pontos de BLEU[20].

Geralmente, uma boa avaliação não deve se concentrar em apenas uma métrica. É primordial ter uma coleção de métricas, cada uma delas capturando diferentes aspectos,

[20]Esse argumento é tecido, por exemplo, nos *blogposts* do professor Ehud Reiter: por que ainda usamos uma métrica de 18 anos atrás, em https://ehudreiter.com/2020/03/02/why-use-18-year-old-bleu/, e pequenas variações no BLEU não têm sentido na realidade, em https://ehudreiter.com/2020/07/28/small-differences-in-bleu-are-meaningless/.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

para serem interpretadas em conjunto; além disso, elas devem estar acompanhadas de análises qualitativas (Thomas; Uminsky, 2022).

A cada novo projeto, devemos sempre refletir: quais métricas são apropriadas para a avaliação do sistema? Qual seu valor mínimo, qual seu valor bom o suficiente para um determinado uso, qual seu valor máximo? Quanto de variação é necessário para que um sistema seja considerado melhor que o outro? É preciso também consultar a literatura atual para ter contato com as métricas e técnicas que vêm sendo usadas naquela área, quais métricas foram propostas mais recentemente. Em geral, começamos usando métricas mais tradicionais. Depois, podemos partir para métricas propostas mais recentemente. Ao longo do projeto, conforme o problema vai tomando forma e ganhamos expertise, podemos então definir ou adaptar métricas que capturem aquilo que é relevante para o uso que buscamos.

Além disso, a busca por otimização cega de uma métrica é prejudicial. Esse problema é expresso pela Lei de Goodhart, que postula que "quando uma medida torna-se uma meta (ou alvo), ela deixa de ser uma boa medida."[21] Thomas; Uminsky (2022) discutem essa questão, constatando que métricas podem ser (e serão) manipuladas, tendem a dar ênfase excessiva a aspectos de curto prazo e por vezes são coletadas em ambientes desfavoráveis ou danosos. Para mitigar ou minimizar esse problema, eles propõem que sejam feitas auditorias no algoritmo, com o envolvimento de diversas partes envolvidas ou impactadas pelo sistema no processo de desenvolvimento e avaliação, bem como com uso de diversas métricas e de análises qualitativas.

É importante não usar métricas sem a compreensão de como elas são computadas. Mas uma vez munidos dessa compreensão, não precisamos implementar tudo do zero. Há bibliotecas como a sklearn[22] e a TorchMetrics[23] que já trazem implementações bem testadas e otimizadas, dando mais garantias de que a computação está correta.

18.5.2 Métricas comuns em PLN

Vamos, agora, conhecer algumas das métricas mais comuns em PLN. Elas estão apresentadas no Quadro 18.3. Devemos conhecer as principais famílias de métricas para termos ferramentas com as quais começar a trabalhar, mas, como já foi exposto, novas métricas sempre podem ser definidas, e cada tarefa tem literatura especializada onde diversas opções são aplicadas. Para escolher quais são apropriadas para um problema, devemos pensar em quais são as relações entre o *input* e o *output*, e considerar suas propriedades, por exemplo, quais são linguísticas, categóricas, contínuas, sequenciais, etc. e como podemos medir se as respostas do sistema são boas ou desejáveis (Resnik; Lin, 2010), como discutimos na Seção 18.2.

Quadro 18.3: Algumas métricas usualmente utilizadas na avaliação de sistemas de PLN.

Área	**Métricas**
agrupamento	índice de Rand, índice de Rand ajustado
aprendizado	curvas de evolução ao longo do tempo
avaliação humana	escala Likert, ranqueamento, pontuação, tempo de reação

[21] https://pt.wikipedia.org/wiki/Lei_de_Goodhart
[22] https://scikit-learn.org/stable/index.html
[23] https://lightning.ai/docs/pytorch/stable/ecosystem/metrics.html

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

dispersão	média, moda, mediana, variância, desvio-padrão
distância	distância Levenshtein, taxa de erro de palavras, distância de Hamming
classificação	acurácia, precisão, revocação, matriz de confusão, medida F, área sob a curva COR
concordância	coeficiente kappa de Cohen, alfa de Krippendorff
geração de texto	BLEU, METEOR, ROUGE
outputs contínuos	erro quadrático médio, erro absoluto médio
probabilidades	perplexidade, entropia, entropia cruzada, divergência Kullback–Leibler, log-verossimilhança
ranqueamento	ranqueamento recíproco médio
recuperação de informação	TF-IDF
relações de informação	informação mútua, informação mútua pontual, correlação, covariância
similaridade, sobreposição	índice Jaccard, coeficiente de Dice, similaridade por cosseno

Essa lista não tem a pretensão de ser exaustiva pois há simplesmente muitas métricas e sempre estão surgindo outras. Além de métricas de aprendizado de máquina, há diversas métricas estatísticas (intervalos de confiança, p-valor), de desempenho de *software* e eficiência de algoritmos (uso de memória, velocidade, complexidade, tempo de resposta), de análise de negócios (satisfação de usuários, engajamento, custos) entre outras. Há, ainda, métricas que controlam para efeitos aleatórios, como o índice de Rand ajustado e as métricas de concordância de anotadores. Todas podem e devem ser levadas em conta em uma avaliação multidisciplinar.

18.6 Uso Responsável e Boas Práticas

Hoje em dia, o PLN é uma área bem empírica, principalmente o PLN baseado em dados e em técnicas de aprendizado profundo. Para fins de compreensão, transparência e documentação, há uma série de boas práticas que devemos seguir, e que facilitam a avaliação do sistema por parte de quem o usa ou regula. Vamos agora abordar alguns desses tópicos.

18.6.1 Como reportar resultados?

Um relatório de avaliação deve levar em conta o público-alvo, ou seja, ajustar o nível de tecnicidade e de detalhe conforme os diferentes consumidores: clientes e usuário/as, órgãos reguladores, pesquisadores, departamentos internos de uma empresa etc. Ainda assim, há informações que sempre devem ser incluídas e tudo deve ser feito de forma transparente. Ou seja, o relatório não deve apenas "vender" as partes boas do sistema, ele deve também discutir os problemas conhecidos, possíveis impactos e formas de mitigação de riscos.

Além de reportar resultados, documentos de avaliação também devem tratar das seguintes questões:

- Por que esse problema existe?
- Quais exemplos e dados estão disponíveis? Mostrar alguns desses exemplos.
- Qual é o *input* e o *output* do sistema (tipo, dimensões, propriedades)?

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

- Como o *input* é mapeado ao *output*, isto é, qual é a função definida matematicamente?
- Qual modelo é proposto?
- Que método de otimização foi utilizado?
- Como os parâmetros e hiper-parâmetros foram selecionados? Quais valores foram usados?
- Quais *corpora* e *datasets* foram utilizados? Qual seu idioma e tamanho? Qual a distribuição dos fenômenos relevantes?
- Porque esse modelo serve a essa tarefa?
- Quais métricas foram usadas? Como interpretá-as? Quais são seus valores, teóricos, de máximo e mínimo, e qual direção indica melhoria?
- Que experimentos foram rodados e com que propósito?
- Quais são as limitações, vieses e riscos?
- Para quais línguas o sistema foi projetado e avaliado?[24]

Lipton; Steinhardt (2019) elencam ainda algumas práticas que parecem ser tendência ao se reportar resultados em aprendizado de máquina mas que são prejudiciais: não distinguir entre o que é explicação e o que é especulação, não identificar as fontes dos ganhos empíricos (ou seja, dizer que um sistema que incorporou diversas modificações é melhor sem saber qual parte dele foi responsável pela melhoria), uso da matemática para ofuscar ou impressionar em vez de esclarecer, e mau uso da linguagem, misturando significados coloquiais com técnicos. É também interessante evitar usar verbos e conceitos que antropomorfizem os sistemas, como entender, decidir, pensar, ponderar, compreender ou sentir. O que quer que os sistemas atuais estejam fazendo para resolver as tarefas, não necessariamente estão seguindo processos cognitivos como os dos humanos, por isso a cautela na escolha do vocabulário é importante. Para uma discussão mais detalhada, ver, por exemplo (Watson, 2019), (Placani, 2024) e (Bender, 2024).

Um relatório de avaliação deve construir uma narrativa sobre o que foi feito, quais eram as hipóteses, até que ponto os resultados confirmam ou refutam essas hipóteses, qual a performance do sistema e quando ele não funciona bem. Descrever tudo isso pode virar uma longa novela, e expressar tudo em texto acaba ficando monótono. Por isso, é essencial usar outros meios complementares. Os principais são tabelas e gráficos. Eles devem ter um propósito, ou seja, fazer parte da narrativa, mostrarem os resultados de forma clara e transparente, sem dar margem a confusão (muito menos deliberadamente!).

Podemos usar as tabelas e gráficos para organizar e expor informações visualmente de forma a facilitar a compreensão do público, dando evidências e respaldos para interpretações dos resultados e das conclusões. Tabelas são mais apropriadas em contextos nos quais os números exatos dos resultados importam. Já gráficos são mais pertinentes quando se quer mostrar relações entre números ou sua evolução ou variação em função de algum parâmetro (por exemplo, tempo ou tamanho).

[24] Recomendação conhecida como Regra de Bender: https://thegradient.pub/the-benderrule-on-naming-the-languages-we-study-and-why-it-matters/. Lembre-se, um sistema que funciona para o inglês ou o português não necessariamente vai funcionar para outras línguas.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Além disso, a estética importa muito! A apresentação e o estilo das tabelas e gráficos devem facilitar sua leitura e interpretação, e, preferencialmente, serem bem organizados e agradáveis aos olhos. Há diversos tipos de gráfico, e cada um funciona melhor para um tipo de informação. O uso de cores deve ser feito levando em conta que existem formas de daltonismo, de modo que o contraste precisa ser mantido. Figuras devem vir acompanhadas de descrições textuais para as pessoas com impedimentos visuais. Um livro de acesso aberto muito útil sobre boas práticas de apresentação está disponível em https://clauswilke.com/dataviz/ por Claus O. Wilke. As regras de formatação da ABNT para trabalhos acadêmicos também podem ser úteis no contexto brasileiro, de modo a deixar as tabelas limpas e bem alinhadas, mesmo em contextos que permitam mais criatividade na apresentação dos resultados. Por exemplo, um relatório institucional vai seguir as regras de imagem da empresa em questão de cores e fontes. Na Figura 18.5, mostramos um exemplo de como apenas detalhes de formatação podem dificultar a visualização de uma tabela, deixando-a confusa e desorganizada.

Figura 18.5: Dois exemplos de formatação de tabelas. No exemplo A, há um excesso de linhas horizontais e verticais, mau uso do espaçamento devido à fonte muito pequena, e falta de uniformização nos comprimentos e alturas das células e no alinhamento. Além disso, a legenda não é nada informativa. O exemplo B traz exatamente as mesmas informações, mas o faz de maneira muito mais bem apresentável, facilitando a interpretação.

Exemplo A

	Tarefa A			Tarefa B		
	métrica 1	métrica 2	métrica 3	métrica 1	métrica 2	métrica 3
Baseline	0,56	123	-3,4	0,13	176	-12,7
Modelo 1	1,45	342	**-2,5**	0,25	76	**-8,5**
Modelo 2	**2,32**	**765**	-12,6	**4,92**	**653**	-21,8
Modelo 3	0,34	234	-24,6	1,55	23	-30,9

Tabela 1: Resultados da avaliação

Exemplo B

	Tarefa A			Tarefa B		
	métrica 1	métrica 2	métrica 3	métrica 1	métrica 2	métrica 3
Baseline	0,56	123	-3,4	0,13	176	-12,7
Modelo 1	1,45	342	**-2,5**	0,25	76	**-8,5**
Modelo 2	**2,32**	**765**	-12,6	**4,92**	**653**	-21,8
Modelo 3	0,34	234	-24,6	1,55	23	-30,9

Tabela 1: Performance dos três modelos avaliados. Para todas as métricas, quanto maior, melhor. Valores em negrito representam o melhor resultado em cada métrica e tarefa. O modelo 2 foi superior aos demais nas métricas 1 e 2, enquanto o modelo 1 teve melhores resultados na métrica 3, na qual os modelos 2 e 3 estão abaixo da *baseline*.

Todos gráficos e tabelas devem vir acompanhados de uma legenda que informe o que está representado neles e quais as principais conclusões podemos tirar dessa informação. A legenda e o gráfico devem formar uma unidade de informação que possa ser compreendida sem o auxílio do texto; o texto deve ser usado para explicações mais longas e elaboradas

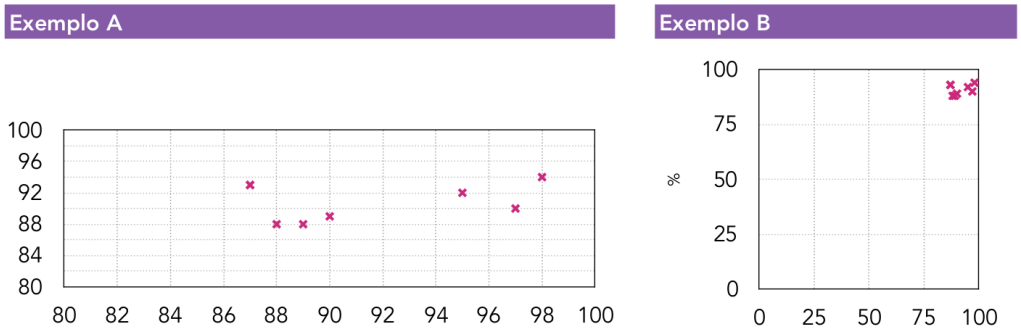
https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

dos detalhes e argumentos.

Evidentemente, esses instrumentos devem ser usados para informar, não para ludibriar. As legendas devem providenciar toda a informação necessária para a compreensão de quem lê, como unidades de medida, nomes das métricas e amostra utilizada. Os eixos de um gráfico devem conter todo o domínio da métrica, não apenas uma porção que convém para fazer os resultados parecerem melhores através de distorções. Se for para dar um *zoom*, os leitores devem ser claramente informados sobre isso, e preferencialmente deve-se usar um gráfico separado, para deixar claro primeiro o panorama geral. Mostramos como essa distorção ocorre na Figura 18.6.

Figura 18.6: Dois exemplos contento exatamente os mesmos dados. No exemplo A, os eixos têm tamanhos diferentes, mesmo sendo o mesmo intervalo. Além disso, não sabemos que métrica está representada neles, pois não há legenda. Como o intervalo começa em 80, temos a impressão de que os dados estão muito dispersos, e poderíamos induzir a ideia errada de que o sistema no canto superior direito é muito superior aos demais. O exemplo B deixa claro que as observações estão muito próximas umas da outras, pois temos o panorama completo de mínimo e máximo da porcentagem e a proporção dos eixos é igual.



18.6.2 Como tornar os resultados reprodutíveis?

Não adianta fazer uma avaliação de um sistema que ninguém consegue checar. Como as pessoas vão acreditar em alguns números em um documento? Para confiarmos nos resultados, é preciso ter acesso ao sistema, para poder verificá-lo, ou ao menos aos *inputs* e *outputs*, e o procedimento de como se deu a avaliação deve estar bem documentado. Em outras palavras, os resultados e sua avaliação devem ser reprodutíveis.

Reprodutibilidade tem se tornado uma dimensão primordial em projetos de PLN. Há tanto a possibilidade de se replicar resultados usando exatamente o mesmo sistema, dados, infraestrutura do original quanto tentar reproduzir resultados ou conclusões usando métodos ou contextos diferentes (Belz et al., 2021; Belz, 2022). Reprodutibilidade pode se referir a vários prismas: reproduzir uma conclusão, um achado e/ou valores de métricas (Cohen et al., 2018).

Há muito tempo já se discutem as dificuldades de se reproduzir experimentos em PLN, inclusive devido a pequenos detalhes como preprocessamento de dados e indisponibilidade de códigos (Fokkens et al., 2013; Wieling et al., 2018). A conferência COLING de 2018

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

propôs algumas diretrizes para melhoria da reprodutibilidade e replicação:[25]

- **Compartilhando dados e códigos**: sempre que possível, é essencial compartilhar dados e código. Isso fomenta a ciência aberta, permitindo que outras pessoas se beneficiem e possam confiar em nossos resultados.
- **Ultrapassar o estado da arte não deve ser um propósito máximo**: apenas tentar melhorar uma métrica não é informativo. Devemos buscar compreensão sobre o que produzimos.
- **Resultados negativos também importam**: é comum só vermos relatórios e artigos mostrando técnicas ou sistemas que funcionam bem (efeito conhecido como viés de publicação). Mas para cada sistema que funciona bem há uma série de tentativas que não funcionaram. Reportar tais esforços também é útil, para distribuir conhecimento e evitar que outras pessoas gastem tempo tentando fazer algo que já sabemos que não dá certo.
- **Avaliação**: a avaliação deve trazer conhecimento sobre como o sistema funciona, que fenômenos ou padrões ele captura, e compará-lo com alternativas.

Nos últimos anos, conferências de PLN também vêm usando *checklists* de reprodutibilidade, com quesitos sobre formalização do problema, escolhas do modelo, variação dos parâmetros, licença dos dados, versões de programas e descrição da infraestrutura utilizada.[26]

18.6.3 Como testar e documentar um sistema?

Há muitas referências em desenvolvimento de *software* sobre como implementar códigos que sejam fáceis de ler, encapsular, adaptar, estender e manter. Não entraremos em muitos detalhes aqui por questão de espaço, mas basta uma tentativa frustrada de tentar utilizar o código de outra pessoa e encontrar um grande labirinto caótico e mal sinalizado para nos convencermos do quão importante é a qualidade do código. Um código limpo e organizado facilita muito a avaliação de um sistema e a reprodução dos resultados.

Em particular, vamos mencionar dois aspectos que são essenciais para se analisar um sistema: sua documentação e a abrangência de seus testes. Documentação é essencial para compreendermos o que o código faz (ou deveria fazer), quais suas dependências e como cada parte se integra ao todo. Quanto aos testes, é evidente que se há um *bug* no código, o funcionamento do sistema vai estar comprometido, e portanto qualquer avaliação de seus resultados vai estar arruinada. No melhor dos casos, a avaliação pode apontar que há algo errado que requer investigação. Mas no pior caso, todos os resultados serão inválidos e incorretos, pois o sistema não está fazendo o que pensamos que ele está fazendo. Desta forma, formular um *script* de testes extensivos, incluindo casos nos extremos, é um pré-requisito para termos garantias de que o sistema está funcionando como deveria e podermos avaliar seus resultados.

Além disso, há recomendações de disponibilização de informação em formatos estruturados para dados e modelos, de modo a fomentar a documentação e a transparência (Bender; Friedman, 2018; Mitchell et al., 2019; Shimorina; Belz, 2022).

[25] http://coling2018.org/slowly-growing-offspring-zigglebottom-anno-2017-guest-post/

[26] Ver a chamada de artigos de EMNLP 2020 (https://2020.emnlp.org/call-for-papers), a proposta de Joelle Pineau (https://www.cs.mcgill.ca/~jpineau/ReproducibilityChecklist.pdf) e a atual política do sistema de revisões da ACL (https://aclrollingreview.org/responsibleNLPresearch/).

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

18.6.4 Como organizar experimentos?

Como vimos, avaliações são baseadas em experimentos. É muito fácil nos perdermos em uma vasta quantidade deles e não conseguirmos comparar e interpretar seus resultados. Recentemente, Ulmer et al. (2022) propuseram uma série de boas práticas para auxiliar nesse processo quanto a dados, códigos e modelos, experimentos e análises e publicações. Recomendamos a leitura do documento, que também traz diversas ferramentas que podem ser empregadas em cada etapa.

Para se comparar algoritmos, é recomendável variar um parâmetro por vez (Musgrave et al., 2020). Se, ao mesmo tempo, usamos mais dados, modificamos dimensões de *embeddings* e alteramos o algoritmo de otimização, não conseguiremos saber qual deles foi responsável por uma melhoria nos resultados.

Reportar apenas os resultados da melhor configuração do modelo não é uma boa prática, especialmente com modelos baseados em redes neurais artificiais, que podem ser sensíveis a pequenas variações nos parâmetros (Bansal et al., 2020). Por isso, é melhor reportar a média e o desvio padrão ou intervalos de confiança com base em uma distribuição de resultados com diferentes valores de um hiper-parâmetro, como a *random seed*, ou a *learning rate* (Dodge et al., 2019)

18.6.5 Atuando com ética

Sistemas em PLN têm impacto na realidade. A vida de outras pessoas pode ser impactada, positiva ou negativamente, pelos efeitos de um sistema. Com a ascensão das tecnologias de linguagem a cada vez mais áreas do cotidiano, faz-se necessário continuar a avaliar seus impactos sociais e regular seus usos para que não se tornem ferramentas nocivas e prejudiciais, especialmente nas populações vulneráveis. O tema de ética, tratado em mais detalhes nos Capítulos 20 e 21, é, portanto, também central na avaliação de sistemas de PLN. A pertinência dessa pauta aumenta cada vez mais (para uma recente revisão de literatura, ver (Madureira; Lasota, 2023)). No código de ética[27] adotado pela Associação de Linguística Computacional, encontramos também algumas diretrizes pertinentes à avaliação, como a honestidade e confiabilidade, busca por justiça, desviar-se de fazer mal, respeito à privacidade. Mais especificamente, há quatro pontos que versam diretamente em como avaliar:

- Entregue avaliações abrangentes e completas de sistemas computacionais e seus impactos, incluindo análise de possíveis riscos.
- Contribua apenas em áreas de sua competência.
- Garanta que o bem público seja a preocupação central durante todo o trabalho profissional de computação.
- Articule, encoraje a aceitação, e avalie o cumprimento de responsabilidades sociais pelos membros da organização ou grupo.

18.7 Para Concluir

Acreditamos termos acumulado argumentos suficientes para nos convencermos de que a avaliação é parte fundamental de qualquer tecnologia de linguagem, especialmente na era

[27] Disponível em https://www.acm.org/code-of-ethics.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

atual, em que há modelos sendo adotados sem haver uma compreensão plena sobre seu funcionamento (Bianchi; Hovy, 2021). Com a ascensão dos grandes modelos de linguagem, a avaliação continua a ser crucial, e é preciso ir além da avaliação baseada em *benchmarks*, focando especialmente na interpretabilidade e diagnóstico dos modelos (Li et al., 2023).

Avaliações devem ser feita de forma rigorosa, sóbria e integral. É muito fácil produzir uma avaliação enviesada, que não demonstre as reais limitações do sistema, mas isso é desleal e contraprodutivo. Não devemos ter receio de compartilhar resultados e recursos para avaliação, mesmo os que indiquem que o sistema não é perfeito. Tampouco devemos hesitar em sermos transparentes sobre os problemas do sistema.

Para concluir, vamos recapitular alguns dos principais pontos tratados neste capítulo em forma de uma lista de dicas e recomendações:

- Use os dados de maneira bem informada e consciente, buscando entender como eles afetam o modelo e os resultados.
- Lembre-se: a linguagem humana não tem fim, então todo conjunto de dados de avaliação é apenas uma pequena amostra da infinitude de possibilidades. Use uma amostra que seja adequada para a tarefa em questão.
- Extrapolações, generalizações e inferências devem ser evitadas a não ser que se tenha evidências suficientes para fazê-las.
- Não selecione apenas a parte boa dos resultados ou apenas alguns exemplos em que o sistema funciona bem na hora de julgar a qualidade ou reportar uma análise.
- Não deixe que o conjunto de dados de teste tenha qualquer influência na construção do modelo.
- Comunique seus resultados com esmero. Construa gráficos e tabelas informativos, compreensíveis e que não distorçam a realidade.
- Desmembre a parte de reportar resultados de sua interpretação, deixando claro o que é "objetivo" nos números e o que é a interpretação de quem os avalia.
- Busque rigor e formalidade, atuando de forma sistemática na hora de rodar experimentos e analisar resultados.
- Preze pela reprodutibilidade. Lembre-se das outras pessoas que vão usar ou testar o sistema e precisam de informação acessível.
- Pense criticamente sobre cada decisão tomada, tanto no desenvolvimento do sistema quanto nos aspectos avaliados. Use métodos pertinentes, verificando que seus pré-requisitos sejam atendidos na situação em questão.
- Reporte os resultados de forma honesta, metódica e clara.
- Reconheça as limitações do sistema e não tente apenas "vender o peixe".
- Não se esqueça da dimensão linguística do PLN.
- Otimizar métricas não deve ser nosso único objetivo.
- Conhecimento se produz com cooperação e colaboração, não com competitividade adversária.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

- Pense, leia, reflita e discuta sobre ética no uso e desenvolvimento de tecnologias de linguagem.

18.8 Exercícios

1. Abra um artigo acadêmico sobre PLN.[28] Examine sua seção de avaliação. Às vezes, há uma seção dedicada a esse tema, outras vezes ela acaba ocorrendo junto com a discussão dos resultados. Anote que tipo de informações você consegue identificar: termos, métricas, procedimentos, recursos. Como as autoras e autores avaliaram o sistema? Você considera a análise convincente para confiar no sistema, ou parece haver aspectos importantes que não foram discutidos? Há clareza e organização na análise? O artigo discute os casos de erro ou apenas foca em mostrar que as métricas melhoraram em relação a outro sistema? Há comparação entre diversas versões do mesmo sistema?

2. Qual a importância da avaliação em um projeto de pesquisa de PLN? E de um produto de PLN na indústria?

3. Suponha que você está em uma entrevista para um estágio ou emprego em uma empresa que oferece serviços de tradução automática de textos. A recrutadora lhe faz algumas perguntas:

   - Como você avaliaria um *software* de tradução automática de textos do inglês, japonês ou hindi para o português?
   - O que muda se forem sistemas separados ou unificados?
   - E se for do português para uma dessas línguas inglês?

   Como você responderia?

4. O que mudaria em sua resposta se fosse um sistema de tradução automática e simultânea de fala (ou seja, a tradução deve ocorrer enquanto a pessoa está falando)?

5. Em que a avaliação de sistemas de PLN se assemelha à avaliação de outros tipos de *software*? E em que ela difere?

6. O que faz uma avaliação ser adequada e justa? Como ela pode se tornar enviesada ou partidária?

7. Por que, como disse Karen Spärck Jones, não há uma forma única de avaliação em PLN? O que isso acarreta, e até que ponto convenções e protocolos podem ser úteis?

8. Escolha uma tarefa de PLN e discuta como avaliá-la conforme a abordagem de decomposição ilustrada na Figura 18.1.

9. Como ocorreriam as seguintes em sistemas de PLN em português:

   - avaliação intrínseca, automática, sumativa e objetiva de um sistema de anotação de *PoS*

[28] Por exemplo, na antologia da ACL (em inglês), disponível em https://aclanthology.org/ ou da Sociedade Brasileira de Computação, disponível em https://sol.sbc.org.br/index.php/stil/issue/archive.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

- avaliação transparente, manual e experimental de um sistema que gera legendas de imagens
- avaliação às escuras, extrínseca, por especialistas e subjetiva de um sistema de análise de agrupamento de textos por tópico

10. Busque um artigo científico e inspecione como os resultados são reportados. Quais das 10 dicotomias você consegue identificar na forma com que os autores e as autoras avaliaram o modelo?

11. Por que não se deve avaliar, e preferencialmente nem ter acesso, aos dados e exemplos de teste durante o desenvolvimento de um sistema?

12. Quais as vantagens e desvantagens de um mesmo conjunto de dados de teste ser usado por diversos pesquisadores ou profissionais da mesma área?

13. Descreva possíveis *baselines* triviais para os seguintes sistemas:

    - um sistema que classifica depoimentos de clientes de um restaurante em três classes: bom, neutro, ruim
    - um sistema que faz desambiguação de significado das palavras corredor, manga, medida, bala
    - um sistema que joga um jogo de adivinhar uma palavra, letra por letra

    Lembre-se, elas devem ser sistemas triviais, podendo fazer uso de variáveis aleatórias, que são usados para definir a performance mínima que qualquer sistema deve ultrapassar para ser melhor que nada.

14. Por que é primordial conhecer bem os dados com os quais trabalhamos?

15. Qual a utilidade da análise de erro?

16. Imagine que você está avaliando um sistema de diálogo que faz reservas de restaurante. O sistema produz respostas com base em cinco fontes de informação: o histórico da conversa atual, as palavras ditas pela usuária no último turno, as preferências da usuária, a localização da usuária, uma base de conhecimento com informações sobre restaurantes. Imagine que você vai fazer um estudo ablativo, removendo cada uma dessas fontes, uma de cada vez. O que você acha que aconteceria com a performance e com o comportamento do sistema em cada caso?

17. Como vieses dos dados podem influenciar a avaliação de um sistema? Que estratégias podem ser usadas para mitigar esse problema?

18. Quais os riscos de se considerar um *gold standard* como infalível? Por que isso é particularmente problemático quando se lida com linguagem natural?

19. Quando uma métrica tradicional deve ser substituída? O que acontece se cada sistema for avaliado com uma métrica diferente? E o que acontece se todos forem avaliados com a mesma métrica?

20. Como podemos mostrar que uma métrica tem correlação com julgamentos humanos?

21. Que práticas você consideraria desleais ou injustas ao conduzir uma avaliação e produzir um relatório dos resultados? Explique.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

22. Existem diversas formas de viés cognitivo. A Wikipedia tem uma lista de exemplos: https://pt.wikipedia.org/wiki/Lista_de_vieses_cognitivos. Escolha três deles e discuta como eles podem influenciar um avaliador ou avaliadora.

23. Qual é a sua parcela de responsabilidade sobre a dupla utilização (*dual use*) de uma tecnologia de linguagem em cada um dos casos abaixo? Como usar a avaliação para mitigar os riscos?

    - você trabalha em uma empresa de PLN e desenvolve um sistema
    - você é consultor(a) de um órgão do governo
    - você é pesquisador(a) e vai publicar um artigo científico
    - você, como indivíduo, disponibiliza um *software* em uma plataforma de *open source*

24. Em seu *blog*, o Prof. Ehud Reiter recomenda uma dose de ceticismo na hora de avaliar um sistema.[29] Quando e por que isso é importante? Como isso se relaciona à reprodutibilidade?

Agradecimentos

Este capítulo se baseia em duas edições do curso Avaliação de Sistemas de PLN, ministrados pela autora na Universidade de Potsdam, Alemanha. Para detalhes, ver (Madureira, 2021). Diversos tópicos foram inspirados nos *blogposts* sobre avaliação do Prof. Ehud Reiter, disponível em https://ehudreiter.com/. Agradeço às colegas Cláudia Freitas e Sheila Castilho, bem como às editoras do livro, pela valiosa revisão deste capítulo.

[29] https://ehudreiter.com/2017/11/21/guidelines-evaluating-ai-systems/

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Capítulo 19

Avaliação conjunta em português

Diana Santos
Cláudia Freitas

Publicado em: 20/11/2024

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

19.1 Apresentação

Avaliação conjunta[1] foi a forma como, na década de 2000, batizamos, em português, o que em inglês estava sendo chamado de *evaluation contest*, *evaluation campaign* ou *joint evaluation* – e, mais tarde, ***shared task*** – uma atividade que reúne vários participantes cujos sistemas ou modelos são comparados ao executar uma tarefa comum.

Ao usar a primeira pessoa, "nós", fazemos referência ao esforço feito pela equipe da Linguateca na época[2]. A Linguateca[3], projeto iniciado em 2000 pelo governo de Portugal com o objetivo de dar força e visibilidade ao processamento computacional da língua portuguesa (entendida como língua internacional), desde seu surgimento teve como uma de suas missões organizar avaliações conjuntas com o objetivo de avançar a área e avaliar seu progresso. Daí as primeiras avaliações conjuntas feitas em/para a língua portuguesa terem sido organizadas pela Linguateca, que também lançou o livro *Avaliação Conjunta: um novo paradigma no processamento computacional da língua portuguesa* (Santos, 2007b) com contribuições da maior parte dos pesquisadores de PLN brasileiros e portugueses da época.

Vejamos um exemplo muito simplificado de uma avaliação conjunta:

> Reunimos todos os interessados na área da resposta automática a perguntas. Combinamos em conjunto – organização e participantes – que tipo de perguntas usar, que tipo de respostas avaliar, e como, qual o conjunto de fontes (ou dados) em que os sistemas vão se basear, e qual o tempo entre a disponibilização das perguntas e o recebimento das respostas. A organização se responsabiliza por criar as perguntas e as (várias) respostas certas, e programas que contabilizam o desempenho dos sistemas. As respostas dos sistemas são processadas por esses programas, produzindo uma ordenação dos sistemas participantes.

Uma característica do modelo de **avaliação conjunta** é o enorme esforço organizacional – como veremos ao longo do capítulo – que nem sempre é possível despender, sobretudo se não houver instituições com esta função e com pessoal dedicado a estas atividades. A Linguateca tentou desempenhar esse papel, embora tivesse um pessoal e um horizonte

[1] Este capítulo é fortemente inspirado em Santos (2007a) e Santos (2021).
[2] Diana Santos é a coordenadora da Linguateca, e Cláudia Freitas integrou a equipe de 2007 a 2023.
[3] www.linguateca.pt

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

temporal reduzidos, não comparáveis aos do NIST (agência americana para a definição e manutenção de padrões) ou da DARPA (agência americana de financiamento de investigação ligada à defesa nacional)[4].

As datas mencionadas nos parágrafos anteriores podem surpreender leitoras/es mais jovens devido à antiguidade. Ao longo do capítulo, utilizamos várias referências que remontam ao século passado. Esta foi uma escolha consciente, motivada, por um lado, pela necessidade de sinalizar que boa parte dos pontos levantados já estão (ou estiveram) no radar do PLN há tempos. Por outro lado, porque, levando em conta os movimentos de idas e vindas do PLN (e da IA), é possível que o tipo de preocupação de uma época venha a ser retomado, ainda que com uma nova e mais moderna roupagem, como diria Karen Sparck-Jones em sua retrospectiva do PLN (Sparck Jones, 2001).

> Às vezes, inovação nada mais é do que antigas ideias que reaparecem com novos figurinos (...). Mas os novos trajes são mais bem feitos, com materiais melhores, e também mais modernos: assim, a pesquisa não está tanto girando em círculos, mas subindo em espiral.

Por fim, ao retomar as avaliações conjuntas realizadas há tempos, bem como avaliações realizadas mais recentemente, buscamos também mostrar a quem chega no PLN de língua portuguesa que nossa língua dispõe de muitos recursos públicos e disponíveis para avaliação. Com este capítulo tentamos trazer mais visibilidade a – e fomentar a utilização de – tanto material e conhecimento já produzido na e para a nossa língua, mas que se encontra disperso por várias publicações.

19.2 Avaliação conjunta: o que é e para que serve

Em uma **avaliação conjunta** pretende-se, acima de tudo, duas coisas:

1. Estabelecer **em conjunto** os objetivos e os fatores de sucesso de uma dada atividade;
2. Fomentar o **diálogo** entre os diversos atores da área.

Além disso, é também possível e desejável que uma avaliação conjunta deixe como legado, como uma espécie de efeito colateral, recursos (dados, métricas, formas de avaliação etc.) que permitam a novos grupos de PLN não precisar começar do zero nessa área e que possibilitem a todos os membros da comunidade a utilização e melhoria do trabalho já feito.

Assim, uma avaliação conjunta difere de uma avaliação "não-conjunta" na dimensão colaborativa/compartilhada que pressupõe. O que se pretende é melhorar em conjunto o estado de uma determinada área, evitando reinventar a roda, aumentando o número de trocas científicas entre grupos distintos, e produzindo padrões de funcionamento que evitem, a partir dali, que grupos novos tenham que começar de novo o processo de pensar a avaliação. O objetivo final é incentivar a investigação a respeito de uma dada área ou tarefa, através da comparação de vários sistemas/modelos e com base em recursos e tarefas comuns.

Como explicam Voorhees; Tice (2000), o **primeiro objetivo** de uma avaliação conjunta é promover a investigação numa dada tarefa. Um **segundo objetivo** é investigar se

[4] Para uma descrição da Linguateca em fases distintas da sua existência, ver Santos (2000) e Santos (2009). Tendo o financiamento acabado em 2011, a Linguateca transformou-se num projeto voluntário.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

a metodologia de avaliação é adequada, e se através dela é possível definir recursos de avaliação reutilizáveis, visto que "coleções de teste reutilizáveis, que permitem que os pesquisadores experimentem com ela e recebam uma resposta rápida sobre a qualidade de métodos alternativos, são fundamentais para avançar o estado da arte." (Voorhees; Tice, 2000, p. 207). É neste sentido que podemos reconhecer o próprio *design* da avaliação como pesquisa.

Finalmente, é preciso notar que só se pode começar a aplicar o modelo de uma avaliação conjunta quando há mais de um grupo interessado, e mais de um sistema numa dada área. Antes disso, a única forma de avaliação possível é a autoavaliação. Esta observação espelha sobretudo uma questão de prioridades: é compreensível que não haja recursos humanos para organizar tal tarefa, se ainda há tanto que fazer em outras áreas já com sistemas a funcionar.

Anterior à popularização das avaliações conjuntas, o PLN contava com as seguintes formas de avaliação (Capítulo 18):

- **autoavaliação**: algum ator (pessoa ou grupo reduzido) criava um sistema para ilustrar um método ou uma teoria e, quando muito, publicava resultados de avaliação segundo as suas próprias premissas, sem garantia de replicabilidade ou de verificação por outros grupos;

- **modelo empresarial**: uma empresa desenvolvia um sistema para um dado negócio, provavelmente com base numa autoavaliação, e a verdadeira avaliação era feita pelos utilizadores na forma de retorno (feedback) a ser incorporada em novas e melhores versões do sistema, caso a empresa decidisse continuar o desenvolvimento.

- avaliação baseada em **quadros de classificação** (***workbenchs***): apesar de estar muito em voga por causa dos grandes modelos de linguagem, é um tipo de avaliação que existe desde antes das avaliações conjuntas. No entanto, diferentemente das avaliações conjuntas, não há controle sobre o que os sistemas já sabem e/ou já treinaram, sendo relativamente fácil melhorar resultados depois de já ter visto os dados. Ou seja, diferentemente das avaliações conjuntas, não há como garantir que todos sistemas/modelos que se utilizam do *workbench* para avaliar estão em igualdade de condições, e que estão vendo o material pela primeira vez, o que torna os resultados artificialmente bons, sem que esse bom desempenho se reflita em outros conjuntos de dados ou em aplicações reais.

- avaliação baseada em **painel de especialistas**: reunia-se um conjunto de especialistas que se pronunciavam individualmente sobre a qualidade de um resultado, como feito nos primórdios da tradução automática (King, 1996)

Uma limitação das duas primeiras maneiras de avaliar é que o conhecimento obtido não é reutilizável por uma comunidade científica (no sentido lato, incorporando também desenvolvedores e testadores, e não apenas pesquisadores). Muitas vezes, esse conhecimento morre nas organizações (universidades ou empresas), devido à mobilidade dos investigadores ou ao encerramento dos projetos. No melhor dos casos, mantém-se a propriedade e o conhecimento (por alguns chamada "vantagem competitiva") no seio de um único grupo. Já na **avaliação conjunta**, como indicamos, é a comunidade como um todo que se beneficia e evolui, somando os esforços.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

19.2.1 Características e ingredientes de uma avaliação conjunta

A seguir listamos as **características** principais de uma avaliação conjunta:

- existência de um **debate** inevitável entre os participantes, buscando um consenso e levando ao conhecimento mútuo de diferentes pontos de vista e de diferentes atores com preocupações distintas;
- identificação de **problemas** a resolver e de problemas já resolvidos;
- definição de um conjunto de **tarefas** objetivas que os sistemas devem efetuar;
- identificação de **diferenças** irredutíveis, e de zonas cinzentas, ambas importantes para a documentação da área e do seu progresso.

De maneira complementar, acreditamos que uma avaliação conjunta tem normalmente os seguintes **ingredientes**:

- vários **sistemas** participantes;
- uma **organização** com conhecimento do assunto e reconhecida idoneidade, a fim de garantir que todos os participantes se encontrem em igualdade de condições;
- uma forma prática de produzir **resultados** e de compartilhá-los com os participantes e com a comunidade;
- a possibilidade prática de **comparar** esses resultados (por outras palavras, os resultados produzidos pelos vários sistemas/modelos têm de ser minimamente comensuráveis).

Quanto ao último ponto, é importante lembrar que uma avaliação conjunta pressupõe uma aplicação num dado contexto. Considerando, por exemplo, uma avaliação de "sistemas de compreensão de texto", estão envolvidos vários tipos de textos, um ou vários domínios fixos. Por isso, sistemas/modelos diferentes, mas igualmente "bons", podem ser incomensuráveis, se acontecer de um deles trabalhar sobre acórdãos da Procuradoria da República, outro sobre resumos de revistas médicas e outro ainda sobre anúncios. Ou seja, pode não ser possível uma tarefa igualmente "neutra" para todos os sistemas/modelos e, ao mesmo tempo, justa para os avaliar. Há, no entanto, a possibilidade de fazer a união dos tipos de texto e problemas relevantes para cada grupo, e medir numa avaliação conjunta quanto os sistemas pioram quando aplicados a outros ambientes. A prática das avaliações conjuntas provou que a especificação, em comum, de um subconjunto de questões e respostas sobre uma mesma base permite uma reflexão sobre os problemas concretos, a forma de avaliação, e as vantagens e desvantagens de diferentes opções, sendo incomparavelmente mais rica do que uma argumentação teórica sobre um conjunto de princípios independentes da aplicação.

Por fim, as seguintes condições são altamente desejáveis:

- os participantes se reúnam em um ou vários **encontros**, presenciais ou virtuais, para que a troca de ideias e a consequente fertilização cruzada possa acontecer da melhor maneira.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

- as **agências de financiamento** considerem os resultados e a participação nestas avaliações não apenas como índice do grau de maturidade da área, mas também (e principalmente?) como uma atividade que devem absolutamente financiar.

O primeiro item é importante porque a documentação publicada online nem sempre é submetida a um escrutínio cuidadoso por parte dos participantes, e é preferível uma exposição oral que permita esclarecer dúvidas imediatamente. O debate entre os participantes e a organização é uma etapa essencial da preparação das avaliações conjuntas, que garante que todas as opções sejam entendidas e aceitas e evita que os sistemas participem com outro modelo conceitual que não o subjacente à própria avaliação conjunta em questão. No entanto, vemos que isso nem sempre é possível, sobretudo quando as avaliações já estão incorporadas em uma infraestrutura maior, como foi o caso do CLEF (Seção 19.8) ou do IberlEF (Seção 19.11).

Do mesmo modo, também é útil estabelecer uma discussão após a própria participação, para reconhecer o que correu bem e o que poderia ser melhorado, e para reforçar a ideia de que uma avaliação conjunta não deve ser considerada como um acontecimento isolado, mas sim como um elo na cadeia de progresso de uma área, muito provavelmente levando a outras edições com um aumento gradual dos desafios oferecidos aos sistemas participantes.

Quanto ao segundo item, reconhecemos que se trata sobretudo de um desejo, e que em geral temos pouca influência sobre isso.

Idealmente, uma avaliação conjunta tem várias edições, como ilustrado na Figura 19.1. Além disso, para ajudar a fixar os parâmetros, pelo menos nas edições iniciais, costuma-se usar um ensaio (*trial* ou *dry-run*, em inglês). Isto significa que uma avaliação conjunta é sempre um acontecimento que decorre num período de tempo alargado.

19.3 Por que participar de uma avaliação conjunta

Ao participar da delimitação de um problema e da discussão relativa às melhores maneiras de aferição desse problema, passa-se, mais tarde, a ter um ambiente que permite separar o desenvolvimento de um sistema (a cargo do grupo) do teste de hipóteses (que é fornecido e gerido por uma organização externa).

Além disso, a existência de prazos externos, o **conhecimento** dos resultados de outros sistemas, e a facilidade de **comparação** e acesso ao estado da arte e ao resultado mínimo garantido (*baseline*)[5], são outros motivos que tornam a participação em avaliações conjuntas uma experiência valiosa.

É também preciso sublinhar que a existência de uma comunidade de pessoas que se reconhecem num mesmo objetivo e com as quais é possível discutir ideias e observar as consequências de opções diferentes, levando a uma fertilização cruzada e a um reconhecimento público, é extremamente importante para transformar uma atividade (neste caso, o processamento computacional da língua portuguesa) numa área científica respeitada e com visibilidade.

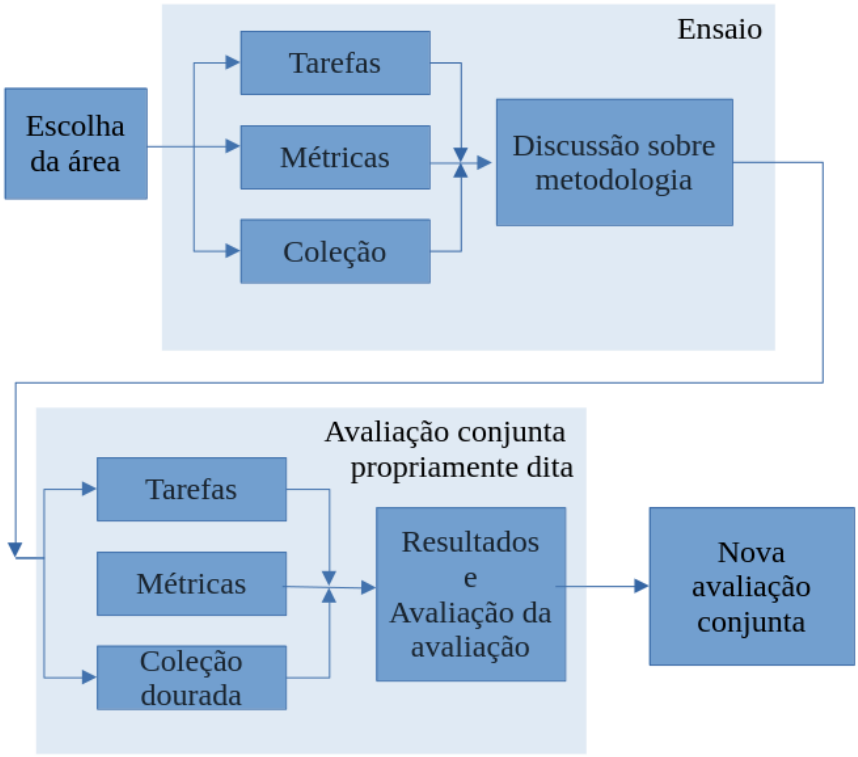
Para que uma avaliação conjunta possa conseguir isto, é no entanto preciso que se baseie em procedimentos rigorosos e consensuais, que abordaremos na próxima seção.

[5] O resultado mínimo garantido é o resultado obtido com um algoritmo ingênuo, fácil de implementar e ao alcance de todos. Só resultados acima desse mínimo são relevantes, visto que, de contrário, o melhor seria usar o tal sistema básico. Ver também 14.4.8.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Figura 19.1: O desenrolar de cada edição de uma avaliação conjunta.



Fonte: Adaptada de Santos (2007a, Fig 1-1)

19.4 Como implementar uma avaliação conjunta

19.4.1 Cartografia do problema e da tarefa

Uma questão pertinente, mas que frequentemente passa despercebida, é a própria **possibilidade de medir o problema**, antes mesmo de avaliar o desempenho na solução.

Sem sabermos a dificuldade (ou entropia, ou perplexidade) de uma dada tarefa, uma avaliação não pode ser útil, e os números são não-interpretáveis. Outra observação extremamente relevante diz respeito à **língua** e à possibilidade de **comparação com outras línguas**. Uma tarefa resolvida para uma língua pode ainda não estar resolvida para o português. Ou, ainda, nem todas as tarefas precisam ser relevantes para todas as línguas. A explicitação de sujeitos omitidos (sujeitos ocultos), pode ser um desafio para a língua portuguesa, mas certamente não o é para a língua inglesa, por exemplo.

A comparação com outras línguas é muitas vezes uma preocupação de pesquisadores e proponentes de avaliações[6], mas pode ter como consequência a desvalorização da originalidade na elaboração dos recursos ou de medidas de avaliação. Na nossa opinião, o importante é tentar medir o problema e a solução na língua portuguesa, em primeiro lugar.

É fundamental, ao organizar uma avaliação conjunta, definir uma tarefa o mais possível neutra e compreensível por pessoas, mesmo que não seja necessariamente de utilidade para o utilizador final. Ou seja, tem de estar **clara** e **bem definida**, e poder ser **compreendida**

[6]Como pode ser observado dada a proliferação de *shared tasks* multilíngues.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

(e repetida) por um conjunto de pessoas diferentes.

Por outro lado, a avaliação de uma tarefa bem definida não tem de ser unívoca, no sentido de comparar cegamente com um conjunto de soluções únicas. Na criação de resumos ou na avaliação de tradução, por exemplo, se, por um lado, é possível avaliar, por outro lado, não é possível definir uma solução única.

19.4.2 Criação dos recursos

Na prática, há duas formas de organizar uma avaliação conjunta:

- criar um **recurso padrão-ouro**, feito total ou parcialmente (por meio da revisão de algum resultado automático) por pessoas desempenhando a tarefa que será feita pelo sistema ao longo da avaliação; ou usar para tal um recurso já criado anteriormente.
- confiar de algum modo na qualidade dos sistemas participantes, cuja soma de resultados, ordenados, leva a um conjunto de respostas (chamado monte ou "pool"), que serão posteriormente "apenas" avaliadas pela organização, de modo a constituir um **recurso de avaliação criado a posteriori** [7]. (Para discussão das formas e consequências desta forma de criação de recursos, ver Zobel (1998) e Braschler; Peters (2003).)

19.4.2.1 Possibilidade de reuso

Uma questão associada à criação de recursos é a possibilidade de reutilização destes. Este é um ponto delicado, relacionado aos pressupostos teóricos de uma avaliação (conjunta ou não), mas, como também se aplica a avaliações conjuntas (que implicam em geral mais trabalho na criação dos recursos), merece ser discutido aqui: recursos criados com uma dada abordagem teórica não podem ser simplesmente transpostos para outra abordagem, sob perigo de prejudicar gravemente os sistemas avaliados.

Para dar um exemplo concreto: os lugares no HAREM (Seção 19.7) eram apenas aqueles que seriam interpretados como lugares por um leitor, não entidades que representassem um país ou uma cidade como organizações ou grupos de pessoas. Se se transferissem os recursos do HAREM para uma avaliação como foi feita no CoNLL (Sang, 2002) ou no ACE (Doddington et al., 2004) para o inglês, em que entidades geopolíticas deveriam ser classificadas como lugares, ter-se-ia de reclassificar todos esses casos, sob pena de prejudicar os sistemas que concorriam com esses pressupostos. Veja-se Santos (2007c) para uma discussão detalhada das diferentes filosofias de avaliação de reconhecimento de entidades mencionadas.

Concluindo, **não há recursos neutros** e, portanto, embora indiquemos a reutilização sempre que possível do muito trabalho que já foi feito, resumido na Seção 19.15, é preciso sempre compreender o contexto e a definição da tarefa subjacente.

Para isso, nunca é demais insistir numa documentação detalhadas das opções tomadas, para que o reuso possa ser feito de maneira informada e consciente.

[7] Esta estratégia pode ser usada em avaliações conjuntas para a tarefa de encontrar documentos e/ou respostas específicos (i) em coleções com milhões de documentos, o que é uma tarefa humanamente impossível, ou (ii) em grandes coleções de textos, mas com uma equipe limitada.

https://braspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

19.4.3 Medidas de avaliação

As medidas por que se pauta uma avaliação conjunta são a sua pedra de toque. Ver também a Seção 18.5.

Conforme já referido, medir só faz sentido se a medida for adequada, isto é, se os valores numéricos apresentarem uma relação forte com as qualidades que se pretende aferir. A relação entre as diferenças nas medidas e a importância ou interesse dos casos resolvidos, ou não resolvidos, pelos sistemas é também pertinente: Voorhees; Tice (2000) notam que **nem todos os casos são iguais**, e que não se faz justiça a um sistema se o mesmo peso é dado às respostas fáceis e às difíceis. Assim, os casos de discordância entre os juízes humanos e o "gabarito" automático que implementaram acabam por ser, precisamente, aqueles em que a resposta é mais complicada, e portanto quando é mais necessário ter um julgamento humano.

Destacamos, todavia, que as questões aqui referidas, embora extremamente pertinentes, só podem ser colocadas – e discutidas – depois de, idealmente, pelo menos uma avaliação conjunta ter ocorrido, para que os dados possam ser trabalhados e as condições da avaliação conjunta avaliadas e melhoradas. Antes disso, contudo, é preciso definir que problema(s) atacar, e como caracterizá-lo(s).

19.5 Como avaliar uma avaliação conjunta

Sugerimos que, de um modo geral, uma avaliação conjunta poderá ser avaliada

- através da **participação** que suscitou e da adesão que teve;
- através do **material** (ferramentas, recursos e publicações) a que deu origem;
- através do **impacto** que teve nos grupos participantes e na área em geral;
- através da **reutilização** (ou não) dos recursos nela ou por ela criados;
- através do **impacto** que esse acontecimento teve no panorama científico nacional ou internacional.

Em uma escala de tempo mais alargada, o **verdadeiro impacto** irá se refletir na constatação de que dali por diante as pessoas passaram a usar os recursos e os conhecimentos produzidos na avaliação conjunta para identificarem os problemas que os seus sistemas/modelos vão ter que resolver, medirem o desempenho dos ditos, e melhorarem o estado global da área em causa.

19.5.1 Críticas e limitações

É preciso, no entanto, esclarecer que há críticas válidas a esta forma de avaliação, ou que o seu abuso pode levar a alguns inconvenientes.

Por um lado, há o perigo de os participantes desenvolverem os seus **sistemas para participar nas competições e não para funcionar na vida real**. Isto tanto pode acontecer pela concentração numa área específica, como considerando cuidadosamente as questões medidas por uma dada avaliação conjunta e desprezando as que não são objeto de comparação.

Por outro lado, a existência dos chamados "cotovelos", que indicam que se chegou a uma fase em que a maior parte dos sistemas já têm um desempenho aceitável, sinaliza que pode

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

ser hora de mudar radicalmente o cenário da avaliação. Hirschman (1998), ao descrever o MUC[8], comenta que, de fato, de uma edição para a próxima, deve-se sempre aumentar os desafios e evoluir na avaliação; Gaizauskas (2003) sugere que o estudo de avaliações anteriores sugere novas métricas que, em si, sugerem novos tipos de avaliação... ou seja, uma atividade de avaliação conjunta deve **progredir ao longo das várias edições**, aprendendo com o passado e com os erros e sucessos dos sistemas/modelos participantes.

Hirschman também chama nossa atenção para quais **partes interessadas** estão envolvidas em uma avaliação específica: o setor, os usuários, os pesquisadores ou os financiadores? As avaliações até agora organizadas para o português privilegiaram a qualidade e os métodos, mas não a eficiência ou o preço, e muito menos a sustentabilidade do ponto de vista de uma agência de financiamento.

O principal perigo de uma avaliação conjunta talvez seja, contudo, que as **medidas ou o cenário não sejam adequadas ao problema** ou privilegiem um certo tipo de funcionamento. Um exemplo seria limitar a priori, em uma tarefa de resposta automática a perguntas (RAP, QA), o tamanho da resposta em *bytes* ou *tokens*[9]. Afinal, que significado pode ter o tamanho de uma resposta?

De modo análogo, é importante verificar a **relação (e relevância) da tarefa para a aplicação** a que está subjacente, sob pena de a avaliação não ser válida. Por exemplo, é justificável definir desambiguação de sentidos como uma tarefa separada do processamento computacional de uma língua? Wilks (2000) critica a própria pertinência de avaliar separadamente esta tarefa, como foi feito no Senseval[10], na esteira do que já Kilgarriff (1997) tinha argumentado.

Outra objeção é a de que as **avaliações conjuntas reduzem o valor de um sistema/modelo a um número**, ou a uma série de números, não necessariamente facilmente interpretáveis, compreensíveis ou mesmo intuitivamente adequados.

As medidas têm de ser intuitivamente interpretáveis para todo o processo fazer sentido. Contudo, nem sempre é fácil eleger uma medida quando há muitas variáveis que é preciso considerar. A este respeito, mencione-se que avaliações conjuntas (canônicas) como o TREC[11] e o CLEF (descrito na Seção 19.8) usam mais de uma dezena de medidas, e em avaliações para o português como o HAREM e o Págico, usamos também mais do que uma.

Do mesmo modo, é muito importante **compreender a junção dos resultados**, quase nunca simples. Tome-se o seguinte exemplo: num campeonato por equipes, temos um conjunto de cinco alunos com nota 1 em uma dada disciplina, e uma outra equipe com uma aluna com nota de 5 mais quatro alunos com nota 0. Embora cumulativamente ambas as equipes atinjam o mesmo valor se a medida for "soma das notas" (5) ou "média das notas" (1), as duas equipes – ou os dois sistemas – estão longe de ser semelhantes.

Nas avaliações conjuntas é importante que a tarefa escolhida – o conjunto de tópicos, o

[8]MUC, *Message Understanding Conference*, é uma avaliação conjunta para PLN em inglês, mais tarde expandida para mais línguas, e que foi fundamental para a definição de avaliação conjunta para o português.

[9]Foi o que aconteceu no TREC QA, uma avaliação conjunta de respostas a perguntas, o que foi amargamente discutido na comunidade de PLN da altura.

[10]O Senseval (https://web.eecs.umich.edu/~mihalcea/senseval/overview.html) foi uma avaliação de sistemas dedicados à tarefa de desambiguação de sentidos, e que deu origem ao SemEval, avaliação dedicada a análises semânticas (https://en.wikipedia.org/wiki/SemEval).

[11]O TREC, *Text REtrieval Conference* (Conferência de Recuperação Textual), teve início em 1992 e, depois disso, passou a desempenhar anualmente o mesmo papel na recuperação de informações que o MUC desempenhou na extração de informação, consulte Harman (1998).

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

conjunto de perguntas – seja suficientemente **variada** e **representativa** para que possam de fato avaliar um sistema. A questão de garantir um número suficiente de perguntas e de confirmar se as diferenças entre sistemas são estatisticamente significativas foram tratadas por exemplo por Buckley; Voorhees (2000) para o TREC, por Chinchor (1992) para o MUC e por Voorhees; Tice (2000) para respostas a perguntas no TREC. Para o português e para o HAREM, ver Cardoso (2006). Ver também a Seção 18.4.7.

Ainda uma última crítica – que embora pertinente, não invalida o modelo de avaliação conjunta –, é que as avaliações conjuntas em geral não consideram a "experiência do utilizador", nem outras qualidades "periféricas" (ou mais dificilmente mensuráveis), tais como qualidade científica, qualidade da documentação, forma de apoio ao utilizador ou maneira como recuperam de erros, inovação, consistência, facilidade de manutenção, legibilidade do código etc.

Esta observação só indica que a aferição do **valor** e da **utilidade de um sistema/modelo não se esgota com uma avaliação conjunta**: outros tipos de avaliação podem também ser necessários. Contudo, em alguns casos, pode se juntar a uma avaliação conjunta um painel de especialistas que pontuam algumas destas propriedades (como nas Morpholympics (Hausser, 1996) organizadas por Hausser em 1994), assim como se podem idealizar avaliações conjuntas com um perfil claro de utilizadores em mente.

Passamos agora à apresentação das avaliações conjuntas para a língua portuguesa que organizamos ou conhecemos, utilizando um recorte estritamente temporal.

19.6 Morfolimpíadas: análise morfológica (2003-2004)

A primeira tarefa para a qual organizamos uma avaliação conjunta foi a **análise morfológica**. As Morfolimpíadas foram motivadas pelo fato de a língua portuguesa ter um paradigma flexional rico, ao mesmo tempo que, na época, havia um número razoável de grupos que tinham algum tipo de processamento morfológico em seus sistemas.

Os participantes receberam os materiais de teste (613 textos) em três formatos: texto corrido, uma unidade ("*token*") por linha, e lista de tipos. Incluímos, nesses textos ou listas, os itens de nossa lista padrão-ouro[12]: 200 formas que, por sua vez, correspondiam a 345 análises diferentes. Os sistemas participantes deveriam então retornar todos os três materiais de teste morfologicamente analisados, sem saber quais palavras seriam avaliadas.

Embora a análise morfológica pareça uma tarefa simples, para produzir uma comparação justa de sistemas com objetivos muito diferentes foi preciso elaborar resultados de acordo com três eixos:

1. análise morfológica propriamente dita
2. verificação ortográfica
3. radicalização[13]

[12] Na Linguateca usamos "lista dourada" (*golden list*) e "coleção dourada" (*golden collection*). É importante explicitar esta terminologia aqui para tornar compreensíveis os artigos anteriores e permitir a sua subsequente procura bibliográfica...

[13] Esta tarefa é chamada em inglês *stemming* e é usada em recuperação de informação; é parecida com lematização mas não pressupõe análise linguística.

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

Além disso, talvez o resultado mais interessante das Morfolimpíadas tenha sido a constatação de que havia – e pensamos que ainda há – uma considerável discordância teórica sobre a forma adequada de fazer a análise morfológica do português, conforme descrito em Santos et al. (2003).

E, de maneira complementar, podemos dizer que as Morfolimpíadas levaram a um grande aprendizado sobre como avaliar, cooperativamente, uma determinada aplicação.

Todos os resultados, dados e programas (em Perl) usados para organizar as Morfolimpíadas foram divulgados na página da Linguateca dedicada a esta avaliação - onde ainda podem ser encontrados[14].

19.7 Primeiro e Segundo HAREM: avaliação de entidades, relações e expressões temporais (2005 e 2007-2008)

O HAREM é um exemplo paradigmático de uma avaliação conjunta que condiz com a maior parte das afirmações genéricas feitas no presente capítulo.

19.7.1 Reconhecimento de entidades mencionadas

HAREM é a sigla (recursiva) de "HAREM - Avaliação de Reconhecimento de Entidades Mencionadas". A tarefa ainda não havia sido tentada por nenhum sistema para a língua portuguesa, portanto, esse foi um caso claro de tentativa de alavancar uma área desconhecida para o português. Conforme indicado na página de motivação do Primeiro HAREM, o reconhecimento de entidades mencionadas (REM, ou NER, em inglês)[15] era uma tarefa "leve" em termos de carga teórica (com isso queremos dizer que não havia campos diferentes com posições irredutíveis entre os interessados, diferentemente do que pode ser dito sobre a análise sintática, por exemplo)[16].

Após um debate extenso e uma experiência de anotação preliminar de textos, que levou à fixação da terminologia e à escolha de quais problemas resolver, resultando em extensas diretivas de anotação, foi criada a Coleção Dourada (recurso padrão-ouro) do Primeiro HAREM, tomando especial cuidado na formalização de várias alternativas possíveis e na definição das tarefas que seriam comparadas entre os sistemas.

Após a avaliação propriamente dita, que incluiu a criação de uma arquitetura de avaliação, os recursos coleção dourada e programas de processamento dos resultados foram tornados públicos.

Todo o processo permitiu medir o estado do reconhecimento de entidades para o português, ao mesmo tempo que se identificaram questões ainda não resolvidas e se produziram estudos e observações – a nível da arquitetura, das métricas, e da influência do gênero textual– cujo interesse extravasa o simples processamento computacional da nossa língua.

Diferentemente de outras avaliações de NER internacionais, como o MUC e o ACE, no HAREM (i) utilizamos um conjunto de classes variado, derivado de uma leitura preliminar de textos de gêneros variados (e que, portanto, continham entidades de natureza semântica variada), e (ii) oferecemos às equipes participantes a possibilidade de escolher em quais classes de entidades gostariam de competir. Ou seja, os sistemas poderiam escolher em

[14] https://www.linguateca.pt/Morfolimpiadas/

[15] Notamos que o termo "entidades nomeadas" também tem sido utilizado em português.

[16] No entanto, embora o reconhecimento de entidades mencionadas pareça uma tarefa simples, a literatura apresenta uma grande quantidade de abordagens, definições e estruturas de avaliação diferentes, ver Jiang et al. (2016).

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

quais classificações gostariam de ser avaliados, por exemplo, apenas LOCAL ou PESSOAS, ou TEMPO – o que obrigou a organização a desenvolver medidas de avaliação específicas, ver Seco et al. (2006).

Além disso, uma característica importante de ambas as edições do HAREM (e que também difere do MUC e do ACE), é a possibilidade de **uma mesma entidade seja classificada, em contexto, de mais de uma maneira**. Assim, em uma frase como

> Eu gosto muito do Brasil

"Brasil" pode estar fazendo referência a um LOCAL; ou aos moradores do local ou à equipa braileira (classe PESSOA), ou, ainda, à organização política/Estado (classe ORGANIZAÇÂO). Se mais do que uma interpretação fosse válida no contexto em que ocorre, todas deveriam estar marcadas.

Essas escolhas, por sua vez, se permitiram uma fotografia rica do que era possível em REM em português, tiveram como consequência a impossibilidade de comparação direta com os resultados de avaliações para outras línguas.

O **Segundo HAREM** (2006-2008) teve como principal diferença com relação ao primeiro a existência de três subtarefas, propostas pela comunidade (duas tarefas novas, além do HAREM "clássico"):

- tarefa de **identificação temporal** (Baptista et al., 2008; Hagège et al., 2009), organizada pelo grupo proponente, ficando a equipe da Linguateca a cargo da anotação e criação do recurso padrão-ouro e da avaliação.
- ReRelEM, tarefa de **identificação de relações semânticas entre as entidades** do HAREM (Freitas et al., 2008a, 2009b), completamente a cargo da equipe da Linguateca.

A Figura 19.2 apresenta as classes e subclasses utilizadas no Segundo HAREM, na tarefa "clássica".

19.7.2 Identificação, classificação e normalização de expressões temporais

Esta tarefa, chamada informalmente do TEMPO, ou de reconhecimento de entidades temporais, foi idealizada por um grupo participante (Baptista et al., 2008) inspirada em avaliações de referência temporal para o inglês, nomeadamente o TempEval (Verhagen et al., 2010), e a proposta de anotação TimeML (Saurí et al., 2006).

É preciso explicar que a tarefa expandia a anotação já existente no Primeiro HAREM (e que tinha tipo e subtipo), definindo mais atributos, que foram chamados atributos estendidos, a saber: sentido, tempo-ref, val-delta e val-norm.

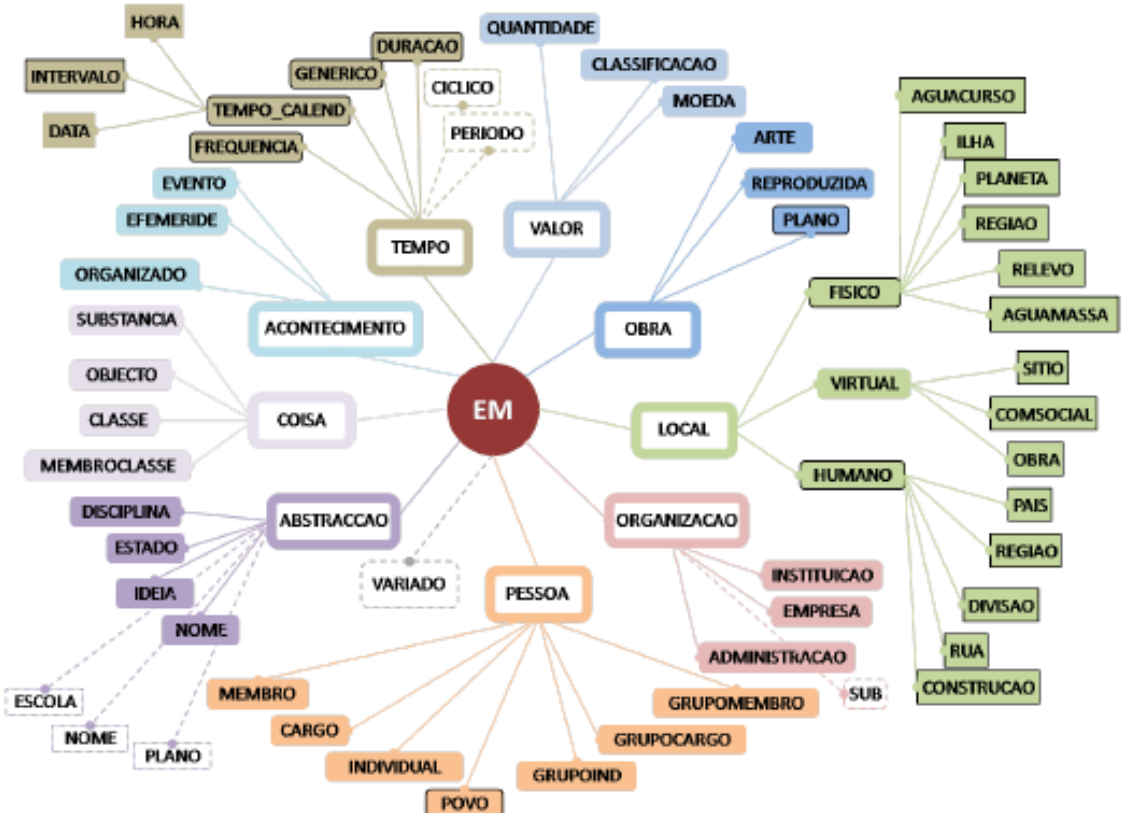
Como relatado em Mota et al. (2008), um subconjunto da coleção dourada do HAREM foi anotado com esta referência temporal mais fina, assim como programas específicos para a avaliar foram desenvolvidos e depois disponibilizados.

Esta tarefa teve sete participantes, embora nem todos tentassem resolver os mesmos problemas (apenas dois tentaram obter a referência temporal e um o sentido e a normalização).

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Figura 19.2: Categorias, tipos e subtipos utilizados no Segundo HAREM.



19.7.3 ReRelEM - relações entre entidades

O objetivo do ReReLEM era identificar um conjunto de relações entre as entidades do HAREM. A organização escolheu – após uma análise minuciosa do material textual – as quatro relações a seguir: IDENTIDADE, INCLUSÃO, LOCALIZAÇÃO e OUTRO.

Uma entidade específica pode obviamente estar relacionada a várias outras, como no Exemplo 19.1, que representa a informação de que a EM b13 está relacionada a b3 e b5 pela relação de `identidade`, e a b11 pela relação de `ocorrência`.

Exemplo 19.1:

> depois de partir em vantagem pontual no `<EM ID="b13" CATEG="ACONTECIMENTO" TIPO="ORGANIZADO" COREL="b3 b5 b11" TIPOREL="ident ident ocorre_em">` **Campeonato do Mundo**`</EM>`

Como é exatamente a mesma coisa dizer que A ocorre em B, ou que B é a localização de A, e se A é idêntico a B e B inclui C, então A também inclui C, e assim por diante, desenvolvemos um conjunto de regras para que todas as relações logicamente válidas possíveis fossem preenchidas automaticamente, inspiradas em Vilain et al. (1995) no MUC.

A coleção padrão-ouro do ReRelEM (um subconjunto da coleção padrão-ouro do HAREM), continha 12 textos e 573 entidades mencionadas. Após a expansão automática, a coleção continha 6.477 relações.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Todo o material das duas edições do HAREM, incluindo programas de avaliação e corpora anotados, estão disponíveis[17].

19.8 CLEF: Recuperação de Informação e QA (2003-2009)

Em 2003, a Linguateca começou a colaborar com o CLEF (acrônimo de *Cross-lingual Evaluation Forum*), um fórum de avaliação multilíngue. Do ponto de vista da língua portuguesa, tomar parte em uma avaliação conjunta com um grande público e muita experiência parecia uma boa ideia, especialmente em tarefas nas quais a Linguateca não tinha nenhuma experiência, como recuperação de informações (RI) e resposta automática a perguntas (RAP, QA). Além disso, o bônus adicional de conseguir participantes não lusófonos para tarefas multilíngues parecia uma maneira de colocar o português no mapa.

Por isso, a Linguateca fez parte da organização do CLEF de 2003 a 2009, e gradualmente, passou para a organização de tarefas mais inovadoras, como o GeoCLEF, o GikiP e o GikiCLEF, além de ajudar no ImageCLEF, no WebCLEF, no ResPubliQA e no LogCLEF.

Ao contrário do que se poderia esperar, contudo, sempre houve menos participantes de processamento em português no CLEF do que no HAREM.

19.8.1 CLEF clássico

A tarefa mais clássica do CLEF é uma tarefa padrão de recuperação de informações em um ambiente multilíngue. Dada uma coleção de documentos e um conjunto de tópicos, os sistemas participantes devem fornecer um conjunto classificado de documentos sobre o tópico. Seja de forma monolíngue (pesquisando tópicos expressos em português em uma coleção de documentos em português) ou multilíngue (por exemplo, pesquisando tópicos expressos em francês em uma coleção em português, ou vice-versa), ou de forma multilíngue em todas as coleções.

Adicionar o português significou, na prática, três questões, discutidas detalhadamente em Santos; Rocha (2003), Rocha; Santos (2007):

- disponibilizar uma coleção de textos – a coleção CHAVE –, com textos completos de dois jornais principais, um de Portugal e outro do Brasil, de 1994 e 1995.
- criar tópicos apropriados para a coleção, bem como para consultas multilíngues. Por exemplo, nos esforçamos por obter tópicos sobre países lusófonos que também estivessem presentes em coleções de outros idiomas e para encontrar, por exemplo, temas finlandeses que ocorressem na coleção portuguesa e também na finlandesa. (Esses tópicos também precisaram ser traduzidos para o inglês, e nós precisamos traduzir os tópicos de outros idiomas para o português.)
- avaliar os resultados na coleção em português, pois vários documentos de cada tópico tiveram que ser marcados como relevantes ou irrelevantes. Para se ter uma ideia do tamanho, em 2004 foram avaliados 22.311 documentos, com uma média de 446 documentos por tópico.

Após a realização da avaliação, tornamos públicos os tópicos e seus julgamentos binários na coleção CHAVE[18], para que os sistemas pudessem usá-los para treinamento.

[17] https://www.linguateca.pt/aval_conjunta/HAREM/
[18] https://www.linguateca.pt/CHAVE/

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

19.8.2 Resposta automática a perguntas (QA@CLEF)

A tarefa de resposta a perguntas[19] seguiu uma configuração de avaliação semelhante na mesma coleção, mas **em vez de tópicos**, foram criadas **perguntas** (classificadas pelo tipo de resposta esperada) e, **em vez de documentos relevantes**, foram avaliadas **respostas específicas**. Antes de sugerir uma pergunta, tivemos que verificar se a resposta poderia ser encontrada na coleção, de preferência em mais de um documento, para que a pergunta pudesse ser usada no QA@CLEF. Todas as respostas a uma determinada pergunta foram agrupadas e disponibilizadas junto com a coleção CHAVE.

Como resultado da organização da trilha[20] QA de 2003 a 2008 (correspondente aos anos de 2004, 2005, 2006, 2007 e 2008), foram criadas e disponibilizadas 4380 perguntas em português, juntamente com suas respostas na coleção CHAVE. (Deve-se observar que as perguntas foram coletadas coletivamente por todos os organizadores, uma vez que QA era tanto monolíngue quanto multilíngue. Essa é a razão pela qual nem todas as perguntas têm respostas na coleção CHAVE, que é em português – elas têm respostas em outras coleções, por exemplo, na italiana.)

É importante relatar que, embora essa tenha sido a configuração no início, a tarefa de QA foi aprimorada e enriquecida de uma edição para a outra, e quem se interessar por QA multilíngue no CLEF pode ler todos os documentos de visão geral, também porque havia várias subtarefas. Aqui, comentamos apenas algumas aspectos relevantes para a parte em português.

Por exemplo, a partir de 2006, além de fornecer uma resposta, era necessário fornecer uma justificativa para ela usando a coleção, para evitar que os sistemas soubessem a resposta de outras fontes, como a Internet ou suas próprias bases de conhecimento. Isso levou a uma classificação quíntupla das respostas: corretas (e justificadas), inexatas, não justificadas, erradas e ausentes (Magnini et al., 2007).

Além disso, as perguntas foram criadas em grupos sobre um tópico, aceitando, por exemplo, correferência entre elas, em vez de apenas perguntas independentes (Giampiccolo et al., 2008).

Por fim, foram adicionadas perguntas de lista, ou seja, perguntas que exigiam uma lista fechada ou aberta como resposta, bem como perguntas temporalmente restritas.

Os recursos para o português do QA@CLEF estão disponíveis na coleção CHAVE[21].

19.8.3 GeoCLEF

O principal objetivo do GeoCLEF (Gey et al., 2007) era desenvolver e avaliar a recuperação de informações geográficas. A ideia era criar tópicos que exigissem **conhecimento geográfico e, portanto, raciocínio**. Dois exemplos de tópicos são

- Cidades em um raio de 100 km de Frankfurt
- Malária nos trópicos

que ilustram, respectivamente, regiões delimitadas com precisão e regiões vagas.

Os recursos para o português do QA@CLEF estão também disponíveis na coleção CHAVE[22].

[19] No CLEF, o nome em inglês é *question answering*, QA.
[20] Uma trilha (*track*) é uma tarefa específica dentro de uma avaliação conjunta maior.
[21] https://www.linguateca.pt/CHAVE/
[22] https://www.linguateca.pt/CHAVE/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

19.8.4 GikiP: Informação geográfica (2008)

O GikiP (Santos et al., 2009) foi uma tarefa piloto do GeoCLEF em 2008. O nome da tarefa inclui G para geográfico, iki para Wikipedia e P para piloto.

Usando as coleções da Wikipédia já disponíveis para a trilha de QA (em português, alemão e inglês), a tarefa era responder a perguntas/solicitações de informações que exigissem algum tipo de raciocínio geográfico, fornecendo uma lista de páginas da Wikipédia como resposta. Isso significava uma espécie de híbrido entre RAP (QA) e IR.

Uma das motivações para usar a Wikipédia em um contexto multilíngue foi o fato de a Wikipédia ser uma mistura interessante de corpora comparáveis e de tradução, considerando os vínculos linguísticos entre diferentes idiomas.

Apenas três sistemas participaram, e todos os tópicos e respostas estão disponíveis na página do GikiP[23].

19.8.5 GikiCLEF

O GikiCLEF (Santos et al., 2010) englobou 9 línguas (e 10 wikipédias) e recebeu 17 corridas (*runs*) de 8 participantes, dos quais um do Brasil e outro de Portugal. Foram preparados 50 tópicos (em 10 "línguas"[24]), e os sistemas precisavam fornecer os IDs das páginas da Wikipédia, além de uma justificativa.

Os tópicos foram escolhidos de modo a refletir aspectos culturais, para que não houvesse uma resposta semelhante em todos os idiomas. Mas, como discutido em Santos; Cabral (2010), isso teve a consequência perversa de que, para qualquer tópico, havia mais respostas em inglês e, como os 50 tópicos cobriam culturas e temas muito diferentes, era difícil fazer generalizações estatísticas. De qualquer forma, o GikiCLEF produziu 1.009 respostas em um conjunto de 10 wikipédias, disponibilizadas no pacote GIRA[25].

Das avaliações organizadas pela Linguateca, esta foi sem dúvida a que exigiu mais esforço. Mas, no que diz respeito à língua portuguesa, ela quase não se materializou em progresso: conforme relatado em Santos; Cabral (2010, p. 219), havia apenas um tópico com tema lusófono: "Estados litorâneos brasileiros". No entanto, é possível rastrear sua influência na próxima avaliação conjunta organizada pela Linguateca, o Págico, em que voltamos a ter desafios somente em português e adicionamos os seres humanos ao circuito, ou melhor, desenvolvemos uma avaliação em que os humanos também podiam competir.

19.9 Págico: RI complexa sobre culturas de língua portuguesa (2012)

O Págico (Mota et al., 2012) foi uma avaliação conjunta na área de recuperação de informação em português cujo objetivo era avaliar sistemas capazes de encontrar respostas não triviais a necessidades de informação complexas, em língua portuguesa. A ideia inicial, no Págico, era, usando **perguntas relacionadas à cultura dos países lusófonos**, tornar a avaliação também relevante para a participação humana - por exemplo, estudantes de português como língua estrangeira ou pesquisadores da cultura portuguesa e brasileira. Hoje,

[23] https://www.linguateca.pt/GikiP/

[24] Há dois padrões de escrita da língua norueguesa, cada um com a sua wikipédia, e daí a disparidade entre 9 e 10.

[25] https://www.linguateca.pt/GikiCLEF/GIRA/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

podemos pensar no Págico como um **precursor de conjuntos de dados (*datasets*) que comparam desempenho humano e das máquinas** ao responder a perguntas.

O nome Págico é uma brincadeira com as palavras "página" e "mágico", e o Págico teve sete participantes (apenas dois com sistemas automáticos). Foi a única avaliação conjunta organizada pela Linguateca em que nenhuma equipe ou pessoa brasileira participou (embora tenha sido co-organizado por uma universidade brasileira, a PUC-Rio, e ironicamente a maioria dos tópicos do Págico diziam respeito ao Brasil). Um dos motivos para a pouca participação talvez tenha sido a dificuldade da tarefa, ou o fato de a tarefa subjacente não ser uma prioridade para a maioria das pessoas que trabalhavam com PLN em português, ou, ainda, a impossibilidade de disponibilizar um conjunto de treino para que as equipes treinassem seus modelos, em um momento em que o aprendizado de máquina já tinha muitos adeptos no PLN.

Especificamente, o objetivo do Págico era responder a 150 perguntas sobre cultura lusófona com base na Wikipédia em português, que tivessem respostas não triviais, no sentido de que não seriam cobertas por uma única página ou hub. Em outras palavras, estávamos visando respostas agregadas: obter uma lista justificada, dada uma necessidade de informação específica, recompensando não só a quantidade mas também a variedade das respostas.

Os tópicos/questões tentaram abranger todos os lugares onde se fala português, e foram distribuídos entre Letras, Artes, Geografia, Cultura, Política, Esportes, Ciências e Economia. Alguns exemplos de perguntas/tópicos podem ser vistos na Quadro 19.1.

Quadro 19.1: Exemplos de perguntas no Págico

Locais mencionados nos Lusíadas
Museus em capitais de países lusófonos
Políticos da África lusófona que estudaram na União Soviética
Compositoras brasileiras de samba
Filmes sobre o Cangaço
Além do samba, que outros gêneros musicais são populares no Carnaval Brasileiro?
Praias de Portugal boas para a prática de surf
Movimentos culturais surgidos no Nordeste do Brasil
Matemáticos de língua portuguesa que estudaram ou trabalharam em Itália

A criação dos tópicos foi concomitante à coleta de possíveis respostas, a fim de permitir, posteriormente, o cálculo de uma "pseudo-abrangência" (e correspondente pseudo-medida F)[26].

No Págico, propusemos várias medidas diferentes para avaliar a participação: além da precisão e da (pseudo)-abrangência habituais, usando todas as respostas corretas agrupadas e as respostas já coletadas pelos organizadores, acrescentamos uma precisão relaxada (sem avaliar a justificativa) e mais duas medidas novas: originalidade e criatividade.

A originalidade, que é medida por corrida e por participante (caso um participante tenha enviado mais de uma participação), recompensa as respostas corretas que foram dadas somente naquela corrida/por aquele participante.

A criatividade, por outro lado, é uma medida que pontua (uma corrida ou um participante) de forma inversamente proporcional ao número de corridas diferentes (ou participantes) que forneceram a mesma resposta.

[26]Pseudo, porque não sabemos todas as respostas que se encontram na Wikipédia. Só conhecemos as que são certas e forem encontradas pelo conjunto de sistemas.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

De forma não surpreendente, os participantes humanos foram mais criativos e originais, mas é também interessante constatar que os sistemas conseguiram encontrar algumas respostas que não foram encontradas pelos participantes humanos.

Todos os dados (incluindo a coleção da Wikipédia de 25 de abril de 2011 em XML, composta por 681.058 documentos), o conjunto de respostas e os resultados foram disponibilizados no recurso que chamamos Cartola[27], e publicamos um volume especial da revista Linguamática com artigos descrevendo a avaliação e os resultados do Págico (Santos et al., 2012).

19.10 ASSIN e ASSIN 2: similaridade semântica e inferência (2016 e 2019)

A primeira ASSIN, Avaliação de Similaridade Semântica e de INferência textual, foi organizada pelo NILC e a sua reunião final decorreu como um satélite do PROPOR 2016 (Fonseca et al., 2016d).

Como o seu nome indica, tinha dois objetivos, nomeadamente avaliar a inferência textual, e a similaridade semântica. A organização propôs, portanto, duas tarefas separadas, ambas sobre pares de frases:

- dadas duas frases, avaliar qual a semelhança de sentido entre elas, numa escala de 1 a 5;
- dadas duas frases, avaliar se uma implica a outra, se é uma paráfrase da outra, ou nenhum destes casos.

Ressalte-se que, embora inspirada por iniciativas relacionadas para o inglês, a ASSIN desenvolveu um método original para obter os pares de frases iniciais (que depois foram selecionados e eventualmente ligeiramente modificados pela organização), com base em notícias do *Google news* em português do Brasil e em português de Portugal (a ASSIN separou as duas variantes em dois recursos distintos). Ou seja, ao contrário do feito anteriormente, basearam-se em pares de frases reais e não em frases criadas por peritos.

Seis sistemas diferentes participaram na tarefa de similaridade, e quatro na de inferência textual.

A ASSIN propôs também sistemas de resultado mínimo garantido (*baselines*) que, inesperadamente, se mostraram bastante bons. Os recursos criados no âmbito da ASSIN, assim como os programas de avaliação, encontram-se acessíveis na página da ASSIN[28].

Alguns anos depois, um conjunto mais alargado de pesquisadores (incluindo o primeiro autor da ASSIN 1) organizou a ASSIN 2 (Real et al., 2019, 2020), que se caracterizou por tornar a tarefa mais simples, de duas maneiras:

- Os pares de frases foram gerados por seres humanos em vez de serem encontrados através da teécnica inovadora da ASSIN 1: mais especificamente, foram gerados através da tradução de um recurso inglês para o português brasileiro seguida de adaptação manual, resultando em frases mais simples, sem referências temporais complexas e discurso indireto, típicas do discurso jornalístico[29].

[27] https://www.linguateca.pt/Cartola
[28] http://nilc.icmc.usp.br/assin/
[29] Seria interessante avaliar se a causa da maior simplicidade vinha sobretudo da diferença entre as línguas, ou da diferença entre os gêneros.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

- Em vez de três alternativas na tarefa de implicação textual, passaram a estar em jogo apenas duas: se havia uma implicação, ou se não havia qualquer relação, mais uma vez no seguimento da tradição internacional.

Segundo a organização, estas duas modificações levaram os sistemas participantes (nove) a aumentar significativamente os números de desempenho, e permitiram que fosse mais simples comparar os resultados com o estado da arte para outras línguas.

Mais uma vez, os recursos – 9.448 pares de frases, divididos em treino, validação e teste (6500/500/2448) – foram tornados públicos após a conferência final, e estão na página da ASSIN2[30]. Os programas de avaliação mantiveram-se inalterados.

Este é um exemplo que vai na contramão da tradição/definição de usar avaliações conjuntas para avançar o estado de arte numa área, visto que a segunda edição da avaliação foi sobre uma tarefa mais fácil, e também mais reduzida. Para a organização, o esforço de simplificar a tarefa foi benéfico, não apenas porque os sistemas puderam ter um desempenho melhor nesta edição, mas também porque o corpus da ASSIN 2 foi feito com uma estratégia de anotação comparável a de outras avaliações conjuntas para o inglês (Real et al., 2019, 2020).

19.11 IberLEF 2019: reconhecimento de entidades e de relações

No âmbito do IberLEF (*Iberian Languages Evaluation Forum*) em 2019 foram propostas três tarefas de avaliação para o português por um grupo de pesquisadores da PUCRS, da Universidade de Évora e da Universidade da Bahia (Collovini et al., 2019), nas duas áreas de reconhecimento de entidades mencionadas e extração de relações.

Em relação às entidades mencionadas, um dos objetivos era aumentar o tipo de dados sobre os quais esta tarefa se aplicava. Documentos clínicos e da polícia foram duas das inovações trazidas. Também foram distribuídos vários recursos já anotados (incluindo algumas das coleções do HAREM) para treinamento dos sistemas (que precisassem).

Enquanto a tarefa do ReRelEM foi alargada para cobrir outros tipos de entidades mencionadas na segunda tarefa, a terceira tarefa, a tarefa de avaliação de extração de quaisquer relações entre entidades (não necessariamente nomes próprios) – chamada em inglês *open information extraction* – foi pioneira para o português.

Mais especificamente, as três tarefas propostas foram:

- reconhecimento de entidades mencionadas (seis categorias: pessoas, organizações, locais, valores, tempos e miscelânea) em vários gêneros textuais, e reconhecimento de pessoas em texto clínico e em documentos da polícia. A avaliação foi feita usando os programas do CoNLL. Uma característica especial desta avaliação foi que os sistemas teriam de ser replicados pela organização (na fase de reprodução) para poderem participar.

- extração de relações entre pessoas, organizações e lugares (ORG-ORG, ORG-PLC, ORG-PER), ou (opção 1) já marcados como tal no texto (reanotando a coleção do ReRelEM), ou (opção 2) sem ter as entidades mencionadas marcadas.

- extração de relações em texto livre, sem que sejam necessariamente apenas entre entidades mencionadas.

[30] https://sites.google.com/view/assin2/

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Houve cinco participantes na primeira tarefa (que enviaram seis corridas), um na segunda e três na terceira (enviando também seis corridas). Ao todo, participaram seis instituições nas três tarefas.

As coleções douradas (ou padrão-ouro) – excetuando os textos clínicos e da polícia, para proteção de dados pessoais –, assim como os programas de avaliação, encontram-se públicos na página do IberLEF 2019[31].

19.12 IDPT: Identificação de ironia (2021)

No IberLEF 2021 houve uma nova tarefa para o português, denominada *Irony Detection in Portuguese* (IDPT), em que os organizadores (Corrêa et al., 2021) criaram dois recursos com tweets e notícias em português, anotados com essa informação (ironia, ou não), baseados em jornais e tweets brasileiros.

Esta tarefa recebeu interesse internacional além do lusófono, tendo seis grupos participantes, um dos quais chinês e outro espanhol. Houve diferenças entre a detecção de ironia em notícias e tweets, com diferentes tecnologias obtendo o melhor desempenho em cada tipo de texto.

Para treinamento foram usados corpora já públicos[32], contendo 15212 tweets, e os conjuntos de textos (*datasets*)[33], provindos de três jornais eletrônicos brasileiros: O Estadão, O sensacionalista e o The Piauí Herald. Os dados de teste encontram-se na página do IDPT2021[34].

19.13 ABSAPT: Avaliação de mineração de opiniões por aspecto (2022)

No IberLEF 2022 aconteceu a tarefa *Aspect-Based Sentiment Analysis in Portuguese* (ABSAPT) (Silva et al., 2022b), que teve como objetivo identificar a opinião (positiva, neutra ou negativa) sobre aspectos (ou características) de produtos em resenhas para o *TripAdvisor* em português, portanto relativos a hotelaria e turismo. Essa tarefa, inspirada por tarefas congêneres no SemEval[35], como a descrita em 2014 por Pontiki et al. (2014), estava dividida em duas subtarefas: identificar primeiro o aspecto, e depois qual a opinião (em inglês, *sentiment*) sobre esse aspecto.

Silva et al. (2022b) listam 40 aspectos diferentes, dos 77 presentes no material de treinamento.

Houve cinco participantes, todos eles brasileiros. O conjunto de dados de treino e teste está disponível para uso acadêmico na página dedicada ao ABSAPT2022[36].

[31] https://github.com/jneto04/iberlef-2019
[32] https://github.com/fabio-ricardo/deteccao-ironia/
[33] https://github.com/schuberty/PLNCrawler
[34] https://sites.google.com/inf.ufpel.edu.br/idpt2021/data
[35] https://semeval.github.io/
[36] https://sites.google.com/inf.ufpel.edu.br/absapt2022/home

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

19.14 DIP: Desafio de Identificação de Personagens (2022-2023)

No âmbito do recente interesse pelas Humanidades Digitais, e mais especificamente pela leitura distante[37], a Linguateca, junto com o NuPILL (Universidade Federal de Santa Catarina e Universidade Estadual do Maranhão) organizou uma avaliação conjunta denominada DIP (Santos et al., 2023), cujo objetivo era, para um conjunto de obras literárias em português, identificar as personagens, as relações familiares entre elas, e algumas suas características (como os vários nomes por que eram tratadas; a sua profissão; e o seu sexo).

Era, portanto, uma tarefa agregadora de várias tarefas de PLN (reconhecimento de entidades mencionadas, reconhecimento de profissões, reconhecimento de relações familiares etc.), mas que se referia a uma obra na sua totalidade e não a partes do texto. Para isso, disponibilizamos 100 obras completas em formato de texto, e 100 obras em formato PDF. Para 20 de cada conjunto tínhamos laboriosamente criado o padrão-ouro.

Embora tivéssemos vários interessados inicialmente, apenas um sistema participou.

De qualquer maneira, criamos um conjunto de recursos interessantes sobre leitura distante em português, assim como propusemos métodos de avaliação inovadores (Willrich; Santos, 2023).

É possível que esses recursos mais tarde possam ser usados por equipes que desenvolvam novos sistemas baseados em treinamento, como foi a estratégia no caso do IberLEF 2019 em relação ao HAREM e ao ReReLEM.

Os recursos criados, assim como os programas de avaliação, encontram-se disponíveis na página do DIP[38].

19.15 Resumo das avaliações conjuntas

Na Tabela 19.1 listamos as avaliações descritas neste capítulo e os recursos existentes.

Tabela 19.1: Panorâmica das avaliações conjuntas para o português: *num* indica o número de participantes. No caso do CLEF, QA@CLEF e GeoCLEF, o número de participantes é a soma dos participantes de todos os anos.

aval conj	ano	num	recursos
Morfolimpíadas	2003-4	7	https://www.linguateca.pt/Morfolimpiadas/
Primeiro HAREM	2004-6	10	https://www.linguateca.pt/primeiroHAREM/
Segundo HAREM	2007-8	10	https://www.linguateca.pt/HAREM/
CLEF RI	2003-7	34	https://www.linguateca.pt/CHAVE
QA@CLEF	2003-9	23	https://www.linguateca.pt/CHAVE
GeoCLEF	2005-8	10	https://www.linguateca.pt/CHAVE
GikiP	2008	3	https://www.linguateca.pt/GikiP/
GikiCLEF	2009	8	https://www.linguateca.pt/GikiCLEF/
Págico	2012	7	https://www.linguateca.pt/Pagico/
ASSIN 1	2016	6	http://nilc.icmc.usp.br/assin/
ASSIN 2	2019	8	https://sites.google.com/view/assin2/
IberlEF	2019	6	https://github.com/jneto04/iberlef-2019
IDPT	2021	6	https://sites.google.com/inf.ufpel.edu.br/idpt2021/data

[37] Uma abordagem computacional dos estudos literários, iniciada por Moretti (2013) e chamada *distant reading* em inglês.

[38] https:/www.linguateca.pt/DIP/

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

ABSAPT	2022	5	https://sites.google.com/inf.ufpel.edu.br/absapt2022/
DIP	2022-3	1	https:/www.linguateca.pt/DIP/

19.16 Considerações finais e lições aprendidas

Neste capítulo tentamos fazer uma panorâmica das avaliações conjuntas (e das "tarefas compartilhadas") que têm havido para o português.

É importante mencionar que existem artigos muito completos que documentam a história de avaliações conjuntas em outras línguas. Citamos apenas alguns aqui: Hirschman (1998) para o MUC, Harman (1998) para o TREC, Braschler; Peters (2004) para o CLEF, Kando (2002) para o NCTIR e Edmonds; Kilgarriff (2002) para o Senseval. Além disso, as avaliações mais atuais têm páginas eletrônicas que são referência fundamental, como as páginas do SemEval, de avaliação semântica e as avaliações associadas às conferências ConLL[39].

Por fim, gostaríamos de comentar algumas lições aprendidas (visto que participamos na organização da maioria das avaliações), assim como aprimorar a diferença entre avaliação conjunta e tarefa compartilhada, com base na panorâmica anterior. Deixamos aos leitores a tarefa de avaliar as avaliações existentes através dos parâmetros que indicamos neste capítulo, e de outros que pareçam relevantes.

Refletindo sobre o papel que a avaliação conjunta tinha para a Linguateca (e que também é consequência de seus objetivos, como indicamos no início do capítulo), e a prática do IberLEF, ou, o que chamamos provisoriamente aqui de "tarefas compartilhadas", parece haver duas diferenças sistemáticas:

1. Tentativa de definição de uma tarefa por um conjunto de pessoas antes da construção dos recursos de avaliação, vs. o compartilhamento (ou partilha) de uma tarefa já conhecida/desempenhada pelos proponentes da tarefa, geralmente com recursos já existentes;

2. Separação estrita entre os organizadores da avaliação e os participantes (no caso de uma avaliação conjunta), e alguma sobreposição (no caso de uma tarefa compartilhada), em que muitas vezes são os próprios proponentes os principais interessados em participar.

Em consequência destas duas diferenças, existe mais interesse em documentar o processo e os recursos quando existe uma instituição ou grupo que apenas organiza, o que é revelado pela quantidade de documentação produzida no âmbito das avaliações conjuntas organizadas pela Linguateca, nomeadamente três livros e dois volumes dedicados de uma revista. Além disso, algo que também percebemos é a inexistência de documentação em português nas tarefas compartilhadas (até agora).

Seja como for, em vez de sugerir uma dicotomia, gostaríamos apenas de chamar a atenção para as opções envolvidas. Estamos na presença de um **contínuo que tem como objetivo avaliar vários sistemas** para a língua portuguesa e, mais do que isso, **avançar o estado da arte e da técnica** nessa área.

[39] https://www.conll.org/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

Agradecimentos

Agradecemos sinceramente a todos os organizadores e participantes nas avaliações conjuntas que descrevemos, que contribuíram significativamente para o avanço do PLN em português.

Agradecemos a Hugo Gonçalo Oliveira a ajuda no esclarecimento de algumas questões relacionadas com a ASSIN 2 e a IDPT.

Finalmente, agradecemos a Graça Nunes e a Helena Caseli o convite para escrever este capítulo, e a Brielen Madureira e a Renata Vieira a sua revisão.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Parte X

Desafios e Perspectivas

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Capítulo 20

Questões éticas em IA e PLN

Maria das Graças Volpe Nunes
Tayane Arantes Soares
Mariza Ferro

Publicado em: 26/09/2023
Atualizado em: 16/04/2026

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

Toda civilização, coletividade ou sociedade surge a partir do compartilhamento de necessidades, bens e valores comuns. A sobrevivência e a prosperidade de uma sociedade dependem de algum tipo de mediação das diferenças e da regulação do comportamento de seus integrantes. A mediação das diferenças e a regulação da conduta de indivíduos partem de pressupostos. A ética refere-se ao comportamento de indivíduos na tomada de decisões e na sua responsabilização por elas, frente aos valores compartilhados pela sociedade em que vivem.

A partir do momento em que os sistemas de inteligência artificial (IA) passam a fazer parte da sociedade, interagindo com humanos e mimetizando seus comportamentos, tomando decisões com certo grau de autonomia e eventualmente colocando pessoas ou sociedades em risco, problemas de natureza ética naturalmente emergem. Nesse contexto, tem havido uma preocupação crescente com as implicações éticas dos atuais sistemas inteligentes, e a sociedade acadêmica de IA tem se movimentado para alertar e promover mudanças para minimizá-los ("AI and Ethics", 2023; Coeckelbergh, 2020).

20.1 Ética em IA

Atualmente (2026), a principal tecnologia de IA para dotar seus programas com inteligência caracteriza-se por fornecer, a algoritmos criados para aprender, grandes quantidades de dados sobre aquilo que deve ser aprendido, ou seja, sobre um conceito ou uma tarefa. E, conforme já discutido no Capítulo 16, se esses dados não forem coletados de maneira criteriosa, podem conter vieses que acabam por provocar comportamentos indesejáveis, incorretos ou inadequados. Isso ocorre muitas vezes por terem sido treinados com dados desbalanceados e sem curadoria, ou por terem aprendido correlações entre os dados que ou são irrelevantes para o conceito que se quer ensinar, ou carregam algum viés indesejado.

Um exemplo concreto de vieses algorítmicos é relatado no documentário "*Coded Bias*" (Kantayya, 2020). A cientista da computação Joy Buolamwini, uma mulher negra, durante a sua pesquisa sobre softwares de visão computacional no MIT (*Massachusetts Institute of Technology*), não conseguia ter seu rosto reconhecido pelo software no qual estava trabalhando. Somente após colocar uma máscara facial branca que o sistema reconheceu a máscara como sendo um rosto. As pesquisadoras Joy Buolamwini e Timnit Gebru constataram que 79,6% dos dados de treinamento desse software eram compostos por pessoas

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

de pele clara (Buolamwini; Gebru, 2018).

Esse fato evidencia o problema estrutural que permeia a criação de ferramentas de IA, tendo em vista a pouca ou nenhuma reflexão por parte de empresas e de pessoas que desenvolvem essas soluções sobre os impactos sociais que essas ferramentas podem causar, além de pouco envolvimento da sociedade no desenvolvimento dessas soluções tecnológicas (Hora, 2021). O'Neil (2021) relata vários outros exemplos de consequências negativas de se utilizar algoritmos de aprendizado de máquina (AM) em tomada de decisões.

Casos de discriminação de raça são frequentes. A ferramenta *Google Fotos* foi acusada de rotular a imagem de um casal negro como "gorilas", e a do *Flickr* rotulou fotos de pessoas negras como "macaco" (Cruz, 2021). A pesquisa de Buolamwini; Gebru (2018) também revelou vieses de raça e de gênero em serviços de IA de empresas como Microsoft, IBM e Amazon. O então Twitter, em 2020, foi denunciado por priorizar rostos de pessoas brancas na exibição de imagens publicadas pelos usuários (INFOBASE, 2021).

A IA também já foi acusada de impulsionar o ódio às minorias e influenciar os resultados de eleições (Cavaliere; Romeo, 2022), explorar fraquezas psicológicas e orientar decisões (Sartori; Theodorou, 2022), causando problemas como a intensa polarização social e ameaças aos princípios democráticos e aos direitos humanos (Artificial intelligence and human rights., 2021; Empoli, 2019).

Outra característica importante desses algoritmos que aprendem (ao menos na abordagem mais utilizada no ano de 2026, que são as redes neurais artificiais) é que aquilo que aprendem não é recuperável de forma compreensível a nós, ou seja, é impossível recuperar exatamente qual conhecimento foi apreendido pela máquina. Aferimos seu conhecimento apenas pelo seu comportamento numa determinada tarefa. Nesse sentido, dizemos que são obscuros, verdadeiras caixas-pretas, ou seja, não explicáveis. Essa característica indesejável deriva da tecnologia chamada *Deep Learning* (Goodfellow et al., 2016), que nada mais é do que um modelo especial de redes neurais. Tais modelos desempenham muito bem, é verdade, mas não conseguimos entender como o fazem. Desse modo, a impossibilidade de se justificar suas decisões, aliada à presença cada vez mais constante desses sistemas no nosso cotidiano, tem gerado um sentimento de insegurança e desconfiança.

O ChatGPT, da OpenAI (OpenAI, 2022), de enorme repercussão no final de 2022, rapidamente teve sua reputação abalada devido à incapacidade de referenciar com exatidão a fonte de suas respostas (até porque, como foi mencionado anteriormente, é extremamente complicado recuperar com precisão o conhecimento apreendido pelo modelo a partir dos dados de treinamento (Heikkilä, 2021)). Consequentemente, torna-se desafiador determinar a fonte exata das respostas geradas, uma vez que o modelo atua essencialmente como um gerador de palavras prováveis com base em uma entrada inicial. Para algumas situações em que essas respostas determinariam decisões ou teriam consequências importantes, a falta de confiança no sistema pode ter afastado alguns usuários. A despeito disso, no entanto, este e outros *chatbots* do mesmo tipo têm sido rapidamente adotados por usuários em suas tarefas de trabalho, seja como uma ferramenta poderosa de criação e edição de textos, seja como uma perigosa fonte de informações muitas vezes incorretas. Para saber mais sobre a tecnologia do ChatGPT, sugerimos a leitura de Capítulo ChatGPT, MariTalk e outros agentes de conversação.

A IA remete a problemas éticos na medida em que são construídos artefatos (sistemas, robôs) que interagem com humanos (de forma direta ou indireta, visível ou ubíqua) e que o resultado dessas interações podem afetar, com algum risco, esses usuários ou a sociedade como um todo. Considerando-se que sistemas inteligentes tendem a ultrapassar barreiras físicas, sociais, temporais e culturais (tendo em vista que são usados em vários lugares

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

do mundo), devemos nos lembrar que seus desenvolvedores sempre estarão inseridos num contexto cultural e moral específicos, o que pode entrar em conflito com os valores éticos de sociedades usuárias distintas.

A fim de evitar problemas dessa natureza, em uma sociedade cada vez mais interativa com máquinas de IA, é fundamental investigar maneiras de construir esses artefatos de maneira responsável (Russel, 2019). Caso contrário, essas novas tecnologias continuarão a perpetuar pontos de vista hegemônicos, reforçando e codificando preconceitos e vieses humanos que ainda lutamos para combater (Bender et al., 2021). Nina da Hora, cientista da computação brasileira e pesquisadora na área de Pensamento Computacional, ressalta que, muitas vezes, a busca global pela ética em IA, por ser baseada na tentativa de manter as tecnologias que estão causando problemas, não aprofunda o entendimento e a investigação dos problemas enfrentados pelas pessoas afetadas por essas tecnologias. Segundo a pesquisadora, é necessário ir além dos aspectos técnicos ao buscar um desenvolvimento ético de sistemas de IA, e investigar também o impacto dessas novas tecnologias na vida das pessoas envolvidas (Hora, 2022).

Existem diversas recomendações e tentativas de regulação dos sistemas de IA ao redor do mundo, sendo que os modelos regulatórios de EUA, China e União Europeia destacam-se como as três principais propostas, apesar de haver outros modelos competitivos, como o do Canadá, por exemplo. Nos EUA, com base nas análises da White & Case[1] sobre o cenário regulatório em 2026, o país consolidou uma abordagem que prioriza a liderança tecnológica e a desregulamentação federal, buscando neutralizar o que o governo atual classifica como "excessos regulatórios estaduais". Diferente do modelo europeu, os EUA não adotaram uma lei nacional abrangente, mas sim uma série de Ordens Executivas que visam remover barreiras à inovação, proteger a segurança nacional e garantir a dominância global do país no setor. A China mantém um modelo de regulação mais impositivo com maior controle estatal, concentrado em um subconjunto de aplicações de IA, mas que trabalha para atualizar a sua regulação para uma nova lei mais abrangente. Entre os dois extremos, a União Europeia, pioneira na discussão sobre IA (Commission, 2021) aprovou em maio de 2024 a a Lei da Inteligência Artificial (*Artificial Intelligence Act* - Regulamento UE 2024/1689), que estabelece regras harmonizadas em matéria de inteligência artificial, sendo o primeiro quadro jurídico abrangente sobre IA a nível mundial. O AI Act é uma lei baseada no risco, ou seja, quanto maior o risco de danos à sociedade, mais rigorosa será a lei. As proibições para aplicações de risco inaceitável, tais como sistemas para pontuação social, policiamento preditivo e reconhecimento de emoções em ambientes de trabalho e instituições de ensino, entraram em vigor em fevereiro de 2025. Já as regras para IA de alto risco tem previsão para entrarem em vigor em agosto de 2026 e agosto de 2027[2].

No Brasil, o projeto de lei (PL 2338/2023) estabelece o marco regulatório da IA no Brasil e o texto original teve sua aprovação no Plenário do Senado Federal em dezembro de 2024 e, no início de 2025, foi encaminhado à Câmara dos Deputados. O projeto estabelece "normas gerais de caráter nacional para o desenvolvimento, implementação e uso responsável de sistemas de IA no Brasil, com o objetivo de proteger os direitos fundamentais e garantir a implementação de sistemas seguros e confiáveis, em benefício da pessoa humana, do regime democrático e do desenvolvimento científico e tecnológico"[3]. A proposta cria um arcabouço regulatório baseado em risco, inspirado no modelo europeu, mas com um perfil misto de

[1] https://www.whitecase.com/insight-our-thinking/ai-watch-global-regulatory-tracker-united-states
[2] https://artificialintelligenceact.eu/ai-act-explorer/
[3] https://legis.senado.leg.br/sdleg-getter/documento?dm=9347622&ts=1720062232582&disposition=inline

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

regulação setorial, a chamada autorregulação regulada, onde as próprias empresas do setor definem normas de conduta, mas sob a supervisão e os limites estabelecidos pelo Estado. Pode-se dizer que todas as questões aqui levantadas são, na letra da lei, contempladas por esse projeto. No entanto, considerando-se o domínio absoluto de poucas e grandes *Big Techs* americanas e a complexidade e subjetividade envolvidas, o grande desafio será fiscalizar o cumprimento da lei. Em termos de regulações setoriais no Brasil, o Conselho Federal de Medicina (CFM) foi pioneiro e publicou em fevereiro de 2026, a Resolução CFM nº 2.454[4], que estabelece o novo marco regulatório para o uso de IA na medicina. A norma entrará em vigor em agosto de 2026 e traz obrigações relevantes para médicos e instituições de saúde.

Até 2023 existiam pelo menos 200 documentos com princípios destinados a fornecer orientações normativas em relação aos sistemas baseados em IA em vários países (Corrêa et al., 2023), nos quais destacam-se os princípios promovidos pela OCDE (Organização para a Cooperação e Desenvolvimento Econômico) para classificação e avaliação de sistemas de IA, que fomenta a universalização de critérios para políticas de IA (OECD, 2022). Vale ainda destacar o documento da UNESCO, aprovado em novembro de 2021, reconhecendo os impactos positivos e negativos da IA nas sociedades e recomendando que os Estados-membros tomem providência quanto à violação de direitos (UNESCO, 2022). O objetivo é sempre recomendar princípios para que os sistemas de IA sejam confiáveis, desenvolvidos e utilizados para o bem da humanidade e do planeta e para preservar os valores por meio da proteção, promoção e respeito aos direitos humanos fundamentais, à liberdade e à igualdade.

A utilização de sistemas de IA pode afetar negativamente vários direitos fundamentais estabelecidos pela Declaração Universal dos Direitos Humanos, adotada e proclamada pela Assembleia Geral das Nações Unidas (UNICEF, 1948) ou por instrumentos particulares de cada país, como a Constituição Brasileira ou a Carta dos Direitos Fundamentais da União Europeia. Os direitos fundamentais são inerentes a todos os seres humanos, independentemente da sua raça, sexo, nacionalidade, etnia, idioma, religião ou qualquer outra condição. Os direitos humanos incluem o direito à vida e à liberdade, liberdade de opinião e expressão, o direito ao trabalho e à educação, entre outros.

Entre os princípios mais comuns que norteiam as regulações e as recomendações para o desenvolvimento e o uso da IA ética e confiável estão a (1) justiça, diversidade e não discriminação, (2) transparência e explicabilidade, (3) robustez técnica e segurança, (4) privacidade e proteção de dados, (5) responsabilidade e prestação de contas[5].

1. Os princípios da justiça, diversidade e não discriminação estão intimamente ligados à promoção da justiça social e a salvaguardar a equidade e a não discriminação de qualquer tipo (gênero, raça, cor, nacionalidade, religião, língua, idade, opinião política etc.), em conformidade com o direito internacional. O objetivo é garantir a distribuição igual e justa de benefícios e custos e garantir que indivíduos e grupos estejam livres de preconceitos, injustiças, discriminação e estigmatização. Ainda, minimizar e evitar reforçar ou perpetuar resultados discriminatórios ou tendenciosos (enviesados), ao longo do ciclo de vida dos sistemas de IA (Smith; Rustagi, 2020). Se preconceitos e injustiças não puderem ser evitados, os sistemas de IA podem aumentar a desigualdade social. Além disso, o uso de sistemas de IA nunca deve

[4] https://www.in.gov.br/web/dou/-/resolucao-cfm-n-2.454-de-11-de-fevereiro-de-2026-689247948

[5] Baseando-se em pontos comuns encontrados nos documentos supracitados da OCDE, UNESCO e regulamentação da UE.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

induzir as pessoas a serem enganadas ou prejudicar sua liberdade de escolha.

Dependendo da forma como é criada e utilizada, a IA tem potencial para criar e/ou reforçar vieses humanos. O viés pode entrar no desenvolvimento e uso de um sistema de IA, especialmente por meio do uso dos algoritmos de aprendizado de máquina durante a geração, a coleta, a rotulagem e o gerenciamento dos dados com os quais o algoritmo aprende; mas também pode ser introduzido durante o design e a avaliação dos algoritmos (Smith; Rustagi, 2020). Já existem muitos exemplos do uso de sistemas que utilizam AM, os quais, com base nos dados que recebem, têm apresentado resultados tendenciosos, imprecisos e injustos, os quais representam riscos imensos para indivíduos e empresas.

São diversas as situações de discriminação de raça e de vieses e discriminação aos grupos minoritários ou culturas, principalmente com o uso de algoritmos de reconhecimento facial ou manipulação de imagens. Por exemplo, a dificuldade em reconhecer rostos de pessoas negras, como os exemplos mencionados no início deste capítulo, com a classificação automática de fotos de pessoas negras como "gorilas"; aplicativo que "desnudava" mulheres mostrando como as *deepfakes* prejudicam os mais vulneráveis (REVIEW, 2022); aplicativos que transformam fotos em caricaturas onde os avatares das mulheres, especialmente orientais, são "pornificadas", enquanto o dos homens são astronautas, exploradores e inventores (Heikkilä, 2022). Além dos casos das mídias sociais, o Brasil vem sofrendo uma profusão de denúncias com o uso de sistemas de reconhecimento facial que levaram a abordagens policiais e até prisões. A Rede de Observatórios da Segurança monitorou, entre março e outubro de 2019, os casos de prisões e abordagens com o uso de reconhecimento facial em cinco estados brasileiros e revelou que 90,5% dos presos por monitoramento facial no Brasil são negros (Nunes, 2019).

Há duas formas para minimizar esses vieses: fazer com que os dados de treinamento reflitam fielmente o universo de situações a que o sistema final será exposto, e detectar, ainda em período de testes, os desvios prováveis e eliminá-los antes de colocar o sistema em uso. Considerando a natureza da tecnologia mais comum hoje em IA – o treinamento de redes neurais – as duas formas são de difícil execução. Quer seja porque nem sempre todos os dados são acessíveis, quer seja porque os testes são incapazes de prever todas as possibilidades, dada a complexidade da tarefa a ser realizada pelo sistema.

2. Os princípios da transparência e explicabilidade são fundamentais para desenvolver e manter a confiança dos usuários nos sistemas de IA. Isso significa que os processos precisam ser transparentes, ou seja, o objetivo dos sistemas de IA deve ser comunicado abertamente e o usuário deve saber que está em contato com um produto ou serviço fornecido diretamente ou com o auxílio de sistemas de IA. Além disso, as decisões tomadas pelo sistema – na medida do possível – devem ser explicáveis aos afetados direta ou indiretamente. Nem sempre é possível explicar por que um modelo gerou uma determinada saída ou decisão (e qual combinação de fatores de entrada contribuiu para isso). Esses casos são chamados de algoritmos "caixa-preta" como os já mencionados algoritmos de redes neurais e suas redes neurais profundas. A comunidade de IA já se mobiliza em direção a tornar seus sistemas mais compreensíveis para seus usuários. A IA explicável (*Explainable AI*, *XAI*) é uma recente área de pesquisa que tem como objetivo propor processos e métodos para tornar

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

os recentes sistemas de aprendizado de máquina mais compreensíveis e confiáveis. Não é tarefa fácil, mas, juntamente com esforços para combinar aprendizado de máquina com métodos simbólicos de representação de conhecimento, é esperado que testemunharemos avanços nessa área.

3. Um componente crucial para alcançar uma IA confiável é a robustez técnica, que está intimamente ligada ao princípio da prevenção de danos. A robustez exige que os sistemas de IA se comportem conforme o planejado, sejam desenvolvidos com uma abordagem preventiva aos riscos, minimizando danos não intencionais e inesperados e evitando danos inaceitáveis. Além disso, vulnerabilidades a ataques (riscos de segurança) devem ser evitadas e eliminadas durante o ciclo de vida dos sistemas de IA para garantir proteção e segurança humana e ambiental.

4. A privacidade e a proteção de dados devem ser respeitadas, protegidas e promovidas ao longo do ciclo de vida dos sistemas de IA. É importante que os dados destinados aos sistemas de IA sejam coletados, utilizados, compartilhados, arquivados e apagados de modo compatível com os marcos jurídicos nacionais, regionais e internacionais relevantes. Por exemplo, na legislação brasileira, a LGPD (Lei Geral de Proteção aos Dados) estabelece diretrizes para o uso dos dados pessoais (LGPD, 2018). A lei similar europeia, GDPR, vai além e trata também do direito de que dados pessoais sejam apagados das bases de dados a qualquer momento, tratado como o direito ao esquecimento pelos modelos de IA.

   O uso de dados pessoais ocorre em diferentes contextos. Exemplos incluem o uso de dados para a identificação de supostas emoções para análise do comportamento do usuário. No caso de um cliente, por exemplo, para criar publicidade direcionada durante as compras com foco nos produtos ou em sua disposição na loja virtual, sem a transparência desejada. A ViaQuatro, empresa que tem a concessão da linha 4-amarela do metrô de São Paulo, foi processada pelo Instituto Brasileiro de Defesa do Consumidor por usar câmeras que coletavam dados referentes às "emoções" dos passageiros, e que seriam usados pela companhia, sem o consentimento dos passageiros. "O sistema inteligente conseguiria identificar se o passageiro está feliz, insatisfeito, surpreso e neutro. Além disso, detectaria o gênero e a faixa etária das pessoas. Os dados capturados seriam usados para a empresa fazer a gestão de seu conteúdo institucional e até de anúncios publicitários" (Cruz, 2018). Mais um exemplo de uso sem transparência e desrespeito à privacidade e a proteção dos dados pessoais.

5. A responsabilização e a prestação de contas complementam os princípios acima e estão intimamente ligadas ao princípio da justiça ao tentar garantir mecanismos que determinem as responsabilidades éticas e jurídicas pelas decisões e ações de alguma forma baseadas em um sistema de IA e seus resultados, antes e depois do seu desenvolvimento, implantação e uso.

   Enquanto se aguarda uma regulamentação direcionada aos sistemas de IA no Brasil, a LGPD tem sido usada para responsabilizar alguns abusos, como no caso citado, da ViaQuatro. A lei diz que a captação de dados sensíveis, como os biométricos, precisa de consentimento do usuário, e a finalidade da coleta também deve ter propósitos legítimos e comunicados aos titulares dos dados. Além disso, o direito à informação dos consumidores está consagrado como um princípio fundamental ao abrigo do Código do Consumidor.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

> Porém, em casos não cobertos por outras leis, a responsabilização por danos causados por sistemas de IA ainda carece de lei própria. O problema é que o avanço tecnológico ocorre em ritmo muito mais rápido do que aquele da política e da justiça. Enquanto se discutem as leis para regular esta IA de hoje, ela continua avançando e modificando nossa forma de interagir com ela e por meio dela, e antes que seja regulamentada por leis, ela pode ter se tornado outra.

A grande questão é se, um dia, um sistema de inteligência artificial estará programado para avaliar adequadamente as informações recebidas e as possíveis consequências que suas ações são capazes de causar ao ambiente e aos seres à sua volta. Será possível programá-lo para embasar suas decisões à luz de valores humanos a fim de exibir um comportamento ético?

Essa possibilidade esbarra em várias dificuldades, como a definição dos valores responsáveis por um comportamento ético, sua representação (seja de forma explícita ou por meio de exemplos), seu processamento por um algoritmo e sua incorporação por um sistema de inteligência artificial.

20.2 Ética em PLN

Não é coincidência que as discussões sobre a regulação de IA se intensificaram após o lançamento do ChatGPT. Os sistemas de PLN, ou seja, aqueles em que a língua tem protagonismo, estão entre aqueles que mais expõem suas semelhanças conosco, suas fragilidades na compreensão da língua humana, seus perigos quando interagem conosco em cenários sensíveis – como para fazer diagnósticos ou pretensas terapias, por exemplo –, os vieses culturais e morais de seus dados, entre outras coisas. O PLN assume, assim, um papel importante no âmbito das questões éticas relacionadas aos sistemas de IA. A fim de evitar vários problemas, é necessário tratar a língua de forma ampla, considerar suas variações, suas mudanças, seu papel na comunicação humana e na sociedade; tudo isso pode nos auxiliar a construir tecnologias melhores e mais inclusivas (Bender, 2020).

Muitas línguas até então desfavorecidas de recursos tecnológicos, como o português, e também línguas minoritárias, têm se beneficiado com sua inserção no mundo digital. No entanto, ainda temos um longo caminho a percorrer. Segundo Emily Bender, 90% das línguas do mundo e suas variedades usadas por mais de um bilhão de pessoas têm pouco ou nenhum suporte em termos de tecnologia linguística, reforçando a ideia de que essas novas tecnologias apresentam um potencial excludente (Bender, 2020).

Nesse sentido, a comunidade de PLN deve ter em mente que vivemos em um mundo linguisticamente diverso e que não é razoável aceitar o inglês, idioma dos dados de treinamento da maioria dos modelos de língua, como representativo de toda variedade linguística e cultural existente. Ao trabalharmos com PLN, não podemos esquecer que linguagem e poder estão atrelados, que a linguagem cria o nosso mundo e molda nossa realidade (Halliday; Matthiessen, 1999). Portanto, temos o desafio de romper com esse monopólio linguístico e desenvolver modelos a partir de dados coletados de forma responsável, validados, balanceados e livres de vieses.

É fato que o PLN tem se beneficiado muito com o uso de aprendizado de máquina. Muitas barreiras foram transpostas ao representar alguns fenômenos linguísticos por meio de exemplos. Para as tarefas clássicas de PLN – *taggers*, *parsers*, reconhecedores de entidades nomeadas, entre outras – os sistemas construídos por AM parecem cada vez

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

melhores à luz de avaliações padronizadas. O problema que se coloca é o uso futuro desses sistemas, suas combinações e suas aplicações fora de qualquer controle (Chandran, 2023).

Nos sistemas de IA mais recentes, a competência linguística é adquirida por meio de treinamento com *corpora* muito grandes, gerando um modelo de língua, ou seja, um sistema capaz de prever qual(is) palavra(s) deve(m) seguir a última palavra vista. São os chamados LLM (*Large Language Models*), apresentados e discutidos no Volume 2. O ChatGPT é um exemplo muito conhecido dessa tecnologia.

Se considerarmos que os *corpora* de treinamento de grandes modelos de língua tendem a ser compostos por uma quantidade massiva de dados linguísticos coletados na internet, e que o acesso à internet é desigual, os dados de treinamento têm grandes chances de não serem representativos e não levarem em conta a diversidade cultural e linguística existentes – desconsiderando aqui o uso não autorizado de dados pessoais. Conforme discutido nos Capítulos Responsabilidade no desenvolvimento e uso de tecnologias de linguagem baseadas em IA e PLN em Redes Sociais, manifestações carregadas de ofensas, preconceitos, discriminação, posturas antiéticas em geral são eventualmente reproduzidas nos textos gerados por esses modelos (Perrigo, 2023), e acabam se perpetuando no modelo gerado.

Além disso, a popularização de modelos de língua, como o ChatGPT, tem suscitado questões importantes no âmbito da ética em PLN, especialmente no que diz respeito à propriedade intelectual e aos direitos autorais. Embora esses dados estejam disponíveis publicamente, uma preocupação fundamental está relacionada à forma como os dados são coletados para o treinamento desses modelos, considerando se as pessoas que produziram esses dados tinham ciência de que suas postagens textuais poderiam ser utilizadas como insumos para modelos de linguagem (Alisson, 2023). Ilustra esse ponto a localização labiríntica, nos produtos da Meta, do formulário que permite ao usuário coibir o uso de seus dados no treinamento de algoritmos[6]. A pergunta que emerge é: como podemos garantir que princípios éticos de transparência, proteção de dados e consentimento serão respeitados nesse processo de coleta?

20.3 Modelos de língua como fonte de conhecimento?

Modelos de língua nos surpreendem ao possibilitarem a geração de textos coerentes, muitas vezes indistinguíveis de textos produzidos por seres humanos. No entanto, esses textos não passam de sequências de palavras prováveis estatisticamente em um dado idioma que foram "cuspidas" pelo modelo a partir de alguma entrada textual. Os modelos em si não entendem os textos gerados. Os modelos apenas apreendem, a partir do *corpus* de treinamento, os padrões (combinações frequentes) linguísticos derivados dos dados, e reproduzem, como bons papagaios estocásticos, esses padrões em novas saídas (Bender, 2023).

Nesse sentido, se a língua é apreendida a partir de um *corpus*, nos modelos de línguas, as características desse *corpus* são determinantes para a qualidade linguística do que será gerado pelo sistema (Capítulo 16). Isso soa óbvio, mas, em se tratando de língua, há uma outra consequência. Em sistemas conversacionais, como os *chatbots*, a linguagem produzida por um sistema tem um efeito: ela atenderá a alguma expectativa do usuário, que pediu uma informação, ou uma sugestão, ou se queixou de algo, ou quer simplesmente dialogar. Não basta, portanto, que a expressão linguística cumpra todos os requisitos de ortografia, gramática, coesão e coerência. É preciso atender a outros critérios. Não

[6] https://www.facebook.com/help/contact/510058597920541

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

é rara a geração de uma expressão linguística correta e elegante, com um conteúdo ou uma informação incorreta ou enviesada pelos dados. Detectar essa imprecisão, no entanto, pode não ser tão fácil. Um interlocutor do *chat*, impressionado pela boa forma do texto, pode aceitar como verdade, sem questionar, o conteúdo expresso por ela. Acontece que um modelo de língua não é capaz de preencher os requisitos relativos à autenticidade e veracidade das expressões que gera. De fato, trata-se de uma ferramenta incrível de geração de texto sendo usada para um fim para o qual não foi projetada. Como são capazes de gerar uma infinidade de expressões linguísticas, tem-se a impressão de que, de fato, têm domínio em várias áreas de conhecimento e tarefas. A consequência é que suas "alucinações" podem ser confundidas com novas "verdades", oferecendo um risco enorme à sociedade, na medida em que a crença nessas verdades pode levar a comportamentos imprevisíveis. Percebe-se aí o perigo de se utilizar um sistema impróprio como se fosse um "gerador de conhecimento".

Tão logo foi disponibilizado o ChatGPT, em 2022, as consequências desse cenário têm sido discutidas por vários setores das sociedades em todo o mundo, incluindo o Brasil. Já se prevê mudanças no trabalho em toda sorte de setores que usam informações para tomada de decisão, bem como aqueles que têm a redação de textos como atividade relevante. Incluem-se, portanto, o jornalismo, a educação formal, a pesquisa, o direito, apenas para citar alguns.

Se mal gerenciadas, essas transformações do mercado de trabalho podem ter custos econômicos e sociais significativos. Segundo a OECD (2023), 27% dos empregos podem ser impactados por automação com tecnologias de IA, e com o advento dos modelos generativos, certamente esses impactos aumentarão, para uma ampla gama de categorias de empregos. Para melhor proteção dos trabalhadores nos próximos anos, a OECD (2023) recomenda, entre outras medidas, que os governos garantam treinamento para lidar com a IA. A UNESCO (UNESCO, 2019) também enfatiza a necessidade de desenvolvimento de valores e habilidades para a vida e para o trabalho na era da IA e recomenda medidas para melhorar a literacia em IA em todas as camadas da sociedade. Capacitar a população com habilidades relacionadas à IA pode melhorar sua empregabilidade e prepará-los para as demandas deste mercado de trabalho em constante evolução. Além disso, compreendemos ser muito importante iniciativas para ampliar a educação em IA para públicos sub-representados em um esforço para aumentar a diversidade da força de trabalho.

Temos testemunhado que sociedades cada vez mais tecnológicas suscitam muitas questões de natureza ética. A velocidade com que os sistemas computacionais evoluem – em especial, os ditos inteligentes – tem nos mostrado que precisamos nos antecipar, de alguma forma, aos riscos que eles podem representar. Para alcançarmos desenvolvimento e utilização ética e responsável de sistemas de IA, precisamos contar com esforços coletivos e transdisciplinares, além de um diálogo constante entre governo, empresas, cientistas, especialistas e sociedade em geral. Nesse sentido, é fundamental promover debates mais amplos e plurais sobre os impactos dessas novas tecnologias, a fim de pensarmos de forma conjunta aplicações positivas dessas ferramentas em nossa sociedade.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Capítulo 21

E agora, PLN?

Maria das Graças Volpe Nunes
Helena de Medeiros Caseli

Publicado em: 26/09/2023
Atualizado em: 16/04/2026

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

Neste último Capítulo elencamos alguns desafios e perspectivas para o PLN em língua portuguesa e finalizamos com uma discussão sobre os limites atuais do PLN.

21.1 Desafios e perspectivas para o PLN-Português

Por razões históricas e econômicas, os sistemas atuais de PLN "estado da arte" são muito mais comuns em inglês do que em qualquer outra língua. Enquanto que outras comunidades têm adaptado para suas línguas os sistemas originalmente criados para o inglês (por meio de novos treinamentos, mas com aproveitamento de parâmetros), comunidades linguísticas minoritárias e comunidades linguísticas de países menos desenvolvidos são invisibilizadas no mundo digital, com consequências negativas e diretas na sua economia e desenvolvimento.

A comunidade de falantes do português no mundo é estimada, em 2025, em cerca de 300 milhões de pessoas (3,7% da população mundial) sendo o quinto idioma mais usado, depois do inglês, mandarim, hindi e espanhol. Contudo, essa representatividade não é contemplada no estado da arte da ciência, que está majoritariamente nas mãos de instituições e organizações não falantes do português. Pesquisadores brasileiros e portugueses têm levantado a necessidade de unir forças para colocar o português no lugar de destaque que ele merece[1].

O processamento do português brasileiro tem avançado de maneira consistente desde meados da década de 1990, principalmente a partir do uso de AM e de abordagens *cross-language* e multilíngue, que facilitam a construção rápida de recursos e soluções, e permitem a geração de uma aplicação em uma língua a partir de uma aplicação em outra língua. Mas ainda é precária a união de esforços entre os países da Comunidade de Países de Língua Portuguesa (CPLP), que inclui Portugal, Angola, Moçambique, Cabo Verde, Guiné-Bissau, São Tomé e Príncipe, além do Brasil. Se as diferenças linguísticas entre os diferentes idiomas representam barreiras para a criação de sistemas comuns, não há dúvida de que a união de esforços trará benefícios para todos. Por ora, o esforço mais visível é aquele entre os mais fortes do grupo, Brasil e Portugal, que realizam um evento científico bianual comum, o PROPOR[2], e mantêm vínculos acadêmicos há várias déca-

[1] https://www.publico.pt/2023/02/09/opiniao/opiniao/lingua-portuguesa-tecnologia-futuro-2038078

[2] CE-PLN. PROPOR (*International Conference on Computational Processing of Portuguese Language*). Disponível em: https://sites.google.com/view/ce-pln/eventos/propor.

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

das. No Brasil, os recursos de PLN compartilhados pela comunidade distribuem-se pelos centros de pesquisa, sendo dois exemplos o NILC[3] e o C4AI[4]. Em Portugal, dois importantes repositórios de recursos e ferramentas para português europeu e brasileiro são a Linguateca[5] e o Portulan Clarin[6].

Em países extensos como o Brasil, onde há uma grande variedade linguística, a exemplo das diferentes línguas indígenas faladas em território nacional[7], das variações dialetais e sociais e dos sotaques regionais do português brasileiro, suas riquezas e diversidades linguísticas dificilmente são representadas nos *corpora*. Essa sub-representação nos dados de treinamento de modelos de aprendizado de máquina é um dos fatores que contribuem para aumentar a codificação de vieses por esses sistemas. Percebe-se, portanto, a importância de os dados linguísticos que alimentam tais sistemas serem coletados de forma responsável, buscando representar as variações linguísticas e idiomáticas das línguas faladas no país.

Um dos primeiros *corpora* em português brasileiro usado para treinar um modelo de língua é o BrWac (*Brazilian Portuguese Web as corpus*), composto por 3,53 milhões de documentos da web, totalizando 2,68 bilhões de *tokens*, com acesso público para pesquisadores[8]. Já o *corpus* Carolina, do Centro de IA, C4AI[9], é, de acordo com os autores, "um *corpus* com um volume robusto de textos em Português Brasileiro contemporâneo (1970-2021), com informações de procedência e tipologia. O *corpus* está disponível em acesso aberto, para download gratuito, desde 8 de março de 2022. A versão atual, Ada 1.2 (8 de março de 2023), tem 823 milhões de *tokens*, mais de dois milhões de textos e mais de 11 GBs". Esse *corpus* é um importante passo para o treinamento de LLM do português brasileiro, e tem o mérito de incluir uma grande variedade de gêneros (jornalismo, literatura, poesia, judiciário, wikis, mídia social, legislativo, acadêmico etc.). Já na área de língua falada, o projeto TaRSila[10] tem como meta a construção de *datasets* robustos para o alcance de resultados no estado da arte para tarefas de reconhecimento e síntese de fala, identificação do falante e clonagem de voz. Destacamos o grande e multipropósito *corpus* de áudios, CORAA[11], alinhado com transcrições e manualmente validado para o treinamento de modelos de reconhecimento e síntese, bem como de análise de sentimentos usando características acústicas.

Na linha dos recursos para estimular o desenvolvimento aberto da IA generativa em português, em 2025, foram apresentados o GigaVerbo, uma concatenação de *corpora* de texto desduplicados em português, totalizando 200 bilhões de *tokens*, e o Tucano, uma série de *decoder-transformers* pré-treinados em português. A descrição completa desses recursos está em (Corrêa et al., 2025), e o código-fonte usado para treinamento e avaliação, entre outros subprodutos do estudo, está disponível no GitHub e no Huggin Faece.

No mesmo C4AI, o projeto PROINDL[12], em parceria com comunidades indígenas, investe no uso da IA para o desenvolvimento de ferramentas que promovam a preservação, revitalização e disseminação de línguas indígenas do Brasil. Entre as técnicas usadas estão

[3] https://sites.google.com/view/nilc-usp/resources-and-tools
[4] https://c4ai.inova.usp.br/pt/pesquisas_2/#NLP2_B_eng
[5] https://www.linguateca.pt/
[6] https://portulanclarin.net/
[7] https://www.gov.br/funai/pt-br/assuntos/noticias/2022-02/brasil-registra-274-linguas-indigenas-diferentes-faladas-por-305-etnias
[8] https://www.inf.ufrgs.br/pln/wiki/index.php?title=BrWaC
[9] https://sites.usp.br/corpuscarolina/
[10] TaRSila. Disponível em: https://sites.google.com/view/tarsila-c4ai.
[11] https://sites.google.com/view/tarsila-c4ai/coraa-versions
[12] https://conexoesoriginarias.github.io/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

as de AM que utilizam poucos dados para criar tradutores automáticos tanto para texto como para fala, além de outras aplicações.

Mesmo com a limitação de variedade e tamanho de *corpora* em português para treinamento de LLMs, grandes modelos de língua para o português podem ser encontrados. São modelos com capacidade multilíngue (ex. os modelos PALM da Google) ou treinados apenas em português (ex. BERTimbau (Souza et al., 2020), Sabiá (Pires et al., 2023b), Albertina[13] e o Bode[14]). Além desses recursos, de caráter geral, vários outros *corpora* de domínios específicos (nas áreas jurídica, médica, científica etc.) têm sido criados visando aplicações mais direcionadas a essas áreas, como os capítulos anteriores evidenciam.

Dessa forma, são claros os avanços em direção a produtos para a língua portuguesa. No entanto, o que pode parecer simples (*corpus* + redes neurais e Transformers + *fine-tuning* = LLM) pode ser, de fato, inviável. O custo de se produzir um LLM de qualidade é extremamente alto. Um estudo[15] mostrou que para treinar um modelo comparável ao GPT-4 é consumida cerca de 20-25 MW de energia durante cerca de 3 meses, o que é suficiente para abastecer cerca de 20 mil lares americanos. Considerando que os usuários do ChatGPT enviam 1 bilhão de mensagens/consultas por dia e estimando o gasto de 0,3 watt-horas por consulta, isso sugere que a inferência geral do ChatGPT requer cerca de 12,5 MW de energia.

São necessárias muitas GPUs para treinar modelos competitivos: quanto maior o número de GPUs, mais parâmetros podem ser usados no modelo, aumentando sua eficácia numa tarefa. Atualmente, poucas instituições públicas ou privadas dispõem de infraestrutura para tal e, ainda assim, com número de GPUs bastante inferior (de 2 a 100) àquela disponível em nuvem (clusters de TPUs[16]) com preços de aluguel que podem chegar a um milhão de dólares. Pesquisadores costumam recorrer a recursos gratuitos e temporários oferecidos pelas gigantes internacionais (ex. Google Cloud).

No ano de 2024, o Governo brasileiro anunciou um plano de incentivo para mitigar essa dependência externa por recursos essenciais ao desenvolvimento tecnológico. O Plano Brasileiro de Inteligência Artificial(PBIA)foi proposto por uma equipe multidisciplinar do Conselho Nacional de Ciência e Tecnologia (CNPq) e tem, como objetivos, "equipar o Brasil com infraestrutura tecnológica avançada com alta capacidade de processamento, incluindo um dos cinco supercomputadores mais potentes do mundo, alimentada por energias renováveis; desenvolver modelos avançados de linguagem em português, com dados nacionais que abarcam nossas características culturais, sociais e linguísticas, para fortalecer a soberania em IA e promover a liderança global do Brasil em IA por meio do desenvolvimento tecnológico nacional e ações estratégicas de colaboração internacional." O texto do plano[17] divulgado e aprovado em 2025, prevê a aplicação de cerca de R$23 bilhões até 2028 em investimentos em infraestrutura (data centers) e pesquisa e desenvolvimento de aplicativos para serviços nas áreas de saúde, educação, segurança, agricultura, entre outros. Como se vê, o PLN é um ator estratégico para o avanço tecnológico nacional.

Com a corrida de EUA e China para liderar a corrida pelo domínio da tecnologia de IA no mundo – i.e. construir recursos que possam ser usados por todo o mundo – levanta-se

[13] Família de modelos treinados para as variantes européia e brasileira do português disponível em: https://huggingface.co/PORTULAN.

[14] https://huggingface.co/recogna-nlp/bode-7b-alpaca-pt-br

[15] https://epoch.ai/gradient-updates/how-much-energy-does-chatgpt-use

[16] TPUs (Unidades de Processamento de Tensor) são aceleradores de treinamento e geração de modelos de machine learning.

[17] https://www.gov.br/mcti/pt-br/acompanhe-o-mcti/noticias/2024/07/plano-brasileiro-de-ia-tera-supercomputador-e-investimento-de-r-23-bilhoes-em-quatro-anos/ia_para_o_bem_de_todos.pdf/view

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

uma questão importante para os demais países consumidores desses recursos, em especial, o Brasil: como alcançar autossuficiência, independência, ou mesmo algum protagonismo nesse área?

Essas questões nos fazem refletir sobre os próximos caminhos a seguir. Nem tudo se resolve com grandes modelos de língua, pois há muitas aplicações interessantes que podem ser desenvolvidas ou com modelos mais modestos ou por meios distintos dos modelos de língua. Considerando tarefas e domínios de conhecimento particulares, é possível construir soluções a partir de modelos treinados apenas nesse domínio. De fato, os resultados tendem a ser melhores do que com o uso de modelos mais genéricos. Além disso, considerar uma tarefa mais específica pode levar a uma solução – qualquer que seja a abordagem – mais eficaz, cujo impacto na sociedade seja significativo.

O PLN vem sendo cada vez mais usado por empresas e startups da área, cujo número vem crescendo muito em nosso país – em 2025 são mais de 18 mil. Certamente isso é fruto da alta demanda por sistemas dessa natureza, mas também do investimento das universidades públicas na formação de recursos humanos nessa área. Estamos vivendo um momento de grande absorção dos profissionais de PLN pelo mercado. Mais um motivo para refletirmos sobre a formação desses profissionais frente aos grandes desafios que essa área (e a IA de modo geral) nos coloca, bem como sobre nosso papel no esclarecimento à sociedade sobre seus limites.

21.2 Há limites para o PLN?

Os sistemas do estado-da-arte em PLN têm impressionado seus usuários com desempenho até melhores que os humanos em alguns cenários (Capítulo ChatGPT, MariTalk e outros agentes de conversação). Isso significa que as metas do PLN já foram alcançadas? E quais seriam essas metas? Ainda perseguimos o objetivo de a máquina processar a língua tal como nós o fazemos? Será mesmo necessário que um sistema de PLN tenha plena capacidade de determinar o significado de expressões linguísticas ou basta que pareçam compartilhar conosco o significado que damos às expressões geradas por ele?

E o que dizer das metas atuais ou futuras? A comunidade tem se esforçado para tornar os sistemas gerativos mais confiáveis e transparentes, e um dos caminhos seria combinar o conhecimento simbólico e justificável com a eficiência dos modelos de redes neurais. Esse cenário parece ser o mais seguro, porém o mercado tecnológico não parece disposto a esperar. Dessa forma, enquanto novos cenários de aplicação festejam as ferramentas de IA, novos desafios vão se apresentando.

O PLN tem sido apresentado como uma área comum a duas disciplinas, Computação e Linguística e durante muito tempo olhar para e processar apenas a parte estrutural da língua pareceu possível ou mesmo suficiente. Com o passar do tempo, a evolução das máquinas e das técnicas, isso mudou e essa língua em uso no cenário digital atual só pode ser tratada de forma transdisciplinar. De fato, a língua tem sido objeto de estudo, análise e fascínio nas mais variadas áreas do conhecimento: Filosofia, Literatura, Linguística, Psicologia, Psicanálise, Ciências Cognitivas, Comunicação Social, entre outras, e do PLN. Isso revela que a língua é um objeto de estudo bastante rico e complexo, e, portanto, não é possível abordá-lo segundo uma ou duas disciplinas apenas.

Os capítulos anteriores evidenciam que PLN é uma área de grande potencial, porém repleta de desafios sobre os quais é difícil fazer previsões. Várias tarefas de IA têm sido solucionadas pelas tecnologias atuais (Redes Neurais, Aprendizado de Máquina), que não são ideias novas; elas ficaram adormecidas até que o hardware das máquinas pudesse

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

processá-las eficientemente. Em se tratando de PLN, no entanto, não é razoável prever que avanços de hardware, ou mesmo de métodos, garantam a solução completa para todas as aplicações que envolvem a língua. A demanda por sistemas que processam a língua não para de crescer. Vale notar que demandas e métodos são interdependentes: enquanto as demandas provocam novos métodos, estes últimos abrem caminho para demandas antes não possíveis.

Este livro também evidencia que os sistemas atuais de PLN espelham aquilo que aprendem a partir dos dados de treinamento dos algoritmos de aprendizado: língua na norma culta, informal, discursos de ódio, misoginia ou racismo; o que quer que tenha sido oferecido ao algoritmo de aprendizado a título de exemplo eventualmente será reproduzido pelo sistema gerado. Como o conhecimento (a língua) adquirido nesses sistemas não é explicitamente representado (ele está imerso em valores probabilísticos ou parâmetros numéricos das redes neurais), não há um controle de quando e como ele será usado. Todos esses efeitos colaterais preocupam a sociedade e trazem para a comunidade de PLN desafios e responsabilidades não existentes antes. E o esforço mundial para a regulamentação da criação e do uso de IA vai nos levar a cenários distintos dos atuais.

Convidamos você a nos acompanhar nessa jornada.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Referências

ABBOTT, B. Presuppositions and common ground. **Linguistics and philosophy**, v. 31, p. 523–538, 2008.

ABERCROMBIE, G. et al. **Mirages. On Anthropomorphism in Dialogue Systems**. (H. Bouamor, J. Pino, K. Bali, Eds.)Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. **Anais**...Singapore: Association for Computational Linguistics, dez. 2023. Disponível em: <https://aclanthology.org/2023.emnlp-main.290>

ABNEY, S. P. Parsing By Chunks. Em: BERWICK, R. C.; ABNEY, S. P.; TENNY, C. (Eds.). **Principle-Based Parsing: Computation and Psycholinguistics**. Dordrecht: Springer Netherlands, 1992. p. 257–278.

ACHIAM, J. et al. Gpt-4 technical report. **arXiv preprint arXiv:2303.08774**, 2023.

ACOSTA, O.; VILLAVICENCIO, A.; MOREIRA, V. **Identification and Treatment of Multiword Expressions Applied to Information Retrieval**. Proceedings of the Workshop on Multiword Expressions: from Parsing and Generation to the Real World. **Anais**...Portland, Oregon, USA: Association for Computational Linguistics, jun. 2011. Disponível em: <https://aclanthology.org/W11-0815>

AFANTENOS, S.; ASHER, N. **Counter-argumentation and discourse: A case study**. Proceedings of the Workshop on Frontiers and Connections between Argumentation Theory and Natural Language Processing. **Anais**...CEUR Workshop Proceedings, 2014.

AFONSO, S. et al. **Floresta sintá(c)tica: a treebank for Portuguese**. (M. G. Rodrigues, C. P. S. Araujo, Eds.)Proceedings of the Third International Conference on Language Resources and Evaluation (LREC 2002). **Anais**...Paris: ELRA, 2002.

AHN, L. VON; KEDIA, M.; BLUM, M. **Verbosity: A Game for Collecting Common-Sense Facts**. Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. **Anais**...: CHI '06.New York, NY, USA: Association for Computing Machinery, 2006. Disponível em: <https://doi.org/10.1145/1124772.1124784>

AI and Ethics. Springer, 2023. Disponível em: <https://link.springer.com/journal/43681/volumes-and-issues>. Acesso em: 7 abr. 2023

AKÇAY, M. B.; OĞUZ, K. Speech emotion recognition: Emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers. **Speech Communication**, v. 116, p. 56–76, 2020.

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

ALAM, T.; KHAN, A.; ALAM, F. **Punctuation Restoration using Transformer Models for High-and Low-Resource Languages**. Proceedings of the Sixth Workshop on Noisy User-generated Text (W-NUT 2020). **Anais**...Online: Association for Computational Linguistics, nov. 2020. Disponível em: <https://aclanthology.org/2020.wnut-1.18>

ALCAIM, A.; SOLEWICZ, J. A.; MORAES, J. A. DE. Freqüência de ocorrência dos fones e listas de frases foneticamente balanceadas no português falado no Rio de Janeiro. **Journal of Communication and Information Systems**, v. 7, n. 1, 1992.

ALEIXO, P.; PARDO, T. A. S. **CSTTool: um parser multidocumento automático para o Português do Brasil**. IV Workshop on MSc Dissertation and PhD Thesis in Artificial Intelligence–WTDIA. **Anais**...a2008.

ALEIXO, P.; PARDO, T. A. S. **CSTNews: um córpus de textos jornalísticos anotados segundo a teoria discursiva multidocumento CST (Cross-document Structure Theory**. [s.l.] Universidade de São Paulo (USP); São Carlos, SP, Brasil., b2008. Disponível em: <http://repositorio.icmc.usp.br//handle/RIICMC/6761>.

ALENCAR, L. F. DE. Donatus: uma interface amigável para o estudo da sintaxe formal utilizando a biblioteca em Python do NLTK. **Alfa: Revista de Linguística (São José do Rio Preto)**, v. 56, n. 2, p. 523–555, jul. 2012.

ALENCAR, L. F. DE; CUCONATO, B.; RADEMAKER, A. MorphoBr: an open source large-coverage full-form lexicon for morphological analysis of Portuguese. **Texto Livre**, v. 11, n. 3, p. 1–25, dez. 2018.

ALENCAR, R. Processos de categorização social: emergência de categorias sociais na fala em interação. **Revista Investigações**, v. 21, n. 2, p. 115–131, 2008.

ALENCAR, V.; ALCAIM, A. **LSF and LPC-derived features for large vocabulary distributed continuous speech recognition in Brazilian Portuguese**. 2008 42nd Asilomar Conference on Signals, Systems and Computers. **Anais**...IEEE, 2008.

ALISSON, S. **Their god is not our god**. Disponível em: <https://www.thecontinent.org/_files/ugd/287178_73f3d2af22614e678f277b631a62e491.pdf>. Acesso em: 11 jun. 2023.

ALLWOOD, J.; TRAUM, D.; JOKINEN, K. Cooperation, dialogue and ethics. **International Journal of Human-Computer Studies**, v. 53, n. 6, p. 871–914, 2000.

ALMASI, M.; SCHIØNNING, A. **Fine-Tuning GPT-3 for Synthetic Danish News Generation**. Proceedings of the 16th International Natural Language Generation Conference. **Anais**...2023.

ALTMANN, G. T.; KAMIDE, Y. Incremental interpretation at verbs: Restricting the domain of subsequent reference. **Cognition**, v. 73, n. 3, p. 247–264, 1999.

ALTMANN, G. T.; MIRKOVIĆ, J. Incrementality and prediction in human sentence

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

processing. **Cognitive science**, v. 33, n. 4, p. 583–609, 2009.

ALTUNYURT, L.; ORHAN, Z.; GÜNGÖR, T. **A Composite Approach for Part of Speech Tagging in Turkish**. 2006. Disponível em: <https://api.semanticscholar.org/CorpusID:9439761>

ALUÍSIO, S. et al. **An Account of the Challenge of Tagging a Reference Corpus for Brazilian Portuguese**. (N. J. Mamede et al., Eds.)Computational Processing of the Portuguese Language. **Anais**...Berlin, Heidelberg: Springer Berlin Heidelberg, 2003.

ALZUBAIDI, L. et al. A survey on deep learning tools dealing with data scarcity: definitions, challenges, solutions, tips, and applications. **Journal of Big Data**, v. 10, abr. 2023.

AMARAL, C. et al. **Priberam's question answering system in qa@ clef 2008**. Workshop of the Cross-Language Evaluation Forum for European Languages. **Anais**...Springer, 2008.

AMARAL, D. O. F. DO. **O reconhecimento de entidades nomeadas por meio de conditional random fields para a língua portuguesa**. Dissertação de Mestrado, Pontifícia Universidade Católica do Rio Grande do Sul, 2013.

ANACLETO, J. et al. **Can Common Sense uncover cultural differences in computer applications?** (M. Bramer, Ed.)Artificial Intelligence in Theory and Practice. **Anais**...Boston, MA: Springer US, 2006.

ANACLETO, J. C. et al. **A Common Sense-Based On-Line Assistant for Training Employees**. (C. Baranauskas et al., Eds.)Human-Computer Interaction – INTERACT 2007. **Anais**...Berlin, Heidelberg: Springer Berlin Heidelberg, 2007.

ANANTHAKRISHNAN, S.; NARAYANAN, S. S. Automatic Prosodic Event Detection Using Acoustic, Lexical, and Syntactic Evidence. **IEEE Transactions on Audio, Speech, and Language Processing**, v. 16, n. 1, p. 216–228, 2008.

ANDROUTSOPOULOS, I.; LAMPOURAS, G.; GALANIS, D. Generating Natural Language Descriptions from OWL Ontologies: The Natural OWL System. **Journal of Artificial Intelligence Research**, v. 48, n. 1, p. 671–715, out. 2013.

ANTONIO, J. D. Proposições relacionais e conversação: uma análise das relações estabelecidas nas trocas de turno. **Acta Scientiarum: Human and social sciences**, v. 25, p. 59, 2003.

ANTUNES, I. **Lutar com palavras: coesão e coerência**. [s.l.] Parábola, 2007.

ANTUNES, I. **Textualidade: noções básicas e implicações pedagógicas**. [s.l.] Editora: Parábola Editorial, 2017.

APPELT, D. E. **Problem Solving Applied to Language Generation**. Proceedings

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

of the 18th Annual Meeting on Association for Computational Linguistics. **Anais**...: ACL'80.Philadelphia, Pennsylvania: Association for Computational Linguistics, 1980. Disponível em: <https://doi.org/10.3115/981436.981455>

ARAÚJO, G. E. ET AL. **Certas Palavras: A 1980s-90s Brazilian Radio Corpus to Test TTS Models in Noisy Multi-Speaker Dialogues**. Proceedings of PROPOR 2026. **Anais**...2026.

ARDILA, R. et al. Common voice: A massively-multilingual speech corpus. **arXiv preprint arXiv:1912.06670**, 2019.

AROYO, L.; WELTY, C. Truth Is a Lie: Crowd Truth and the Seven Myths of Human Annotation. **AI Magazine**, v. 36, n. 1, p. 15–24, 2015.

Artificial intelligence and human rights. 1. ed. [s.l.] Dykinson, S.L., 2021.

ASHER, N. et al. **Discourse structure and dialogue acts in multiparty dialogue: the STAC corpus**. 10th International Conference on Language Resources and Evaluation (LREC 2016). **Anais**...2016.

ASHER, N.; LASCARIDES, A. **Logics of conversation**. [s.l.] Cambridge University Press, 2003.

ASHER, N.; VIEU, L. Subordinating and coordinating discourse relations. **Lingua**, v. 115, n. 4, p. 591–610, 2005.

AUER, S. et al. **DBpedia: A Nucleus for a Web of Open Data**. (K. Aberer et al., Eds.)The Semantic Web. **Anais**...Berlin, Heidelberg: Springer Berlin Heidelberg, 2007.

AVELAR, M.; FERRARI, L. Integração experiencial e dêixis locativa: O papel discursivo dos gestos. **Cadernos de Estudos Linguísticos**, v. 59, n. 1, p. 73–89, 2017.

AZEVEDO, R. R. DE. **Um sistema de diálogo inteligente baseado em lógica de descrições**. tese de doutorado—[s.l.] Universidade Federal de Pernambuco, 2015.

BAADER, F. et al. **The Description Logic Handbook: Theory, Implementation and Applications**. Cambridge, Reino Unido: Cambridge University Press, 2003.

BÄCKSTRÖM, T. et al. **Introduction to Speech Processing**. 2. ed. [s.l: s.n.].

BADENE, S. et al. **Learning Multi-party Discourse Structure Using Weak Supervision**. 25th International conference on computational linguistics and intellectual technologies (Dialogue 2019). **Anais**...2019.

BAEVSKI, A. et al. **wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations**., 2020. Disponível em: <https://arxiv.org/abs/2006.11477>

BAGGA, A.; BALDWIN, B. **Algorithms for Scoring Coreference Chains**. Pro-

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

ceedings of the first International Conference on Language Resources and Evaluation Workshop on Linguistics Coreference. **Anais**...Granada, Spain: 1998.

BAHDANAU, D.; CHO, K.; BENGIO, Y. **Neural Machine Translation by Jointly Learning to Align and Translate**. (Y. Bengio, Y. LeCun, Eds.)3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings. **Anais**...San Diego, California.: 2015. Disponível em: <http://arxiv.org/abs/1409.0473>

BAKER, C. F.; FILLMORE, C. J.; LOWE, J. B. **The Berkeley FrameNet Project**. 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics, Volume 1. **Anais**...Montreal, Quebec, Canada: Association for Computational Linguistics, ago. 1998. Disponível em: <https://aclanthology.org/P98-1013>

BAKER, C.; FELLBAUM, C.; PASSONNEAU, R. Semantic Annotation of MASC. Em: **Handbook of Linguistic Annotation**. [s.l.] Springer Netherlands, 2017. p. 699–717.

BALDWIN, T.; KIM, S. N. Multiword Expressions. Em: INDURKHYA, N.; DAMERAU, F. J. (Eds.). **Handbook of Natural Language Processing**. 2. ed. Boca Raton, FL, USA: CRC Press, Taylor; Francis Group, 2010. p. 267–292.

BALLOCCU, S. et al. **Leak, Cheat, Repeat: Data Contamination and Evaluation Malpractices in Closed-Source LLMs**. (Y. Graham, M. Purver, Eds.)Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers). **Anais**...St. Julian's, Malta: Association for Computational Linguistics, mar. 2024. Disponível em: <https://aclanthology.org/2024.eacl-long.5/>

BANARESCU, L. et al. **Abstract Meaning Representation for Sembanking**. Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse. **Anais**...Sofia, Bulgaria: Association for Computational Linguistics, 2013. Disponível em: <http://aclweb.org/anthology/W13-2322>

BANERJEE, S.; LAVIE, A. **METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments**. (J. Goldstein et al., Eds.)Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization. **Anais**...Ann Arbor, Michigan: Association for Computational Linguistics, jun. 2005. Disponível em: <https://aclanthology.org/W05-0909>

BANSAL, N.; AGARWAL, C.; NGUYEN, A. **SAM: The Sensitivity of Attribution Methods to Hyperparameters**. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). **Anais**...2020. Disponível em: <https://doi.org/10.1109/CVPRW50498.2020.00009>

BAPTISTA, J.; HAGÈGE, C.; MAMEDE, N. Identificação, classificação e normalização de expressões temporais do português: A experiência do Segundo HAREM e o futuro. Em:

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

MOTA, C.; SANTOS, D. (Eds.). **Desafios na avaliação conjunta do reconhecimento de entidades mencionadas**. [s.l.] Linguateca, 2008. p. 33–54.

BAPTISTA, J.; MAMEDE, N.; REIS, S. **Support Verb Constructions across the Ocean Sea**. (A. Bhatia et al., Eds.)Proceedings of the 18th Workshop on Multiword Expressions @LREC2022. **Anais**...Marseille, France: European Language Resources Association, jun. 2022. Disponível em: <https://aclanthology.org/2022.mwe-1.6>

BARREIRA, R.; PINHEIRO, V.; FURTADO, V. **FrameFOR – Uma Base de Conhecimento de Frames Semânticos para Perícias de Informática (FrameFOR - a Knowledge Base of Semantic Frames for Digital Forensics)[In Portuguese]**. Proceedings of the 11th Brazilian Symposium in Information and Human Language Technology. **Anais**...Uberlândia, Brazil: Sociedade Brasileira de Computação, out. 2017. Disponível em: <https://aclanthology.org/W17-6620>

BARREIRO, A. et al. When Multiwords Go Bad in Machine Translation. **MT Summit workshop Proceedings on Multi-word Units in Machine Translation and Transla tion Technology**, p. 10, 2013.

BARROS, D. L. P. DE. Procedimentos e recursos discursivos da conversação. **Estudos de língua falada: variações e confrontos**, v. 3, p. 47, 1999.

BARROS, D. L. P. DE. Introdução à Linguística II: princípios de análise. Em: FIORIN, J. L. (Ed.). 5. ed. São Paulo: Contexto, 2021. p. 187–219.

BARZILAY, R.; LAPATA, M. **Collective Content Selection for Concept-to-text Generation**. Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing. **Anais**...: HLT'05.Vancouver, British Columbia, Canada: Association for Computational Linguistics, 2005. Disponível em: <https://doi.org/10.3115/1220575.1220617>

BARZILAY, R.; LAPATA, M. **Aggregation via Set Partitioning for Natural Language Generation**. Proceedings of the Main Conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics. **Anais**...: HLT-NAACL'06.New York, New York: Association for Computational Linguistics, 2006. Disponível em: <https://doi.org/10.3115/1220835.1220881>

BASILE, V. et al. **It's the end of the gold standard as we know it. On the impact of pre-aggregation on the evaluation of highly subjective tasks**. CEUR Workshop Proceedings. **Anais**...CEUR-WS, 2020. Disponível em: <https://iris.unito.it/handle/2318/1770149>

BASILE, V. et al. **We Need to Consider Disagreement in Evaluation**. Proceedings of the 1st Workshop on Benchmarking: Past, Present and Future. **Anais**...Online: Association for Computational Linguistics, ago. 2021. Disponível em: <https://aclanthology.org/2021.bppf-1.3>

BASSO, R. M. **A Semântica das Relações Anafóricas entre Eventos**. tese de

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

doutorado—[s.l.] Universidade Estadual de Campinas, SP, 2009.

BATES, M. et al. Research in Knowledge Representation for Natural Language Understanding: Bolt, Beranek, and Newman. **SIGART Bull.**, n. 79, p. 30–31, jan. 1982.

BATISTA, C.; DIAS, A. L.; NETO, N. Free resources for forced phonetic alignment in Brazilian Portuguese based on Kaldi toolkit. **EURASIP Journal on Advances in Signal Processing**, v. 2022, n. 1, p. 11, 19 fev. 2022.

BAVELAS, J. B. et al. Interactive gestures. **Discourse Processes**, v. 15, n. 4, p. 469–489, 1992.

BAVELAS, J. B. **Face-to-face dialogue: theory, research, and applications**. [s.l.] Oxford University Press, 2022.

BAVELAS, J. B.; COATES, L.; JOHNSON, T. Listener responses as a collaborative process: The role of gaze. **Journal of communication**, v. 52, n. 3, p. 566–580, 2002.

BAVELAS, J. B.; GERWING, J. Conversational hand gestures and facial displays in face-to-face dialogue. Em: **Social communication**. [s.l.] Psychology Press, 2007. p. 283–308.

BAYYARAPU, H. S. **Efficient algorithm for Context Sensitive Aggregation in Natural Language generation**. Proceedings of the International Conference Recent Advances in Natural Language Processing. **Anais**...: RANLP'11.Hissar, Bulgaria: Association for Computational Linguistics, 2011. Disponível em: <http://aclanthology.coli.uni-saarland.de/pdf/R/R11/R11-1012.pdf>

BEATTIE, G. W. Sequential Temporal Patterns of Speech and Gaze in Dialogue. **Semiotica**, v. 23, n. 1/2, 1978.

BECKMAN, M. E.; HIRSCHBERG, J.; SHATTUCK-HUFNAGEL, S. The original ToBI system and the evolution of the ToBI framework. Em: JUN, S.-A. (Ed.). **Prosodic typology: the phonology of intonation and phrasing**. Oxford: Oxford University Press, 2005. p. 9–54.

BEJČEK, E.; STRAŇÁK, P.; PECINA, P. **Syntactic Identification of Occurrences of Multiword Expressions in Text using a Lexicon with Dependency Structures**. Proceedings of the 9th Workshop on Multiword Expressions. **Anais**...Atlanta, Georgia, USA: Association for Computational Linguistics, jun. 2013. Disponível em: <https://aclanthology.org/W13-1016>

BELHASIN, O.; BAR-SHALOM, G.; EL-YANIV, R. **TransBoost: improving the best ImageNet performance using deep transduction**. Proceedings of the 36th International Conference on Neural Information Processing Systems. **Anais**...: NIPS '22.Red Hook, NY, USA: Curran Associates Inc., 2022.

BELINKOV, Y.; GLASS, J. Analysis Methods in Neural Language Processing: A Survey.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Transactions of the Association for Computational Linguistics, v. 7, p. 49–72, 2019.

BELZ, A. Last Words: That's Nice ... What Can You Do With It? **Computational Linguistics**, v. 35, n. 1, mar. 2009.

BELZ, A. et al. **A Systematic Review of Reproducibility Research in Natural Language Processing**. Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume. **Anais**...Online: Association for Computational Linguistics, abr. 2021. Disponível em: <https://aclanthology.org/2021.eacl-main.29>

BELZ, A. A Metrological Perspective on Reproducibility in NLP*. **Computational Linguistics**, v. 48, n. 4, p. 1125–1135, dez. 2022.

BELZ, A. et al. **Non-Repeatable Experiments and Non-Reproducible Results: The Reproducibility Crisis in Human Evaluation in NLP**. Findings of the Association for Computational Linguistics: ACL 2023. **Anais**...Toronto, Canada: Association for Computational Linguistics, jul. a2023. Disponível em: <https://aclanthology.org/2023.findings-acl.226>

BELZ, A.; THOMSON, C.; REITER, E. **Missing Information, Unresponsive Authors, Experimental Flaws: The Impossibility of Assessing the Reproducibility of Previous Human Evaluations in NLP**. The Fourth Workshop on Insights from Negative Results in NLP. **Anais**...Dubrovnik, Croatia: Association for Computational Linguistics, b2023. Disponível em: <https://aclanthology.org/2023.insights-1.1>

BENDER, E. M. **Linguistic Fundamentals for Natural Language Processing: 100 Essentials from Morphology and Syntax**. Springer Nature Switzerland AG 2013: Springer Cham, 2013. p. XVII–166

BENDER, E. M. **The Power of Linguistics - Unpacking Natural Language Processing Ethics with Emily M. Bender. [Podcast]**. Disponível em: <https://www.radicalai.org/e16-emily-bender>. Acesso em: 7 abr. 2023.

BENDER, E. M. et al. **On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?** . Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency. **Anais**...: FAccT '21.New York, NY, USA: Association for Computing Machinery, 2021. Disponível em: <https://doi.org/10.1145/3442188.3445922>

BENDER, E. M. **You Are Not a Parrot And a chatbot is not a human. And a linguist named Emily M. Bender is very worried what will happen when we forget this**. Disponível em: <https://nymag.com/intelligencer/article/ai-artificial-intelligence-chatbots-emily-m-bender.html>. Acesso em: 9 abr. 2023.

BENDER, E. M. Resisting Dehumanization in the Age of "AI". **Current Directions in Psychological Science**, v. 0, n. 0, p. 09637214231217286, 2024.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

BENDER, E. M.; FRIEDMAN, B. Data Statements for Natural Language Processing: Toward Mitigating System Bias and Enabling Better Science. **Transactions of the Association for Computational Linguistics**, v. 6, p. 587–604, 2018.

BENDER, E. M.; KOLLER, A. **Climbing towards NLU: On Meaning, Form, and Understanding in the Age of Data**. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. **Anais**...Online: Association for Computational Linguistics, jul. 2020. Disponível em: <https://aclanthology.org/2020.acl-main.463>

BENGIO, Y. et al. A Neural Probabilistic Language Model. **J. Mach. Learn. Res.**, v. 3, n. null, p. 1137–1155, mar. 2003.

BENNETT, A. **Interruptions and the interpretation of conversation**. Annual Meeting of the Berkeley Linguistics Society. **Anais**...1978. Disponível em: <https://doi.org/10.1080/01638538109544513>

BENOTTI, L.; BLACKBURN, P. **Grounding as a Collaborative Process**. Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume. **Anais**...Online: Association for Computational Linguistics, abr. 2021. Disponível em: <https://aclanthology.org/2021.eacl-main.41>

BERG-KIRKPATRICK, T.; BURKETT, D.; KLEIN, D. **An Empirical Investigation of Statistical Significance in NLP**. Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning. **Anais**...Jeju Island, Korea: Association for Computational Linguistics, jul. 2012. Disponível em: <https://aclanthology.org/D12-1091>

BERNARDO, S. Episódio e evento na organização tópica da conversa informal. **Soletras**, v. 1, p. 34–49, 2001.

BERNARDO, S. Então e agora na conversa informal. **Soletras**, v. 5-6, p. 65–81, 2003.

BERNARDO, S. Papel das formas O? H e O? H em turnos conversacionais. **Revista do GELNE**, v. 7, n. 1/2, p. 73–88, 2005.

BERNARDO, S. P. Foco e ponto de vista na organização conversacional. **Pesquisas em Lingüística e Literatura: Descrição, Aplicação, Ensino - ISBN: 85-906478-0-3**, 2002.

BERNARDO, S. P.; VELOZO, N. DE A.; ABREU, J. C. DE. Espaços mentais na conceptualização de conversa: dois modelos em análise. **Revista do GELNE**, v. 23, n. 1, p. 201–216, 2021.

BERNSEN, N. O.; DYBKJÆR, H.; DYBKJÆR, L. Cooperativity in human-machine and human-human spoken dialogue. **Discourse processes**, v. 21, n. 2, p. 213–236, 1996.

BERTAGLIA, T. F. C.; NUNES, M. DAS G. V. **Exploring Word Embeddings for Unsupervised Textual User-Generated Content Normalization**. Proceedings of

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

the 2nd Workshop on Noisy User-generated Text (WNUT). **Anais**...Osaka, Japan: The COLING 2016 Organizing Committee, dez. 2016. Disponível em: <https://aclanthology.org/W16-3916>

BERTAGLIA, T. F. C.; NUNES, M. DAS G. V. **Normalização textual de conteúdo gerado por usuário**. mathesis—[s.l.] Universidade de São Paulo, 2017.

BERTOLDI, A. **Os Limites da Criação Automática de Léxicos Computacionais Baseados em Frames: Um Estudo Contrastivo do Frame Criminal_process (The Limits of the Automatic Creation of Frame-based Computational Lexicons: a Contrastive Study of the Criminal_process Frame) [in Portuguese]**. Proceedings of the 8th Brazilian Symposium in Information and Human Language Technology. **Anais**...2011. Disponível em: <https://aclanthology.org/W11-4510>

BIANCHI, F.; HOVY, D. **On the Gap between Adoption and Understanding in NLP**. Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021. **Anais**...Online: Association for Computational Linguistics, ago. 2021. Disponível em: <https://aclanthology.org/2021.findings-acl.340>

BICK, E. **The Parsing System "Palavras": Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework**. tese de doutorado—[s.l.] Aarhus University Press, Denmark; University of Arhus, 2000.

BICK, E. **A dependency-based approach to anaphora annotation**. Proceedings of th 9th International Conference on Computational Processing of the Portuguese Language. **Anais**...Porto Alegre, Brazil: 2010.

BICK, E. S. PFN-PT: A Framenet Annotator for Portuguese: Anotação semântica automática: um novo Framenet para o português. **Domínios de Linguagem**, v. 16(4)7, p. 1401–1435, 2009.

BIDERMAN, M. T. C. **Teoria linguística: linguística quantitativa e computacional**. Rio de Janeiro: Martins Fontes, 1978.

BIRD, S. **Decolonising Speech and Language Technology**. Proceedings of the 28th International Conference on Computational Linguistics. **Anais**...Barcelona, Spain (Online): International Committee on Computational Linguistics, dez. 2020. Disponível em: <https://aclanthology.org/2020.coling-main.313>

BIRD, S.; LOPER, E. **NLTK: The Natural Language Toolkit**. Proceedings of the ACL Interactive Poster and Demonstration Sessions. **Anais**...Barcelona, Spain: Association for Computational Linguistics, jul. 2004. Disponível em: <https://aclanthology.org/P04-3031>

BIRON, T. et al. Automatic detection of prosodic boundaries in spontaneous speech. **PLoS ONE**, v. 16, n. 5, p. 1–21, maio 2021.

BIZER, C. et al. DBpedia: A crystallization point for the Web of Data. **Web Semantics**,

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

2009.

BLACKBURN, P.; BOS, J. **Representation and Inference for Natural Language: A First Course in Computational Semantics**. [s.l.] Center for the Study of Language; Information, 2005.

BOBROW, D. G. et al. GUS, a frame-driven dialog system. **Artificial Intelligence**, v. 8, n. 2, p. 155–173, 1977.

BOERSMA, P.; WEENINK, D. **Praat: doing phonetics by computer [Computer program]. Version 6.3.10**., 2023. Disponível em: <http://www.praat.org/>

BOGANTES, D. et al. **Towards Lexical Encoding of Multi-Word Expressions in Spanish Dialects**. Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16). **Anais**...Portorož, Slovenia: European Language Resources Association (ELRA), 2016. Disponível em: <https://aclanthology.org/L16-1358>

BOJANOWSKI, P. et al. Enriching Word Vectors with Subword Information. **Transactions of the Association for Computational Linguistics**, v. 5, p. 135–146, 2017.

BOND, F.; FOSTER, R. **Linking and extending an open multilingual wordnet**. Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). **Anais**...Sofia, Bulgaria: Association for Computational Linguistics, ago. 2013. Disponível em: <https://aclanthology.org/P13-1133>

BOS, J. et al. Survey of existing interactive systems. **Trindi (Task Oriented Instructional Dialogue) report**, v. D1, p. 3, 1999.

BOTELHO, J. M. Conversação: Mudança e desvio de tópico conversacional. **Revista Philologus**, v. 17, n. 50, 2011.

BOTT, S. et al. **GhoSt-PV: A Representative Gold Standard of German Particle Verbs**. (M. Zock, A. Lenci, S. Evert, Eds.)Proceedings of the 5th Workshop on Cognitive Aspects of the Lexicon (CogALex - V). **Anais**...Osaka, Japan: The COLING 2016 Organizing Committee, dez. 2016. Disponível em: <https://aclanthology.org/W16-5318>

BOTZER, N. et al. **TK-KNN: A Balanced Distance-Based Pseudo Labeling Approach for Semi-Supervised Intent Classification**. (H. Bouamor, J. Pino, K. Bali, Eds.)Findings of the Association for Computational Linguistics: EMNLP 2023. **Anais**...Singapore: Association for Computational Linguistics, dez. 2023. Disponível em: <https://aclanthology.org/2023.findings-emnlp.429/>

BOUAMOR, D.; SEMMAR, N.; ZWEIGENBAUM, P. **Identifying bilingual Multi-Word Expressions for Statistical Machine Translation**. (N. C. (Conference. Chair) et al., Eds.)Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12). **Anais**...Istanbul, Turkey: European Language Resources Association (ELRA), maio 2012.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

BOUAYAD-AGHA, N. et al. **Content selection from semantic web data**. INLG 2012 Proceedings of the Seventh International Natural Language Generation Conference. **Anais**...2012.

BOWMAN, S. R. et al. **A large annotated corpus for learning natural language inference**. Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. **Anais**...Lisbon, Portugal: Association for Computational Linguistics, set. 2015. Disponível em: <https://aclanthology.org/D15-1075>

BOXER, D. **Applying sociolinguistics: Domains and face-to-face interaction**. [s.l.] John Benjamins Publishing, 2002. v. 15

BRAGGAAR, A. et al. Evaluating Task-oriented Dialogue Systems: A Systematic Review of Measures, Constructs and their Operationalisations. **arXiv preprint arXiv:2312.13871**, 2023.

BRANDOM, R. B. **Articulating Reasons: An Introduction to Inferentialism**. Cambridge, Massachusetts, EUA: Harvard University Press, 2001.

BRASCHLER, M.; PETERS, C. CLEF 2002 Methodology and Metrics. Em: PETERS, C. (Ed.). **Advances in Cross-Language Information Retrieval: Results of the CLEF 2002 Evaluation Campaign**. [s.l.] Springer, 2003. p. 512–525.

BRASCHLER, M.; PETERS, C. Cross-Language Evaluation Forum: Objectives, Results, Achievements. **Information Retrieval**, v. 7, n. 1-2, p. 7–31, 2004.

BRAUDE, D. A.; SHIMODAIRA, H.; YOUSSEF, A. B. **Template-warping based speech driven head motion synthesis.** Interspeech. **Anais**...2013.

BREEN, J. **JMdict: a Japanese-Multilingual Dictionary**. Proceedings of the Workshop on Multilingual Linguistic Resources. **Anais**...Geneva, Switzerland: COLING, 2004. Disponível em: <https://aclanthology.org/W04-2209>

BRENNAN, S. E.; GALATI, A.; KUHLEN, A. K. Two minds, one dialog: Coordinating speaking and understanding. Em: **Psychology of learning and motivation**. [s.l.] Elsevier, 2010. v. 53p. 301–344.

BREWSTER, C.; WILKS, Y. **Ontologies, taxonomies, thesauri:learning from texts**. (M. Deegan, Ed.)Proceedings of Use of Computational Linguistics in the Extraction of Keyword Information from Digital Library Content Workshop. **Anais**...2004. Disponível em: <http://www.cbrewster.com/papers/KeyWord_FMO.pdf>

BRILL, E. **A Simple Rule-Based Part of Speech Tagger**. Proceedings of the Third Conference on Applied Natural Language Processing. **Anais**...: ANLC '92.USA: Association for Computational Linguistics, 1992. Disponível em: <https://doi.org/10.3115/974499.974526>

BRITO, I. A. et al. **ToxSyn-PT: A Large-Scale Synthetic Dataset for Hate Speech**

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Detection in Portuguese., 2025. Disponível em: <https://arxiv.org/abs/2506.10245>

BROWN, T. B. et al. **Language Models are Few-Shot Learners**. (H. Larochelle et al., Eds.)Advances in Neural Information Processing Systems. **Anais**...Curran Associates, Inc., 2020. Disponível em: <https://proceedings.neurips.cc/paper/2020/hash/1457c0d6bfcb4967418bfb8ac142f64a-Abstract.html>

BRUNEAU, T. J. Communicative silences: Forms and functions. **Journal of communication**, v. 23, n. 1, p. 17–46, 1973.

BUCKLEY, C.; VOORHEES, E. **Evaluating Evaluation Measure Stability**. Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. **Anais**...2000. Disponível em: <https://sigir.org/wp-content/uploads/2017/06/p235.pdf>

BULHÕES, J. DO S. U. et al. **Levantamento, análise e descrição de elementos paralinguísticos do português espontâneo**. mathesis—[s.l.] Universidade Federal do Pará, 2006.

BUOLAMWINI, J.; GEBRU, T. **Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification**. (S. A. Friedler, C. Wilson, Eds.)Proceedings of the 1st Conference on Fairness, Accountability and Transparency. **Anais**...: Proceedings of Machine Learning Research.PMLR, 2018. Disponível em: <https://proceedings.mlr.press/v81/buolamwini18a.html>

BUTNARIU, C. et al. **SemEval-2 Task 9: The Interpretation of Noun Compounds Using Paraphrasing Verbs and Prepositions**. Proceedings of the 5th International Workshop on Semantic Evaluation. **Anais**...Uppsala, Sweden: Association for Computational Linguistics, jul. 2010. Disponível em: <https://aclanthology.org/S10-1007>

CABRAL, L. et al. **FakeWhastApp.BR: NLP and Machine Learning Techniques for Misinformation Detection in Brazilian Portuguese WhatsApp Messages**. Proceedings of the 23rd International Conference on Enterprise Information Systems (ICEIS 2021) - Volume 1. **Anais**...2021.

CABRÉ, M. T. **La terminología: representación y comunicación**. [s.l.] Editora: Documenta Universitaria, 1999.

CAÇÃO, F. N. et al. **DEEPAGÉ: Answering Questions in Portuguese About the Brazilian Environment**. (A. Britto, K. Valdivia Delgado, Eds.)Intelligent Systems. **Anais**...Cham: Springer International Publishing, 2021.

CAFFERKEY, C.; HOGAN, D.; GENABITH, J. VAN. **Multi-word units in treebank-based probabilistic parsing and generation**. Proc. of RANLP 2007. **Anais**...Borovets: 2007.

CALLISON-BURCH, C. et al. **Findings of the 2010 joint workshop on statistical**

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

machine translation and metrics for machine translation. Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR. **Anais**...2010.

CALZOLARI, N. et al. **Towards best Practice for Multiword Expressions in Computational Lexicons**. proc of the Third lrecconf (LREC 2002). **Anais**...Las Palmas, Canary Islands, Spain: elra, 2002.

CAMPOS, J. et al. **Towards Fully Automated News Reporting in Brazilian Portuguese**. Anais do XVII Encontro Nacional de Inteligência Artificial e Computacional. **Anais**...Porto Alegre, RS, Brasil: SBC, 2020. Disponível em: <https://sol.sbc.org.br/index.php/eniac/article/view/12158>

CAMPRESS (ED.). **Cambridge International Dictionary of Phrasal Verbs**. Cambridge, UK: campress, 1997.

CANDIDO JUNIOR, A. et al. CORAA: a large corpus of spontaneous and prepared speech manually validated for speech recognition in Brazilian Portuguese. **CoRR**, v. abs/2110.15731, 2021.

CANDIDO JUNIOR, A. et al. CORAA ASR: a large corpus of spontaneous and prepared speech manually validated for speech recognition in Brazilian Portuguese. **Language Resources & Evaluation**, 2022.

CANDITO, M. et al. A French corpus annotated for multiword expressions and named entities. **Journal of Language Modelling**, v. 8, n. 2, p. 415–479, 2021.

CANDITO, M.; CONSTANT, M. **Strategies for Contiguous Multiword Expression Analysis and Dependency Parsing**. Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). **Anais**...Baltimore, Maryland: Association for Computational Linguistics, jun. 2014. Disponível em: <https://aclanthology.org/P14-1070>

CAP, F. et al. **How to Produce Unseen Teddy Bears: Improved Morphological Processing of Compounds in SMT**. Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics (EACL). **Anais**...Goteborg, Sweden: 2014.

CARAPINHA, C.; PLAG, C. **A interação verbal em sala de audiências: turn design**. Actas do XIII Congreso Internacional de Lingüística Xeral: Vigo, 13-15 de xuño de 2018. **Anais**...Universidade de Vigo, 2018. Disponível em: <http://cilx2018.uvigo.gal/actas/pdf/661468.pdf>

CARDOSO, N. **Avaliação de Sistemas de Reconhecimento de Entidades Mencionadas**. mathesis—[s.l.] Faculdade de Engenharia da Universidade do Porto, 2006.

CARDOSO, N. **Rembrandt - a named-entity recognition framework**. Proceedings of the Eighth International Conference on Language Resources and Evaluation. **Anais**...Istanbul, Turkey: 2012. Disponível em: <http://www.lrec-conf.org/proceed

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

ings/lrec2012/summaries/409.html>

CARDOSO, P. C. F. et al. **CSTNews-a discourse-annotated corpus for single and multi-document summarization of news texts in Brazilian Portuguese**. Proceedings of the 3rd RST Brazilian Meeting. **Anais**...2011.

CARDOSO, P. C. F. **Exploração de métodos de sumarização automática multi-documento com base em conhecimento semântico-discursivo**. tese de doutorado—[s.l.] (Doutorado em Ciências de Computação e Matemática Computacional) - Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, 2014.

CARLSON, L.; MARCU, D. Discourse tagging reference manual. **ISI Technical Report ISI-TR-545**, v. 54, n. 2001, p. 56, 2001.

CARPUAT, M.; DIAB, M. **Task-based Evaluation of Multiword Expressions: a Pilot Study in Statistical Machine Translation**. Proceedings of HLT: The 2010 Annual Conference of the North American Chapter of the ACL (NAACL 2003). **Anais**...Los Angeles, California: ACL, jun. 2010.

CARVALHO, G.; MATOS, D. M. DE; ROCIO, V. **IdSay: Question Answering for Portuguese**. (C. Peters et al., Eds.)Evaluating Systems for Multilingual and Multimodal Information Access. **Anais**...Berlin, Heidelberg: Springer Berlin Heidelberg, 2009.

CARVALHO, M. W. P. L.; ACIOLI, M. D. Entre falas simultâneas, tomadas de turno e sobreposição de vozes: quem tem a palavra no debate? **Revista do GELNE**, v. 19, p. 155–165, 2017.

CASANOVA, E. **Síntese de voz aplicada ao português brasileiro usando aprendizado profundo**. {B.S.} thesis—[s.l.] Universidade Tecnológica Federal do Paraná, 2019.

CASANOVA, E. et al. TTS-Portuguese Corpus: a corpus for speech synthesis in Brazilian Portuguese. **Language Resources and Evaluation**, v. 56, n. 3, p. 1043–1055, 2022.

CASANOVA, E.; SHULBY, C. D.; ALUÍSIO, S. M. Deep learning approaches for speech synthesis and speaker verification. **Acoustic communication: an interdisciplinary approach**, 2021.

CASELI, H. DE M.; FREITAS, C.; VIOLA, R. Processamento de Linguagem Natural. Em: **Tópicos em Gerenciamento de Dados e Informações: Minicursos do SBBD 2022**. [s.l.] Sociedade Brasileira de Computação, 2022. p. 1–28.

CASTILHO, A. T. DE. O português culto falado no Brasil: história do Projeto NURC. Em: PRETI, D.; URBANO, H. (Eds.). **A linguagem falada culta na cidade de São Paulo**. São Paulo, SP: TAQ/Fapesp, 1990. v. 4 – Estudosp. 141–292.

CASTILHO, A. T. DE. Gramática do Português Brasileiro: fundamentos, perspectivas. **Cadernos de Linguística**, v. 2, n. 1, p. e252, abr. 2021.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

CASTRO FERREIRA, T. et al. **NeuralREG: An end-to-end approach to referring expression generation**. Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). **Anais**...Melbourne, Australia: Association for Computational Linguistics, 2018. Disponível em: <http://aclweb.org/anthology/P18-1182>

CASTRO FERREIRA, T. et al. **Neural data-to-text generation: A comparison between pipeline and end-to-end architectures**. Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). **Anais**...Hong Kong, China: Association for Computational Linguistics, nov. 2019. Disponível em: <https://www.aclweb.org/anthology/D19-1052>

CASTRO FERREIRA, T. et al. **Evaluating Recognizing Question Entailment Methods for a Portuguese Community Question-Answering System about Diabetes Mellitus**. (R. Mitkov, G. Angelova, Eds.)Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021). **Anais**...Held Online: INCOMA Ltd., set. 2021. Disponível em: <https://aclanthology.org/2021.ranlp-1.28>

CASTRO FERREIRA, T.; PARABONI, I. Referring Expression Generation: Taking Speakers' Preferences into Account. Em: SOJKA, P. et al. (Eds.). **Text, Speech and Dialogue**. Lecture Notes em Computer Science. [s.l.] Springer International Publishing, 2014a. v. 8655p. 539–546.

CASTRO FERREIRA, T.; PARABONI, I. **Classification-based Referring Expression Generation**. Computational Linguistics and Intelligent Text Processing (CICLing-2014), Lecture Notes in Computer Science 8403. **Anais**...Kathmandu, Nepal: Springer, b2014.

CAVALIERE, P.; ROMEO, G. From Poisons to Antidotes: Algorithms as Democracy Boosters. **European Journal of Risk Regulation**, v. 13, n. 3, p. 421–442, 2022.

CERVONE, A.; STEPANOV, E.; RICCARDI, G. **Coherence Models for Dialogue**. Proc. Interspeech 2018. **Anais**...2018. Disponível em: <https://10.21437/Interspeech.2018-2446>

CHALAMALASETTI, K. et al. **clembench: Using Game Play to Evaluate Chat-Optimized Language Models as Conversational Agents**. (H. Bouamor, J. Pino, K. Bali, Eds.)Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. **Anais**...Singapore: Association for Computational Linguistics, dez. 2023. Disponível em: <https://aclanthology.org/2023.emnlp-main.689>

CHANDRAN, R. **Indigenous groups in NZ, US fear colonisation as AI learns their languages**. Disponível em: <https://www.context.news/ai/nz-us-indigenous-fear-colonisation-as-bots-learn-their-languages>. Acesso em: 7 abr. 2023.

CHANG, K.-W. et al. **Illinois-Coref: The UI system in the CoNLL-2012 shared**

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

task. Joint Conference on EMNLP and CoNLL-Shared Task. **Anais**...Association for Computational Linguistics, 2012.

CHAPELLE, O.; SCHOLKOPF, B.; ZIEN, A. **Semi-Supervised Learning**. [s.l.] The MIT Press, 2006.

CHARPENTIER, F.; STELLA, M. **Diphone synthesis using an overlap-add technique for speech waveforms concatenation**. ICASSP'86. IEEE International Conference on Acoustics, Speech, and Signal Processing. **Anais**...IEEE, 1986.

CHE, X. et al. **Punctuation Prediction for Unsegmented Transcript Based on Word Vector**. Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16). **Anais**...Portorož, Slovenia: European Language Resources Association (ELRA), 2016. Disponível em: <https://aclanthology.org/L16-1103>

CHEN, K.; HASEGAWA-JOHNSON, M. A. **How prosody improves word recognition**. Speech Prosody 2004. **Anais**...2004.

CHEN, L.-W.; RUDNICKY, A. **Exploring Wav2vec 2.0 Fine Tuning for Improved Speech Emotion Recognition**. ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). **Anais**...IEEE, 2023.

CHINCHOR, N. **The statistical significance of the MUC-4 results**. Proceedings of the Fourth Message Understanding Conference (MUC-4). **Anais**...Morgan Kaufmann Publ., 1992. Disponível em: <https://dl.acm.org/doi/pdf/10.3115/1072064.1072068>

CHISHOLM, A.; RADFORD, W.; HACHEY, B. **Learning to generate one-sentence biographies from Wikidata**. Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers. **Anais**...: EACL'17.Valencia, Spain: Association for Computational Linguistics, 2017. Disponível em: <http://aclanthology.coli.uni-saarland.de/pdf/E/E17/E17-1060.pdf>

CHOUEKA, Y. **Looking for Needles in a Haystack or Locating Interesting Collocational Expressions in Large Textual Databases**. (C. Fluhr, D. E. Walker, Eds.)Proceedings of the 2nd International Conference on Computer-Assisted Information Retrieval (Recherche d'Information et ses Applications - RIA 1988). **Anais**...Cambridge, MA, USA: CID, 1988.

CHOVIL, N. Discourse-oriented facial displays in conversation. **Research on Language & Social Interaction**, v. 25, n. 1-4, p. 163–194, 1991.

CHUNG, Y.-A.; GLASS, J. **Speech2Vec: A Sequence-to-Sequence Framework for Learning Word Embeddings from Speech**. Proc. Interspeech 2018. **Anais**...2018.

CHURCH, K. How many multiword expressions do people know? **tslp Special Issue on mwes: from theory to practice and use, part 1 (TSLP)**, v. 10, n. 2, 2013.

CIERI, C.; MILLER, D.; WALKER, K. **The Fisher Corpus: a Resource for the Next**

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Generations of Speech-to-Text. Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04). **Anais**...Lisbon, Portugal: European Language Resources Association (ELRA), 2004. Disponível em: <http://www.lrec-conf.org/proceedings/lrec2004/pdf/767.pdf>

CLARK, E. V. Conversational repair and the acquisition of language. **Discourse Processes**, v. 57, n. 5-6, p. 441–459, 2020.

CLARK, H. H. **Arenas of language use**. [s.l.] University of Chicago Press, 1992.

CLARK, H. H. **Using language**. [s.l.] Cambridge University Press, 1996a.

CLARK, H. H. Communities, commonalities, and communication. Em: **Rethinking linguistic relativity**. [s.l.] Cambridge University Press, 1996b. v. 17p. 324–355.

CLARK, H. H. How to talk with children. Em: **Language in Interaction**. [s.l.] John Benjamins, 2014. p. 333–352.

CLARK, H. H.; BRENNAN, S. E. Grounding in communication. Em: **Perspectives on socially shared cognition.** [s.l.] American Psychological Association, 1991. p. 127–149.

CLARK, H. H.; SCHAEFER, E. F. Collaborating on contributions to conversations. **Language and cognitive processes**, v. 2, n. 1, p. 19–41, 1987.

CLARK, H. H.; TREE, J. E. F. Using uh and um in spontaneous speaking. **Cognition**, v. 84, n. 1, p. 73–111, 2002.

CLARK, H. H.; WILKES-GIBBS, D. Referring as a collaborative process. **Cognition**, v. 22, n. 1, p. 1–39, 1986.

CLERWALL, C. Enter the Robot Journalist. **Journalism Practice**, v. 8, n. 5, p. 519–531, 2014.

COECKELBERGH, M. Artificial Intelligence, Responsibility Attribution, and a Relational Justification of Explainability. **Science and Engineering Ethics**, v. 26, p. 2051–2068, 2020.

COELHO DA SILVA, T.; FERNANDES DE MACÊDO, J.; MAGALHÃES, R. **Tracking the Evolution of Covid-19 Symptoms through Clinical Conversations**. Proceedings of the 5th Clinical Natural Language Processing Workshop. **Anais**...Toronto, Canada: Association for Computational Linguistics, jul. 2023. Disponível em: <https://aclanthology.org/2023.clinicalnlp-1.6>

COELHO, G. E.; SERRALHEIRO, A. J.; NETO, J. P. **A spoken dialog system speech interface based on a microphone array**. Computational Processing of the Portuguese Language: 8th International Conference, PROPOR 2008 Aveiro, Portugal, September 8-10, 2008 Proceedings 8. **Anais**...Springer, 2008. Disponível em: <https://doi.org/10.1007/978-3-540-85980-2_3>

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

COHEN, J. A Coefficient of Agreement for Nominal Scales. **Educational and Psychological Measurement**, v. 20, n. 1, p. 37–46, 1960.

COHEN, K. B. et al. **Three Dimensions of Reproducibility in Natural Language Processing**. Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018). **Anais**...Miyazaki, Japan: European Language Resources Association (ELRA), 2018. Disponível em: <https://aclanthology.org/L18-1025>

COHEN, P. R.; HOWE, A. E. How Evaluation Guides AI Research: The Message Still Counts More than the Medium. **AI Magazine**, v. 9, n. 4, p. 35, 1988.

COLLOVINI, S. et al. **Summ-it: Um Corpus Anotado com Informações Discursivas Visando a Sumarização Automática**. Proceedings of V Workshop em Tecnologia da Informação e da Linguagem Humana. **Anais**...Rio de Janeiro, Brasil: 2007.

COLLOVINI, S. et al. **Extraction of Relation Descriptors for Portuguese Using Conditional Random Fields**. Proceedings of the 14th Ibero-American Conference on Advances in Artificial Intelligence. **Anais**...Santiago de Chile: 2014.

COLLOVINI, S. et al. **IberLEF 2019 Portuguese Named Entity Recognition and Relation Extraction Tasks**. Proceedings of the Iberian Languages Evaluation Forum co-located with 35th Conference of the Spanish Society for Natural Language Processing. **Anais**...2019. Disponível em: <http://ceur-ws.org/Vol-2421/NER/_Portuguese/_overview.pdf>

COMMISSION, E. **Proposal for a Regulation laying down harmonised rules on artificial intelligence**. Disponível em: <https://digital-strategy.ec.europa.eu/en/library/proposal-regulation-laying-down-harmonised-rules-artificial-intelligence >. Acesso em: 28 ago. 2023.

CONSTANT, M. et al. Multiword Expression Processing: A Survey. **Computational Linguistics**, 2017.

CONSTANT, M.; NIVRE, J. **A Transition-Based System for Joint Lexical and Syntactic Analysis**. Proc. of ACL 2016. **Anais**...Berlin: 2016.

COPESTAKE, A. et al. Minimal recursion semantics: An introduction. **Research on language and computation**, v. 3, p. 281–332, 2005.

CORDEIRO, S. R. et al. Unsupervised Compositionality Prediction of Nominal Compounds. **Computational Linguistics**, v. 45, n. 1, p. 1–57, 2019.

COREIXAS, T. **Resolução De Correferência E Categorias De Entidades Nomeadas**. Dissertação de Mestrado, Pontifícia Universidade Católica do Rio Grande do Sul, 2010.

CORRÊA, N. K. et al. Worldwide AI ethics: A review of 200 guidelines and recommendations for AI governance. **Patterns**, v. 4, n. 10, p. 100857, 2023.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

CORRÊA, N. K. et al. Tucano: Advancing neural text generation for Portuguese. **Patterns**, p. 101325, jul. 2025.

CORRÊA, U. B. et al. Overview of the IDPT Task on Irony Detection in Portuguese at IberLEF 2021. **Procesamiento del Lenguaje Natural**, v. 67, p. 269–276, 2021.

CORTES, E. et al. **An Empirical Comparison of Question Classification Methods for Question Answering Systems**. (N. Calzolari et al., Eds.)Proceedings of the Twelfth Language Resources and Evaluation Conference. **Anais**...Marseille, France: European Language Resources Association, 2020. Disponível em: <https://aclanthology.org/2020.lrec-1.665>

CORTES, E. G.; WOLOSZYN, V.; BARONE, D. A. C. **When, Where, Who, What or Why? A Hybrid Model to Question Answering Systems**. (A. Villavicencio et al., Eds.)Computational Processing of the Portuguese Language. **Anais**...Cham: Springer International Publishing, 2018.

COSTA, L. F. **Using Answer Retrieval Patterns to Answer Portuguese Questions**. (C. Peters et al., Eds.)Evaluating Systems for Multilingual and Multimodal Information Access. **Anais**...Berlin, Heidelberg: Springer Berlin Heidelberg, 2009.

COSTA, L. F.; CABRAL, L. M. **Answering Portuguese Questions**. (A. Teixeira et al., Eds.)Computational Processing of the Portuguese Language. **Anais**...Berlin, Heidelberg: Springer Berlin Heidelberg, 2008.

COSTA, P. B. DA; PARABONI, I. Transferência de estilo textual arbitrário em português. **Linguamática**, v. 15, n. 2, p. 19–36, 2023.

COSTA, P. B. DA; PARABONI, I. **Sequence-to-sequence and transformer approaches to Portuguese text style transfer**. Proceedings of the 16th International Conference on Computational Processing of Portuguese. **Anais**...Santiago de Compostela, Galicia/Spain: Association for Computational Lingustics, mar. 2024. Disponível em: <https://aclanthology.org/2024.propor-1.54>

COSTA, P.; PARABONI, I. **Personality-dependent Neural Text Summarization**. Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019). **Anais**...2019.

COUCKE, A. et al. Snips voice platform: an embedded spoken language understanding system for private-by-design voice interfaces. **arXiv preprint arXiv:1805.10190**, 2018.

COUILLAULT, A. et al. **Evaluating corpora documentation with regards to the Ethics and Big Data Charter**. Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14). **Anais**...Reykjavik, Iceland: European Language Resources Association (ELRA), 2014. Disponível em: <http://www.lrec-conf.org/proceedings/lrec2014/pdf/424_Paper.pdf>

CRESTI, E. et al. **The C-ORAL-ROM CORPUS. A Multilingual Resource of**

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Spontaneous Speech for Romance Languages. Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04). **Anais**...Lisbon, Portugal: European Language Resources Association (ELRA), 2004. Disponível em: <http://www.lrec-conf.org/proceedings/lrec2004/pdf/357.pdf>

CRISTEA, D.; IDE, N.; ROMARY, L. **Veins theory: A model of global discourse cohesion and coherence**. Coling-ACL Conference. **Anais**...1998.

CROCKER, M. W. Computational psycholinguistics. **The handbook of computational linguistics and natural language processing**, p. 482–513, 2010.

CRUSE, D. A. **Lexical Semantics**. Cambridge, UK: campress, 1986.

CRUZ, B. S. **Concessionária do Metrô de SP é processada por ter câmeras que leem nossas emoções**. Disponível em: <https://www.uol.com.br/tilt/noticias/redacao/2018/08/31/concessionaria-do-metro-de-sp-e-processada-por-ter-cameras-que-leem-emocoes.htm >. Acesso em: 29 ago. 2023.

CRUZ, B. S. **Racismo Calculado**. Disponível em: <https://www.uol.com.br/tilt/reportagens-especiais/como-os-algoritmos-espalham-racismo/#cover >. Acesso em: 29 ago. 2023.

CRUZ, J. A. DA et al. **Creating an Academic Conversational Agent for Dynamic Information Retrieval**. Proceedings of the XVI Brazilian Symposium on Information Systems. **Anais**...: SBSI '20.New York, NY, USA: Association for Computing Machinery, 2020. Disponível em: <https://doi.org/10.1145/3411564.3411647>

CUNHA RECUERO, R. DA. Elementos para a análise da conversação na comunicação mediada pelo computador. **Verso e Reverso**, v. 22, 2008.

DAHL, V. Natural language processing and logic programming. **Journal of Logic Programming**, v. 19-20, n. 1, p. 681–714, 1994.

DALE, R.; HADDOCK, N. **Generating referring expressions involving relations**. Proceedings of the fifth conference on European chapter of the Association for Computational Linguistics. **Anais**...: EACL'91.Berlin, Germany: Association for Computational Linguistics, 1991.

DANESCU-NICULESCU-MIZIL, C. et al. **Echoes of power: Language effects and power differences in social interaction**. Proceedings of the 21st international conference on World Wide Web. **Anais**...2012. Disponível em: <https://doi.org/10.1145/2187836.2187931>

DANTAS, A. C. et al. **AstroBot: Um chatbot com inteligência artificial para auxiliar no processo de ensino e aprendizagem de física**. Anais dos Workshops do Congresso Brasileiro de Informática na Educação. **Anais**...2019. Disponível em: <https://doi.org/10.5753/cbie.wcbie.2019.1196>

DE PAIVA, V. et al. **An overview of Portuguese wordnets**. Proceedings of the 8th

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Global WordNet Conference (GWC). **Anais**...2016.

DE PAIVA, V.; RADEMAKER, A.; MELO, G. DE. **OpenWordNet-PT: An Open Brazilian Wordnet for Reasoning**. Proceedings of COLING 2012: Demonstration Papers. **Anais**...2012.

DEEMTER, K. VAN. **Designing Algorithms for Referring with Proper Names**. Proceedings of the 9th International Natural Language Generation conference. **Anais**...: INLG'16.Edinburgh, UK: Association for Computational Linguistics, a2016. Disponível em: <http://www.aclweb.org/anthology/W16-6605>

DEEMTER, K. VAN. **Computational Models of Referring. A Study in Cognitive Science**. Cambridge, Massachusetts, USA: MIT Press, 2016b.

DEMPSEY, P. The teardown: Google Home personal assistant. **Engineering & Technology**, v. 12, n. 3, p. 80–81, 2017.

DERIU, J. et al. Survey on evaluation methods for dialogue systems. **Artificial Intelligence Review**, v. 54, p. 755–810, 2021.

DEVLIN, J. et al. **BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding**. (J. Burstein, C. Doran, T. Solorio, Eds.)Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019. **Anais**...Minneapolis, MN, USA: Association for Computational Linguistics, 2019. Disponível em: <https://doi.org/10.18653/v1/n19-1423>

DHUMAL DESHMUKH, R.; KIWELEKAR, A. **Deep Learning Techniques for Part of Speech Tagging by Natural Language Processing**. 2020 2nd International Conference on Innovative Mechanisms for Industry Applications (ICIMIA). **Anais**...mar. 2020.

DI FELIPPO, A. et al. The DANTEStocks Corpus: an analysis of the distribution of Universal Dependencies-based Part-of-Speech tags. **Revista da ABRALIN**, v. 22, n. 2, p. 249–271, 2024.

DI GANGI, M. A. et al. **MuST-C: a Multilingual Speech Translation Corpus**. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). **Anais**...Minneapolis, Minnesota: Association for Computational Linguistics, jun. 2019. Disponível em: <https://aclanthology.org/N19-1202>

DIAS-DA-SILVA, B. C. **A face tecnológica dos estudos da linguagem: o processamento automático das línguas naturais. 1996. 272f**. tese de doutorado—[s.l.] Tese (Doutorado em Lingüística e Língua Portuguesa)–Faculdade de Ciências e ..., 1996.

DIAS-DA-SILVA, B. C. **Wordnet.Br: An Exercise of Human Language Technology Research**. Proceedings of the Third International WordNet Conference. **Anais**...2005. Disponível em: <http://semanticweb.kaist.ac.kr/conference/gwc/pdf2006

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

/6.pdf>

DODDINGTON, G. et al. **The Automatic Content Extraction (ACE) Program: Tasks, Data, and Evaluation**. (M. T. Lino et al., Eds.)Proceedings of LREC'2004, Fourth International Conference on Language resources and Evaluation (Lisboa, 26-28 May 2004). **Anais**...2004. Disponível em: <http://www.lrec-conf.org/proceedings/lrec2004/pdf/5.pdf>

DODGE, J. et al. **Show Your Work: Improved Reporting of Experimental Results**. Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). **Anais**...Hong Kong, China: Association for Computational Linguistics, nov. 2019. Disponível em: <https://aclanthology.org/D19-1224>

DONG, L. et al. **Learning to Generate Product Reviews from Attributes**. Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers. **Anais**...: EACL'17.Valencia, Spain: Association for Computational Linguistics, 2017. Disponível em: <http://aclanthology.coli.uni-saarland.de/pdf/E/E17/E17-1059.pdf>

DROR, R. et al. **The Hitchhiker's Guide to Testing Statistical Significance in Natural Language Processing**. Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). **Anais**...Melbourne, Australia: Association for Computational Linguistics, jul. 2018. Disponível em: <https://aclanthology.org/P18-1128>

DROR, R. et al. **Statistical significance testing for natural language processing**. [s.l.] Springer, 2020.

DU BOIS, J. W. et al. **Santa Barbara corpus of spoken American English. Parts 1–4**. Philadelphia: Linguistic Data Consortium, 2000--2005.

DU BOIS, J. W. et al. **Discourse transcription**. Santa Barbara: Department of Linguistics, University of California, 1992. v. 4

DUMA, D.; KLEIN, E. **Generating Natural Language from Linked Data: Unsupervised template extraction**. Proceedings of the 10th International Conference on Computational Semantics (IWCS 2013) – Long Papers. **Anais**...Potsdam, Germany: Association for Computational Linguistics, 2013. Disponível em: <http://www.aclweb.org/anthology/W13-0108>

DUNCAN, S. Some signals and rules for taking speaking turns in conversations. **Journal of personality and social psychology**, v. 23, n. 2, p. 283, 1972.

DUNIETZ, J. The field of natural language processing is chasing the wrong goal. **MIT Technology Review**, 2020.

DURAN, M. S. et al. **The Dawn of the Porttinari Multigenre Treebank: In-**

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

troducing its Journalistic Portion. Anais do XIV Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana. **Anais**...Porto Alegre, RS, Brasil: SBC, 2023. Disponível em: <https://sol.sbc.org.br/index.php/stil/article/view/25443>

DURAN, M. S.; ALUÍSIO, S. M. **Propbank-Br: a Brazilian Treebank Annotated with Semantic Role Labels**. Proceedings of the 8th International Conference on Language Resources and Evaluation - LREC. **Anais**...2012.

DUŠEK, O.; NOVIKOVA, J.; RIESER, V. Findings of the E2E NLG challenge. **arXiv preprint arXiv:1810.01170**, 2018.

EBDEN, P.; SPROAT, R. The Kestrel TTS text normalization system. **Natural Language Engineering**, v. 21, p. 333–353, maio 2014.

EDMONDS, P.; KILGARRIFF, A. Introduction to the special issue on evaluating word sense disambiguation systems. **Natural Language Engineering**, v. 8, n. 4, p. 279–291, 2002.

EIJCK, J. VAN; UNGER, C. **Computational Semantics with Functional Programming**. [s.l.] Cambridge University Press, 2010.

EKMAN, P. An argument for basic emotions. **Cognition and Emotion**, v. 6, n. 3-4, p. 169–200, 1992.

EL AYADI, M.; KAMEL, M. S.; KARRAY, F. Survey on speech emotion recognition: Features, classification schemes, and databases. **Pattern recognition**, v. 44, n. 3, p. 572–587, 2011.

EMPOLI, G. DA. **Os engenheiros do caos: Como as fake news, as teorias da conspiração e os algoritmos estão sendo utilizados para disseminar ódio, medo e influenciar eleições**. [s.l.] Vestígio Editora, 2019.

ENGELMANN, D. C. et al. **A conversational agent to support hospital bed allocation**. Brazilian Conference on Intelligent Systems. **Anais**...Springer, 2021. Disponível em: <https://doi.org/10.1007/978-3-030-91702-9_1>

ERYIĞIT, G. et al. **Annotation and Extraction of Multiword Expressions in Turkish Treebanks**. Proceedings of the 11th Workshop on Multiword Expressions. **Anais**...Denver, Colorado: Association for Computational Linguistics, jun. 2015. Disponível em: <https://aclanthology.org/W15-0912>

ESSENFELDER, R.; RODRIGUES, V. P. Seqüências inseridas: fluência e disfluência em uma conversação espontânea. **Revista Virtual de Estudos da Linguagem–ReVEL**, v. 3, n. 4, 2005.

ETHAYARAJH, K.; JURAFSKY, D. **Utility is in the Eye of the User: A Critique of NLP Leaderboards**. Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). **Anais**...Online: Association for Computational

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Linguistics, nov. 2020. Disponível em: <https://aclanthology.org/2020.emnlp-main.393>

EVALDO LEAL, S. et al. **MuPe Life Stories Dataset: Spontaneous Speech in Brazilian Portuguese with a Case Study Evaluation on ASR Bias against Speakers Groups and Topic Modeling**. (O. Rambow et al., Eds.)Proceedings of the 31st International Conference on Computational Linguistics. **Anais**...Abu Dhabi, UAE: Association for Computational Linguistics, jan. 2025. Disponível em: <https://aclanthology.org/2025.coling-main.407/>

EVERT, S. Corpora and collocations. Em: LÜDELING, A.; KYTÖ, M. (Eds.). **Corpus Linguistics: An International Handbook**. [s.l.] De Gruyter Mouton, 2009. v. 2p. 1212–1248.

EVERT, S.; KRENN, B. **Methods for the Qualitative Evaluation of Lexical Association Measures**. Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics. **Anais**...Toulouse, France: Association for Computational Linguistics, jul. 2001. Disponível em: <https://aclanthology.org/P01-1025>

FARAHMAND, M.; SMITH, A.; NIVRE, J. **A Multiword Expression Data Set: Annotating Non-Compositionality and Conventionalization for English Noun Compounds**. proc of the 11th Workshop on mwes (MWE 2015). **Anais**...Denver, Colorado, USA: acl, 2015. Disponível em: <http://aclweb.org/anthology/W15-0904>

FÁVERO, L. L.; ANDRADE, M. L. DA C. V. DE O.; AQUINO, Z. G. O. DE. Perguntas e respostas como mecanismos de coesão e coerência no texto falado. **Gramática do português falado**, v. 4, p. 473–508, 1996.

FÁVERO, L. L.; ANDRADE, M. L. DA C. V. DE O.; AQUINO, Z. G. O. DE. Discurso e interação: a reformulação nas entrevistas. **DELTA: Documentação de Estudos em Lingüística Teórica e Aplicada**, v. 14, p. 91–103, 1998.

FAYEK, H. M. **Speech Processing for Machine Learning: Filter banks, Mel-Frequency Cepstral Coefficients (MFCCs) and What's In-Between.**, 2016. Disponível em: <https://haythamfayek.com/2016/04/21/speech-processing-for-machine-learning.html>

FELLBAUM, C. **WordNet: An Electronic Lexical Database**. [s.l.] The MIT Press, 1998.

FERLA, J. R. Discurso reportado em narrativas: a construção colaborativa de histórias na fala-em-interação. **Trabalho de conclusão de curso. Universidade do Vale do Rio dos Sinos**, 2020.

FERNANDES, E. R.; SANTOS, C. N. DOS; MILIDIÚ, R. L. Latent trees for coreference resolution. **Computational Linguistics**, 2014.

FERNANDES, U. DA S. et al. **Analyzing MoLIC's Applicability to Model the Interaction of Conversational Agents: A Case Study on ANA Chatbot**. Procee-

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

dings of the XX Brazilian Symposium on Human Factors in Computing Systems. **Anais**...: IHC '21.New York, NY, USA: Association for Computing Machinery, 2021. Disponível em: <https://doi.org/10.1145/3472301.3484367>

FERRADEIRA, J. E. DE S. **Resolução de anáfora pronominal**. mathesis—[s.l.] Universidade Nova de Lisboa; Dissertação de Mestrado, Universidade Nova de Lisboa, 1993.

FERREIRA, A. et al. **Agentes de conversação para idosos, plataforma Guardião**. Anais Estendidos do XVIII Simpósio Brasileiro sobre Fatores Humanos em Sistemas Computacionais. **Anais**...Porto Alegre, RS, Brasil: SBC, 2019. Disponível em: <https://sol.sbc.org.br/index.php/ihc_estendido/article/view/8386>

FERREIRA, F.; SWETS, B. How incremental is language production? Evidence from the production of utterances requiring the computation of arithmetic sums. **Journal of Memory and Language**, v. 46, n. 1, p. 57–84, 2002.

FERREIRA, T. C. **Advances in Natural Language Generation: Generating varied outputs from semantic inputs**. tese de doutorado—[s.l.] Tilburg University, 2018.

FERREIRA, T. C. et al. **The 2020 bilingual, bi-directional webnlg+ shared task overview and evaluation results (webnlg+ 2020)**. Proceedings of the 3rd International Workshop on Natural Language Generation from the Semantic Web (WebNLG+). **Anais**...2020.

FILLMORE, C. J. et al. **Frame semantics and the nature of language**. Annals of the New York Academy of Sciences: Conference on the origin and development of language and speech. **Anais**...New York, 1976.

FILLMORE, C. J.; KAY, P.; O'CONNOR, M. C. Regularity and Idiomaticity in Grammatical Constructions: The Case of Let Alone. **Language**, v. 64, p. 501–538, 1988.

FINCH, S. E.; CHOI, J. D. **Towards Unified Dialogue System Evaluation: A Comprehensive Analysis of Current Evaluation Protocols**. Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue. **Anais**...1st virtual meeting: Association for Computational Linguistics, jul. 2020. Disponível em: <https://aclanthology.org/2020.sigdial-1.29>

FINLAYSON, M.; KULKARNI, N. **Detecting Multi-Word Expressions Improves Word Sense Disambiguation**. Proc. of the ACL 2011 Workshop on MWEs. **Anais**...Portland, OR: 2011.

FIRTH, J. R. The technique of semantics. **Transactions of the philological society**, v. 34, n. 1, p. 36–73, 1957.

FLAKE, J. K.; FRIED, E. I. Measurement Schmeasurement: Questionable Measurement Practices and How to Avoid Them. **Advances in Methods and Practices in Psychological Science**, v. 3, n. 4, p. 456–465, 2020.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

FLEISS, J. L. Measuring nominal scale agreement among many raters. **Psychological Bulletin**, v. 76, n. 5, p. 378–382, 1971.

FOKKENS, A. et al. **Offspring from Reproduction Problems: What Replication Failure Teaches Us**. Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). **Anais**...Sofia, Bulgaria: Association for Computational Linguistics, ago. 2013. Disponível em: <https://aclanthology.org/P13-1166>

FONSECA, E. B. **Resolução de correferências em língua portuguesa: pessoa, local e organização**. Dissertação de Mestrado, Pontifícia Universidade Católica do Rio Grande do Sul, 2014.

FONSECA, E. B. et al. **Summ-it++: an enriched version of the summ-it corpus**. Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16). **Anais**...a2016.

FONSECA, E. B. **Resolução de correferência nominal usando semântica em língua portuguesa**. tese de doutorado—[s.l.] Pontifícia Universidade Católica do Rio Grande do Sul; Pontifícia Universidade Católica do Rio Grande do Sul, 2018.

FONSECA, E. B.; VIEIRA, R.; VANIN, A. **Dealing With Imbalanced Datasets For Coreference Resolution**. Proceedings of The Twenty-Eighth International Flairs Conference. **Anais**...a2015.

FONSECA, E. B.; VIEIRA, R.; VANIN, A. **Adapting an Entity Centric Model for Portuguese Coreference Resolution**. Portorož, Slovenia, b2016.

FONSECA, E. B.; VIEIRA, R.; VANIN, A. **CORP: Coreference Resolution for Portuguese**., c2016.

FONSECA, E. B.; VIEIRA, R.; VANIN, A. A. Coreference Resolution In Portuguese: Detecting Person, Location And Organization. **Journal of the Brazilian Computational Intelligence Society**, v. 12, n. 2, p. 86–97, 2014.

FONSECA, E. R. et al. Visão geral da avaliação de similaridade semântica e inferência textual. **Linguamática**, v. 8, n. 2, p. 3–13, d2016.

FONSECA, E. R.; ROSA, J. L. G. **Mac-Morpho Revisited: Towards Robust Part-of-Speech Tagging**. Proceedings of the 9th Brazilian Symposium in Information and Human Language Technology. **Anais**...2013. Disponível em: <https://aclanthology.org/W13-4811>

FONSECA, E. R.; ROSA, J. L.; ALUÍSIO, S. M. Evaluating word embeddings and a revised corpus for part-of-speech tagging in Portuguese. **Journal of the Brazilian Computer Society**, v. 21, n. 1, p. 32–38, fev. b2015.

FONSECA, E.; VANIN, A.; VIEIRA, R. **Mention clustering to improve portuguese**

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

semantic coreference resolution. International Conference on Applications of Natural Language to Information Systems. **Anais**...Springer, 2018.

FORT, K.; ADDA, G.; COHEN, K. B. Last Words: Amazon Mechanical Turk: Gold Mine or Coal Mine? **Computational Linguistics**, v. 37, n. 2, p. 413–420, jun. 2011.

FORTE MARTINS, A. D. et al. **Detection of misinformation about covid-19 in Brazilian Portuguese WhatsApp messages**. International Conference on Applications of Natural Language to Information Systems. **Anais**...Springer, 2021.

FREGE, G. Über Sinn und Bedeutung. **Zeitschrift für Philosophie und philosophische Kritik**, v. 100, p. 25–50, 1892/19601892/1960.

FREIRE, P. **Pedagogia do Oprimido**. Rio de Janeiro: Paz e Terra/SA, 1989.

FREITAS, C. et al. Relações semânticas do ReRelEM: além das entidades no Segundo HAREM. Em: MOTA, C.; SANTOS, D. (Eds.). **Desafios na avaliação conjunta do reconhecimento de entidades mencionadas**. [s.l.] Linguateca, 2008a. p. 77–96.

FREITAS, C. et al. **Relation detection between named entities: report of a shared task**. Proceedings of the Workshop on Semantic Evaluations: Recent Achievements and Future Directions. **Anais**...Boulder, Colorado: a2009.

FREITAS, C. et al. **Detection of relations between named entities: report of a shared task**. Proceedings of the NAACL HLT Workshop on Semantic Evaluations: Recent Achievements and Future Directions, SEW-2009. **Anais**...Boulder, Colorado, USA: b2009. Disponível em: <https://comum.rcaap.pt/bitstream/10400.26/20504/1/FreitasetalSEW2009.pdf>

FREITAS, C. et al. **Second HAREM: Advancing the State of the Art of Named Entity Recognition in Portuguese**. Proceedings of the International Conference on Language Resources and Evaluation. **Anais**...Valletta, Malta: 2010. Disponível em: <http://www.lrec-conf.org/proceedings/lrec2010/summaries/412.html>

FREITAS, C. et al. O que é uma resposta? Notas de uns avaliadores estafados. **Linguamática**, v. 4, n. 1, p. 67–75, 2012.

FREITAS, C. et al. **Tagsets and Datasets: Some Experiments Based on Portuguese Language**. (A. Villavicencio et al., Eds.)Computational Processing of the Portuguese Language. **Anais**...Cham: Springer International Publishing, 2018.

FREITAS, C. **Linguística Computacional**. [s.l.] Parábola Editorial, 2022.

FREITAS, C. et al. Recursos linguísticos para o PLN específico de domínio: o Petrolês. **Linguamática**, v. 15, n. 2, p. 51–68, 2023.

FREITAS, C.; PARDO, T. A. S. PropBanks e representações semânticas: o que temos, o que queremos e o que podemos. **Linguamática**, v. 17, n. 2, 2025.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

FREITAS, C.; ROCHA, P.; BICK, E. **Floresta sintá(c)tica: bigger, thicker and easier**. International Conference on Computational Processing of the Portuguese Language. **Anais**...Springer, b2008.

FREITAS, C.; SANTOS, D. **Gender Depiction in Portuguese: Distant reading Brazilian and Portuguese literature**. 2nd Annual Conference of Computational Literary Studies. **Anais**...2023. Disponível em: <https://www.linguateca.pt/Diana/download/FreitasSantos2023-2ndCCLS.pdf>

FREITAS, C.; SOUZA, E. Sujeito oculto às claras: uma abordagem descritivo-computacional / Omitted subjects revealed: a quantitative-descriptive approach. **REVISTA DE ESTUDOS DA LINGUAGEM**, v. 29, n. 2, p. 1033–1058, 2021.

FREITAS, C.; SOUZA, E. DE. A study on methods for revising dependency treebanks: in search of gold. **Language Resources and Evaluation**, v. 58, n. 1, p. 11–131, 2024.

FRESCHI, A. C. **A avaliação por pares no teletandem institucional integrado: um estudo de caso sobre o feedback linguístico nas sessões orais em português**. mathesis—[s.l.] Universidade Estadual Paulista (Unesp), 2017.

FRIEDMAN, B. et al. Value sensitive design and information systems. **Early engagement and new technologies: Opening up the laboratory**, p. 55–95, 2013.

GAGO, P. C. Questões de transcrição em análise da conversa. **Veredas-Revista de Estudos Linguísticos**, v. 6, n. 2, 2002.

GAIZAUSKAS, R. **Evaluating Language Processing Applications and Components**., 2003. Disponível em: <https://www.linguateca.pt/Repositorio/rgaizauskasPROPOR2003.pdf>

GAMMERMAN, A.; VOVK, V.; VAPNIK, V. **Learning by Transduction**. : UAI'98.San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1998.

GAO, Y. et al. **Retrieval-Augmented Generation for Large Language Models: A Survey**., 2024. Disponível em: <https://arxiv.org/abs/2312.10997>

GARCEZ, P. DE M. A organização da fala-em-interação na sala de aula: controle social, reprodução de conhecimento, construção conjunta de conhecimento. **Calidoscópio**, v. 4, n. 1, p. 66–80, 2006.

GARCEZ, P. M.; LODER, L. L. Reparo iniciado e levado a cabo pelo outro na conversa cotidiana em português do Brasil. **DELTA: Documentação de Estudos em Lingüística Teórica e Aplicada**, v. 21, p. 279–312, 2005.

GARCIA, M. et al. **Probing for idiomaticity in vector space models**. Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume. **Anais**...Online: Association for Computational Linguistics, abr. 2021. Disponível em: <https://aclanthology.org/2021.eacl-main.310>

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

GARCIA, M.; GAMALLO, P. **An Entity-Centric Coreference Resolution System for Person Entities with Rich Linguistic Information**. Proceedings of 25th International Conference on Computational Linguistics. **Anais**...Dublin, Ireland: 2014. Disponível em: <http://aclweb.org/anthology/C/C14/C14-1070.pdf>

GARDENT, C. et al. **The WebNLG Challenge: Generating Text from RDF Data**. Proceedings of the 10th International Conference on Natural Language Generation. **Anais**...: INLG'17.Santiago de Compostela, Spain: Association for Computational Linguistics, 2017. Disponível em: <http://aclweb.org/anthology/W17-3518>

GARGETT, A. et al. **The GIVE-2 Corpus of Giving Instructions in Virtual Environments**. Proceedings of LREC-2010. **Anais**...Valletta, Malta: ELRA, 2010.

GATT, A.; BELZ, A. Introducing Shared Tasks to NLG: The TUNA Shared Task Evaluation Challenges. Em: KRAHMER, E.; THEUNE, M. (Eds.). **Empirical Methods in Natural Language Generation**. Berlin, Heidelberg: Springer-Verlag, 2010. p. 264–293.

GATT, A.; KRAHMER, E. Survey of the State of the Art in Natural Language Generation: Core tasks, applications and evaluation. **Journal of Artificial Intelligence Research**, v. 61, n. 1, p. 65–170, 2018.

GATT, A.; SLUIS, I. VAN DER; DEEMTER, K. VAN. **Evaluating algorithms for the generation of referring expressions using a balanced corpus**. Proceedings of ENLG-07. **Anais**...Schloss Dagstuhl, Germany: Association for Computational Linguistics, 2007.

GAUY, M. M.; FINGER, M. **Pretrained audio neural networks for Speech emotion recognition in Portuguese**. Proceedings of the Workshop on Automatic Speech Recognition for Spontaneous and Prepared Speech & Speech Emotion Recognition in Portuguese co-located with 15th edition of the International Conference on the Computational Processing of Portuguese (PROPOR 2022). **Anais**...2022.

GEERAERT, K.; BAAYEN, R. H.; NEWMAN, J. "Spilling the bag" on idiomatic variation. Em: MARKANTONATOU, S. et al. (Eds.). **Multiword expressions at length and in depth: Extended papers from the MWE 2017 workshop**. Berlin: Language Science Press., 2018. p. 1–33.

GEHRMANN, S. et al. The gem benchmark: Natural language generation, its evaluation and metrics. **arXiv preprint arXiv:2102.01672**, 2021.

GEY, F. et al. GeoCLEF 2006: the CLEF 2006 Cross-Language Geographic Information Retrieval Track Overview. Em: PETERS, C. et al. (Eds.). **Evaluation of Multilingual and Multi-modal Information Retrieval - 7th Workshop of the Cross-Language Evaluation Forum, CLEF 2006. Alicante, Spain, September, 2006. Revised Selected papers**. Lecture Notes em Computer Science. Berlin / Heidelberg: Springer, 2007. v. 4730p. 852–876.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

GIAMPICCOLO, D. et al. Overview of the CLEF 2007 Multilingual Question Answering Track. Em: PETERS, C. et al. (Eds.). **Advances in Multilingual and Multimodal Information Retrieval: 8th Workshop of the Cross-Language Evaluation Forum, CLEF 2007, Budapest, Hungary, September 19-21, 2007, Revised Selected Papers**. Lecture Notes em Computer Science. Berlin: Springer, 2008. v. 5152p. 200–236.

GILES, H. Communication Accommodation Theory. **The International Encyclopedia of Communication Theory and Philosophy**, p. 1–7, 2016.

GINZBURG, J. **The Interactive Stance**. [s.l.] Oxford University Press, 2012.

GINZBURG, J.; FERNÁNDEZ, R. M.; SCHLANGEN, D. Disfluencies as intra-utterance dialogue moves. **Semantics and Pragmatics**, v. 7, n. 9, p. 64, 2014.

GOLDBERG, A. **Constructions at Work: The Nature of Generalization in Language**. [s.l.] Oxford University Press, 2005.

GOLDBERG, A. E. Compositionality. Em: RIEMER, N. (Ed.). **The Routledge Handbook of Semantics**. [s.l.] Routledge, 2015.

GOLDBERG, E.; DRIEDGER, N.; KITTREDGE, R. I. Using Natural-Language Processing to Produce Weather Forecasts. **IEEE Expert: Intelligent Systems and Their Applications**, p. 45–53, 1994.

GOLUB, G. H.; REINSCH, C. **Singular Value Decomposition and Least Squares Solutions**. [s.l.] Numer. Math 14, 1970. p. 403–420

GOMES, D. S.; COELHO, O.; MORGADO, C. As implicações da espacialização como categoria analítica da conversa na Língua Brasileira de Sinais e na Língua Gestual Portuguesa. **Sensos-e**, v. 7, n. 3, p. 57–69, 2020.

GONÇALO OLIVEIRA, H. **Beyond the automatic construction of a lexical ontology for Portuguese: resources developed in the scope of Onto.PT**. Proceedings of Workshop on Tools and Resources for Automatically Processing Portuguese and Spanish. **Anais**...: TorPorEsp.São Carlos, SP, Brasil: BDBComp, 2014. Disponível em: <http://www.lbd.dcc.ufmg.br/colecoes/torporesp/2014/004.pdf>

GONÇALO OLIVEIRA, H. et al. **Using Lucene for Developing a Question-Answering Agent in Portuguese**. (R. Rodrigues et al., Eds.)8th Symposium on Languages, Applications and Technologies (SLATE 2019). **Anais**...: Open Access Series em Informatics (OASIcs).Dagstuhl, Germany: Schloss Dagstuhl – Leibniz-Zentrum für Informatik, 2019. Disponível em: <https://drops.dagstuhl.de/entities/document/10.4230/OASIcs.SLATE.2019.2>

GONÇALO OLIVEIRA, H. et al. **A Brief Survey of Textual Dialogue Corpora**. Proceedings of the Thirteenth Language Resources and Evaluation Conference. **Anais**...Marseille, France: European Language Resources Association, jun. 2022.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Disponível em: <https://aclanthology.org/2022.lrec-1.135>

GONÇALO OLIVEIRA, H.; GOMES, P. ECO and Onto-PT: a flexible approach for creating a Portuguese Wordnet automatically. **Language Resources and Evaluation**, v. 48, n. 2, p. 373–393, 2014.

GONÇALVES, M. et al. Avaliação de recursos computacionais para o português. **Linguamática**, v. 12, n. 2, p. 51–68, 2020.

GONÇALVES, S. C. L. Projeto ALIP (Amostra Linguística do Interior Paulista) e banco de dados Iboruna: 10 anos de contribuição com a descrição do português brasileiro. **Estudos Linguísticos (São Paulo. 1978)**, v. 48, n. 1, p. 276–297, 2019.

GOODFELLOW, I.; BENGIO, Y.; COURVILLE, A. **Deep Learning**. [s.l.] MIT Press, 2016. v. 1

GOODING, S.; TASLIMIPOOR, S.; KOCHMAR, E. **Incorporating Multiword Expressions in Phrase Complexity Estimation**. Proceedings of the 1st Workshop on Tools and Resources to Empower People with REAding DIfficulties (READI). **Anais**...Marseille, France: European Language Resources Association, 2020. Disponível em: <https://aclanthology.org/2020.readi-1.3>

GORMAN, K.; BEDRICK, S. **We Need to Talk about Standard Splits**. (A. Korhonen, D. Traum, L. Màrquez, Eds.)Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. **Anais**...Florence, Italy: Association for Computational Linguistics, jul. 2019. Disponível em: <https://aclanthology.org/P19-1267/>

GRALIŃSKI, F. et al. **Computational Lexicography of Multi-Word Units. How Efficient Can It Be?** Proceedings of the 2010 Workshop on Multiword Expressions: from Theory to Applications. **Anais**...Beijing, China: Coling 2010 Organizing Committee, ago. 2010. Disponível em: <https://aclanthology.org/W10-3702>

GREEN, S.; MARNEFFE, M.-C. DE; MANNING, C. D. Parsing Models for Identifying Multiword Expressions. **Computational Linguistics**, v. 39, n. 1, p. 195–227, mar. 2013.

GRÉGOIRE, N. DuELME: a Dutch electronic lexicon of multiword expressions. **Language Resources and Evaluation**, v. 44, p. 23–39, 2010.

GREGOROMICHELAKI, E. et al. Incrementality and intention-recognition in utterance processing. **Dialogue & Discourse**, v. 2, n. 1, p. 199–233, 2011.

GRIES, S. C. **Estatística com R para a Linguística**. [s.l.] FALE/ UFMG, 2019.

GRIS, L. R. S. et al. **Bringing NURC/SP to digital life: the role of open-source automatic speech recognition models**. Anais do XIX Encontro Nacional de Inteligência Artificial e Computacional. **Anais**...Porto Alegre, RS, Brasil: SBC, 2022. Disponível em: <https://sol.sbc.org.br/index.php/eniac/article/view/22793>

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

GRIS, L. R. S. et al. **Evaluating OpenAI's Whisper ASR for Punctuation Prediction and Topic Modeling of life histories of the Museum of the Person.**, 2023. Disponível em: <https://arxiv.org/abs/2305.14580>

GROSS, M. **Lexicon - Grammar The Representation of Compound Words**. Coling 1986 Volume 1: The 11th International Conference on Computational Linguistics. **Anais**...1986. Disponível em: <https://aclanthology.org/C86-1001>

GROSZ, B. J.; JOSHI, A. K.; WEINSTEIN, S. Centering: A framework for modelling the local coherence of discourse. **IRCS Technical Reports Series**, 1995.

GROSZ, B. J.; SIDNER, C. L. Attention, intentions, and the structure of discourse. **Computational linguistics**, v. 12, n. 3, p. 175–204, 1986.

GROUP, E. E. W. et al. **EAGLES Evaluation of Natural Language Processing Systems - Final Report**. ISSCO, 1996. Disponível em: <https://www.issco.unige.ch/en/research/projects/eagles/index.html>

GRUBER, T. R. **Siri, A Virtual Personal Assistant-Bringing Intelligence to the Interface**. Semantic Technologies Conference. **Anais**...2009.

GUERINO, G.; VALENTIM, N. **"Is anybody there?": Exploring the use and difficulties of Brazilians with Conversational Systems**. Anais do XIX Simpósio Brasileiro sobre Fatores Humanos em Sistemas Computacionais. **Anais**...Porto Alegre, RS, Brasil: SBC, 2020. Disponível em: <https://sol.sbc.org.br/index.php/ihc/article/view/13835>

GULATI, A. et al. Conformer: Convolution-augmented Transformer for Speech Recognition. **CoRR**, v. abs/2005.08100, 2020.

GUO, Q. et al. **P2: A Plan-and-Pretrain Approach for Knowledge Graph-to-Text Generation**. Proceedings of the 3rd International Workshop on Natural Language Generation from the Semantic Web (WebNLG+). **Anais**...2020.

GURURANGAN, S. et al. **Annotation Artifacts in Natural Language Inference Data**. Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers). **Anais**...New Orleans, Louisiana: Association for Computational Linguistics, jun. 2018. Disponível em: <https://aclanthology.org/N18-2017>

GURURANGAN, S. et al. **Don't Stop Pretraining: Adapt Language Models to Domains and Tasks**. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. **Anais**...Online: Association for Computational Linguistics, jul. 2020. Disponível em: <https://aclanthology.org/2020.acl-main.740>

HAGÈGE, C.; BAPTISTA, J.; MAMEDE, N. **Portuguese Temporal Expressions Recognition: from TE characterization to an effective TER module implementation**. The 7th Brazilian Symposium in Information and Human Language Technology

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

(STIL 2009). **Anais**...São Carlos, Brasil: 2009. Disponível em: <http://www.nilc.icmc.usp.br/til/stil2009_English/Proceedings/stil/Hagege-57697_1.pdf>

HAGEMEIJER, T. et al. **The PALMA Corpora of African Varieties of Portuguese**. Proceedings of the Thirteenth Language Resources and Evaluation Conference. **Anais**...Marseille, France: European Language Resources Association, jun. 2022. Disponível em: <https://aclanthology.org/2022.lrec-1.539>

HALL, J. **A Probabilistic Part-of-Speech Tagger with Suffix Probabilities**. tese de doutorado—[s.l: s.n.].

HALLIDAY, M. A. K.; MATTHIESSEN, C. M. I. M. **Construing Experience Through Meaning: A Language Based Approach to Cognition**. [s.l.] Continuum, 1999.

HAPKE, H.; HOWARD, C.; LANE, H. **Natural Language Processing in Action: Understanding, analyzing, and generating text with Python**. [s.l.] Manning, 2019.

HARMAN, D. The Text Retrieval Conferences (TRECs): Providing a Test-Bed for Information Retrieval Systems. **Bulletin of the American Society for Information Science**, v. 24, n. 4, p. 11–13, 1998.

HARRIS, Z. S. Distributional Structure. **Word**, v. 10, n. 2-3, p. 146–162, 1954.

HAUSSER, R. The coordinator's final report on the first Morpholympics. Em: HAUSSER, R. (Ed.). **Linguistische Verifikation: Dokumentation zur Ersten Morpholympics 1994**. [s.l.] Max Niemeyer Verlag, 1996. p. 167–181.

HAVASI, C.; SPEER, R.; ALONSO, J. **ConceptNet 3: a Flexible, Multilingual Semantic Network for Common Sense Knowledge**. Recent Advances in Natural Language Processing. **Anais**...Borovets, Bulgaria: To appear, 2007.

HAVIV, A. et al. **Understanding Transformer Memorization Recall Through Idioms**. Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics. **Anais**...Dubrovnik, Croatia: Association for Computational Linguistics, 2023. Disponível em: <https://aclanthology.org/2023.eacl-main.19>

HAYES, P. **Expanding the Horizons of Natural Language Interfaces**. 18th Annual Meeting of the Association for Computational Linguistics. **Anais**...Philadelphia, Pennsylvania, USA: Association for Computational Linguistics, jun. 1980. Disponível em: <https://aclanthology.org/P80-1019>

HAYES, P. J.; REDDY, D. R. Steps toward graceful interaction in spoken and written man-machine communication. **International Journal of Man-Machine Studies**, v. 19, n. 3, p. 231–284, 1983.

HEALEY, P. G.; MILLS, G. J. **A Dialogue Experimentation Toolkit**. Proceedings of the Annual Meeting of the Cognitive Science Society. **Anais**...2009. Disponível em:

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

<https://dialoguetoolkit.github.io/chattool/>

HEEMAN, P. A. **Dialogue transcription tools - TRAINS Technical Note 94-1**. [s.l.] University of Rochester, 1995. Disponível em: <https://dl.acm.org/doi/abs/10.5555/898276>.

HEEMAN, P. A.; HIRST, G. Collaborating on Referring Expressions. **Computational Linguistics**, v. 21, n. 3, p. 351–382, 1995.

HEIKKILÄ, M. **Why you shouldn't trust AI search engines**. Disponível em: <https://www.technologyreview.com/2023/02/14/1068498/why-you-shouldnt-trust-ai-search-engines/>. Acesso em: 9 abr. 2023.

HEIKKILÄ, M. **The viral AI avatar app Lensa undressed me—without my consent**. Disponível em: <https://www.technologyreview.com/2022/12/12/1064751/the-viral-ai-avatar-app-lensa-undressed-me-without-my-consent/>. Acesso em: 28 ago. 2023.

HEIM, I. File Change Semantics and the Familiarity Theory of Definiteness. Em: **Formal Semantics**. [s.l.] Wiley-Blackwell, 2008. p. 223–248.

HELDNER, M.; EDLUND, J. Pauses, gaps and overlaps in conversations. **Journal of Phonetics**, v. 38, n. 4, p. 555–568, 2010.

HENDERSON, P. et al. **Ethical challenges in data-driven dialogue systems**. Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society. **Anais**...2018. Disponível em: <https://doi.org/10.1145/3278721.3278777>

HENDRICKX, I. et al. **SemEval-2010 Task 8: Multi-Way Classification of Semantic Relations between Pairs of Nominals**. Proceedings of the 5th International Workshop on Semantic Evaluation. **Anais**...2010. Disponível em: <http://www.aclweb.org/anthology/S10-1006>

HENDRYCKS, D. et al. **Measuring Massive Multitask Language Understanding**. Proceedings of the International Conference on Learning Representations (ICLR). **Anais**...a2021.

HENDRYCKS, D. et al. **Aligning AI With Shared Human Values**. Proceedings of the International Conference on Learning Representations (ICLR). **Anais**...b2021.

HICKE, Y. et al. Assessing the efficacy of large language models in generating accurate teacher responses. **arXiv preprint arXiv:2307.04274**, 2023.

HILGERT, J. G. A construção do sentido e da compreensão na conversa, mostrada em procedimentos meta-enunciativos. **Linha D'Água**, v. 25, n. 2, p. 107–129, 2012.

HILGERT, J. G. A emergência da compreensão na conversa, mostrada no trabalho colaborativo de otimização de enunciados. **Todas as Letras-Revista de Língua e Literatura**,

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

v. 16, n. 1, 2014.

HIRSCHMAN, L. The evolution of Evaluation: Lessons from the Message Understanding Conferences. **Computer Speech and Language**, v. 12, n. 4, p. 281–305, 1998.

HIRSCHMAN, L.; THOMPSON, H. S. Overview of Evaluation in Speech and Natural Language Processing. Em: **Survey of the State of the Art in Human Language Technology**. USA: Cambridge University Press, 1997. p. 409–414.

HORA, N. DA. **Coded Bias: linguagem acessível para entender vieses em algoritmos**. Disponível em: <https://mittechreview.com.br/coded-bias-linguagem-acessivel-para-entender-vieses-em-algoritmos/>. Acesso em: 7 abr. 2023.

HORA, N. DA. **Ética em IA: a pergunta que não estamos fazendo**. Disponível em: <https://mittechreview.com.br/etica-em-ia-a-pergunta-que-nao-estamos-fazendo/>. Acesso em: 7 abr. 2023.

HORSMANN, T.; ZESCH, T. **Assigning Fine-grained PoS Tags based on High-precision Coarse-grained Tagging**. Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers. **Anais**...Osaka, Japan: The COLING 2016 Organizing Committee, dez. 2016. Disponível em: <https://aclanthology.org/C16-1032>

HOU, Y.; MARKERT, K.; STRUBE, M. **A Rule-Based System for Unrestricted Bridging Resolution: Recognizing Bridging Anaphora and Finding Links to Antecedents**. Proceedings of the Conference on Empirical Methods in Natural Language Processing. **Anais**...Doha, Qatar: 2014. Disponível em: <http://aclweb.org/anthology/D/D14/D14-1222.pdf>

HSU, W.-N. et al. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. **IEEE/ACM Transactions on Audio, Speech, and Language Processing**, v. 29, p. 3451–3460, 2021.

HUANG, J.-T.; HASEGAWA-JOHNSON, M.; SHIH, C. **Unsupervised prosodic break detection in Mandarin speech**. Proc. Speech Prosody 2008. **Anais**...2008.

HUANG, X.; ACERO, A.; HON, H. W. **Spoken Language Processing: A Guide to Theory, Algorithm, and System Development**. [s.l.] Prentice Hall PTR, 2001.

HWANG, A.; HIDEY, C. **Confirming the Non-compositionality of Idioms for Sentiment Analysis**. Proceedings of the Joint Workshop on Multiword Expressions and WordNet (MWE-WN 2019). **Anais**...Florence, Italy: Association for Computational Linguistics, ago. 2019. Disponível em: <https://aclanthology.org/W19-5114>

ILARI, R.; GERALDI, J. W. **Semântica**. [s.l.] Ética, 1985.

INFOBASE. **Inteligência Artificial e a perpetuação do racismo**. Disponível em: <https://infobase.com.br/inteligencia-artificial-e-a-perpetuacao-do-racismo/>. Acesso

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

em: 28 ago. 2023.

ISCEN, A. et al. **Label Propagation for Deep Semi-Supervised Learning**. jun. 2019.

ITO, K. **The LJ speech dataset**. https://keithito.com/LJ-Speech-Dataset/, 2017.

JACINTHO, F.; PENHA, A. Interfaces conversacionais: Análise de tarefas para Siri e Google Now. **Ergodesign & HCI**, v. 4, n. 2, p. 72–81, 2016.

JAFARLOU, M.; KUBEK, M. M. **Reducing Labeling Costs in Sentiment Analysis via Semi-Supervised Learning**. Proceedings of the 2024 8th International Conference on Natural Language Processing and Information Retrieval. **Anais**...: NLPIR '24.New York, NY, USA: Association for Computing Machinery, 2025. Disponível em: <https://doi.org/10.1145/3711542.3711570>

JEON, J. H.; LIU, Y. **Semi-supervised Learning for Automatic Prosodic Event Detection Using Co-training Algorithm**. Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP. **Anais**...Suntec, Singapore: Association for Computational Linguistics, ago. 2009. Disponível em: <https://aclanthology.org/P09-1061>

JI, Z. et al. Survey of Hallucination in Natural Language Generation. **ACM Comput. Surv.**, v. 55, n. 12, mar. 2023.

JIANG, R.; BANCHS, R. E.; LI, H. **Evaluating and Combining Named Entity Recognition Systems**. Proceedings of the Sixth Named Entity Workshop, joint with 54th ACL. **Anais**...2016. Disponível em: <https://www.aclweb.org/anthology/W16-2703.pdf>

JOHNSON, K. **Acoustic and Auditory Phonetics**. [s.l.] Wiley, 2011.

JOOS, M. Description of language design. **Journal of Acoustical Society of America - JASA**, v. 22, p. 701–708, 1950.

JOSÉ, M. M. et al. **Integrating Question Answering and Text-to-SQL in Portuguese**. (V. Pinheiro et al., Eds.)Computational Processing of the Portuguese Language. **Anais**...Cham: Springer International Publishing, 2022.

JOSHI, M. et al. **BERT for Coreference Resolution: Baselines and Analysis**. Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). **Anais**...Hong Kong, China: Association for Computational Linguistics, nov. 2019. Disponível em: <https://aclanthology.org/D19-1588>

JURAFSKY, D.; MARTIN, J. H. **Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech**

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Recognition. 2. ed. [s.l.] Prentice Hall, 2008. p. 988

JURAFSKY, D.; MARTIN, J. H. **Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition**. 3rd. ed. USA: Prentice Hall PTR, 2023.

KAHANE, S.; COURTIN, M.; GERDES, K. **Multi-word annotation in syntactic treebanks - Propositions for Universal Dependencies**. Proceedings of the 16th International Workshop on Treebanks and Linguistic Theories. **Anais**...Prague, Czech Republic: 2017. Disponível em: <https://aclanthology.org/W17-7622>

KANDO, N. NTCIR and Its Background – Evaluation Workshop on Information Access Technologies and Test Collections. **Journal of the Japanese Society for Artificial Intelligence**, v. 17, n. 3, p. 296–300, 2002.

KANITZ, A.; FRANK, I. Aprendizagem enquanto produção conjunta de conhecimento: avançando tarefas e alcançando entendimentos satisfatórios na fala-em-interação. **Revista Brasileira de Linguística Aplicada**, v. 14, p. 111–140, 2014.

KANITZ, A.; LUZ, R. L. Letramento multimodal e construção conjunta de conhecimento na fala-em-interação. **Revista Brasileira de Linguística Aplicada**, v. 19, p. 603–633, 2019.

KANTAYYA, S. **Coded Bias**. Disponível em: <https://www.codedbias.com>. Acesso em: 7 abr. 2023.

KARPAS, E. et al. **MRKL Systems: A modular, neuro-symbolic architecture that combines large language models, external knowledge sources and discrete reasoning**., 2022. Disponível em: <https://arxiv.org/abs/2205.00445>

KATO, A.; SHINDO, H.; MATSUMOTO, Y. **Construction of an English Dependency Corpus incorporating Compound Function Words**. Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16). **Anais**...Portorož, Slovenia: European Language Resources Association (ELRA), 2016. Disponível em: <https://aclanthology.org/L16-1263>

KENEDY, E.; OTHERO, G. DE Á. **Para conhecer sintaxe**. São Paulo: Contexto, 2018.

KILGARRIFF, A. I Don't Believe in Word Senses. **Computers and the Humanities**, 1997.

KILGARRIFF, A. **Thesauruses for Natural Language Processing**. Proceedings of Natural Language Processing and Knowledge Engineering. **Anais**...2003. Disponível em: <https://www.kilgarriff.co.uk/Publications/2003-K-Beijing-thes4NLP.pdf>

KIM, J. et al. Glow-TTS: A Generative Flow for Text-to-Speech via Monotonic Alignment Search. **arXiv preprint arXiv:2005.11129**, 2020.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

KIM, J.; KONG, J.; SON, J. **Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech**. International Conference on Machine Learning. **Anais**...PMLR, 2021.

KIM, S. N. et al. **SemEval-2010 Task 5 : Automatic Keyphrase Extraction from Scientific Articles**. Proceedings of the 5th International Workshop on Semantic Evaluation. **Anais**...Uppsala, Sweden: Association for Computational Linguistics, jul. 2010. Disponível em: <https://aclanthology.org/S10-1004>

KING, M. Evaluating Natural Language Processing Systems. **Communications of the ACM**, v. 39, n. 1, p. 73–79, jan. 1996.

KIPPER, K.; DANG, H. T.; PALMER, M. **Class-Based Construction of a Verb Lexicon**. Proceedings of the Seventeenth National Conference on Artificial Intelligence and Twelfth Conference on Innovative Applications of Artificial Intelligence. **Anais**...AAAI Press, 2000.

KIRK, H. et al. **Handling and Presenting Harmful Text in NLP Research**. Findings of the Association for Computational Linguistics: EMNLP 2022. **Anais**...Abu Dhabi, United Arab Emirates: Association for Computational Linguistics, dez. 2022. Disponível em: <https://aclanthology.org/2022.findings-emnlp.35>

KIRSTAIN, Y.; RAM, O.; LEVY, O. **Coreference Resolution without Span Representations**. Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers). **Anais**...2021.

KITZINGER, C. Repair. Em: **The handbook of conversation analysis**. [s.l.] Wiley Online Library, 2012. p. 229–256.

KLATT, D. H. Software for a cascade/parallel formant synthesizer. **the Journal of the Acoustical Society of America**, v. 67, n. 3, p. 971–995, 1980.

KLIE, J.-C. et al. **The INCEpTION Platform: Machine-Assisted and Knowledge-Oriented Interactive Annotation**. Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations. **Anais**...Santa Fe, USA: Association for Computational Linguistics, 2018. Disponível em: <http://tubiblio.ulb.tu-darmstadt.de/106270/>

KOCH, I. G. V. **O texto e a construção do sentido**. 7. ed. Campinas, SP: Contexto, 2003.

KOCH, I. G. V. Digressão e Relevância Conversacional. **Cadernos de Estudos Linguísticos**, v. 37, p. 81–91, 2012.

KOCH, I. G. V.; TRAVAGLIA, L. **Texto e coerência**. 13. ed. [s.l.] Cortez, 2012.

KÖHN, A. **Incremental Natural Language Processing: Challenges, Strategies,**

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

and Evaluation. Proceedings of the 27th International Conference on Computational Linguistics. **Anais**...Santa Fe, New Mexico, USA: Association for Computational Linguistics, ago. 2018. Disponível em: <https://aclanthology.org/C18-1253>

KOIZUMI, Y. et al. Miipher: A Robust Speech Restoration Model Integrating Self-Supervised Speech and Text Representations. **arXiv preprint arXiv:2303.01664**, b2023.

KOIZUMI, Y. et al. LibriTTS-R: A Restored Multi-Speaker Text-to-Speech Corpus. **arXiv preprint arXiv:2305.18802**, a2023.

KONRAD, P. G. **A busca vs. o resguardo de informações acerca dos crimes em interrogatórios policiais: um olhar sob a perspectiva da fala-em-interação**. mathesis—[s.l.] Universidade do Vale do Rio dos Sinos, 2018.

KOPPATZ, M. et al. Automatic generation of factual news headlines in finnish. **arXiv preprint arXiv:2212.02170**, 2022.

KORKONTZELOS, I. **Unsupervised Learning of Multiword Expressions**. tese de doutorado—York, UK: University of York, 2011.

KORSKO, P. **The narrative shape of two-party complaints in Portuguese: A discourse analytic study**. tese de doutorado—[s.l.] Teachers College, Columbia University, 2004.

KOWAL, S.; O'CONNELL, D. C. Transcription as a crucial step of data analysis. **The SAGE handbook of qualitative data analysis**, p. 64–79, 2014.

KRAHMER, E.; DEEMTER, K. VAN. Computational generation of referring expressions: A survey. **Computational Linguistics**, v. 38, n. 1, p. 173–218, 2012.

KRAHMER, E.; ERK, S. VAN; VERLEG, A. Graph-Based Generation of Referring Expressions. **Computational Linguistics**, v. 29, n. 1, p. 53–72, 2003.

KRAHMER, E.; THEUNE, M. Efficient context-sensitive generation of referring expressions. Em: DEEMTER, K. VAN; KIBBLE, R. (Eds.). **Information sharing: Reference and presupposition in language generation and interpretation**. Stanford, CA: CSLI, 2002. p. 223–264.

KRIPPENDORFF, K. Estimating the Reliability, Systematic Error and Random Error of Interval Data. **Educational and Psychological Measurement**, v. 30, n. 1, p. 61–70, 1970.

KRUSE, J. S.; BARBOSA, P. A. Alinha-PB: a phonetic aligner for Brazilian Portuguese. **Journal of Communication and Information Systems**, v. 36, n. 1, p. 192–199, dez. 2021.

KUDO, T. **Subword Regularization: Improving Neural Network Translation**

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Models with Multiple Subword Candidates. Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). **Anais**...Melbourne, Australia: Association for Computational Linguistics, jul. 2018. Disponível em: <https://aclanthology.org/P18-1007>

KUKICH, K. **Design of a Knowledge-based Report Generator**. Proceedings of the 21st Annual Meeting on Association for Computational Linguistics. **Anais**...: ACL'83.Cambridge, Massachusetts: Association for Computational Linguistics, 1983. Disponível em: <https://doi.org/10.3115/981311.981340>

KUMAWAT, D.; JAIN, V. POS Tagging Approaches: A Comparison. **International Journal of Computer Applications**, v. 118, n. 6, p. 32–38, maio 2015.

KUO, Y. et al. **Community-Based Game Design: Experiments on Social Games for Commonsense Data Collection**. Proceedings of the ACM SIGKDD Workshop on Human Computation. **Anais**...: HCOMP '09.New York, NY, USA: Association for Computing Machinery, 2009. Disponível em: <https://doi.org/10.1145/1600150.1600154>

KYLE, K. K. J. F. S.; JOSE, K. A. C. Y. B.; SOTELO, S. M. **Char2wav: End-to-end speech synthesis**. International Conference on Learning Representations, workshop. **Anais**...2017.

LACERDA, A. R. T. DE; AGUIAR, C. S. R. **FLOSS FAQ Chatbot Project Reuse: How to Allow Nonexperts to Develop a Chatbot**. Proceedings of the 15th International Symposium on Open Collaboration. **Anais**...: OpenSym '19.New York, NY, USA: Association for Computing Machinery, 2019. Disponível em: <https://doi.org/10.1145/3306446.3340823>

LAKHOTIA, K. et al. On Generative Spoken Language Modeling from Raw Audio. **Transactions of the Association for Computational Linguistics**, v. 9, p. 1336–1354, 2021.

LARSSON, S. **User-initiated Sub-dialogues in State-of-the-art Dialogue Systems**. Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue. **Anais**...Saarbrücken, Germany: Association for Computational Linguistics, ago. 2017. Disponível em: <https://aclanthology.org/W17-5503>

LAVIE, A.; AGARWAL, A. **Meteor: An Automatic Metric for MT Evaluation with High Levels of Correlation with Human Judgments**. Proceedings of the Second Workshop on Statistical Machine Translation. **Anais**...: StatMT'07.Prague, Czech Republic: 2007. Disponível em: <http://dl.acm.org/citation.cfm?id=1626355.1626389>

LEBRET, R.; GRANGIER, D.; AULI, M. **Neural Text Generation from Structured Data with Application to the Biography Domain**. Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing. **Anais**...: EMNLP'16.Austin, Texas: Association for Computational Linguistics, 2016. Disponível em: <http://aclanthology.coli.uni-saarland.de/pdf/D/D16/D16-1128.pdf>

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

LEE, C. VAN DER et al. Human evaluation of automatically generated text: Current trends and best practice guidelines. **Computer Speech & Language**, v. 67, p. 101151, 2021.

LEE, C. VAN DER; KRAHMER, E.; WUBBEN, S. **PASS: A Dutch data-to-text system for soccer, targeted towards specific audiences**. Proceedings of INLG-2017. **Anais**...Santiago de Compostela, Spain: Association for Computational Linguistics, a2017. Disponível em: <http://aclweb.org/anthology/W17-3513>

LEE, H. et al. **Stanford's multi-pass sieve coreference resolution system at the CoNLL-2011 shared task**. Proceedings of the Fifteenth Conference on Computational Natural Language Learning: Shared Task. **Anais**...2011.

LEE, H. et al. Deterministic coreference resolution based on entity-centric, precision-ranked rules. **Computational Linguistics**, v. 39, n. 4, p. 885–916, 2013.

LEE, K. et al. End-to-end neural coreference resolution. **arXiv preprint arXiv:1707.07045**, b2017.

LEIDNER, J. L.; PLACHOURAS, V. **Ethical by Design: Ethics Best Practices for Natural Language Processing**. Proceedings of the First ACL Workshop on Ethics in Natural Language Processing. **Anais**...Valencia, Spain: Association for Computational Linguistics, abr. 2017. Disponível em: <https://aclanthology.org/W17-1604>

LENAT, D. B.; GUHA, R. V. **Building large knowledge-based systems: representation and inference in the Cyc project**. [s.l.] Addison-Wesley, 1989.

LEONARDELLI, E. et al. **Agreeing to Disagree: Annotating Offensive Language Datasets with Annotators' Disagreement**. (M.-F. Moens et al., Eds.)Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. **Anais**...Online; Punta Cana, Dominican Republic: Association for Computational Linguistics, nov. 2021. Disponível em: <https://aclanthology.org/2021.emnlp-main.822/>

LEVELT, W. J. **Speaking: From intention to articulation**. [s.l.] MIT press, 1993.

LEVINSON, S. C. Speech Acts. Em: **The Oxford Handbook of Pragmatics**. [s.l.] Oxford University Press, 2017.

LEWIS, D. Scorekeeping in a language game. Em: **Semantics from different points of view**. [s.l.] Springer, 1979. p. 172–187.

LEWIS, M. et al. **BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension**. (D. Jurafsky et al., Eds.)Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020. **Anais**...Association for Computational Linguistics, 2020. Disponível em: <https://doi.org/10.18653/v1/2020.acl-main.703>

LGPD. **Lei Geral de Proteção de Dados Pessoais (LGPD)**. Disponível em: <https:

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

//www.planalto.gov.br/ccivil_03/_ato2015-2018/2018/lei/l13709.htm>. Acesso em: 9 abr. 2023.

LI, J. et al. Molweni: A challenge multiparty dialogues-based machine reading comprehension dataset with discourse structure. **arXiv preprint arXiv:2004.05080**, 2020.

LI, S. et al. **Defining a New NLP Playground**. (H. Bouamor, J. Pino, K. Bali, Eds.)Findings of the Association for Computational Linguistics: EMNLP 2023. **Anais**...Singapore: Association for Computational Linguistics, dez. 2023. Disponível em: <https://aclanthology.org/2023.findings-emnlp.799>

LI, X.; ROTH, D. **Learning question classifiers**. COLING 2002: The 19th International Conference on Computational Linguistics. **Anais**...2002.

LIESENFELD, A.; LOPEZ, A.; DINGEMANSE, M. **The timing bottleneck: Why timing and overlap are mission-critical for conversational user interfaces, speech recognition and dialogue systems**. Proceedings of the 24th Meeting of the Special Interest Group on Discourse and Dialogue. **Anais**...Prague, Czechia: Association for Computational Linguistics, set. 2023. Disponível em: <https://aclanthology.org/2023.sigdial-1.45>

LIMA, R. et al. **A Large Dataset of Spontaneous Speech with the Accent Spoken in São Paulo for Automatic Speech Recognition Evaluation**. Anais da XXXIV Brazilian Conference on Intelligent Systems. **Anais**...Porto Alegre, RS, Brasil: SBC, 2024. Disponível em: <https://sol.sbc.org.br/index.php/bracis/article/view/33550>

LIN, C.-H. et al. **Rich prosodic information exploration on spontaneous Mandarin speech**. 2016 10th International Symposium on Chinese Spoken Language Processing (ISCSLP). **Anais**...Tianjin: 2016.

LIN, C.-H. et al. Hierarchical prosody modeling for Mandarin spontaneous speech. **The Journal of the Acoustical Society of America**, v. 145, n. 4, p. 2576–2596, 2019.

LIN, C.-Y. **ROUGE: A Package for Automatic Evaluation of Summaries**. Text Summarization Branches Out. **Anais**...Barcelona, Spain: Association for Computational Linguistics, jul. 2004. Disponível em: <https://aclanthology.org/W04-1013>

LIN, D. **Automatic Identification of Non-compositional Phrases**. Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics. **Anais**...College Park, Maryland, USA: Association for Computational Linguistics, jun. 1999. Disponível em: <https://aclanthology.org/P99-1041>

LINARDATOS, P.; PAPASTEFANOPOULOS, V.; KOTSIANTIS, S. Explainable AI: A Review of Machine Learning Interpretability Methods. **Entropy**, v. 23, n. 1, 2021.

LIPTON, Z. C.; STEINHARDT, J. Troubling Trends in Machine Learning Scholarship: Some ML papers suffer from flaws that could mislead the public and stymie future research. **Queue**, v. 17, n. 1, p. 45–77, 2019.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

LITMAN, D. J.; ALLEN, J. F. A plan recognition model for subdialogues in conversations. **Cognitive science**, v. 11, n. 2, p. 163–200, 1987.

LIU, C.-W. et al. **How NOT To Evaluate Your Dialogue System: An Empirical Study of Unsupervised Evaluation Metrics for Dialogue Response Generation**. Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing. **Anais**...Austin, Texas: Association for Computational Linguistics, nov. 2016. Disponível em: <https://aclanthology.org/D16-1230>

LIU, H.; SINGH, P. **Commonsense Reasoning in and Over Natural Language**. (M. Gh. Negoita, R. J. Howlett, L. C. Jain, Eds.)Knowledge-Based Intelligent Information and Engineering Systems. **Anais**...Berlin, Heidelberg: Springer Berlin Heidelberg, 2004.

LIU, N. F. et al. **Lexical Semantic Recognition**. Proceedings of the 17th Workshop on Multiword Expressions (MWE 2021). **Anais**...Online: Association for Computational Linguistics, 2021. Disponível em: <https://aclanthology.org/2021.mwe-1.6>

LODER, L. L.; GONZALEZ, P. C.; GARCEZ, P. M. Reparo em terceira posição e intersubjetividade na fala-em-interação em português brasileiro. **Veredas-Revista de Estudos Linguísticos**, v. 6, n. 2, 2002.

LOPE, J.; GRAÑA, M. An ongoing review of speech emotion recognition. **Neurocomputing**, 2023.

LOPES, L. et al. **PortiLexicon-UD: a Portuguese Lexical Resource according to Universal Dependencies Model**. Proceedings of the Language Resources and Evaluation Conference. **Anais**...Marseille, France: European Language Resources Association, jun. 2022. Disponível em: <https://aclanthology.org/2022.lrec-1.715>

LOPES, L. et al. **Disambiguation of Universal Dependencies Part-of-Speech Tags of Closed Class Words in Portuguese**. (A. Britto, K. V. Delgado, Eds.)Proceedings of the 12th Brazilian Conference on Intelligent Systems (BRACIS). **Anais**...2023.

LOPES, L.; PARDO, T. **Towards Portparser - a highly accurate parsing system for Brazilian Portuguese following the Universal Dependencies framework**. (P. Gamallo et al., Eds.)Proceedings of the 16th International Conference on Computational Processing of Portuguese - Vol. 1. **Anais**...Santiago de Compostela, Galicia/Spain: Association for Computational Lingustics, mar. 2024. Disponível em: <https://aclanthology.org/2024.propor-1.41>

LOSNEGAARD, G. S. et al. **PARSEME Survey on MWE Resources**. (N. C. (Conference Chair) et al., Eds.)Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016). **Anais**...Paris, France: European Language Resources Association (ELRA), 2016.

LUCY, L.; BAMMAN, D. **Gender and Representation Bias in GPT-3 Generated Stories**. Proceedings of the Third Workshop on Narrative Understanding. **Anais**...Virtual: Association for Computational Linguistics, jun. 2021. Disponível em:

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

<https://aclanthology.org/2021.nuse-1.5>

LUDUSAN, B.; SYNNAEVE, G.; DUPOUX, E. **Prosodic boundary information helps unsupervised word segmentation**. Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. **Anais**...2015.

LUO, X. **On Coreference Resolution Performance Metrics**. Proceedings of the Conference on Empirical Methods in Natural Language Processing. **Anais**...Vancouver, Canada: 2005.

LYONS, J. **Semantics: Volume 2**. [s.l.] Cambridge university press, 1977. v. 2

MADUREIRA, B. **Flamingos and Hedgehogs in the Croquet-Ground: Teaching Evaluation of NLP Systems for Undergraduate Students**. Proceedings of the Fifth Workshop on Teaching NLP. **Anais**...Online: Association for Computational Linguistics, jun. 2021. Disponível em: <https://aclanthology.org/2021.teachingnlp-1.14>

MADUREIRA, B.; ÇELIKKOL, P.; SCHLANGEN, D. **Revising with a Backward Glance: Regressions and Skips during Reading as Cognitive Signals for Revision Policies in Incremental Processing**. (J. Jiang, D. Reitter, S. Deng, Eds.)Proceedings of the 27th Conference on Computational Natural Language Learning (CoNLL). **Anais**...Singapore: Association for Computational Linguistics, dez. a2023. Disponível em: <https://aclanthology.org/2023.conll-1.22>

MADUREIRA, B.; KAHARDIPRAJA, P.; SCHLANGEN, D. **The Road to Quality is Paved with Good Revisions: A Detailed Evaluation Methodology for Revision Policies in Incremental Sequence Labelling**. Proceedings of the 24th Meeting of the Special Interest Group on Discourse and Dialogue. **Anais**...Prague, Czechia: Association for Computational Linguistics, set. b2023. Disponível em: <https://aclanthology.org/2023.sigdial-1.14>

MADUREIRA, B.; LASOTA, L. Das Inquietudes em Tecnologias de Linguagem. Em: **Novas Tecnologias**. [s.l.] Editora Casa do Direito, 2023.

MADUREIRA, B.; SCHLANGEN, D. **Incremental Processing in the Age of Non-Incremental Encoders: An Empirical Assessment of Bidirectional Models for Incremental NLU**. Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). **Anais**...Online: Association for Computational Linguistics, nov. 2020. Disponível em: <https://aclanthology.org/2020.emnlp-main.26>

MAGNINI, B. et al. Overview of the CLEF 2006 Multilingual Question Answering Track. Em: PETERS, C. et al. (Eds.). **Evaluation of Multilingual and Multi-modal Information Retrieval - 7th Workshop of the Cross-Language Evaluation Forum, CLEF 2006. Alicante, Spain, September, 2006. Revised Selected papers**. Lecture Notes em Computer Science. Berlin / Heidelberg: Springer, 2007. v. 4730p. 223–256.

MAHAJAN, K.; SHAIKH, S. **On the Need for Thoughtful Data Collection for**

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Multi-Party Dialogue: A Survey of Available Corpora and Collection Methods. Proceedings of the 22nd Annual Meeting of the Special Interest Group on Discourse and Dialogue. **Anais**...Singapore; Online: Association for Computational Linguistics, jul. 2021. Disponível em: <https://aclanthology.org/2021.sigdial-1.36>

MALINGAN, N. **Attention Mechanism in Deep Learning**., 2024. Disponível em: <https://www.scaler.com/topics/deep-learning/attention-mechanism-deep-learning/>

MANN, W. C.; THOMPSON, S. A. Rhetorical structure theory: Toward a functional theory of text organization. **Text-interdisciplinary Journal for the Study of Discourse**, v. 8, n. 3, p. 243–281, 1988.

MANNING, C. D.; SCHÜTZE, H. **Foundations of statistical natural language processing**. Cambridge, USA: mitpress, 1999.

MARCACINI, R. M. et al. Cross-domain aspect extraction for sentiment analysis: A transductive learning approach. **Decision Support Systems**, v. 114, p. 70–80, 2018.

MARCACINI, R. M.; CANDIDO JUNIOR, A.; CASANOVA, E. **Overview of the Automatic Speech Recognition for Spontaneous and Prepared Speech & Speech Emotion Recognition in Portuguese (SE&R) Shared-tasks at PROPOR 2022**. Proceedings of the Workshop on Automatic Speech Recognition for Spontaneous and Prepared Speech & Speech Emotion Recognition in Portuguese co-located with 15th edition of the International Conference on the Computational Processing of Portuguese (PROPOR 2022). **Anais**...2022.

MARCHAL, M. et al. **Establishing Annotation Quality in Multi-label Annotations**. (N. Calzolari et al., Eds.)Proceedings of the 29th International Conference on Computational Linguistics. **Anais**...Gyeongju, Republic of Korea: International Committee on Computational Linguistics, out. 2022. Disponível em: <https://aclanthology.org/2022.coling-1.322>

MARCU, D. **From local to global coherence: A bottom-up approach to text planning**. AAAI/IAAI. **Anais**...Citeseer, 1997.

MARCU, D.; CARLSON, L.; WATANABE, M. **The automatic translation of discourse structures**. 1st Meeting of the North American Chapter of the Association for Computational Linguistics. **Anais**...2000.

MARCUSCHI, L. A. Atos de referenciação na interação face a face. **Cadernos de Estudos Linguísticos**, v. 41, p. 37–54, 2001.

MAREGA, L. M. P.; JUNG, N. M. A sobreposição de falas na conversa cotidiana: disputa pela palavra? **Revista Veredas**, v. 15, n. 1, 2011.

MARKANTONATOU, S. et al. **IDION: A database for Modern Greek multiword expressions**. Proceedings of the Joint Workshop on Multiword Expressions and WordNet (MWE-WN 2019). **Anais**...Florence, Italy: Association for Computational Linguistics,

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

ago. 2019. Disponível em: <https://aclanthology.org/W19-5115>

MARKANTONATOU, S. et al. **PMWE conventions for examples containing multiword expressions**., 2021. Disponível em: <https://gitlab.com/parseme/pmwe/-/raw/master/Conventions-for-MWE-examples/PMWE_series_conventions_for_multilingual_examples.pdf>

MARNEFFE, M.-C. DE et al. Universal Dependencies. **Computational Linguistics**, v. 47, n. 2, p. 255–308, jun. 2021.

MARSLEN-WILSON, W. Linguistic structure and speech shadowing at very short latencies. **Nature**, v. 244, n. 5417, p. 522–523, 1973.

MARTINS, E. J. **Enunciação e diálogo**. tese de doutorado—[s.l.] Universidade Estadual de Campinas, Instituto de Estudos da Linguagem, Campinas, SP, 1987.

MARTINS, H. Sobre a estabilidade do significado em Wittgenstein. **Veredas**, v. 4, n. 2, p. 19–42, 2000.

MARTINS, H. Três Caminhos na Filosofia da Linguagem. Em: **Introdução à Linguística. Volume III**. [s.l.] Editora Cortez, 2004.

MARTINS, R.; NUNES, M. DAS G. V.; HASEGAWA, R. **Curupira: A Functional Parser for Brazilian Portuguese**. (N. J. Mamede et al., Eds.)Computational Processing of the Portuguese Language. **Anais**...Berlin, Heidelberg: Springer Berlin Heidelberg, 2003.

MARTSCHAT, S.; STRUBE, M. Latent Structures for Coreference Resolution. **Transactions of the Association for Computational Linguistics**, v. 3, p. 405–418, 2015.

MATOS, V. B. et al. **Coordination within Conversational Agents with Multiple Sources**. Anais do XX Encontro Nacional de Inteligência Artificial e Computacional. **Anais**...SBC, 2023. Disponível em: <https://doi.org/10.5753/eniac.2023.234533>

MATTOS, L. DE et al. **Contribuições para o desenvolvimento de Agentes Pedagógicos Conversacionais e sua integração a Ambientes Virtuais de Aprendizagem**. Anais do XXXIII Simpósio Brasileiro de Informática na Educação. **Anais**...SBC, 2022. Disponível em: <https://doi.org/10.5753/sbie.2022.225088 >

MAZIERO, E. G. et al. **A base de dados lexical e a interface web do TeP 2.0: thesaurus eletrônico para o Português do Brasil**. Proceedings of the XIV Brazilian Symposium on Multimedia and the Web. **Anais**...Salvador, Brazil: 2008.

MAZIERO, E. G. **Análise retórica com base em grande quantidade de dados**. tese de doutorado—[s.l.] Universidade de São Paulo, 2016.

MAZIERO, E. G.; HIRST, G.; PARDO, T. A. S. **Adaptation of discourse parsing models for the Portuguese language**. 2015 Brazilian Conference on Intelligent Systems

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

(BRACIS). **Anais**...IEEE, 2015.

MAZIERO, E. G.; JORGE, M. L. DEL R. C.; PARDO, T. A. S. Identifying Multidocument Relations. **NLPCS**, v. 7, p. 60–69, 2010.

MAZIERO, E. G.; PARDO, T. A. S. Automatic Identification of Multi-document Relations. **Proceedings of the PROPOR 2012 PhD and MSc/MA Dissertation Contest**, p. 1–8, 2012.

MAZUMDER, M. et al. **Multilingual spoken words corpus**. Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2). **Anais**...2021.

MCCRAE, J. P. et al. **English WordNet 2019 – An Open-Source WordNet for English**. Proceedings of the 10th Global Wordnet Conference. **Anais**...Wroclaw, Poland: Global Wordnet Association, jul. 2019. Disponível em: <https://aclanthology.org/2019.gwc-1.31>

MCDONALD, R. et al. **Universal Dependency Annotation for Multilingual Parsing**. Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). **Anais**...Sofia, Bulgaria: Association for Computational Linguistics, ago. 2013. Disponível em: <https://aclanthology.org/P13-2017>

MCGUIRE, J. et al. The reputational and ethical consequences of deceptive chatbot use. **Scientific Reports**, v. 13, n. 1, 2023.

MEER, M. VAN DER et al. **Annotator-Centric Active Learning for Subjective NLP Tasks**. (Y. Al-Onaizan, M. Bansal, Y.-N. Chen, Eds.)Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing. **Anais**...Miami, Florida, USA: Association for Computational Linguistics, nov. 2024. Disponível em: <https://aclanthology.org/2024.emnlp-main.1031/>

MEI, H.; BANSAL, M.; WALTER, M. R. **What to talk about and how? Selective Generation using LSTMs with Coarse-to-Fine Alignment**. Proceedings of NAACL-2016. **Anais**...: HLT-NAACL'16.San Diego, California: Association for Computational Linguistics, 2016. Disponível em: <http://aclanthology.coli.uni-saarland.de/pdf/N/N16/N16-1086.pdf>

MEL'ČUK, I. **General Phraseology: Theory and Practice**. Amsterdam/Philadelphia: John Benjamins, 2023. v. 36

MEL'ČUK, I.; CLAS, A.; POLGUÈRE, A. **Introduction à la lexicologie explicative et combinatoire**. Louvain la Neuve, Belgium: Editions Duculot, 1995.

MEL'ČUK, I.; POLGUÈRE, A. A Formal Lexicon In The Meaning-Text Theory Or (How To Do Lexica With Words). **cl**, v. 13, n. 3-4, p. 261–275, 1987.

MELLO, H.; RASO, T.; ALMEIDA FERRARI, L. DE. **C-ORAL–Brasil II: Corpus**

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

de referência do português brasileiro falado informal., no prelono prelo.

MELO, G. DE; WEIKUM, G. **Towards a universal wordnet by learning from combined evidence**. Proceedings of the 18th ACM conference on Information and knowledge management. **Anais**...2009.

MENDES, R. B.; OUSHIRO, L. **Mapping Paulistano Portuguese: the SP2010 Project**. Proceedings of the VIIth GSCP International Conference: Speech and Corpora. **Anais**...Firenze, Italy: Fizenze University Press, 2012.

MIKOLOV, T. et al. **Efficient Estimation of Word Representations in Vector Space**., a2013. Disponível em: <https://arxiv.org/abs/1301.3781>

MIKOLOV, T. et al. **Distributed Representations of Words and Phrases and their Compositionality**. (C. J. Burges et al., Eds.)Advances in Neural Information Processing Systems. **Anais**...Curran Associates, Inc., b2013. Disponível em: <https://proceedings.neurips.cc/paper_files/paper/2013/file/9aa42b31882ec039965f3c4923ce901b-Paper.pdf>

MINSKY, M. A framework for representing knowledge. **The psychology of computer vision**, 1975.

MITCHELL, M. et al. **Model cards for model reporting**. Proceedings of the conference on fairness, accountability, and transparency. **Anais**...2019.

MITKOV, R. **The Oxford handbook of Computational Linguistics**. [s.l.] Oxford University Press, 2003.

MITKOV, R. 21 Discourse Processing. **The handbook of computational linguistics and natural language processing**, p. 599, 2010.

MOL, L. et al. The communicative import of gestures: Evidence from a comparative analysis of human–human and human–machine interactions. **Gesture**, v. 9, n. 1, p. 97–126, 2009.

MONTI, J. et al. (EDS.). **Proceedings of The 3rd Workshop on Multi-word Units in Machine Translation and Translation Technology (MUMTTT 2017)**. Geneva, Switzerland: Editions Tradulex, 2017.

MOORE, R. K. Spoken language processing: Piecing together the puzzle. **Speech Communication**, v. 49, n. 5, p. 418–435, 2007.

MORAES GARCEZ, P. DE; STEIN, F. Organização da fala-em-interação: o dispositivo para o gerenciamento de fala sobreposta na conversa cotidiana em dados de português brasileiro. **Revista de Estudos da Linguagem**, v. 23, n. 1, p. 159–194, 2015.

MORETTI, F. **Distant Reading**. [s.l.] Verso, 2013.

MOTA, C. et al. É tempo de avaliar o tempo. Em: MOTA, C.; SANTOS, D. (Eds.).

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Desafios na avaliação conjunta do reconhecimento de entidades mencionadas. [s.l.] Linguateca, 2008. p. 55–75.

MOTA, C. et al. **Págico: Evaluating Wikipedia-based information retrieval in Portuguese**. (N. Calzolari et al., Eds.)Proceedings of the Eigth International Conference on Language Resources and Evaluation (LREC'12). **Anais**...Istambul: 2012. Disponível em: <http://www.lrec-conf.org/proceedings/lrec2012/pdf/590_Paper.pdf>

MULLER, P. et al. **Manuel d'annotation en relations de discours du projet annodis**., 2012.

MUNIZ, M. C. M. **A construção de recursos linguístico-computacionais para o português do Brasil: o projeto Unitex-PB**. mathesis—[s.l.] Instituto de Ciências Matemáticas e de Computação - Universidade de São Paulo - ICMC/USP, 2004.

MURTARELLI, G.; GREGORY, A.; ROMENTI, S. A conversation-based perspective for shaping ethical human–machine interactions: The particular challenge of chatbots. **Journal of Business Research**, v. 129, p. 927–935, 2021.

MUSGRAVE, K.; BELONGIE, S.; LIM, S.-N. **A metric learning reality check**. Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXV 16. **Anais**...Springer, 2020. Disponível em: <https://doi.org/10.1007/978-3-030-58595-2_41>

NAMIUTI, C. O Corpus Anotado do Português Histórico: um avanço para as pesquisas em Linguística Histórica do Português. **Revista Virtual de Estudos da Linguagem**, v. 2, p. 1–9, ago. 2004.

NASCIMENTO, M. F. B. DO; GONÇALVES, J. B. Corpus de Referência do Português Contemporâneo (CRPC) - desenvolvimento e aplicações. **Actas do XI Encontro Nacional da Associação Portuguesa de Lingüística**, v. 1, p. 143–150, 1996.

NAVIGLI, R.; PONZETTO, S. P. BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. **Artificial intelligence**, v. 193, p. 217–250, 2012.

NEEDHAM, M.; HODLER, A. E. **Graph Algorithms: Practical Examples in Apache Spark and Neo4j**. [s.l.] O'Reilly Media, 2019.

NETO, J. P. et al. **Design of a multimodal input interface for a dialogue system**. Computational Processing of the Portuguese Language: 7th International Workshop, PROPOR 2006, Itatiaia, Brazil, May 13-17, 2006. Proceedings 7. **Anais**...Springer, 2006. Disponível em: <https://doi.org/10.1007/11751984_18>

NEVES, M. H. DE M. **Texto e gramática**. [s.l.] Contexto, 2013.

NEWELL, A. A tutorial on speech understanding systems. **Speech recognition**, p. 4–54, 1975.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

NG, V.; CARDIE, C. **Improving machine learning approaches to coreference resolution**. Proceedings of the 40th Annual Meeting on Association for Computational Linguistics. **Anais**...Association for Computational Linguistics, 2002.

NIVRE, J. et al. **The CoNLL 2007 Shared Task on Dependency Parsing**. Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL). **Anais**...Prague, Czech Republic: Association for Computational Linguistics, jun. 2007. Disponível em: <https://aclanthology.org/D07-1096>

NIVRE, J.; FANG, C.-T. **Universal Dependency Evaluation**. (M.-C. de Marneffe, J. Nivre, S. Schuster, Eds.)Proceedings of the NoDaLiDa 2017 Workshop on Universal Dependencies (UDW 2017). **Anais**...Gothenburg, Sweden: Association for Computational Linguistics, 2017. Disponível em: <https://aclanthology.org/W17-0411>

NIVRE, J.; NILSSON, J. Multiword units in syntactic parsing. **Proceedings of Methodologies and Evaluation of Multiword Units in Real-World Applications (MEMURA)**, 2004.

NOORALAHZADEH, F.; ØVRELID, L. **Syntactic Dependency Representations in Neural Relation Classification**. Proceedings of the Workshop on the Relevance of Linguistic Structure in Neural Architectures for NLP. **Anais**...Melbourne, Australia: Association for Computational Linguistics, jul. 2018. Disponível em: <https://aclanthology.org/W18-2907>

NOVIKOVA, J. et al. **Why We Need New Evaluation Metrics for NLG**. Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. **Anais**...: EMNLP'17.Copenhagen, Denmark: Association for Computational Linguistics, 2017. Disponível em: <http://aclweb.org/anthology/D17-1237>

NOZAKI, J. et al. **End-to-end Speech-to-Punctuated-Text Recognition**. Proc. Interspeech 2022. **Anais**...2022.

NUNES, A. S. **A coconstrução do conhecimento através de jogos de linguagem em uma aula de língua portuguesa: um estudo das estratégias de leitura a partir da análise dos enquadres interacionais.** mathesis—[s.l.] Programa de Pós-Graduação em Letras, Mestrado Profissional (PROFLETRAS); Universidade do Estado do Rio de Janeiro, 2016.

NUNES, E. G. Os marcadores conversacionais na constituição do texto falado. **Verbum. Cadernos de Pós-Graduação. ISSN 2316-3267**, v. 6, n. 2, p. 120–125, 2017.

NUNES, P. **LEVANTAMENTO REVELA QUE 90,5% DOS PRESOS POR MONITORAMENTO FACIAL NO BRASIL SÃO NEGROS**. Disponível em: <https://www.intercept.com.br/2019/11/21/presos-monitoramento-facial-brasil-negros/>. Acesso em: 28 ago. 2023.

O'NEIL, C. **Algoritmos de Destruição em Massa**. [s.l.] Editora Rua do Sabão, 2021.

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

OECD. **The OECD Framework for the Classification of AI systems**. Disponível em: <https://wp.oecd.ai/app/uploads/2022/02/Classification-2-pager-1.pdf>. Acesso em: 28 ago. 2023.

OECD. **OECD Employment Outlook 2023**. [s.l: s.n.]. p. 267

OLIVEIRA, F. S. et al. **CML-TTS: A Multilingual Dataset for Speech Synthesis in Low-Resource Languages**. International Conference on Text, Speech, and Dialogue. **Anais**...Springer, 2023.

OLIVEIRA, L. E. S. E. et al. **Experiments on Portuguese Clinical Question Answering**. (A. Britto, K. Valdivia Delgado, Eds.)Intelligent Systems. **Anais**...Cham: Springer International Publishing, 2021.

OLIVEIRA, L. F. A. DE et al. **Challenges In Annotating A Treebank Of Clinical Narratives In Brazilian Portuguese**. Computational Processing of the Portuguese Language: 15th International Conference, PROPOR 2022, Fortaleza, Brazil, March 21–23, 2022, Proceedings. **Anais**...Berlin, Heidelberg: Springer-Verlag, 2022. Disponível em: <https://doi.org/10.1007/978-3-030-98305-5_9>

OLIVEIRA, L. M. DE; DIAS, J. G. O autorreparo como estratégia adaptativa na fala em interação de um afásico. **Linguagem em (Dis) curso**, v. 18, p. 49–68, 2018.

OLIVEIRA, M. R. DE et al. Repetição em diálogos: análise funcional da conversação. **Série Ensaios**, v. 9, 1998.

OLIVEIRA, M. R. DE. Manual de Linguística. Em: MARTELOTTA, M. E. (Ed.). São Paulo: Contexto, 2008. p. 193–204.

OLIVER, A. et al. **Realistic evaluation of deep semi-supervised learning algorithms**. Proceedings of the 32nd International Conference on Neural Information Processing Systems. **Anais**...: NIPS'18.Red Hook, NY, USA: Curran Associates Inc., 2018.

OLIVIERA JR., M. NURC Digital: um protocolo para a digitalização, anotação, arquivamento e disseminação do material do Projeto da Norma Urbana Linguística Culta (NURC). **CHIMERA: Revista de Corpus de Lenguas Romances y Estudios Lingüísticos**, v. 3, n. 2, p. 149–174, set. 2016.

OPENAI. **ChatGPT: OpenA's conversational AI model**. Disponível em: <https://openai.com/blog/chatgpt/>. Acesso em: 7 abr. 2023.

ORENGO, V. M.; HUYCK, C. **A Stemming Algorithmm for the Portuguese Language**. Proceedings Eighth Symposium on String Processing and Information Retrieval. **Anais**...IEEE Computer Society, 2001.

OSBORNE, D. M. The realization of speech acts of refusals of an invitation among Brazilian friends. **Revista de estudos da linguagem**, v. 18, n. 2, p. 61–85, 2010.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

OSBORNE, T.; GERDES, K. The status of function words in dependency grammar: A critique of Universal Dependencies (UD). **Glossa: a journal of general linguistics (2016-2021)**, jan. 2019.

OSGOOD, C. E.; SUCI, G. J.; TENENBAUM, P. H. **The Measurement of meaning**. Urbana: University of Illinois Press, 1957.

OSTENDORF, M.; PRICE, P.; SHATTUCK-HUFNAGEL, S. **The Boston University Radio news corpus**., 1995. Disponível em: <https://doi.org/10.35111/Z7XK-Z229>

OSTERMANN, A. C.; ANDRADE, D. N. P.; FREZZA, M. A prosódia como componente de formação e de atribuição de sentido a ações na fala-em-interação: o caso de formulações no tribunal. **DELTA: Documentação de Estudos em Lingüística Teórica e Aplicada**, v. 32, p. 481–513, 2016.

OUSHIRO, L. Wh-interrogatives in Brazilian Portuguese: the influence of common ground. **University of Pennsylvania Working Papers in Linguistics**, v. 17, n. 2, p. 17, 2011.

OUSHIRO, L.; MENDES, R. B. A Variação em interrogativas de constituinte no fluxo conversacional. **Signum: Estudos da Linguagem**, v. 15, n. 3, p. 273–292, 2012.

OUYANG, L. et al. **Training language models to follow instructions with human feedback**. (A. H. Oh et al., Eds.)Advances in Neural Information Processing Systems. **Anais**...2022. Disponível em: <https://openreview.net/forum?id=TG8KACxEON>

OVCHINNIKOVA, E. **Integration of World Knowledge for Natural Language Understanding**. [s.l.] Atlantis Press, 2012.

ÖZSEVEN, T. Investigation of the effect of spectrogram images and different texture analysis methods on speech emotion recognition. **Applied Acoustics**, v. 142, p. 70–77, 2018.

PĂIŞ, V.; TUFIŞ, D. Capitalization and punctuation restoration: a survey. **Artificial Intelligence Review**, v. 55, p. 1681--1722, 2022.

PALMER, M.; FININ, T.; WALTER, S. M. **Workshop on the Evaluation of Natural Language Processing Systems**. [s.l.] Air Force Systems Command; Rome Air Development Center, 1988. Disponível em: <https://ebiquity.umbc.edu/paper/html/id/1074>.

PALMER, M.; GILDEA, D.; KINGSBURY, P. **The Proposition Bank: An Annotated Corpus of Semantic Roles**. Computational Linguistics, 31: 1. **Anais**...The MIT PressJournals, 2005.

PAPINENI, K. et al. **BLEU: A Method for Automatic Evaluation of Machine Translation**. Proceedings of the 40th Annual Meeting on Association for Computational Linguistics. **Anais**...: ACL '02.USA: Association for Computational Linguistics, 2002. Disponível em: <https://doi.org/10.3115/1073083.1073135>

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

PARABONI, I.; GALINDO, M.; IACOVELLI, D. Stars2: a corpus of object descriptions in a visual domain. **Language Resources and Evaluation**, v. 51, n. 2, p. 439–462, 2017.

PARDO, T. et al. **Porttinari - a Large Multi-genre Treebank for Brazilian Portuguese**. Anais do XIII Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana. **Anais**...Porto Alegre, RS, Brasil: SBC, 2021. Disponível em: <https://sol.sbc.org.br/index.php/stil/article/view/17778>

PARDO, T. A. S. **Métodos para análise discursiva automática**. tese de doutorado— [s.l.] Universidade de São Paulo, 2005.

PARK, D. S. et al. **SpecAugment: A Simple Data Augmentation Method for Automatic Speech Recognition**. Interspeech 2019. **Anais**...ISCA, set. 2019. Disponível em: <https://doi.org/10.21437%2Finterspeech.2019-2680>

PARMAR, M. et al. **Don't Blame the Annotator: Bias Already Starts in the Annotation Instructions**. Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics. **Anais**...Dubrovnik, Croatia: Association for Computational Linguistics, 2023. Disponível em: <https://aclanthology.org/2023.eacl-main.130>

PAROUBEK, P.; CHAUDIRON, S.; HIRSCHMAN, L. Principles of Evaluation in Natural Language Processing. **Revue TAL**, v. 48, n. 1, p. 7–31, 2007.

PARRA ESCARTÍN, C. et al. **Ethical Considerations in NLP Shared Tasks**. Proceedings of the First ACL Workshop on Ethics in Natural Language Processing. **Anais**...Valencia, Spain: Association for Computational Linguistics, abr. 2017. Disponível em: <https://aclanthology.org/W17-1608>

PARRA ESCARTÍN, C.; NEVADO LLOPIS, A.; SÁNCHEZ MARTÍNEZ, E. Spanish multiword expressions: Looking for a taxonomy. Em: **Multiword expressions: Insights from a multi-lingual perspective**. [s.l.] Language Science Press, 2018. p. 271–323.

PASCHOAL, A. F. et al. **Pirá: A bilingual Portuguese-English dataset for question-answering about the ocean**. Proceedings of the 30th ACM International Conference on Information & Knowledge Management. **Anais**...2021. Disponível em: <https://doi.org/10.1145/3459637.3482012>

PASCHOAL, L. N. et al. **Towards a Conversational Agent to Support the Software Testing Education**. Proceedings of the XXXIII Brazilian Symposium on Software Engineering. **Anais**...: SBES '19.New York, NY, USA: Association for Computing Machinery, 2019. Disponível em: <https://doi.org/10.1145/3350768.3352456>

PASQUER, C. et al. **Verbal Multiword Expression Identification: Do We Need a Sledgehammer to Crack a Nut?** Proceedings of the 28th International Conference on Computational Linguistics. **Anais**...Barcelona, Spain (Online): International Committee

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

on Computational Linguistics, dez. 2020.

PAULLADA, A. et al. Data and its (dis) contents: A survey of dataset development and use in machine learning research. **Patterns**, v. 2, n. 11, 2021.

PENNINGTON, J.; SOCHER, R.; MANNING, C. **GloVe: Global Vectors for Word Representation**. Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). **Anais**...Doha, Qatar: Association for Computational Linguistics, out. 2014. Disponível em: <https://aclanthology.org/D14-1162>

PEREIRA, A. et al. Systematic review of question answering over knowledge bases. **IET Software**, v. 16, n. 1, p. 1–13, 2022.

PEREZ-BELTRACHINI, L. et al. **Content selection as semantic-based ontology exploration**. (C. Gardent, A. Gangemi, Eds.)Proceedings of the 2nd International Workshop on Natural Language Generation and the Semantic Web (WebNLG 2016). **Anais**...Edinburgh, Scotland: Association for Computational Linguistics, set. 2016. Disponível em: <https://aclanthology.org/W16-3508>

PERRAULT, C. R.; ALLEN, J. F. **Speech Acts as a Basis for Understanding Dialogue Coherence**. Theoretical Issues in Natural Language Processing-2. **Anais**...1978. Disponível em: <https://aclanthology.org/T78-1017>

PERRIGO, B. Disponível em: <https://time.com/6247678/openai-chatgpt-kenya-workers/>. Acesso em: 9 abr. 2023.

PIAI, L. **Anotação de Corpus: Caracterização de Entidades Nomeadas em Tweets do Mercado Financeiro**. Mestrado em Linguística—São Paulo: Programa de Pós-Graduação em Linguística da Universidade Federal de São Carlos, 2025.

PIMENTEL, C. L. **A elaboração de um corpus oral: a etapa de transcrição da interação na sala de aula de português como língua adicional**. mathesis—[s.l.] Pontifícia Universidade Católica do Rio Grande do Sul, 2016.

PING, W. et al. Deep voice 3: 2000-speaker neural text-to-speech. **arXiv preprint arXiv:1710.07654**, 2017.

PINHEIRO, V. et al. **InferenceNet.Br: Expression of Inferentialist Semantic Content of the Portuguese Language**. (T. A. S. Pardo et al., Eds.)Computational Processing of the Portuguese Language. **Anais**...Berlin, Heidelberg: Springer Berlin Heidelberg, 2010.

PIRES, I.; CASELI, H.; NERIS, V. **Design de um chatbot para o diálogo com universitários com possível perfil depressivo**. Anais Estendidos do XXIII Simpósio Brasileiro de Computação Aplicada à Saúde. **Anais**...Porto Alegre, RS, Brasil: SBC, a2023. Disponível em: <https://sol.sbc.org.br/index.php/sbcas_estendido/article/view/25323>

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

PIRES, R. et al. **Sabiá: Portuguese Large Language Models**. (M. C. Naldi, R. A. C. Bianchi, Eds.)Intelligent Systems. **Anais**...Cham: Springer Nature Switzerland, b2023.

PLACANI, A. Anthropomorphism in AI: hype and fallacy. **AI and Ethics**, p. 1–8, 2024.

PLANK, B.; HOVY, D.; SØGAARD, A. **Linguistically debatable or just plain wrong?** (K. Toutanova, H. Wu, Eds.)Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). **Anais**...Baltimore, Maryland: Association for Computational Linguistics, jun. 2014. Disponível em: <https://aclanthology.org/P14-2083>

POESIO, M.; STUCKARDT, R.; VERSLEY, Y. **Anaphora Resolution: Algorithms, Resources, and Applications**. 1. ed. [s.l.] Springer, 2016.

PONTIKI, M. et al. **SemEval-2014 Task 4: Aspect Based Sentiment Analysis**. Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014). **Anais**...Association for Computational Linguistics, 2014. Disponível em: <https://aclanthology.org/S14-2004/>

POPIEL, S. J.; MCRAE, K. The figurative and literal senses of idioms, or all idioms are not used equally. **Journal of Psycholinguistic Research**, v. 17, n. 6, p. 475–487, 1 nov. 1988.

POPOVIĆ, M. **chrF++: words helping character n-grams**. Proceedings of the second conference on machine translation. **Anais**...2017.

PORTER, M. F. An algorithm for suffix stripping. **Program**, v. 14, n. 3, p. 130–137, 1980.

PORTET, F. et al. Automatic generation of textual summaries from neonatal intensive care data. **Artificial Intelligence**, v. 173, n. 7–8, p. 789–816, 2009.

POSNER, J.; RUSSELL, J. A.; PETERSON, B. S. The circumplex model of affect: An integrative approach to affective neuroscience, cognitive development, and psychopathology. **Development and psychopathology**, v. 17, n. 3, p. 715–734, 2005.

PRABHAKARAN, V.; RAMBOW, O. **Written Dialog and Social Power: Manifestations of Different Types of Power in Dialog Behavior**. Proceedings of the Sixth International Joint Conference on Natural Language Processing. **Anais**...Nagoya, Japan: Asian Federation of Natural Language Processing, out. 2013. Disponível em: <https://aclanthology.org/I13-1025>

PRADHAN, S. et al. **CoNLL-2011 shared task: Modeling unrestricted coreference in ontonotes**. Proceedings of the Fifteenth Conference on Computational Natural Language Learning: Shared Task. **Anais**...Portland, Oregon: Association for Computational Linguistics, 2011.

PRADHAN, S. et al. **CoNLL-2012 shared task: Modeling multilingual unrestric-**

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

ted coreference in OntoNotes. Proceedings of Joint Conference on Empirical Methods in Natural Language Processing and Conference on Natural Language Learning - Shared Task. **Anais**...Jeju Island, Korea: 2012.

PRADHAN, S. et al. **Scoring Coreference Partitions of Predicted Mentions: A Reference Implementation**. Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics. **Anais**...Baltimore, MD, USA: 2014. Disponível em: <http://aclweb.org/anthology/P/P14/P14-2006.pdf>

PRATAP, V. et al. **Massively Multilingual ASR: 50 Languages, 1 Model, 1 Billion Parameters**., a2020. Disponível em: <https://arxiv.org/abs/2007.03001>

PRATAP, V. et al. MLS: A Large-Scale Multilingual Dataset for Speech Research. **Proc. Interspeech 2020**, p. 2757–2761, b2020.

PRZEPIÓRKOWSKI, A. et al. **Extended phraseological information in a valence dictionary for NLP applications**. Proceedings of Workshop on Lexical and Grammatical Resources for Language Processing. **Anais**...Dublin, Ireland: Association for Computational Linguistics; Dublin City University, ago. 2014. Disponível em: <https://aclanthology.org/W14-5811>

PURINGTON, A. et al. **" Alexa is my new BFF" Social Roles, User Satisfaction, and Personification of the Amazon Echo**. Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems. **Anais**...2017.

PURVER, M. et al. **Split Utterances in Dialogue: a Corpus Study**. Proceedings of the SIGDIAL 2009 Conference. **Anais**...London, UK: Association for Computational Linguistics, set. 2009. Disponível em: <https://aclanthology.org/W09-3937>

PURVER, M. R. J. **The theory and use of clarification requests in dialogue**. tese de doutorado—[s.l.] University of London, 2004.

QUARESMA, P.; RODRIGUES, I. **Using dialogues to access semantic knowledge in a web IR system**. Computational Processing of the Portuguese Language: 6th International Workshop, PROPOR 2003 Faro, Portugal, June 26–27, 2003 Proceedings 6. **Anais**...Springer, 2003. Disponível em: <https://doi.org/10.1007/3-540-45011-4_32>

QUARESMA, P.; RODRIGUES, I. **A Question-Answering System for Portuguese Juridical Documents**. Proceedings of the 10th International Conference on Artificial Intelligence and Law. **Anais**...: ICAIL '05.New York, NY, USA: Association for Computing Machinery, 2005. Disponível em: <https://doi.org/10.1145/1165485.1165536>

QUINTANILHA, I. M.; NETTO, S. L.; BISCAINHO, L. W. P. An open-source end-to-end ASR system for Brazilian Portuguese using DNNs built from newly assembled corpora. **Journal of Communication and Information Systems**, v. 35, n. 1, p. 230–242, 2020.

QUINTANO, L.; RODRIGUES, I. **Managing dialog and access control in natural language querying**. Computational Processing of the Portuguese Language: 6th

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

International Workshop, PROPOR 2003 Faro, Portugal, June 26–27, 2003 Proceedings 6. **Anais**...Springer, 2003. Disponível em: <https://doi.org/10.1007/3-540-45011-4_33>

RABINER, L. R.; JUANG, B. H. **Fundamentals of Speech Recognition**. [s.l.] Pearson Education, 1993.

RADEMAKER, A. et al. **Universal Dependencies for Portuguese**. Proceedings of the Fourth International Conference on Dependency Linguistics (Depling 2017). **Anais**...Pisa,Italy: Linköping University Electronic Press, set. 2017. Disponível em: <https://aclanthology.org/W17-6523>

RADEV, D. R. **A common theory of information fusion from multiple text sources step one: cross-document structure**. 1st SIGdial workshop on Discourse and Dialogue. **Anais**...2000.

RADFORD, A. et al. Robust speech recognition via large-scale weak supervision. **arXiv preprint arXiv:2212.04356**, 2022.

RAFFEL, C. et al. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. **Journal of Machine Learning Research**, v. 21, n. 140, p. 1–67, 2020.

RAHMAN, A.; NG, V. **Coreference Resolution with World Knowledge**. Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies. **Anais**...Portland, Oregon, USA: a2011. Disponível em: <http://www.aclweb.org/anthology/P11-1082>

RAHMAN, A.; NG, V. Narrowing the modeling gap: a cluster-ranking approach to coreference resolution. **Journal of Artificial Intelligence Research**, p. 469–521, b2011.

RAJI, D. et al. **AI and the Everything in the Whole Wide World Benchmark**. (J. Vanschoren, S. Yeung, Eds.)Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks. **Anais**...Curran, 2021. Disponível em: <https://datasets-benchmarks-proceedings.neurips.cc/paper_files/paper/2021/file/084b6fbb10729ed4da8c3d3f5a3ae7c9-Paper-round2.pdf>

RAJPURKAR, P. et al. **SQuAD: 100,000+ Questions for Machine Comprehension of Text**. (J. Su, K. Duh, X. Carreras, Eds.)Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing. **Anais**...Austin, Texas: Association for Computational Linguistics, nov. 2016. Disponível em: <https://aclanthology.org/D16-1264>

RAMISCH, C. **Multiword Expressions Acquisition: A Generic and Open Framework**. [s.l.] Springer, 2015. v. XIVp. 230

RAMISCH, C. et al. **DeQue: A Lexicon of Complex Prepositions and Conjunctions in French**. Proceedings of LREC 2016. **Anais**...Portoroz, Slovenia: ELRA, a2016.

RAMISCH, C. et al. **How Naked is the Naked Truth? A Multilingual Lexicon of**

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Nominal Compound Compositionality. Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). **Anais**...Berlin, Germany: ACL, b2016.

RAMISCH, C. et al. **Edition 1.1 of the PARSEME Shared Task on Automatic Identification of Verbal Multiword Expressions**. Proceedings of the Joint Workshop on Linguistic Annotation, Multiword Expressions and Constructions (LAW-MWE-CxG-2018). **Anais**...Santa Fe, NM, USA: ACL, a2018.

RAMISCH, C. et al. **A Corpus Study of Verbal Multiword Expressions in Brazilian Portuguese**. Computational Processing of the Portuguese Language 13th International Conference, PROPOR 2018, Canela, Brazil, September 24–26, 2018, Proceedings. **Anais**...: Lecture Notes em Artificial Intelligence.Cham, Switzerland: Springer International Publishing, b2018.

RAMISCH, C. et al. **Edition 1.2 of the PARSEME Shared Task on Semi-supervised Identification of Verbal Multiword Expressions**. Proceedings of the Joint Workshop on Multiword Expressions and Electronic Lexicons. **Anais**...online: Association for Computational Linguistics, 2020. Disponível em: <https://www.aclweb.org/anthology/2020.mwe-1.14>

RAMISCH, C. **Multiword expressions in computational linguistics: down the rabbit hole and through the looking glass**. tese de doutorado—Marseille, France: Aix Marseille University, 2023.

RAMISCH, C.; BESACIER, L.; KOBZAR, A. **How hard is it to automatically translate phrasal verbs from English to French?** MT Summit 2013 Workshop on Multiword Units in Machine Translation and Translation Technology. **Anais**...Nice, France: 2013.

RAMISCH, C.; VILLAVICENCIO, A. Computational Treatment of Multiword Expressions. Em: MITKOV, R. (Ed.). **The Oxford Handbook of Computational Linguistics**. 2nd. ed. [s.l.] Oxford University Press, 2018.

RANCHHOD, E.; MOTA, C.; BAPTISTA, J. **A Computational Lexicon of Portuguese for Automatic Text Parsing**. SIGLEX99: Standardizing Lexical Resources. **Anais**...1999. Disponível em: <https://aclanthology.org/W99-0511>

RAO, K. S.; KOOLAGUDI, S. G.; VEMPADA, R. R. Emotion recognition from speech using global and local prosodic features. **International journal of speech technology**, v. 16, p. 143–160, 2013.

RASO, T. et al. O projeto C-ORAL-BRASIL. **CHIMERA: Revista de Corpus de Lenguas Romances y Estudios Lingüísticos**, v. 1, p. 31–67, 2015.

RASO, T.; MELLO, H. **C-ORAL–BRASIL I: corpus de referência do português brasileiro falado informal**. Belo Horizonte: Editora UFMG, 2012a.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

RASO, T.; MELLO, H. C-ORAL–BRASIL I: corpus de referência do português brasileiro falado informal. A general presentation. **Speech and Corpora**, p. 16, b2012.

RASO, T.; TEIXEIRA, B.; BARBOSA, P. Modelling automatic detection of prosodic boundaries for Brazilian Portuguese spontaneous speech. **Journal of Speech Sciences**, v. 9, p. 105–128, set. 2020.

READ, J. et al. **Sentence Boundary Detection: A Long Solved Problem?** Proceedings of COLING 2012: Posters. **Anais**...Mumbai, India: The COLING 2012 Organizing Committee, dez. 2012. Disponível em: <https://aclanthology.org/C12-2096>

REAL, L.; FONSECA, E.; GONÇALO OLIVEIRA, H. **Organizing the ASSIN 2 Shared Task**. Proceedings of the ASSIN 2 Shared Task: Evaluating Semantic Textual Similarity and Textual Entailment in Portuguese: co-located with XII Symposium in Information and Human Language Technology (STIL 2019). **Anais**...2019. Disponível em: <https://ceur-ws.org/Vol-2583/1_ASSIN-2.pdf>

REAL, L.; FONSECA, E.; GONÇALO OLIVEIRA, H. **The ASSIN 2 Shared Task: A Quick Overview**. Computational Processing of the Portuguese Language: 14th International Conference, PROPOR 2020, Evora, Portugal, March 2–4, 2020, Proceedings. **Anais**...Berlin, Heidelberg: Springer-Verlag, 2020. Disponível em: <https://doi.org/10.1007/978-3-030-41505-1_39>

RECASENS, M.; HOVY, E. H. BLANC: Implementing the Rand index for coreference evaluation. **Natural Language Engineering**, v. 17, n. 4, p. 485–510, 2011.

REDDY, S.; MCCARTHY, D.; MANANDHAR, S. **An Empirical Study on Compositionality in Compound Nouns**. Proceedings of 5th International Joint Conference on Natural Language Processing. **Anais**...Chiang Mai, Thailand: Asian Federation of Natural Language Processing, nov. 2011. Disponível em: <https://aclanthology.org/I11-1024>

REI, R. et al. **COMET: A Neural Framework for MT Evaluation**. Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). **Anais**...Online: Association for Computational Linguistics, nov. 2020. Disponível em: <https://aclanthology.org/2020.emnlp-main.213>

REIS, E. S.; SILVA, L. A. DA. Planejamento e replanejamento dos turnos conversacionais. **Cadernos do CNLF**, v. 17, n. 2, p. 1–2013, 2013.

REITER, E. et al. Choosing Words in Computer-generated Weather Forecasts. **Artificial Intelligence**, v. 167, n. 1-2, p. 137–169, set. 2005.

REITER, E. **An Architecture for Data-to-text Systems**. Proceedings of ENLG-2007. **Anais**...: ENLG'07.Germany: Association for Computational Linguistics, 2007. Disponível em: <http://dl.acm.org/citation.cfm?id=1610163.1610180>

REITER, E. A Structured Review of the Validity of BLEU. **Computational Linguistics**,

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

v. 44, n. 3, p. 393–401, 2018.

REITER, E.; DALE, R. **Building natural language generation systems**. New York, NY, USA: Cambridge University Press, 2000.

REITER, E.; ROBERTSON, R.; OSMAN, L. M. Lessons from a failure: Generating tailored smoking cessation letters. **Artificial Intelligence**, v. 144, n. 1, p. 41–58, 2003.

RESNIK, P.; LIN, J. Evaluation of NLP systems. Em: **The handbook of computational linguistics and natural language processing**. [s.l.] Wiley Online Library, 2010. p. 271–295.

REVIEW, M. T. **Um aplicativo de Inteligência Artificial que "desnudava" mulheres mostra como as deepfakes prejudicam os mais vulneráveis**. Disponível em: <https://mittechreview.com.br/um-aplicativo-de-inteligencia-artificial-que-desnudava-mulheres-mostra-como-as-deepfakes-prejudicam-os-mais-vulneraveis/>. Acesso em: 28 ago. 2023.

RIBEIRO, M. T. et al. **Beyond Accuracy: Behavioral Testing of NLP Models with CheckList**. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. **Anais**...Online: Association for Computational Linguistics, jul. 2020. Disponível em: <https://aclanthology.org/2020.acl-main.442>

RIESER, V.; LEMON, O. **Reinforcement learning for adaptive dialogue systems: a data-driven methodology for dialogue management and natural language generation**. [s.l.] Springer Science & Business Media, 2011.

RIZZOLATTI, G.; ARBIB, M. A. Language within our grasp. **Trends in Neurosciences**, v. 21, n. 5, p. 188–194, 1998.

ROBERTS, A. et al. **Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer**. [s.l.] Google, 2019.

ROBERTS, F.; FRANCIS, A. L.; MORGAN, M. The interaction of inter-turn silence with prosodic cues in listener perceptions of "trouble" in conversation. **Speech communication**, v. 48, n. 9, p. 1079–1093, 2006.

ROCHA, E. B.; PIMENTEL, M.; DINIZ, M. C. **Desenvolvimento de um Modelo da Participação em Bate papo seguindo a abordagem Design Science Research**. Anais do X Simpósio Brasileiro de Sistemas de Informação. **Anais**...Porto Alegre, RS, Brasil: SBC, 2014. Disponível em: <https://sol.sbc.org.br/index.php/sbsi/article/view/6099>

ROCHA, M. A corpus-based study of anaphora in English and Portuguese, Corpus-based and Computational Approaches to Discourse Anaphora. Em: [s.l.] John Benjamins Publishing Company, 2000. p. 81–94.

ROCHA, P.; SANTOS, D. CLEF: Abrindo a porta à participação internacional em ava-

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

liação de RI do português. Em: SANTOS, D. (Ed.). **Avaliação conjunta: um novo paradigma no processamento computacional da língua portuguesa**. Lisboa, Portugal: IST Press, 2007. p. 143–158.

RODRIGUES, I. M. G. **Fala e movimentos do corpo na interacção face a face: estratégias de reparação e de (des) focalização e co-funções conversacionais na manutenção de vez**. tese de doutorado—[s.l.] Universidade do Porto, 2003.

RODRIGUES, R.; GOMES, P. **RAPPORT — A Portuguese Question-Answering System**. (F. Pereira et al., Eds.)Progress in Artificial Intelligence. **Anais**...Cham: Springer International Publishing, 2015.

ROGERS, A. **Changing the World by Changing the Data**. Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). **Anais**...Online: Association for Computational Linguistics, ago. 2021. Disponível em: <https://aclanthology.org/2021.acl-long.170>

ROHANIAN, O. et al. **Verbal Multiword Expressions for Identification of Metaphor**. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. **Anais**...Online: Association for Computational Linguistics, jul. 2020. Disponível em: <https://aclanthology.org/2020.acl-main.259>

ROLLER, S.; SCHULTE IM WALDE, S. **Feature Norms of German Noun Compounds**. (V. Kordoni et al., Eds.)Proceedings of the 10th Workshop on Multiword Expressions (MWE). **Anais**...Gothenburg, Sweden: Association for Computational Linguistics, abr. 2014. Disponível em: <https://aclanthology.org/W14-0818>

ROLLER, S.; SCHULTE IM WALDE, S.; SCHEIBLE, S. **The (Un)expected Effects of Applying Standard Cleansing Models to Human Ratings on Compositionality**. Proceedings of the 9th Workshop on Multiword Expressions. **Anais**...Atlanta, Georgia, USA: Association for Computational Linguistics, jun. 2013. Disponível em: <https://aclanthology.org/W13-1005>

RONCARATI, C. **As cadeias do texto: construindo sentidos.** [s.l.] Parábola, 2010.

ROSÉN, V. et al. **MWEs in Treebanks: From Survey to Guidelines**. Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16). **Anais**...Portorož, Slovenia: European Language Resources Association (ELRA), 2016. Disponível em: <https://aclanthology.org/L16-1368>

ROSSANO, F. Gaze in conversation. **The handbook of conversation analysis**, p. 308–329, 2012.

ROSSI, D. et al. **Identifying pedagogical intervention in MOOCs learning processes: a conversational agent proposal**. Anais do XXXII Simpósio Brasileiro de Informática na Educação. **Anais**...Porto Alegre, RS, Brasil: SBC, 2021. Disponível em: <https://sol.sbc.org.br/index.php/sbie/article/view/18112>

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

ROTHE, S.; NARAYAN, S.; SEVERYN, A. Leveraging Pre-trained Checkpoints for Sequence Generation Tasks. **Transactions of the Association for Computational Linguistics**, v. 8, p. 264–280, 2020.

RUANE, E.; BIRHANE, A.; VENTRESQUE, A. **Conversational AI: Social and Ethical Considerations.** AICS. **Anais**...2019. Disponível em: <https://ceur-ws.org/Vol-2563/aics_12.pdf>

RUITER, J. P. DE. Turn-Taking. Em: **The Oxford Handbook of Experimental Semantics and Pragmatics**. [s.l.] Oxford University Press, 2019.

RUPPENHOFER, J. et al. **FrameNet II: Extended theory and practice**. [s.l: s.n.].

RUSSEL, S. **Human Compatible Artificial Intelligence and the Problem of Control**. [s.l.] Penguin Books, 2019.

RUSSELL, S.; NORVIG, P. **Artificial Intelligence: A Modern Approach**. 3rd. ed. USA: Prentice Hall Press, 2009.

SACKS, H.; SCHEGLOFF, E. A.; JEFFERSON, G. A simplest systematics for the organization of turn taking for conversation. Em: **Studies in the organization of conversational interaction**. [s.l.] Elsevier, 1978. p. 7–55.

SADOCK, J. Speech acts. Em: **The handbook of pragmatics**. [s.l.] Wiley Online Library, 2006. p. 53–73.

SAEKI, T. et al. **Virtuoso: Massive Multilingual Speech-Text Joint Semi-Supervised Learning for Text-To-Speech**., 2023. Disponível em: <https://arxiv.org/abs/2210.15447>

SAG, I. A. et al. **Multiword Expressions: A Pain in the Neck for NLP**. Conference on Intelligent Text Processing and Computational Linguistics. **Anais**...2002. Disponível em: <https://api.semanticscholar.org/CorpusID:1826481>

SALESKY, E. et al. The multilingual tedx corpus for speech recognition and translation. **arXiv preprint arXiv:2102.01757**, 2021.

SALOMÃO, M. M. M. FrameNet Brasil: A work in progress. **Calidoscópio**, v. 7, p. 171–182, 2009.

SALTON, G.; ALLAN, J. Text retrieval using the vector processing model. dez. 1994.

SANCHES, M. F. et al. **Textual Datasets For Portuguese-Brazilian Language Models**. Anais do IV Dataset Showcase Workshop. **Anais**...SBC, 2022. Disponível em: <https://doi.org/10.5753/dsw.2022.224294>

SANG, E. F. T. K. **Introduction to the CoNLL-2002 Shared Task: Language-Independent Named Entity Recognition**. Proceedings of CoNLL-2002. **Anais**...Taipei,

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

Taiwan: 2002. Disponível em: <https://aclanthology.org/W02-2024/>

SANTANA, B. P. **Morfologia ornamental: as vogais temáticas do português brasileiro o Unitex-PB**. mathesis—Curitiba, PR: Universidade Federal do Paraná, Setor de Ciências Humanas, Programa de Pós-Graduação em Letras, 2019.

SANTHANAM, S.; SHAIKH, S. A survey of natural language generation techniques with a focus on dialogue systems-past, present and future directions. **arXiv preprint arXiv:1906.00500**, 2019.

SANTOS, D. **O projecto Processamento Computacional do Português: Balanço e perspectivas**. (M. das Graças Volpe Nunes, Ed.)V Encontro para o processamento computacional da língua portuguesa escrita e falada (PROPOR 2000). **Anais**...São Paulo: ICMC/USP, 2000. Disponível em: <https://www.linguateca.pt/Diana/download/SantosPROPOR2000.pdf>

SANTOS, D. **Evaluation in natural language processing**., c2007. Disponível em: <http://www.linguateca.pt/Diana/download/EvaluationESSLLI07.pdf>

SANTOS, D. Avaliação conjunta. Em: SANTOS, D. (Ed.). **Avaliação conjunta: um novo paradigma no processamento computacional da língua portuguesa**. Lisboa, Portugal: IST Press, 2007a. p. 1–12.

SANTOS, D. (ED.). **Avaliação conjunta: um novo paradigma no processamento computacional da língua portuguesa**. Lisboa, Portugal: IST Press, 2007b.

SANTOS, D. Caminhos percorridos no mapa da portuguesificação: A Linguateca em perspectiva. **Linguamática**, v. 1, n. 1, p. 25–59, 2009.

SANTOS, D. et al. GikiP at GeoCLEF 2008: Joining GIR and QA forces for querying Wikipedia. Em: PETERS, C. et al. (Eds.). **Evaluating Systems for Multilingual and Multimodal Information Access 9th Workshop of the Cross-Language Evaluation Forum, CLEF 2008, Aarhus, Denmark, September 17-19, 2008, Revised Selected Papers**. [s.l.] Springer, 2009. p. 894–905.

SANTOS, D. et al. **GikiCLEF: Crosscultural issues in multilingual information access**. (N. Calzolari et al., Eds.)Proceedings of the International Conference on Language Resources and Evaluation (LREC 2010). **Anais**...Valletta, Malta: European Language Resources Association, 2010. Disponível em: <http://www.lrec-conf.org/proceedings/lrec2010/pdf/272_Paper.pdf>

SANTOS, D. et al. (EDS.). **Edição especial Págico - português mágico**. [s.l.] Linguamática, 2012. v. 4

SANTOS, D. et al. Comparando anotações linguísticas na Gramateca: filosofia, ferramentas e exemplos. **Domínios de Lingu@gem**, v. 9, n. 2, p. 11–26, 2015.

SANTOS, D. **Evaluation contests in Portuguese: Linguateca's contribution**., 2021.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Disponível em: <https://www.linguateca.pt/Diana/download/AvalConjLRE16May2021.pdf>

SANTOS, D. et al. DIP - Desafio de Identificação de Personagens: objectivo, organização, recursos e resultados. **Linguamática**, v. 15, n. 1, p. 3–30, 2023.

SANTOS, D.; CABRAL, L. M. GikiCLEF : Expectations and lessons learned. Em: PETERS, C. et al. (Eds.). **Multilingual Information Access Evaluation, VOL I**. [s.l.] Springer, 2010. p. 212–222.

SANTOS, D.; COSTA, L.; ROCHA, P. **Cooperatively evaluating Portuguese morphology**. (J. Baptista et al., Eds.)Computational Processing of the Portuguese Language: 6th International Workshop, PROPOR 2003. Faro, Portugal, June 2003 (PROPOR 2003). **Anais**...Berlin/Heidelberg: Springer Verlag, 2003.

SANTOS, D.; ROCHA, P. **AvalON: uma iniciativa de avaliação conjunta para o português**. (A. Mendes, T. Freitas, Eds.)Actas do XVIII Encontro Nacional da Associação Portuguesa de Linguística (APL 2002). **Anais**...Lisboa: APL, 2003. Disponível em: <https://www.linguateca.pt/Diana/download/SantosRochaAPL2002.pdf>

SANTOS, F. R. DOS et al. **EDUARDO - A Semantic Model for Automatic Content Integration with an Conversational Intelligent Agent**. Anais do XXII Simpósio Brasileiro de Sistemas Multimídia e Web. **Anais**...Porto Alegre, RS, Brasil: SBC, 2016. Disponível em: <https://sol.sbc.org.br/index.php/webmedia/article/view/5372>

SANTOS, F.; FREITAS, T. **CORP-ORAL: Spontaneous Speech Corpus for European Portuguese**. Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08). **Anais**...Marrakech, Morocco: European Language Resources Association (ELRA), 2008. Disponível em: <http://www.lrec-conf.org/proceedings/lrec2008/pdf/331_paper.pdf>

SANTOS, J.; ALVES, A.; GONÇALO OLIVEIRA, H. **Leveraging on Semantic Textual Similarity for developing a Portuguese dialogue system**. International Conference on Computational Processing of the Portuguese Language. **Anais**...Springer, 2020. Disponível em: <https://doi.org/10.1007/978-3-030-41505-1_13>

SANTOS SILVA, D. DOS; PARABONI, I. Generating Spatial Referring Expressions in Interactive 3D Worlds. **Spatial Cognition & Computation**, v. 15, n. 03, p. 186–225, 2015.

SANTOS, V. G. et al. **CORAA NURC-SP Minimal Corpus: a manually annotated corpus of Brazilian Portuguese spontaneous speech**. Proc. IberSPEECH 2022. **Anais**...2022.

SARMENTO, C. DA S. **Da Abordagem do Léxico em Livros Didáticos de Língua Portuguesa: os Anos Finais do Ensino Fundamental**. mathesis—Brasília: UnB, 2019.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

SARMENTO, L.; PINTO, A. S.; CABRAL, L. **REPENTINO – a wide-scope gazetteer for entity recognition in portuguese**. Proceedings of International Workshop on Computational Processing of the Portuguese Language. **Anais**...Springer, 2006.

SARTORI, L.; THEODOROU, A. A Sociotechnical Perspective for the Future of AI: Narratives, Inequalities, and Human Control. **Ethics and Inf. Technol.**, v. 24, n. 1, mar. 2022.

SAURÍ, R. et al. **TimeML Annotation Guidelines, Version 1.2.1**., 2006. Disponível em: <https://nilsreiter.de/assets/2017-10-01-howto-annotation/timeml-1.2.1.pdf>

SAVARY, A. et al. Literal Occurrences of Multiword Expressions: Rare Birds That Cause a Stir. **The Prague Bulletin of Mathematical Linguistics**, v. 112, p. 5–54, 2019b2019b.

SAVARY, A. et al. **PARSEME – parsing and multiword Expressions within a European multilingual network**. Proc. of LTC 2015. **Anais**...Poznań: 2015.

SAVARY, A. et al. **The PARSEME Shared Task on Automatic Identification of Verbal Multiword Expressions**. Proceedings of the 13th Workshop on MWEs. **Anais**...Valencia, Spain: ACL, 2017.

SAVARY, A. et al. PARSEME multilingual corpus of verbal multiword expressions. Em: MARKANTONATOU, S. et al. (Eds.). **Multiword expressions at length and in depth: Extended papers from the MWE 2017 workshop**. Phraseology e Multiword Expressions. Berlin, Germany: Language Science Press, 2018. v. 2.

SAVARY, A. et al. Object-oriented lexical encoding of multiword expressions: Short and sweet. **Lexique**, n. 27, p. 87–120, 2020.

SAVARY, A. et al. PARSEME Meets Universal Dependencies: Getting on the Same Page in Representing Multiword Expressions. **Northern European Journal of Language Technology**, v. 9, p. 14, b2023.

SAVARY, A. et al. **PARSEME corpus release 1.3**. Proceedings of the 19th Workshop on Multiword Expressions (MWE 2023). **Anais**...Dubrovnik, Croatia: Association for Computational Linguistics, a2023.

SAVARY, A.; CORDEIRO, S.; RAMISCH, C. **Without lexicons, multiword expression identification will never fly: A position statement**. Proceedings of the Joint Workshop on Multiword Expressions and WordNet (MWE-WN 2019). **Anais**...Florence, Italy: Association for Computational Linguistics, 2019a2019a. Disponível em: <https://aclanthology.org/W19-5110>

SCARTON, C. E.; ALUISIO, S. M. **Towards a cross-linguistic VerbNet-style lexicon for Brazilian portuguese**. Workshop on Creating Cross-language Resources for Disconnected Languages and Styles - CREDISLAS. **Anais**...ELRA, 2012.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

SCHEGLOFF, E. A. Overlapping talk and the organization of turn-taking for conversation. **Language in society**, v. 29, n. 1, p. 1–63, 2000.

SCHEGLOFF, E. A.; JEFFERSON, G.; SACKS, H. The preference for self-correction in the organization of repair in conversation. **Language**, v. 53, n. 2, p. 361–382, 1977.

SCHEGLOFF, E. A.; SACKS, H. Opening up closings. **Semiotica**, 1973.

SCHLANGEN, D. Language tasks and language games: On methodology in current natural language processing research. **arXiv preprint arXiv:1908.10747**, 2019.

SCHLANGEN, D. **Targeting the Benchmark: On Methodology in Current Natural Language Processing Research**. Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers). **Anais**...Online: Association for Computational Linguistics, ago. 2021. Disponível em: <https://aclanthology.org/2021.acl-short.85>

SCHLANGEN, D. **Norm Participation Grounds Language**. Proceedings of the 2022 CLASP Conference on (Dis)embodiment. **Anais**...Gothenburg, Sweden: Association for Computational Linguistics, set. 2022. Disponível em: <https://aclanthology.org/2022.clasp-1.7>

SCHLANGEN, D. **What A Situated Language-Using Agent Must be Able to Do: A Top-Down Analysis**., a2023. Disponível em: <https://arxiv.org/abs/2302.08590>

SCHLANGEN, D. Dialogue games for benchmarking language understanding: Motivation, taxonomy, strategy. **arXiv preprint arXiv:2304.07007**, b2023.

SCHLANGEN, D. **On General Language Understanding**. (H. Bouamor, J. Pino, K. Bali, Eds.)Findings of the Association for Computational Linguistics: EMNLP 2023. **Anais**...Singapore: Association for Computational Linguistics, dez. c2023. Disponível em: <https://aclanthology.org/2023.findings-emnlp.591>

SCHLANGEN, D.; SKANTZE, G. A general, abstract model of incremental dialogue processing. **Dialogue & Discourse**, v. 2, n. 1, p. 83–111, 2011.

SCHMID, H. **Part-of-Speech Tagging with Neural Networks**., 1994. Disponível em: <https://arxiv.org/abs/cmp-lg/9410018>

SCHNEIDER, N. et al. **SemEval-2016 Task 10: Detecting Minimal Semantic Units and their Meanings (DiMSUM)**. Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016). **Anais**...San Diego, California: Association for Computational Linguistics, 2016. Disponível em: <https://aclanthology.org/S16-1084>

SCHNEIDER, N.; SMITH, N. A. **A Corpus and Model Integrating Multiword Expressions and Supersenses**. Proceedings of the 2015 Conference of the North American

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Chapter of the Association for Computational Linguistics: Human Language Technologies. **Anais**...Denver, Colorado: Association for Computational Linguistics, 2015. Disponível em: <https://www.aclweb.org/anthology/N15-1177>

SCHONE, P.; JURAFSKY, D. **Is Knowledge-Free Induction of Multiword Unit Dictionary Headwords a Solved Problem?** (L. Lee, D. Harman, Eds.)Proceedings of the 2001 Conference on Empirical Methods in Natural Language Processing. **Anais**...2001.

SCHRÖDER, U. The interplay of verbal, vocal, and visual cues in the co-construction of the experience of alterity in exchange students' talk. **Journal of Pragmatics**, v. 81, p. 21–35, 2015.

SCHULTE IM WALDE, S. et al. **GhoSt-NN: A Representative Gold Standard of German Noun-Noun Compounds**. Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16). **Anais**...Portorož, Slovenia: European Language Resources Association (ELRA), 2016. Disponível em: <https://aclanthology.org/L16-1362>

SCHUSTER, M.; NAKAJIMA, K. **Japanese and Korean voice search**. 2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). **Anais**...2012.

SCHUSTER, M.; PALIWAL, K. K. Bidirectional recurrent neural networks. **IEEE transactions on Signal Processing**, v. 45, n. 11, p. 2673–2681, 1997.

SEARA, I. **Estudo Estatístico dos Fonemas do Português Brasileiro Falado na Capital de Santa Catarina para Elaboração de Frases Foneticamente Balanceadas**. tese de doutorado—[s.l.] Dissertação de Mestrado, Universidade Federal de Santa Catarina ..., 1994.

SECO, N. et al. **A Complex Evaluation Architecture for HAREM**. (R. Vieira et al., Eds.)Computational Processing of the Portuguese Language: 7th International Workshop, PROPOR 2006. **Anais**...Springer, 2006.

SELLAM, T.; DAS, D.; PARIKH, A. P. **BLEURT: Learning Robust Metrics for Text Generation**. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020. **Anais**...2020. Disponível em: <https://doi.org/10.18653/v1/2020.acl-main.704>

SELLARS, W. Inference and Meaning. **Mind**, v. 62, n. 247, p. 313–338, 1953.

SENNRICH, R.; HADDOW, B.; BIRCH, A. **Neural Machine Translation of Rare Words with Subword Units**. Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). **Anais**...Berlin, Germany: Association for Computational Linguistics, ago. 2016. Disponível em: <https://aclanthology.org/P16-1162>

SENO, E. R. M. **RHeSumaRST: um sumarizador automático de estruturas RST.**

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

mathesis—[s.l.] Universidade Federal de São Carlos, 2005.

SERBAN, I. et al. **Building end-to-end dialogue systems using generative hierarchical neural network models**. Proceedings of the AAAI conference on artificial intelligence. **Anais**...2016. Disponível em: <https://doi.org/10.48550/arXiv.1507.04808>

SERBAN, I. V. et al. A Survey of Available Corpora For Building Data-Driven Dialogue Systems: The Journal Version. **Dialogue & Discourse**, v. 9, n. 1, p. 1–49, 2018.

SERETAN, V. **Syntax-Based Collocation Extraction**. 1st. ed. Dordrecht, Netherlands: springer, 2011. v. 44

SERRA, C. R. **Realização e percepção de fronteiras prosódicas no português do Brasil: fala espontânea e leitura**. tese de doutorado—Rio de Janeiro: Universidade Federal do Rio de Janeiro, 2009.

SETTLES, B. **Active Learning Literature Survey**., 2010. Disponível em: <https://burrsettles.com/pub/settles.activelearning.pdf>

SHAPIRO, S. C. SNePS: A Logic for Natural Language Understanding and Commonsense Reasoning. Em: **Natural Language Processing and Knowledge Representation: Language for Knowledge and Knowledge for Language**. Cambridge, MA, USA: MIT Press, 2000. p. 175–195.

SHEN, J. et al. **Natural tts synthesis by conditioning wavenet on mel spectrogram predictions**. 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). **Anais**...IEEE, 2018.

SHIMORINA, A.; BELZ, A. **The Human Evaluation Datasheet: A Template for Recording Details of Human Evaluation Experiments in NLP**. Proceedings of the 2nd Workshop on Human Evaluation of NLP Systems (HumEval). **Anais**...Dublin, Ireland: Association for Computational Linguistics, 2022. Disponível em: <https://aclanthology.org/2022.humeval-1.6>

SHMUELI, B. et al. **Beyond Fair Pay: Ethical Implications of NLP Crowdsourcing**. Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. **Anais**...Online: Association for Computational Linguistics, jun. 2021. Disponível em: <https://aclanthology.org/2021.naacl-main.295>

SHRIBERG, E. **Preliminaries to a theory of speech disfluencies**. tese de doutorado—[s.l.] University of California at Berkele, 1994.

SHRIBERG, E. To "errrr" is human: ecology and acoustics of speech disfluencies. **Journal of the International Phonetic Association**, v. 31, n. 1, p. 153–169, 2001.

SHUMAILOV, I. et al. AI models collapse when trained on recursively generated data. **Nature**, v. 631, n. 8022, p. 755–759, 2024.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

SI, S. et al. Sentence Similarity Computation in Question Answering Robot. **Journal of Physics: Conference Series**, v. 1237, n. 2, p. 022093, jun. 2019.

SIDDHI, D.; VERGHESE, J. M.; BHAVIK, D. Survey on various methods of text to speech synthesis. **International Journal of Computer Applications**, v. 165, n. 6, 2017.

SIDNELL, J. Turn-continuation by self and by other. **Discourse Processes**, v. 49, n. 3-4, p. 314–337, 2012.

SIDNER, C. **A progress report on the discourse and reference components of PAL**. [s.l.] Massachusetts Institute of Tech Cambridge Artificial Intelligence LAB, 1978.

SILVA, E. DA; LATERZA, J.; FALEIROS, T. **New State-of-the-Art for Question Answering on Portuguese SQuAD v1.1**. Anais do X Symposium on Knowledge Discovery, Mining and Learning. **Anais**...Porto Alegre, RS, Brasil: SBC, a2022. Disponível em: <https://sol.sbc.org.br/index.php/kdmile/article/view/24974>

SILVA, E.; PARDO, T.; ROMAN, N. **Etiquetagem morfossintática multigênero para o português do Brasil segundo o modelo Üniversal Dependencieš**. Anais do XIV Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana. **Anais**...Porto Alegre, RS, Brasil: SBC, 2023. Disponível em: <https://sol.sbc.org.br/index.php/stil/article/view/25438>

SILVA, F. L. V. DA et al. ABSAPT 2022 at IberLEF: Overview of the Task on Aspect-Based Sentiment Analysis in Portuguese. **Procesamiento del Lenguaje Natural**, v. 69, p. 199–205, b2022.

SILVA, J. F. DA. **Resolução de correferência em múltiplos documentos utilizando aprendizado não supervisionado**. Dissertação de Mestrado, Universidade de São Paulo, 2011.

SIMMONS, R.; SLOCUM, J. Generating English Discourse from Semantic Networks. **Commun. ACM**, v. 15, n. 10, p. 891–905, out. 1972.

SIMÕES, A.; GUINOVART, X. G. **Bootstrapping a Portuguese WordNet from Galician, Spanish and English Wordnets**. IberSPEECH Conference. **Anais**...2014. Disponível em: <https://api.semanticscholar.org/CorpusID:10377782>

SINCLAIR, J. (ED.). **Collins COBUILD Dictionary of Phrasal Verbs**. London, UK: Collins COBUILD, 1989.

SINGH, P. et al. **Open Mind Common Sense: Knowledge Acquisition from the General Public**. (R. Meersman, Z. Tari, Eds.)On the Move to Meaningful Internet Systems 2002: CoopIS, DOA, and ODBASE. **Anais**...Berlin, Heidelberg: Springer Berlin Heidelberg, 2002.

SINGH, Y. B.; GOEL, S. A systematic literature review of speech emotion recognition

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

approaches. **Neurocomputing**, 2022.

SKANTZE, G. **Error Handling in Spoken Dialogue Systems: Managing Uncertainty, Grounding and Miscommunication**. tese de doutorado—[s.l.] KTH, 2007.

SKANTZE, G. Turn-taking in conversational systems and human-robot interaction: a review. **Computer Speech & Language**, v. 67, p. 101178, 2021.

SKANTZE, G.; DOĞRUÖZ, A. S. **The Open-domain Paradox for Chatbots: Common Ground as the Basis for Human-like Dialogue**. Proceedings of the 24th Meeting of the Special Interest Group on Discourse and Dialogue. **Anais**...Prague, Czechia: Association for Computational Linguistics, set. 2023. Disponível em: <https://aclanthology.org/2023.sigdial-1.57>

SMADJA, F. A. Retrieving Collocations from Text: Xtract. **cl**, v. 19, n. 1, p. 143–177, 1993.

SMILEY, C. et al. **When to Plummet and When to Soar: Corpus Based Verb Selection for Natural Language Generation**. Proceedings of the 9th International Natural Language Generation conference. **Anais**...: INLG'16.Edinburgh, UK: Association for Computational Linguistics, 2016. Disponível em: <http://anthology.aclweb.org/W16-6606>

SMITH, G.; RUSTAGI, I. **Mitigating Bias in Artificial Intelligence: An Equity Fluent Leadership Playbook**. [s.l.] Berkeley Haas Center for Equity, Gender; Leadership, 2020.

SØGAARD, A. et al. **What's in a p-value in NLP?** Proceedings of the Eighteenth Conference on Computational Natural Language Learning. **Anais**...Ann Arbor, Michigan: Association for Computational Linguistics, jun. 2014. Disponível em: <https://aclanthology.org/W14-1601>

SOHN, K. et al. **FixMatch: simplifying semi-supervised learning with consistency and confidence**. Proceedings of the 34th International Conference on Neural Information Processing Systems. **Anais**...: NIPS '20.Red Hook, NY, USA: Curran Associates Inc., 2020.

SOON, W. M.; NG, H. T.; LIM, C. Y. A Machine Learning Approach to Coreference Resolution of Noun Phrases. **Computational Linguistics**, v. 27, n. 4, p. 521–544, 2001.

SOUSA, A. G. DE et al. **Using a Domain Ontology to Bridge the Gap between User Intention and Expression in Natural Language Queries.** ICEIS (1). **Anais**...2020.

SOUSA, C. S. C.; ANDRADE, I. M.; ALMEIDA, T. G. DE. A Monopolização de uma conversa informal: Uma descrição dos movimentos de continua ção a partir da linguística sistêmico-funcional. **EntreLetras**, v. 13, n. 1, p. 158–183, 2022.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

SOUZA, B. B. DE. A interpretação de línguas de sinais como ação conjunta: uma análise da interação entre o intérprete de turno e o intérprete de apoio. **Trabalho de conclusão de curso. Universidade Federal de São Carlos**, 2021.

SOUZA, E. DE. **Construção e avaliação de um treebank padrão ouro**. Mestrado—[s.l.] PUC-Rio, 2023.

SOUZA, E. DE; FREITAS, C. **Explorando variações no tagset e na anotação Universal Dependencies (UD) para Português: Possibilidades e resultados com base no treebank PetroGold**. Anais do XIV Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana. **Anais**...Association for Computational Linguistics, 2023.

SOUZA, F.; NOGUEIRA, R.; LOTUFO, R. **BERTimbau: pretrained BERT models for Brazilian Portuguese**. (R. Cerri, R. C. Prati, Eds.)Proceedings of the 2020 Brazilian Conference on Intelligent Systems. **Anais**...Springer International Publishing, 2020.

SOUZA, J. W. DA C. **Descrição linguística da complementaridade para a sumarização automática multidocumento**. mathesis—[s.l.] Universidade Federal de São Carlos, 2015.

SOUZA, J. W. DA C. **Aprofundamento da caracterização linguístico-computacional da complementaridade em um corpus jornalístico multidocumento**. tese de doutorado—[s.l.] (Doutorado em Linguística) - Programa de Pós-Graduação em Linguística, Universidade Federal de São Carlos, 2019.

SPARCK JONES, K. **Towards Better NLP System Evaluation**. Human Language Technology: Proceedings of a Workshop held at Plainsboro, New Jersey, March 8-11, 1994. **Anais**...1994. Disponível em: <https://aclanthology.org/H94-1018>

SPARCK JONES, K. Natural language processing: a historical review. Em: **Current issues in computational linguistics: in honour of Don Walker**. [s.l.] Springer, 2001. p. 3–16.

SPARCK JONES, K.; GALLIERS, J. R. Evaluating Natural Language Processing Systems: An Analysis and Review. **Lecture Notes in Computer Science**, 1995.

SPEER, R.; CHIN, J.; HAVASI, C. ConceptNet 5.5: An Open Multilingual Graph of General Knowledge. **CoRR**, v. abs/1612.03975, 2016.

SPINDOLA, S. et al. **Interpretability of Attention Mechanisms in a Portuguese-Based Question Answering System about the Blue Amazon**. Anais do XVIII Encontro Nacional de Inteligência Artificial e Computacional. **Anais**...Porto Alegre, RS, Brasil: SBC, 2021. Disponível em: <https://sol.sbc.org.br/index.php/eniac/article/view/18302>

SRIPADA, S.; GAO, F. **Summarizing Dive Computer Data: A Case Study in Integrating Textual and Graphical Presentations of Numerical Data**. Workshop

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

on Multimodal Output Generation. **Anais**...: MOG'07.Association for Computational Linguistics, 2007.

SRIPADA, S.; REITER, E.; DAVY, I. SumTime-Mousam: Configurable marine weather forecast generator. **Expert Update**, v. 6, n. 3, p. 4–10, fev. 2004.

STAB, C. et al. **Argumentation Mining in Persuasive Essays and Scientific Articles from the Discourse Structure Perspective.** ArgNLP. **Anais**...2014.

STEIN, F. O dispositivo para o gerenciamento de sobreposições de vozes na conversa cotidiana em português brasileiro. **Salão de Iniciação Científica (22.: 2010 out. 18-22: Porto Alegre, RS). Livro de resumos. Porto Alegre: UFRGS.**, 2010.

STEVENS, S. S. A Scale for the Measurement of the Psychological Magnitude Pitch. **Acoustical Society of America Journal**, v. 8, n. 3, p. 185, jan. 1937.

STIVERS, T. Sequence Organization. Em: **The handbook of conversation analysis**. [s.l.] Wiley Online Library, 2013. v. 191.

STYMNE, S.; CANCEDDA, N.; AHRENBERG, L. Generation of Compound Words in Statistical Machine Translation into Compounding Languages. **Computational Linguistics**, p. 1——42, 2013.

SUCHANEK, F. M.; KASNECI, G.; WEIKUM, G. **Yago: a core of semantic knowledge**. Proceedings of the 16th international conference on World Wide Web. **Anais**...2007.

SUNDAR, A.; HECK, L. **Multimodal Conversational AI: A Survey of Datasets and Approaches**. Proceedings of the 4th Workshop on NLP for Conversational AI. **Anais**...Dublin, Ireland: Association for Computational Linguistics, 2022. Disponível em: <https://aclanthology.org/2022.nlp4convai-1.12>

SUNKARA, M. et al. **Multimodal Semi-Supervised Learning Framework for Punctuation Prediction in Conversational Speech**. Proc. Interspeech 2020. **Anais**...2020.

SUNKARA, M. et al. Neural Inverse Text Normalization. **CoRR**, v. abs/2102.06380, 2021.

SUTSKEVER, I.; VINYALS, O.; LE, Q. V. **Sequence to Sequence Learning with Neural Networks**. (Z. Ghahramani et al., Eds.)Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada. **Anais**...2014. Disponível em: <https://proceedings.neurips.cc/paper/2014/hash/a14ac55a4f27472c5d894ec1c3c743d2-Abstract.html>

TABOADA, M.; MANN, W. C. Rhetorical structure theory: Looking back and moving ahead. **Discourse studies**, v. 8, n. 3, p. 423–459, 2006.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

TACHIBANA, H.; UENOYAMA, K.; AIHARA, S. Efficiently Trainable Text-to-Speech System Based on Deep Convolutional Networks with Guided Attention. **arXiv preprint arXiv:1710.08969**, 2017.

TAN, L.; PAL, S. **Manawi: Using Multi-Word Expressions and Named Entities to Improve Machine Translation**. Proceedings of the 14th Machine Translation Summint. Workshop on Multi-word units in Machine Translation and Translation Technologies. **Anais**...2014.

TAN, X. et al. A survey on neural speech synthesis. **arXiv preprint arXiv:2106.15561**, 2021.

TASLIMIPOOR, S.; ROHANIAN, O.; HA, L. A. **Cross-lingual Transfer Learning and Multitask Learning for Capturing Multiword Expressions**. Proceedings of the Joint Workshop on Multiword Expressions and WordNet (MWE-WN 2019). **Anais**...Florence, Italy: Association for Computational Linguistics, ago. 2019. Disponível em: <https://aclanthology.org/W19-5119>

TAYYAR MADABUSHI, H. et al. **AStitchInLanguageModels: Dataset and Methods for the Exploration of Idiomaticity in Pre-Trained Language Models**. Findings of the Association for Computational Linguistics: EMNLP 2021. **Anais**...Punta Cana, Dominican Republic: Association for Computational Linguistics, nov. 2021. Disponível em: <https://aclanthology.org/2021.findings-emnlp.294>

TAYYAR MADABUSHI, H. et al. **SemEval-2022 Task 2: Multilingual Idiomaticity Detection and Sentence Embedding**. Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022). **Anais**...Seattle, United States: Association for Computational Linguistics, jul. 2022. Disponível em: <https://aclanthology.org/2022.semeval-1.13>

TEDESCHI, S. et al. **What's the Meaning of Superhuman Performance in Today's NLU?** Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). **Anais**...Toronto, Canada: Association for Computational Linguistics, jul. 2023. Disponível em: <https://aclanthology.org/2023.acl-long.697>

TEIXEIRA, A. L. R. et al. **DaMata: A robot-journalist covering the Brazilian Amazon deforestation**. Proceedings of the 13th International Conference on Natural Language Generation. **Anais**...2020.

TEIXEIRA, B. H. F. **Detecção automática de fronteiras prosódicas na fala espontânea**. tese de doutorado—Belo Horizonte: Universidade Federal de Minas Gerais, 2022.

TEIXEIRA, B. H. F.; MITTMAN, M. M. Acoustic Models for the Automatic Identification of Prosodic Boundaries in Spontaneous Speech. **Revista de Estudos da Linguagem**, v. 26, n. 4, p. 1455–1488, 2018.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

TEIXEIRA, B.; BARBOSA, P.; RASO, T. **Automatic Detection of Prosodic Boundaries in Brazilian Portuguese Spontaneous Speech**. (A. Villavicencio et al., Eds.)Computational Processing of the Portuguese Language. **Anais**...Cham: Springer International Publishing, 2018.

TEIXEIRA, J. P. et al. **Phonetic Events from the Labeling the European Portuguese DataBase for Speech Synthesis, FEUP/IPBDB**. Seventh European Conference on Speech Communication and Technology. **Anais**...2001.

TEIXEIRA, J. P.; FREITAS, D.; FUJISAKI, H. **Prediction of Fujisaki model's phrase commands**. Eighth European Conference on Speech Communication and Technology. **Anais**...2003.

TENNANT, H. R. **Evaluation of Natural Language Processors**. tese de doutorado—[s.l.] University of Illinois Urbana-Champaign, 1980.

TESCH, L. M. O uso de digressões em textos orais. **Filologia e Linguística Portuguesa**, v. 17, n. 2, p. 273–293, 2015.

TESNIÈRE, L. **Eléments de Syntaxe Structurale**. Paris: Klincksieck, 1959.

THAKKAR, M.; PISE, N. **Survey of Available Datasets for Designing Task Oriented Dialogue Agents**. 2019 International Conference on Mechatronics, Remote Sensing, Information Systems and Industrial Information Technologies (ICMRSISIIT). **Anais**...2019. Disponível em: <https://doi.org/10.1109/ICMRSISIIT46373.2020.9405898>

THEUNE, M. et al. From data to speech: a general approach. **Natural Language Engineering**, v. 7, n. 1, p. 47–86, 2001.

THOMAS, R. L.; UMINSKY, D. Reliance on metrics is a fundamental challenge for AI. **Patterns**, v. 3, n. 5, 2022.

TJOA, E.; GUAN, C. A Survey on Explainable Artificial Intelligence (XAI): Toward Medical XAI. **IEEE Transactions on Neural Networks and Learning Systems**, v. 32, n. 11, p. 4793–4813, 2021.

TOKUDA, K. et al. **Speech parameter generation algorithms for HMM-based speech synthesis**. 2000 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No. 00CH37100). **Anais**...IEEE, 2000.

TOMASELLO, M. The usage-based theory of language acquisition. Em: BAVIN, E. L.; NAIGLES, L. R. E. (Eds.). **The Cambridge Handbook of Child Language**. Cambridge Handbooks em Language e Linguistics. 2. ed. [s.l.] Cambridge University Press, 2015. p. 89–106.

TORRENT, T. T. et al. **Copa 2014 FrameNet Brasil: a frame-based trilingual electronic dictionary for the Football World Cup**. Proceedings of COLING 2014,

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

the 25th International Conference on Computational Linguistics: System Demonstrations. **Anais**...Dublin, Ireland: Dublin City University; Association for Computational Linguistics, ago. 2014. Disponível em: <https://aclanthology.org/C14-2003>

TORRENT, T. T.; ELLSWORTH, M. Behind the Labels: Criteria for Defining Analytical Categories in FrameNet Brasil. **Veredas-Revista de Estudos Linguisticos**, v. 17, n. 1, p. 44–66, 2013.

TOSCANO, M. E. S. As relações interpessoais e a correção na língua falada. **Cadernos de Linguagem e Sociedade**, v. 5, p. 119–119, 2001.

TRANCOSO, I. et al. Corpus de diálogo CORAL. **PROPOR'98**, 1998.

TRUESWELL, J. C.; TANENHAUS, M. K. **Approaches to studying world-situated language use: Bridging the language-as-product and language-as-action traditions**. [s.l.] MIT Press, 2005.

TSVETKOV, Y.; WINTNER, S. **Identification of Multi-word Expressions by Combining Multiple Linguistic Information Sources**. Proceedings of the Conference on Empirical Methods in Natural Language Processing. **Anais**...: EMNLP '11.Stroudsburg, PA, USA: Association for Computational Linguistics, 2011.

TSVETKOV, Y.; WINTNER, S. Extraction of multi-word expressions from small parallel corpora. **Natural Language Engineering**, v. 18, n. 04, p. 549–573, 2012.

TURNER, R. et al. **Generating Spatio-temporal Descriptions in Pollen Forecasts**. Proceedings of the Eleventh Conference of the European Chapter of the Association for Computational Linguistics: Demonstrations. **Anais**...: EACL'06.Trento, Italy: Association for Computational Linguistics, 2006. Disponível em: <http://dl.acm.org/citation.cfm?id=1608974.1608998>

ULMER, D. et al. **Experimental Standards for Deep Learning in Natural Language Processing Research**. Findings of the Association for Computational Linguistics: EMNLP 2022. **Anais**...Abu Dhabi, United Arab Emirates: Association for Computational Linguistics, dez. 2022. Disponível em: <https://aclanthology.org/2022.findings-emnlp.196>

UNESCO. **Beijing consensus on artificial intelligence and education**. UNESCO Paris, 2019.

UNESCO, D. G. **Recomendação sobre a Ética da Inteligência Artificial**. Disponível em: <https://unesdoc.unesco.org/ark:/48223/pf0000381137_por >. Acesso em: 28 ago. 2023.

UNICEF. **Declaração Universal dos Direitos Humanos**. Disponível em: <https://www.unicef.org/brazil/declaracao-universal-dos-direitos-humanos>. Acesso em: 28 ago. 2023.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

UZÊDA, V. R.; PARDO, T. A. S.; NUNES, M. G. V. A comprehensive comparative evaluation of RST-based summarization methods. **ACM Transactions on Speech and Language Processing (TSLP)**, v. 6, n. 4, p. 1–20, 2010.

VALLE, R. et al. Flowtron: an Autoregressive Flow-based Generative Network for Text-to-Speech Synthesis. **arXiv preprint arXiv:2005.05957**, 2020.

VAN DEEMTER, K. et al. Toward a computational psycholinguistics of reference production. **Topics in cognitive science**, v. 4, n. 2, p. 166–183, 2012.

VAPNIK, V. N. **Statistical Learning Theory**. [s.l.] Wiley, 1998.

VAPNIK, V. N. **The Nature of Statistical Learning Theory**. 2. ed. New York, NY: Springer, 2000. p. 314

VARGAS, F. A.; PARDO, T. A. S. **Aspect clustering methods for sentiment analysis**. Proceedings of International conference on computational processing of the Portuguese language. **Anais**...Springer, 2018.

VASWANI, A. et al. **Attention is All you Need**. (I. Guyon et al., Eds.)Advances in Neural Information Processing Systems. **Anais**...Curran Associates, Inc., 2017. Disponível em: <https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html>

VEAUX, C. et al. CSTR VCTK corpus: English multi-speaker corpus for CSTR voice cloning toolkit. **University of Edinburgh. The Centre for Speech Technology Research (CSTR)**, 2017.

VERHAGEN, M. et al. **SemEval-2010 Task 13: TempEval-2**. Proceedings of the 5th International Workshop on Semantic Evaluation, SemEval. **Anais**...2010. Disponível em: <https://www.aclweb.org/anthology/S10-1010.pdf>

VIEIRA, F. E.; FARACO, C. A. **Texto e discurso. Escrever na universidade.** [s.l.] Parábola, 2019.

VIEIRA, R. et al. Coreference and anaphoric relations of demonstrative noun phrases in multilingual corpus. **Anaphora Processing: linguistic, cognitive and computational modeling**, p. 385–403, 2005.

VIEIRA, R.; GONÇALVES, P. N.; SOUZA, J. G. C. DE. Processamento computacional de anáfora e correferência. **Revista de Estudos da Linguagem**, v. 16, n. 1, 2012.

VIETHEN, J.; DALE, R. **GRE3D7: A Corpus of Distinguishing Descriptions for Objects in Visual Scenes**. UCNLG+Eval: Language Generation and Evaluation Workshop. **Anais**...Edinburgh, UK: Association for Computational Linguistics, 2011.

VILAIN, M. et al. **A model-theoretic coreference scoring scheme**. Proceedings of the 6th Message Understanding Conference (MUC-6). **Anais**...Los Altos, CA, EUA:

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

Morgan Kaufmann, 1995. Disponível em: <http://acl.ldc.upenn.edu/M/M95/M95-1005.pdf>

VINCIARELLI, A. et al. Open challenges in modelling, analysis and synthesis of human behaviour in human–human and human–machine interactions. **Cognitive Computation**, v. 7, p. 397–413, 2015.

VINCZE, V.; NAGY T., I.; BEREND, G. **Multiword Expressions and Named Entities in the Wiki50 Corpus**. Proceedings of the International Conference Recent Advances in Natural Language Processing 2011. **Anais**...Hissar, Bulgaria: Association for Computational Linguistics, set. 2011. Disponível em: <https://aclanthology.org/R11-1040>

VINCZE, V.; NAGY T., I.; FARKAS, R. **Identifying English and Hungarian Light Verb Constructions: A Contrastive Approach**. Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). **Anais**...Sofia, Bulgaria: Association for Computational Linguistics, ago. 2013. Disponível em: <https://aclanthology.org/P13-2046>

VINYALS, O.; LE, Q. **A Neural Conversational Model**., 2015. Disponível em: <https://doi.org/10.48550/arXiv.1506.05869>

VOGEL, L. H. **Um olhar para além do verbo: os usos do olho na fala-em-interação**. Universidade Federal do Rio Grande do Sul, 2018.

VOORHEES, E. M.; TICE, D. M. **Building a Question Answering Test Collection**. Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. **Anais**...2000. Disponível em: <https://dl.acm.org/doi/10.1145/345508.345577>

VRANDEČIĆ, D.; KRÖTZSCH, M. Wikidata: a free collaborative knowledgebase. **Communications of the ACM**, v. 57, n. 10, p. 78–85, 2014.

WAGNER, J. et al. Dawn of the transformer era in speech emotion recognition: closing the valence gap. **IEEE Transactions on Pattern Analysis and Machine Intelligence**, 2023.

WAGNER, P.; MALISZ, Z.; KOPP, S. **Gesture and speech in interaction: An overview**. **Speech Communication**Elsevier, 2014. Disponível em: <https://doi.org/10.1016/j.specom.2013.09.008>

WAGSTAFF, K. L. **Machine learning that matters**. Proceedings of the 29th International Coference on International Conference on Machine Learning. **Anais**...2012. Disponível em: <https://doi.org/10.48550/arXiv.1206.4656>

WALKER, M. A. et al. **PARADISE: A Framework for Evaluating Spoken Dialogue Agents**. 35th Annual Meeting of the Association for Computational Linguistics and 8th Conference of the European Chapter of the Association for Computational Linguistics. **Anais**...Madrid, Spain: Association for Computational Linguistics, jul. 1997. Disponível

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

em: <https://aclanthology.org/P97-1035>

WALLIS, S. Completing Parsed Corpora. Em: ABEILLÉ, A. (Ed.). **Treebanks: Building and Using Parsed Corpora**. Dordrecht: Springer Netherlands, 2003. p. 61–71.

WALTER, E. (ED.). **Cambridge Idioms Dictionary**. 2. ed. Cambridge, UK: campress, 2006.

WANG, A. et al. **GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding**. Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP. **Anais**...Brussels, Belgium: Association for Computational Linguistics, nov. 2018. Disponível em: <https://aclanthology.org/W18-5446/>

WANG, A. et al. SuperGLUE: A Stickier Benchmark for General-Purpose Language Understanding Systems. **Advances in Neural Information Processing Systems**, v. 32, p. 3261–3275, 2019.

WANG, C. et al. Covost: A diverse multilingual speech-to-text translation corpus. **arXiv preprint arXiv:2002.01320**, a2020.

WANG, C. et al. Voxpopuli: A large-scale multilingual speech corpus for representation learning, semi-supervised learning and interpretation. **arXiv preprint arXiv:2101.00390**, 2021.

WANG, C. et al. Neural Codec Language Models are Zero-Shot Text to Speech Synthesizers. **arXiv preprint arXiv:2301.02111**, a2023.

WANG, C.; WU, A.; PINO, J. Covost 2 and massively multilingual speech-to-text translation. **arXiv preprint arXiv:2007.10310**, b2020.

WANG, W. Y.; GEORGILA, K. **Automatic detection of unnatural word-level segments in unit-selection speech synthesis**. 2011 IEEE Workshop on Automatic Speech Recognition & Understanding. **Anais**...IEEE, 2011.

WANG, Y. et al. Tacotron: A fully end-to-end text-to-speech synthesis model. **arXiv preprint arXiv:1703.10135**, 2017.

WANG, Y. et al. **DSPM-NLG: A Dual Supervised Pre-trained Model for Few-shot Natural Language Generation in Task-oriented Dialogue System**. Findings of the Association for Computational Linguistics: ACL 2023. **Anais**...b2023.

WANI, T. M. et al. A comprehensive review of speech emotion recognition systems. **IEEE Access**, v. 9, p. 47795–47814, 2021.

WATSON, D. The rhetoric and reality of anthropomorphism in artificial intelligence. **Minds and Machines**, v. 29, n. 3, p. 417–440, 2019.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

WEN, T.-H. et al. **Semantically Conditioned LSTM-based Natural Language Generation for Spoken Dialogue Systems**. Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. **Anais**...: EMNLP'15.Lisbon, Portugal: Association for Computational Linguistics, 2015. Disponível em: <http://aclweb.org/anthology/D15-1199>

WEN, T.-H. et al. **Multi-domain Neural Network Language Generation for Spoken Dialogue Systems**. Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. **Anais**...: HLT-NAACL'16.San Diego, California: Association for Computational Linguistics, 2016. Disponível em: <https://aclanthology.info/pdf/N/N16/N16-1015.pdf>

WIELING, M.; RAWEE, J.; NOORD, G. VAN. Squib: Reproducibility in Computational Linguistics: Are We Willing to Share? **Computational Linguistics**, v. 44, n. 4, p. 641–649, dez. 2018.

WIGHTMAN, C. W.; OSTENDORF, M. Automatic recognition of prosodic phrases. **[Proceedings] ICASSP 91: 1991 International Conference on Acoustics, Speech, and Signal Processing**, v. 1, p. 321–324, 1991.

WILKENS, R. et al. **LexSubNC: A Dataset of Lexical Substitution for Nominal Compounds**. Proceedings of the 12th International Conference on Computational Semantics (IWCS 2017). **Anais**...Montpellier, France: 2017.

WILKS, Y. Is Word Sense Disambiguation Just One More NLP Task? **Computers and the Humanities**, v. 34, n. 1-2, p. 235–243, 2000.

WILLIAMS, I. et al. **Contextual speech recognition in end-to-end neural network systems using beam search**. 2018. Disponível em: <https://www.isca-speech.org/archive/Interspeech_2018/pdfs/2416.pdf>

WILLIAMS, J. D.; RAUX, A.; HENDERSON, M. The dialog state tracking challenge series: A review. **Dialogue & Discourse**, v. 7, n. 3, p. 4–33, 2016.

WILLRICH, R.; SANTOS, D. Avaliação no DIP. **Linguamática**, v. 15, n. 1, p. 69–87, 2023.

WILSON, T. P.; ZIMMERMAN, D. H. The structure of silence between turns in two-party conversation. **Discourse processes**, v. 9, n. 4, p. 375–390, 1986.

XIE, Z.; COHN, T.; LAU, J. H. **The Next Chapter: A Study of Large Language Models in Storytelling**., 2023. Disponível em: <https://arxiv.org/abs/2301.09790>

YANG, F.; HEEMAN, P. A.; KUN, A. L. An Investigation of Interruptions and Resumptions in Multi-Tasking Dialogues. **Computational Linguistics**, v. 37, n. 1, p. 75–104, mar. a2011.

YANG, J.-H. et al. **Enriching Mandarin speech recognition by incorporating a**

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

hierarchical prosody model. 2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). **Anais**...b2011. Disponível em: <https://doi.org/10.1109/ICASSP.2011.5947492>

YANG, M. et al. **Learning ASR pathways: A sparse multilingual ASR model**., 2023. Disponível em: <https://arxiv.org/abs/2209.05735>

YANG, X. et al. **An Entity-Mention Model for Coreference Resolution with Inductive Logic Programming.** Proceeding of Association for Computational Linguistics. **Anais**...2008.

YAO, S. et al. **ReAct: Synergizing Reasoning and Acting in Language Models**., 2023. Disponível em: <https://arxiv.org/abs/2210.03629>

YAZDANI, M.; FARAHMAND, M.; HENDERSON, J. **Learning Semantic Composition to Detect Non-compositionality of Multiword Expressions**. Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. **Anais**...Lisbon, Portugal: Association for Computational Linguistics, set. 2015. Disponível em: <https://aclanthology.org/D15-1201>

YEH, Y.-T.; ESKENAZI, M.; MEHRI, S. **A Comprehensive Assessment of Dialog Evaluation Metrics**. The First Workshop on Evaluations and Assessments of Neural Conversation Systems. **Anais**...Online: Association for Computational Linguistics, nov. 2021. Disponível em: <https://aclanthology.org/2021.eancs-1.3>

YI, J.; TAO, J. Self-attention Based Model for Punctuation Prediction Using Word and Speech Embeddings. **ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)**, p. 7270–7274, 2019.

YNGVE, V. H. **Random generation of English sentences**. [s.l.] Massachusetts Inst. of Technology, 1961.

YNGVE, V. H. **On getting a word in edgewise**. Papers from the sixth regional meeting Chicago Linguistic Society, April 16-18, 1970, Chicago Linguistic Society, Chicago. **Anais**...1970.

YU, J. et al. Choosing the content of textual summaries of large time-series data sets. **Natural Language Engineering**, v. 13, n. 1, p. 25–49, 2007.

ZAMPIERI, N.; ILLINA, I.; FOHR, D. **Multiword Expression Features for Automatic Hate Speech Detection**. (E. Métais et al., Eds.)Natural Language Processing and Information Systems - 26th International Conference on Applications of Natural Language to Information Systems, NLDB 2021, Saarbrücken, Germany, June 23-25, 2021, Proceedings. **Anais**...: Lecture Notes em Computer Science.Springer, 2021. Disponível em: <https://doi.org/10.1007/978-3-030-80599-9/_14>

ZANINELLO, A.; BIRCH, A. **Multiword Expression aware Neural Machine Translation**. Proceedings of the 12th Language Resources and Evaluation Conference.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Anais...Marseille, France: European Language Resources Association, 2020. Disponível em: <https://aclanthology.org/2020.lrec-1.471>

ZE, H.; SENIOR, A.; SCHUSTER, M. **Statistical parametric speech synthesis using deep neural networks**. 2013 ieee international conference on acoustics, speech and signal processing. **Anais**...IEEE, 2013.

ZELASKO, P. et al. **Punctuation Prediction Model for Conversational Speech**. (B. Yegnanarayana, Ed.)Interspeech 2018, 19th Annual Conference of the International Speech Communication Association, Hyderabad, India, 2-6 September 2018. **Anais**...ISCA, 2018. Disponível em: <https://doi.org/10.21437/Interspeech.2018-1096>

ZEMAN, D. **Reusable Tagset Conversion Using Tagset Drivers**. Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08). **Anais**...Marrakech, Morocco: European Language Resources Association (ELRA), 2008. Disponível em: <http://www.lrec-conf.org/proceedings/lrec2008/pdf/66_paper.pdf>

ZEMAN, D.; RESNIK, P. **Cross-Language Parser Adaptation between Related Languages**. Proceedings of the IJCNLP-08 Workshop on NLP for Less Privileged Languages. **Anais**...2008. Disponível em: <https://aclanthology.org/I08-3008>

ZEN, H. et al. LibriTTS: A Corpus Derived from LibriSpeech for Text-to-Speech. **Proc. Interspeech 2019**, p. 1526–1530, 2019.

ZEWDU, A.; YITAGESU, B. Part of speech tagging: a systematic review of deep learning and machine learning approaches. **Journal of Big Data**, v. 9, jan. 2022.

ZHANG, T. et al. **BERTScore: Evaluating Text Generation with BERT**. 8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020. **Anais**...OpenReview.net, 2020. Disponível em: <https://openreview.net/forum?id=SkeHuCVFDr>

ZHANG, Z.; STRUBELL, E.; HOVY, E. **A Survey of Active Learning for Natural Language Processing**. (Y. Goldberg, Z. Kozareva, Y. Zhang, Eds.)Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing. **Anais**...Abu Dhabi, United Arab Emirates: Association for Computational Linguistics, dez. 2022. Disponível em: <https://aclanthology.org/2022.emnlp-main.414/>

ZHAO, J. et al. **Gender Bias in Coreference Resolution: Evaluation and Debiasing Methods**. Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers). **Anais**...New Orleans, Louisiana: Association for Computational Linguistics, jun. 2018. Disponível em: <https://aclanthology.org/N18-2003>

ZHOU, N. et al. CDGAN-BERT: Adversarial constraint and diversity discriminator for semi-supervised text classification. **Knowledge-Based Systems**, v. 284, p. 111291, 2024.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

ZHU, X. et al. **Introduction to Semi-Supervised Learning**. [s.l.] Morgan; Claypool Publishers, 2009.

ZIN, K. K. **Hidden Markov model with rule based approach for part of speech tagging of Myanmar language**. International Conference on Intelligent Cloud Computing. **Anais**...2009. Disponível em: <https://api.semanticscholar.org/CorpusID:63473605>

ZOBEL, J. **How reliable are the results of large-scale information retrieval experiments?** Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval. **Anais**...ACM, 1998.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Apêndice A

Alguns pressupostos para o processamento de fala

Apêndice do Capítulo 2

Camila de Araújo Azevedo
Heliana Ribeiro de Mello
Priscila Osório Côrtes

Publicado em: 26/09/2023

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

A.1 Alguns pressupostos:[1] estatística[2], probabilidade, teoria da informação

Aleatoriedade e incerteza desempenham um papel importante em muitas disciplinas científicas. A maioria dos problemas de processamento de linguagem falada pode ser caracterizada em um contexto probabilístico. A teoria da probabilidade e estatística fornecem a linguagem matemática para descrever e analisar tais sistemas.

Os critérios e métodos usados para estimar as probabilidades desconhecidas e densidades de probabilidade formam a base para a teoria da estimativa. A teoria da estimativa forma as bases para a aprendizagem de parâmetros no reconhecimento de padrões.

O teste de significância também é importante em estatística, que lida com a confiança da inferência estatística, como saber se a estimativa de algum parâmetro pode ser aceita com confiança. No reconhecimento de padrões, o teste de significância é extremamente importante para determinar se a diferença observada entre dois classificadores diferentes é real.

A teoria da informação foi originalmente desenvolvida para sistemas de comunicação eficientes e confiáveis. Ela evoluiu para uma teoria matemática preocupada com a essência do processo de comunicação. Ela fornece uma estrutura para o estudo de questões fundamentais, como a eficiência da representação da informação e as limitações na transmissão confiável de informação através de um canal de comunicação. Muitos desses problemas são fundamentais para o processamento de fala.

Abordamos brevemente nesta seção algumas dessas questões, a fim de fornecer um panorama de conhecimentos fundamentais para o profissional que trabalha com o processamento da fala.

[1] Esta seção é amplamente baseada nas discussões em (Gries, 2019; Jurafsky; Martin, 2023)

[2] Recomendamos o ambiente de programação R para a computação estatística. Para acessá-lo, cf. An Introduction to R (r-project.org) (https://cran.r-project.org/doc/manuals/R-intro.html)

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

A.1.1 Probabilidade

A ideia de incerteza e probabilidade remonta a cerca de 3500 a.C., quando jogos de azar com objetos ósseos foram desenvolvidos no Egito. Dados cúbicos com marcações virtualmente idênticas aos dados modernos foram encontrados em túmulos egípcios datados de aproximadamente 2000 a.C. O jogo de dados desempenhou um papel importante no desenvolvimento inicial da teoria da probabilidade. A teoria matemática moderna da probabilidade acredita-se ter sido iniciada pelos matemáticos franceses Blaise Pascal (1623-1662) e Pierre Fermat (1601-1665) quando eles trabalharam em certos problemas de jogo envolvendo dados. O matemático inglês Thomas Bayes (1702-1761) foi o primeiro a usar a probabilidade indutivamente e estabeleceu uma base matemática para a inferência de probabilidade, levando ao que é agora conhecido como teorema de Bayes. A teoria da probabilidade tem se desenvolvido constantemente desde então e tem sido amplamente aplicada em diversos campos de estudo.

A teoria da probabilidade lida com as médias de fenômenos em massa que ocorrem sequencial ou simultaneamente. Frequentemente usamos expressões probabilísticas em nosso dia a dia, como quando dizemos: "É muito provável que o Dow (índice Dow Jones Industrial) atinja 12.000 pontos no próximo mês", ou "A chance de chuvas dispersas em Seattle neste fim de semana é alta". Cada uma dessas expressões se baseia no conceito de probabilidade, ou seja, a probabilidade de que algum evento específico ocorra.

A probabilidade pode ser usada para representar o grau de confiança no resultado de algumas ações (observações) que não são definitivas. Na teoria da probabilidade, o termo espaço amostral, S, é usado para se referir à coleção (conjunto) de todos os possíveis resultados. Um evento refere-se a um subconjunto do espaço amostral ou uma coleção de resultados. A probabilidade do evento A, denotada como P(A), pode ser interpretada como a frequência relativa com que o evento A ocorreria se o processo fosse repetido um grande número de vezes sob condições semelhantes. Com base nessa interpretação, P(A) pode ser calculado simplesmente contando o número total, S_N, de todas as observações e o número de observações A_N cujo resultado pertence ao evento A. Ou seja,

$$P(A) = \frac{N_A}{N_S}$$

P(A) está limitado entre zero e um, ou seja, $0 \leq P(A) \leq 1$ para todos os A.

O limite inferior da probabilidade P(A) é zero quando o conjunto de eventos A é um conjunto vazio. Por outro lado, o limite superior da probabilidade P(A) é um quando o conjunto de eventos A acontece de ser S.

A.1.2 Variáveis aleatórias

Os elementos em um espaço amostral podem ser numerados e referidos pelos números atribuídos. Uma variável X que especifica a quantidade numérica em um espaço amostral é chamada de variável aleatória. Portanto, uma variável aleatória X é uma função que mapeia cada resultado possível s no espaço amostral S para números reais X(s). Como cada evento é um subconjunto do espaço amostral, um evento é representado como um conjunto de {s} que satisfaz $\{s \mid X\ (s) = x\}$. Usamos letras maiúsculas para denotar variáveis aleatórias e letras minúsculas para denotar valores fixos da variável aleatória. Assim, a probabilidade de X = x é denotada como:

$P(X = x) = P(s \mid X\ (s) = x)$

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Uma variável aleatória X é uma variável aleatória discreta, ou X tem uma distribuição discreta, se X pode assumir apenas um número finito n de valores diferentes x_1, x_2, ..., x_n, ou no máximo, uma sequência infinita de valores diferentes x_1, x_2, ... Se a variável aleatória X é uma variável aleatória discreta, a função de probabilidade (f.p.) ou função de massa de probabilidade (f.m.p.) de X é definida como a função p tal que, para qualquer número real x:

$Px^{(x)} = P(X = x)$

A.1.3 Média e Variância

A média é uma medida de tendência central que representa o valor médio de um conjunto de dados. É calculada somando todos os valores do conjunto e dividindo pelo número total de valores. A média é útil para determinar um valor representativo do conjunto de dados e fornecer uma estimativa de seu centro.

A variância, por outro lado, é uma medida de dispersão que quantifica a variabilidade ou a dispersão dos valores em relação à média. Ela indica o quão afastados os valores individuais estão da média. A variância é calculada encontrando a diferença entre cada valor e a média, elevando ao quadrado essas diferenças, somando-as e dividindo pelo número total de valores. A variância fornece uma ideia da dispersão dos dados e é utilizada para avaliar a variabilidade de um conjunto de dados.

A.1.3.1 A Lei dos Grandes Números

A Lei dos Grandes Números é um conceito fundamental na estatística que descreve o comportamento dos resultados médios de uma sequência de experimentos aleatórios. Ela estabelece que, à medida que o número de observações aumenta, a média dessas observações se aproxima do valor esperado teórico ou verdadeiro.

Em termos mais simples, a Lei dos Grandes Números afirma que, quando repetimos um experimento aleatório um grande número de vezes, a média dos resultados observados se aproximará cada vez mais da média esperada ou valor esperado. Isso implica que, quanto mais dados são coletados, mais confiáveis e representativos se tornam os resultados.

Essa lei é de fundamental importância na estatística, pois permite que façamos inferências e tomemos decisões com base em amostras representativas dos dados. Ela estabelece uma relação entre o tamanho da amostra e a precisão das estimativas estatísticas, proporcionando uma base sólida para a análise e interpretação de dados em diversas áreas, inclusive em PLN.

A.1.4 Covariância e Correlação

A covariância e a correlação são medidas estatísticas que descrevem o relacionamento entre duas variáveis.

A covariância mede a relação linear entre duas variáveis aleatórias. Ela indica como as duas variáveis variam juntas. Se a covariância for positiva, isso significa que as variáveis tendem a aumentar ou diminuir juntas. Por outro lado, se a covariância for negativa, indica que as variáveis têm uma relação inversa, ou seja, uma tende a aumentar quando a outra diminui. Uma covariância de zero indica que não há uma relação linear aparente entre as variáveis.

No entanto, a covariância por si só não fornece uma medida padronizada para a força e a direção da relação entre as variáveis. É aí que a correlação entra em cena. A correlação

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

é uma medida padronizada que varia entre -1 e 1, que indica a força e a direção da relação linear entre as variáveis.

Uma correlação de 1 indica uma relação linear positiva perfeita, onde as variáveis aumentam ou diminuem juntas na mesma proporção. Uma correlação de -1 indica uma relação linear negativa perfeita, onde as variáveis têm uma relação inversa perfeita. Uma correlação de 0 indica que não há uma relação linear aparente entre as variáveis.

A correlação é uma medida mais útil que a covariância pois é independente da escala das variáveis e fornece uma medida padronizada para a força da relação linear entre elas. Ela é amplamente utilizada na análise de dados, no planejamento de experimentos e na modelagem estatística para entender e quantificar o relacionamento entre variáveis.

A.1.5 Vetores aleatórios e distribuições multivariadas

Um vetor aleatório é uma coleção de variáveis aleatórias que são agrupadas como um único objeto. Cada elemento do vetor aleatório representa uma variável aleatória diferente, e o vetor como um todo pode ser usado para descrever simultaneamente o comportamento dessas variáveis aleatórias.

Uma distribuição multivariada é uma distribuição de probabilidade que descreve conjuntamente as probabilidades de ocorrência de múltiplas variáveis aleatórias. Ela fornece informações sobre as relações e dependências entre as variáveis aleatórias em um conjunto de dados.

Ao lidar com vetores aleatórios, é comum usar distribuições multivariadas para modelar a relação conjunta entre as variáveis. Uma distribuição multivariada especifica a forma como as variáveis aleatórias se relacionam entre si e como a probabilidade é distribuída em seu espaço de valores conjunto.

Essas distribuições são amplamente utilizadas em áreas como estatística, ciência de dados e análise de dados para analisar e entender as relações entre múltiplas variáveis aleatórias. Elas permitem realizar análises conjuntas, como calcular a probabilidade conjunta de eventos, estimar parâmetros e fazer previsões baseadas nas relações entre as variáveis.

A.1.6 Algumas distribuições úteis

Existem várias distribuições relevantes para aplicações de probabilidades e estatística em sistemas de língua falada. A escolha da distribuição adequada depende do contexto e dos dados específicos sendo analisados. Aqui estão alguns exemplos:

- Distribuição Binomial: É usada para modelar o número de sucessos em um número fixo de tentativas independentes, onde cada tentativa tem apenas dois resultados possíveis (sucesso ou fracasso). Por exemplo, pode ser usada para modelar a probabilidade de acerto em um teste de múltipla escolha, onde cada pergunta tem apenas duas opções de resposta.
- Distribuição de Poisson: É usada para modelar a ocorrência de eventos raros em um intervalo de tempo ou espaço. Por exemplo, pode ser usada para modelar a taxa de ocorrência de palavras específicas em um discurso ou texto.
- Distribuição Gaussiana (Normal): É uma das distribuições mais comuns e amplamente utilizadas. Ela descreve dados simétricos ao redor de uma média, com a maioria dos valores concentrados perto da média e uma cauda que se estende para ambos os lados. É frequentemente usada para modelar características acústicas de fala, como duração de fonemas ou intensidade de som.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

- Distribuição de Bernoulli: É uma distribuição específica da distribuição binomial, usada quando há apenas duas possibilidades de resultado (sucesso ou fracasso) em um único evento. Pode ser usada para modelar a probabilidade de uma determinada palavra aparecer em uma frase ou ocorrência de um determinado evento em um diálogo.
- Distribuição de Dirichlet: É uma distribuição multivariada usada para modelar a distribuição de probabilidades em um espaço de múltiplas categorias. É frequentemente usada em processamento de linguagem natural para modelar a distribuição de palavras em um *corpus* de texto.

A.1.7 Teoria da Estimação e Teste de Significância

A Teoria da Estimação é um ramo da estatística que lida com métodos e técnicas para estimar parâmetros desconhecidos com base em dados amostrais. Ela envolve a utilização de informações amostrais para fazer inferências sobre características de uma população maior. A teoria da estimação fornece procedimentos e medidas para determinar o valor estimado de um parâmetro desconhecido, bem como sua precisão e confiabilidade.

No contexto de sistemas de língua falada, a teoria da estimação pode ser aplicada de várias maneiras. Por exemplo, pode ser usada para estimar a taxa de erro de reconhecimento de fala em um sistema de reconhecimento automático de fala. Com base em uma amostra de dados de entrada e saída do sistema, é possível estimar a taxa de erro geral do sistema e a variação dessa estimativa.

Existem vários tipos de estimação utilizados na teoria da estimação. Aqui estão alguns exemplos:

- Estimação de ponto: Nesse tipo de estimação, um único valor é estimado para o parâmetro desconhecido. Por exemplo, se quisermos estimar a média de altura de uma população com base em uma amostra, podemos usar a média amostral como uma estimativa pontual desse parâmetro.
- Estimação por intervalo: Ao contrário da estimação de ponto, a estimação por intervalo fornece um intervalo de valores dentro do qual o parâmetro desconhecido provavelmente está contido. Por exemplo, podemos usar uma estimativa intervalar para estimar a proporção de falantes nativos de uma língua específica em uma determinada região, fornecendo um intervalo de confiança em torno dessa estimativa.
- Estimação por máxima verossimilhança: Nesse método, o objetivo é encontrar o valor do parâmetro que maximiza a probabilidade de obter os dados observados. A estimativa de máxima verossimilhança é frequentemente usada para estimar os parâmetros de distribuições de probabilidade em sistemas de língua falada, como a probabilidade de ocorrência de palavras em um *corpus* de texto.
- Estimação por mínimos quadrados: Esse método é amplamente utilizado na regressão linear, onde a linha de melhor ajuste é encontrada minimizando a soma dos quadrados das diferenças entre os valores observados e os valores previstos pela linha. Por exemplo, na modelagem da fala, a estimação por mínimos quadrados pode ser usada para encontrar os coeficientes que melhor ajustam um modelo de predição acústica aos dados de fala.

O teste de significância, por sua vez, é uma técnica estatística utilizada para avaliar a força da evidência contra uma hipótese nula. Ele permite determinar se os resultados observados são estatisticamente significativos ou se podem ser atribuídos ao acaso. O

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

teste de significância envolve a comparação dos dados observados com uma distribuição de probabilidade teórica esperada sob a hipótese nula.

Em sistemas de língua falada, o teste de significância pode ser usado para determinar se existe uma diferença significativa na taxa de reconhecimento entre dois algoritmos de reconhecimento de fala. Realizando testes estatísticos apropriados, é possível avaliar se a diferença observada na taxa de reconhecimento é estatisticamente significativa ou se pode ser atribuída ao acaso.

Existem diferentes tipos de testes de significância estatística que podem ser aplicados em diversos cenários. Aqui estão alguns exemplos:

- Teste t de Student: É usado para testar se há diferença significativa entre as médias de duas amostras independentes. Por exemplo, pode ser aplicado para determinar se a pontuação média de fluência em uma língua é significativamente diferente entre dois grupos de alunos.
- Teste qui-quadrado: É utilizado para testar a independência entre duas variáveis categóricas. Por exemplo, pode ser usado para verificar se existe uma associação significativa entre o gênero dos falantes e o uso de determinados fonemas em um idioma.
- Teste de ANOVA: É aplicado para testar se há diferenças significativas entre as médias de três ou mais grupos independentes. Por exemplo, pode ser usado para comparar as médias de desempenho em fala entre diferentes grupos de falantes nativos e não nativos.
- Teste de correlação: É utilizado para avaliar se existe uma relação significativa entre duas variáveis contínuas. Por exemplo, pode ser aplicado para verificar se há uma correlação significativa entre a frequência fundamental da voz e a altura percebida da fala.
- Teste de qui-quadrado de ajuste: É usado para verificar se uma distribuição observada se ajusta a uma distribuição teórica específica. Por exemplo, pode ser aplicado para verificar se a distribuição de frequência de palavras em um *corpus* de texto se ajusta à distribuição de Zipf.

Essas técnicas estatísticas, teoria da estimação e teste de significância, são essenciais para analisar dados de sistemas de língua falada, avaliar o desempenho de algoritmos e fazer inferências estatísticas relevantes. Elas permitem tomar decisões informadas com base em evidências estatísticas sólidas.

A.1.8 Teoria da Informação, Entropia e Informação Mútua

A Teoria da Informação é um campo da matemática e da ciência da computação que se preocupa em quantificar e estudar a transmissão de informações. Foi desenvolvida por Claude Shannon na década de 1940 e tem uma ampla gama de aplicações, incluindo o processamento da língua natural falada.

A entropia é um conceito fundamental na Teoria da Informação. Ela mede a quantidade média de informação contida em uma fonte de dados. Quanto maior a entropia, maior a incerteza e, portanto, maior a quantidade de informação necessária para descrever ou transmitir os dados. A entropia é calculada com base na distribuição de probabilidade dos eventos. Em processamento da língua natural falada, a entropia pode ser usada para medir a previsibilidade das palavras em um *corpus* de fala, ou seja, quanto elas são esperadas ou inesperadas com base nas frequências de ocorrência.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

A informação mútua é uma medida de dependência entre duas variáveis aleatórias. Ela quantifica a quantidade de informação que uma variável fornece sobre a outra. A informação mútua é calculada com base nas distribuições de probabilidade conjuntas das variáveis. Em termos simples, a informação mútua mede quanto a informação de uma variável reduz a incerteza sobre a outra variável. Em processamento da língua natural falada, a informação mútua pode ser usada para medir a associação entre diferentes palavras em um *corpus* de fala ou a dependência entre características acústicas e fonéticas.

A relação entre esses três conceitos é que a entropia é a medida fundamental da quantidade de informação contida em uma fonte de dados, enquanto a informação mútua mede a dependência ou associação entre duas fontes de informação. A entropia pode ser usada para calcular a informação mútua entre duas variáveis aleatórias, fornecendo uma medida da quantidade de informação compartilhada por essas variáveis.

Em relação às aplicações em processamento da língua natural falada, esses conceitos têm várias utilidades:

- Modelagem de linguagem: A entropia pode ser usada para medir a complexidade ou a incerteza da sequência de palavras em um *corpus* de fala, ajudando a desenvolver modelos de linguagem mais eficientes e precisos.
- Codificação de voz: A Teoria da Informação fornece princípios para a compressão e transmissão eficiente de sinais de voz, reduzindo a taxa de bits necessária para transmitir a informação de forma confiável.
- Detecção de padrões: A informação mútua pode ser aplicada para identificar padrões relevantes em dados de fala, como a associação entre determinados fonemas ou palavras em um contexto específico.
- Reconhecimento de fala: A informação mútua pode ajudar a melhorar a precisão dos sistemas de reconhecimento de fala, fornecendo informações sobre a dependência entre as características acústicas e os fonemas ou palavras correspondentes.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Apêndice B

Diretrizes de avaliação

Apêndice do Capítulo 18

Brielen Madureira

Publicado em: 13/03/2024

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

B.1 Diretrizes de avaliação para um modelo de cinco estágios

Critérios extraídos de (Cohen; Howe, 1988), traduzidos com substituição de "pesquisa" para "projeto", de forma a para torná-los mais genéricos, ou seja, se referir ao desenvolvimento de projetos em geral e não apenas aos de pesquisa em inteligência artificial.

Avaliando a tarefa ou objetivo do projeto

1. A tarefa é significante? Por quê?
    - Se o problema já foi definido antes, sua reformulação traz melhorias?
2. Há chances de seu projeto contribuir de forma significativa para o problema? A tarefa é manejável?
3. Conforme a tarefa se torna mais específica para seu projeto, ela ainda é representativa de uma classe de tarefas?
4. Algum aspecto interessante do problema foi simplificado ou abstraído?
    - Se o problema já estava previamente definido, algum aspecto dessa definição anterior foi abstraído ou simplificado?
5. Quais são os objetivos intermediários do projeto? Quais tarefas-chave serão ou já foram resolvidas como parte do projeto?
6. Como você saberá quando tiver demonstrado satisfatoriamente uma solução para o problema? É possível, para a tarefa em questão, demonstrar que se chegou a uma solução?

Critérios para avaliação de métodos

1. Que melhorias o método traz sobre tecnologias existentes?
    - Ele cobre mais situações (*input*)?
    - Ele produz mais variedade de comportamentos desejáveis (*outputs*)?
    - Espera-se que ele seja mais eficiente em termos de memória, velocidade, tempo de desenvolvimento, etc.?

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

   - Ele é promissor para futuros desenvolvimentos (por exemplo, se um novo paradigma for introduzido)?
2. Há uma métrica já estabelecida para avaliar a performance do método (por exemplo, ela é normativa, tem validade cognitiva)?
3. O método depende de outros métodos? Ou seja, há necessidade de pré-processamento de dados, acesso a certas bases de conhecimento ou rotinas?
4. Quais são as premissas ou suposições necessárias?
5. Qual é o escopo do método?
   - Quão extensível ele é? Seria simples aumentar sua escala com uma base de conhecimento maior?
   - Como ele aborda a tarefa? E partes dela? E uma classe de tarefas?
   - Ele ou suas partes poderiam ser aplicados a outros problemas?
   - Ele é transferível para tarefas mais complicadas (talvez com mais conhecimento ou mais ou menos limitações ou com interações complexas)?
6. Quando o método não consegue dar uma boa resposta, ele se abstém, dá uma resposta ruim ou dá a melhor resposta possível dados os recursos disponíveis?
7. Quão bem esse método é compreendido?
   - Por que ele funciona?
   - Em quais circunstâncias ele não funciona?
   - As limitações são inerentes ao modelos ou simplesmente não foram tratadas?
   - As decisões de *design* estão bem justificadas?
8. Qual a relação entre o problema e o método? Por que ele funciona para esta tarefa específica?

Critérios para avaliação da implementação do método

1. Quão demonstrativo é o programa?
   - Podemos avaliar seu comportamento externo?
   - Quão transparente ele é? Podemos avaliar seu comportamento interno?
   - A classe de habilidades necessárias para a tarefa pode ser demonstrada por um conjunto bem definido de casos de teste?
   - Quantos casos de teste ele demonstra?
2. Ele é refinado especialmente para um exemplo em particular?
3. Quão bem o programa implementa o método?
   - Você consegue determinar as limitações do programa?
   - Houve partes que foram deixadas de fora ou foram feitos quebra galhos? Por que e com qual efeito?
   - A implementação forçou uma definição detalhada ou mesmo uma re-avaliação do método? Como essa re-avaliação foi feita?
4. A performance do programa é previsível?

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Critérios para avaliação do *design* do experimento

1. Quantos exemplos podem ser demonstrados?
   - Eles são qualitativamente diferentes?
   - Esses exemplos ilustram todas a habilidades propostas? Eles ilustram limitações?
   - O número de exemplos é suficiente para justificar generalizações indutivas?
2. A performance do programa deve ser comparada com um *standard*, como um outro programa, *experts* ou novatos, com com sua própria performance refinada? O *standard* deve ser normativo, ter validade cognitiva, basear-se em eventos do mundo real ou em simulações?
3. Quais são os critérios de uma boa performance? Quem os define?
4. O programa pode ser generalizado (ou seja, é independente do domínio)?
   - Ele pode ser testado em diversos domínios?
   - Os domínios são qualitativamente diferentes?
   - Eles representam uma classe de domínios?
   - A performance no domínio inicial deve ser comparada à performance em outros domínios? Você espera que o programa está customizado para ter uma performance melhor no(s) domínio(s) usado(s) para *debugging*?
   - O conjunto de domínios é suficiente para justificar uma generalização indutiva?
5. Uma série de programas relacionados está sendo avaliada?
   - Você consegue determinar como as diferenças entre os programas se manifestam em diferentes comportamentos?
   - Se o método foi implementado de forma distinta em cada programa da série, como essas diferenças afetam as generalizações?
   - Houve dificuldades na implementação do método em outros programas?

Critérios para avaliação dos achados do experimento

1. Qual foi a performance do programa em comparação com o *standard* selecionado (por exemplo, outros programas, humanos, comportamento normativo)?
2. A performance do programa foi diferente das predições de como o método deveria funcionar?
3. Quão eficiente é o programa em termos de requerimentos de memória e de conhecimento?
4. O programa demonstrou boa performance?
5. Você aprendeu o que você queria com o programa e os experimentos?
6. É fácil para os usuários entenderem-no?
7. Você consegue definir as limitações de performance do programa?

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

8. Você entende por que o programa funciona ou não funciona?
    - Qual o impacto de mudanças sutis no programa?
    - Sua performance está de acordo com o esperado em exemplos que não foram usados para *debugging*?
    - O efeito de diferentes estratégias de controle pode ser determinado?
    - Como o programa responde se o *input* é rearranjado, tem ruído ou está incompleto?
    - Qual a relação entre as características dos problemas de teste e da performance (tanto externos, ou internos caso os registros do programa estejam disponíveis)?
    - A compreensão do programa pode ser generalizada para o método? Para características do método? Para uma tarefa maior?

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Apêndice C

EAGLES

Apêndice do Capítulo 18

Brielen Madureira

Publicado em: 13/03/2024

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

C.1 EAGLES

Receita de sete passos para avaliação de sistema de PLN, proposta pelo grupo de trabalho EAGLES em 1999.[1]

1. **Por que esta avaliação está sendo feita?**
    - Qual é o propósito desta avaliação? Todas as partes envolvidas têm o mesmo entendimento do propósito?
    - O que exatamente está sendo avaliado? Trata-se de um sistema ou de um componente? De forma isolada ou em um contexto específico? Quais são os limites do sistema?
2. **Elaboração de um modelo de tarefa**
    - Identificar todos os papeis e agentes relevantes.
    - Para que o sistema será usado?
    - Quem vai usá-lo? O que vão fazer com ele? Como são essas pessoas?
3. **Definição de características do nível máximo de qualidade**
    - Quais aspectos do sistema serão avaliados? São todos igualmente importantes?
4. **Produção de requisitos detalhados para o sistema sob avaliação, com base nos itens 2 e 3**
    - Para cada aspecto que foi identificado como importante, é possível encontrar uma medida válida e confiável de como o objeto avaliado performa nesse aspecto? Caso negativo, os aspectos precisam ser subdivididos de forma válida em sub-atributos que possam ser mensuráveis. Este ponto deve ser repetido até que se atinja um ponto onde todos os atributos são mensuráveis.
5. **Definição das métricas a serem aplicadas ao sistema para os requisitos produzidos no item 4**

[1] Disponível em https://www.issco.unige.ch/en/research/projects/eagles/ewg99/7steps.html

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

- Tanto a medida quanto o método para se obter essa medida precisam ser definidos para cada atributo.
- Para cada atributo mensurável, o que vai ser considerado uma pontuação boa, satisfatória ou insatisfatória dado o modelo da tarefa do item 2? Quais são os pontos de corte?
- Geralmente, um atributo tem mais de um sub-atributo. Como os valores dos diferentes sub-atributos serão combinados em um valor para o nível acima de modo a refletir sua importância relativa (novamente, dado o modelo da tarefa)?

6. **Projeto da execução da avaliação**
   - Desenvolvimento de materiais de teste para dar suporte ao teste do sistema em si.
   - Quem exatamente vai conduzir as diferentes medições? Quando? Em quais circunstâncias? Que forma terá o resultado final?
7. **Execução da avaliação**
   - Fazer as medições.
   - Comparar com as pontuações pré-definidas como satisfatórias.
   - Resumir os resultados em um relatório de avaliação, conforme o item 1.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Sobre as/os autoras/es

Helena de Medeiros Caseli

Helena Caseli é formada em Ciência da Computação pela UFU, com pós-graduação no ICMC/USP. É professora titular na UFSCar, onde atua desde 2008 e coordena o LALIC (Laboratório de Linguística e Inteligência Computacional). Suas áreas de interesse incluem aprendizado de máquina, tradução automática e PLN aplicado a redes sociais em domínios como política e saúde mental. Foi a idealizadora, é uma das organizadoras deste livro e colaborou com os Capítulos 1, 4, 21, Tradução Automática e Detecção de Transtornos de Saúde Mental a partir de Texto.
Email: helenacaseli@ufscar.br

Maria das Graças Volpe Nunes

Maria das Graças Volpe Nunes é formada em Ciência da Computação pela UFSCar, mestre pela USP e doutora pela PUC-Rio. Foi docente no ICMC-USP, São Carlos, de 1981 a 2013, onde agora é professora sênior. Atualmente é membro do C4AI, da USP, IBM e FAPESP. É uma das fundadoras do NILC, onde atuou em várias áreas do PLN para português: revisão ortográfica e gramatical, tradução automática, sumarização, análise de sentimentos, entre outras. É uma das organizadoras deste livro e colaborou com os Capítulos 1, 20 e 21.
Email: gracan@icmc.usp.br

Adriana Silvina Pagano

Adriana Pagano é formada em Letras pela Universidad Nacional de La Plata. Fez seu mestrado no Programa de Inglês e Literatura Correspondente da UFSC e seu doutorado no Programa de Estudos Linguísticos da UFMG. É professora titular da Faculdade de Letras da UFMG desde 1995. Suas áreas de interesse incluem tradução, produção textual multilíngue e modelagem sistêmico-funcional da linguagem. Neste livro, colaborou com os Capítulos 1, 6, 7 e PLN na Saúde.
Email: apagano@ufmg.br

Aline Aver Vanin

Aline Aver Vanin é licenciada em Letras pela UCS, com mestrado e doutorado em Linguística pela PUCRS. Atuou em estágio pós-doutoral no Laboratório de PLN do Programa de Pós-Graduação em Computação da PUCRS. É docente do Departamento de Educação e Humanidades da UFCSPA desde 2014 e colaboradora do Programa de Pós-Graduação em Letras da UNISC desde 2019. Tem interesse na interface discurso e PLN. Neste livro, colaborou com o Capítulo 12.
Email: alinevanin@ufcspa.edu.br

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Aline Villavicencio

Aline Villavicencio é formada em Ciência da Computação pela PUC-RS e doutora pela Universidade de Cambridge (Inglaterra). É professora titular da Universidade de Exeter (Inglaterra) onde é diretora do Institute of Data Science and Artificial Intelligence. Sua pesquisa é em Processamento de Linguagem Natural (PLN) inspirado em modelos cognitivo computacionais de processamento de linguagem e com foco em processamento de expressões multipalavras. Neste livro, colaborou com o Capítulo 5.
Email: a.villavicencio@exeter.ac.uk

Amanda Pontes Rassi

Amanda Rassi é bacharel e mestre em Linguística pela Universidade Federal de Goiás (UFG) e doutora na mesma área pela Universidade Federal de São Carlos (UFSCar). Tem experiência com correção automática de redações, tendo atuado na Redação Nota 1000 e atualmente na Somos Educação. Suas áreas de interesse incluem sintaxe, semântica, lexicologia, lexicografia e PLN. Neste livro, colaborou com os Capítulos 4, 6 e Correção Automática de Redação.
Email: amanda.rassi@somoseducacao.com.br

Ana Clara Souza Pagano

Ana Clara Souza Pagano é formada em Linguística Teórica e Descritiva pela Universidade Federal de Minas Gerais (UFMG). Fez seu Trabalho de Conclusão de Curso em Estilometria computacional. Suas áreas de interesse incluem: Aprendizado de Máquina (AM), Inteligência Artificial (IA), tradução e Processamento de Linguagem Natural (PLN). Neste livro, colaborou com os Capítulos 6 e 7.
Email: anapagano.ufmg@gmail.com

Arnaldo Candido Junior

Arnaldo Candido Junior possui graduação em Ciência da Computação pela Universidade Estadual Paulista Júlio de Mesquita Filho (2005), mestrado em Ciências da Computação e Matemática Computacional pelo Instituto de Ciências Matemáticas e de Computação (ICMC) da Universidade de São Paulo (USP) (2008) e doutorado em Ciências da Computação e Matemática Computacional também pelo ICMC/USP. É professor da Universidade Estadual Paulista. Atua nas áreas de linguagens de programação e inteligência artificial (PLN e processamento da fala). Neste livro, colaborou com o Capítulo 3.
Email: arnaldo.candido@unesp.br

Brielen Madureira

Brielen Madureira é formada em Matemática Aplicada e Computacional pela Universidade de São Paulo (USP). Tem mestrado em Ciência e Tecnologia da Linguagem pela Universidade de Saarland (Alemanha) e doutorado em Linguística Computacional pela Universidade de Potsdam (Alemanha), onde participou do grupo de pesquisa em Diálogo. Mais amplamente, tem interesse na avaliação de tecnologias de linguagem e nas urgentes

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

considerações éticas que elas demandam. Atualmente é pesquisadora em fase de pós-doutorado na Universidade de Leipzig (Alemanha), no grupo Climate Discourse. Neste livro, colaborou com os Capítulos 18, 15 e Responsabilidade no desenvolvimento e uso de tecnologias de linguagem baseadas em IA.
Email: brielen.madureira@uni-leipzig.de

Camila de Araújo Azevedo

Camila Azevedo foi aluna da primeira turma do Bacharelado em Linguística da Universidade Federal de São Carlos (UFSCar), atua há mais de 10 anos com síntese de voz e atualmente trabalha no time de processamento de fala do SiDi. Neste livro, colaborou com o Capítulo 2.
Email: ca.araujo.azevedo@gmail.com

Carlos Ramisch

Carlos Ramisch é formado em ciência da computação pela Universidade Federal do Rio Grande do Sul (UFRGS) e concluiu o doutorado em 2012, em co-tutela entre a UFRGS e a Université de Grenoble (França). Desde 2013, ele é professor na Universidade de Aix-Marseille e pesquisador afiliado ao Laboratoire d'Informatique et Systèmes, na França. Sua pesquisa é focada na área de semântica computacional, em especial no processamento de expressões multipalavras. Neste livro, colaborou com o Capítulo 5.
Email: carlos.ramisch@lis-lab.fr

Cláudia Freitas

Cláudia Freitas é formada em Letras pela PUC-Rio, onde concluiu o doutorado em Estudos da Linguagem (2007) sobre a construção de ontologias a partir de *corpus*. De 2012 a 2022 foi docente do Programa de Pós-Graduação em Estudos da Linguagem da PUC-Rio, onde fundou o grupo ComCorHD (Linguística Computacional, *Corpus* e Humanidades Digitais). De 2007 a 2022 foi também pesquisadora da Linguateca. Ao longo de 2023 e 2024 foi pesquisadora do ICMC/USP, vinculada ao C4AI, e atualmente é pesquisadora independente. Suas áreas de interesse incluem construção e avaliação de *datasets*/*corpora*, semântica/significado, e articulação entre abordagens simbólicas e neurais. Neste livro, colaborou com os Capítulos 8, 16, 19 e ChatGPT, MariTalk e outros agentes de conversação.
Email: claudiafreitas@usp.br

Daniela Barreiro Claro

Daniela Barreiro Claro é professora titular da Universidade Federal da Bahia (UFBA), fez seu Mestrado em Ciências da Computação pela Universidade Federal de Santa Catarina (UFSC) e o seu Doutorado na Université d'Angers (França). Em 2009, ela fundou o Grupo de Pesquisa FORMAS na área de Semântica e PLN. Suas principais áreas de pesquisa incluem extração de informação, aprendizado de máquina e aprendizado multimodal. Neste livro, colaborou com os Capítulos 10 e Extração de Informação.
Email: dclaro@ufba.br

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Dante Augusto Barone

Dante Augusto Barone é doutor em Ciência da Computação pelo Instituto Nacional Politécnico de Grenoble(INPG), França e tem mestrado e graduação em Engenharia Elétrica, pela USP e pela UFRGS, respectivamente. Atualmente coordena o Programa de Pós-Graduação em Informática na Educação da UFRGS. Suas áreas de interesse incluem Aprendizado de Máquina, Processamento de Linguagem Natural e Ética em Inteligência Artificial. Contribuiu e liderou projeto de cooperação com a Petrobrás para o desenvolvimento de robô, o qual recebeu o prêmio de inovação em 2021. Neste livro, colaborou com o Capítulo 14.
Email: dante.barone@gmail.com

Diana Santos

Diana Santos é formada em Engenharia Eletrotécnica e de Computadores em 1985, com mestrado na mesma área com tese em tradução automática em 1988, e doutorada em Engenharia Informática em 1996 com uma tese em semântica computacional, todos pelo Instituto Superior Técnico, Lisboa. Desde 1998 que trabalha para o processamento computacional da língua portuguesa no mundo, sendo líder da Linguateca desde 2000. Desde 2011 que é professora na Universidade de Oslo em Linguística Portuguesa, e também leciona ou lecionou Tradução, Estatística, e Humanidades Digitais. Os seus interesses são quase todas as áreas do PLN, e avaliação em geral. Neste livro, colaborou com o Capítulo 19.
Email: dianamsmpsantos@gmail.com

Edresson Casanova

Edresson Casanova é bacharel em Ciência da Computação pela UTFPR, com pós-graduação em inteligência artificial e PLN com foco em processamento de fala no ICMC/USP. Atualmente, é Deep Learning engineer na coqui.ai. Suas áreas de interesse incluem síntese de fala, clonagem de voz, verificação de locutores e reconhecimento automático de fala. Neste livro, colaborou com o Capítulo 3.
Email: edresson1@gmail.com

Eduardo Cortes

Eduardo Cortes é doutor em Ciência da Computação pela Universidade Federal do Rio Grande do Sul (UFRGS), com experiência na área de Inteligência Artificial, com ênfase em Sistemas de Question Answering, Processamento de Linguagem Natural (PLN) e Aprendizado de Máquina (AM). Atualmente, atua como pesquisador em um projeto conjunto entre a Unisinos e a SAP, focado em PLN. Neste livro, colaborou com o Capítulo 14.
Email: eduardogcortes8@gmail.com

Elisa Terumi Rubel Schneider

Elisa Terumi é formada em Informática pela Universidade Tecnológica Federal do Paraná (UTFPR), com especialização em Desenvolvimento Web (PUCPR), mestrado em Bioinformática (UFPR) e doutorado em Informática atuando na linha de pesquisa de Inteligência

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Artificial (PUCPR). Tem interesse em PLN aplicado na área de saúde. Neste livro, colaborou com os Capítulos 7 e PLN na Saúde.
Email: lisa.terumi@gmail.com

Eloize Rossi Marques Seno

Eloize Seno é Bacharel em Análise e Desenvolvimento de Sistemas pela UNILINS, com mestrado em inteligência artificial e PLN pelo Programa de Pós-graduação em Ciência da Computação da UFSCar e doutorado em inteligência artificial e PLN pelo ICMC/USP. É docente no Instituto Federal de São Paulo desde 2011. Suas áreas de interesse incluem sumarização automática, tradução automática, análise semântica e mineração de opinião. Neste livro, colaborou com os Capítulos 9 e 10.
Email: eloizeseno@gmail.com

Evandro Fonseca

Evandro Fonseca possui graduação em Ciência da Computação (2011) pela Universidade Católica de Pelotas (UCPEL), mestrado (2014) e doutorado (2018) em Ciência da Computação, com foco em PLN, pela Pontifícia Universidade Católica do Rio Grande do Sul (PUCRS). Atualmente atua como Team Lead | AI/NLP Specialist na Blip. Neste livro, colaborou com o Capítulo 12.
Email: metalmorphy@gmail.com

Flaviane Romani Fernandes Svartman

Flaviane Svartman possui graduação e doutorado em Linguística pela Unicamp, com período de estágio de doutorado na Faculdade de Letras da Universidade de Lisboa e pós-doutorado também pela Unicamp. Atua como docente na área de Filologia e Língua Portuguesa na Faculdade de Filosofia, Letras e Ciências Humanas da USP desde 2010. Suas investigações concernem ao estudo da fonologia e da fonética da língua portuguesa, com especial interesse na prosódia, na interface sintaxe-fonologia, na comparação entre variedades africanas, brasileiras e europeias do português e no PLN. Neste livro, colaborou com os Capítulos 3 e Classificação de Áudio aplicada à Saúde.
Email: flaviane@gmail.com

Heliana Mello

Heliana Mello é formada em Letras pela Universidade Federal de Minas Gerais (UFMG), com doutorado em Linguística pela CUNY. É docente na UFMG desde 1998. Suas áreas de interesse incluem métodos empíricos e quantitativos para a análise da linguagem, compilação e estudo de *corpora* de fala. Neste livro, colaborou com o Capítulo 2.
Email: hmello@ufmg.br

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Ivandré Paraboni

Ivandré Paraboni é doutor em Ciência da Computação pela University of Brighton (Inglaterra, 2003), com estágio de pós-doutorado junto à University of Aberdeen (Escócia, 2012). É professor associado (livre docente) e pesquisador da USP/EACH. Possui experiência na área de Ciência da Computação, atuando principalmente nas áreas de PLN e IA. Interesses de pesquisa abrangem o processamento da língua humana em um sentido amplo do termo, incluindo métodos computacionais derivados das Ciências Cognitivas e aplicações práticas de classificação de documentos extraídos da WEB. Neste livro, colaborou com os Capítulos 13 e Detecção de Transtornos de Saúde Mental a partir de Texto.
Email: ivandre@usp.br

Jackson Wilke da Cruz Souza

Jackson Souza é bacharel (2012), mestre (2015) e doutor (2019) em Linguística pela Universidade Federal de São Carlos (UFSCar). Atualmente, é professor da Universidade Federal da Bahia (UFBA), atuando junto ao Departamento de Ciência, Tecnologia e Inovação e ao Programa de Pós-Graduação em Língua e Cultura da UFBA. Tem interesse nas áreas de análise de discurso, sumarização automática e análise e descrição linguística. Neste livro, colaborou com os Capítulos 11 e Sumarização Automática.
Email: jackcruzsouza@gmail.com

Laila Mota

Laila Mota é formada em Ciência e Tecnologia pela Universidade Federal da Bahia (UFBA), com pós-graduação em Tecnologia da Informação pela UFBA. Suas áreas de interesse incluem aprendizado de máquina, linguística computacional e modelos de linguagem. Neste livro, colaborou com o Capítulo 10.
Email: lailapmota@gmail.com

Lucelene Lopes

Lucelene Lopes é doutora em Ciência da Computação pela PUCRS (2012). Consultora Empresarial na área de Processamento de Linguagem Natural (PLN) desde 2016, e desde 2021 é pesquisadora no ICMC/USP, vinculada ao C4AI. Suas áreas de interesses incluem PLN, aprendizado de máquina, *parsing* e modelos de linguagem. Neste livro, colaborou com os Capítulos 4 e 7.
Email: lucelene@gmail.com

Maria José Bocorny Finatto

Maria José Bocorny Finatto é linguista especializada em Estudos do Léxico e Terminologia. Investiga a história das terminologias médicas e modos de aplicar técnicas da Linguagem Simples para maior acessibilidade das informações escritas para diferentes perfis de pessoas no Brasil. Análise textual com recursos PLN é seu foco. Orientadora de mestrado e doutorado junto à Universidade Federal do Rio Grande do Sul (UFRGS). Neste livro, colaborou com os Capítulos 4, PLN no Direito e PLN e Humanidades Digitais.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Email: mariafinatto@gmail.com

Mariza Ferro

Mariza Ferro é formada em Ciência da Computação pela UNIOESTE, com mestrado em Inteligência Artificial no ICMC/USP e doutorado em Modelagem Computacional pelo LNCC. É docente do Instituto de Computação da Universidade Federal Fluminense (UFF) e coordenadora do Núcleo de Referência em IA Ética e Confiável desde 2020. Suas áreas de interesse incluem IA ética, IA verde e sustentável, IA para o estar social e aprendizado de máquina para predição de eventos extremos. Neste livro, colaborou com o Capítulo 20.
Email: mariza@ic.uff.br

Marli Quadros Leite

Marli Quadros Leite é Professora Titular do Departamento de Letras Clássicas e Vernáculas (FFLCH | USP) e atua como pesquisadora nas áreas de análise do Discurso Oral e na de História das Ideias Linguísticas. Atualmente é Pró-Reitora de Cultura e Extensão Universitária da USP. Neste livro, colaborou com o Capítulo 3.
Email: mqleite@usp.br

Nataly Leopoldina Patti da Silva

Nataly Patti é bacharel e mestre em Sistemas de Informação pela Universidade de São Paulo (USP), onde desenvolveu pesquisas em PLN com foco em Aprendizado Transdutivo e classificação de Atos de Fala. Atualmente, é aluna de doutorado na USP e atua como Cientista de Dados no Instituto de Pesquisas SiDi, contribuindo com projetos de PLN, especialmente no Samsung Search, Samsung Health e Bixby. Suas áreas de interesse incluem PLN, aprendizado de máquina e recuperação de informação. Neste livro, colaborou com o Capítulo 17.
Email: natalypatti@gmail.com

Norton Trevisan Roman

Norton Trevisan Roman é Bacharel em Física, com Mestrado e Doutorado em Ciência da Computação com foco em PLN, todos pela Unicamp, possuindo também Livre-Docência pela USP. Docente na EACH-USP desde 2010, suas áreas de interesse incluem Processamento de Língua Natural e aplicações de PLN e IA a diversas tarefas, indo desde o tratamento de emoção e sentimento à predição do mercado financeiro. Neste livro, colaborou com o Capítulo 17.
Email: nortontr@gmail.com

Paula Christina Figueira Cardoso

Paula Figueira Cardoso tem graduação em Informática pelo ILES de Santarém, mestrado em Engenharia Elétrica pela Universidade Federal do Pará (UFPA) e doutorado pelo Instituto de Ciências Matemáticas e de Computação (ICMC) – USP/São Carlos. Atualmente é

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

docente na UFPA. Suas áreas de interesse incluem geração de textos, sumarização automática, tecnologias educacionais e pesquisas aplicadas na área de interface homem-máquina. Neste livro, colaborou com os Capítulos 11 e Sumarização Automática.
Email: paulastm@gmail.com

Priscila Osório Côrtes

Priscila Osório Côrtes é formada em Letras Alemão/Português pela Universidade Federal de Minas Gerais (UFMG) e mestre em linguística pela Albert Ludwigs-Universität Freiburg, Alemanha. Trabalha há mais de 7 anos na área de Processamento de Linguagem Natural (PLN), e há mais de 5 com processamento de fala na empresa SiDi Campinas. Neste livro, colaborou com o Capítulo 2.
Email: pritico@gmail.com

Renata Ramisch

Renata Ramisch é formada em Letras pela Universidade Federal do Rio Grande do Sul (UFRGS) e mestre em Linguística pela Universidade Federal de São Carlos (UFSCar). É linguista na Redação Nota 1000, que faz parte do grupo Somos Educação. Suas áreas de interesse incluem correção automática de redações e tratamento computacional de expressões multipalavras. Neste livro, colaborou com o Capítulo 5.
Email: renata.ramisch@gmail.com

Renata Vieira

Renata Vieira é PhD em Informática e Ciências Cognitivas pela Universidade de Edimburgo, Escócia. É pesquisadora da área de Processamento de Linguagem Natural (PLN) atuando em diversas áreas de aplicação. É investigadora principal convidada do Centro de História, Culturas e Sociedades da Universidade de Évora (Portugal) e conselheira científica do Instituto de Inteligência Artificial na Saúde. Neste livro, colaborou com os Capítulos 12, 14, Extração de Informação e PLN e Humanidades Digitais.
Email: renata.vieira@gmail.com

Ricardo Marcacini

Ricardo Marcacini é formado em Informática pela Universidade de São Paulo (USP), com pós-graduação em ciência de computação e matemática computacional no ICMC/USP. Atualmente é docente do Departamento de Ciências de Computação (SCC) do ICMC/USP. Realiza pesquisa na área de aprendizado de máquina, com foco em mineração de textos, análise de eventos e análise de sentimentos. Neste livro, colaborou com o Capítulo 3.
Email: ricardo.marcacini@icmc.usp.br

Roana Rodrigues

Roana Rodrigues é formada em Letras, mestra e doutora em Linguística pela Universidade Federal de São Carlos (UFSCar). É docente do Departamento de Letras Estrangeiras da

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Universidade Federal de Sergipe (UFS). Suas áreas de interesse incluem sintaxe, semântica, estudos do léxico e Processamento de Linguagem Natural (PLN). Neste livro, colaborou com o Capítulo 11.
Email: roana@academico.ufs.br

Sandra Maria Aluísio

Sandra Maria Aluísio é formada em Ciência da Computação pela UFSCar, com Doutorado em IA e PLN, pela USP. Foi docente de 1988 a 2018, no ICMC/USP, onde hoje atua no Programa Professor Sênior. As áreas de interesse incluem linguística de *corpus*, simplificação textual, recursos semânticos, detecção automática da estrutura do discurso científico e ferramentas de suporte à escrita, análise automática de distúrbios de linguagem, e processamento de fala. Neste livro, colaborou com os Capítulos 3 e Complexidade Textual e suas Tarefas Relacionadas.
Email: sandra@icmc.usp.br

Solange Rezende

Solange Rezende é docente e pesquisadora no Laboratório de Inteligência Computacional (LABIC) do Departamento de Ciências de Computação do ICMC/USP desde 1991. Tem experiência na área de IA, atuando principalmente nos temas relacionados com mineração de dados e textos, análise de sentimentos e sistemas de recomendação. Possui graduação em licenciatura em Ciências Habilitação em Matemática pela UFU (1986), mestrado em Ciências de Computação e Matemática Computacional pela USP (1990) e doutorado em Engenharia Mecânica – São Carlos pela USP (1993). Realizou pós-doutorado na Universidade de Minnesota, USA (1995-1996). Neste livro, colaborou com o Capítulo 3.
Email: solange@icmc.usp.br

Tayane Soares

Tayane Soares é Cientista de Dados na empresa Blip, formada em Letras pela UFMG e mestra em Linguística Aplicada pelo PosLin-UFMG. No mestrado, seu trabalho de pesquisa teve como foco a avaliação de viés de gênero em modelos de tradução automática, especificamente no par linguístico inglês-português. Seu interesse de pesquisa abrange a interseção entre linguagem e tecnologia, incluindo linguística computacional, processamento de língua natural e aprendizado de máquina. Além disso, também se interessa em pesquisar e avaliar como a aplicação dessas técnicas em tecnologias digitais afeta a sociedade e as pessoas que as utilizam. Neste livro, colaborou com o Capítulo 20.
Email: arantesoares@gmail.com

Thiago Castro Ferreira

Thiago Castro Ferreira é Bacharel (2008-2012) e Mestre em Ciências (2012-2014) pelo curso de Sistemas de Informação da Universidade de São Paulo. Com foco em Geração de Língua Natural, possui doutorado pela Universidade de Tilburg (2015-2018), e pós-doutorados também pela Universidade de Tilburg (2018-2019) e Universidade Federal de Minas Gerais (2019-2021). Thiago é atualmente um Cientista Aplicado na aiXplain com

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

interesses em Engenharia de Software, Processamento de Linguagem Natural, Aprendizado de Máquina, Modelos de Linguagem e Agentes. Neste livro, colaborou com o Capítulo 13.
Email: thiago.castro.ferreira@gmail.com

Valéria de Paiva

Valeria de Paiva é formada em Matemática pela PUC-Rio, com pós-graduação em Algebra na PUC e e mestrado/doutorado na Universidade de Cambridge, UK. É pesquisadora principal (Principal Researcher) no Topos Institute desde 2021. Suas áreas de interesse incluem semântica de linguagem natural, inferência em linguagem natural, recursos léxico-semânticos e lógica, especialmente lógica categórica. Gosta de sistemas híbridos de PLN, open software, dependências universais e da OpenWordNet-PT. Fundadora de *Women in Logic* e *Lógicas Brasileiras*. Neste livro, colaborou com o Capítulo 9.
Email: valeria@topos.institute

Vinícius G. Santos

Vinícius G. Santos é graduado, mestre e doutor em Letras pela Universidade de São Paulo. Atualmente realiza pesquisa de pós-doutorado vinculada ao C4AI–USP. Atua na área de descrição e análise linguística de variedades africanas e brasileiras da língua portuguesa, com especial interesse em fonologia, prosódia, entoação e elaboração de recursos para PLN. Neste livro, colaborou com o Capítulo 3.
Email: vinicius.santos@alumni.usp.br

Vládia Pinheiro

Vládia Pinheiro é formada em Ciência da Computação pela Universidade Federal do Ceará (UFC), com doutorado em Ciência da Computação na mesma universidade. É docente do Programa de Pós-Graduação em Informática Aplicada, na UNIFOR, desde 2011. Suas áreas de interesse incluem recursos semânticos, *parsers* semânticos, extração de informação, sumarização de textos, *chatbots*, aplicações de inteligência artificial e ciência de dados. Neste livro, colaborou com o Capítulo 9.
Email: vladiacelia@gmail.com

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol1/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Processamento de Linguagem Natural

Conceitos, Técnicas e Aplicações em Português

Organizado por:
Helena de Medeiros Caseli
Maria das Graças Volpe Nunes

4ª Edição
Volume 2: Modelos de Linguagem
Abril 2026

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol2/

Conteúdo

Volume 2: Modelos de Linguagem 1
Sobre o Volume 2 . . . 1
Como citar . . . 1

I Modelos de Linguagem 3

1 Modelos de linguagem 4
1.1 Relembrando a Hipótese Semântica e Definindo Modelos . . . 4
1.2 Modelos de Linguagem Probabilísticos . . . 6
1.2.1 Estimando as probabilidades a partir de *corpora* . . . 7
1.2.2 Usando o modelo probabilístico . . . 8
1.3 Modelos de Linguagem Neurais . . . 9
1.3.1 Um Contexto Histórico . . . 9
1.3.2 Modelos de Linguagem Neurais Modernos . . . 10
1.3.3 Treinamento e Ajustes em Modelos de Linguagem Neurais . . . 35
1.3.4 Avaliação de Modelos de Linguagem Neurais . . . 41
1.4 Tendências . . . 44
1.4.1 A Era dos *Large Language Models* (LLMs) . . . 44
1.4.2 Treinamento Eficiente de Modelos de Linguagem Neurais . . . 48
1.4.3 Estratégias de Treinamento para Agentes de Conversação: alinhamento e feedback humano . . . 50

2 O papel dos dados no pré-treinamento de Grandes Modelos de Linguagem 53
2.1 Introdução . . . 53
2.2 Por que e quando pré-treinar? . . . 55
2.3 Conjuntos de dados para pré-treinamento em português . . . 57
2.3.1 Recursos gerais . . . 57
2.3.2 Recursos para o português . . . 58
2.4 A importância da curadoria de dados . . . 60
2.5 O impacto da qualidade dos dados . . . 61
2.5.1 O custo da qualidade . . . 62
2.5.2 Considerações finais desta seção . . . 63
2.6 O impacto da diversidade dos dados . . . 64
2.7 A questão da proveniência dos dados . . . 66
2.8 O *corpus* Carolina como um estudo de caso . . . 67
2.9 Considerações finais . . . 68

3 Treinamento de Grandes Modelos de Linguagem na prática 70

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol2/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

3.1 Introdução . . . 70
3.2 Continuação de pré-treino . . . 71
3.2.1 Mecanismos de treinamento autossupervisionado . . . 71
3.2.2 Contextos e objetivos da continuação de pré-treino . . . 72
3.3 Ajuste fino de modelos . . . 73
3.3.1 Ajuste fino completo . . . 74
3.3.2 PEFT . . . 76
3.3.3 *Few-shot learning* . . . 79
3.4 *AI alignment* . . . 81
3.4.1 Fundamentos de aprendizado por reforço . . . 82
3.4.2 *Reinforcement Learning from Human Feedback* (RLHF) . . . 83
3.4.3 *Direct Preference Optimization* (DPO) . . . 84
3.4.4 Discussão comparativa: RLHF vs. DPO . . . 85
3.5 Exemplo prático . . . 86
3.5.1 Preparando o ambiente . . . 86
3.5.2 O ponto de partida . . . 87
3.5.3 Ajuste fino supervisionado com LoRA . . . 90
3.5.4 Refinamento do comportamento com DPO . . . 93
3.5.5 Considerações finais . . . 95

4 Avaliação de Grandes Modelos de Linguagem 96
4.1 A importância e os desafios intrínsecos da avaliação de LLMs . . . 96
4.1.1 Tipos de avaliação . . . 97
4.2 Avaliações tradicionais . . . 98
4.2.1 Descompasso com a percepção humana . . . 98
4.2.2 Dificuldade em avaliar aspectos complexos . . . 99
4.2.3 Exemplos práticos . . . 100
4.2.4 Avaliação humana: O padrão ouro? . . . 100
4.2.5 Importância e metodologias comuns . . . 100
4.2.6 Desafios da avaliação humana . . . 101
4.3 *Benchmarks* e *datasets* de avaliação . . . 102
4.3.1 *Benchmarks* genéricos (General-Purpose) . . . 102
4.3.2 *Benchmarks* para tarefas específicas . . . 102
4.3.3 *Benchmarks* para segurança e confiabilidade . . . 104
4.3.4 A importância de ir além do placar . . . 104
4.4 Avaliação baseada em modelos (*Model-based evaluation*) . . . 104
4.4.1 Vantagens . . . 105
4.4.2 Abordagens . . . 105
4.4.3 Desafios e Limitações . . . 106
4.4.4 Exemplos de *frameworks* de *LLM as a Judge* . . . 107
4.5 Aspectos emergentes e avançados da avaliação . . . 108
4.5.1 Robustez e Ataques adversariais . . . 108
4.5.2 Avaliação de justiça, viés e toxicidade . . . 108
4.5.3 Factualidade, alucinações e citabilidade . . . 109
4.5.4 Eficiência computacional e custos . . . 109
4.6 Considerações éticas e o Futuro da avaliação de LLMs . . . 110
4.6.1 Responsabilidade na avaliação . . . 110
4.6.2 Transparência e Interpretabilidade . . . 110

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

4.6.3 Dualidade do uso . . . 110
4.6.4 O papel da regulação e de padrões na avaliação . . . 110
4.6.5 Direções futuras . . . 110
4.7 Considerações finais . . . 111

5 Geração Aumentada por Recuperação (RAG) 113
5.1 Fundamentos da Geração Aumentada por Recuperação . . . 113
5.1.1 Introdução à RAG . . . 114
5.1.2 Arquitetura RAG . . . 115
5.1.3 O componente Recuperador (*Retriever*) . . . 116
5.1.4 O componente Gerador (*Generator*) . . . 119
5.2 RAG vs. *Fine-tuning* . . . 120
5.2.1 *Fine-tuning* . . . 120
5.2.2 RAG . . . 120
5.2.3 Abordagens híbridas . . . 121
5.3 RAG - Aspectos avançados . . . 121
5.3.1 RAG em duas etapas: Re-ordenando resultados . . . 121
5.3.2 RAG adaptiva e o Modelo agêntico . . . 124
5.3.3 Indo além do texto . . . 126
5.3.4 Avaliando RAG . . . 128
5.3.5 Otimização de desempenho: Latência e Escalabilidade . . . 129
5.3.6 Ética em RAG: Viés, privacidade e responsabilidade . . . 130
5.4 RAG no mundo real: Estudos de caso por indústria . . . 131
5.4.1 Setor financeiro: Conformidade regulatória e Análise de risco . . . 131
5.4.2 Saúde e Ciências da vida: Suporte à decisão clínica . . . 132
5.4.3 Setor jurídico: Revolucionando a pesquisa e a *Due Diligence* . . . 132
5.4.4 Contexto científico: Rastreabilidade da informação . . . 132
5.5 O futuro da RAG . . . 133
5.5.1 O debate: RAG vs. Janelas de contexto amplas (LCW) . . . 133
5.5.2 RAG como um mecanismo de atenção inteligente . . . 134
Agradecimentos . . . 135

6 Responsabilidade no desenvolvimento e uso de tecnologias de linguagem baseadas em IA 136
6.1 Introdução . . . 136
6.2 Quais facetas da responsabilidade devem ser consideradas? . . . 137
6.2.1 Dilemas nos dados . . . 138
6.2.2 Impactos e danos . . . 139
6.2.3 Riscos e maus usos . . . 142
6.2.4 Inconveniências operacionais . . . 145
6.2.5 Hegemonia de interesses comerciais e econômicos . . . 148
6.3 Como atuar de forma responsável? . . . 152
6.3.1 Para saber mais . . . 153
6.3.2 Formas de resistência . . . 154
6.4 Considerações finais . . . 154
Agradecimentos . . . 155

7 ChatGPT, MariTalk e outros agentes de conversação 156

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol2/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

7.1 Introdução . . . 156
7.2 Jogos de Linguagem . . . 159
7.3 Jogos que os agentes parecem jogar bem . . . 161
7.3.1 Sumarização . . . 161
7.3.2 Criação de história . . . 162
7.3.3 Tradução automática . . . 168
7.3.4 Escrever e-mails . . . 171
7.4 Jogos que os agentes melhoraram . . . 172
7.4.1 Simplificação textual . . . 172
7.5 Jogos que os agentes parecem jogar mal . . . 176
7.5.1 Perguntas sobre conteúdo pouco popular na internet . . . 176
7.5.2 Conte uma piada . . . 180
7.5.3 Inferências e raciocínio . . . 181
7.6 Quando o jogo é perigoso . . . 181
7.7 Conclusões provisórias . . . 185

Referências **186**

Sobre as/os autoras/es **215**
Aline Paes . . . 215
Altigran Soares Silva . . . 215
André Carvalho . . . 215
Brielen Madureira . . . 215
Cláudia Freitas . . . 215
Cristiane Namiuti . . . 216
Daniela Vianna . . . 216
Danilo Samuel Jodas . . . 216
Douglas Rodrigues . . . 216
Gabriel Lino Garcia . . . 216
Gabriela Alves Lachi . . . 216
Felipe Ribas Serras . . . 217
Jessica Rodrigues . . . 217
João Paulo Papa . . . 217
João Renato Ribeiro Manesco . . . 217
João Vitor Mariano Correia . . . 217
Livy Real . . . 217
Lucas Lasota . . . 218
Marcelo Finger . . . 218
Mariana Lourenço Sturzeneker . . . 218
Mayara Feliciano Palma . . . 218
Miguel de Mello Carpi . . . 218
Pedro Henrique Paiola . . . 218
Vanessa Martins do Monte . . . 218

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Volume 2: Modelos de Linguagem

Sobre o Volume 2

Dada a relevância atual do emprego de técnicas de aprendizado de máquina com redes neurais profundas, Transformers e Grandes Modelos de Linguagem (LLM) na área de Inteligência Artificial e, em particular, em PLN, reservamos este volume aos capítulos que tratam desses temas.

No Capítulo 1 são discutidas as redes neurais profundas e os Transformers, as técnicas de *fine-tuning* e *prompt engineering*. Essa leitura é essencial para o leitor ficar atualizado sobre o que tem sido usado para a criação das principais aplicações de PLN, assim como de toda inteligência artificial.

Os Capítulos 2 e 3 discorrem sobre o papel dos dados no pré-treinamento de modelos e sobre as principais técnicas de treinamento de grandes modelos de linguagem, respectivamente.

O importante tema de avaliação dos LLMs é tratado no Capítulo 4, que introduz seus fundamentos, métricas, *benchmarks* e considerações de ordem ética.

O desafio para mitigar o efeito de alucinação dos LLMs é discutido no Capítulo 5, onde é apresentada a técnica de geração aumentada de recuperação (*Retrieval-Augmented Generation*, RAG), que suporta acesso a fontes de dados externas para gerar respostas factuais e relevantes, reduzindo alucinações dos modelos.

O Capítulo 6 apresenta um amplo panorama das discussões envolvendo aspectos importantes do desenvolvimento responsável de tecnologias de linguagem baseadas em inteligência artificial.

Finalmente, no Capítulo 7, o leitor terá oportunidade de conhecer o caminho percorrido pelo PLN até chegar aos agentes atuais, como ChatGPT e MariTalk, que são analisados em profundidade.

Como citar

Caseli, H.M.; Nunes, M.G.V. (org.) **Processamento de Linguagem Natural: Conceitos, Técnicas e Aplicações em Português**. 4 ed. BPLN, 2026. v. 2. ISBN 978-65-02-00158-5. Disponível em: https://brasileiraspln.ufscar.br/livro-pln-4ed-vol2.

Note 1: Bibref

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol2/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

```
@Book{BPLN_livro_4ed_vol2:2026,
   title     = {Processamento de Linguagem Natural: Conceitos,
                Técnicas e Aplicações em Português},
   editor    = {Caseli, H. M. and Nunes, M. G. V.},
   year      = {2026},
   publisher = {BPLN},
   edition   = {4},
   volume    = {2},
   isbn      = {978-65-02-00158-5},
   url       = {https://brasileiraspln.ufscar.br/livro-pln-4ed-vol2}
}
```

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol2/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Parte I

Modelos de Linguagem

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol2/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Capítulo 1

Modelos de linguagem

Aline Paes
Daniela Vianna
Jessica Rodrigues

Publicado em: 26/09/2023

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol2/

1.1 Relembrando a Hipótese Semântica e Definindo Modelos

Da segunda década em diante do século XXI, testemunhamos um avanço significativo no desenvolvimento e popularização do aprendizado de representações numéricas para linguagem. Na época de escrita deste capítulo, **modelos de linguagem computacionais**, em particular os gerados por redes neurais são utilizados para representar textos escritos, fala, e até mesmo especificações que não são consideradas como parte da "linguagem natural", como, por exemplo, formalizações matemáticas (Geva et al., 2020; Gong et al., 2022; Li et al., 2023c; Piękos et al., 2021), código (Li et al., 2023b; Wang et al., 2021) [1][2], e até codificação de informações genéticas e moleculares (Brandes et al., 2022; Nijkamp et al., 2022). Os modelos de linguagem produzidos por redes neurais tanto geram como consomem textos mapeados para representações numéricas.

Mas por que seria importante representar informações essencialmente simbólicas em um formato numérico? A resposta mais simples e direta é que os computadores gostam de números. Seguindo ao porquê, a pergunta que segue é **"como representar tais informações simbólicas em um formato numérico, de forma a capturar sua semântica ?"** A segunda parte da pergunta – a tentativa de captura da semântica – é o ponto-chave, uma vez que simplesmente representar os componentes da língua em um formato numérico poderia guiar para uma simples representação por indexação. Ou seja, cada caractere ou palavra – ou cada componente léxico – poderia ser mapeado para um número distinto. Entretanto, tais números não teriam nenhuma conotação semântica. Assim, o arcabouço adotado de forma mais ampla para resolver este problema é mapear os componentes da língua para vetores em um espaço semântico, seguindo a hipótese distribucional. Como melhor detalhado no Capítulo Semântica com técnicas simbólicas, a hipótese distribucional tem como mote inferir significado a partir do contexto em que as palavras ocorrem. Apenas para ter uma ideia, considere, por exemplo o texto a seguir, em que a palavra "*bruble*" não pertence à língua portuguesa (até onde sabemos):

[1] https://github.com/features/copilot/
[2] https://openai.com/blog/openai-codex

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Exemplo 1.1:

> Sempre deixo as notificações do meu *bruble* desligadas. Mas no outro dia, estava escrevendo uma mensagem no meu *bruble* e as notificações de vários aplicativos apareceram na tela.

Pelo contexto, podemos inferir que a palavra "*bruble*" seria "celular" e é justamente nesta motivação que a semântica distribucional se coloca. Indo além, segundo Firth, o significado de uma palavra pode ser depreendido pelas palavras que coocorrem com ela, ideia difundida pelo slogan "*you shall know a word by the company it keeps*" (Firth, 1957), que, no contexto do significado das palavras, podemos adaptar para algo como "Diga-me com quem andas, e te direi quem és".

Embora contexto possa contemplar diversas definições, para a geração de modelos semânticos distribucionais, contexto é definido pela coocorrência de itens. A coocorrência pode ser traduzida para: itens que aparecem próximos uns dos outros ou ainda itens que aparecem em contextos similares.

Os modelos de linguagem mais recentes apresentam uma significativa sinergia com a hipótese distribucional. Por um lado, eles se fundamentam na hipótese distribucional, uma vez que assumem que o contexto pode ser usado para a predição de uma ou mais palavras; por outro lado, modelos de linguagem podem gerar as representações numéricas que sumarizem os contextos em que as palavras ocorrem, permitindo a investigação da hipótese distribucional em termos de similaridade. Nesta sinergia, os modelos de linguagem mais recentes que geram representações vetoriais de forma dinâmica se sobrepõem às limitações dos métodos distribucionais estáticos mais clássicos, uma vez que os vetores de um mesmo item podem ser diferentes dependendo do contexto em que ele aparece.

Mas antes de entrarmos em detalhes sobre os modelos de linguagem atuais, temos uma pergunta ainda mais básica a ser respondida: **O que é um modelo?** Um modelo é uma simplificação de um fenômeno complexo, no nosso caso, uma simplificação da língua que possa ser representada por ferramentas computacionais. Embora um modelo tente capturar as nuances do fenômeno real, justamente por ser uma simplificação, ele não tem a intenção de substituir o fenômeno real, mas representá-lo para auxiliar o nosso entendimento ou resolver algumas tarefas. Porém, idealmente, o modelo deve manter alguma consistência com o fenômeno real. Por isso, um modelo de linguagem deveria respeitar os princípios léxicos, sintáticos e semânticos, componentes essenciais de qualquer linguagem, natural ou não.

Também, um modelo deveria considerar o mesmo funcionamento do fenômeno real. Mas como a questão de como nosso cérebro processa e produz linguagem continua em aberto (Berwick; Chomsky, 2017), nos modelos de linguagem computacionais, assume-se que um texto escrito ou falado é oriundo de um processo de completação. Em suas primeiras abordagens, definia-se que um modelo de linguagem computacional deveria ser capaz de completar a próxima palavra em uma sequência, considerando todas as palavras que vieram antes. Por exemplo, considerando a sentença "Vamos completar o texto com a próxima …", um modelo poderia completá-la com "palavra". Atualmente, alguns modelos também podem considerar completar partes de uma sequência considerando palavras (ou *tokens*) que vieram antes ou depois do elemento que se deseja completar, seguindo uma abordagem inspirada no teste Cloze (Santos et al., 2002; Taylor, 1953). Por exemplo, seguindo o caso anterior, poderíamos ter "Vamos …o …com a próxima palavra", onde ... poderiam ser preenchidos com palavras. Um modelo de linguagem computacional não precisa

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol2/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

estar restrito a completar uma única palavra, mas sim uma sequência delas, independente de serem as próximas palavras, ou palavras em outras posições da sequência.

Nas próximas seções, vamos entender melhor como essas tarefas são abordadas em termos computacionais[3].

1.2 Modelos de Linguagem Probabilísticos

Em termos computacionais, a modelagem probabilística de linguagem é a tarefa que atribui uma probabilidade a uma sequência de palavras. Ou seja, o modelo assume que existe uma probabilidade associada à existência de uma sequência de palavras $p_{1:i}$, representada por $P(p_{1:i})$, onde i representa a posição da última palavra na sequência considerada. Usando a regra da cadeia da probabilidade, a fórmula pode ser definida como:

$$P(p_{1:i}) = P(p_1)P(p_2|p_1)P(p_3|p_{1:2})P(p_4|p_{1:3}) \dots P(p_i|p_{1:i-1}) \tag{1.1}$$

Apenas uma observação: multiplicações de valores menores que um podem fazer com que o resultado seja zero, considerando a limitação dos computadores em manipularem números em ponto flutuante. Chamamos esse problema de *underflow*. Para aliviá-lo, podemos usar log e somar os termos, ao invés de multiplicar:

$$\begin{aligned} \log P(p_{1:i}) = & \\ & \log P(p_1) + \log P(p_2|p_1) + \log P(p_3|p_{1:2}) + \\ & + \log P(p_4|p_{1:3}) + \dots + \log P(p_i|p_{1:i-1}) \end{aligned} \tag{1.2}$$

Observe que na fórmula, temos uma sequência de tarefas de predição de palavra, onde o objetivo é predizer uma palavra condicionando-a às palavras precedentes. Assim, pensando na completação discutida anteriormente, assumimos que a tarefa de completar uma sequência de palavras com uma próxima palavra é definida por uma distribuição de probabilidade condicional[4] das palavras que poderiam completar a sequência, dadas as palavras que vieram antes na sequência, ou seja:

$$P(p_i|p_1, \dots, p_{i-1}) \tag{1.3}$$

onde p_i é uma palavra do vocabulário, i é a sua posição na sequência, p_1 é a primeira palavra da sequência e p_{i-1} é a última palavra da sequência. Modelos de língua que seguem esta formulação são chamados de *modelos autorregressivos ou causais* e são frequentemente empregados para tarefas que envolvem geração de texto. A ideia é simples: (1) use o modelo probabilístico para escolher o próximo *token*; (2) adicione o *token* gerado na sequência de entrada; (3) repita. Mas no passo (1), quando falamos que um *token* é gerado pelo modelo, o que acontece, na verdade, é que um *token* é escolhido de acordo com uma distribuição de probabilidade aprendida pelo modelo. Tal distribuição de probabilidade é definida para um vocabulário, o conjunto de *tokens* que o modelo conhece.

Voltando ao nosso exemplo anterior, ele seria modelado pela seguinte distribuição de probabilidade condicional

$$P(p_i|\text{Vamos, completar, o, texto, com, a, próxima})$$

[3]Neste vídeo https://www.youtube.com/watch?v=Xz_qPyizGEY&t=1s o leitor pode se informar mais sobre o uso de aprendizado de máquina em PLN.

[4]Probabilidade condicional é a probabilidade de um evento ocorrer, dado que outro evento já ocorreu. A probabilidade condicional é representada por $P(A|B)$, o que significa a probabilidade do evento A ocorrer, dado que o evento B já ocorreu.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol2/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

onde

$$P(\text{palavra}|\text{Vamos, completar, o, texto, com, a, próxima})$$

poderia ter um valor de, digamos, $0,88$. No caso de uma palavra pouco provável, digamos, CHUTEIRA, esse valor poderia ser bem pequenino, digamos, $0,00001$ (por enquanto, assuma que esses valores vieram do além).

Entretanto, não é computacionalmente eficiente considerar toda a sequência anterior para predizer a próxima palavra na sequência. Embora modelos probabilísticos sejam apelativos, principalmente pela sua simplicidade, eles sofrem da "maldição da dimensionalidade": modelar a distribuição conjunta de, digamos, sequências de 10 palavras, com um vocabulário de 100.000 palavras, traz a enorme quantidade de $100.000^{10} - 1$ parâmetros.

Então, podemos simplificar ainda mais o modelo, assumindo a suposição de Markov (Markov, 1954), que dita, informalmente, que apenas o passado mais recente é importante para o futuro. Assim, considerando a suposição de Markov, assume-se que predizer a próxima palavra é independente das outras palavras na sequência, dada a última palavra vista. Ou seja,

$$P(p_i|p_{1..i-1}) \approx P(p_i|p_{i-1})$$

No nosso exemplo, consideraríamos apenas a palavra PRÓXIMA para predizer a palavra PALAVRA (desculpem a redundância), ou seja, $P(\text{palavra}|\text{próxima})$. Este modelo é conhecido como *bigrama*, por considerar apenas um par de palavras na probabilidade condicional. Generalizando, um unigrama consiste em considerar a probabilidade a priori de apenas uma palavra, $P(p_i)$, um bigrama consiste em considerar duas palavras $P(p_i|p_{i-1})$, um trigrama consiste em considerar as duas palavras anteriores $P(p_i|p_{i-1}, p_{i-2})$, e assim por diante. Generalizando ainda mais, um modelo n-grama é representado por $P(p_i|p_1, ..., p_{i-n})$.

Perceba que existe uma troca na decisão de que valor de n considerar. Enquanto valores menores de n tornam o modelo probabilístico mais eficiente de ser computado, por outro lado, eles perdem precisão. Considerando nosso exemplo, é mais fácil de predizer que p_i seria PALAVRA se pensarmos na sequência anterior completa. Olhando apenas para PRÓXIMA, a gama de palavras que fariam sentido vir depois é muito maior. Entretanto, conforme veremos a seguir, essas probabilidades precisam vir de algum lugar (não do além), e esse lugar são textos existentes (ou melhor dizendo, o *corpora*). Quanto maior for a sequência considerada, mais rara será a sua aparição no *corpora*, o que pode prejudicar o cálculo do valor de probabilidade para uma determinada sequência.

1.2.1 Estimando as probabilidades a partir de *corpora*

Para estimar as probabilidades das sequências de palavras, usaremos um conjunto de textos. Quanto maior e mais diverso o conjunto, maior é a chance dele conter muitas variações de sequências. Mas também, mais demorado será o seu processamento. Considerando a probabilidade frequentista, para calcular a probabilidade condicional $P(p_i|p_{1:i})$, podemos simplesmente usar contagem. Vamos começar de um modelo bigrama para depois generalizarmos. Neste caso,

$$P(p_i|p_{i-1}) = \frac{c(p_{i-1}, p_i)}{c(p_{i-1})}$$

onde $c(p_{i-1}, p_i)$ representa quantas vezes a sequência formada pelas duas palavras p_i e p_{i-1} apareceram nos textos, mais precisamente, quantas vezes p_{i-1} aparece antes de p_i, e $c(p_{i-1})$ representa quantas vezes a palavra p_{i-1} aparece no texto.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol2/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

De forma similar, para estimar as probabilidades de um modelo trigrama, temos que

$$P(p_i|p_{i-1}, p_{i-2}) = \frac{c(p_{i-2}, p_{i-1}, p_i)}{c(p_{i-2}, p_{i-1})}$$

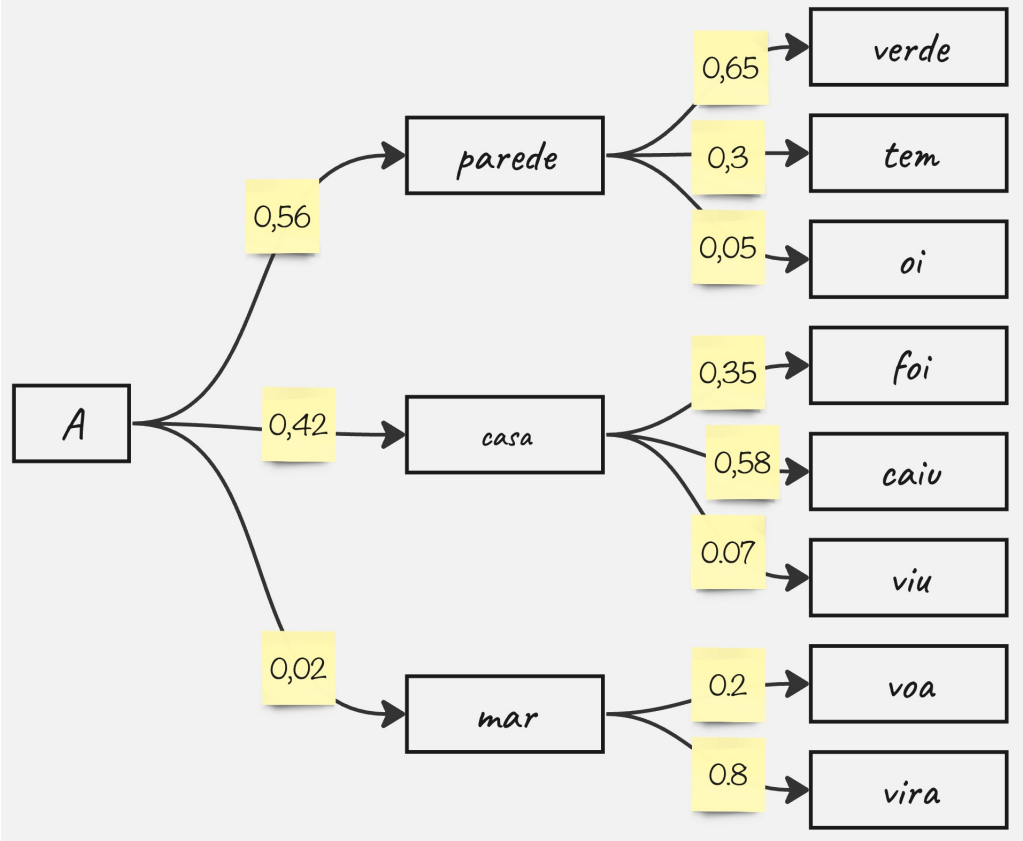
Ou seja, $c(\dots)$ representa quantas vezes um dado n-grama ocorreu no texto.

O modelo n-grama também serve para calcular as probabilidades mesmo de sequências de palavras que não apareceram no conjunto de treinamento. Ou seja, a probabilidade de uma sequência não vista de palavras será obtida a partir da concatenação de gramas menores que formam a sequência. Entretanto, perde-se informação ao não considerar contextos maiores. Outro problema é desconsiderar a similaridade entre palavras, que poderia servir para devolver probabilidades de palavras ou sequências não vistas durante o treinamento. Modelos de linguagem neurais tentam abordar esses problemas com métodos mais sofisticados de aprendizado de máquina do que simplesmente contagem.

1.2.2 Usando o modelo probabilístico

A escolha do próximo *token* conforme a probabilidade pode seguir diferentes algoritmos. O primeiro que pode nos vir à mente é um processo guloso, ou seja, a cada iteração, escolhemos o *token* com a maior probabilidade. Assim, para escolher a palavra p_t na iteração t, usamos $p_t = \arg\max_p P(p|p_{1:t-1})$. No exemplo da Figura 1.1, supondo que o *token* "A" já foi emitido, a busca gulosa escolherá a palavra "casa" e depois "caiu".

Figura 1.1: Exemplo de geração de sentença com a busca gulosa



https://brasileiraspln.ufscar.br/livro-pln-4ed-vol2/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Entretanto, um processo guloso pode trazer um sério problema, que é a falta de diversidade. Também, pode ser que uma escolha conjunta seja melhor do que uma escolha individual. Mas a escolha individual deixa a busca um pouco míope em relação ao que ainda está por vir. Assim, é bem comum utilizar outros mecanismos. Um desses outros mecanismos é a busca em feixe, que armazena possíveis escolhas, desde que esse armazenamento não ultrapasse o limite máximo de feixes. No exemplo da Figura 1.1 e assumindo um feixe de tamanho dois, a busca, além de guardar "(A, casa)" também guardaria "(A, parede)". Na próxima iteração, teríamos as seguintes possibilidades: (A, PAREDE, VERDE) e (A, CASA, CAIU), com a primeira opção tendo probabilidade $0.4 \times 0.9 = 0.36$ e a segunda possibilidade com probabilidade $0.5 \times 0.4 = 0.2$.

Ainda assim, pode ser difícil de garantir muita diversidade, tanto nas palavras geradas na mesma execução como na geração em diferentes execuções. Uma escolha aleatória do tipo $p_t \sim P(p_t|p_{1:t-1})$ poderia trazer a desejada diversidade. Porém, os resultados também podem ficar bem incoerentes (Holtzman et al., 2020). Um procedimento que ajuda um pouco é usar um valor de *temperatura* parametrizável. Valores mais altos de temperatura produzem saídas mais aleatórias, enquanto valores menores fazem com que as saídas sejam mais similares, ou seja, mais determinísticas. Outras abordagens também existem, incluindo amostragem das top-k palavras, com uma redistribuição da massa de probabilidade apenas entre essas k (Fan et al., 2018), ou amostragem baseada em um limiar p, que escolhe o menor conjunto de palavras cuja probabilidade acumulada exceda p (Holtzman et al., 2020).

1.3 Modelos de Linguagem Neurais

O uso de *n-grams* discutido na seção anterior é uma forma de generalizar e tornar eficiente o cálculo da probabilidade de uma sequência de palavras. Outra forma de atender às necessidades de generalização – ou seja, calcular uma probabilidade para uma sequência de palavras ao usar um modelo, mesmo que a sequência não tenha aparecido durante o treinamento do modelo[5] – é considerar que a probabilidade associada a um modelo de linguagem é uma função e "aprender" tal função. Redes Neurais (Goodfellow et al., 2016) são métodos de aprendizado de máquina conhecidos por sua propriedade de aproximação universal de funções. Ou seja, dada uma rede neural com ao menos uma camada escondida e um número suficiente de neurônios, ela é um aproximador universal de funções contínuas no espaço de interesse (Hornik et al., 1989). Caso você queira entender melhor como funciona uma rede neural, os capítulos 5 e 6 de (Goodfellow et al., 2016) são uma boa introdução (dentre muitas outras referências).

1.3.1 Um Contexto Histórico

A ideia de usar redes neurais para aprender funções que representem modelos de linguagem pode parecer recente, mas não é. Na verdade, as primeiras tentativas datam do início da década de 90, com o trabalho de Miikkulainen; Dyer (1991). Ainda na década de 90, também foram propostas técnicas baseadas em redes neurais para prever a probabilidade do próximo caractere (Schmidhuber; Heil, 1996). Mas os modelos que mais se assemelham aos modelos de linguagem neurais da era de *Deep Learning* (redes neurais profundas) foram propostos no início dos anos 2000, de forma independente, com os trabalhos *Can artificial*

[5] Lembre que, em um modelo de linguagem probabilístico, treinar o modelo é encontrar as probabilidades por contagem.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol2/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

neural network learn language models? (Xu; Rudnicky, 2000) e *A neural probabilistic language model* (Bengio et al., 2003).

Enquanto o primeiro caso usava uma forma limitada de rede neural, sem camadas escondidas e limitando a predição a apenas uma palavra, ou seja, modelando apenas unigramas e bigramas, o segundo caso já apresentava várias características e fundamentos encontrados nos modelos de linguagem neurais modernos. A proposta do primeiro trabalho era aprender (i) funções de representações distribuídas para cada palavra $P(w)$, que consideraria a vizinhança das palavras nos textos de treinamento, bem na linha do que vimos no Capítulo Representação vetorial e semântica distribucional. Mas além da probabilidade das palavras, o modelo também aprenderia de forma simultânea (ii) a função de probabilidade associada a uma sequência de palavras, a partir das probabilidades das palavras. Assim, mesmo que no momento de usar o modelo aparecesse uma sequência de palavras não vista durante o treinamento, ainda seria possível obter a probabilidade da sequência, a partir das palavras e sequências similares vistas durante o treinamento.

No modelo proposto, a rede neural é utilizada para predizer a próxima palavra, dadas as palavras anteriores. Para tanto, seus pesos são treinados para aprender a função de probabilidade do modelo de linguagem a partir da maximização da log-verossimilhança dos dados de treinamento. De forma mais específica, um exemplo de treinamento é uma sequência de palavras de tamanho T, $p_1 \ldots p_T$, com cada palavra $p_i \in V$, onde V é um vocabulário finito de palavras de uma língua. O objetivo da rede neural é representar uma função $f(p_i, \ldots, p_{i-n+1}) = \hat{P}(p_i|p_1^{i-1})$, onde n é o tamanho de uma janela de contexto. Assim como no modelo puramente probabilístico, pode-se obter um modelo da probabilidade conjunta de sequências de palavras a partir do produto destas probabilidades condicionais. A função $f(p_i, \ldots, p_{i-n+1})$ é decomposta em duas partes: (i) um mapeamento C de qualquer palavra $p_i \in V$ para um vetor $C(p_i) \in \mathbb{R}$ – um ***embedding***[6] da palavra; e (ii) uma função g que mapeia uma sequência de vetores, capturados a partir de C, ou seja, $C(p_{i-n+1}, \ldots, C(p_{i-1})$ para uma distribuição de probabilidade condicional da próxima palavra p_i. A saída da função g é um vetor cujo k-ésimo elemento estima a probabilidade $\hat{P}(p_i = k|p_1^{i-1})$. Ou seja, a função f é uma composição das funções g e C: $f(k, p_{i-1}, \ldots, p_{i-n+1}) = g(k, C(p_{i-1}), \ldots, C(p_{i-n+1}))$. A função g será parametrizada pelos pesos ω aprendidos pela rede neural. Os parâmetros do modelo são $\Theta = (C; \omega)$, descobertos a partir da minimização da função de custo $L = \frac{1}{I} \sum_i \log f(p_i, p_{i-1}, \ldots, p_{i-n+1}; \Theta) + R(\Theta)$.

O que aconteceu com esse modelo para ele não ficar tão famoso como os modelos de linguagem neurais atuais? O treinamento da tal rede neural era extremamente ineficiente e impraticável na época, um problema que começou a ser resolvido alguns anos depois com o advento das Unidades de Processamento Gráfico (GPUs). As GPUs ajudaram a impulsionar a era do *Deep Learning* (Goodfellow et al., 2016) ao focarem na realização de cálculos matriciais (tudo que uma rede neural quer) em tempos muito menores do que se o mesmo cálculo fosse feito em uma Unidade Central de Processamento (CPU).

1.3.2 Modelos de Linguagem Neurais Modernos

Considerando as limitações que discutimos no Capítulo Representação vetorial e semântica distribucional ao se definir *embeddings* de forma estática, vários métodos desenvolvidos a partir de 2017 passaram a construir *embeddings* de forma **dinâmica**, considerando o contexto da sentença no momento do uso, e por isso comumente denominados de **embeddings**

[6] Uma representação vetorial densa de baixa dimensionalidade, aprendida a partir de textos. Veja mais no Capítulo Representação vetorial e semântica distribucional.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol2/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

contextualizados. Isso quer dizer que as unidades de representação (*tokens*) podem ter *embeddings* distintos, definidos no momento em que eles são aplicados. Considere, por exemplo, as sentenças do Exemplo 1.2:

Exemplo 1.2:

1. Sentei no banco da praça.
2. O banco estava sem notas de R$ 200,00.
3. O banco estava super cheio hoje!

A palavra "banco" na sentença 1) evoca mais o sentido de assento, embora também seja possível pensar em outros significados. A sentença 2) evoca mais o sentido de estabelecimento comercial financeiro. A sentença 3), apesar de evocar mais o segundo sentido, também poderia estar falando de um assento cheio de pessoas. Sendo assim, uma lista estática de palavras e seus *embeddings* falharia em retornar representações distintas para estas diferentes interpretações.

Considere ainda o exemplo Exemplo 1.3:

Exemplo 1.3:

> Em frente à agência do banco de Pineapólis, existe um banco amarelo que data da década de 50, onde várias pessoas famosas já pararam para descansar e algumas vezes entoar uma melodia.

Observe que a palavra "banco" aparece duas vezes na mesma sentença, com dois significados distintos. Ainda assim, um método de geração de *embeddings* contextualizados deve ter a habilidade de devolver representações vetoriais distintas para os dois *tokens*.

Para tanto, a unidade de representação é associada a um *embedding* a partir do **contexto** corrente em que ela aparece, onde contexto, em geral, é definido nos modelos de linguagem por uma sequência de *tokens* que aparecem antes e depois do *token* em questão. No exemplo anterior, teríamos *embedding* distintos para os diversos "bancos" mencionados. Na verdade, o *embedding* poderia diferir até mesmo para *tokens* do tipo "banco" com a mesma semântica, devido aos diferentes outros *tokens* que aparecem em seus contextos. Entretanto, ainda se espera que quando mais próxima for a semântica do *token*, mais próximos fiquem os vetores no espaço vetorial.

Uma outra vantagem associada aos *embeddings* contextualizados é a possibilidade de representar informação que vai além do idioma. Esses *embeddings* são chamados de *cross-lingual* (Agirre, 2020). Ou seja, é possível que os *embeddings* associados às palavras "mãe" e "*mother*" estejam próximos no espaço vetorial, mesmo que ambas as palavras estejam em idiomas distintos.

Para que os *embeddings* de um *token* sejam gerados conforme o contexto dinâmico em que aparecem, a forma de recuperação e de armazenamento precisam ser diferentes daquelas que discutimos com os *embeddings* estáticos. Lá, poderíamos armazená-los em uma tabela e recuperá-los pela indexação da palavra. Já os *embeddings* contextualizados são recuperados a partir de uma função que tem como entrada a sequência em que a unidade de representação de interesse está inserida. Por exemplo, para devolver como saída o *embedding* da palavra *"banco"* a partir da sentença *"Sentei no banco da praça."*, teremos $\text{emb}_{\text{banco}} = f(e_{\text{sentei}}, e_{\text{no}}, e_{\text{banco}}, e_{\text{da}}, e_{\text{praça}})$, onde f é a função de geração do *embedding* e $\text{emb}_{palavra}$ é a sua saída. Cada palavra que será entrada da função precisa primeiro ser transformada para uma representação vetorial ($e_{palavra}$). Outra observação importante é

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol2/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

que, no nosso exemplo, a própria palavra é entrada da função. Nem sempre isso acontece, para evitar a influência da própria palavra na representação gerada.

A função f pode assumir diferentes formas. Uma possibilidade seria simplesmente recuperar os *embeddings* estáticos de cada palavra no contexto e executar alguma forma de agregação, conforme discutido no Capítulo Representação vetorial e semântica distribucional.

Porém, temos alguns problemas em simplesmente usar uma função de agregação. Um deles é que não teríamos diferença entre os *embeddings* de um *token* e os *embeddings* da sentença em que ele está inserido. Com essa forma simplificada de simplesmente agregar os *embeddings* estáticos, tanto a palavra "banco" como a palavra "praça" na sentença acima, teriam a mesma representação final. Mesmo que os *tokens* de interesse fossem removidos da entrada, outros problemas surgiriam, incluindo a falta de consideração com a ordem das palavras e palavras semanticamente distintas, porém lexicalmente idênticas, tendo a mesma representação em sentenças distintas.

Assim, torna-se necessário considerar outras funções mais elaboradas. Mas na dificuldade de se definir que função seria essa, por que não descobri-la **automaticamente**? Essa é a ideia da geração de *embeddings* contextualizados a partir de redes neurais.

A aplicação de *embeddings* contextualizados para abordar tarefas de PLN inclui dois aspectos: a geração dos *embeddings* e a sua utilização em tarefas finais. Dois principais métodos para a geração de *embeddings* contextualizados se destacaram entre 2017 e 2023: as redes neurais recorrentes – incluindo CoVe (McCann et al., 2017) (*Context Vectors*) e ELMo[7] (*Embeddings from Language Models*) (Peters et al., 2018)) – e os **Transformers** (Vaswani et al., 2017a) – incluindo BERT [8](*Bidirectional Encoder Representations from Transformers*) (Devlin et al., 2019) e GPT [9](*Generative Pre-trained Transformer*) (Brown et al., 2020). Vamos aqui seguir a ordem cronológica, primeiro falando dos modelos baseados em redes recorrentes, que surgiram primeiro, para depois falarmos dos modelos baseados em Transformers.

1.3.2.1 Modelos de Linguagem com Redes Recorrentes

Representar sequências de tamanhos variáveis é uma habilidade essencial para modelar a linguagem: sentenças não são obrigadas a conterem uma quantidade fixa de palavras; e, a ordem em que as palavras aparecem pode fazer toda a diferença para a sua sintaxe e sua semântica. Redes Neurais recorrentes abordam ambos os aspectos: aceitam entradas de tamanho variável e consideram a ordem dos componentes da entrada para induzir um vetor que represente uma sequência. Veja um esquema de uma rede neural recorrente na Figura 1.2.

De forma abstrata, um modelo de linguagem baseado em redes recorrentes opera gerando uma palavra a partir de uma sequência de palavras anteriores, seguindo os passos abaixo:

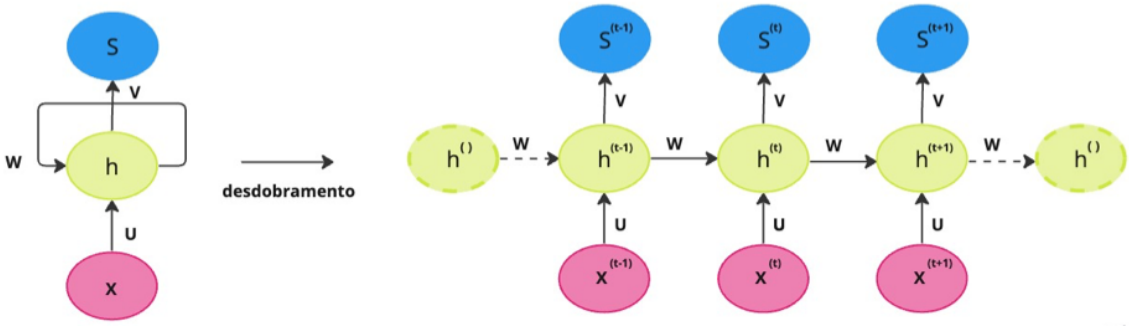
1. Calcula-se o vetor de *embedding* $h_t^0 = X_t\mathbf{E}$, onde $\mathbf{E}$ é uma matriz de dimensão $|V \times N|$, X_t é um vetor *one-hot*[10] do tamanho do vocabulário, ou seja, $|1 \times V|$ representando uma palavra, e t representa a t-ésima palavra da sequência sendo gerada

[7]Inspirado pelo nome de um dos personagens de Vila Sésamo.
[8]Também inspirado pelo nome de um dos personagens de Vila Sésamo.
[9]Até onde sabemos, nenhum personagem de Vila Sésamo tem esse nome.
[10]Um vetor *one-hot* tem uma posição apenas igual a 1 e as demais são zero. A posição 1 se refere a exatamente um elemento.

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol2/

Figura 1.2: Esquema clássico de uma rede recorrente, com X representando a entrada, h representando o estado oculto e S representando a saída da rede. À esquerda da figura, temos o esquema físico da rede, demonstrando como ela é implementada. Observe que a entrada se conecta ao estado oculto por meio de uma matriz de pesos U, o estado recorrente se conecta a si mesmo por meio de uma matriz de pesos W e o estado oculto também se conecta à saída por meio da matriz de pesos V. À direita da figura, temos a versão da rede conforme a sua operação em tempo de execução: X^t representa uma unidade da entrada (por exemplo, um *token*) na posição t da sequência de entrada. As matrizes de peso são fixas para todas as posições.



2. Calcula-se a saída da camada escondida $h_t^1 = fn\left(\mathbf{W_h}\left[\begin{smallmatrix} h_t^0 \\ h_{t-1}^1 \end{smallmatrix}\right]\right)$, onde $\left[\begin{smallmatrix} h_t^0 \\ h_{t-1}^1 \end{smallmatrix}\right]$ representa a concatenação dos vetores associados à saída da camada escondida fisicamente anterior (h_t^0) e da camada escondida do instante anterior (logicamente anterior) (h_{t-1}^1), $\mathbf{W_h}$ é a matriz de pesos da camada escondida, e fn é uma função de ativação, por exemplo, a tangente hiperbólica. Este passo pode se repetir diversas vezes, dependendo de quantas camadas escondidas a rede tiver. O sobrescrito indica a camada da rede.
3. Calcula-se a saída $y_t = \mathbf{W_o} h_t^1$, onde $\mathbf{W_o}$ representa a matriz de pesos da camada de saída.
4. Calcula-se a distribuição de probabilidade $p_t = \text{softmax} y_t$.
5. Resgata-se a palavra com o maior valor de probabilidade na tabela *one-hot*.
6. O processo continua até encontrar um *token* de fim de sequência, ou até alcançar uma saída máxima.

Pensando em uma geração *token* a *token*, é necessário ter algum *token* de início, que represente a camada anterior, para o primeiro *token*. Ele servirá para indicar a camada logicamente anterior usada, (h_{t-1}^1). As matrizes de pesos são os componentes aprendidos na rede. Para o aprendizado, pode-se considerar um conjunto de textos e fazer a tarefa de predição ser devolver a palavra correta na t-ésima posição, para t de 1 até um valor qualquer.

1.3.2.1.1 *Embeddings from Language Models* – ELMo

O ELMo é um modelo de linguagem que opera em uma rede neural recorrente [11] com várias camadas. Assim, cada camada pode ser usada para gerar uma representação con-

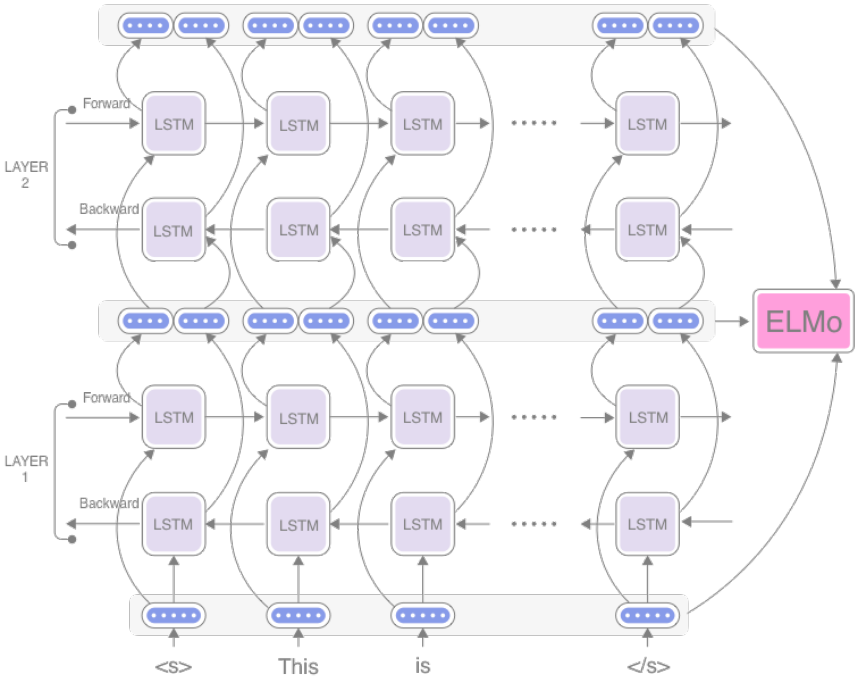
[11] O ELMo, assim como vários outros métodos de PLN, utiliza um tipo especial de rede recorrente chamada de *Long Short-Term Memory* (Hochreiter; Schmidhuber, 1997). LSTMs conseguem manipular entradas de tamanhos maiores do que redes recorrentes tradicionais, pois o que será aprendido, esque-

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol2/

textualizada de um *token*. As camadas de redes recorrentes do ELMo olham para a frente e para trás na sentença (são chamadas de redes recorrentes bidirecionais), dando origem a duas representações, uma para cada direção. Então, cada *token* pode ter um conjunto de representações, mais precisamente $2L + 1$ representações, onde L é a quantidade de camadas da rede. A multiplicação por 2 é devido às duas direções. E de onde vem o 1? É que o ELMo também inclui nesse conjunto de representações a entrada não contextualizada do *token*. Lembra que falamos antes que de todo modo os geradores de *embeddings* contextualizados devem iniciar por alguma representação vetorial? Mesmo os métodos que geram representações contextualizadas, precisam ter de onde começar. Então, o ELMo também inclui a representação "descontextualizada" e_{t_i}, que é a representação de entrada do *token*, como uma possível representação. Ou seja, cada camada $j \in \{1, 2, \ldots, L\}$ da rede produz as representações $\text{emb}_{t_i,j} = (\overleftarrow{h}_{t_i,j}, \overrightarrow{h}_{t_i,j})$ para o *token* t_i. Assim, o *token* terá um conjunto $\text{emb}_{t_i} = (e_{t_i}, \{\overleftarrow{h}_{t_i,1}, \overrightarrow{h}_{t_i,1}, \ldots, \overleftarrow{h}_{t_i,L}, \overrightarrow{h}_{t_i,L}\})$ de possíveis representações. A Figura 1.3 exibe um esquema da arquitetura do ELMo. O aprendizado de um modelo de linguagem baseado em redes recorrentes segue o algoritmo *backpropagation through time* (Werbos, 1990) a partir de um conjunto enorme de textos.

Figura 1.3: Esquema da arquitetura do ELMo.



Fonte: http://www.realworldnlpbook.com/blog/improving-sentiment-analyzer-using-elmo.html

1.3.2.1.2 Utilização dos *embeddings* do ELMo

Mas como podemos usar esses *embeddings* para resolver uma tarefa? Por exemplo,

cido ou devolvido como saída é controlado durante o aprendizado por um tipo especial de rede neural chamada de *gate*. As redes recorrentes tradicionais diferentes das LSTMs frequentemente falham em manipular sequências longas, pois uma informação de longa distância pode se perder ao longo das várias recorrências, um problema conhecido como "gradiente que some".

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol2/

suponha que a tarefa seja classificar uma publicação em uma rede social como sendo um comentário tóxico ou não[12]. Essa é uma tarefa de **classificação**. Uma forma de resolvê-la é treinar um classificador, que receberá um texto e informará se esse texto possui conteúdo tóxico ou não. Tanto para treinar o classificador como para usá-lo, a entrada textual precisa ser transformada para uma informação numérica, que é a língua que o computador entende. No caso que estamos falando aqui, a informação numérica será obtida justamente a partir dos *embeddings*. Com o ELMo, podemos obter esses *embeddings* de duas formas: (i) juntando todos os elementos acima em um único vetor, por exemplo, os somando, ou seguindo uma operação mais simples, (ii), por exemplo, selecionando somente aqueles que estão na última camada, ou seja, considerando somente $(\overleftarrow{h}_{t_i,L}, \overrightarrow{h}_{t_i,L})$. Mais precisamente, o *embedding* que o ELMo gera para um *token* é definido por

$$emb'_{t_i} = \gamma^{\text{tarefa}} \sum_{j=0}^{L} \mathcal{S}_j^{\text{tarefa}} [\overleftarrow{h}_{t_i,j}; \overrightarrow{h}_{t_i,j}]$$

onde variamos j de 0 até L, para incluir o *embedding* da primeira camada (o descontextualizado), [_;_] indica uma operação de concatenação, γ^{tarefa} é um hiperparâmetro relacionado à tarefa, e $\mathcal{S}_j^{\text{tarefa}}$ são os pesos da camada, normalizados por uma função Softmax[13].

Observe que você pode experimentar outras variações. Por exemplo, podemos usar as camadas mais próximas da saída apenas, fazendo o somatório começar em $j = k$, onde k é uma posição intermediária na rede. Também é possível concatenar o *embedding* descontextualizado com a saída da última camada.

1.3.2.1.3 *Embeddings* de sentenças

Até agora falamos de *embeddings* de *tokens*. Mas a maioria das tarefas considera entradas que são frases, ou um texto, ou seja, uma sequência de *tokens*. Na verdade, embora seja possível recuperar os *embeddings* de qualquer tipo de unidade de representação a partir do ELMo, incluindo caracteres, palavras, frases, textos, a saída default das implementações mais comuns[14] são os *embeddings* de uma sentença. Eles são obtidos a partir de uma operação de amostragem por média (*mean pooling*) dos *embeddings* de *tokens* da última camada da rede, conforme discutimos antes. Perceba que isso é bem diferente do que a função simples que mencionamos antes, uma vez que as representações vetoriais passam por várias transformações matemáticas dentro da rede neural.

1.3.2.1.4 ELMo para português

Como usual, o modelo ELMo foi originalmente treinado e avaliado na língua inglesa. Mas existem versões deste modelo treinadas para as variantes brasileira e europeia do português (Rodrigues et al., 2020), disponibilizadas na biblioteca oficial do ELMo, a Allen NLP [15]. O modelo foi treinado em tarefas de similaridade sintática e comparado com sucesso a representações estáticas também treinadas para o português.

[12] Um texto é considerado um comentário tóxico quando ele contém expressões ou intenções indesejáveis de desagrado, violência psíquica, física e emocional, crueldade, negativismo, manipulação, egoísmo etc. Veja mais sobre esse assunto no Capítulo PLN em Redes Sociais.

[13] A função de Softmax, também conhecida como regressão logística multinomial, é definida como $\sigma(\mathbf{z})_i = \frac{e^{z_i}}{\sum_{j=1}^{|V|} e^{z_j}}$ e tem como característica principal receber |V| valores reais e convertê-los para uma distribuição de probabilidade de |V| saídas possíveis, onde |V| é o tamanho do vocabulário.

[14] Por exemplo, usando o TensorFlow: https://tfhub.dev/google/elmo/2.

[15] https://allenai.org/allennlp/software/elmo

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol2/

1.3.2.2 Modelos de Linguagem baseados em Transformers

Embora as redes recorrentes possam resolver tarefas sequenciais e não demandem entradas de tamanho fixo, o que parece perfeito para tarefas de PLN, elas têm um grande problema: sua característica sequencial faz com que elas não sejam paralelizáveis, ou seja, uma rede recorrente não pode ser separada em vários componentes para serem treinados em paralelo. Tal característica torna o treinamento das redes recorrentes bem ineficiente, o que acarreta em um outro problema: as entradas não podem ser muito grandes e nem exigirem uma dependência de longa distância. Mesmo que as redes do tipo *Long Short-Term Memory* (LSTMs) tenham aliviado o problema da dependência de longa distância com o uso dos mecanismos de *gate*, eles acarretam em redes com mais parâmetros para serem treinados, o que de novo nos leva à questão da ineficiência.

Bem, esse é um problema para modelar línguas com redes neurais recorrentes, uma vez que um texto pode ser enorme e ainda assim trazer componentes importantes lá no início para serem usados lá no fim. Considere, por exemplo, a frase do Exemplo 1.4:

Exemplo 1.4:

> A garota de blusa amarela com uma frase em que os verbos estavam em letras pretas, que andava tão rápido e nunca em linha reta, a ponto de passar pelas nossas vistas como se fosse quase um furacão, tinha na parte de trás da sua blusa uma frase atribuída a Gandhi: "Acreditar em algo e não vivê-lo, é desonesto".

Caso quiséssemos saber qual é a cor das letras em que a palavra "Acreditar" foi escrita, teríamos que conectar "Acreditar" com "verbo" e ver no início da frase que eles são escritos em preto. Claro que esse é um exemplo exagerado, mas pare para contar quantas palavras estão entre a cor da letra e o primeiro verbo da frase de Ghandi. Uma rede recorrente teria que aprender tais conexões, apesar da distância.

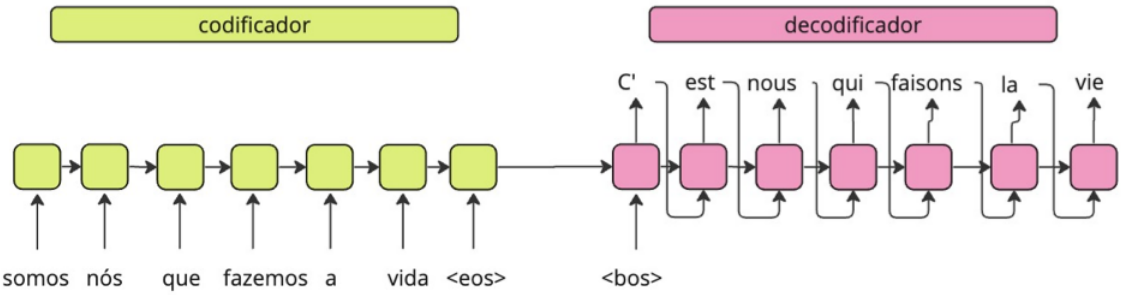
Outro ponto que precisamos mencionar antes de chegar aos Transformers (Vaswani et al., 2017a) do título, que não são os mesmos dos filmes e brinquedos, mas que guardam muitas semelhanças, são as tarefas de PLN em que modelos de linguagem são costumeiramente usados: as tarefas de geração de sequências, no nosso caso, sequências de letras, palavras, textos. Tais sequências não são meramente concatenações de palavras, pois elas devem obedecer a princípios sintáticos e semânticos. Ainda, a geração de sequências não envolve apenas gerar textos do zero, ou completar frases, mas também gerar sequências a partir de outras sequências. Neste caso, a tarefa é chamada de forma genérica na literatura de *sequence-to-sequence* ou "seq2seq" (Cho et al., 2014; Sutskever et al., 2014). Por exemplo, as tarefas de tradução automática, sumarização, respostas a consultas complexas, entre outras, requerem que a entrada seja um texto (uma sequência) e que a saída também seja um texto (outra sequência).

Do ponto de vista da modelagem da arquitetura de uma rede neural para resolver tarefas "seq2seq", o mais comum é considerar dois grandes componentes: o primeiro, chamado de *encoder* ou codificador, é responsável por processar a sequência de entrada – para nós a sequência de letras, *tokens*, palavras, frases, e codificá-la como um vetor de números (as redes neurais gostam de números), chamado de *vetor de contexto*; o segundo componente, chamado de *decoder* ou decodificador, é responsável por receber e processar o vetor de contexto e transformá-lo na sequência de saída – uma sequência de letras, *tokens*, palavras, frases. Veja um diagrama de alto nível deste processo na Figura 1.4, que exemplifica uma tarefa de tradução automática.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol2/

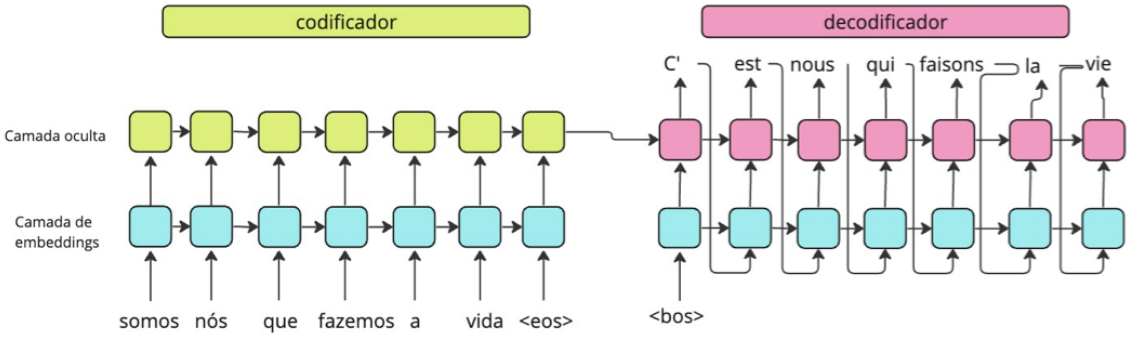
©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Figura 1.4: Exemplo do esquema básico de uma arquitetura seq2seq, exemplificada com uma tarefa de tradução de uma sentença em português para o francês. Os *tokens* "<eos>" e "<bos>" indicam o fim e início de sentença. Perceba que, assim como em uma rede recorrente tradicional, a saída da iteração anterior serve como entrada para a próxima iteração. No codificador, cada iteração apenas armazena informação, enquanto no decodificador, existe uma entrada e uma emissão de *token* a cada iteração.



O codificador é uma rede neural – ou várias delas – e o mesmo vale para o decodificador. Logo, as redes neurais têm seus parâmetros aprendidos com o foco de receber uma sequência-fonte e devolver a sequência-alvo desejada. Outro ponto importante é que a rede precisa da representação numérica dos itens na sequência de entrada. Assim, ou podemos ter uma camada inicial que faz a transformação de um vetor *one-hot* para um vetor de *embeddings*, ou podemos recuperar os *embeddings* das palavras a partir de um modelo pré-treinado. A Figura 1.5 traz um exemplo da arquitetura anterior, detalhando a camada de *embeddings*.

Figura 1.5: Esquema básico de uma arquitetura seq2seq, exemplificada com uma tarefa de tradução de uma sentença em português para o francês. Os *tokens* "<eos>" e "<bos>" indicam o fim e início de sentença. Aqui, a camada de *embeddings* está explícita.



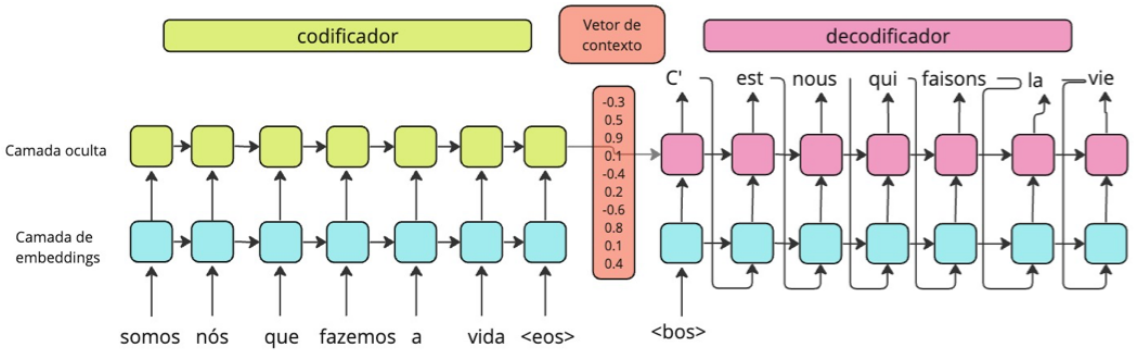
Bem, o codificador e o decodificador podem muito bem ser redes recorrentes, com uma ou mais camadas, do tipo LSTM, ou alguma outra variação. Em tais casos, o último estado escondido da rede, no sentido lógico, ou seja, obtido após processar o último item da sequência, será o vetor de contexto. A Figura 1.6 explicita o vetor de contexto, que antes estava representado de forma implícita como a seta de ligação entre o codificador e o decodificador.

Aqui temos um problema: é complicado assumir que esse último estado escondido,

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol2/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Figura 1.6: Esquema básico de uma arquitetura seq2seq, exemplificada com uma tarefa de tradução de uma sentença em português para o francês, com o vetor de contexto conectando o estado final do codificador e a entrada do decodificador.



codificado como o vetor de contexto, conseguirá capturar todos os aspectos necessários para resolver a tarefa, ainda mais se a sequência de entrada for grande. Para lidar com este problema, pesquisadores elaboraram uma nova estratégia, chamada de *mecanismo de atenção*[16] (Bahdanau et al., 2015; Luong et al., 2015).

1.3.2.2.1 Atenção!

O objetivo do mecanismo de atenção – na verdade, um conjunto adicional de parâmetros para a rede – é que os itens mais relevantes da entrada recebam uma valoração maior no vetor de contexto. Mas outros itens também podem receber algum valor. A ideia é mais ou menos assim, e deixe de lado a língua natural só por um minuto, para um exemplo mais abstrato: suponha que você quer aprender a assar um bolo de chocolate. Vamos chamar "Assar o bolo de chocolate" de consulta. Você pode pegar o livro de receitas da sua avó para te ajudar. O livro é composto de diversas receitas, que vamos chamar de chaves. O que você quer é encontrar a receita mais adequada, e para isso, todas as receitas vão receber alguma valoração. A receita do bolo de chocolate perfeito deve ter um valor maior em relação aos demais, mas uma receita de bolo de chocolate com morango, também pode receber alguma relevância. Mas uma receita de Tiramissu deveria ter uma relevância bem pequenininha. O mecanismo de atenção segue essa ideia: a saída é a consulta, a informação que precisa ser gerada a partir da entrada, as chaves. Para definir a chave mais relevante, são calculados pesos de atenção, que definirão o vetor de contexto.

Para considerar a relevância de diferentes itens na entrada, o codificador não considera que apenas o último estado escondido da rede será o vetor de contexto, mas que todos os estados escondidos, ou seja, todos os estados obtidos após o processamento de cada item da sequência, também podem participar do vetor de contexto. Mas agora o decodificador terá mais trabalho, pois ele precisará decidir o que fazer com esses vários vetores antes de gerar os itens da saída, e ainda considerando que é necessário focar nas partes mais relevantes para a resolução da tarefa. Assim, antes de gerar a saída pelo decodificador, são executados os seguintes passos:

1. Computar uma pontuação para cada estado escondido, também chamada de pontu-

[16] Alguns consideram que o mecanismo de atenção é uma forma de explicação, enquanto outros discordam veementemente disso. Veja as discussões: (Bibal et al., 2022; Jain; Wallace, 2019; Wiegreffe; Pinter, 2019).

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol2/

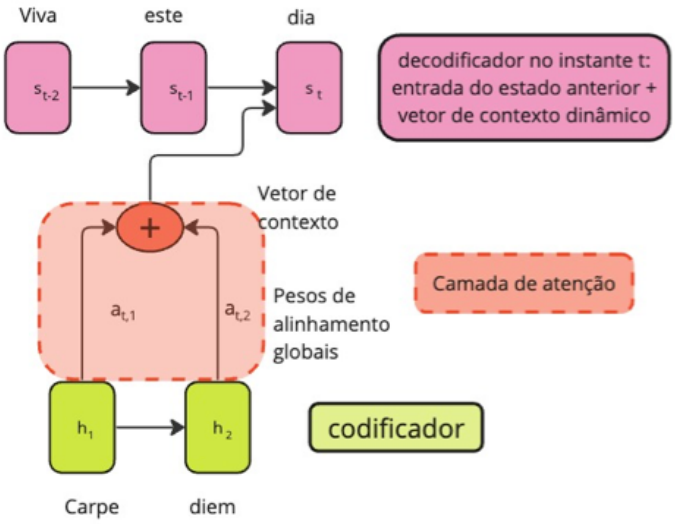
ação de alinhamento, seguindo a Equação 1.4.

2. Passar as pontuações combinadas – por concatenação, em geral, ou o vetor resultante da equação anterior, pensando em termos de representações matriciais – por uma função de "softmax", para capturar alguma noção de probabilidade da relevância, produzindo os pesos de atenção.
3. Multiplicar cada estado escondido (lembrando que ele é representando por um vetor) pelos pesos de atenção, de forma a tornar os estados escondidos mais relevantes com valores ainda maiores, e obter o efeito oposto para os estados menos relevantes. O resultado deste passo será o vetor de contexto.

$$att = \mathbf{W}_{\text{combinado}} \times \tanh(\mathbf{W}_{\text{decod}} \times \mathbf{H}_{\text{decod}} + \mathbf{W}_{\text{codif}} \times \mathbf{H}_{\text{codif}}) \tag{1.4}$$

onde $\mathbf{W}_{\text{codif}}$ e $\mathbf{W}_{\text{decod}}$ representam matrizes de pesos (parâmetros) aprendidos e $\mathbf{H}_{\text{decod}}$ representam estados escondidos. Observe a semelhança com uma rede neural de uma camada escondida com a função de ativação de tangente hiperbólica. Veja um exemplo ilustrativo na Figura 1.7.

Figura 1.7: Ilustração do mecanismo de atenção aditivo para a tradução da expressão em latim *carpe diem* para a expressão em português "Viva este dia".



Fonte: Adaptado de (Bahdanau et al., 2015)

O mecanismo de atenção calculado desta forma também é chamado de mecanismo aditivo ou mecanismo de Bahdanau, proposto em Bahdanau et al. (2015). Existe um segundo tipo de atenção, proposto em (Luong et al., 2015), chamado de mecanismo de atenção multiplicativo. As diferenças principais são que o decodificador produz um estado intermediário a partir do estado escondido anterior antes de calcular os pesos de atenção, e as pontuações de alinhamento podem ser de três tipos: (i) multiplicando os estados escondidos do codificador e decodificador apenas, ou seja, $att = \mathbf{H}_{decod} \times \mathbf{H}_{codif}$, (ii) multiplicando uma matriz de pesos aprendidos ao resultado da multiplicação em (i), ou seja, $att = \mathbf{W} \times \mathbf{H}_{decod} \times \mathbf{H}_{codif}$, e (iii) somando os estados escondidos do codificador e decodificador, que são multiplicados por uma matriz de pesos, passam pela função de ativação da tangente hiperbólica e são finalmente multiplicadas a uma matriz de pesos, $att = \mathbf{W} \times tanh(\mathbf{W}_{combinado}(\mathbf{H}_{decod} + \mathbf{H}_{codif}))$. Este último caso é o mais similar ao mecanismo aditivo, mas os estados escondidos compartilham uma matriz de pesos, diferente da

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol2/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Equação 1.4. Ao final, o vetor de contexto é concatenado com o estado do decodificador no instante anterior, para produzir uma nova saída.

O mecanismo de atenção apresentado até agora é chamado de mecanismo de atenção geral, uma vez que ele tenta encontrar os componentes da entrada que são mais relevantes para gerar a saída. Transformers fazem uso de um mecanismo de atenção adicional, chamado de *auto-atenção*, em que a captura da relevância é feita entre os elementos de uma mesma sequência, usualmente da entrada.

1.3.2.2.2 Finalmente os Transformers

Mas vamos finalmente entender o que são esses tais Transformers, uma arquitetura de rede neural proposta em 2017 e que faz uso do mecanismo de atenção, entre outros componentes conhecidos de redes neurais, e cujo esquema está representado na Figura 1.9. Um Transformer tem dois componentes principais, adivinhem só, um componente de codificação e um componente de decodificação. Entretanto, diferente do que falamos antes nos modelos seq2seq, Transformers não são constituídos por redes neurais recorrentes. Com isso, é possível paralelizá-los e alcançar tempos de treinamento mais eficientes para modelos de linguagem do que usando redes recorrentes. Mas não é só isso: o uso extensivo de mecanismos de atenção, combinados a outros componentes, fez dos Transformers e suas diversas variações o estado da arte em diversas tarefas de PLN, ao menos até o momento de escrita deste livro (Wolf et al., 2020). Eles são o componente principal dos modelos de linguagem em larga escala (em inglês, *large language models* ou LLMs) que deram o que falar no início do ano de 2023, principalmente com a vasta disponibilidade de agentes de conversação e suas interfaces de programação de aplicações[17].

Vamos entender do que esses codificadores e decodificadores são compostos, já que não são redes recorrentes. O codificador é, na verdade, uma pilha de sub-codificadores, enquanto o decodificador é uma pilha de sub-decodificadores. No artigo original, essas pilhas tinham seis componentes, mas poderia ser qualquer outra quantidade. Os sub-codificadores possuem estruturas idênticas e são constituídos de dois outros componentes: um mecanismo de auto-atenção e uma rede neural completamente conectada de uma camada.

Antes de explorar os demais componentes, vamos observar como funciona o mecanismo de auto-atenção. Considere que cada item da sequência (um *token*, uma palavra) é representado por um *embedding*. Como antes, o vetor de *embedding* pode ter sido pré-treinado. Considerando o ponto de entrada de um transformer como sendo o tal vetor de *embeddings*, são criados três vetores a partir de cada palavra ou *token*. A implementação é toda matricial, para fazer bom uso das GPUs, mas podemos abstrair para vetores, para facilitar o entendimento. Vamos então considerar que temos a frase *viva este dia*. Transformers fazem uso de *tokens* de subpalavras, para aliviar o problema das palavras que estariam fora de um vocabulário pré-treinado, conforme apresentado no Capítulo Sequência de caracteres e palavras. Mas, para simplificar, vamos assumir que cada palavra é um *token*. Temos então três *tokens* na frase, que serão representados pelos vetores x_1 – viva, x_2 – este, e x_3 – dia, que são os *embeddings* de cada palavra. A partir de cada um deles, criamos três outros vetores, q, k e v, de *query* (consulta), *keys* (chaves) e *values* (valores)[18], respectivamente (lembra do exemplo do bolo?). O vetor q se refere a um item de interesse que está sendo codificado. O vetor k se refere aos demais itens da sentença. O vetor v representa a codificação do valor dado a cada item, considerando o item de interesse. Para

[17] https://openai.com/blog/introducing-chatgpt-and-whisper-apis

[18] Para manter as letras clássicas que referenciam os vetores, vamos manter os nomes em inglês.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol2/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

o nosso exemplo, temos então, os vetores q_1, k_1 e v_1 para a palavra "viva", q_2, k_2 e v_2 para a palavra "este" e q_3, k_3 e v_3 para a palavra "dia".

Como os vetores são obtidos? Como é de praxe com redes neurais, usando matrizes de pesos aprendidas com os dados. Assim, multiplicando o vetor x_1 pela matriz de pesos associadas às *queries*, $\mathbf{W}_Q$, temos o vetor q_1. O mesmo vale para os demais itens, ou seja, para obter o vetor k_1, multiplicamos x_1 por uma outra matriz de pesos $\mathbf{W}_k$, e para obter v_1, multiplicamos x_1 por uma outra matriz de pesos $\mathbf{W}_v$. Vamos agora calcular o peso de atenção, para, dada uma palavra que está fazendo às vezes de *query*, identificarmos quais são as *keys* mais relevantes para produzir o vetor de valoração. Observe que essa intuição está inserida nas matrizes de peso aprendidas. Então, assumindo inicialmente que a query é a palavra "viva", multiplicamos seu vetor q_1 por cada uma das keys, k_1, k_2 e k_3. Observe que depois o mesmo será feito para as demais palavras. Os valores multiplicados são divididos por 8 [19]. A seguir, como antes, os valores passam por uma função de softmax, para que eles sejam transformados em probabilidades. A soma de todas as probabilidades vai ser sempre igual a um. Ou seja, ficamos com a ideia de que cada *key* é mais ou menos importante para cada palavra query, de acordo com o valor computado pela softmax. Agora aparecem os vetores de valoração. Os valores calculados pelo softmax são multiplicados por cada um dos vetores de *values*, para codificar a importância das demais palavras na valoração. Finalmente, esses valores são somados, produzindo um valor final, chamado de z_1 para a primeira palavra, que será passado adiante para a rede neural completamente conectada. Veja um esquema do processo para a primeira palavra na Figura 1.8.

1.3.2.2.3 Atenção em múltiplas versões

A linguagem é sujeita a várias complexidades que podem fazer o mecanismo de atenção não ser suficiente para representá-las. Dependendo da sentença de entrada, podemos ter variações na atenção. Isso é bem comum em problemas de correferência. Por exemplo, na frase "Eles não levaram os livros nos compartimentos porque eles eram muito pequenos", o segundo "eles" pode se referir a vários outros pronomes e substantivos na frase. Ainda, dependendo do contexto, as palavras podem ter vários significados, um conceito chamado de polissemia. Então, para capturar diferentes representações de uma palavra, bem como diferentes interações da palavra com os demais componentes, Transformers incluem um nível de paralelismo no processamento da entrada, a partir de um componente adicional chamado de *multi-heads attention*. No artigo original, os Transformers incluem oito versões paralelas do mecanismo de atenção, fazendo com que tenhamos oito vezes três matrizes de *keys*, *values*, *queries* inicializadas aleatoriamente, o que permite que tais matrizes capturem diferentes aspectos da entrada.

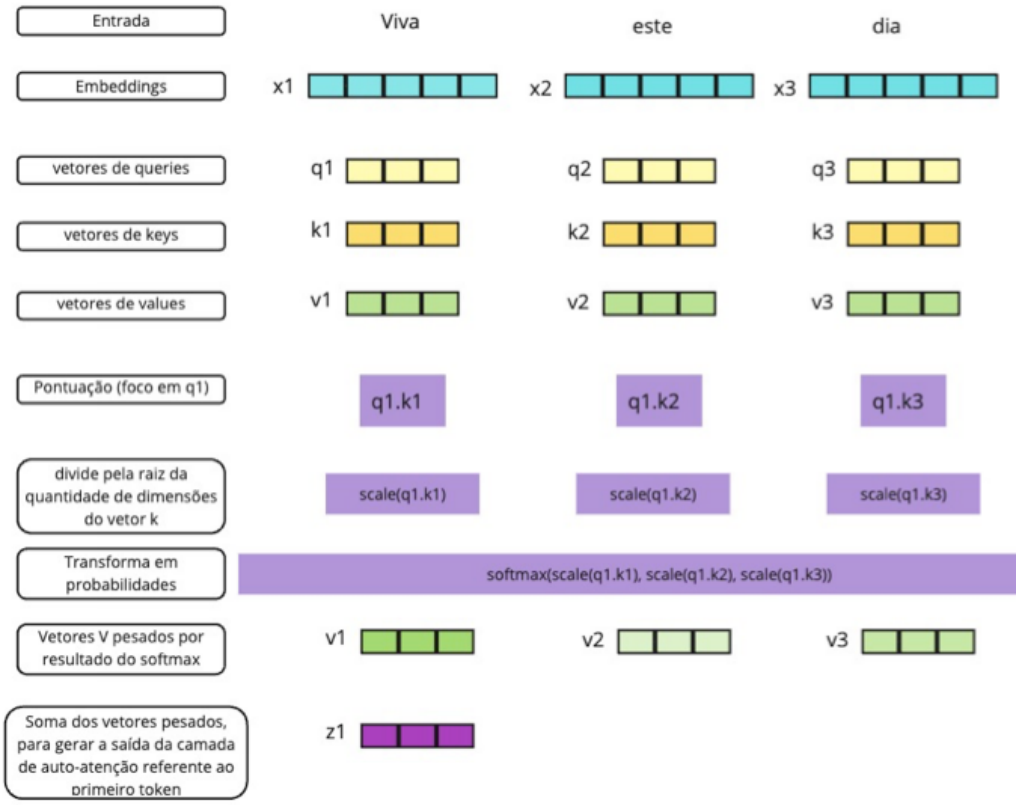
Lembre que o mecanismo de atenção produz uma matriz final Z após processar a entrada a partir das diferentes matrizes de peso Q, K, V. Anteriormente, falamos que o resultado do passo de atenção é submetido a uma rede neural completamente conectada. Para que a rede completamente conectada consiga lidar com as várias matrizes Z geradas em paralelo, elas são concatenadas e multiplicadas por uma outra matriz de pesos adicional, gerando, finalmente, uma única matriz Z que representa o resultado do mecanismo de atenção em suas várias versões. Como essa matriz de pesos adicional também é aprendida,

[19] A explicação no artigo é que 8 é a raiz quadrada da dimensão dos vetores de *keys*, que no artigo é 64. Essa divisão é uma espécie de normalização, para deixar os cálculos dos gradientes mais estáveis. Outros valores poderiam ser usados, especialmente com vetores de outras dimensões.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol2/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Figura 1.8: Exemplo do mecanismo de auto-atenção para codificar a frase “Viva este dia”.



Fonte: Adaptado de http://jalammar.github.io/illustrated-transformer/

Transformers dão a chance de alguma das versões do mecanismo de atenção paralelo ter mais ou menos relevância que algum outro, dependendo dos dados de treinamento.

Mas quantas matrizes são treinadas, não é mesmo? Pare um momento para fazer uma conta de quantos pesos um Transformer precisa treinar, considerando os componentes que apresentamos até aqui. O que isso pode fazer com o meio ambiente, se um décimo das pessoas do planeta resolvessem treinar seu próprio Transformer?

1.3.2.2.4 E as posições das palavras??

Duas motivações foram apresentadas para construir modelos de linguagem a partir de redes recorrentes: (i) permitir entradas de tamanho variável e (ii) permitir que o aprendizado tenha acesso à ordem das palavras e absorva a diferença que vem de ordens distintas, bem como a importância da ordem para tarefas sintáticas e semânticas. A recorrência é o mecanismo utilizado para atender a estas duas motivações no ELMo, por exemplo. Mas Transformers não incluem nada de recorrência. E agora?

A bem da verdade, para permitir o treinamento de forma eficiente, as implementações das redes recorrentes já não deixavam a entrada ser tão variável assim. Para que os tensores sejam formados e manipulados de forma eficiente, é comum que algumas implementações preencham frases com símbolos nulos e organizem frases que tenham o tamanho mais aproximado o possível, para que eles fiquem nos mesmos lotes e ajudem na manipulação dos tensores. De certa forma, Transformers possuem uma entrada de tamanho pré-definido. O tamanho pré-definido, em geral, é até bem menor do que gostaríamos para manipular

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol2/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

textos um pouco mais longos, por questões de desempenho. Na Seção 1.4 falaremos de como este problema tem sido abordado. Mas é possível lidar com sentenças de tamanhos distintos nos Transformers, adotando alguma das abordagens abaixo:

- Quando a sentença de entrada tem menos *tokens* que a quantidade de *tokens* de entrada esperada pelo modelo: esse é o caso mais fácil, quando a sentença é preenchida com valores nulos, um processo chamado de *padding*.
- Quando a sentença de entrada tem mais *tokens* que a quantidade de *tokens* de entrada esperada pelo modelo: duas soluções podem ser adotadas. A mais simples é truncar a entrada, removendo elementos do início ou do fim da sentença. Outra forma mais elaborada é quebrar a sentença em janelas com elementos sobressalentes entre elas e passar esses pedaços ou janelas várias vezes no modelo.

O outro problema, a ordem das palavras, exige a inclusão de um componente adicional no modelo, uma vez que a ordem é de extrema relevância para a sintaxe e a semântica, e portanto também para modelos que tentam aprender a resolver tarefas sintáticas ou semânticas. Assim, Transformers incluem um tipo especial de *embedding*, chamado de *codificador de posição* (*positional encoding*), para contemplar alguma informação sobre as posições dos *tokens* durante o aprendizado. O codificador posicional é um vetor a mais somado ao vetor de *embeddings* de entrada de cada *token*. Embora, em um primeiro pensamento, possa parecer mais direto considerar um valor simples de posição, como por exemplo, um índice, essa abordagem traria alguns problemas. O primeiro é que o valor pode ficar muito grande, dependendo de quantas palavras temos, e o modelo poderia se confundir achando que esses valores altos têm alguma importância. Mesmo se o valor fosse normalizado entre 0 e 1, diferentes tamanhos de sentenças trariam diferentes valores, o que também atrapalharia a generalização do aprendizado.

Assim, o codificador de posição define um vetor de valores contínuos do mesmo tamanho do *embedding* de entrada do *token*, para que seja possível somá-los. Para incorporar mais uma ideia de distância entre as palavras, ou de posição relativa, do que uma ideia rígida de posição, o vetor é obtido a partir de uma função que intercala entre a aplicação de seno ou cosseno. Mais precisamente, para codificar a informação de posição de um *token* que está em uma posição k na sequência de entrada, considerando cada posição i do vetor posicional, fazemos:

$$p(k, 2i) = \sin\left(\frac{k}{n^{\frac{2i}{d}}}\right) p(k, 2i+1) = \cos\left(\frac{k}{n^{\frac{2i}{d}}}\right)$$

onde d é a dimensão do vetor posicional e n é um valor pré-definido[20]. Para posições pares do vetor de saída, aplica-se o seno e para posições ímpares, aplica-se o cosseno. A Tabela 1.1 apresenta um exemplo simplificado da aplicação do codificador posicional[21].

[20] No artigo original, foi definido o valor de 10.000.

[21] Para quem quiser entender a derivação do porquê essas funções fazem sentido para representar a informação de posição, podem debruçar nesta detalhada explicação: https://timodenk.com/blog/linear-relationships-in-the-transformers-positional-encoding/ ou para uma ideia mais geral, aqui: https://kazemnejad.com/blog/transformer_architecture_positional_encoding/.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol2/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Tabela 1.1: Exemplo da computação do codificador posicional, considerando um vetor de saída de quatro dimensões apenas e n=10

Token	**índice na sentença**	**i=0**	**i=1**	**i=2**	**i=3**
viva	0	0	1	0	1
este	1	0,8415	0,5403	0,3109	0,9504
dia	2	0,9093	-0,4161	0,5911	0,8607

1.3.2.2.5 Mecanismo residual e normalização

O último subcomponente dos Transformers que precisamos falar é a inclusão de duas conexões residuais (He et al., 2016) dentro da subcamada de codificação. A conexão residual surgiu na área de visão computacional, com a motivação que redes neurais com muitas camadas podem esquecer uma informação importante de entrada após ela passar por muitos processamentos. Em geral, esse esquecimento se dá pelo problema do gradiente que vira zero, depois de muitas multiplicações de valores menores que um (Hochreiter, 1991) durante o *backpropagation*[22]. A conexão residual justamente evita parte dessas transformações multiplicativas, pulando algumas delas.

No caso dos Transformers, além de evitar que o treinamento se perca com multiplicações de valores muito pequenos, a motivação é que os *embeddings* de representação de palavras também continuem a ser aproveitados de alguma forma, trazendo uma ideia de representação local dos *tokens* para a subcamada de codificação. Ou seja, ao permitir que a informação sem ser processada pela camada de auto-atenção e que a informação sem ser processada pela camada completamente conectada sejam consideradas, é como se a rede estivesse lembrando da representação original do *token*, quando necessário. Para tanto, a saída da camada de auto-atenção é somada com a entrada original, que por sua vez é somada com a saída da camada completamente conectada, preservando, e repassando para a frente de alguma forma, a entrada original.

A arquitetura Transformer original possui em seu componente de codificação seis subcamadas de codificadores, com as camadas internas completamente conectadas tendo 512 neurônios artificiais intermediários (na camada oculta, ou escondida) e oito *heads* de atenção.

Só mais um detalhe para fecharmos o nível interno de uma camada de codificação: para ajudar no aprendizado dos gradientes, antes da camada de atenção e antes da camada completamente conectada, temos uma camada de normalização. Na verdade, tem uma pequena confusão com essa camada. A figura original dos Transformers (Figura 1.9) aponta que a normalização acontece após o processamento da camada de atenção e após o processamento da camada completamente conectada (pós-normalização). Entretanto, no código original está diferente[23]: na verdade, no código temos uma pré-normalização, que ocorre antes do cálculo dos valores de atenção. Argumenta-se que isto ajuda a lidar melhor com os gradientes (Xiong et al., 2020). Mas esta discussão, de onde inserir a camada de normalização, e onde ela apresenta mais vantagens, ainda é um ponto de investigação em aberto.

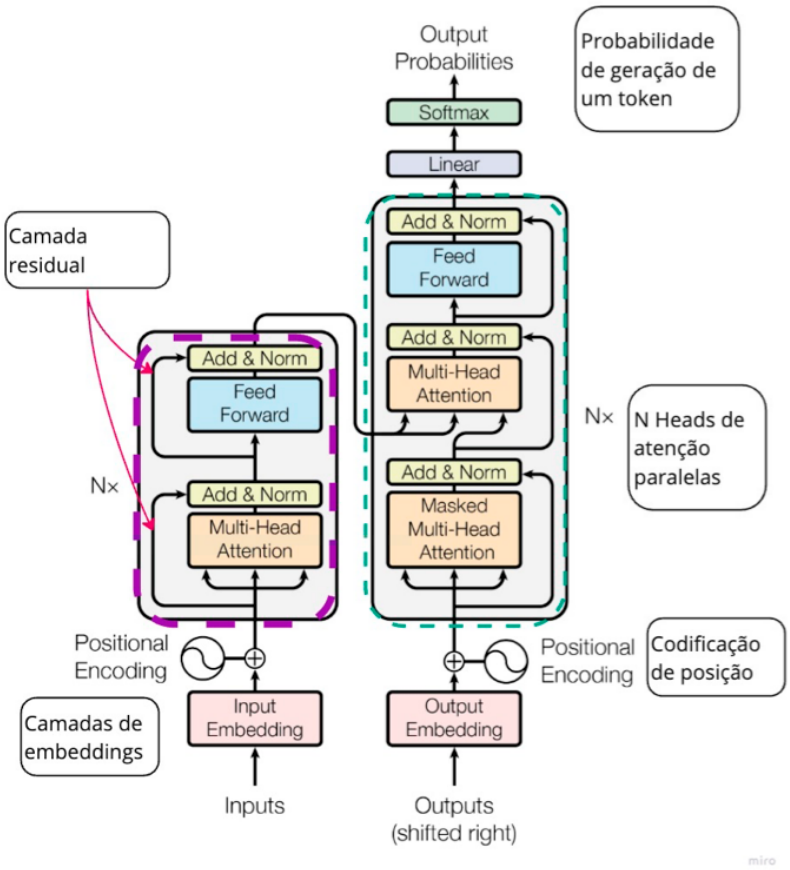
[22] O algoritmo *backpropagation* é o mais utilizado para aprender os pesos de uma rede neural, usando o método de otimização de gradiente descendentes visando minimizar o erro do modelo. Veja mais detalhes em https://www.deeplearningbook.com.br/algoritmo-backpropagation-parte-2-treinamento-de-redes-neurais/.

[23] Veja mais em https://sebastianraschka.com/blog/2023/self-attention-from-scratch.html.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol2/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Figura 1.9: Arquitetura Transformer. O pontilhado violeta representa o codificador e o pontilhado verde representa o decodificador.



Fonte: Modificado a partir de (Vaswani et al., 2017a).

1.3.2.2.6 Decodificador e comunicação entre codificador e decodificador

A arquitetura do componente Decodificador é bastante parecida com a arquitetura do Codificador. Ambas são uma pilha de sub-camadas, com cada sub-camada incluindo uma camada de auto-atenção, de normalização, um componente residual e uma camada de rede completamente conectada. Porém, aqui temos um componente a mais, que é uma camada de atenção convencional, que se comunica com o Codificador. Assim, após o processo de codificação, as matrizes de atenção K e V serão a entrada para a camada de atenção convencional do decodificador. Já a matriz Q vem mesmo da camada anterior, justamente para que essa camada ajude o modelo a decidir o que ele precisa considerar para gerar a saída.

Uma outra diferença é que a camada de auto-atenção do decodificar é chamada de mascarada, uma vez que ela não tem acesso aos *tokens* que estão em uma posição posterior a um certo *token*. Assim, no decodificador, considerando o processamento de um *token* t_i, a camada de auto-atenção do decodificador só terá acesso aos *tokens* $\{t_0, t_1, \ldots, t_{i-1}\}$ para calcular o valor da auto-atenção. Este comportamento tem a ver com o que se espera de um decodificador: que ele gere um próximo *token*, dados os *tokens* anteriores a ele, mas sem saber o futuro de antemão. Assim, o decodificador lembra muito o processo do modelo probabilístico que discutimos antes: (1) o processamento parte dos *tokens* "anteriores", que inicialmente é apenas um *token* especial de início de sentença; a entrada do decodificador

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol2/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

considera esses *tokens* anteriores e a camada de atenção convencional considera as matrizes geradas pelo codificador; (2) um *token* é gerado pelo modelo; (3) o processo se repete.

Entretanto, nos modelos probabilísticos, fica clara a existência das probabilidades, enquanto até agora só falamos em vetores. Transformers incluem uma última camada de rede neural, justamente para resolver tal discrepância. Assim, a última camada da arquitetura recebe a saída do decodificar (um vetor) e a processa com uma camada linear completamente conectada. A saída da camada linear completamente conectada é um vetor de logits[24] do tamanho do vocabulário, que representam uma pontuação associada a cada palavra do vocabulário. Finalmente, tais valores de pontuação passam por uma operação de SOFTMAX, para converter esses valores reais em valores que fiquem entre 0 e 1, representando a probabilidade de que o decodificador emita cada uma das palavras do vocabulário.

1.3.2.2.7 Arquiteturas que instanciam Transformers

Tarefas que lidam com linguagem têm sido abordadas por diferentes instanciações de Transformers: podemos considerar a arquitetura completa, podemos considerar apenas o componente codificador, ou podemos considerar apenas o componente decodificador. Ainda, é possível não usar todas as camadas existentes no modelo original, mas subconjuntos (ou até mesmo superconjuntos) delas.

1.3.2.2.8 Codificador: BERT e seus amigos

A arquitetura mais utilizada que considera apenas o componente codificador chama-se BERT[25], de *Bidirectional Encoder Representations for Transformers* (Devlin et al., 2019). O BERT foi treinado em duas versões, uma chamada *base* e outra chamada *large*. A versão base possui 12 subcamadas de codificação, que por sua vez incluem camadas completamente conectadas com 768 unidades de neurônios artificiais intermediários e 12 *heads* de atenção. A versão *large* é composta de 24 subcamadas de codificadores, com suas camadas completamente intermediárias tendo 1024 neurônios artificiais intermediários e as camadas de atenção com 16 *heads*.

Em ambas as arquiteturas, a entrada para o BERT tem uma limitação de 512 *tokens*, devido, principalmente, ao processamento quadrático do mecanismo de atenção, que considera todos os *tokens* para cada *token* em seus cálculos. O primeiro *token* é um especial chamado de [CLS], cujo uso ficará mais claro quando falarmos do processo de treinamento e inferência com modelos de linguagem. O BERT também pode receber duas sentenças (também ficará mais claro daqui a pouco), e nesse caso elas são separadas com um outro *token* especial chamado [SEP]. Para cada *token* da entrada, o BERT produz um vetor de saída de 768 ou 1024 posições, dependendo da configuração *base* ou *large*. A saída completa de um modelo BERT é um tensor de quatro dimensões, com a primeira representando a quantidade de subcamadas de codificação (12) mais a camada dos *embeddings* de entrada, totalizando 13, a segunda representando a quantidade de lotes (voltaremos nele ao falar do treinamento), a terceira a quantidade de *tokens* na entrada, e a quarta o tamanho da camada escondida. Veja um exemplo abaixo, onde a quantidade de *tokens* é 26 devido ao processo de tokenização em *subtokens* (na verdade, são 24 *tokens*, pois o primeiro e o último são os *tokens* especiais, [CLS] e [SEP]). O código do exemplo segue o *framework* HuggingFace (Wolf et al., 2020).

[24] Neste contexto, logits são basicamente logaritmos de probabilidades não normalizados.

[25] Também é uma brincadeira com um personagem do famoso desenho infantil *Sesame Street*, no ar desde 1969, e que no Brasil ganhou uma versão chamada Vila Sésamo.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol2/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

```
import torch
# carregando os módulos do framework HuggingFace
from transformers import AutoTokenizer, AutoModel

# carregando o tokenizador
tokenizer = AutoTokenizer.from_pretrained("bert-base-uncased")

# carregando o modelo BERT BASE
model = AutoModel.from_pretrained("bert-base-uncased")

# frase de Epicteto, um filósofo estoico
s = "a riqueza não consiste em ter grandes posses, mas em ter poucas
necessidades"

# frase tokenizada
input_sentence = torch.tensor(tokenizer.encode(s)).unsqueeze(0)

# saída do modelo
# output_hidden_states=True faz com que tenhamos acesso a todas as
camadas na posição 2 da variável de saída
out = model(input_sentence,output_hidden_states=True)

print("Numero de camadas: ", len(out[2]))
print("Numero de lotes: ", len(out[2][0]))
print("Numero de tokens: ", len(out[2][0][0]))
print("Numero de neurônios artificiais: ", len(out[2][0][0][0]))

Numero de camadas:  13
Numero de lotes:  1
Numero de tokens:  26
Numero de neurônios artificiais:  768
```

Por curiosidade, os *tokens* gerados são como seguem abaixo. Observe que os espaços não são *tokens*, mas estão aqui apenas para conseguirmos separar um *token* de outro.

```
a ri ##que ##za [UNK] consist ##e em ter grande ##s posse ##s , mas
em ter po ##uca ##s nec ##ess ##idad ##es
```

onde as duas tralhas indicam que o *token* é, originalmente, parte de uma palavra e [UNK] é um *token* para indicar que não foi possível encontrar aquele componente no vocabulário. No caso acima, provavelmente é devido à presença de uma palavra com til, o que não existe na língua inglesa.

Um modelo de linguagem segue um vocabulário e o vocabulário do BERT original segue a língua inglesa. Embora o processo de tokenização em subpalavras consiga identificar algumas palavras, toda a definição dos pesos durante o processo de aprendizado será feito com base em um vocabulário de um idioma distinto. E quando falamos de uma língua, não é apenas o vocabulário que é relevante, mas as regras sintáticas, seus significados, e até mesmo características culturais e da sociedade.

Então, temos duas opções: ou considerar um modelo que tenha se deparado com um vocabulário de vários idiomas, ou treinar um modelo separado para uma língua. Para o primeiro caso, temos, por exemplo, o modelo **BERT multilingual** (chamado de mBERT), treinado com textos da Wikipedia de 104 idiomas. Com o mesmo teste feito antes, teríamos agora 18 *tokens*, ao invés de 26, mostrando que, ao menos, mais palavras são reconhecidas. Em particular, usando o mBERT, o processo de tokenização devolve

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol2/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

```
a riqueza não consiste em ter grandes posse ##s , mas em ter poucas
necessidade ##s
```

Para o segundo caso, temos alguns modelos treinados especificamente para o português (outras línguas também, mas nosso interesse aqui é na nossa linda língua materna). O modelo **BERTPT** (Feijó; Moreira, 2020) foi treinado com um vocabulário de tamanho 30.000, assim como o modelo BERT original, porém mantendo a configuração original de maiúsculas e minúsculas e sinais diacríticos (os acentos). Foram usados 4,8GB de textos, considerando textos em português do Brasil e europeu, tanto mais formais, como Wikipedia-PT[26] e EuroParl[27], como textos mais informais, como Open Subtitles[28]. No total, foram considerados 992 milhões de *tokens*. A arquitetura utilizada foi a base. O modelo BERTPT apresentou resultados melhores em bases de dados compostas por textos mais informais.

O modelo **BERTimbau**[29] (Souza et al., 2020) também partiu da arquitetura do BERT, mas treinou duas versões, uma a partir da arquitetura *base* e outra a partir da arquitetura *large*. Assim como no BERTPT, são mantidas as letras maiúsculas e minúsculas e acentos e o tamanho do vocabulário também é de 30.000 *tokens*. O conjunto de textos usados para treinar os modelos foi o brWaC (Wagner Filho et al., 2018), que é composto de textos em português do Brasil, contendo 2,68 bilhões de *tokens* e 3,53 milhões de documentos, e após uma fase de pré-processamento ficou com 17,5GB de textos. Outra diferença em relação ao BERTPT é que a tokenização utilizou o algoritmo BPE, enquanto o BERTPT e o BERT original seguem o algoritmo WordPiece, ambos mencionados no Capítulo Sequência de caracteres e palavras. Apenas para fins de comparação com o exemplo anterior, abaixo temos o resultado da tokenização usando o tokenizador do BERTimbau, que fica com um *token* a menos que o mBERT. O resultado foi gerado com a versão disponibilizada no hub de modelos HuggingFace[30] (Wolf et al., 2020). Na maioria dos resultados apontados no artigo, o BERTimbau supera o mBERT.

```
a riqueza não consiste em ter grandes posse ##s , mas em ter poucas
necessidades
```

O último modelo para português baseado no BERT que citaremos aqui é o **Albertina** (Rodrigues et al., 2023), treinado em duas variantes, português europeu (Albertina PT-PT) e português do Brasil (Albertina PT-BR). A versão PT-BR também foi treinada com o brWaC. Já a versão PT-PT foi treinada com um subconjunto de textos em português extraídos da versão de Janeiro de 2023 do *corpus* Oscar (Abadji et al., 2022b) e de outros três *corpora* constituídos de documentos do parlamento europeu e português. No total, foram utilizados oito milhões de documentos contendo 2,2 bilhões de *tokens*.

Uma diferença crucial do Albertina para os modelos anteriores é que a arquitetura base não é a do BERT, mas sim uma versão estendida com duas novas técnicas, chamada DeBERTa (do inglês, *Decoding enhanced BERT with disentangled attention*) (He et al., 2021). A primeira modificação diz respeito ao mecanismo de atenção. Lembre que nos Transformers, um *token* é representado pela soma do seu vetor inicial de *embeddings* e do seu vetor de codificação de posição. No DeBERTa, e consequentemente no Albertina,

[26] https://dumps.wikimedia.org/backup-index.html
[27] https://www.europarl.europa.eu/
[28] http://opus.nlpl.eu/OpenSubtitles-v2016.php
[29] Uma brincadeira com o berimbau.
[30] https://huggingface.co/neuralmind/bert-base-portuguese-cased

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol2/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

temos dois vetores que são processados separadamente (daí o *disentangled*, ou desemaranhado em português), onde o vetor de codificação de posição representa a posição relativa de um *token* i em relação a um *token* j. O valor de atenção cruzada de dois *tokens* é calculado como $A_{i,j} = \{\mathbf{H_i}, \mathbf{P_{i|j}}\} \times \{\mathbf{H_j}, \mathbf{P_{j|i}}\}^\top$, onde $\mathbf{H_i}$ representa o vetor de *embeddings* do *token* i e $\mathbf{P_{i|j}}$ representa a posição relativa do *token* i em relação ao *token* j. A outra modificação tem a ver com a tarefa de treinamento genérica das arquiteturas Transformers e voltaremos nela na seção seguinte.

Para não perder o costume, veja abaixo o resultado do processo de tokenização usando a versão PT-BR disponibilizada no HuggingFace[31]. Observe que a representação é diferente das anteriores: cada *token* que é o início de uma palavra recebe um '_' como prefixo. O tokenizador também tem a diferença de tratar espaços somo se eles fossem parte do *token*.

```
_a _ rique za _não _consist e _em _ter _grande s _posses , _mas _em
_ter _pou cas _necess idades
```

Também temos alguns modelos BERT treinados para tweets em português[32]. Certamente, existem vários outros modelos treinados para o português que não apareceram aqui. Observem que esta não é para ser mesmo uma lista exaustiva.

Existem diversas outras arquiteturas que estendem, melhoram, modificam, ou treinam com mais dados ou com outros parâmetros o componente codificador dos Transformers. Exemplos incluem ROBERTa (Liu et al., 2019a), que incluiu modificações no treinamento e usa o algoritmo de tokenização BPE ao invés do WordPiece; DistillBERT (Sanh et al., 2019), que se vale de um processo de destilação de conhecimento para aproximar os pesos do modelo original e obter um modelo menor que o BERT; AlBERT (Lan et al., 2020), que introduz três mecanismos – fatorização das matrizes de *embeddings*, compartilhamento de pesos e uma nova forma de treinamento – para obter um modelo mais eficiente que o BERT; ELECTRA (Clark et al., 2020), que também muda a forma de treinamento do BERT para obter *embeddings* melhores com um processo mais eficiente; dentre muitos outros.

1.3.2.2.9 Decodificador: GPT e seus vizinhos

A saída do codificador em um modelo Transformer é uma representação vetorial, que para resolver uma tarefa final ainda precisa passar por algum outro processo. Mas essa representação pode ser bem robusta, uma vez que ela olha ambos os lados direito e esquerdo de um *token* ao construir sua representação vetorial. Já o componente decodificador tem uma característica autorregressiva, e a sua saída pode ser mesmo um texto, mas cada *token* só pode olhar para aqueles que vieram antes dele na sequência. Como falamos antes, esse componente é o que remonta de fato ao propósito original de um modelo de linguagem computacional: gerar o próximo *token*, dados os *tokens* anteriores a ele, ou seja, gerar um texto. A família de arquiteturas GPT (Radford; Narasimhan, 2018) (do inglês, *Generative Pre-trained Transformer*, ou Transformer Gerativo Pré-treinado) usa blocos decodificadores da arquitetura Transformer para funcionar como um modelo autorregressivo de geração de texto. Claramente, sem o componente codificador, não temos o mecanismo de atenção tradicional como parte das entradas do decodificador, como acontece na arquitetura Transformer.

[31] https://huggingface.co/PORTULAN/albertina-ptbr

[32] https://huggingface.co/melll-uff/bertweetbr e https://huggingface.co/pablocosta/bertabaporu-base-uncased, por exemplo

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol2/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Em sua primeira versão, o GPT era bem similar ao componente decodificador do Transformer, sendo composto de 12 subcamadas de decodificadores com 12 *heads* de auto-atenção mascaradas de dimensão 768, e camadas escondidas completamente conectadas de 3072 dimensões. A tokenização também é baseada em subpalavras, mas segue o algoritmo BPE ao invés do WordPiece, como o BERT. A segunda versão, chamada criativamente de GPT-2 (Radford et al., 2019a) veio em quatro versões de tamanhos variados: GPT-2 *small*, com 12 subcamadas de decodificadores e dimensão dos *embeddings* de 768, a versão GPT-2 *medium*, com 24 subcamadas de decodificadores e dimensão dos *embeddings* de 1024, a versão GPT-2 *large*, com 36 subcamadas de decodificadores e dimensão dos *embeddings* de 1280, e a versão GPT-2 *extra large*, com 48 subcamadas de decodificadores e dimensão dos *embeddings* de 1600. A camada de normalização passou a estar na entrada de cada subcamada e adicionou-se uma outra camada de normalização após o último bloco de auto-atenção. O código a seguir, que também usa o HugginFace, apresenta exemplos de geração de texto usando o GPT2.

```
# modelo multilingual
model_name = "sberbank-ai/mGPT"
tokenizer = GPT2Tokenizer.from_pretrained(model_name)
model = GPT2LMHeadModel.from_pretrained(model_name)
model.cuda()
model.eval()

# geração de texto default
def cond_gen(tokenizer, model, prefix):
    # encode context the generation is conditioned on
    input_ids = tokenizer.encode(prefix, return_tensors='pt').cuda()

    # generate text until the output length (which includes the context
    # length reaches 50)
    greedy_output = model.generate(input_ids, max_length=50)
    return list(map(tokenizer.decode, greedy_output))[0]

# imprimir a saída do modelo
def print_output(output):
    print("Output:\n" + 100 * '-')
    print(output)

# usando beam search
def cond_gen_beam(tokenizer, model, prefix, ngram=1):
    input_ids = tokenizer.encode(prefix, return_tensors='tf')
    beam_output = model.generate(
                        input_ids,
                        max_length=50,
                        num_beams=5,
                        no_repeat_ngram_size=ngram,
                        early_stopping=True
                )
    return beam_output[0]

# usando top-k sampling
def cond_gen_sample(tokenizer, model, prefix):
    input_ids = tokenizer.encode(prefix, return_tensors='tf')
    sample_output = model.generate(
        input_ids,
        do_sample=True,
        max_length=50,
```

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol2/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

```
        top_k=0
    )
    return sample_output[0]

# manipulando o parâmetro de temperatura
def cond_gen_sample_temp(tokenizer, model, prefix, temp=0.5):
    input_ids = tokenizer.encode(prefix, return_tensors='tf')
    sample_output = model.generate(
        input_ids,
        do_sample=True,
        max_length=50,
        top_k=0,
        temperature=temp
    )
    return sample_output[0]

prefix = 'Eu gosto de'
output = cond_gen(tokenizer_pt, model_pt, prefix)
print_output(output)

Output:
----------------------------------------------------------------------
Eu gosto de fazer o que gosto, mas não sou muito de fazer o que não
gosto.

output = cond_gen_beam(tokenizer, model, prefix)
print_output(output)
Output:
----------------------------------------------------------------------
Eu gosto de pensar que a vida é muito curta, mas eu não posso viver
sem ela.
Não importa o quanto você se preocupe com as coisas boas e ruins do
mundo; ninguém pode ser fel


output = cond_gen_sample_temp(tokenizer, model, prefix, 0.7)
print_output(output)

Output:
----------------------------------------------------------------------
Eu gosto de ler, mas não sou leitora compulsiva.
Eu gosto de livros que me dão vontade de ter saudades, e quando eu
vejo uma resenha que me encanta, eu leio uma história
```

As mudanças arquiteturais da versão original para a versão 2 não foram profundas, exceto pelo tamanho e consequente quantidade de parâmetros treinados. Porém, no artigo do GPT-2 começou-se a vislumbrar um modelo mais geral, que pudesse executar várias tarefas (aprendizado de múltiplas tarefas, ou agnóstico de tarefas (Collobert; Weston, 2008)), mesmo sem ser treinado novamente para cada uma delas (configuração (Romera-Paredes; Torr, 2015)), e usando apenas a geração de texto como uma abstração de qualquer outra tarefa mais específica. O argumento era que o pré-treinamento em um conjunto grande e diverso de textos seria suficiente para que o modelo pudesse lidar com problemas com os quais não havia sido explicitamente treinado. Por exemplo, o artigo exemplifica que algumas sentenças nos textos usados para o pré-treinamento já eram exemplos de tradução de uma língua para a outra, o que faria o modelo aprender a traduzir naturalmente.

Este foi o principal motivador para o desenvolvimento da versão 3 da arquitetura GPT,

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol2/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

chamada de GPT-3 (Brown et al., 2020). A ideia seria que durante o pré-treinamento (que vamos entender na próxima seção) o modelo consegue desenvolver indiretamente habilidades de geração de texto que poderiam ser usadas para resolver diversas tarefas, como tradução e resposta a perguntas, entre outras. Tais habilidades poderiam ser resgatadas em tempo de execução de acordo com a tarefa pedida, um processo chamado de *aprendizado em um contexto* (Dong et al., 2023). Três configurações foram discutidas no artigo, que já eram objeto de estudo de outros trabalhos voltados para o aprendizado a partir de poucos exemplos. A primeira configuração se chama *zero-shot* e explora cenários em que o modelo recebe como contexto uma descrição da tarefa (que até poderia ser opcional, dependendo da tarefa) e um ***prompt***[33], e espera-se que o modelo responda a partir destes dois componentes apenas, sem nenhum tipo de ajuste nos seus pesos. Ou seja, nenhum exemplo é dado para o modelo[34]. Por exemplo, abaixo temos uma descrição e *prompt* para tradução automática[35].

```
Traduza de português para francês: # descrição da tarefa
penso, logo, existo: # prompt
---
je pense, donc je suis # saída do modelo
```

A segunda configuração chama-se *one-shot* e, neste caso, um exemplo completo é fornecido para o modelo como parte do contexto.

```
Traduza de português para francês: # descrição da tarefa
penso, logo, existo: je pense, donc je suis # exemplo fornecido
conhece-te a ti mesmo: # prompt
---
Connais-toi toi-même. # saída do modelo
```

Em um outro exemplo (bem hipotético, o GPT-3 não foi testado para tais habilidades em português), perceba que a descrição da tarefa e o exemplo podem estar juntos:

```
concordância incorreta em português: há bastante alunos reprovados
conjugação correta em português: há bastantes alunos reprovados # exm
conjugação incorreta em português:  vejo muito alunos no corredor
conjugação correta em português: # prompt
---
vejo muitos alunos no corredor # saída do modelo
```

A terceira configuração chama-se *few-shot* e a diferença para as anteriores é apenas que mais de um exemplo são fornecidos para o modelo ter como base. Intuitivamente, tais exemplos, a descrição da tarefa e o *prompt* direcionarão o modelo para certos pesos que

[33] Um *prompt* é um texto em linguagem natural (em oposição à linguagem de programação) que especifica uma instrução do que deve ser feito. Um *prompt* pode ser formulado como uma pergunta, uma observação, um questionamento, ou ainda representar uma tarefa específica, por exemplo, classificar um texto. *Prompts* podem ter um formato livre, mas modelos de linguagem são bastante sensíveis ao conteúdo textual dele. Afinal, modelos como GPT estão completando um texto (neste caso, o *prompt*), condicionados ao que tem neste texto.

[34] Note que chamamos de exemplo, mas para quem está acostumado com aprendizado de máquina, este não seria um exemplo para ajustar os parâmetros do modelo. Seria apenas um exemplo mesmo.

[35] Estamos usando exemplos em português, mas na época, as habilidades do modelo em lidar com várias línguas não foi demonstrada no artigo. Mas mencionou-se que no treinamento do GPT-2, os textos foram filtrados para que ficassem apenas aqueles em inglês. Para o GPT-3, esse requerimento foi relaxado e 7% dos textos estavam em outras línguas.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol2/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

ativarão as distribuições de probabilidade de geração de textos para o ponto correto. É mesmo como se ele estivesse sempre completando textos que tenham alguma coerência com o que foi visto antes. Vamos lembrar que modelos gerativos são desenhados para predizer o próximo *token* a partir dos *tokens* anteriores e é isso que está sendo feito aqui. No nosso último exemplo, *tokens* anteriores são desde "concordância incorreta ..." até "... correta em português:", removendo os comentários (o que vem depois da tralha).

Em termos de arquitetura, foi usada a mesma do GPT-2, mas com uma alteração no mecanismo de atenção para fatorizar de forma esparsa as matrizes de atenção (Child et al., 2019) e reduzir a complexidade do mecanismo de atenção de $\mathcal{O}n^2$ para $\mathcal{O}n\sqrt{n}$ [36]. Foram treinados oito modelos de diferentes tamanhos, variando de 12 a 96 camadas e dimensão de *embeddings* de 768 a 12.288. O maior deles, com 175 bilhões de parâmetros foi o que obteve os melhores resultados, em geral, e que ganhou o privilégio de ser o GPT-3. O artigo também apontou as limitações correntes do GPT-3 e potenciais aplicações perigosas, que valem a pena a leitura.

A partir de 2022, a empresa que desenvolveu o GPT, chamada OpenAI[37] começou gradativamente a disponibilizar versões melhoradas do GPT-3, introduzindo novas formas de treinamento que se valem de cada vez mais textos e modelos cada vez maiores. Por exemplo, a empresa disponibilizou novas versões e interfaces de programação de aplicações (APIs) para os modelos que estenderam o GPT-3, chamados de "text-davinci-003" e "code-davinci-002", que passaram a ser genericamente chamados de GPT-3.5. O GPT-3.5 é a base para o famoso ChatGPT[38], veja mais no Capítulo 7, um agente de conversação genérico e que foi disponibilizado para o mundo testar no finzinho de 2022. Na época da escrita deste capítulo, a versão mais nova do GPT era o GPT-4.

Um outro modelo interessante e que também se vale da geração autorregressiva dos decodificadores é o XLNet (Yang et al., 2019). Ele melhora a modelagem autorregressiva, que tem como desvantagem não ter um olhar bidirecional, ao considerar todas as possíveis permutações de uma sequência durante o aprendizado. Ele também usa um esquema de codificação posicional relativo, proposto na arquitetura Transformer-XL (Dai et al., 2019), mas com algumas reparametrizações com foco em remoção de ambiguidade. Assim como ocorre com as arquiteturas codificadoras, muitas outras abordagens têm sido propostas a todo momento. Algumas delas voltarão ao nosso radar quando falarmos das tendências correntes da área na Seção 1.4.

1.3.2.2.10 Instâncias de um Transformer inteiro: T5 e seus aliados

Os Transformers também podem ser utilizados ou instanciados com seus dois componentes, o codificador e o decodificador. A desvantagem é que estes modelos costumam ser maiores e precisam de mais dados para serem treinados. A vantagem é que o modelo tanto terá os benefícios da codificação de atenção bidirecional, como de ter pronto um modelo de linguagem para geração de texto. Mas um dos grandes benefícios é permitir que os problemas sejam tratados sempre do ponto de vista da geração de textos, permitindo que um mesmo modelo possa resolver várias tarefas. Ou seja, se queremos fazer tradução, vamos receber um texto inicial e completá-lo com a tradução em outra língua. Se queremos responder a consultas, a pergunta é a entrada e a resposta é a saída. E por aí vai. E o uso de instruções ajudam o modelo a completar o texto da maneira apropriada para resolver

[36] A notação de Big O descreve complexidade de um código em termos algébricos (Cormen et al., 2001; Knuth, 1997).
[37] https://openai.com/blog/chatgpt
[38] https://chat.openai.com/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol2/

a tarefa específica.

Uma arquitetura que desempenha esse papel de ser um consumidor e produtor de texto (*text-to-text*) é o T5 (Raffel et al., 2020a). A arquitetura do T5 segue os Transformers, mas o *embedding* de posição é relativo, seguindo o deslocamento de posição entre as matrizes de chave e de consulta que estão sendo comparadas. Outros pontos de mudança são a remoção do valor de viés da camada de normalização e a inclusão da camada de normalização após o caminho residual. O modelo foi avaliado com diversos conjuntos de dados para tarefas variadas, incluindo análise de sentimentos, similaridade de sentenças, desambiguação de palavras, resolução de correferências, entre outras. Assim como já falamos no GPT, para permitir que diversas tarefas pudessem ser abordadas pelo mesmo modelo, os autores usaram uma instrução limitada, que funcionava quase como um hiperparâmetro (no artigo, chama-se **prefixo específico de tarefa**).

Um outro modelo também bastante utilizado e discutido na literatura é o BART (*Bidirectional and Auto-Regressive Transformers*) (Lewis et al., 2020a) [39]. Seu objetivo é mapear um texto com ruídos (corrompido) para o texto original, uma tarefa que lembra reconstrução de imagens. Duas arquiteturas são apresentadas no artigo: uma com seis camadas de codificadores e decodificadores cada e outra com 12.

A exemplo do que já vimos antes, o T5 também foi treinado para lidar com português, dando origem ao modelo denominado de PTT5 (Carmo et al., 2020). O modelo foi avaliado em uma base de similaridade de sentenças (ASSIN 2) e de reconhecimento de entidades nomeadas (HAREM). Observe que o modelo é limitado às instruções dessas duas tarefas. Além do modelo monolingual, temos a versão multilingual do T5, o mT5 (Xue et al., 2021), treinada por pesquisadores da Google com o *corpus* mC4, uma versão do *corpus* Common Crawl para 101 idiomas, incluindo o português. E temos ainda a versão multilingual do BART (Liu et al., 2020), treinado inicialmente com um subconjunto do Common Crawl para 25 idiomas (mBART) que não incluiu o português, mas sua extensão para 50 idiomas (mBART-50 (Tang et al., 2020)), sim.

1.3.2.2.11 *Embeddings* de Sentenças com Transformers

Podemos recuperar *embeddings* contextualizados tanto para *tokens*, como para combinações de *tokens*, o que inclui sentenças e textos. Quando uma combinação de *tokens* resulta em uma palavra, estamos falando de um *embedding* de palavra. Quando recuperamos os *embeddings* de uma frase, estamos falando de um *embedding* de sentença.

Modelos baseados em Tranformers permitem recuperar *embeddings* de sentenças basicamente de duas formas: construindo os *embeddings* a partir das médias dos *embeddings* de cada *token* na sentença, usando uma ou mais camadas da arquitetura (em geral, usamos as quatro últimas camadas), ou usando os *embeddings* de saída do primeiro *token*, o [CLS]. Perceba que a agregação por média não é o mesmo processo discutido no Capítulo Representação vetorial e semântica distribucional para obter *embeddings* de sentenças a partir de representações estáticas. Com Transformers, além dos *tokens* influenciarem uns aos outros com o mecanismo de atenção, também temos o codificador de posição, que influenciará na saída final. Uma outra possibilidade é treinar modelos que consigam devolver *embeddings* de sentenças de entrada, o que é feito costumeiramente tendo como base a tarefa de similaridade semântica e arquiteturas siamesas, como no modelo *sentence-transformers*[40] (Reimers; Gurevych, 2019, 2020), ou treinamentos contrastivos, como a abordagem SIM-CSE (Gao et al., 2021).

[39] Afinal, além de Sesame Street, também gostamos dos Simpsons.
[40] https://www.sbert.net/

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol2/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Mas como todos esses modelos podem ser usados, treinados, aprendidos, refinados? É o que vamos discutir a seguir.

1.3.3 Treinamento e Ajustes em Modelos de Linguagem Neurais

Na seção anterior, vimos como funcionam os modelos de linguagem neurais, além de abordarmos diversos modelos em português, ressaltando detalhes de suas arquiteturas e *corpora* usados para treinar os pesos iniciais de cada um dos modelos. Esse passo inicial de treinamento, que é parte do modelo disponibilizado em arcabouços como TensorFlow e HuggingFace, é chamado de **pré-treinamento**. O pré-treinamento (ou *training from scratch*, ou treinamento do zero) refere-se a uma técnica de treinamento de redes neurais profundas (*deep neural network*) que no caso de modelos de linguagem usa uma quantidade expressiva de textos sem nenhum rótulo ou anotação, com o intuito de gerar um modelo de propósito geral capaz de "entender" linguagem.

O fato dos textos não terem rótulos ou anotações é o que permite usar uma quantidade enorme de textos, pois sabemos que anotar exemplos é uma tarefa custosa e que requer um tempo precioso de especialistas. Ainda assim, o pré-treinamento de um modelo de linguagem é uma tarefa desafiadora em muitos aspectos, incluindo a necessidade do uso intensivo de uma quantidade significativa de recursos computacionais por longos períodos de tempo. Além do alto custo envolvido no pré-treinamento desses modelos, ainda é preciso levar em consideração os impactos ambientais resultantes do alto consumo de energia.

De modo geral, o pré-treinamento de um modelo de linguagem engloba os seguintes passos:

1. **Escolher *corpora*** (Capítulo Conjunto de dados, *dataset* e *corpus*): a escolha dos *corpora* ideais é um passo importante no treinamento de modelos de linguagem. Mas quais seriam as características desses *corpora* ideais? Essa é uma pergunta difícil de responder, uma vez que a escolha dos *corpora* vai depender muito dos objetivos finais desse modelo, além é claro, da disponibilidade de tais *corpora*. A escolha dos *corpora* pode ser baseada no domínio a ser explorado. Por exemplo, um modelo pré-treinado com textos do Twitter, pode ser mais equipado para resolver tarefas que envolvam textos com uma linguagem mais informal. A escolha também pode basear-se na língua. O objetivo pode ser treinar um modelo monolingual ou até mesmo multilingual. Um outra possibilidade, muito adotada com os modelos de linguagem grandes, é usar *corpora* bem variados, formados por *corpus* de diferentes domínios e línguas. Veja mais sobre as particularidades e requerimentos da criação e escolha de *corpora* no Capítulo Conjunto de dados, *dataset* e *corpus*.
2. **Limpar e pré-processar os textos**: embora modelos de linguagem neurais não precisem de muitos passos de limpeza e pré-processamento, como costumava ser feito para o treinamento de modelos de aprendizado de máquina anteriores, ainda é necessário executar uma normalização dos textos, ainda que simples. Tal passo inclui remover caracteres especiais, remover URLs e remover textos que tenham apenas poucos caracteres. Como muitos dos textos são coletados da Web, também costuma-se remover etiquetas HTML, para que estas não sejam confundidas com palavras ou caracteres importantes.
3. **Treinar o tokenizador** (Capítulo Sequência de caracteres e palavras): como vimos no Capítulo Sequência de caracteres e palavras, tokenização é o processo de dividir o texto em unidades menores, chamadas de *tokens*. Esse é um passo importante no treinamento de modelos de linguagem, podendo impactar o desempenho final de tais

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol2/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

modelos. Neste passo, podemos optar em usar um tokenizador pré-treinado, como por exemplo o tokenizador do GPT-3, ou podemos treinar um tokenizador do zero, assim como fazemos com os modelos de linguagem. O treinamento do tokenizador também requer a escolha de *corpora*. Neste caso, podemos usar os mesmos *corpora* escolhidos para treinar os pesos do modelo de linguagem. Além da escolha dos *corpora*, é necessário definir o tipo de tokenizador a ser treinado. Escolhas populares são o *Byte-pair encoding* (BPE) (Provilkov et al., 2020), usados por modelos como o GPT-3, e o WordPiece (Schuster; Nakajima, 2012), usado por modelos como o BERT. Ambos dividem as palavras em sub-palavras, para acomodar melhor palavras que não tenham aparecido durante o treinamento do tokenizador. Aliás, treinamento aqui é justamente definir como as palavras serão "quebradas" (ou não) para definir os *tokens* do modelo. Veja mais sobre o processo de tokenização no Capítulo Sequência de caracteres e palavras.

4. **Definir a arquitetura do modelo**: a escolha da arquitetura adotada para treinar um modelo de linguagem depende de muitos fatores, entre eles a disponibilidade de recursos de alto-desempenho para treinamento do modelo. A arquitetura pode ser uma rede neural recorrente ou uma rede neural baseada em Transformers. A arquitetura envolve quais serão os componentes em termos de camadas, *heads* e funções de ativação. Como podemos ter uma quantidade combinatória de tipos de componentes, costuma-se usar uma arquitetura pré-definida, como BERT ou GPT. Mas nada impede de uma pessoa definir a sua arquitetura do zero.
5. **Definir a função objetivo ou tarefa intermediária**: a tarefa intermediária é responsável por guiar o aprendizado do modelo. Algumas das tarefas mais utilizadas em modelos atuais serão exploradas na Seção 1.3.3.1.
6. **Definir os hiperparâmetros**: nas seções anteriores, falamos muito em "matriz de pesos" e "pesos da camada de atenção"; esses pesos são considerados parâmetros do modelo e são aprendidos com auxílio dos dados de treinamento. Já os hiperparâmetros são parâmetros que ajudam a controlar o processo de treinamento. Eles podem influenciar na qualidade final do modelo, como também na velocidade do treinamento. Esses são alguns dos hiperparâmetros mais comuns usados para o treinamento e ajuste de modelos de linguagem:

    - Taxa de aprendizagem (*learning rate*): a taxa de aprendizagem está relacionada ao algoritmo de otimização usado para atualizar os pesos do modelo a partir dos dados de treinamento. A grosso modo, a medida que os dados de treinamento circulam pela rede, os pesos do modelo são atualizados até que se alcancem pesos ideais. A taxa de aprendizagem controla o tamanho dessas atualizações e, consequentemente, afeta diretamente a convergência do modelo e o tempo de convergência.
    - *batch size*: número de amostras dos dados de treinamento, ou seja, número de sequênciasde texto, que são processadas ao mesmo tempo antes de cada atualização dos pesos do modelo. O valor ideal do *batch size* vai depender da arquitetura e tarefa alvo. É importante ressaltar que quanto maior o *batch size*, maior o consumo de memória, o que pode tornar o treinamento proibitivo em muitos casos.
    - Número de épocas (*number of epochs*): é o número total de vezes que todos os exemplos de treinamento passarão pelo modelo.
    - Taxa de regularização (*dropout rate*): é usada para controlar o problema de

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol2/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

sobreajuste, ou seja, evitar que o modelo se ajuste perfeitamente ao conjunto de treinamento perdendo sua capacidade de generalização na presença de novos dados.

7. **Avaliação**: após o treinamento, é importante avaliar a qualidade e coerência dos textos gerados pelo modelo. Um modelo também pode ser avaliado em relação ao seu desempenho na realização de uma tarefa final de PLN, como as tarefas de sumarização e análise de sentimentos. Veremos mais detalhes na Seção 1.3.4.

O pré-treinamento considera um objetivo genérico, de geração ou preenchimento de texto, que não requer nenhuma anotação por parte de especialistas. Entretanto, este modelo genérico pode ser ainda melhorado tendo em vista uma tarefa final. Assim, modelos pré-treinados podem ser ajustados de acordo com um domínio ou uma tarefa específica, o que chamamos de **continuado ou ajuste fino (*fine-tuning*)**, como veremos na Seção 1.3.3.2.

Resumindo, o treinamento de um modelo envolve a seleção dos *corpora*, o pré-processamento e limpeza desses dados, a seleção de uma arquitetura e tarefa intermediária, o treinamento em si e a avaliação do modelo pré-treinado. Em seguida, visitaremos algumas das tarefas intermediárias mais populares.

1.3.3.1 Tarefa Intermediária para o Pré-treinamento

Durante o pré-treinamento de um modelo de linguagem, uma ou mais funções objetivo ou tarefas intermediárias são utilizadas para guiar o aprendizado do modelo a gerar texto, ou, de forma mais genérica, a predizer partes do texto que estejam faltando. O intuito é que o modelo passe a ter uma compreensão estatística da(s) língua(s) em que foi treinado. Vários objetivos foram propostos na literatura, alguns a nível de *token* e outros a nível da sentença. Todos eles têm em comum o intuito de se basearem em uma tarefa de treinamento auto-supervisionada, ou seja, em que as saídas dos exemplos podem ser geradas de forma automática. Discutiremos alguns deles a seguir, começando pelas duas tarefas mais amplamente adotadas na literatura:

1. **Modelagem de linguagem mascarada (em inglês, *Masked Language Modeling* ou MLM)** (Devlin et al., 2019): esta tarefa é inspirada no teste Cloze (Taylor, 1953) [41] e foi proposta para treinar modelos bidirecionais, como o BERT. Neste caso, os textos de entrada são alterados para que em cada uma das sequências, uma porcentagem dos *tokens* seja substituída pelo *token* especial [MASK]. O objetivo é estimar os *tokens* mascarados levando em consideração o contexto dos demais *tokens* da sequência. Por exemplo, suponha a sentença mascarada do Exemplo 1.5, em que a original é atribuída a Sêneca:

   Exemplo 1.5:

   > Apressa-te a viver [MASK] e pensa que cada [MASK] é, por si [MASK], uma vida.

[41] O teste Cloze, segundo o dicionário Cambridge, consiste em um texto com algumas palavras removidas, e o leitor deve preencher tais palavras faltantes com algo que faça sentido. Ele foi proposto para medir o nível de leiturabilidade de um indivíduo.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol2/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

O objetivo do modelo seria encontrar as palavras mais adequadas para entrar no lugar de [MASK][42]. Perceba que o modelo poderia encontrar palavras adequadas diferentes das originais, mas que ainda seriam plausíveis. Por exemplo, a primeira máscara poderia ser substituída por "muito", embora no texto original (ao menos na versão traduzida para o português, a palavra seja "bem"). Por isso, avaliar o resultado de modelos de linguagem com tarefas de predição de texto é tão complexo.

2. **Modelagem de linguagem causal ou autorregressiva (em inglês, *Casual Language Modeling* ou CLM)**: Esta é a tarefa que mais se assemelha à tarefa de modelagem de linguagem como definimos no início deste capítulo, ou seja, o objetivo é completar o próximo *token* em uma sequência considerando apenas os *tokens* anteriores. Diferente da tarefa anterior, em que a sequência é vista como um todo, com apenas as posições com máscara faltando, aqui o modelo só pode atender aos *tokens* da esquerda, diferentemente dos modelos bidirecionais como o BERT.

As duas tarefas acima são as mais comumente empregadas como tarefas intermediárias na literatura. Entretanto, outras também já foram exploradas:

1. *Replaced Token Detection* (RTD) (Clark et al., 2020): quando falamos de MLM, vimos que a entrada do modelo é corrompida pela substituição de *tokens* originais da sentença, pelo *token* especial [MASK]. No caso do RTD, um gerador, que pode ser um modelo de linguagem menor, é utilizado para gerar *tokens* ambíguos que serão usados no lugar do *token* [MASK]. Esses *tokens* ambíguos, embora incorretos, são próximos do significado semântico do *token* original. Agora, ao invés de ter que prever o *token* mascarado, como ocorre quando usamos a MLM, o objetivo é identificar se um *token* é o *token* original da sentença de entrada ou se ele é um *token* gerado pelo gerador. Um exemplo seria: dada a sentença original "A professora ensinou o novo conteúdo", uma sentença após o RTD poderia ser "A professora aprendeu o novo conteúdo";
2. *Shuffled Word Detection* (Shuffle) (Yamaguchi et al., 2021): nesta tarefa, uma porcentagem dos *tokens* de entrada são aleatoriamente embaralhados antes de serem processados pelo modelo. O objetivo do modelo é identificar dentre os *tokens* da sequência de entrada, aqueles que foram inicialmente embaralhados. Considerando a sentença "O gato sentou no tapete da sala", uma sentença embaralhada seria "O gato no tapete da sala sentou";
3. *Token Order Permutations*: é a tarefa utilizada para treinar o modelo XLNet (Yang et al., 2019). Como nos modelos de linguagem autorregressivos, o objetivo é prever um *token* com base no contexto dos *tokens* anteriores, só que agora, a probabilidade de um *token* é condicionada a todas as permutações de *tokens* em uma sequência. Assim, o modelo consegue aprender o contexto de forma bidirecional, mas sem se restringir à ordem original da sequência, como nos modelos baseados no BERT. Na teoria, são geradas todas as sentenças possíveis a partir da permutação dos *tokens* da sentença original. Na prática, apenas uma amostra dessas sentenças permutadas são usadas durante o treinamento. Exemplos de sentenças seriam "Eu amo chocolate / amo eu chocolate / amo chocolate eu / chocolate eu amo / chocolate amo eu etc.";
4. *Next Sentence Prediction* (NSP): é uma função objetivo que foi usada para treinar o modelo BERT em conjunto com a função MLM. O objetivo da NSP é aprender a

[42] Em tradução livre, a sentença original é "Apressa-te a viver bem e pensa que cada dia é, por si só, uma vida".

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol2/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

relação entre duas sentenças, ou seja, se elas são sentenças contíguas ou não. Exemplos positivos são criados através da extração de sentenças consecutivas presentes nos *corpora* usados para treinar o modelo. Já os exemplos negativos são criados através do pareamento de duas sentenças oriundas de diferentes documentos dos *corpora*. Alguns estudos (Joshi et al., 2020; Liu et al., 2021) mostraram que NSP não funciona bem, ou é desnecessária para algumas tarefas. Por essa razão, modelos como o RoBERTa (Liu et al., 2021), removeram NSP do seu pré-treinamento;

5. *Sentence-Order Prediction* (SOP): essa tarefa tenta estimar se duas sentenças consecutivas estão na ordem correta ou não, ou seja, se elas tiveram ou não sua ordem invertida (Lan et al., 2020). Ao contrário da tarefa NSP, que cria exemplos negativos através da concatenação de sentenças extraídas de documentos diferentes, na SOP os exemplos negativos são criados usando duas sentenças consecutivas extraídas do mesmo documento, só que agora elas terão suas ordens invertidas. Os exemplos positivos são criados usando a mesma técnica adotada por NSP. Essa pequena alteração na construção dos exemplos negativos força o modelo a fazer uma distinção mais refinada com relação a ordem e coerência das sentenças.
6. *Translation language modeling* (TLM): foi proposto em (Conneau; Lample, 2019) e utilizado em conjunto com as funções objetivo MLM e CLM para treinar o modelo XLM. A TLM é uma extensão da MLM, uma vez que também usa o *token* especial [MASK] para mascarar *tokens* da sequência original. Só que agora, ao invés de usar sequências na mesma língua, o modelo XLM concatena duas sentenças de línguas diferentes, como por exemplo uma sentença em inglês e outra em português. Depois, *tokens* das duas sequências concatenadas são aleatoriamente substituídos pelo *token* [MASK]. Para prever um *token* mascarado na sentença em português, o modelo pode atender (mecanismos de atenção, Seção 1.3.2.2 tanto a outros *tokens* da sentença em português quanto a *tokens* da sequência em inglês.

A escolha de funções objetivo não é o único desafio para o pré-treinamento de modelos de linguagem. Outro fator relevante e com grande impacto na capacidade e qualidade final do modelo é a escolha dos *corpora*. Modelos de linguagem devem ser treinados com uma grande quantidade de dados de alta qualidade. Mesmo que não seja necessário anotar os dados, montar essas grandes coleções de dados deve ser um tarefa cuidadosa, ainda que exaustiva e demorada. O ideal é garantir que esses *corpora* sejam o mais diversos possível e sem enviesamentos, polarização e textos maliciosos, o que requer um grande esforço de filtragem e pré-processamento dos dados. Hoje em dia, vários esforços são feitos no sentido de minimizar os efeitos do uso de *corpora* contendo textos maliciosos. Por exemplo, técnicas como treinamento adversarial (Kianpour; Wen, 2020) são utilizadas para expor os modelos a textos maliciosos com o intuito de ensinar esses modelos a reconhecer tais tipos de textos. Outra técnica que tem se tornado frequente, é o uso de humanos para revisar e moderar o texto gerado por modelos de língua. Assim, esse tipo de informação pode ser utilizada para melhorar o modelo de forma iterativa.

Apesar de os modelos de linguagem serem treinados usando coleções vastas e diversas de textos, eles podem se tornar obsoletos, uma vez que essas coleções são estáticas. Com o tempo, o modelo pode não ser capaz de gerar e reconhecer textos sobre eventos atuais. Por exemplo, um modelo treinado com textos anteriores a Setembro de 2022 pode não ser capaz de reconhecer que o atual monarca da Inglaterra é o Rei Charles III, uma vez que sua mãe, a Rainha Elizabeth II, faleceu em Setembro de 2022. Além disso, dada a variedade de domínios existentes e a forma dinâmica como novas tendências e culturas

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol2/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

emergem ao longo dos anos, é muito difícil garantir que um modelo de linguagem será capaz de entender e resolver de forma precisa as mais diversas tarefas do PLN. Na seção seguinte discutiremos formas de usar novas coleções de dados para ajustar um modelo a tarefas e domínios específicos.

1.3.3.2 Ajustes em Modelos de Linguagem Neurais

Uma das formas encontradas para atualizar modelos de linguagem é o que chamamos de treinamento continuado (em inglês, *continued pre-training*) (Gururangan et al., 2020; Jin et al., 2022; Ke et al., 2023).

No treinamento continuado, o modelo é treinado por mais algumas iterações ou épocas usando uma coleção de textos diferente dos *corpora* utilizados no pré-treinamento, mas mantendo a mesma tarefa intermediária. Ou seja, o treinamento continuado, assim como o pré-treinamento, é um processo de treinamento auto-supervisionado. Tradicionalmente, o treinamento continuado pode ser dividido em dois tipos: o treinamento continuado com foco na adaptação da tarefa (*Task Adaptative Pre-Training*, TAPT) e o treinamento continuado com foco na adaptação do domínio (*Domain Adaptative Pre-Training*, DAPT). No caso do TAPT, o treinamento continuado ocorre com a utilização de uma coleção de textos não-rotulados relacionados a uma tarefa específica, por exemplo a tarefa de análise de sentimentos. A coleção não precisa ser grande, mas precisa representar bem diferentes aspectos da tarefa alvo. Já no DAPT, o foco não é a tarefa, mas sim o domínio. Neste caso, o modelo é treinado por mais algum tempo utilizando uma coleção de textos que tratam de algum domínio específico. Por exemplo, um domínio pode ser a biomedicina ou até mesmo artigos científicos sobre inteligência artificial. Mais recentemente, foi proposto o treinamento continuado baseado em instruções (*Prompt-based Continued Pre-training*, PCP), que seria uma combinação do treinamento continuado tradicional (TAPT) com o ajuste de instruções (*instruction tuning*) (Shi; Lipani, 2023) (veja mais na Seção 1.4. Assim como no TAPT e DAPT, no PCP a função objetivo original, ou tarefa intermediária, é utilizada durante o treinamento continuado. Mas neste caso, teremos dois tipos de entrada: os textos não-rotulados relacionados a tarefa alvo, como no TAPT; e, os *prompts* ou instruções também relacionadas a tarefa alvo.

O treinamento continuado é um dos métodos utilizados para adaptar um modelo pré-treinado a alguma tarefa (TAPT) ou domínio específico (DAPT). Outro método que também permite a adaptação de modelos é o método do *ajuste fino* (*fine-tuning*) (Howard; Ruder, 2018a). Enquanto o treinamento continuado não requer textos rotulados e usa a mesma tarefa intermediária adotada durante o pré-treinamento do modelo, no ajuste fino usamos textos rotulados e uma função objetivo específica da tarefa alvo, por exemplo, a tarefa de classificação. Os dois métodos resultam no ajuste dos pesos do modelos. Entretanto, por ser mais específico e focar totalmente na tarefa alvo, através de dados rotulados e o uso de uma função objetivo específica, o ajuste fino costuma requerer menos dados para promover o ajuste dos pesos do modelo pré-treinado, além de resultar em um modelo altamente ajustado ao contexto da tarefa final. Com isso, podemos dizer que o ajuste fino resulta em um tempo de treinamento menor do que o treinamento continuado, o que também vai impactar no custo final de geração do modelo.

Se pensarmos no treinamento dos modelos de linguagem como um processo que pode ocorrer em duas etapas, o pré-treinamento do modelo seria a primeira etapa e o treinamento continuado e/ou ajuste fino, seria a segunda etapa. Nessa primeira etapa, o modelo é treinado depois de serem definidas a arquitetura e a tarefa intermediária, além da sele-

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol2/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

ção e pré-processamento de grandes coleções de textos a serem utilizados no aprendizado. Já na segunda etapa, os pesos do modelo são ajustados para um domínio e/ou tarefa específica. Apesar da possibilidade de um ajuste dos modelos pré-treinados, essa etapa de ajuste não é obrigatória. Tanto o modelo pré-treinado, como o modelo ajustado podem ser utilizados em diversas tarefas de PLN.

Um exemplo de tarefa é a análise de sentimentos. Aqui, vamos considerar a tarefa de análise de sentimentos como um problema de classificação binária com dois sentimentos possíveis: positivo e negativo. Dada uma coleção de treinamento composta por sentenças rotuladas, o objetivo final é treinar um classificador capaz de classificar novas sentenças em um desses dois sentimentos, positivo ou negativo. Um exemplo de sentença rotulada seria: "Maria gostou muito do computador", sentimento positivo. Neste caso, podemos usar um modelo de linguagem pré-treinado para gerar representações vetoriais dessas sentenças, os *embeddings*. Dados os *embeddings* e os rótulos, podemos usar qualquer algoritmo de classificação, como máquina de vetores de suporte (SVM, *Support-Vector Machine*) (Cortes; Vapnik, 1995) ou regressão logística (Tolles; Meurer, 2016), para treinar um classificador capaz de categorizar novas sentenças não rotuladas em um dos dois sentimentos. A extração de *features* (*feature extraction*) ou *embeddings*, pode ser feita usando tanto um modelo pré-treinado, como também um modelo ajustado.

Ainda considerando a tarefa de análise de sentimentos, poderíamos ajustar um modelo de linguagem de diversas formas. No caso do TAPT, poderíamos usar uma coleção de dados de análise de sentimentos sem a necessidade dos rótulos. No caso do DAPT, precisaríamos de uma coleção de dados associada ao domínio em questão, mas também sem a necessidade dos rótulos. Por exemplo, se a tarefa é analisar o sentimento dos consumidores em relação a marcas de carros, poderíamos então usar no ajuste uma coleção de dados contendo opiniões de consumidores sobre marcas de carros. Note que aqui, o foco não é a tarefa de análise de sentimento, mas sim o domínio. Outro tipo de ajuste possível seria o ajuste fino. Neste caso, usaríamos um coleção de dados de análise de sentimentos para opiniões de consumidores sobre marcas de carros. No ajuste fino, a coleação precisa ser rotulada, uma vez que o modelo de linguagem será ajustado usando a tarefa final. Como estamos tratando a análise de sentimentos como uma tarefa de classificação binária, a tarefa final usada nos ajustes é a tarefa de classificação. Para realizar o ajuste, podemos adicionar uma camada de classificação à arquitetura do modelo e então ajustar os pesos do modelo usando os textos da coleção rotulada.

1.3.4 Avaliação de Modelos de Linguagem Neurais

Com o crescente número de modelos de linguagem disponíveis, é bem desafiante decidir qual a melhor maneira de avaliar a qualidade ou capacidade desses modelos[43]. Tradicionalmente, modelos de linguagem são avaliados por métricas como perplexidade (*perplexity*), entropia cruzada (*cross-entropy*) e bits-por-caracter (*bits-per-character*, BPC). Esse tipo de avaliação é comumente chamada de avaliação intrínseca, com o modelo sendo avaliado através do seu desempenho na tarefa intermediária. No caso dos modelos de linguagem, a tarefa intermediária é prever o próximo *token* de uma sequência. Uma outra forma de avaliar modelos de linguagem é aplicá-los diretamente na resolução de uma tarefa final e então avaliar o quanto a qualidade da solução melhorou. Por exemplo, se estamos na dúvida entre adotar o modelo A ou o modelo B para resolver uma tarefa de classificação ou uma tarefa de reconhecimento de voz, podemos aplicar os dois modelos, A e B, e então

[43]Veja mais sobre Avaliação de sistemas de PLN no Capítulo Avaliação de tecnologias de linguagem.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol2/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

medir qual das duas soluções produziu os melhores resultados. Esse tipo de avaliação é chamada de avaliação extrínseca. Apesar da avaliação extrínseca ser considerada a melhor maneira de avaliar a capacidade de um modelo de linguagem em resolver uma tarefa específica, ele é um processo de alto custo e que envolve longos tempos de execução.

Como foi dito no parágrafo anterior, a avaliação intrínseca não depende de nenhuma tarefa final específica, ela considera apenas a qualidade do modelo na geração do próximo *token* da sequência. Para entender os conceitos de perplexidade, entropia cruzada e bits-por-caracter, precisamos primeiro falar de entropia. A ideia de entropia foi proposta em 1951 por C. E. Shannon (Shannon, 1951) para medir a quantidade média de informação que é transmitida por cada letra de um texto. Shannon também definiu entropia da seguinte forma: "Se a linguagem for traduzida em dígitos binários (0 e 1) da forma mais eficiente, a entropia é o número médio de dígitos binários necessários por letra da linguagem original". No contexto de linguagem, entropia é a quantidade de informação contida em um caractere em uma sequência de texto infinita. A entropia (H) é definida como:

$$H = -\sum p(i)log(p(i)))$$

onde i é o próximo *token* a ser gerado pelo modelo e $p(i)$ é a probabilidade do *token* i ser escolhido como o próximo *token* da sequência, dados os *tokens* anteriores. Podemos dizer que, se um modelo captura bem a estrutura de uma língua, consequentemente a entropia do modelo deve ser baixa.

Nós vimos na Seção 1.2, Equação 1.3, que a tarefa de completar uma sequência de palavras com uma próxima palavra é definida por uma distribuição de probabilidade condicional das palavras que poderiam completar a sequência, dadas as palavras que vieram antes na sequência. Assim sendo, o modelo de língua tem como objetivo aprender uma distribuição Q, a partir de uma amostra de texto, que seja próxima da distribuição P, que é a distribuição empírica da língua. Para medir o quão próximas são essas duas distribuições, muitas vezes usamos a entropia cruzada, definida como:

$$H(P,Q) = -\sum_i P(i)logQ(i) = H(P) + D_{KL}(P \parallel Q)$$

onde $H(P)$ é a entropia da distribuição empírica P e $D_{KL}(P \parallel Q)$ é a divergência de Kullback-Leibler de Q para P, ou seja, a entropia relativa de P com relação a Q. A divergência de Kullback-Leibler (Joyce, 2011) é uma medida estatística que, neste caso, mede o quão diferente é a distribuição de probabilidade Q da distribuição de probabilidade de referência P.

O conceito de perplexidade está totalmente relacionado ao conceito de entropia e entropia cruzada. A perplexidade é entendida como uma medida de incerteza e é definida como a exponencial da entropia cruzada:

$$PPL(P,Q) = 2^H(P,Q)$$

Teoricamente, quanto menor a perplexidade, melhor o desempenho do modelo em prever o próximo *token* da sequência.

Também seguindo a linha da entropia, temos a métrica bits-por-caracter que mede o número médio de bits necessários para representar um caracter. Ou seja, seguindo a definição de entropia dada por Shannon, podemos dizer que a entropia é o número médio de BPC.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol2/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Até aqui, temos falado muito em "número médio de bits" e entropia a nível de caractere. Mas quando revisitamos as seções anteriores, notamos que dependendo do tokenizador adotado pelo modelo de língua, o texto de entrada pode ser quebrado em palavras, subpalavras e até caracteres. Sendo assim, sempre que vamos comparar modelos de linguagem diferentes, é importante atentar para o tipo de tokenização usada pelo modelo e então, ajustar as métricas de acordo. Outro detalhe importante que precisa ser observado, é o tamanho máximo de contexto permitido por um modelo de linguagem, uma vez que, em geral, modelos de linguagem com comprimento de contexto mais longo costumam ter um valor de entropia cruzada menor quando comparado com modelos com comprimento de contexto menores.

Outra forma de avaliar e comparar diferentes modelos de linguagem, é através do uso de *benchmarks*, conforme discutido de forma extensiva no Capítulo Conjunto de dados, *dataset* e *corpus*. *Benchmarks* para modelos de linguagem são conjuntos de dados referentes a várias tarefas linguísticas que ajudam a avaliar a capacidade dos modelos no entendimento e geração de texto. O uso de *benchmarks* permite uma padronização com relação aos dados e métricas, o que é fundamental para que experimentos possam ser replicados e comparados em diferentes estudos. Além disso, o uso de *benchmarks* permite o monitoramento da evolução dos modelos com o passar do tempo. Entre os *benchmarks* mais populares está o GLUE (*General Language Understanding Evaluation*)[44] (Wang et al., 2018) e o SuperGLUE[45] (Wang et al., 2019a), ambos focados na língua inglesa. Para a língua portuguesa, temos o Poeta (*Portuguese Evaluation Tasks*), que inclui 14 bases de dados de tarefas finais, incluindo similaridade textual, análise de sentimentos, perguntas e respostas, entre outros. Apesar dos benefícios trazidos pelo uso de *benchmarks*, é preciso ficar atento a possíveis limitações, como a existência de vieses nas coleções de dados e a falta de representatividade e diversidade nos textos, o que pode impactar a generalização dos resultados dos modelos, além da escassez de coleções multilíngues. Dentre as métricas disponíveis no GLUE e SuperGLUE estão a acurácia, o F1-score e o Coeficiente de Correlação de Matthews (em inglês, *Matthews Correlation Coefficient* ou MCC).

No contexto de tarefas de classificação, a acurácia é a fração de previsões que o modelo acertou e pode ser definida como:

$$\text{Acurácia} = \frac{vp + vn}{vp + vn + fp + fn}$$

onde vp (verdadeiro positivo) é o número de amostras positivas que foram classificadas corretamente, vn (verdadeiro negativo) é o número de amostras negativas que foram classificadas corretamente, fp (falso positivo) é o número de amostras negativas que foram classificadas como positivas e fn (falso negativo) é o número de amostras positivas que foram classificadas como negativas.

O F1-score é a média harmônica da precisão e revocação, podendo ser definida como:

$$\text{F1-score} = 2\frac{\text{precisão} \times \text{revocação}}{\text{precisão} + \text{revocação}}$$

A precisão e revocação podem ser definidas como:

$$\text{precisão} = \frac{vp}{vp + fp}$$

$$\text{revocação} = \frac{vp}{vp + fn}$$

[44] https://gluebenchmark.com/
[45] https://super.gluebenchmark.com/

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol2/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Embora as fórmula estejam focada em tarefa binárias, considerando exemplos positivos e negativos, é possível generalizá-la para qualquer quantidade de classes. De forma similar, é sempre possível reduzir uma tarefa com qualquer quantidade de classes para uma avaliação binária.

O Coeficiente de Correlação de Matthews (*Matthews Correlation Coefficient*, MCC) é outra métrica que baseia-se nos números de verdadeiro positivo (vp), verdadeiro negativo (vn), falso positivo (fp) e falso negativo (fn). MCC foi proposta com a classificação binária em mente (Matthews, 1975) e pode ser definida como:

$$MCC = \frac{vp \times vn - fp \times fn}{\sqrt{(vp + fp)(vp + fn)(vn + fp)(vn + fn)}}$$

Métricas de avaliação automática para geração de texto não é algo novo. Uma das métricas mais utilizadas é o ROUGE (*Recall-Oriented Understudy for Gisting Evaluation*), proposta em 2004 por Chin-Yew Lin (Lin, 2004) com o intuito de avaliar resumos gerados por técnicas de sumarização de texto. O ROUGE calcula o número de sobreposições de unidades, como n-gramas, entre uma referência e o texto candidato a ser avaliado. Também no contexto da avaliação automática para geração de texto, foram propostas as métricas BERTScore (Zhang et al., 2020) e BARTScore (Yuan et al., 2021). Ambas usam modelos de linguagem pré-treinados para tentar avaliar a qualidade do texto gerado. BERTScore usa os *embeddings* gerados por modelos como o BERT para calcular a similaridade (similaridade de cosseno) entre os *tokens* da sequência gerada e os *tokens* da sequência de referência. Então, métricas como precisão, revocação e *F-measure* são calculadas. Já o BARTScore usa um modelo pré-treinado baseado na arquitetura *encoder-decoder* (BART) para avaliar o texto gerado em diferentes perspectivas, incluindo coerência e factualidade. A ideia é tratar a avaliação da geração de texto como um problema de geração de texto, ou seja, usar o próprio modelo *encoder-decoder* para converter o texto de entrada no texto de saída e vice-versa.

1.4 Tendências

1.4.1 A Era dos *Large Language Models* (LLMs)

O termo ***Large Language Models*** (LLMs), que podemos traduzir como Modelos de Linguagem Grandes, Modelos de Linguagem Enormes, ou Modelos de Linguagem de Larga-Escala, tem se popularizado para referenciar qualquer modelo de linguagem neural. Entretanto, neste livro consideramos que LLMs se diferenciam dos demais modelos pré-treinados devido a:

- a sua **quantidade enorme de parâmetros**. Embora não exista um limite inferior universalmente aceito, tipicamente, modelos que são chamados de LLMs na literatura possuem mais de um bilhão de parâmetros, mas podendo alcançar centenas de bilhões (Zhao et al., 2023).
- o seu enquadramento na categoria de métodos de **IA Gerativa** (ou generativa). Tais modelos têm como função primária a geração de conteúdo, que no caso dos LLMs traduz-se em geração de texto.
- as suas **habilidades emergentes**, que não costumam ser observadas em modelos menores (Wei et al., 2022a). Argumenta-se que tais habilidades não poderiam ser observadas ao examinar sistemas menores, um fenômeno similar à transição de fase

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol2/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

observada em sistemas físicos. A habilidade emergente mais comumente observada é a possibilidade de **utilizar LLMs sem nenhum treinamento adicional** que vá atualizar seus parâmetros por meio de otimização de gradientes. Ao invés deste ajuste específico, eles podem aproveitar seu pré-treinamento e serem utilizados a partir de instruções em linguagem natural – os *prompts* – e/ou demonstrações da tarefa a partir de um ou mais exemplos. Esta habilidade é conhecida como **aprendizado em contexto** (em inglês, *in-context learning* ou *few-shot prompt* (Brown et al., 2020), manifestando-se de forma curiosa com os modelos abordando tarefas para as quais não foram explicitamente treinados. Neste caso, o modelo recebe ou não uma instrução e pares de exemplos de entrada e saída, com o teste no final, conforme discutimos com o GPT. A tarefa do modelo será predizer os próximos *tokens* após a última entrada de teste. A Tabela 1.2 traz um exemplo, considerando a tarefa de análise de sentimentos, mas assumindo um modelo pré-treinado que não foi ajustado para ela.

Uma outra habilidade emergente é a estratégia de **cadeia de pensamento** (CoT, do inglês, *chain-of-thought*) (Wei et al., 2022b). Discutivelmente, tal estratégia exibiria habilidades de "raciocínio" dos LLMs, embora esta seja uma terminologia polêmica. A estratégia CoT permite que os modelos retornem passos intermediários da resposta final, em tarefas que requerem múltiplos passos de raciocínio para serem resolvidas, usando instruções do tipo "Explique passo a passo ..." ou similares.

Tabela 1.2: Exemplo de teste da habilidade emergente de aprendizado de contexto em LLMs. Os exemplos de 1 a 3 constituem em pares de entrada e saída para o modelo, enquanto a última linha, chamada de teste, apresenta apenas a entrada para o modelo. O modelo completa tal entrada com a saída em verde. Exemplos extraídos da base de dados TweetSentBR (Brum; Nunes, 2018)

Exemplo	Entrada	Saída
1	"Vitor é gracinha demais #MasterChefBR"	Positivo
2	"O #MasterChefBR tá na mesma vibe do #BBB: odeio todos."	Negativo
Teste	"Que tensooooooooo cozinhar com plateia! #MasterChefBR"	Negativo

Alguns desses modelos são: BLOOM (Scao et al., 2022), Chinchilla (Hoffmann et al., 2022), Galactica (Taylor et al., 2022), Gopher (Rae et al., 2021), GPT-3 (Brown et al., 2020), LaMDA (Thoppilan et al., 2022), LLaMA (Touvron et al., 2023a), Sabiá (Pires et al., 2023) (sim, a língua portuguesa também tem um LLM para chamar de seu), PaLM (Chowdhery et al., 2022), entre vários outros[46].

Considerando a língua portuguesa, o modelo Sabiá foi treinado com um subconjunto em português do *corpora* ClueWeb 2022 (Overwijk et al., 2022). O ClueWeb foi tokenizado com o tokenizador do GPT-2 e o processo resultou em 7,8 bilhões de *tokens*. O Sabiá usou uma estratégia de treinamento continuado a partir dos modelos LLaMA – nas versões com 7 bilhões e 65 bilhões de parâmetros – e do modelo GPT-J (Wang; Komatsuzaki, 2021), que contém 6 bilhões de parâmetros.

[46]Veja um catálogo em https://amatriain.net/blog/transformer-models-an-introduction-and-catalog-2d1e9039f376/.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol2/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Observe que a maioria desses modelos são a base de algum agente de conversação que surgiu no fim de 2022 e início de 2023: o GPT é o modelo utilizado pelo famoso agente de conversação ChatGPT[47] (GPT-3.5 e GPT-4), o LaMDA é o modelo utilizado pelo BARD[48], LLaMA é o modelo do Vicuna[49] e o Sabiá é o modelo utilizado pela MariTalk[50]. Vamos falar mais um pouco destes agentes no Capítulo 7[51].

Algumas habilidades emergentes têm sido observadas popularmente nos agentes de conversação mencionados. Por exemplo, a Figura 1.10 exibe a saída de um programa em Python "implementado pelo" ChatGPT. Observe, entretanto que, embora o modelo de linguagem não tenha sido treinado explicitamente para escrever programas em Python, ele pode ter se deparado com situações como essa em seu pré-treinamento, uma vez que foi treinado com textos da Internet. Ainda, os demais passos do processo de desenvolvimento do ChatGPT envolvem o ajuste de instruções e o alinhamento com feedback humano por meio do aprendizado por reforço (Ouyang et al., 2022; Stiennon et al., 2020a; Ziegler et al., 2019). Assim, pode ser que durante estes passos o modelo tenha sido ajustado para lidar com este tipo de instrução. Mas ele não tem como ter sido ajustado para todas as tarefas possíveis com as quais ele tem se deparado.

A Figura 1.11 exibe um exemplo da outra habilidade emergente que mencionamos, a estratégia CoT para resolver um problema matemático simples.

Sempre é bom ressaltar que os modelos podem apresentar comportamentos diferentes de acordo com a entrada apresentada a eles. E ainda que usar o mesmo *prompt* mais de uma vez não é garantia de retorno da mesma resposta, dada a natureza probabilística dos modelos gerativos. Por exemplo, considere a interação com o BARD representada na Figura 1.12, onde o objetivo era traduzir uma frase famosa do latim para o português. Ambas foram tentativas frustradas, em que ele nem se deu ao trabalho de responder em português. Mas observe o que acontece com uma instrução diferente, exibida na Figura 1.13. Embora a segunda instrução seja mais informativa, não necessariamente é este o motivo da tentativa ter sido bem sucedida. Este tipo de estratégia é chamado de *hard prompt tuning* ou **engenharia/desenho de *prompts*** (Liang et al., 2022; Schick; Schütze, 2021) e consiste em modificar as entradas para tentar obter saídas distintas.

Prompts também são usados para mapear exemplos de tarefas distintas para uma entrada em linguagem natural, na tentativa de se obter uma resposta também em linguagem natural por parte do modelo. Por exemplo, suponha o exemplo a seguir, extraído do *dataset* de reconhecimento de emoções apresentado em (Cortiz et al., 2021):

- "o que eu acho incrível nesse filme é que o Harry Potter é a própria referência à mágica" [52]

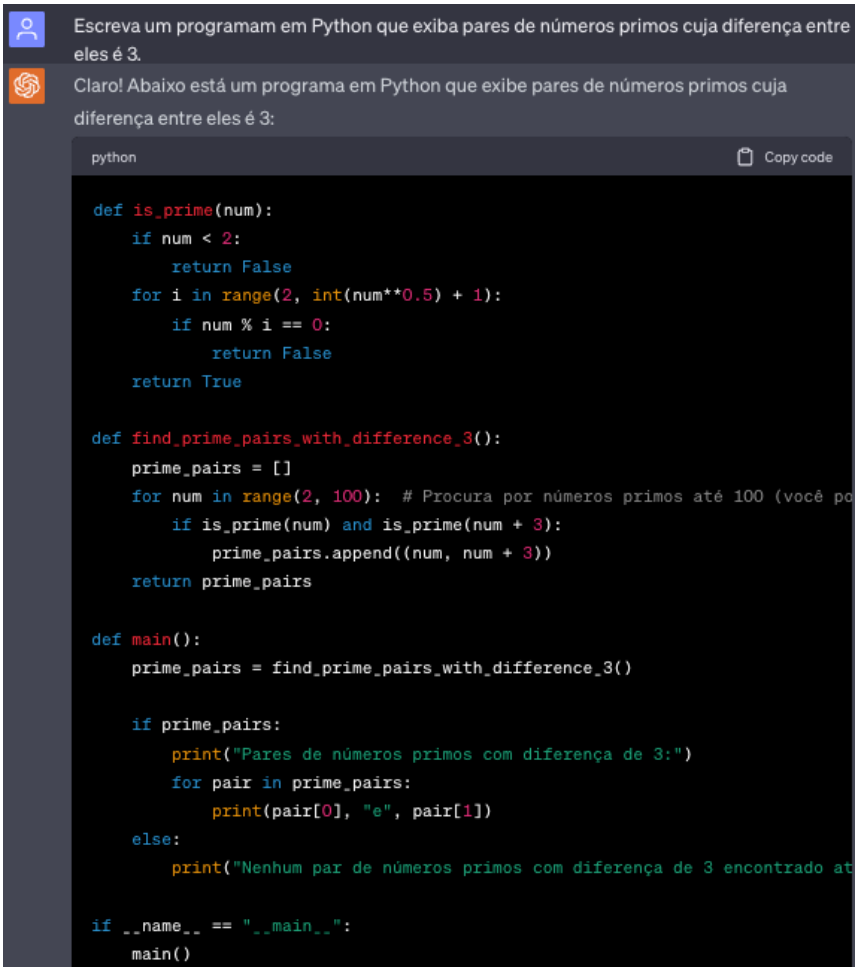
Para que exemplos como esse possam ser classificados por um modelo de linguagem autorregressivo, podemos embuti-lo na seguinte instrução:

[47] https://openai.com/blog/chatgpt
[48] https://bard.google.com/
[49] https://lmsys.org/blog/2023-03-30-vicuna/
[50] https://chat.maritaca.ai/
[51] Neste vídeo https://www.youtube.com/watch?v=prJrQ8XL-AY&t=9s o leitor encontra mais detalhes sobre a adaptação de modelos de linguagem para o português.
[52] https://github.com/diogocortiz/PortugueseEmotionRecognitionWeakSupervision/blob/main/README.md

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol2/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Figura 1.10: Saída de um programa em Python escrito pelo ChatGPT. Ele ainda explica ao final da saída o que são as funções e alerta que o programa foi configurado para funcionar apenas até o número 100, e que o usuário poderia fazer eventuais ajustes.



- "o que eu acho incrível nesse filme é que o Harry Potter é a própria referência a mágica." A emoção expressa nesta mensagem é | tristeza | raiva | admiração | confusão | curiosidade

onde o texto em azul veio do *dataset* (o exemplo) e o texto em vermelho é a instrução.

Mas esta é apenas uma entre muitas formas possíveis de se escrever o *prompt* para a tarefa de reconhecimento de emoções. Ademais, cada tarefa distinta pode ter quantidades e formatos distintos de entrada. Por exemplo, uma tarefa de inferência textual precisa incluir dois componentes, a premissa e a hipótese, ambas em azul no texto abaixo, extraídas da base de dados ASSIN-2 (Real et al., 2020):

- "Suponha a frase:" "Uma mulher está dirigindo um carro e está conversando animadamente com o carinha que está sentado ao lado dela." Podemos inferir que A mulher e o carinha estão viajando de carro. Sim ou Não?

Existem arcabouços que podem nos ajudar na criação de *prompts*. Um dos mais com-

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol2/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Figura 1.11: Exemplo da estratégia CoT com o agente de conversação ChatGPT (realizado em 01 de agosto de 2023)

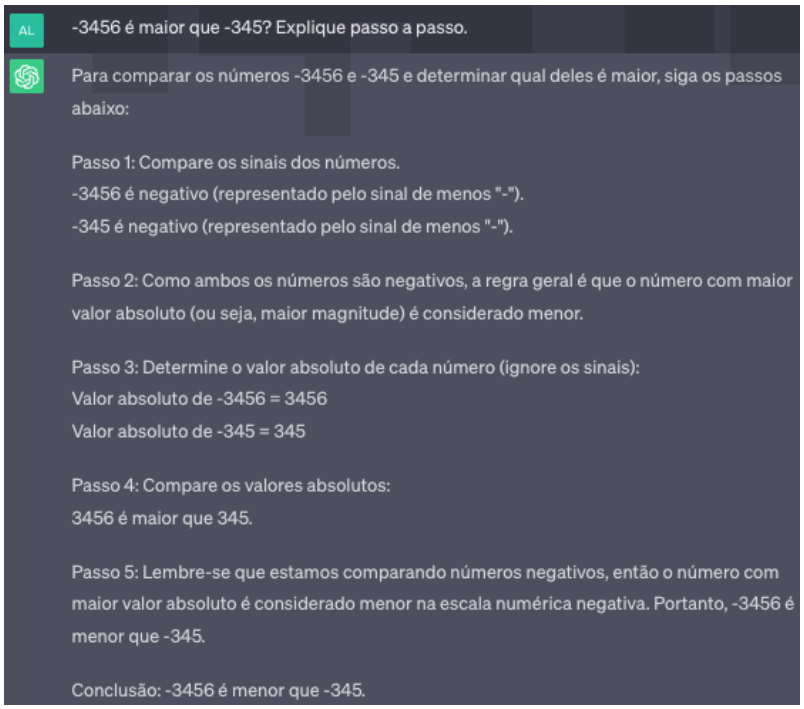
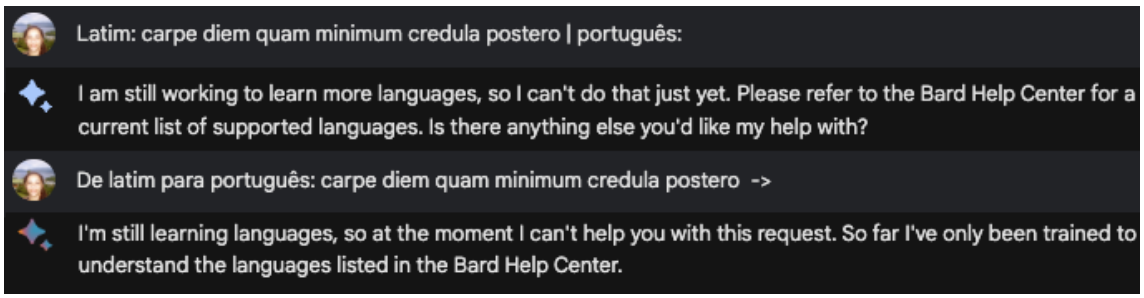


Figura 1.12: Tentativa de tradução do latim para o português com o BARD.



pletos é o PromptSource[53], que inclui a coleção P3 (*Public Pool of Prompts*). O P3 é composto de mais de 2000 opções de *prompts* para diversas tarefas de PLN, porém tudo em inglês. Entender como os LLMs produzem as saídas de acordo com as entradas que são dadas para eles por meio de *prompts* é um campo de estudo recente, porém bastante ativo, desde os primeiros resultados dos LLMs (Xie et al., 2022; Xu et al., 2023).

1.4.2 Treinamento Eficiente de Modelos de Linguagem Neurais

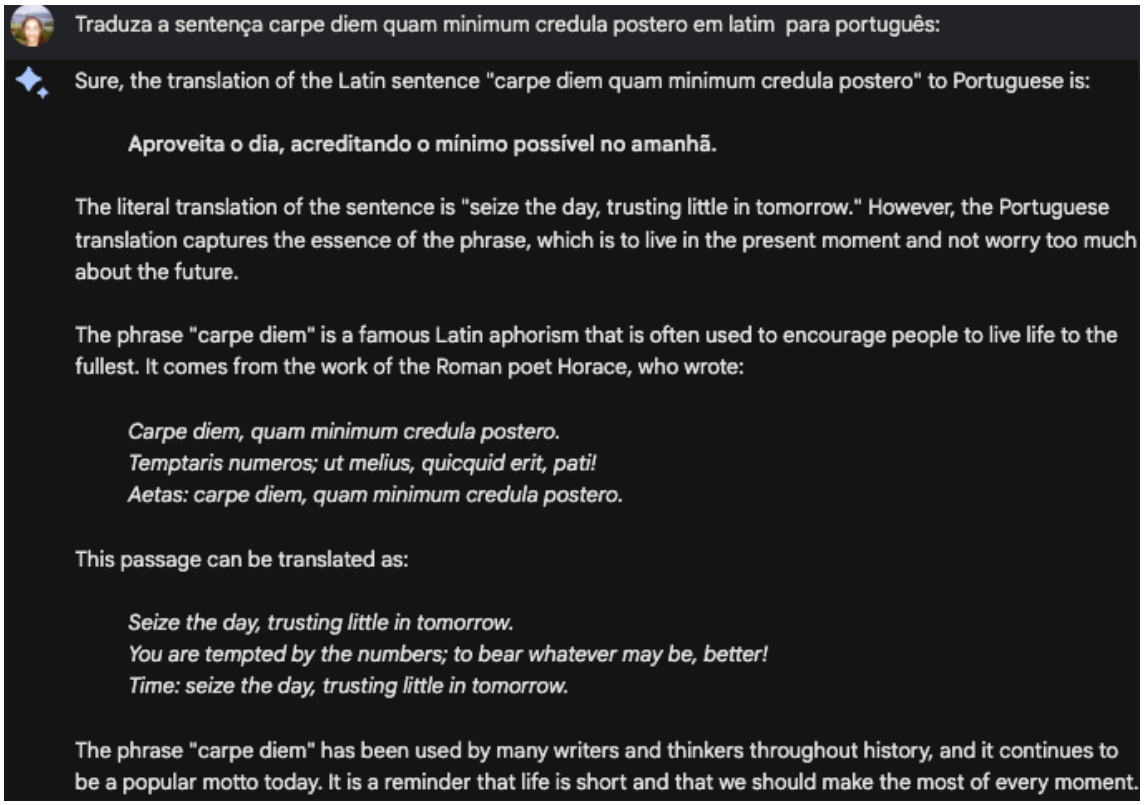
Embora LLMs possam ser usados sem nenhum ajuste em seus pesos, eles acabam por ficar muito dependentes dos *prompts* e da exposição implícita que o modelo teve para uma certa tarefa durante o seu pré-treinamento. Assim, o desempenho de modelos que se baseiam apenas na habilidade emergente gerativa pode ficar bem abaixo do desempenho de um outro modelo, ainda que menor, que é ajustado para uma tarefa específica (Raffel et al.,

[53] https://github.com/bigscience-workshop/promptsource

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol2/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Figura 1.13: Tentativa de tradução do latim para o português com o BARD, com um *prompt* diferente, mais informativo.



2020a). Mas como ajustar um modelo de bilhões de parâmetros de forma razoavelmente eficiente? Para responder a esta pergunta, novas abordagens sugerem que o treinamento seja feito em apenas partes dos modelos, ou com estratégias baseadas em reparametrização das matrizes de pesos.

Para o primeiro caso, podemos mencionar três estratégias:

1. *Soft prompt tuning*, ou apenas *prompt tuning* (Lester et al., 2021). Neste caso, o modelo fica congelado, exceto por uma quantidade adicional de k parâmetros numéricos ajustáveis – por isso o *soft* – que são concatenados no início dos *embeddings* do texto de entrada. Esses k parâmetros serão treinados de acordo com a tarefa-alvo, usando o algoritmo clássico de retro-propagação. Observe a diferenca entre as versões *hard* e *soft*: a primeira não tem ajuste de parâmetros, se baseando apenas na troca de palavras na instrução, enquanto a segunda é diferenciável, ou seja, o *prompt* é composto por um conjunto de pesos ajustáveis.
2. *Prefix tuning* (Li; Liang, 2021). Nesta estratégia, pesos ajustáveis são acrescentados no início de cada bloco dos Transformers. Observe que o modelo "original" permanece congelado, sem ajustes, assim como na abordagem de *soft prompt tuning*. Porém, enquanto lá pesos ajustáveis aparecem apenas no início dos *embeddings* de entrada, que seriam mesmo o local de inserção das instruções, aqui eles são concatenados no início de cada bloco do Transformer. Ainda, antes da concatenação, eles passam por duas camadas de redes neurais completamente conectadas, para garantir que o prefixo esteja em um mesmo espaço de representação vetorial que a entrada

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol2/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

do bloco. Ou seja, o processo de adaptação de prefixos, teoricamente, é mais custoso que o processo de adaptação de *prompts*. Assim, a ordem de processamento do bloco do Transformer se torna: CAMADA COMPLETAMENTE CONECTADA PARA PROCESSAMENTO DOS *PROMPTS* -> CONCATENAÇÃO DA SAÍDA ANTERIOR COM A ENTRADA DO MODELO -> AUTO-ATENÇÃO -> NORMALIZAÇÃO -> CAMADA COMPLETAMENTE CONECTADA DO TRANSFORMER -> NORMALIZAÇÃO (desconsiderando as conexões residuais).

3. Adaptadores (Houlsby et al., 2019). Adaptadores também acrescentam pesos ajustáveis adicionais a cada bloco do Transformer, mas não no início do bloco e sim no meio do bloco. Assim, os adaptadores são camadas de rede neural completamente conectadas, com uma função de ativação não-linear entre elas, introduzidas imediatamente antes da camada de normalização. Ou seja, a ordem de processamento se torna: AUTO-ATENÇÃO -> ADAPTADOR -> NORMALIZAÇÃO -> CAMADA COMPLETAMENTE CONECTADA DO TRANSFORMER -> ADAPTADOR -> NORMALIZAÇÃO (desconsiderando as conexões residuais, para facilitar a comparação com os *prompts*).

Já as abordagens baseadas em reparametrização tem o método de adaptação baseado no posto das matrizes de peso, LoRA (do inglês, *Low-Rank Adaptation*) (Hu et al., 2022) como seu principal representante. A motivação principal vem de um estudo anterior, que apontou que modelos ajustados para uma nova tarefa possuem uma dimensão menor que os modelos pré-treinados (Aghajanyan et al., 2021), ou seja, que eles poderiam ser decompostos para matrizes menores sem perder informação. Dessa forma, o método aprende como decompor as matrizes de atualização dos gradientes para postos menores.

LoRA também se motiva no espaço de memória necessário para armazenar as mudanças nas matrizes de peso durante o seu treinamento. Nesta mesma direção, abordagens baseadas em quantização, que guardam os pesos de treinamento em variáveis tipadas com menos precisão, também têm sido foco de investigação recentemente[54] (Dettmers et al., 2023).

Um outro ponto a ser considerado com o uso de LLMs (e até mesmo LMs) é o tamanho da entrada. Considerando que a o método de atenção tem uma complexidade de ordem quadrática, a maioria dos modelos baseados em Transformers usualmente limitam a sua entrada em cerca de 500 a 1024 *tokens*. Entradas maiores, em geral, precisam ser truncadas. Mesmo abordagens que consideram matrizes de atenção esparsas, como Longformer (Beltagy et al., 2020), ainda têm limitações. Abordagens recentes armazenam e recuperam camadas de decodificação ou separam entradas em pedaços de tamanho menores, possibilitando lidar com textos de tamanhos até 500 mil *tokens* (Bertsch et al., 2023; Ivgi et al., 2023; Wu et al., 2022).

1.4.3 Estratégias de Treinamento para Agentes de Conversação: alinhamento e feedback humano

Um modelo de linguagem não é treinado explicitamente para interagir com usuários, apenas para completar sentenças. Para criar um agente de conversação tendo como base um modelo de linguagem, é necessário incluir no modelo a habilidade de tentar responder ao usuário de acordo com a sua intenção expressa nas instruções, ou seja se alinhar ao diálogo, acompanhando a conversa. Idealmente, o agente de conversação também deve evitar respostas indevidas que poderiam levar a comportamentos nocivos.

[54] https://huggingface.co/blog/hf-bitsandbytes-integration

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol2/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Assim, em (Ouyang et al., 2022) os autores tinham como motivação "tornar modelos de linguagem **úteis** – ajudando os usuários a resolver tarefas – **honestos** – não inventando informação ou levando o usuário para uma falsidade – e **inofensivos** – não causando algum mal físico, psicológico ou social". É menos desafiador indicar se os modelos atuais conseguem ter a primeira característica. Entretanto, as duas últimas não podemos afirmar com convicção que foram alcançadas, nem mesmo pelos modelos mais atuais.

Para tornar possível tal alinhamento entre a saída de um modelo de linguagem e a intenção do usuário, recorre-se a uma outra camada de aprendizado, o **Aprendizado por Reforço com Feedback Humano**, ou a partir de preferências humanas, (do inglês, *Reinforcement learning from Human Feedback* ou RLHF) (Christiano et al., 2017) uma abordagem que já havia se mostrado frutífera em visão computacional. O aprendizado por reforço se vale de uma função de recompensa: caso a saída seja adequada, a recompensa é positiva; caso contrário, devolve-se uma penalidade. Ou seja, se o modelo estiver devolvendo uma resposta adequada – e aqui o adequado seria a resposta que obedecesse aos três princípios acima, de utilidade, honestidade e inofensibilidade – então a recompensa seria positiva.

Acontece que definir tal valor de recompensa não é trivial e este tem sido um dos grandes desafios da área de aprendizado por reforço. Nos agentes de conversação, o que é feito, é aprender um modelo de recompensa a partir de exemplos. Pares de *prompts* e respostas são gerados, usualmente de forma automática pelos modelos, por questões de escala. Mas nada impede que esses pares também sejam curados em outros conjuntos de dados ou definidos por pessoas (Zhou et al., 2023a). A partir daí, anotadores vão dizer quais são as suas saídas preferidas, geralmente usando mais de um modelo para ter alguma base de comparação. Então, considerando essa saída, o modelo de recompensa pode ser treinado.

Embora os agentes de conversação baseados em modelos de linguagem já estejam sendo usados para resolver problemas reais, muito ainda precisa ser alcançado. Em particular, duas limitações podem fazer com que tais modelos ainda não estejam prontos para serem adotados em larga escala e em aplicações sensíveis: (i) ainda são poucas as línguas que tais modelos conseguem lidar, se compararmos com a quantidade de línguas que temos no mundo; e (ii) ainda existe um viés social negativo embutido em tais modelos. Para ver um exemplo, considere os exemplos das Figuras 1.14 e 1.15 e perceba a diferença ao completar uma frase para o gênero masculino e feminino.

Isso no faz pensar, à medida que fechamos este capítulo, que é nossa responsabilidade como pesquisadores e desenvolvedores de tecnologia, que devemos considerar o potencial positivo e negativo dos modelos de linguagem de revolucionarem a nossa compreensão e interação com a tecnologia, e de estabelecerem novas formas de comunicação. Devemos, sim, celebrar os progressos que foram feitos, os insights que foram obtidos e as barreiras que foram rompidas. Mas à medida que a pesquisa em modelos de linguagem avança, e novas ferramentas a partir deles surgem e se tornam tão populares, precisamos ter em mente os limites que continuam a ser desafiados e a responsabilidade ao se usar e aplicar modelos cujas saídas ainda fogem da nossa compreensão. Assim, relembrando Alan Turing, mas do que nunca podemos afirmar que *"We can only see a short distance ahead, but we can see plenty there that needs to be done"*[55], em particular se considerarmos a língua portuguesa e tantas outras milhares espalhadas no planeta que ainda carecem da nossa atenção (sem trocadilhos) e compreensão.

[55]Tradução nossa: Somos capazes de ver apenas uma curta distância à frente, mas já podemos ver que muito ainda precisa ser feito.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol2/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Figura 1.14: Exemplo de viés de gênero para o português no agente de conversação BARD. Aqui, solicitou-se ao agente completar a frase “Ele trabalha no hospital como”.

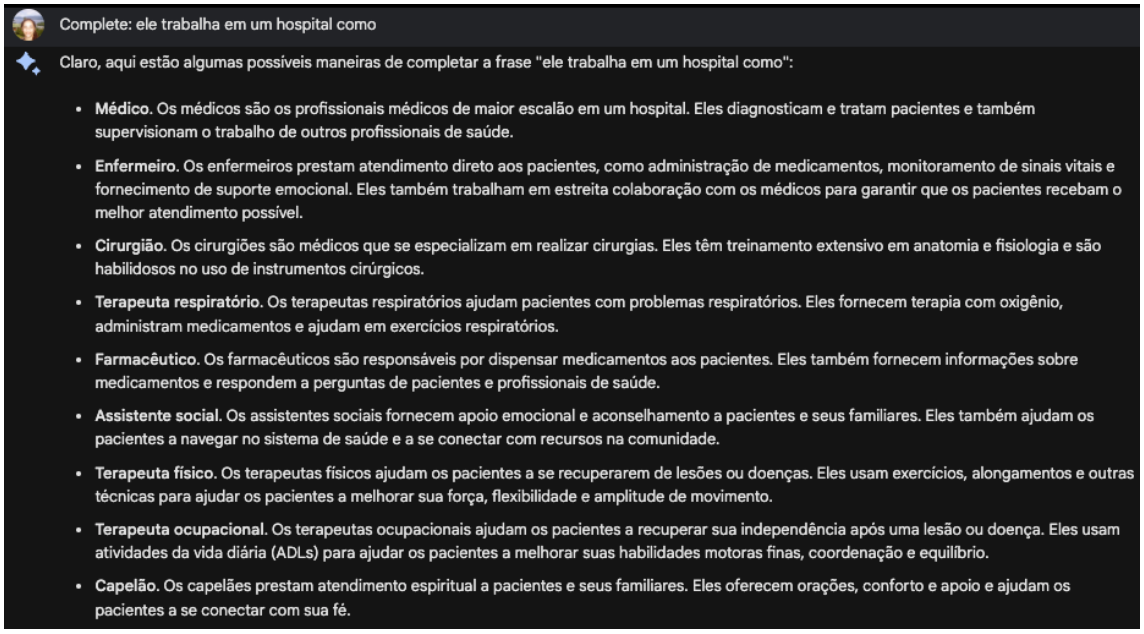
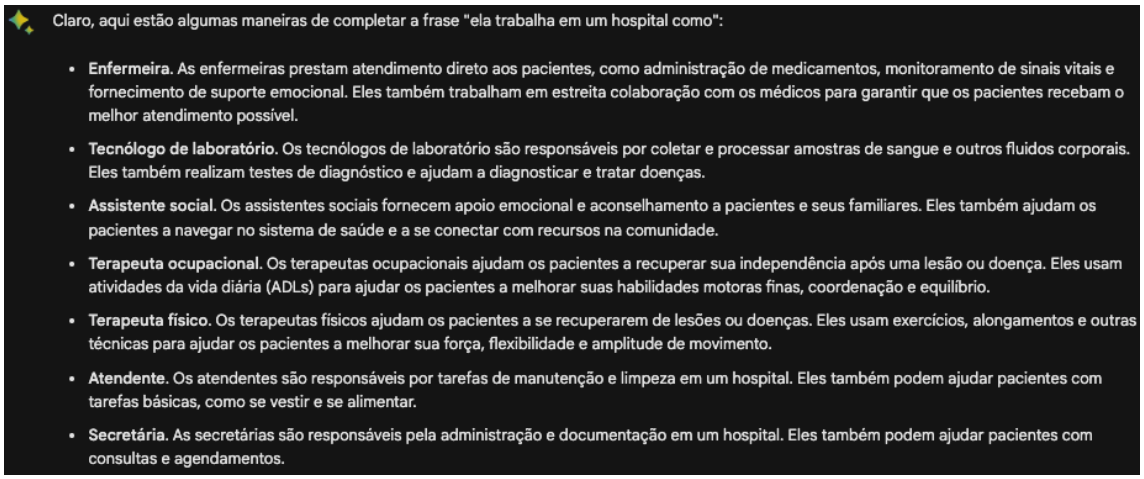


Figura 1.15: Exemplo de viés de gênero para o português no agente de conversação BARD. Aqui, solicitou-se ao agente completar a frase “Ela trabalha no hospital como”.



https://brasileiraspln.ufscar.br/livro-pln-4ed-vol2/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Capítulo 2

O papel dos dados no pré-treinamento de Grandes Modelos de Linguagem

Miguel de Mello Carpi
Felipe Ribas Serras
Mariana Lourenço Sturzeneker
Mayara Feliciano Palma
Gabriela Alves Lachi
Aline Silva Costa
Maria Clara Paixão de Sousa
Cristiane Namiuti
Vanessa Martins do Monte
Marcelo Finger

Publicado em: 16/04/2026

> "Em suma, tanto naquelas leituras se enfrascou, que as noites se lhe passavam a ler desde o sol posto até à alvorada, e os dias, desde o amanhecer até fim da tarde. E assim, do pouco dormir e do muito ler se lhe secou o cérebro, de maneira que chegou a perder o juízo.
>
> Encheu-se-lhe a fantasia de tudo que achava nos livros, assim de encantamentos, como pendências, batalhas, desafios, feridas, requebros, amores, tormentas, e disparates impossíveis; e assentou-se-lhe de tal modo na imaginação ser verdade toda aquela máquina de sonhadas invenções que lia, que para ele não havia história mais certa no mundo."
>
> — *Dom Quixote de La Mancha*, Miguel de Cervantes (Cervantes, 1605)

2.1 Introdução

As palavras que iniciam este capítulo foram extraídas do célebre romance espanhol *Dom Quixote de la Mancha*, no qual Miguel de Cervantes narra a vida do fidalgo Alonso Quixano, um homem nobre de meia idade, que, de tanto ler romances de cavalaria, acreditou-se ele próprio um cavaleiro e engajou-se nas mais insanas desventuras.

Com sua obra, Cervantes nos alerta sobre o poder que a informação que consumimos tem sobre nós e sobre como, providos dos contextos certos (ou errados), podemos tomar um moinho por um gigante ou um gigante por um moinho.

Guardadas as devidas proporções, alguém interessado em treinar um grande modelo de linguagem, se depara com o mesmo problema: de que forma o enorme volume de textos usado no processo de treinamento pode afetar o modelo resultante? Quais aspectos desses

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

textos influenciarão o modelo e suas capacidades? O que podemos (ou não) fazer em relação a isso?

Os grandes modelos de linguagem que temos a nossa disposição atualmente — baseados na tecnologia Transformer e suas variantes — possuem, pelo menos, duas etapas principais de treinamento: o **pré-treinamento** e o **ajuste fino**.

O **ajuste fino** corresponde ao processo que usualmente visualizamos, quando pensamos em treinar alguém para a realização de uma tarefa: a exposição do aluno a um conjunto representativo de exemplos de como realizar aquela tarefa corretamente.

Mas, para podermos ensinar a um ser humano uma tarefa específica, usualmente pressupomos que ele possua um amplo arcabouço de capacidades gerais prévias, inclusive as que permitam a ele se comunicar conosco e aprender a realizar tarefa em si.

Quando falamos de grandes modelos de linguagem, essas informações contextuais básicas são adquiridas na fase do **pré-treinamento**. É nessa fase que o modelo será exposto a grandes volumes de dados de linguagem — não anotados para uma tarefa específica, mas sim amostras abrangentes de como a linguagem humana funciona.

Durante o pré-treinamento, o modelo obtém valores para os seus parâmetros internos que vão permitir a ele usar palavras de acordo com seu significado apropriado, formar sentenças seguindo as regras sintáticas da língua, imitar estratégias discursivas como as usadas pelos seres humanos.

Muito mais do que isso, entretanto, os parâmetros dos modelos passam a encapsular, também, os vieses e conhecimento de mundo presentes nos textos, as estruturas das inferências neles realizadas, a maneira como eles se estruturam, variam, se adaptam, e as estratégias segundo as quais os textos armazenam, fornecem e transformam a informação neles contida.

Para o bem ou para o mal, as surpreendentes capacidades que os modelos de linguagem têm demonstrado nos últimos anos, são testemunhos de que eles aproximam — não só a gramática de uma língua e de como gerá-la — mas uma gramática de mundo, uma gramática das operações linguísticas que realizamos sobre a realidade e que a realidade realiza sobre a nossa língua.

Neste capítulo focamos, justamente, nessa fase do pré-treinamento de um grande modelo de linguagem. Tentamos, de forma breve, fornecer uma noção geral de como os dados usados influenciam no processo de pré-treinamento, nos modelos resultantes e nas suas capacidades e características. Como no restante do livro, daremos um foco especial para a língua portuguesa e para recursos para ela disponíveis.

Para tal, começaremos discutindo em que situações o pré-treinamento de um modelo se faz necessário (Seção 2.2), apresentando em seguida os principais conjuntos de dados disponíveis para o pré-treinamento em português (Seção 2.3). Iniciaremos então discussões acerca da natureza dos dados e quais principais aspectos devemos considerar na sua seleção, incluindo tópicos como a importância da curadoria (Seção 2.4), da qualidade (Seção 2.5) e da diversidade (Seção 2.6) dos dados usados no pré-treinamento, além das principais problemáticas referentes à proveniência dos dados (Seção 2.7). Por fim, apresentaremos brevemente o *corpus* Carolina, como um estudo de caso de como tratar várias dessas questões em tempo de montagem de um *corpus* ou *dataset* para treinamento de modelos de linguagem (Seção 2.8).

Para acompanhar o texto, recomendamos a leitura prévia ou complementar do Capítulo 1, que apresenta os modelos de linguagem e como eles funcionam, do Capítulo Conjunto de dados, *dataset* e *corpus*, que versa sobre *corpora* e *datasets* anotados para tarefas específicas — muitas vezes usados na etapa posterior do ajuste fino desses modelos — e do

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol2/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Capítulo 3, que se aprofunda em técnicas e detalhes do treinamento e como essas podem impactar os produtos finais.

2.2 Por que e quando pré-treinar?

Como mencionado anteriormente, no paradigma vigente, o treinamento é constituído de dois estágios principais: o estágio do **pré-treinamento** envolve treinar uma grande rede neural em um conjunto vasto de textos de domínios e origens diversas — usualmente extraídos da web — para que o modelo aprenda a estrutura da língua (ou das línguas), além de obter conhecimento geral do mundo (Radford et al., 2018). O segundo estágio, do **ajuste fino**, envolve adaptar o modelo geral obtido para uma tarefa específica — como análise de sentimentos (Sun; Dredze, 2025) — usando dados anotados para aquela tarefa.

A decisão de pré-treinar um modelo de linguagem do zero ou de realizar o ajuste fino de um modelo pré-treinado já existente é uma das mais importantes no desenvolvimento de modelos de linguagem baseados em grandes redes neurais. Essa decisão representa uma troca entre eficiência e eficácia, já que o pré-treinamento de um modelo requer grandes volumes de dados e *clusters* de GPUs para processá-los, ao passo que o ajuste fino requer menos dados e menos processamento, a custa do controle sobre o modelo pré-treinado e o conteúdo usado para tal.

Em algumas situações, uma solução intermediária entre pré-treinar um modelo do zero e apenas ajustar um modelo já pré-treinado, é a de realizar um **pré-treinamento continuado**. Como o nome sugere, o pré-treinamento continuado consiste em continuar com o tipo de pré-treinamento inicial, mas com dados de linguagem mais específicos para a tarefa fim (Kerner, 2024). É pouco provável, por exemplo, que um modelo pré-treinado com dados da internet tenha recebido um vasto conhecimento de literatura biomédica, dessa forma se o objetivo final é classificar textos biomédicos, faz sentido, antes, apresentar ao modelo textos desse domínio. Entretanto, há possíveis desafios na introdução dessa nova etapa: ao treinar o modelo com novos dados ele pode acabar se esquecendo do conteúdo já aprendido, um fenômeno conhecido como esquecimento catastrófico (Mecklenburg et al., 2024); além disso, o tokenizador, peça fundamental para o modelo conseguir ler os textos, pode não estar ajustado para o léxico do domínio em questão, sendo necessário adaptá-lo primeiro, o que pode culminar na necessidade de um pré-treinamento completo (Gu et al., 2021).

Além da incapacidade de se adaptar o léxico, outra possível razão para realizar o pré-treinamento ocorre quando você precisa que o seu modelo seja executado utilizando baixos recursos computacionais, inferiores aos exigidos pelos modelos pré-treinados disponíveis, o que leva a necessidade de escolher uma arquitetura neural que ainda não foi nem pré-treinada para o idioma escolhido e muito menos adaptada para a tarefa desejada (Yao et al., 2022).

Uma situação semelhante surge quando a tarefa fim ou a arquitetura do modelo são fundamentalmente novas. O ajuste fino é um processo de adaptação, não de criação; ele refina comportamentos existentes, mas dificilmente consegue instilar capacidades inteiramente novas. Estudos mostram, por exemplo, que modelos pré-treinados podem ter dificuldades com tarefas que exigem raciocínio lógico complexo, como operações *booleanas*, mesmo após o ajuste fino, enquanto modelos treinados do zero para aquela tarefa específica podem apresentar um desempenho superior (Jensen; Plank, 2022).

A própria estrutura dos dados da tarefa pode, também, justificar um novo pré-treinamento. Pesquisas recentes demonstram que alinhar os dados de pré-treinamento

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

com os da tarefa final desde o início é computacionalmente mais eficiente, podendo atingir o mesmo desempenho com uma fração dos recursos (Mizrahi et al., 2025).

O pré-treinamento do zero também se justifica quando é necessário alterar a **arquitetura** para comportar parametrizações distintas, como um contexto maior, por exemplo (Li et al., 2022). Além da arquitetura, os próprios **objetivos de treino** podem justificar um novo pré-treinamento. Os objetivos padrão mais comuns são o **MLM (*Masked Language Model*, ou Modelo de Linguagem Mascarado)**, no qual o modelo aprende a prever *tokens* que foram aleatoriamente ocultados (mascarados) no texto de entrada, e o **CLM (*Causal Language Model*, ou Modelo de Linguagem Causal)**, onde o modelo é treinado para prever o próximo *token* em uma sequência. Se a tarefa ou o domínio exigir um objetivo de treino fundamentalmente novo, que não se encaixe nesses moldes, o pré-treinamento do zero pode ser a única opção viável.

Talvez a razão mais comum e empiricamente validada para o pré-treinamento, no entanto, seja a incompatibilidade fundamental entre o domínio da aplicação e o conhecimento contido nos modelos de propósito geral. Um caso de estudo exemplar é o do PubMedBERT, um modelo pré-treinado do zero exclusivamente com textos da literatura biomédica. O primeiro problema que ele resolve é o da tokenização: modelos gerais quebram termos técnicos como "naloxona" em múltiplos fragmentos sem significado ([na, ##lo, ##xon, ##a]), diluindo sua representação semântica. Com um vocabulário próprio, o PubMedBERT trata "naloxona" como um único *token*, aprendendo uma representação mais rica e eficiente. Além disso, o conhecimento de um domínio geral pode ser irrelevante ou até mesmo prejudicial em um contexto especializado, um fenômeno conhecido como "transferência negativa". Em uma vasta gama de tarefas de processamento de linguagem natural biomédica, o PubMedBERT superou não apenas os modelos de domínio geral, mas também modelos que passaram por pré-treinamento continuado, demonstrando o valor de construir uma nova base de conhecimento quando o domínio é suficientemente distinto (Gu et al., 2021).

Nas demais situações, a realização de um ajuste fino para a tarefa fim, com ou sem a etapa intermediária do pré-treinamento continuado, tende a ser a escolha mais razoável. Resumindo a discussão desta seção, o Quadro 2.1 apresenta os principais critérios para ajudar na decisão de pré-treinar ou não.

Se o leitor se encontrar numa situação onde o pré-treinamento é necessário, ele se deparará com um novo desafio: o da seleção de quais conjuntos de dados usar no pré-treinamento e de quais aspectos levar em conta para essa seleção. Para ajudar nesse processo, apresentamos, na seção seguinte alguns dos principais conjuntos de dados disponíveis para o pré-treinamento em português e, nas seções seguintes, discutimos quais aspectos levar em conta nessa seleção.

Quadro 2.1: Guia para avaliação se o pré-treinamento é necessário ou não.

Critérios de avaliação	Ajuste fino / pré-treinamento continuado	Pré-treinar do zero
1. Domínio e Dados		

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol2/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

1.1 Vocabulário	Linguagem padrão, jargão limitado. Tokenizador existente funciona bem.	Léxico vasto, único e especializado. Tokenizador padrão não fragmenta bem os termos.
1.2 Dados disponíveis	*Dataset* rotulado pequeno/médio.	*Corpus* com pelo menos 1B *tokens*.
1.3 Conhecimento geral	Conhecimento de mundo é útil.	Conhecimento de mundo não é necessário.
2. Tarefa e Capacidades		
2.1 Novidade da tarefa	Tarefa padrão de PLN em um novo domínio.	Tarefa exige raciocínio novo e complexo. Conhecimento médico, jurídico, ou matemático.
2.2 Teto de desempenho	Ajuste fino gera um desempenho "bom o suficiente".	Ajuste fino gera modelos que falham em casos essenciais.
3. Arquitetura e Tecnologia		
3.1 Adequação da arquitetura	Arquitetura Transformer padrão é suficiente.	Requer uma arquitetura para desafios específicos.
3.2 Objetivos de treino	Objetivos padrão (MLM, CLM) são suficientes.	Requer novos objetivos para o domínio.
4. Recursos e Estratégia		
4.1 Orçamento e prazo	Orçamento e tempo limitados (dias/semanas).	Orçamento alto e bastante tempo (meses). Acesso a cluster de GPUs.
4.2 Importância estratégica	Desempenho "bom o suficiente" é aceitável. Será necessário realizar mais de uma tarefa.	Dominar o nicho é um objetivo central. O modelo é um ativo estratégico chave.

2.3 Conjuntos de dados para pré-treinamento em português

Nesta seção, apresentaremos alguns dos conjuntos de dados disponíveis que podem ser usados para o pré-treinamento de modelos em português. Nossa apresentação visa fornecer um panorama geral ao leitor, sem pretender ser uma revisão exaustiva dos recursos disponíveis. Com esse objetivo, dividimos a apresentação em duas grandes categorias, acompanhando a criação de recursos para a língua portuguesa em diferentes momentos ao longo dos últimos vinte anos: (i) recursos gerais de conjuntos de dados multilíngues, com parte dos dados disponíveis em português; (ii) *datasets* desenvolvidos especificamente em – e para o – português.

2.3.1 Recursos gerais

Atualmente existem vários recursos multilíngues de propósito geral que podem ser usados para o pré-treinamento de grandes modelos de linguagem e aplicações similares. Eles podem ser úteis para pesquisadores interessados em treinar modelos multilíngues, mas também para aqueles que desejam focar em uma língua específica – no nosso caso, o português –, bastando extrair a parte dos dados referentes ao idioma de interesse. A seguir, apresentamos brevemente os seguintes recursos desse tipo: ***Common Crawl***, **OSCAR**,

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol2/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

HPLT, e os recursos da ***TenTen Corpus Family***.

A partir dos anos 2000, vimos um esforço de projetos voltados ao acesso a dados da web. Um dos principais representantes dessa categoria, o ***Common Crawl*** (Common Crawl, [s.d.]), fornece um repositório aberto com "fotografias" do conteúdo de páginas web desde 2008, chegando a *petabytes* de dados. A ideia é disponibilizar o conteúdo da maneira como foi publicado, focando especificamente em preservar o código HTML das fontes, e disponibilizá-lo no formato WARC (*Web ARChive*) para que informações sobre a requisição e seus metadados estejam também acessíveis, armazenando inclusive as interações entre o servidor e o(s) *web crawlers*. A organização geral desses dados não é orientada pela língua em que estão escritos, mas pelo ano em que determinado *batch* foi coletado da internet.

De 2013, o ***TenTen Corpus Family*** (Jakubček et al., 2013) é uma família de *corpora* que também parte da proposta da coleta de uma grande quantidade de dados disponíveis na web, mas com o diferencial de incorporar apenas conteúdo com valor linguístico para mais de 50 línguas. O *corpus* voltado para o português inclui as variedades europeia e brasileira e possui quase 17 bilhões de palavras coletadas entre agosto de 2020 e novembro de 2023. Além da anotação morfossintática (*part-of-speech*), possui também anotação de gêneros do discurso e classificação em tópicos. Diferente do *Common Crawl*, não é um *corpus* aberto, sua versão completa pode ser acessada mediante pagamento.

Com sua primeira publicação em 2019, o **OSCAR (*Open Super-large Crawled Aggregated coRpus*)** (Abadji et al., 2022a) também é um grande projeto de coleta de dados da web por meio de *crawlers*. Trata-se de um *corpus* multilíngue de código aberto, voltado para aplicações em inteligência artificial e aprendizado de máquina. O OSCAR busca dar especial atenção a comunidades linguísticas com poucos recursos disponíveis e atualmente disponibiliza dados em 166 línguas, incluindo o português.

O projeto ***High Performance Language Technologies (HPLT)*** (Burchell et al., 2025), que teve sua primeira publicação de *datasets* em 2024, tem como objetivo reunir um grande volume de dados em diversas línguas para computação de alta performance, visando produzir modelos de linguagem e tradução. A versão mais recente de seus conjuntos de dados está disponível sob a licença *Creative Commons CC0* e conta com 4,5 *petabytes* de dados da web comprimidos, majoritariamente advindos do *Internet Archive* e *Common Crawl*. O foco da nova versão está na maior qualidade dos textos para treinamento de modelos, buscando diminuir ruído e duplicação. Dados em português estão disponíveis no *dataset* monolíngue do *HPLT*, *monoHPLT-PT*, contando com 470,63 milhões de documentos e 203,71 bilhões de palavras no subconjunto deduplicado e 237,81 milhões de documentos e 146,27 bilhões de palavras no subconjunto deduplicado e com artefatos ruidosos removidos.

2.3.2 Recursos para o português

Agora, apresentaremos *corpora* desenvolvidos exclusivamente para a língua portuguesa. Apesar de não cobrirem outras línguas, eles oferecem um panorama muito mais diverso do próprio português. Assim como nos recursos gerais, passaremos brevemente por recursos construídos ao longo dos últimos anos, como a **Linguateca**, o ***Corpus* do Português**, o **brWaC** e o **BlogSET-BR**, para depois mencionarmos recursos mais recentes, que já surgem em resposta à revolução dos grandes modelos de linguagem.

A **Linguateca** (Santos, 2000) é uma organização virtual que concentra recursos disponíveis para o processamento computacional da língua portuguesa. Por ser um conjunto bastante diverso de recursos, destacaremos aqui apenas alguns dos *subcorpora* mais rele-

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol2/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

vantes no que tange ao volume de dados:

- O ***Corpus* do Português** é um *corpus* anotado que permite buscas por palavra, frase, classe de palavra (part-of-speech), entre outros, e se divide em três *subcorpora*: (1) "*Genre / Historical*" é o braço histórico, criado em 2006, com textos de 1200 a 1900 e 45 milhões de palavras; (2) "*Web / Dialects*" (Davies; Ferreira, 2016) contém 1 bilhão de palavras obtidas a partir de páginas da web de quatro países lusófonos, seu diferencial sendo a possibilidade de comparar dialetos; (3) "*NOW (2012 - 2019)*" (Davies; Ferreira, 2018) tem textos dos mesmos países, mas cobrindo o período de 2012 a 2019.

- O ***Corpus* Brasileiro** (Sardinha et al., 2010) é um conjunto de aproximadamente um bilhão de palavras do português brasileiro escrito e falado, com anotação sintática e alguma anotação semântica, além de atributos de gênero discursivo.

O ***Brazilian Portuguese Web as Corpus* (brWaC)** (Wagner Filho et al., 2018) é um *corpus* construído seguindo a estrutura *WaCky*. O *WaCky* (*The Web-As-Corpus Kool Yinitiative*) consitui ferramentas e interfaces que facilitam o uso de *web crawlers* por linguistas, para processar, indexar e pesquisar dados disponíveis na web. Na sua versão mais recente, disponibilizada em Janeiro de 2017 e disponível para download e navegação no *NoSketch Engine* (versão aberta do *Sketch Engine*), o brWaC possui 3,53 milhões de documentos e 2,68 bilhões de *tokens*, com anotação de 134 atributos textuais diferentes e classificação de acordo com o nível de legibilidade feito por modelos de aprendizagem de máquina.

O **BlogSET-BR** (Santos et al., 2018) é um *dataset* para o português brasileiro, voltado para a linguagem gerada por usuários para o ambiente virtual: é uma grande coleção de publicações feitas em *blogs*, com 2,1 bilhão de palavras extraídas de 7,4 milhões de publicações em mais de 808 *blogs* brasileiros. Um diferencial desse *corpus* é a informação sobre autoria vinculada aos textos.

Os conjuntos de dados listados até aqui, são conjuntos mais clássicos, desenvolvidos ao longo das últimas décadas, que permaneceram relevantes até os dias de hoje. Com a revolução dos grandes modelos baseados na tecnologia Transformer, entretanto, cresceram também os recursos disponíveis para o português. Nesse sentido, diversos conjuntos têm sido desenvolvidos voltados especificamente para o treinamento desses tipos de modelos, entre as contribuições mais recentes destacamos:

- **Carolina** (Crespo et al., 2023; Finger et al., 2020): um *corpus* aberto do português contemporâneo (1970-2021) com informações de procedência e tipologia. A versão mais recente, 2.0 Bea, possui 2,1 milhões de documentos e 1,2 bilhões de *tokens*. O *corpus* busca construir uma base de dados a partir de critérios definidos previamente, com controle sobre as fontes, gêneros textuais, data de publicação e metadados associados. Falaremos do Carolina com um pouco mais de profundidade nas próximas seções.

- **Aroeira** (Lira et al., 2024): um *corpus* curado, que contém 35,5 milhões de documentos e mais de 15,1 bilhões de *tokens*, com foco na qualidade dos dados, a serem usados para a tarefa de treinamento de modelos;

- **GigaVerbo** (Corrêa et al., 2025): um *dataset* que contém 145 milhões de documentos e concatena uma grande quantidade de outros *corpora*, inclusive alguns menci-

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol2/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

onados ao longo deste texto. Foi desenvolvido especificamente para o treinamento dos modelos da Família Tucano;

- **LegalPT** (Garcia et al., 2024): é um agregador de *datasets* jurídicos disponíveis publicamente em português, com 24.194.918 documentos em sua versão atual, que se difere dos demais por essa especialização em documentos jurídicos. Vemos também, em decorrência dessa especialização, o esforço de endereçar a questão da repetição dos documentos em corpora jurídicos com a construção de sua versão deduplicada, o LegalPT *(deduplicated)*.

2.4 A importância da curadoria de dados

Na seção anterior, apresentamos alguns dos principais conjuntos de dados disponíveis atualmente para o pré-treinamento em português. Dependendo da aplicação de interesse, um pesquisador pode adotar uma ou mais dessas fontes, ou desenvolver um conjunto de dados próprio, que pode incluí-las ou não. Em qualquer um desses casos, é importante refletir sobre o processo de curadoria de dados que serão usados, adaptando-o ao uso ou projeto específico. Vale, portanto, pararmos para discutir, mesmo que brevemente, como entendemos a curadoria de dados e qual a sua importância.

A construção de conjuntos de dados para o português é uma etapa essencial no desenvolvimento de tecnologias de processamento da linguagem, como discutido no Capítulo Conjunto de dados, *dataset* e *corpus*. No entanto, mais do que simplesmente acumular grandes volumes de texto, é fundamental considerar a curadoria desses dados. A promessa de uma curadoria cuidadosa é justamente a de garantir não apenas quantidade, mas também qualidade, diversidade e representatividade nos *corpora* utilizados para treinar e avaliar modelos.

Embora seja tecnicamente possível extrair textos em grande escala da internet, esse caminho apresenta uma série de desafios. A coleta bruta de dados traz consigo artefatos ruidosos (como *tags HTML* no meio do texto), duplicação, vieses negativos, problemas de direitos autorais e excessos ou lacunas linguísticas que podem comprometer a confiabilidade do conjunto final — o que, por sua vez, levanta questões sobre qualquer modelo treinado sobre ele.

Além disso, a ausência de filtros rigorosos pode fazer com que textos de baixa qualidade ou pouco relevantes ocupem espaço significativo no *corpus* resultante da coleta, prejudicando análises e aplicações práticas, além de acarretar maiores custos de desenvolvimento e manutenção.

Nos dias de hoje, entretanto, essa preocupação vai além dos problemas imediatos de qualidade e ganha contornos ainda mais críticos com descobertas recentes: um estudo publicado em julho 2024 na revista Nature (Shumailov et al., 2024) analisou os impactos de treinar ferramentas de Inteligência Artificial com dados gerados a partir delas. Conduzido por pesquisadores de instituições dos EUA, Reino Unido e Canadá, o trabalho mostrou que, após alguns ciclos de treinamento com esse tipo de dado, os modelos começam a cometer erros significativos e, em estágios mais avançados, passam a produzir informações incoerentes, num processo denominado "colapso" do modelo.

Segundo Ilia Shumailov (Shumailov et al., 2024), isso ocorre porque os dados não "compreendidos" pelo modelo original ficam sub-representados na linguagem gerada, então esse fenômeno de sub-representação vai se propagando, até que os erros se multiplicam a ponto de inutilizar os modelos. Esse risco ilustra a gravidade de se trabalhar com *corpora* não

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol2/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

controlados, já que qualquer aplicação que faça uso desses dados pode estar comprometida. Em especial, com a massificação do acesso a ferramentas de IA gerativa, torna-se cada vez mais difícil garantir que os textos coletados sejam de autoria exclusivamente humana — sobretudo aqueles produzidos a partir de 2022.

A curadoria atua, então, como um mecanismo de controle da qualidade dos dados: assegura a representatividade dos diferentes registros do português, reduz a presença de ruídos e vieses, e garante que os textos utilizados respeitem critérios éticos e metodológicos. Mais do que um simples repositório, o conjunto de dados se torna um recurso confiável para pesquisa e desenvolvimento em PLN.

Os fatores mencionados demonstram que, no cenário atual, a curadoria meticulosa não é um luxo, mas uma condição fundamental para o avanço da área. Nas seções seguintes nos aprofundamos em alguns desses aspectos.

2.5 O impacto da qualidade dos dados

Na Seção 2.2, discutimos como o pré-treinamento de um modelo de linguagem de ponta requer uma grande quantidade de dados textuais que, como visto na Seção 2.3, é muitas vezes proveniente da web. Sabemos que a web abriga conteúdos diversos, produzidos por diferentes públicos para diferentes fins e pode conter informações contraditórias ou desnecessárias para um determinado objetivo. Logo, no processo de preparação dos dados, é necessário garantir a qualidade dos textos que serão usados para treinar os modelos.

Mas o que é exatamente um texto de qualidade? Para os objetivos desta seção consideramos que um texto de qualidade é um texto escrito por um ser humano e que passou por um processo de seleção (Albalak et al., 2024). A seleção, por sua vez, costuma se orientar em três dimensões principais: a de **relevância**, a de **eficiência**, e a de **segurança**.

Essas dimensões de qualidade podem ser avaliadas tanto de forma **extrínseca** quanto **intrínseca**. A avaliação extrínseca mede o impacto dos dados no desempenho de um modelo já treinado, como, por exemplo, avaliar se um modelo treinado nesse *corpus* se torna mais propenso a gerar textos tóxicos. Já a avaliação intrínseca analisa características diretas do *corpus*; por exemplo, calcular a porcentagem de documentos duplicados (uma medida de eficiência) ou a frequência de termos de uma lista de bloqueio (uma medida de segurança). Nesta seção, focaremos apenas nas formas intrínsecas, que melhor se encaixam no escopo deste capítulo.

A primeira dimensão de qualidade é a da relevância. Um texto é dito relevante se ele está relacionado com a distribuição da tarefa a ser desempenhada pelo modelo, e se ele pode aumentar a sua capacidade. No contexto médico, uma lista de nomes de medicamentos pode apresentar termos que estão na distribuição desejada, porém ter somente o nome deles dificilmente aumentará a capacidade do modelo. Textos onde os nomes dos medicamentos aparecem acompanhados de uma descrição de efeitos colaterais são melhores já que apresentam mais informação relevante. Da mesma forma, um texto extraído da internet pode conter trechos de rodapé, de cabeçalho, de margens, ou de botões que não tem nada a ver com o conteúdo principal. E, antes de tudo isso, o texto está escrito em português?

Boa parte da adequação de idioma pode ser feita por meio dos metadados. Com eles é possível filtrar documentos cujo endereço pertença a um país lusófono, como feito no brWaC (Wagner Filho et al., 2018). De maneira alternativa ou complementar, classificadores automáticos podem ser utilizados (Albalak et al., 2024), como o FastText (Joulin et al., 2017), usado na construção do *corpus* Aroeira (Lira et al., 2024). Já para garantir

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol2/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

que não há ruído semântico, pode-se utilizar heurísticas (Albalak et al., 2024) como a presença de textos comuns em cabeçalhos e rodapés (ex: "Leia mais..."), a razão entre símbolos e palavras, o comprimento das sentenças e dos parágrafos, ou a repetição de *n-gramas* (Weber et al., 2024).

Além das heurísticas, a relevância semântica também pode ser avaliada com modelos. Uma abordagem comum é treinar um classificador para distinguir entre documentos de um *corpus* de referência de alta qualidade (com um "sinal positivo") e documentos de um *corpus* mais ruidoso (com um "sinal negativo") (Albalak et al., 2024; Wenzek et al., 2020). Outra alternativa é usar um modelo de linguagem para medir a perplexidade do texto (Touvron et al., 2023b), uma medida de quão surpreendente ele é, dado o conhecimento prévio de como os elementos linguísticos usualmente se distribuem dentro de um texto naquela língua. Nesse modelo, parágrafos com baixa perplexidade (menos surpreendentes) são considerados de maior qualidade (Wenzek et al., 2020).

Na segunda dimensão, a de eficiência, o objetivo é garantir que o conjunto dos dados torne o treinamento mais eficaz possível por meio do controle da redundância dos dados. Para isso costuma-se remover textos duplicados e trechos muito frequentes. Essa remoção faz com que o tempo de treinamento seja melhor usado, focando em conteúdo não-repetido (Lee et al., 2022) e, crucialmente, reduz a contaminação entre os conjuntos de treinamento e de teste (Longpre et al., 2025), o que permite uma avaliação mais robusta do modelo.

Há várias maneiras de remover conteúdo duplicado. Para uma remoção exata, o *hashing* pode ser utilizado para acelerar a computação (Wenzek et al., 2020). Também é possível utilizar o número de *n-gramas* em comum, e até mesmo a similaridade semântica calculada com modelos já treinados para remover documentos "quase idênticos" (Albalak et al., 2024; Lee et al., 2022).

Por fim, a terceira dimensão, a de segurança, tem como objetivo garantir que o modelo final siga os padrões de observância designados (Albalak et al., 2024; Longpre et al., 2025). A ideia é que um modelo é tão seguro quanto os dados utilizados para treiná-lo. As principais motivações estão em garantir que o conteúdo proteja a privacidade, siga padrões de observância, e que não seja tóxico. Aqui diremos que é tóxico qualquer conteúdo rude, desrespeitoso ou que contenha linguajar que faça uma pessoa desistir de um diálogo (Longpre et al., 2024, 2025; Rae et al., 2022). Já o conteúdo explícito pode incluir material pornográfico, violento, discriminatório, ou com viés de gênero e racial (Albalak et al., 2024; Lira et al., 2024).

As formas mais comuns para remover conteúdos indesejados consistem na utilização de listas de bloqueio de endereços web (fontes conhecidas por esse tipo de conteúdo) ou listas de bloqueio de palavras e expressões (Albalak et al., 2024). Esta última contém termos que, se presentes no texto, farão com que o documento inteiro seja descartado, como feito no T5 (Raffel et al., 2023) e no Aroeira (Lira et al., 2024). Como para as outras dimensões, também é possível utilizar aqui classificadores automáticos treinados para identificar toxicidade (Albalak et al., 2024). Já para garantir a privacidade, há ferramentas para remoção de nomes próprios, de endereço postal ou de e-mail, e de números de identificação pessoal.

2.5.1 O custo da qualidade

É importante ressaltar que a busca por qualidade dos dados não é absoluta e nem deve ser. Esta busca apresenta dois desafios centrais: a impossibilidade prática de uma filtragem perfeita, e o fato de que a remoção total de certos fenômenos pode ser indesejável para os

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol2/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

objetivos do modelo.

2.5.1.1 Quanto à impossibilidade

Em primeiro lugar, independente do nível de esforço dos desenvolvedores, todas as etapas de tratamento mencionadas utilizam modelos computacionais propensos a falhar, em especial quando estamos falando de um volume de textos tão imenso e diverso, que não pode ser verificado manualmente.

Pensando em heurísticas, um filtro que elimina textos com linhas muito curtas pode fazer com que textos como canções e poemas sejam descartados erroneamente (Abadji et al., 2022a; Albalak et al., 2024). Já listas de bloqueio de palavras podem fazer com que conteúdos legítimos de domínios como jurídico, médico ou produzido pela comunidade LGBTQIAPN+ sejam removidos por não levar em conta o contexto no qual as palavras aparecem (Kreutzer et al., 2022; Xu et al., 2021).

Os classificadores automáticos também podem cometer erros de classificação. Um modelo para reconhecer português, por exemplo, pode acabar descartando um texto regional caso não tenha sido treinado com dados da região em questão (Albalak et al., 2024), ou seja, se os textos apresentarem uma distribuição linguística distinta daquela dos dados de treinamento (Blodgett et al., 2020b).

2.5.1.2 Quanto à totalidade

Em segundo lugar, como discutimos anteriormente, quando estamos falando de pré-treinamento, estamos falando sobre ensinar a um modelo como a língua é usada, para que depois ele possa usar essa habilidade básica na resolução de tarefas específicas. Repetições e redundâncias são em algum nível parte da linguagem humana e, muitas vezes, nós mesmos usamos esse tipo de fenômeno para entender quais frases e expressões são mais comuns em determinados contextos e devem ser utilizadas em contextos similares. Há evidências de que manter um certo nível de duplicação é importante para o desempenho do modelo (Albalak et al., 2024; Lee et al., 2022).

De forma análoga, quando o objetivo final do modelo é uma tarefa específica, a remoção total de certos dados pode ser contraproducente. Se estivermos interessados em desenvolver um modelo que seja capaz de detectar textos tóxicos, a presença de amostras desse tipo de texto no pré-treinamento pode ser útil (Longpre et al., 2024, 2025). Isso permitiria ao modelo aprender os padrões de linguagem associados à toxicidade, que ele poderá usar posteriormente para aprender a detectar esse tipo de texto.

2.5.2 Considerações finais desta seção

Nesta seção, foram abordadas algumas dimensões referentes à qualidade dos dados. Em resumo, um dado de qualidade é aquele que é relevante, eficiente, e seguro, mas qual o nível de relevância, eficiência e segurança que podemos garantir pode variar bastante em função dos objetivos e das ferramentas que temos à disposição. No processo de falar sobre qualidade, acabamos tocando em tópicos relacionados a diversidade e a proveniência dos dados. Discutiremos esses aspectos nas próximas seções.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol2/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

2.6 O impacto da diversidade dos dados

Além da qualidade dos dados e suas várias dimensões, que discutimos na seção anterior, outro aspecto importante de considerarmos ao selecionar *datasets* ou *corpora* para o pré-treinamento de modelos de linguagem é o da diversidade linguística.

Muitas vezes, quando falamos de diversidade linguística, a primeira coisa que nos vem à mente são as diferenças entre línguas, como a inglesa e a portuguesa, ou de forma mais ampla, entre famílias linguísticas distintas, como a indo-europeia — que contém as línguas inglesa e portuguesa — e a família afro-asiática — que inclui línguas como o hebraico e o egípcio.

Por conta dessa intuição, podemos tomar a preocupação com diversidade linguística como um problema restrito a modelos multilinguísticos, treinados com dados de diversas línguas simultaneamente.

De fato, essa é uma questão fundamental no desenho e construção de modelos multilinguísticos, entretanto também o é na construção dos modelos monolinguísticos, voltados para uma língua específica, como no caso dos modelos especializados em português.

Isso porque o que entendemos como a língua portuguesa é, na verdade, um conjunto bastante diverso de manifestações linguísticas distintas que — apesar de encontrarem coesão nas regras básicas do português — estão submetidas a padrões e convenções significativamente diferentes.

Essas variações da língua podem se dar sobre diferentes dimensões. É comum que as pessoas deem maior atenção às dimensões de variação que elas entendam como marcas que as distinguam de outros grupos de usuários da língua. Esse é o caso de variações em função da distribuição geográfica, do grupo social e do período histórico (ver Wolfram (2006)).

Contudo, esse tipo de variação representa apenas uma parte do potencial de variação linguística de um idioma. Algumas dimensões de variação são mais dinâmicas, de forma que um único falante usualmente alterna entre variedades para fins distintos: esse é o caso da variação em **registro**, **gênero** e **estilo**, que definimos de acordo com Biber; Finegan (1994) e Biber; Conrad (2009):

- **Registro**: variação segundo a situação em que a linguagem é produzida, refletindo situações comunicativas que se repetem de forma regular dentro de uma estrutura social (Exemplos são: registro da conversação pessoal, registro da comunicação em fóruns virtuais, registro da escrita em livros didáticos).

- **Gênero**: variação segundo o tipo de mensagem e as convenções associadas à como aquele tipo de mensagem é construído (Exemplos são: entrada de enciclopédia, carta, manual, receita culinária.)

- **Estilo**: variação segundo as preferências estéticas de quem produz, ou consome a linguagem, podendo essas preferências serem mais ou menos coletivamente alinhadas (Exemplos são: o estilo da terceira fase do romantismo, o estilo pessoal de Machado de Assis)

Quando olhamos com atenção, vemos que uma pessoa usa a linguagem de forma bem diferente quando no tribunal , ou quando conversando com os amigos num bar (variação em registro); quando está escrevendo uma carta formal ou uma receita de bolo (variação em gênero); quando procura escrever um poema árcade ou cubista (variação em estilo).

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol2/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

O mais interessante é que, em todos esses casos, a linguagem não apenas varia de forma dinâmica, mas exerce um papel contextual. Basta percebermos como usar a linguagem apropriada para um dos contextos mencionados em outro desses contextos causaria estranhamento.

Do mesmo modo que um ser humano precisa ser capaz de alternar entre essas diversas variedades, de forma a cumprir tarefas de naturezas distintas com as quais ele se depara no seu dia a dia, um modelo de linguagem — que tenha o propósito geral de auxiliar os seres humanos nessas mesmas tarefas corriqueiras — precisa, também ele, apresentar essa plasticidade linguística, e utilizar-se da capacidade de reconhecer e produzir a variedade da linguagem apropriada para o contexto em que ela será usada, com funções, convenções e marcadores estéticos apropriados.

O leitor pode, entretanto, concordar conosco em nosso argumento sobre a importância de expor um modelo de linguagem a esses níveis de variação linguística, mas se perguntar se o pré-treinamento é a fase apropriada para tal. Afinal, o pré-treinamento é o momento em que o modelo deveria capturar as estruturas linguísticas gerais do idioma e, para realizar tarefas específicas — como por exemplo, redigir uma petição jurídica — ele passará, posteriormente, por um ajuste fino específico para essa tarefa.

De fato, para a realização de uma tarefa-alvo, os modelos precisam passar por essa etapa de ajuste fino com dados específicos para aquela tarefa, como discutimos nas seções anteriores. E, quando falamos de tarefas voltadas para contextos bastante particulares — como é o caso do exemplo mencionado da escrita da petição jurídica — é de se esperar que o ajuste fino também forneça um direcionamento ao modelo sobre o uso específico da linguagem para a realização daquela tarefa.

Contudo, essa etapa de ajuste fino dificilmente será suficiente para a adequação da linguagem, sem que tenha havido representação dessa forma de usar a língua durante o pré-treinamento. Isso porque, as diferenças entre tais variedades intralinguísticas podem ser bastante sutis e pervasivas.

Sardinha et al. (2014) realizaram uma análise da variação da língua portuguesa em nível de registro, sobre um conjunto diverso de propriedades gramaticais anotadas de forma semi-automática, baseados na metodologia de Biber (1988). Sua análise identificou que os registros no português variam ao longo de seis dimensões funcionais distintas, caracterizadas por propriedades como uso de pronomes, preposições e artigos, classes verbais, nominalização do sujeito, entre outras.

Já Serras et al. (2024) investigaram a diferenciabilidade de domínios discursivos do português brasileiro, realizando uma análise estatística individual sobre as distribuições de várias propriedades, incluindo propriedades morfológicas, sintáticas e semânticas. Eles observaram que, apesar de alto grau de sobreposição, várias das propriedades estudadas eram estatisticamente distintivas entre os diferentes domínios discursivos analisados.

Resultados como esses são testemunhos empíricos de que a língua portuguesa não é um corpo homogêneo, mas sim, um conjunto de manifestações distintas com funções e histórias específicas, que se sobrepõe em diferentes graus para constituir o que entendemos como o nosso idioma. Essa heterogeneidade é muitas vezes sutil, mas transversal a diferentes níveis de análise, estando entranhada em diversas propriedades que definem a estrutura gramatical da língua.

Se o pré-treinamento é a etapa em que o modelo cria uma representação interna da língua e da gramática que a rege, é sim fundamental que essa etapa inclua um nível de diversidade linguística apropriado. É isso que permitirá encapsular nessa representação os elementos das várias sublínguas que compõe o idioma a ser dominado, elementos esses

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol2/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

que estão entranhados em vários níveis da gramática da língua e que poderão ser posteriormente aproveitados pelo modelo na construção de habilidades específicas para contextos particulares.

2.7 A questão da proveniência dos dados

Proveniência, é o processo de rastreamento dos dados, ou seja, de suas origens e histórico de processamento[1]. Muitas vezes, em grandes *corpora*, é difícil de identificar de qual fonte um texto foi retirado, principalmente se os dados foram extraídos da web e processados em larga escala. Isso pode não só gerar um descumprimento acidental de direitos autorais dos textos originais, mas também uma dificuldade de manipulação desses dados por parte dos usuários. Ou seja, se por exemplo alguém não quiser utilizar textos de algum *website* específico ou tenha detectado um viés no modelo treinado, é mais difícil encontrar e retirar essas fontes indesejadas do conjunto textual caso não haja dados de proveniência marcados no *corpus*.

A questão dos direitos autorais e do respeito às licenças dos textos tem se tornado cada vez mais importante nos últimos anos, em especial depois da emergência de *chatbots* baseados em LLMs tão eficientes, como o ChatGPT, o Gemini e outras alternativas. Se antes os frutos desse desenvolvimento tecnológico tinham um papel mais restrito e indireto na vida das pessoas, eles passaram a estar presentes no dia-a-dia, gerar grandes receitas e ameaçar a substituição de muitos ofícios.

Essa possibilidade gerou um sentimento geral e justo de questionamento acerca dos impactos dessas tecnologias e da adequação da coleta de dados para seu desenvolvimento e treinamento. As questões envolvendo autoria de dados gerados por inteligência artificial e o uso adequado dos dados de treinamento invadiram tanto os tribunais como as mesas de bar.

É de se esperar que nos próximos anos, a sociedade encontre uma solução para esses debates. Até lá, o cuidado com a proveniência dos dados deve ser redobrado, tanto no sentido de garantir que os dados utilizados no treinamento tenham uma licença apropriada para a utilização a ser feita do modelo, como que sua fonte e licença sejam conhecidas, caso seja necessário remover esses dados no futuro, em função da evolução do nosso entendimento coletivo acerca de como tais dados podem ou não ser utilizados.

Um outro debate importante é o da licença dos próprios modelos produzidos a partir desses dados. Apesar de restrições diversas, os desenvolvedores de um modelo para fins de pesquisa usualmente têm a oportunidade de escolher a licença sob a qual o modelo será distribuído. Embora um debate nesse sentido esteja fora do escopo deste capítulo, recomendamos que os desenvolvedores de tais modelos pensem nessa questão com esmero, e sempre que possível, priorizem licenças de uso abertas que, ao mesmo tempo, visem coibir quaisquer usos nocivos.

Além disso, outra questão importante que vem à tona com o uso de textos da web para treinamento de modelos é o problema relacionado a dados pessoais e sensíveis e como tratá-los. Essa questão se tornou ainda mais relevante com o advento da Lei Geral de Proteção de Dados Pessoais (LGPD)[2], que busca justamente regulamentar o tratamento desses dados. Apesar de não conseguirmos abordar com profundidade essa discussão neste

[1] Tratamos com mais detalhes a respeito de proveniência, direitos autorais e métodos de extração de textos da web em (Crespo et al., 2023).

[2] O texto da lei pode ser conferido em: https://www.planalto.gov.br/ccivil_03/_ato2015-2018/2018/lei/l13709.htm

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol2/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

capítulo[3], vale destacar que a proveniência também se torna relevante nestas questões, pois sabendo de onde os dados vêm, é mais fácil identificar quais são os processos que precisam ser realizados para garantir a anonimização e remoção de dados pessoais dos textos.

2.8 O *corpus* Carolina como um estudo de caso

Nas seções anteriores, apresentamos diversos aspectos a serem considerados na construção de um conjunto de dados visando o pré-treinamento de modelos de linguagem. Gostaríamos de encerrar o capítulo fazendo uma breve apresentação do *corpus* Carolina. Esse é um *corpus* dual, construído visando tanto o treinamento de modelos de linguagem quanto a pesquisa linguística. Ele é uma tentativa destes autores e de seus colaboradores de conjugar vários dos aspectos discutidos neste capítulo na construção de um bom conjunto de dados. Essa não é uma empreitada fácil, já que muitas dessas restrições e objetivos competem entre si — vide como pode ser difícil garantir diversidade linguística e se ater a dados de licença aberta ao mesmo tempo — entretanto, acreditamos que o Carolina tem se tornado tanto um recurso útil por si só, como um protótipo para a construção de futuros recursos semelhantes.

O *corpus* Carolina[4] (Córpus Geral do Português Brasileiro com Informações de Procedência e Tipologia) é um *corpus* aberto, desenvolvido por uma equipe multidisciplinar de Linguistas e Cientistas da Computação, que faz parte do desafio de NLP2[5] do C4AI (Center for Artificial Intelligence)[6]. O Carolina está em construção contínua desde setembro de 2020, sendo lançadas anualmente novas versões com melhorias desde março de 2022, data da publicação da versão 1.0 (Ada). A versão atual, 2.0 (Bea), é composta por mais de 2 milhões de textos e cerca de 1,3 bilhões de *tokens*. Todos os textos do *corpus* foram extraídos da web e são posteriores a 1970.

O *corpus* Carolina foi criado com o intuito de preservar informações de proveniência e de tipologia textual. Para alcançar esse objetivo, apresentamos um cabeçalho com metadados anotados em lotes de textos[7], que conservam o máximo de informações que podemos obter de nossas fontes[8] e seguem o padrão TEI (*Text Encoding Initiative*) (TEI Consortium, 2021).

Assim, informações como a licença e URL da fonte são obrigatoriamente anotadas de modo a garantir a rastreabilidade dos textos originais, bem como a conformidade com os direitos autorais de cada um. Na Figura 2.1 a *tag* "<license>" contém as informações da licença da fonte e "<media>" a URL do texto.

Também anotamos informações das duas tipologias textuais do *corpus*, a Tipologia Ampla, ou Tipologia do Carolina, que organiza os textos dentro da estrutura do *corpus* e a Tipologia da Fonte, que contém os termos utilizados pelas próprias fontes para se referirem aos seus textos. Além disso, anotamos o Domínio de cada texto, que entendemos como a esfera em que circulam (por exemplo: Acadêmico, Pedagógico, Comercial etc.). Na

[3]Para uma discussão mais ampla a respeito da LGPD e a Inteligência Artificial cf. (Lazzarini Moretti; Maltese Zuffo, 2025).

[4]O nome do *corpus* é uma homenagem a Carolina Michaelis de Vasconcelos (1851-1925), filóloga e linguista alemã, radicada em Portugal, e primeira mulher a atuar como docente na Faculdade de Letras da Universidade de Lisboa, em 1911.

[5]https://dev-c4ai.webhostusp.sti.usp.br/?page_id=6120&lang=pt

[6]https://sites.google.com/icmc.usp.br/nlp2

[7]Para mais informações sobre a metodologia de extração e anotação dos dados do *corpus* Carolina cf. (Sturzeneker et al., 2022).

[8]A lista completa de metadados que anotamos encontra-se em (Crespo et al., 2023, p. 13–14).

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol2/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Figura 2.1: Alguns metadados contendo informações de proveniência textual.

```
<sourceDesc>
    <biblFull>
        <fileDesc>
            <titleStmt>
                <title>
                    <name>Discussão:Mutirão:Página principal</name>
                    <media mimeType="text/xml" url="https://br.wikimedia.org/wiki?curid=13" source="
                    brwikimedia-20210701-pages-meta-current.xml" />
                </title>
                <author />
                <editor role="translator" />
                <sponsor>Unknown</sponsor>
            </titleStmt>
            <extent>
                <measure unit="bytes" quantity="14471932" />
                <measure unit="tokens" quantity="1586470" />
                <measure unit="pages" quantity="-1" />
            </extent>
            <publicationStmt>
                <publisher />
                <authority>Wikimedia Foundation</authority>
                <date>2014-09-16</date>
                <availability status="free">
                    <license target="https://creativecommons.org/licenses/by-sa/3.0/deed.pt_BR">Creative Commons - Atribuição -
                    Compartilha Igual 3.0 Não Adaptada (CC BY-SA 3.0)</license>
                </availability>
```

Figura 2.2 as *tags* "<catRef>" contêm os tipos amplos e da fonte e "<domain>" apresenta o domínio. Esses metadados nos fornecem informações importantes acerca do registro dos textos que compõem o Carolina e podem ser usados para mensurar sua diversidade linguística.

Figura 2.2: Alguns metadados contendo informações de tipologia textual.

```
                <profileDesc>
                    <creation resp="user" />
                    <textDesc>
                        <channel mode="w" />
                        <constitution type="single" />
                        <derivation />
                        <domain>Virtual Forum</domain>
                        <factuality />
                        <interaction />
                        <preparedness type="monitored" />
                        <purpose />
                    </textDesc>
                    <textClass>
                        <catRef scheme="#Source_typology" target="#DISCUSSION_VIR_W" />
                    </textClass>
                    <langUsage>
                        <language ident="pt-BR" />
                    </langUsage>
                </profileDesc>
            </biblFull>
        </sourceDesc>
    </fileDesc>
    <profileDesc>
        <textClass>
            <catRef scheme="#Carolina_typology" target="#WIKIS" />
        </textClass>
    </profileDesc>
```

2.9 Considerações finais

Neste capítulo, apresentamos uma discussão acerca do papel dos dados no pré-treinamento de grandes modelos de linguagem. Discutimos as situações em que o pré-treinamento se

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol2/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

faz necessário, alguns dos principais conjuntos de dados disponíveis para o pré-treinamento em português e os principais aspectos que alguém deve considerar tanto ao selecionar um conjunto de dados, quanto ao construir um conjunto de dados para pré-treinamento.

Esperamos que os pontos apresentados sejam úteis aos leitores que estejam interessados em desbravar essa árdua tarefa do pré-treinamento dos grandes modelos de linguagem. Se em algum momento você obtiver um modelo cheio de alucinações, apesar de todo o seu esforço na seleção de dados, não perca a esperança: lembre-se que até Dom Quixote acabou são no final.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol2/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Capítulo 3

Treinamento de Grandes Modelos de Linguagem na prática

João Vitor Mariano Correia
João Renato Ribeiro Manesco
Danilo Samuel Jodas
Douglas Rodrigues
Gabriel Lino Garcia
Pedro Henrique Paiola
João Paulo Papa

Publicado em: 16/04/2026

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol2/

3.1 Introdução

Grandes Modelos de Linguagem (LLMs) representam uma mudança de paradigma no Processamento de Linguagem Natural, impulsionada pela arquitetura de *Transformers* (Vaswani et al., 2017b), que permitiu um avanço significativo na capacidade dos modelos de ponderar contexto. Trabalhos pioneiros como o BERT (Devlin et al., 2019), com sua abordagem de pré-treinamento bidirecional, validaram a eficácia de treinar modelos massivos em vastos conjuntos textuais, dando um passo fundamental para a compreensão de linguagem via redes neurais. O resultado é uma série de modelos com uma capacidade sem precedentes de processar e gerar texto de forma fluente e coerente.

Essa fundação de entendimento de linguagem é adquirida durante a fase de pré-treinamento, em que o modelo absorve conhecimento de fontes como a internet e grandes acervos de livros, permitindo que diversas aplicações de ampla relevância possam surgir, tais como sistemas já discutidos anteriormente nesta edição, como modelos de linguagem introduzidos no Capítulo 1, tradução automática (Capítulo Tradução Automática), sumarização automática (Capítulo Sumarização Automática), abrangendo diferentes contextos de uso (Minaee et al., 2024).

Apesar da eficácia dos LLM, sua aplicabilidade em cenários do mundo real depende criticamente de sua especialização, em especial tratando-se de modelos menores. A base generalista, embora ampla, é insuficiente para tarefas que exigem mais precisão ou conhecimento de nicho, tornando o processo de adaptação uma necessidade para determinados contextos (Anisuzzaman et al., 2025). Essa necessidade de adaptação pode ser analisada a partir de duas dimensões principais e complementares: a especialização para domínios de conhecimento e o alinhamento para contextos linguísticos e culturais específicos.

A primeira dimensão busca fazer com que o corpo de conhecimento dos modelos de linguagem seja composto por termos linguísticos e conhecimentos específicos de um campo nichado, facilitando sua aplicação em tarefas direcionadas, sendo crucial para aplicações

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

de alto impacto. No campo jurídico, por exemplo, um modelo refinado é capaz de interagir mais precisamente com o vernáculo próprio de documentos legais para realizar sumarizações ou revisões (Lai et al., 2024). Na medicina, a especialização permite que um modelo interprete literatura científica e registros de saúde para que seja capaz de apoiar decisões clínicas (Zhou et al., 2023b). Da mesma forma, no setor financeiro, um modelo ajustado pode identificar nuances em relatórios de mercado que seriam invisíveis a um sistema generalista (Lee et al., 2024). Nesses contextos, o refinamento é o que distingue uma ferramenta funcional de um sistema verdadeiramente confiável e eficaz.

A segunda dimensão busca realizar a adaptação linguística e cultural dos modelos, um desafio fundamental na construção de modelos equitativos. A grande maioria dos dados utilizados no pré-treinamento de modelos publicados abertamente está em inglês, o que cria um desequilíbrio cultural e representativo a nível de linguagem nesses modelos (AlKhamissi et al., 2024). Desta forma, modelos pouco expostos à língua portuguesa sofrem dificuldade em entender certas nuances linguísticas, bem como podem estar expostos a certos vieses culturais dos dados, como é o caso de modelos médicos que ignoram a possibilidade de diagnóstico de certas doenças endêmicas como a dengue pelos dados de treinamento serem de locais não expostos a este problema (Paiola et al., 2024).

Para que a tecnologia seja efetivamente democrática e útil em contextos não anglófonos, como o brasileiro, a tradução da resposta nem sempre é a solução. Exigindo um processo de ajuste fino ou continuação de pré-treino com dados no idioma local. Garantindo que o modelo compreenda as particularidades culturais, expressões idiomáticas e o contexto social do idioma refinado.

Atingir um alto nível de robustez com LLMs, no entanto, requer um processo que vai além do treinamento em escala. É necessária uma estratégia de adaptação cuidadosa, que considere tanto a profundidade técnica de um domínio quanto a riqueza de um contexto linguístico. Para abordar essa necessidade, este capítulo irá explorar o espectro de técnicas de especialização utilizadas no treinamento de modelos de linguagem. Iniciaremos discutindo a continuação de pré-treino, seguiremos para o ajuste fino supervisionado em suas diversas formas, incluindo técnicas de parametrização eficiente e aprendizado *few-shot*, e concluiremos com métodos de alinhamento, que refinam o comportamento do modelo com base em preferências humanas. Ao final do capítulo, apresentamos uma implementação prática do refinamento de modelos.

3.2 Continuação de pré-treino

A Continuação de Pré-treino (CPT) consiste em uma técnica de adaptação realizada antes do ajuste fino supervisionado, visando ampliar o conhecimento de um LLM em um domínio específico. Essa etapa intermediária entre o pré-treinamento generalista e o *fine-tuning* retoma o treinamento original do modelo, de forma autossupervisionada, através de técnicas como a Modelagem de Linguagem Mascarada (MLM) ou Causal (CLM), aplicadas em um *corpus* especializado de grande escala. Esta estratégia resulta na incorporação de vocabulário, conceitos e fatos do domínio desejado. Neste capítulo serão apresentados os fundamentos técnicos, cenários de aplicação e distinções relevantes dessa abordagem.

3.2.1 Mecanismos de treinamento autossupervisionado

Para modelos autorregressivos, como os das famílias GPT e Qwen, a técnica predominante é a de Modelagem de Linguagem Causal (CLM) (Bai et al., 2023; Radford et al., 2019b),

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol2/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

cujo objetivo é treinar o modelo para prever o próximo *token* em uma sequência com base nos *tokens* anteriores. De forma mais precisa, a probabilidade de uma sequência de *tokens* é decomposta em uma cadeia de probabilidades condicionais, onde cada *token* é previsto condicionando-se nos *tokens* que o precedem.

O objetivo da CLM é maximizar essa probabilidade nos textos observados, o que equivale a minimizar a soma dos erros associados a cada previsão individual, ou seja, a entropia cruzada entre as distribuições preditas e os *tokens* reais. A natureza estritamente unidirecional, da esquerda para a direita, do CLM flui naturalmente com a arquitetura do tipo *decoder* em arquiteturas de *transformers*, que empregam máscaras nos mecanismos de atenção. Essas máscaras garantem que, ao prever um *token* em uma determinada posição, o modelo utilize apenas informações dos *tokens* que o antecedem na sequência, e não dos posteriores.

Em contrapartida, arquiteturas bidirecionais como o BERT utilizam a Modelagem de Linguagem Mascarada (MLM) (Devlin et al., 2019), cujo funcionamento, ao invés de prever o próximo *token* de uma sequência, trabalha corrompendo uma fração dos *tokens* no texto de entrada, criando uma espécie da "máscara" no texto, desta forma, o modelo deve prever quais os *tokens* originais que deveriam compor a sequência. Nesse processo, para uma sequência de entrada T, um subconjunto de posições M é selecionado para mascaramento. Para cada posição $i \in M$, um *token* t_i pode ser substituído pelo símbolo de máscara, por um *token* aleatório ou mantido inalterado, de modo a incentivar o modelo a aprender representações contextuais mais robustas. A função de perda é calculada apenas sobre os *tokens* mascarados, já que o modelo se concentra em predizer esses elementos, e como a predição é condicionada à sequência inteira corrompida, o modelo aprende representações contextuais, considerando tanto o contexto à esquerda quanto à direita.

Outra estratégia, chamada de Previsão da Próxima Sentença (NSP), introduzida no pré-treinamento do BERT, consiste em classificar se uma sentença sucede outra no texto. Diferente da Modelagem CLM, que realiza predição *token* a *token* de forma unidirecional, o NSP busca classificar sentenças completas. Essa técnica, no entanto, caiu em desuso no estado da arte, devido a estudos indicarem benefício limitado, e foi descartada em modelos recentes (Liu et al., 2019b).

3.2.2 Contextos e objetivos da continuação de pré-treino

A aplicação da continuação de pré-treino é uma decisão estratégica que visa aprimorar o modelo de maneiras que o ajuste fino supervisionado por si só não consegue alcançar. O principal desses cenários é a adaptação para aplicações específicas, onde um LLM generalista é infundido com o conhecimento de um campo técnico (Shi et al., 2024). Um exemplo prático seria treinar um LLM com uma biblioteca de artigos jurídicos, de tal forma que ele se torne fluente no jargão e nos conceitos do direito, ou até mesmo em conhecimento médico, como será trabalhado ao longo deste capítulo. Outro uso importante é a adaptação temporal, que atualiza o conhecimento do modelo para além de sua data de corte, utilizando um conjunto de textos e notícias recentes. Adicionalmente, a CPT é útil para adaptações de idioma, melhorando a proficiência de um modelo em uma língua menos representada nos dados de treinamento originais, como o português, aprimorando sua fluência e compreensão de nuances culturais.

É importante reforçar a distinção entre a CPT e o Ajuste Fino Supervisionado, já que a CPT tem o objetivo de expandir a base de conhecimento geral de um modelo, através da utilização de dados não estruturados. Já o ajuste fino busca impor um comportamento

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol2/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

específico ao modelo, utilizando dados rotulados preparados, para uma ou mais tarefas específicas. A CPT opera sobre conjuntos massivos de texto bruto, o resultado é um modelo com vocabulário e conhecimento de domínio aprofundados, mas não necessariamente com os detalhes específicos das tarefas em que deve atuar.

3.3 Ajuste fino de modelos

Conforme explorado na seção anterior, na fase de pré-treinamento os Grandes Modelos de Linguagem adquirem um vasto conhecimento de mundo e uma profunda compreensão de padrões linguísticos, adquiridos a partir da exposição a *corpora* textuais massivos. Essa base de conhecimento confere aos modelos a capacidade de realizar uma gama de tarefas de forma emergente, sem treinamento explícito. Contudo, para alcançar excelência e confiabilidade em aplicações especializadas, é necessário um processo de adaptação direcionado. Este processo é conhecido como ajuste fino (do inglês, *fine-tuning*).

O ajuste fino representa uma importante mudança de paradigma na história do Processamento de Linguagem Natural. Ele pode ser definido como a especialização de um modelo generalista, transformando-o em uma ferramenta especialista capaz de executar tarefas com alta precisão e adaptar-se a domínios de conhecimento específicos (Wu et al., 2025). A intuição por trás desta técnica pode ser ilustrada por uma analogia com a formação médica: o pré-treinamento é análogo aos anos de faculdade, onde se adquire um conhecimento abrangente. O ajuste fino, por sua vez, equivale à residência médica, na qual o profissional aprofunda seu conhecimento em uma área específica, como a cardiologia, para diagnosticar e tratar com uma proficiência que um clínico geral não possui. De forma similar, o ajuste fino não ensina o LLM a gerar linguagem do zero; ele refina suas competências preexistentes para uma nova aplicação.

Formalmente, a técnica se fundamenta no princípio da aprendizagem por transferência (do inglês, *transfer learning*), no qual um modelo desenvolvido para uma tarefa inicial é reutilizado como ponto de partida para treinar outro modelo em uma segunda tarefa, relacionada à primeira. O conhecimento capturado durante o pré-treinamento em um domínio de origem e tarefa de origem é transferido para melhorar o desempenho e acelerar a aprendizagem em um domínio de destino e uma tarefa de destino. Essa abordagem mostrou-se fundamental para superar um dos maiores desafios históricos da área: a necessidade de vastos conjuntos de dados rotulados e o custo computacional proibitivo do treinamento de modelos de ponta a partir do zero (Howard; Ruder, 2018b).

O principal método para essa especialização é o Ajuste Fino Supervisionado (do inglês, *Supervised Fine-Tuning* ou SFT) (Devlin et al., 2019). Neste processo, o modelo pré-treinado tem seu treinamento estendido, mas desta vez em um conjunto de dados menor e de alta qualidade, composto por exemplos rotulados que demonstram o comportamento desejado — geralmente em um formato de pares de entrada e saída (por exemplo, no caso da tarefa de *Question Answering* (QA), o par `pergunta -> resposta esperada`). O mecanismo é uma continuação direta do treinamento de redes neurais:

1. **Seleção do modelo base**: A escolha de um modelo pré-treinado apropriado é o primeiro passo. A decisão deve levar em conta a arquitetura, o tamanho e a afinidade do domínio de pré-treinamento com a tarefa final.

2. **Preparação do *dataset***: A qualidade do ajuste é diretamente proporcional à qualidade do conjunto de dados. Esta etapa envolve a coleta, limpeza e formatação de

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol2/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

dados rotulados e representativos da tarefa, geralmente estruturados em pares de *prompt*-resposta.

3. **Ajuste de hiperparâmetros**: A seleção de hiperparâmetros como taxa de aprendizado, tamanho do lote (do inglês, *batch size*) e número de épocas é crucial. Um ajuste inadequado pode levar a um treinamento instável ou a um sobreajuste.

4. **Treinamento e validação**: O modelo é treinado em lotes de dados, e seu desempenho é continuamente monitorado em um conjunto de validação separado. O monitoramento da perda de validação é essencial para detectar sobreajuste e aplicar estratégias como a parada antecipada, que interrompe o treinamento quando a generalização para dados não vistos começa a piorar.

Este processo iterativo refina gradualmente o conhecimento do modelo, especializando-o para as nuances da tarefa alvo. Neste ponto, é crucial diferenciar o ajuste fino de outras técnicas de adaptação:

- **Continuação de pré-treino**: Enquanto a continuação de pré-treino visa expandir a base de conhecimento geral do modelo em um novo domínio (ex: alimentar um LLM com uma biblioteca de artigos jurídicos), utilizando dados não estruturados, o ajuste fino visa ensinar uma habilidade ou comportamento específico (ex: ensinar o modelo a sumarizar processos judiciais), utilizando dados rotulados.

- **Engenharia de *prompts***: Esta técnica não modifica nenhum dos parâmetros do modelo. Em vez disso, ela guia o comportamento do modelo no momento da inferência, formulando cuidadosamente a entrada (*prompt*) para induzir a resposta desejada. Métodos como *one-shot* ou *few-shot learning*, onde um ou alguns exemplos da tarefa são incluídos no *prompt*, são formas de engenharia de *prompts*. É uma forma de adaptação sem treinamento, que depende inteiramente das capacidades de aprendizado no contexto do modelo base.

Em suma, o ajuste fino representa um caminho poderoso para a especialização de LLMs, capacitando-os a ir além da fluência geral para entregar desempenho de ponta em aplicações de nicho. Contudo, a estratégia pela qual essa adaptação de parâmetros é realizada varia significativamente, levando a um espectro de métodos com diferentes *trade-offs* entre desempenho, custo computacional e complexidade. Nas subseções a seguir, exploraremos as principais abordagens práticas: desde o ajuste fino completo, que modifica todos os parâmetros do modelo, passando pelas eficientes técnicas de PEFT (do inglês, *Parameter-Efficient Fine-Tuning*), que ajustam apenas um pequeno subconjunto de parâmetros, até as estratégias de *Few-Shot Learning*, que operam na fronteira entre o ajuste e a engenharia de *prompts*. As seções seguintes exploram técnicas avançadas como *Reinforcement Learning from Human Feedback* (RLHF) e *Direct Preference Optimization* (DPO), que estendem o SFT para alinhamento com preferências humanas.

3.3.1 Ajuste fino completo

O ajuste fino completo (do inglês, *Full Fine-Tuning* ou FFT) representa a forma mais direta e conceitualmente mais simples de especialização de um modelo de linguagem. Nesta abordagem, a totalidade dos parâmetros do modelo pré-treinado é descongelada e se torna treinável. Isso concede ao modelo máxima flexibilidade para que todo o seu conhecimento

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol2/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

generalista, encapsulado em bilhões de parâmetros, seja reconfigurado para se adaptar profundamente aos padrões do novo conjunto de dados (Lv et al., 2024; Raffel et al., 2020b).

A principal vantagem desta abordagem é seu potencial para alcançar o máximo desempenho possível na tarefa alvo. Ao permitir que todas as camadas da rede neural sejam atualizadas, desde as representações iniciais de *embeddings* até os mecanismos de atenção e as camadas *feed-forward*, o modelo pode aprender nuances extremamente complexas e específicas do domínio, algo que abordagens mais restritivas poderiam não capturar. No entanto, este potencial de desempenho vem acompanhado de custos e desafios significativos, que devem ser cuidadosamente ponderados:

- **Alto custo computacional e de memória**: O FFT é extremamente exigente em termos de recursos. Treinar todos os parâmetros de um modelo com bilhões de pesos requer acesso a múltiplas GPUs de ponta com alta capacidade de VRAM, não apenas para armazenar os pesos do modelo, mas também os gradientes, os estados do otimizador e as ativações intermediárias durante o treinamento. Isso torna o processo financeiramente proibitivo para muitas organizações.
- **Custo de armazenamento e implantação**: Como o FFT modifica todos os pesos, cada nova tarefa especializada resulta em uma cópia completa e independente do modelo, que pode ocupar dezenas ou centenas de gigabytes. Gerenciar, armazenar e servir múltiplas instâncias de modelos tão grandes em um ambiente de produção é logisticamente complexo e caro.
- **Esquecimento catastrófico**: Este é um dos riscos mais proeminentes. Ao se especializar intensamente na nova tarefa, o modelo pode sobrescrever e perder drasticamente as capacidades gerais adquiridas durante o pré-treinamento (Howard; Ruder, 2018b). Por exemplo, um modelo ajustado para gerar código em Python pode esquecer como manter um diálogo coerente sobre história.
- **Risco de sobreajuste**: Especialmente quando o conjunto de dados de ajuste fino é pequeno ou pouco diverso, o modelo pode memorizar os exemplos de treinamento em vez de aprender a generalizar o padrão subjacente. Isso resulta em um excelente desempenho nos dados de validação, mas um baixo desempenho em dados novos e não vistos.
- **Amplificação de viés**: Se o conjunto de dados utilizado para o ajuste fino contiver vieses sociais ou demográficos, o processo de FFT pode não apenas replicá-los, mas também reforçá-los, resultando em um modelo final ainda mais enviesado do que sua versão pré-treinada. Por exemplo, um modelo pode ser induzido a associar automaticamente certas profissões a um gênero, reforçando estereótipos sociais.

Apesar de seus desafios, o FFT permanece uma ferramenta relevante e poderosa para cenários específicos. A decisão de utilizá-lo deve ser pragmática, considerando as seguintes condições:

1. **Quando o desempenho máximo é crítico**: Em aplicações de alto risco, como diagnóstico médico ou análise financeira, onde cada ponto percentual de acurácia é crucial e o orçamento permite o investimento.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol2/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

2. **Quando há uma mudança substancial de domínio**: Se a tarefa alvo for radicalmente diferente dos dados de pré-treinamento, uma adaptação profunda pode ser necessária.

3. **Quando recursos e dados são abundantes**: O FFT só deve ser considerado quando se tem acesso a recursos computacionais massivos e, fundamentalmente, a um conjunto de dados de ajuste fino que seja grande, diverso e de altíssima qualidade para mitigar os riscos.

Em conclusão, o ajuste fino completo é uma faca de dois gumes. Ele oferece o caminho para o mais alto nível de especialização, mas a um custo extremo e com riscos significativos de degradar propriedades valiosas do modelo original. Para a grande maioria dos casos de uso práticos, o balanço entre desempenho, custo e risco pende favoravelmente para abordagens mais eficientes. Essas desvantagens impulsionaram a pesquisa e o desenvolvimento de uma nova família de técnicas, conhecidas como ajuste eficiente de parâmetros (do inglês, *Parameter-Efficient Fine-Tuning* ou PEFT).

3.3.2 PEFT

O desenvolvimento de metodologias voltadas à redução do custo computacional associado a LLMs tornou-se um requisito crítico para sua adaptação a tarefas especializadas. Essa demanda torna-se relevante no contexto do *fine-tuning* sobre conjuntos de dados de domínio específico, onde o custo computacional impacta diretamente a viabilidade prática da aplicação. Nesse cenário, busca-se empregar estratégias de treinamento que priorizem a atualização seletiva de camadas ou módulos do modelo, ao mesmo tempo minimizando o número de parâmetros ajustados. Tal abordagem tem o objetivo de não apenas reduzir o consumo de recursos e o tempo de processamento, mas também preservar, ou até aprimorar, a acurácia de predição visando garantir que a especialização do modelo não comprometa seu desempenho geral.

Neste contexto, surge uma abordagem denominada *Parameter-Efficient Fine-Tuning* (PEFT) (Ding et al., 2023), cujo objetivo é oferecer um conjunto de técnicas para adaptar grandes modelos de linguagem previamente pré-treinados, como o LLaMA (Gajulamandyam et al., 2025) ou BERT (Li, 2024), para tarefas mais específicas. Em vez de re-treinar integralmente o modelo, que pode conter bilhões de parâmetros e exigir recursos computacionais significativos, o PEFT se concentra em ajustar apenas uma pequena fração desses parâmetros. O objetivo primário consiste em manter o conhecimento geral que o modelo adquiriu durante seu treinamento inicial. O PEFT, então, age como uma "camada extra", adicionando um pequeno número de novos parâmetros treináveis. Esses novos parâmetros são ajustados para a tarefa específica, enquanto a maioria do modelo original permanece "congelada".

A motivação fundamental por trás do PEFT é a observação de que, durante o *fine-tuning*, nem todos os parâmetros de um modelo precisam ser atualizados para alcançar a melhor adaptação. Modelos pré-treinados são compostos por representações que generalizam o conhecimento linguístico. Neste sentido, apenas pequenos ajustes direcionais são necessários para especializar o modelo em domínios ou tarefas específicas. Este conceito levou ao desenvolvimento de técnicas que preservam o conhecimento original do modelo enquanto introduzem capacidades adaptativas através de estruturas paramétricas eficientes.

Técnicas de PEFT têm demonstrado melhorias significativas no que tange à eficácia dos novos modelos desenvolvidos, tanto do ponto de análise quantitativa quanto da avaliação

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol2/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

qualitativa dos resultados. Estes avanços indicam a consolidação de tais técnicas no contexto de adaptação de modelos em diversos cenários, permitindo uma redução que abrange parcelas minúsculas dos parâmetros do modelo original, enquanto mantém performance similar com o *fine-tuning* completo. Li (Li, 2024), por exemplo, apresentou um refinamento seletivo de 0,2% dos parâmetros de um modelo BERT usando uma abordagem baseada na atualização dos parâmetros de *bias* e normalização, demonstrando resultados competitivos com técnicas consolidadas na literatura.

3.3.2.1 Low-Ranking Adaptation

No contexto do PEFT, a técnica de *Low-Rank Adaptation* (LoRA) (Hu et al., 2021) emergiu como uma das mais conhecidas e amplamente utilizadas. Seu funcionamento consiste em aplicar uma decomposição matemática para atualizar os pesos durante o processo de *fine-tuning*. Quando o *fine-tuning* tradicional é aplicado, cada matriz de peso do modelo original, W_0, é atualizada para $W_0 + \Delta W$. No entanto, o LoRA reformula essa atualização ΔW como o produto de duas matrizes menores:

$$\Delta W = BA, \tag{3.1}$$

onde B e A são matrizes de baixo *rank*, onde B possui dimensões $d \times r$ e A possui dimensões $r \times k$, sendo d a dimensão da matriz original de pesos. Esta abordagem é fundamentada no princípio de que as mudanças necessárias para adaptar um modelo a uma nova tarefa podem ser representadas de forma eficiente em um espaço de menor dimensão.

Durante o processo de treinamento com LoRA, as matrizes de peso originais, W_0, são mantidas completamente congeladas para preservar o conhecimento adquirido durante o pré-treinamento. As novas matrizes A e B são inicializadas de forma que seu produto BA seja inicialmente zero. Isso assegura que o modelo, no início do treinamento, seja idêntico ao modelo pré-treinado original. Conforme o treinamento avança, as matrizes A e B são as únicas que aprendem as adaptações necessárias para a tarefa específica. A saída de cada camada é então calculada como $h = W_0x + BAx$, onde o primeiro termo mantém o comportamento original do modelo e o segundo adiciona a especialização aprendida.

O parâmetro de *rank*, r, é um hiperparâmetro essencial no LoRA, que governa o equilíbrio entre eficiência e capacidade expressiva do modelo. *Ranks* menores resultam em maior eficiência computacional e menor uso de memória, mas podem limitar a capacidade de adaptação do modelo. Por outro lado, *ranks* maiores aumentam a flexibilidade, mas também elevam o custo computacional.

3.3.2.2 Quantized Low-Rank Adaptation

A combinação de quantização e LoRA é uma técnica avançada e extremamente eficiente para otimizar o ajuste fino de LLMs. A quantização é o processo de converter os pesos e ativações de uma rede neural de um formato de alta precisão (como ponto flutuante de 32 bits - **float32**) para um formato de baixa precisão (como inteiros de 4 ou 8 bits - **int8** ou **int32**). A principal motivação para isso é a economia de memória, já que um **float32** ocupa 4 bytes, enquanto um **int8** ocupa apenas 1 byte. Apesar de a diferença de memória na quantização de um único parâmetro ser pequena, ela se torna muito significativa quando aplicada a modelos de bilhões de parâmetros. Além de economizar memória, a quantização também acelera os cálculos, pois processadores e GPUs podem realizar operações com inteiros de forma muito mais rápida e eficiente. Contudo, o principal desafio

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol2/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

dessa abordagem é a possível perda de precisão, que pode degradar o desempenho do modelo se não for tratada com cautela.

A técnica de *Quantization* LoRA (Dettmers et al., 2023), denominada QLoRA, fornece uma solução para o desafio da perda de precisão ao aplicar a quantização seletivamente. O processo é dividido basicamente em duas etapas. Primeiro, o modelo pré-treinado original é quantizado para um formato de baixa precisão, usando uma técnica denominada **NormalFloat** de 4 bits. Todos os parâmetros são convertidos e armazenados em menos bits, reduzindo o modelo de dezenas de *gigabytes* para apenas alguns.

Em seguida, o processo de *fine-tuning* é realizado apenas nas matrizes de baixo *rank* do LoRA, que são mantidas em um formato de alta precisão, ou seja, em f*loats* de 16 ou 32 bits. Esses poucos parâmetros em alta precisão representam o ponto chave para evitar a degradação de desempenho, pois o otimizador precisa de precisão para calcular os gradientes e atualizar os pesos de forma eficaz. Se as matrizes LoRA fossem quantizadas, o processo de aprendizado seria instável e ineficaz.

Mais especificamente, o QLoRA é baseado nas seguintes técnicas:

1. **NormalFloat 4 (NF4)**: A maioria dos pesos em uma rede neural profunda segue uma distribuição próxima de uma gaussiana ou normal, onde a maioria dos valores se concentra perto de zero. O NF4 define 16 pontos de quantização (já que $2^4 = 16$) de forma não uniforme, alocando mais pontos de precisão para os valores próximos de zero. Isso significa que os valores mais comuns e críticos do modelo são representados com maior precisão, enquanto os valores extremos, que são menos frequentes, são representados com menor precisão. O resultado é uma perda de informação muito menor em comparação com a quantização uniforme.

2. ***Double Quantization***: Executa a quantização das constantes de quantização NF4 que auxiliam na reconstrução dos valores originais. Essa segunda etapa de quantização visa aplicar uma segunda redução do consumo de memória da rede, aplicando uma quantificação de ponto flutuante de 32 bits dessas constantes para 8 bits.

3. ***Paged Optimizers***: é uma técnica de gerenciamento de memória cuja função é evitar os picos de uso de VRAM (memória da GPU) que ocorrem durante o treinamento, especialmente para armazenar parâmetros como gradiente, o momento do gradiente e a variância. Para solucionar esse problema, os *Paged Optimizers* alocam o estado do otimizador na memória da CPU. Quando a GPU precisa de espaço específico desses dados para o cálculo de gradientes, ela o aloca na VRAM. Após isso, esse espaço é realocado de volta para a CPU. Essa troca dinâmica entre a memória da CPU e da GPU garante que o treinamento continue de forma estável, sem ultrapassar o limite de VRAM.

Na prática, durante o treinamento, o fluxo de dados em uma camada com LoRA quantizado funciona da seguinte maneira: a entrada de um *token*, $\boldsymbol{x}$, é multiplicada pela matriz de peso quantizada do modelo base, W_0^{int8}. Em paralelo, essa mesma entrada é multiplicada pelas matrizes de alta precisão do LoRA, A e B. O resultado do cálculo com LoRA, isto é, BAx, é então somado ao resultado do cálculo com o modelo base (W_0^{int8}), e tudo é feito em alta precisão antes de ser passado para a próxima camada.

A principal vantagem reside no fato de que a computação de maior custo, que envolve a maior parte dos parâmetros (W_0^{int8}), é feito de forma rápida e com pouca memória. O cálculo com os parâmetros do LoRA, que é a parte que realmente "aprende', é mantido

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol2/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

em alta precisão para garantir a qualidade do *fine-tuning* e evitar perda de precisão no cálculo dos gradientes. Como resultado, a combinação de *quantization* e LoRA preserva o desempenho do modelo, mas de maneira mais eficiente no que concerne ao uso de memória.

3.3.2.3 Módulos adapters

Outra técnica que se destaca no contexto de PEFT são os módulos *Adapters*, propostos por Houlsby et al. (Houlsby et al., 2019), cuja estrutura é formada por pequenos módulos de rede neural que são incorporados nas camadas de um modelo *Transformer* pré-treinado. O princípio consiste em manter fixos os parâmetros do modelo base e treinar apenas os parâmetros desses novos módulos adicionados. Essa abordagem permite que o modelo original preserve o conhecimento já adquirido durante o pré-treinamento, enquanto os *Adapters* aprendem as adaptações específicas para a nova tarefa durante o ajuste fino do modelo.

A arquitetura de um módulo *Adapter*, em sua forma original, segue o padrão de um gargalo (*bottleneck*). Essa estrutura consiste em:

1. ***Down-projection*:** A entrada de uma camada do modelo original (com dimensão d) é projetada para uma dimensão menor, m, usando uma matriz de peso W_{down}. Isso cria um "gargalo" na arquitetura para comprimir a representação do dado.

2. ***Up-projection*:** A representação comprimida (h_{down}) é então projetada de volta para a dimensão original d usando uma segunda matriz de peso W_{up}, dando origem à uma representação de dimensão original h_{up}.

O valor da dimensão m é um hiperparâmetro, e geralmente é menor que d. Isso faz com que o número de parâmetros nos *Adapters* seja menor que o número de parâmetros do modelo original.

Para garantir a estabilidade do treinamento, é comum que a saída do módulo *Adapter* seja somada à entrada da camada, seguindo uma conexão residual (*skip-connection*). A saída final da camada com o *Adapter* é então representada por:

$$h_{final} = \mathbf{x} + h_{up}, \tag{3.2}$$

onde $\mathbf{x}$ é a entrada original do *Adapter*. A matriz W_{up} é geralmente inicializada com valores próximos de zero, para que o módulo *Adapter* comece com uma função de identidade. Isso garante que, no início do treinamento, a saída do *Adapter* seja zero e o comportamento do modelo seja idêntico ao original.

Em suma, as técnicas de PEFT consolidaram-se no contexto de ajuste fino de grandes modelos de linguagens ao oferecerem uma alternativa com a introdução de módulos com pouco impacto em suas arquiteturas originais. Entretanto, tais técnicas possuem suas próprias limitações, sendo a principal delas a de não poder replicar completamente o desempenho do *fine-tuning* tradicional em todos os cenários. Em tarefas que exigem uma adaptação mais complexa e mudanças significativas no modelo, o ajuste completo de todos os parâmetros ainda pode ser a abordagem mais eficaz.

3.3.3 *Few-shot learning*

Complementar às estratégias de ajuste fino completo e às técnicas de PEFT, as abordagens de aprendizagem com poucas amostras (do inglês, *few-shot learning*) buscam maximizar

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol2/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

ainda mais a eficiência computacional, reduzindo a quantidade de exemplos necessários para a especialização de um modelo. No campo do Processamento de Linguagem Natural, este termo abrange duas metodologias distintas para adaptar modelos a novas tarefas, nomeadamente *In-Context Learning* e Ajuste Fino com Poucas Amostras. Embora ambas partam da premissa de utilizar uma quantidade mínima de exemplos, elas diferem fundamentalmente no mecanismo de aprendizado, onde uma ocorre sem a atualização dos pesos do modelo, enquanto a outra envolve um ajuste fino deliberado.

3.3.3.1 Aprendizagem no contexto (*In-Context Learning*)

A primeira abordagem, e a que mais popularizou o termo no contexto de PLN. Este método não envolve uma fase de treinamento com atualização de gradientes. Em vez disso, o modelo é condicionado a realizar uma nova tarefa no momento da inferência. O aprendizado é transitório e guiado pelo *prompt* de entrada, que contém uma descrição da tarefa e um pequeno número de exemplos ilustrativos (os *shots*).

O trabalho de demonstrou que modelos de grande escala possuem capacidades emergentes de generalização a partir de poucos exemplos, viabilizando a aprendizagem no contexto de forma eficaz. Os autores avaliam três configurações:

- ***Zero-shot learning*:** O modelo recebe apenas a descrição em linguagem natural da tarefa, sem nenhum exemplo de como executá-la.
- ***One-shot learning*:** O modelo recebe a descrição da tarefa acompanhada de um único exemplo.
- ***Few-shot learning*:** O modelo recebe a descrição da tarefa e múltiplos exemplos (tipicamente entre 2 e 32).

A eficácia desta técnica está intrinsecamente ligada à qualidade e ao formato do *prompt*. A engenharia de *prompts* torna-se uma disciplina crucial, pois a maneira como os exemplos são apresentados pode influenciar drasticamente a precisão e a coerência da saída do modelo. A seleção de exemplos diversificados e representativos da tarefa em questão é fundamental para guiar o modelo de forma adequada.

A principal vantagem desta abordagem reside na sua eficiência computacional e na economia de recursos. Como não há treinamento, os custos associados à esta etapa são completamente eliminados. No entanto, a performance embora muitas vezes competitiva, pode não alcançar a mesma robustez de um modelo que passou pelo processo de atualização de parâmetros, especialmente em tarefas altamente especializadas ou que requerem conhecimento de um domínio muito específico. Além disso, a janela de contexto do modelo impõe um limite físico à quantidade de exemplos que podem ser fornecidos, e o custo da inferência pode aumentar.

3.3.3.2 Ajuste fino com poucas amostras

A segunda abordagem alinha-se mais com a definição clássica de *few-shot learning* em outras áreas de IA. Ela consiste em realizar o ajuste fino supervisionado do modelo utilizando um conjunto de dados deliberadamente pequeno, mas de alta qualidade. A hipótese central é que, para modelos que já possuem um vasto conhecimento de mundo adquirido durante o pré-treinamento, não é a quantidade de dados de ajuste fino que importa, mas sim a sua qualidade.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol2/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Um exemplo proeminente desta técnica é apresentado no trabalho de (Ye et al., 2025). Os autores exploram a premissa de que o raciocínio complexo exige volumes massivos de dados de treinamento. Eles propõem a hipótese LIMO (do inglês, *Less-Is-More Reasoning Hypothesis*) que postula que, em modelos onde o conhecimento de um domínio já foi extensivamente codificado o raciocínio sofisticado pode emergir a partir de demonstrações mínimas, mas estrategicamente desenhadas, de processos cognitivos. Segundo a hipótese, existem dois fatores principais que determinam o ponto de ativação para o raciocínio complexo: (1) a presença de conhecimento prévio dentro dos parâmetros do modelo, adquirido durante o pré-treinamento e (2) a efetividade de poucos exemplos em demonstrar processos de solução de problemas que estimulem uma reflexão mais extensa.

Para validar essa hipótese, os autores desenvolvem um rigoroso processo de curadoria de dados. Partindo de um conjunto com dezenas de milhões de problemas matemáticos, eles aplicam um sistema de filtragem em múltiplos estágios para selecionar apenas 800 amostras de treinamento de alta qualidade. Os critérios de seleção priorizaram problemas desafiadores e cadeias de raciocínio que exibiam características como elaboração detalhada, autoverificação e clareza lógica. Utilizando apenas 800 amostras para realizar um ajuste fino supervisionado simples o modelo alcançou resultados notáveis e amplamente superiores a outros modelos, tanto dentro quanto fora do domínio, confirmando que, para certas tarefas, a curadoria meticulosa de um pequeno conjunto de dados é mais crucial do que a quantidade massiva, permitindo o desenvolvimento de capacidades de raciocínio robustas de forma muito mais eficiente

Em suma, as duas vertentes do *few-shot learning* oferecem um contraponto claro em termos de mecanismo e objetivo. A aprendizagem no contexto é uma técnica de inferência: ela não altera os pesos do modelo e depende da qualidade do *prompt* para guiar o comportamento do modelo de forma temporária. Por outro lado, o ajuste fino com poucas amostras é uma técnica de treinamento que modifica permanentemente o modelo. Porém, seu sucesso não está no volume, mas na qualidade e na densidade informacional dos dados. O investimento é deslocado da coleta massiva para a curadoria meticulosa, com o objetivo de alcançar um desempenho robusto e de ponta em domínios específicos.

3.4 *AI alignment*

O alinhamento de sistemas de IA aos valores humanos (do inglês, *AI Alignment*) refere-se ao desafio de garantir que agentes artificiais otimizem objetivos que estejam em conformidade com as intenções humanas, mesmo em contextos complexos, ambíguos ou não totalmente especificados. Esse problema é frequentemente enquadrado no contexto mais amplo do *controle de sistemas inteligentes*, conforme discutido por Russell (Russel, 2019) e Bostrom (Bostrom, 2014), e está no cerne das preocupações de segurança da IA contemporânea (Hendrycks et al., 2021; Ngo et al., 2023).

À medida que LLMs, agentes de decisão e sistemas autônomos adquirem maior expressividade e poder de generalização, torna-se cada vez mais crítico mitigar riscos decorrentes de comportamentos inesperados, otimizações mal direcionadas ou divergências entre metas explícitas e consequências práticas. Trabalhos como Gabriel (Gabriel, 2020) destacam que o alinhamento não se restringe a uma questão técnica, mas envolve também aspectos normativos e éticos sobre quais valores humanos devem orientar a otimização algorítmica.

Ao contrário de tarefas com métricas bem definidas, como acurácia em classificação, muitos problemas reais envolvem critérios subjetivos, valores éticos e decisões com múltiplas dimensões. Nesses casos, especificar manualmente uma função de recompensa ou

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol2/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

um conjunto de regras explicitamente alinhadas com os valores humanos é frequentemente impraticável. Assim, cresce o interesse por abordagens que incorporam diretamente o julgamento humano no processo de treinamento, não apenas como fonte de dados supervisionados, mas como sinal de alinhamento.

É útil esclarecer o que se entende por agentes artificiais. São sistemas de IA capazes de interagir com um ambiente e tomar decisões de forma autônoma, com base em objetivos explícitos ou implícitos. Eles operam em um ciclo de percepção, decisão e ação, recebendo observações do ambiente (por exemplo, entradas textuais ou sensores), executando ações (como gerar respostas, mover-se fisicamente ou chamar APIs) e otimizando comportamentos de acordo com critérios de desempenho. Exemplos incluem assistentes conversacionais interativos, robôs autônomos, sistemas de recomendação adaptativos ou LLMs integrados a agentes de planejamento e execução como o AutoGPT. À medida que o grau de autonomia e poder de ação desses agentes aumenta, cresce também o potencial de desalinhamento, tornando o *AI Alignment* uma prioridade crítica.

O tema ganhou relevância crescente à medida que sistemas de IA passaram a atuar de forma autônoma em cenários críticos, como decisões financeiras, diagnósticos médicos e controle de veículos autônomos, onde pequenos desalinhamentos entre a função de recompensa e os valores humanos podem gerar comportamentos perigosos ou não intencionais (Amodei et al., 2016; Carlsmith, 2022). Além disso, fenômenos como *goal misgeneralization* e *specification gaming* ilustram a dificuldade de definir objetivos de forma robusta (Krakovna et al., 2020; Shah et al., 2022).

O primeiro ocorre quando o modelo aprende corretamente o comportamento desejado durante o treinamento, mas generaliza a meta subjacente de forma incorreta em novos contextos, perseguindo intenções aparentes em vez de objetivos reais. Por exemplo, um agente treinado para coletar moedas em um jogo pode, em cenários novos, priorizar qualquer objeto circular, interpretando erroneamente o conceito de "moeda". Já o *specification gaming* refere-se à exploração de falhas na função de recompensa ou nas restrições impostas pelo projetista, levando o agente a atingir pontuações elevadas sem realizar a tarefa de maneira adequada. Um exemplo clássico é o de agentes que maximizam métricas intermediárias, como distância percorrida, sem cumprir o objetivo final de forma útil. Esses fenômenos reforçam a necessidade de métodos de alinhamento que incorporem o julgamento humano de forma mais direta e adaptativa.

Duas das abordagens mais representativas desse paradigma são o *Reinforcement Learning from Human Feedback* (RLHF) (Christiano et al., 2017; Leike et al., 2018; Ouyang et al., 2022; Stiennon et al., 2020b) e o *Direct Preference Optimization* (DPO) (Rafailov et al., 2023), descritos a seguir com maior detalhamento.

3.4.1 Fundamentos de aprendizado por reforço

Antes de descrever RLHF e DPO, é importante revisar brevemente conceitos fundamentais do aprendizado por reforço (RL). O RL modela a interação entre um agente e um ambiente como um processo de decisão de Markov (MDP), definido por um conjunto de estados $\mathcal{S}$, ações $\mathcal{A}$, uma função de transição $P(s' \mid s, a)$ e uma função de recompensa $r(s, a)$.

Uma política $\pi_\phi(a \mid s)$ descreve a probabilidade de o agente tomar a ação a no estado s, parametrizada por ϕ. O objetivo do RL é maximizar a recompensa esperada acumulada ao longo do tempo:

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

$$J(\pi_\phi) = \mathbb{E}_{\pi_\phi}\left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t)\right] \tag{3.3}$$

onde $\gamma \in [0, 1)$ é um fator de desconto temporal. Tradicionalmente, assume-se que a função de recompensa $r(s, a)$ é totalmente especificada pelo projetista do sistema. No entanto, em tarefas abertas e de alto nível, como diálogo natural, decisões éticas ou planejamento estratégico, definir manualmente $r(s, a)$ de forma alinhada aos valores humanos é virtualmente impossível. Essa limitação motiva métodos que estimam recompensas indiretamente a partir de preferências humanas ou substituem sua modelagem explícita por técnicas baseadas em sinais comportamentais observados.

3.4.2 *Reinforcement Learning from Human Feedback* (RLHF)

O RLHF (Christiano et al., 2017; Leike et al., 2018; Ouyang et al., 2022) é um procedimento de alinhamento que modela o problema como uma tarefa de controle sequencial, na qual um agente otimiza uma política $\pi_\phi(y \mid x)$, a probabilidade de gerar uma resposta y dado um *prompt* x, com o objetivo de maximizar uma função de recompensa derivada de preferências humanas.

3.4.2.1 Modelo de recompensa

Dado um conjunto de comparações humanas entre respostas y^+ (preferida) e y^- (rejeitada) para um mesmo x, o objetivo é aprender uma função de recompensa $r_\theta(x, y)$ tal que:

$$P(y^+ \succ y^-) = \frac{\exp(r_\theta(x, y^+))}{\exp(r_\theta(x, y^+)) + \exp(r_\theta(x, y^-))} \tag{3.4}$$

A função de perda associada é a entropia cruzada negativa entre as preferências humanas e a distribuição induzida por r_θ:

$$\mathcal{L}_{\text{RM}}(\theta) = -\mathbb{E}_{(x,y^+,y^-)}\left[\log \frac{\exp(r_\theta(x, y^+))}{\exp(r_\theta(x, y^+)) + \exp(r_\theta(x, y^-))}\right] \tag{3.5}$$

3.4.2.2 Aprendizado por reforço

Com r_θ fixado, ajusta-se a política π_ϕ para maximizar a recompensa esperada:

$$\max_\phi \quad \mathbb{E}_{x\sim\mathcal{D}}\left[\mathbb{E}_{y\sim\pi_\phi(\cdot|x)}\left[r_\theta(x, y)\right]\right] \tag{3.6}$$

Para garantir estabilidade e evitar *drift*, impõe-se uma penalidade de divergência Kullback-Leibler (KL) em relação à política original π_0:

$$\mathcal{L}_{\text{RLHF}}(\phi) = -\mathbb{E}_{x\sim\mathcal{D}}\left[\mathbb{E}_{y\sim\pi_\phi(y|x)}\left[r_\theta(x, y) - \beta \cdot \log \frac{\pi_\phi(y \mid x)}{\pi_0(y \mid x)}\right]\right] \tag{3.7}$$

A função da Equação 3.6 pode ser reinterpretada como a maximização de uma vantagem regularizada:

$$\mathcal{L}_{\text{RLHF}}(\phi) = -\mathbb{E}_{\pi_\phi}\left[A^{r_\theta}(x, y) - \beta D_{\text{KL}}(\pi_\phi \parallel \pi_0)\right]$$

onde $A^{r_\theta}(x, y)$ representa a vantagem induzida pelo modelo de recompensa. Essa formulação evidencia que o treinamento busca não apenas maximizar a recompensa, mas também restringir a divergência da política em relação ao comportamento original do modelo.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol2/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Essa função é comumente otimizada via *Proximal Policy Optimization* (PPO) (Schulman et al., 2017), um algoritmo de aprendizado por reforço que atualiza a política por gradientes com restrições de divergência KL, garantindo que mudanças bruscas não degradem o desempenho.

3.4.2.3 Discussão sobre instabilidades

A separação entre a estimativa de preferências (r_θ) e a otimização da política (π_ϕ) pode introduzir instabilidades devido ao fenômeno conhecido como *reward hacking*. Quando o modelo de recompensa é imperfeito, a política pode explorar regiões do espaço de respostas onde r_θ é superestimado, produzindo comportamentos desalinhados com as intenções humanas. Além disso, a escolha do coeficiente β na Eq. (6) afeta a regularização: valores baixos permitem derivações excessivas da política, enquanto valores altos podem restringir a exploração, levando à subotimização.

3.4.3 *Direct Preference Optimization* (DPO)

O DPO (Rafailov et al., 2023) propõe um modelo mais direto e principiado para treinamento de políticas alinhadas, evitando a estimativa explícita de uma função de recompensa. A ideia central é modelar as preferências humanas como uma distribuição *a posteriori* sobre políticas, baseada na inferência bayesiana.

3.4.3.1 Formulação bayesiana

Assume-se que os dados de preferência $y^+ \succ y^-$ são amostrados de uma distribuição preferencial induzida por uma política ótima π^*, tal que:

$$P(y^+ \succ y^- \mid x) = \frac{\exp(\beta \cdot \log \pi^*(y^+ \mid x))}{\exp(\beta \cdot \log \pi^*(y^+ \mid x)) + \exp(\beta \cdot \log \pi^*(y^- \mid x))} \tag{3.8}$$

onde $\beta > 0$ controla a rigidez da preferência. Ao aproximar $\pi^* \approx \pi_\phi$, pode-se obter uma função de perda supervisionada para o aprendizado direto da política:

$$\mathcal{L}_{\text{DPO}}(\phi) = -\mathbb{E}_{(x,y^+,y^-)} \left[\log \frac{\exp(\beta \cdot \Delta \log \pi_\phi)}{\exp(\beta \cdot \Delta \log \pi_\phi) + 1} \right] \tag{3.9}$$

com $\Delta \log \pi_\phi = \log \pi_\phi(y^+ \mid x) - \log \pi_\phi(y^- \mid x)$.

A função de perda da Eq. (8) é análoga à perda logística em tarefas de classificação binária, onde o modelo aprende a distinguir respostas preferidas de rejeitadas. Sob essa perspectiva, o DPO pode ser interpretado como uma forma de inferência bayesiana aproximada, na qual a política ótima é tratada como distribuição posterior sobre ações condicionada às preferências humanas observadas.

3.4.3.2 Relação com aprendizado por reforço

A perda DPO pode ser interpretada como uma aproximação supervisionada do gradiente de política que surge da maximização da preferência empírica. Em vez de utilizar amostragem ou atualização por episódios, o DPO permite otimização por *backpropagation* puro, pois todas as quantidades relevantes são diretamente diferenciáveis.

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

3.4.3.3 Vantagens e limitações

Dentre as principais vantagens do DPO, temos:

- **Eficiência computacional:** não requer RL nem estimativa de r_θ;
- **Estabilidade:** treinamento supervisionado direto evita instabilidades de PPO;
- **Simplicidade:** menos componentes, menos fontes de erro, maior interpretabilidade.

Apesar de seus méritos, o DPO depende fortemente da fidelidade das distribuições de preferência observadas. Essa dependência significa que a política tende a reproduzir fielmente os padrões nos dados disponíveis, mas pode falhar em capturar estruturas de longo prazo que só emergem após várias etapas de decisão. Um modelo de recompensa explícito, ao inferir padrões latentes e recompensas cumulativas, pode capturar dependências temporais e objetivos hierárquicos que o DPO, por sua natureza supervisionada, tende a ignorar.

3.4.4 Discussão comparativa: RLHF vs. DPO

Ambas as abordagens compartilham o objetivo de incorporar preferências humanas ao treinamento de políticas, mas diferem em aspectos fundamentais. O RLHF separa o aprendizado da função de recompensa do processo de otimização, o que oferece maior flexibilidade e capacidade de generalização, porém introduz complexidade e risco de instabilidades devido ao *reward hacking*. O DPO, por outro lado, adota uma formulação mais direta e eficiente, reduzindo o custo computacional e a instabilidade, ao preço de depender fortemente da qualidade dos dados de preferência. Assim, o RLHF tende a ser mais adequado para cenários com dados escassos, mas onde a generalização é essencial, enquanto o DPO se mostra vantajoso quando há abundância de *feedback* humano confiável e rápido treinamento é desejável. Revisões abrangentes recentes (Hendrycks et al., 2021; Ngo et al., 2023) destacam que abordagens híbridas e multiestágio, combinando aprendizado por preferências, interpretabilidade e segurança formal, representam um caminho promissor para sistemas de IA alinhados e robustos. O Quadro 3.1 resume comparativamente as principais diferenças entre o RLHF e o DPO, destacando seus fundamentos teóricos, características de treinamento e limitações práticas.

Quadro 3.1: Comparação entre *Reinforcement Learning from Human Feedback* (RLHF) e *Direct Preference Optimization* (DPO)

Aspecto	RLHF	DPO
Formulação	Baseado em aprendizado por reforço com modelagem explícita de recompensa e atualização da política via PPO (Christiano et al., 2017; Ouyang et al., 2022; Schulman et al., 2017).	Baseado em otimização supervisionada direta de preferências humanas, sem estimar explicitamente a função de recompensa (Rafailov et al., 2023).

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol2/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Objetivo de treinamento	Maximizar a recompensa aprendida $r_\theta(x, y)$ regularizada pela divergência KL entre políticas.	Maximizar diretamente a probabilidade de respostas preferidas y^+ sobre rejeitadas y^- em uma formulação logística.
Tipo de atualização	Aprendizado por reforço com amostragem de trajetórias e gradientes de política (*policy gradients*).	Treinamento supervisionado via *backpropagation* padrão, sem interação com o ambiente.
Custo computacional	Alto, devido à necessidade de amostragem, atualização iterativa da política e estimativa de recompensas.	Baixo, pois elimina o ciclo de RL e opera inteiramente em dados estáticos de preferências.
Estabilidade de treinamento	Sensível ao ajuste do coeficiente de regularização β e à qualidade do modelo de recompensa, sujeito a *reward hacking*.	Mais estável, pois não requer modelagem de recompensas nem otimização sequencial via PPO.
Interpretação teórica	Aproximação regularizada do gradiente de vantagem em um MDP com recompensa aprendida.	Inferência bayesiana aproximada sobre políticas condicionadas a preferências humanas.
Generalização	Pode extrapolar preferências para novos contextos, desde que o modelo de recompensa capture relações latentes.	Limitado à distribuição observada de preferências, podendo falhar em extrapolações de longo prazo.
Aplicações típicas	Treinamento de LLMs (por exemplo, InstructGPT) e agentes de decisão com *feedback* humano em tempo real.	Ajuste fino supervisionado de LLMs com grandes conjuntos de preferências curadas (*offline alignment*).
Principais limitações	Complexidade computacional, dependência de modelo de recompensa robusto e risco de instabilidade.	Dependência da qualidade das preferências e dificuldade em capturar dependências temporais complexas.

3.5 Exemplo prático

Nesta seção, materializamos os conceitos teóricos discutidos ao longo do capítulo em uma aplicação prática de ponta a ponta. O objetivo é demonstrar o processo de especialização de um grande modelo de linguagem para uma tarefa de nicho no idioma português. Para isso, vamos criar o Dr. Bode, um assistente virtual de saúde, partindo de um modelo generalista e aplicando as técnicas de ajuste fino supervisionado (com PEFT) e otimização por preferência direta para adaptá-lo a um domínio altamente especializado, como a medicina.

3.5.1 Preparando o ambiente

A reprodutibilidade é um pilar da pesquisa e desenvolvimento em IA. Para garantir a consistência dos resultados, nosso ambiente de trabalho deve ser configurado corretamente. Este guia foi desenvolvido e testado com `Python 3.9`. As bibliotecas e suas respectivas

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol2/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

versões são listadas abaixo. Destacamos as principais: `transformers` e `datasets` para o ecossistema *Hugging Face*; `peft` para a aplicação de *Parameter-Efficient Fine-Tuning*; `trl` para os treinadores de alto nível (SFT e DPO); e `bitsandbytes` para a quantização do modelo.

```
transformers==4.53.1
torch==2.7.1
datasets==3.6.0
tqdm==4.67.1
peft==0.16.0
accelerate==1.8.1
trl==0.9.6
bitsandbytes==0.41.1
```

Ainda, definimos as seguintes variáveis globais, utilizadas durante a execução das células a seguir

```
NOME_MODELO = "Qwen/Qwen3-1.7B"
DATASET_AVALIACAO = "recogna-nlp/drbodebench"
DATASET_SFT = "recogna-nlp/drbode240_dataset"
DATASET_DPO = "recogna-nlp/dpo-mmlu-ptbr"
```

3.5.2 O ponto de partida

Todo projeto de ajuste fino começa com a escolha de um modelo de propósito geral. Para este estudo de caso, nosso ponto de partida será o `Qwen/Qwen3-1.7B`, um modelo com 1.7 bilhão de parâmetros. Este modelo foi escolhido por seu sólido desempenho em tarefas de raciocínio e seguimento de instruções, aliado a um tamanho que permite o treinamento em hardware acessível. Contudo, ele carece de conhecimento especializado em saúde, e nosso objetivo é justamente preencher essa lacuna. O treinamento será realizado com o *dataset* `recogna-nlp/drbode240_dataset`[1].

Antes de qualquer modificação, é crucial estabelecer uma linha de base (*baseline*) para quantificar nosso progresso. Avaliaremos o desempenho do modelo original no *benchmark* `recogna-nlp/drbodebench`[2], um conjunto composto por questões de múltipla escolha extraídas de provas de residência médica no Brasil. O resultado nos mostrará a capacidade inicial do modelo de responder a questões médicas complexas, servindo como ponto de partida para a jornada de especialização. O script a seguir realiza três etapas fundamentais:

- **Formatação do *prompt*:** Estrutura a questão no formato esperado pelo modelo.
- **Geração e extração da resposta:** Envia o *prompt* ao modelo e extrai a alternativa prevista.
- **Cálculo de acurácia:** Compara a resposta do modelo com o gabarito e calcula a acurácia total.

```
import re
import json
import time
import torch
```

[1] https://huggingface.co/datasets/recogna-nlp/drbode240_dataset
[2] https://huggingface.co/datasets/recogna-nlp/drbodebench

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol2/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

```
from datasets import load_dataset
from transformers import AutoModelForCausalLM, AutoTokenizer
from tqdm import tqdm


def formatar_prompt_medico(questao: dict) -> str:
    """
    Cria um prompt formatado a partir de um item do dataset.

    Args:
        questao (dict): Um dicionário contendo 'enunciado', 'alternativas'
                        e opcionalmente 'img_description'.

    Returns:
        str: O prompt completo e formatado, pronto para ser enviado ao modelo.
    """

    prompt = f"""Você é um especialista médico. Responda APENAS com a letra da
    alternativa correta (A, B, C, D ou E). Não escreva nada mais.

Questão:
{questao['enunciado']}

"""
    # Trata de forma segura a existência de imagens na questão.
    if questao.get("contains_img", False):
        prompt += f"Descrição da imagem: {questao.get('img_description', '')}\n\n"

    for letra, texto in questao["alternativas"].items():
        prompt += f"{letra}) {texto}\n"

    prompt += "\nResposta (apenas a letra):"
    return prompt.strip()


def obter_resposta_modelo(prompt: str, modelo: AutoModelForCausalLM,
    tokenizer: AutoTokenizer) -> str:
    """
    Gera uma resposta do modelo para um determinado prompt.

    Args:
        prompt (str): O prompt formatado.
        modelo (AutoModelForCausalLM): O modelo de linguagem carregado.
        tokenizer (AutoTokenizer): O tokenizer correspondente.

    Returns:
        str: A resposta textual gerada pelo modelo.
    """
    # O template de chat é específico de cada família de modelos.
    # Esta etapa garante que a entrada seja formatada como o modelo espera.
    messages = [{"role": "user", "content": prompt}]
    input_ids = tokenizer.apply_chat_template(
        messages, add_generation_prompt=True, return_tensors="pt",
        enable_thinking=False
    ).to(modelo.device)

    # Gera a resposta. `max_new_tokens` é baixo pois esperamos apenas uma letra.
    outputs = modelo.generate(
```

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol2/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

```
        input_ids,
        max_new_tokens=10,
        pad_token_id=tokenizer.eos_token_id
    )

    # Decodifica os tokens gerados, ignorando o prompt de entrada.
    resposta = tokenizer.decode(
        outputs[0][input_ids.shape[1] :],
        skip_special_tokens=True
    )

    return resposta.strip()


def extrair_alternativa(texto_resposta: str) -> str:
    """
    Extrai a primeira letra de alternativa (A-E) de um texto.

    Args:
        texto_resposta (str): O texto bruto gerado pelo modelo.

    Returns:
        str: A letra da alternativa em maiúsculo, ou uma string vazia se não for
        encontrada.
    """
    # A regex [A-Ea-e] busca a primeira ocorrência de uma das letras.
    match = re.search(r"[A-Ea-e]", texto_resposta)
    if match:
        return match.group(0).upper()
    return ""


def executar_avaliacao(modelo: AutoModelForCausalLM, tokenizer: AutoTokenizer,
    dataset: load_dataset, arquivo_saida: str):
    """
    Orquestra o processo de avaliação em todo o dataset.
    """

    start_time = time.time()
    resultados = []

    for item in tqdm(dataset, desc="Avaliando Questões"):
        prompt = formatar_prompt_medico(item)
        resposta_bruta = obter_resposta_modelo(prompt, modelo, tokenizer)
        predicao_processada = extrair_alternativa(resposta_bruta)

        resultados.append({
            "origem": item["origem"],
            "prova": item["prova"],
            "numero_questao": item["numero"],
            "enunciado": item["enunciado"],
            "resposta_correta": item["resposta"],
            "predicao_modelo": predicao_processada,
            "resposta_bruta_modelo": resposta_bruta,
        })


    # Cálculo de métricas e salvamento
```

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol2/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

```
    end_time = time.time()
    duracao_min = (end_time - start_time) / 60
    print(f"\nAvaliação finalizada em {duracao_min:.2f} minutos.")

    corretos = sum(1 for res in resultados if res["predicao_modelo"] ==
    res["resposta_correta"])
    total = len(resultados)
    acuracia = (corretos / total) * 100
    print(f"\nAcurácia final: {acuracia:.2f}% ({corretos}/{total})")

    with open(arquivo_saida, "w", encoding="utf-8") as f:
        json.dump(resultados, f, ensure_ascii=False, indent=4)
    print(f"Resultados salvos em: {arquivo_saida}")


# Carregando modelo e tokenizer
tokenizer = AutoTokenizer.from_pretrained(
    NOME_MODELO,
    trust_remote_code=True
)
model = AutoModelForCausalLM.from_pretrained(
    NOME_MODELO,
    trust_remote_code=True,
    torch_dtype=torch.bfloat16,
    device_map="auto"
)

# Carregando dataset de avaliação
eval_dataset = load_dataset(DATASET_AVALIACAO, split="train")

# Executar a avaliação
output_file = "resultado_modelo_base.json"
executar_avaliacao(model, tokenizer, eval_dataset, output_file)
```

Após a avaliação, o modelo base acerta 281 de 1301 questões, resultando em uma acurácia de 21,60%. Este valor, pouco acima de uma escolha aleatória (20% para 5 alternativas), confirma a ausência de conhecimento médico especializado e estabelece nosso ponto de partida.

3.5.3 Ajuste fino supervisionado com LoRA

Com a linha de base estabelecida, o próximo passo é ensinar ao modelo o conhecimento específico do domínio proposto. Utilizaremos a técnica de ajuste fino supervisionado, aplicando QLoRA para otimizar o uso de memória e reduzir os requisitos computacionais. Para o treinamento, seguiremos os passos de (1) carregar o modelo base quantizado, (2) configurar os adaptadores LoRA e (3) executar o treinamento com o `SFTTrainer`.

- **Configuração e carregamento do modelo:** Primeiro, configuramos a quantização e carregamos o modelo base. A quantização de 4 bits, gerenciada pela biblioteca `bitsandbytes`, é o que nos permite carregar um modelo de bilhões de parâmetros em uma GPU com VRAM limitada. A biblioteca `accelerate` auxilia no mapeamento do modelo para o dispositivo correto de forma automática.
- **Configuração do LoRA**: Definimos a configuração do LoRA usando a biblioteca `peft`. Em vez de treinar todos os 1.7 bilhão de parâmetros, vamos inserir adaptadores em camadas específicas do modelo. Apenas esses adaptadores, que representam uma pequena fração do total de parâmetros, serão treinados, resultando em uma economia massiva de recursos.
- **Treinamento com SFTTrainer:** Com o modelo preparado, utilizamos o `SFTTrainer` da biblioteca `trl`. Esta classe é um facilitador de alto nível que abstrai a complexidade do loop de treinamento. Ela lida com a formatação dos dados, o processamento em lotes e a otimização do treinamento de forma eficiente, exigindo apenas a configuração dos hiperparâmetros através da classe `SFTConfig`.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol2/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

```python
import os

import torch
from accelerate import PartialState
from peft import LoraConfig, get_peft_model, prepare_model_for_kbit_training
from transformers import AutoModelForCausalLM, AutoTokenizer, BitsAndBytesConfig
from datasets import load_dataset
from trl import SFTConfig, SFTTrainer


def encontrar_modulos_lineares(modelo: AutoModelForCausalLM) -> list:
    """
    Encontra todos os módulos lineares no modelo para aplicar o LoRA.

    Esta função auxiliar percorre todas as camadas do modelo e identifica aquelas
    que são instâncias de `torch.nn.Linear`, que são os alvos ideais para
    a inserção dos adaptadores LoRA. A camada 'lm_head' é geralmente excluída
    porque sua adaptação pode levar a instabilidade no treinamento.

    Args:
        modelo (AutoModelForCausalLM): O modelo a ser inspecionado.

    Returns:
        list: Uma lista de nomes dos módulos lineares alvo.
    """
    cls = torch.nn.Linear
    modulos_lora = set()
    for nome, modulo in modelo.named_modules():
        if isinstance(modulo, cls):
            nomes = nome.split(".")
            modulos_lora.add(nomes[0] if len(nomes) == 1 else nomes[-1])

    if "lm_head" in modulos_lora:
        modulos_lora.remove("lm_head")

    return list(modulos_lora)


def preprocessar_dataset(exemplo):
    """
    Formata uma entrada do dataset para o formato de texto único esperado pelo
    SFTTrainer.
    """
    partes_texto = []
    for entrada in exemplo["data"]:
        partes_texto.append(f"{entrada['role'].capitalize()}: {entrada['content']}")
    return {"data": "\n".join(partes_texto)}


# Determina o dispositivo para treinamento distribuído, se aplicável
string_dispositivo = PartialState().process_index

# Configuração da Quantização (QLoRA)
# QLoRA reduz o uso de memória ao carregar o modelo com pesos de 4 bits.
config_quantizacao = BitsAndBytesConfig(
    load_in_4bit=True,
    bnb_4bit_quant_type="nf4",  # Tipo de quantização (Normal Float 4)
    bnb_4bit_compute_dtype=torch.bfloat16,  # Tipo de dado para os cálculos
```

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol2/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

```
)

# Carregamento do modelo com a configuração de quantização
modelo_base = AutoModelForCausalLM.from_pretrained(
    NOME_MODELO,
    torch_dtype=torch.bfloat16,
    trust_remote_code=True,
    quantization_config=config_quantizacao,
    device_map={"": string_dispositivo},
)
# Desativa o cache para treinamento
modelo_base.config.use_cache = False
modelo_base = prepare_model_for_kbit_training(modelo_base)

# Carregamento do tokenizer
tokenizer = AutoTokenizer.from_pretrained(
    NOME_MODELO,
    trust_remote_code=True
)
tokenizer.pad_token = tokenizer.eos_token  # Token de padding como fim de sentença.
tokenizer.padding_side = "right" # Garante que o padding seja adicionado à direita.


# Configuracao do PEFT (LoRA)
# LoRA congela os pesos do modelo original e injeta
# matrizes treináveis (adapters) em camadas específicas.
peft_config = LoraConfig(
    r=128,  # O posto (rank) das matrizes de adaptação.
    lora_alpha=256,  # Fator de escala para os pesos do LoRA.
    target_modules=encontrar_modulos_lineares(modelo_base),  # Aplica LoRA em
                                                             #todas as camadas lineares.
    lora_dropout=0.05,  # Dropout para regularização dos adaptadores LoRA.
    bias="none",  # Não treina os vieses das camadas.
    task_type="CAUSAL_LM", # Especifica a tarefa como modelagem de linguagem causal.
)

# Aplica o wrapper PEFT para preparar o modelo para o treinamento LoRA.
modelo_peft = get_peft_model(modelo_base, peft_config)

# Carregamento do dataset
sft_dataset = load_dataset(DATASET_SFT, split="train")
sft_dataset = sft_dataset.map(preprocessar_dataset, remove_columns=["data"])

# Configuração do treinamento
# A classe SFTConfig armazena todos os hiperparâmetros para o SFTTrainer
output_dir_sft = "./modelo_sft"
argumentos_treinamento = SFTConfig(
    per_device_train_batch_size=8,
    gradient_accumulation_steps=4,
    gradient_checkpointing=True,
    max_grad_norm=0.3,
    num_train_epochs=1,
    learning_rate=2e-5,
    bf16=True,
    save_total_limit=3,
    logging_steps=1,
    output_dir=output_dir_sft,
    optim="paged_adamw_32bit",
```

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol2/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

```
    lr_scheduler_type="cosine",
    overwrite_output_dir=True,
    warmup_ratio=0.05,
    dataset_text_field="data",
)

# Instanciação e execução do treinador
treinador = SFTTrainer(
    model=modelo_peft,
    train_dataset=sft_dataset,
    peft_config=peft_config,
    args=argumentos_treinamento,
    processing_class=tokenizer,
)
treinador.train()

# Salva apenas os pesos dos adaptadores LoRA, que é um arquivo muito pequeno.
diretorio_checkpoint_final = os.path.join(output_dir_sft, "final_checkpoint")
treinador.model.save_pretrained(diretorio_checkpoint_final)
tokenizer.save_pretrained(diretorio_checkpoint_final)
```

Após o treinamento por uma única época, realizamos uma nova avaliação no *benchmark*. O modelo ajustado responde corretamente a 325 questões, alcançando uma acurácia de 24,98%, um aumento absoluto de 3,38 pontos percentuais. O ganho, embora positivo, é modesto. Isso ilustra um ponto importante: o SFT é altamente dependente da qualidade, diversidade e volume dos dados de treinamento. Fatores como o tamanho limitado do conjunto e o treinamento por uma única época podem explicar a magnitude do ganho. Para melhorias mais substanciais, seria necessário um conjunto de dados mais extenso, de maior qualidade ou treinando por múltiplas épocas de treinamento.

3.5.4 Refinamento do comportamento com DPO

Nosso modelo agora possui conhecimento especializado, adquirido durante o ajuste fino supervisionado. O próximo passo é refinar seu comportamento para que ele não apenas gere respostas corretas, mas também prefira um estilo de resposta a outro. O objetivo do DPO não é primariamente ensinar novo conhecimento, mas alinhar o comportamento do modelo a um estilo preferível.

O DPO requer um *dataset* de preferências no formato `(prompt, chosen, rejected)`. A criação desses datasets é um dos grandes gargalos em AI Alignment, pois geralmente exige anotação humana especializada e cara. Para fins didáticos, utilizamos um pequeno *dataset* de exemplo. O mecanismo do DPO ajusta o modelo para que a probabilidade da resposta chosen seja maximizada em relação à rejected, aprendendo implicitamente uma função de recompensa. O ponto de partida para o DPO é o nosso modelo já ajustado com SFT. Utilizamos o `DPOTrainer` da biblioteca `trl`, que ajustará os pesos do modelo com base no *dataset* de preferências para alinhar seu comportamento.

```
import torch
from peft import PeftModel
from transformers import AutoModelForCausalLM, AutoTokenizer
from trl import DPOConfig, DPOTrainer


# Carregamento do dataset de preferência
dataset_dpo = load_dataset(DATASET_DPO, split="train")
dataset_dpo = dataset_dpo.train_test_split(test_size=0.1)


# Carregamento do modelo ajustado com SFT
caminho_adaptadores = "./modelo_sft/final_checkpoint"

modelo_base = AutoModelForCausalLM.from_pretrained(
    NOME_MODELO,
```

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol2/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

```
    trust_remote_code=True,
    torch_dtype=torch.bfloat16,
    device_map="auto",
)

# Anexa os adaptadores LoRA do SFT e os torna treináveis novamente para a fase de
# DPO.
modelo_dpo = PeftModel.from_pretrained(
    modelo_base,
    caminho_adaptadores,
    is_trainable=True
)

tokenizer = AutoTokenizer.from_pretrained(caminho_adaptadores,
trust_remote_code=True)

# Configura o token de padding, se necessário.
if tokenizer.pad_token is None:
    tokenizer.pad_token = tokenizer.eos_token

# A classe DPOConfig armazena todos os hiperparâmetros para o treinamento.
config_dpo = DPOConfig(
    per_device_train_batch_size=2,
    per_device_eval_batch_size=2,
    gradient_accumulation_steps=8,
    num_train_epochs=1,
    learning_rate=5e-7,
    logging_steps=10,
    eval_strategy="steps",
    eval_steps=100,
    save_strategy="steps",
    save_steps=500,
    save_total_limit=2,
    report_to="none",
    remove_unused_columns=False,
    max_length=512,
    padding_value=None,
    beta=0.5,
    loss_type="sigmoid",
)

# Instanciação e execução do treinador DPO
treinador_dpo = DPOTrainer(
    model=modelo_dpo,
    ref_model=None,
    args=config_dpo,
    processing_class=tokenizer,
    train_dataset=dataset_dpo["train"],
    eval_dataset=dataset_dpo["test"],
)
treinador_dpo.train()

# Salva os adaptadores LoRA refinados pelo DPO.
output_dir_dpo = "./modelo_dpo"
treinador_dpo.save_model(output_dir_dpo)
tokenizer.save_pretrained(output_dir_dpo)
```

Após a etapa de DPO, a acurácia do modelo foi de 25,06% (326 acertos). O ganho de 0,08 pontos percentuais é marginal, o que é esperado. O DPO não foi projetado para melhorar a performance em

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol2/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

testes de conhecimento de múltipla escolha, mas sim para moldar o estilo da geração de texto. Uma avaliação qualitativa, analisando a cautela, clareza e segurança das respostas abertas do modelo, seria necessária para medir o verdadeiro impacto desta etapa em cenários reais de avaliação.

3.5.5 Considerações finais

Neste estudo de caso, transformamos o modelo generalista `Qwen/Qwen3-1.7B` em um protótipo de especialista em saúde. Partimos de uma acurácia de 21,60% e, por meio de um processo incremental de duas fases, aprimoramos tanto seu conhecimento quanto seu comportamento.

1. Ajuste fino supervisionado: Com QLoRA, infundimos conhecimento médico específico, elevando a acurácia para 24,98%. Esta etapa foi crucial para ensinar ao modelo o conteúdo factual do domínio.
2. Otimização por preferência direta: Com um *dataset* de preferências, alinhamos o comportamento do modelo, resultando em uma acurácia final de 25,06% e refinando o estilo e a forma da resposta.

A análise comparativa revela os pontos fortes e fracos inerentes a cada técnica. O SFT é a principal fonte de ganho de conhecimento, enquanto o DPO oferece um controle mais refinado sobre o comportamento. Este exemplo prático demonstra que a especialização de LLMs não é um ato único, mas um processo iterativo, onde a combinação de técnicas permite criar modelos mais capazes, seguros e alinhados. O Dr. Bode, desenvolvido neste capítulo, embora ainda um protótipo, ilustra o potencial e os desafios práticos dessa jornada.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol2/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Capítulo 4

Avaliação de Grandes Modelos de Linguagem

Fundamentos, métodos tradicionais e desafios atuais

Livy Real
André Carvalho
Altigran Soares da Silva

Publicado em: 16/04/2026

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol2/

4.1 A importância e os desafios intrínsecos da avaliação de LLMs

Talvez você já tenha ouvido falar do caso em que uma LLM (Large Language Model, Grande Modelo de Linguagem) vendeu um carro por 1 dólar[1]. Ou de quando um advogado estado-unidense, ex-advogado do próprio presidente, foi processado por inventar precedentes jurídicos[2]. Se não ouviu, certamente você já ficou preso em conversas intermináveis e inúteis de *chatbots* de diversas empresas. Todos esses casos aconteceram e acontecem todos os dias porque os responsáveis por tais sistemas, mesmo que inadvertidamente, desconsideram a necessidade de uma avaliação robusta de seus sistemas baseados em LLMs. A ampla divulgação comercial desses sistemas pode levar à percepção equivocada de que já atingimos a era da Inteligência Artificial Geral (AGI), o tão esperado momento em que as IAs terão 'inteligência' semelhante à humana e serão autodidatas. Isso ainda não corresponde à realidade. Hoje, em 2025, é possível afirmar que as saídas das LLMs sempre parecem boas, parecem verossímeis e factuais, mas com frequência não o são. Esse capítulo aborda a necessidade da avaliação de LLMs, considerando técnicas tradicionais e técnicas mais recentes, e vai defender a integração da avaliação das LLMs dentro do próprio ciclo de desenvolvimento de um sistema ou produto.

A avaliação de LLMs constitui uma questão fundamental no ciclo de vida de pesquisa e desenvolvimento desses modelos e de aplicações construídas com base neles, incluindo aí os agentes inteligentes. A sua importância transcende a mera verificação de desempenho, sendo crucial para o avanço da área a comparação entre diferentes arquiteturas, a identificação e mitigação de riscos e a garantia de aplicações práticas robustas e seguras (Meva; Kukadiya, 2025). A complexidade inerente à linguagem natural, combinada com a natureza gerativa e as capacidades emergentes dos LLMs, introduz desafios avaliativos que superam as abordagens tradicionais de Processamento de Linguagem Natural (PLN) (Minaee et al., 2025). A falta de interpretabilidade destes modelos ainda contribui à longa lista de novos desafios enfrentados na área atualmente.

LLMs, como modelos de linguagem estatísticos pré-treinados em larga escala, exibem "habilidades" interessantes que vão além da mera previsão de palavras. Incluem-se o aprendizado em contexto (in-context learning), em que os modelos assimilam novas tarefas a partir de poucos exemplos fornecidos no *prompt* , e o seguimento de instruções (*instruction following*), que permite que os modelos executem tarefas sem exemplos explícitos após o *instruction tuning*[3] (Minaee et al., 2025). Estas capacidades, embora revolucionárias, trazem novos desafios à avaliação, pois as métricas clássicas usadas para mensurar sistemas de PLN frequentemente não conseguem capturar a profundidade semântica, a criatividade ou a factualidade das respostas geradas. Como as LLMs são treinadas para a simples tarefa de predizer a próxima palavra de

[1] https://the-decoder.com/people-buy-brand-new-chevrolets-for-1-from-a-chatgpt-chatbot/

[2] https://www.npr.org/2023/12/30/1222273745/michael-cohen-ai-fake-legal-cases

[3] *Instruction tuning* é o processo de re-treinar um modelo de linguagem já pré-treinado usando pares (instrução, resposta) para alinhar o modelo a seguir instruções (instruction following) (Zhang et al., 2025c).

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

uma sequência, mas, a partir desse treinamento, surgem capacidades emergentes (habilidades dos modelos que não foram explicitamente treinadas), não sabemos de antemão quais habilidades devemos avaliar.

A avaliação do raciocínio passo a passo, por exemplo, é fundamental para compreender e aprimorar a capacidade dos LLMs em problemas complexos, exigindo critérios que vão além da simples precisão da resposta final, como factualidade, validade, coerência e utilidade. A necessidade de *frameworks* de avaliação mais sofisticados é cada vez mais evidente, especialmente à medida que os LLMs são integrados em domínios de alta criticidade, como a saúde, na qual segurança e eficácia são imperativas. A avaliação deve, portanto, ser multifacetada, abrangendo não apenas o desempenho técnico, mas também a segurança, a eficiência e o impacto social dos modelos (Bender et al., 2021).

A evolução do paradigma de avaliação é uma resposta direta à forma como os LLMs operam. Sistemas de PLN anteriores frequentemente se propunham a tarefas mais simples e as métricas tradicionais projetadas para tais tarefas frequentemente assumem a existência de uma única resposta correta (Minaee et al., 2025). *Benchmarks* estáticos e métricas rígidas guiaram por décadas o avanço das técnicas de PLN, mesmo que as limitações dessas metodologias fossem conhecidas e frequentemente apontadas (Bowman; Dahl, 2021; Kalouli et al., 2023; Ribeiro et al., 2020). No entanto, a natureza gerativa dos LLMs, que podem produzir diversas saídas válidas para a mesma entrada, torna o problema da avaliação mais evidente, na medida em que a própria definição de uma resposta "boa" ou "correta" se torna uma definição fluida e multidimensional (Meva; Kukadiya, 2025; Minaee et al., 2025). Esta observação implica uma transição: saímos de métricas puramente objetivas e baseadas em correspondência de strings e vamos em direção a abordagens de avaliação mais diversificadas, buscando diferentes relações semânticas e alinhadas com a percepção humana. O desafio não é apenas medir a correção, mas sim a qualidade em um espaço subjetivo, criativo e dependente do contexto, o que demanda uma evolução contínua dos *frameworks* de avaliação.

Adicionalmente, a avaliação não é um processo *post-hoc*, mas uma parte integral do ciclo de desenvolvimento dos LLMs. O aprendizado por reforço (Reinforcement Learning, RL) utiliza pontuações de avaliadores humanos para aprimorar o treinamento de um LLM. Isso estabelece um vínculo causal: métricas de avaliação falhas ou incompletas, especialmente aquelas que não se alinham com os valores humanos ou com as capacidades emergentes dos modelos, podem levar à otimização dos sistemas para a métrica em si, em vez de levá-los a novas versões realmente úteis e seguras. Assim, a qualidade e o alinhamento das pontuações de avaliação impactam diretamente a trajetória de aprendizado do modelo e seu desempenho subsequente, especialmente em esforços de alinhamento como o RLHF (Reinforcement Learning from Human Feedback).

4.1.1 Tipos de avaliação

A avaliação de LLMs pode ser categorizada de diversas maneiras para abordar os múltiplos aspectos de seu desempenho. Uma distinção fundamental é entre avaliação *intrínseca* e *extrínseca*[4]. A avaliação intrínseca foca na qualidade inerente do modelo, frequentemente medida por métricas que avaliam a probabilidade de sequências de texto, como a perplexidade, ou a sobreposição lexical com textos de referência. A avaliação extrínseca, por outro lado, mede o desempenho do LLM em uma tarefa finalística ou *downstream*[5] ou em um cenário de aplicação real, como a capacidade de um LLM de melhorar a experiência do usuário em um *chatbot* ou sua eficácia em tarefas de sumarização de notícias.

Outra categorização crucial diferencia a *avaliação humana* da *automática*. A avaliação automática tradicional emprega algoritmos e métricas computacionais para comparar a saída do modelo com referências ou para inferir a qualidade com base em características textuais. Este tipo inclui métricas estatísticas tradicionais (como BLEU e ROUGE). Atualmente, também há a avaliação automática baseada em modelos de linguagem, como a utilização do BERTScore e o paradigma LLM-as-a-Judge) (Gao et al., 2024). Vamos visitá-las nas próximas páginas.

Já a avaliação humana é amplamente considerada o "padrão ouro" especialmente para tarefas de Geração de Linguagem Natural (NLG) devido à sua capacidade de capturar nuances e critérios de qualidade subjetivos que as métricas automáticas não conseguem capturar (Dierk et al., 2025; Gao et al., 2024; Wang et al., 2025).

[4] A Seção Paradigmas: Tipos de avaliação deste livro traz outras formas de categorizar possíveis avaliações de sistemas de PLN. Veja as Seções Avaliação intrínseca e Avaliação extrínseca para entender a aplicação de avaliação intrínseca e extrínseca a *datasets* específicos.

[5] Uma tarefa *downstream* é uma tarefa 'mais para o fim do fluxo'. Considere que um LLM é treinado para prever a próxima palavra (essa seria a *upstream task*, no início do fluxo). Os usos reais, análise de sentimento, classificação textual, agente conversacional, são as *downstream tasks*, as tarefas reais nas quais aquele modelo será empregado (Schnabel et al., 2015).

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol2/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Há uma complementaridade e compromissos (*trade-offs*) inerentes entre a avaliação humana e a automática. Embora a avaliação humana seja o "padrão ouro", ela apresenta desafios significativos em termos de custo e escalabilidade, sendo um processo caro e demorado (Dierk et al., 2025; Gao et al., 2024). Em contraste, as métricas automáticas são eficientes e rápidas de calcular, mas muitas vezes carecem da capacidade de capturar a profundidade semântica e a coerência contextual que os humanos podem discernir (Minaee et al., 2025). Esta observação implica que um *framework* de avaliação holístico deve alavancar os pontos fortes de ambas as abordagens, em vez de depender exclusivamente de uma.

A combinação de verificações automáticas rápidas para qualidade básica e consistência com revisões humanas direcionadas para aspectos subjetivos, segurança e raciocínio complexo sugere um futuro no qual as abordagens de avaliação híbridas se tornarão o padrão. O trade-off entre custo/velocidade e profundidade/nuance possivelmente impulsionará as escolhas metodológicas na implantação prática de LLMs. A avaliação humana é essencial para validar textos gerados por máquinas e garantir o alinhamento com requisitos específicos do domínio e expectativas do usuário. A dificuldade intrínseca da avaliação de tarefas gerativas significa que a avaliação humana é sempre necessária. Além de não existir uma "resposta única correta" para muitas tarefas de geração de linguagem (Gao et al., 2024), é apenas a avaliação humana que garante o alinhamento de LLMs ou de qualquer outro tipo de avaliação (meta-avaliação) a valores humanos. Por valores humanos, entende-se não apenas ética e responsabilidade social, mas justamente a qualidade da máquina em agir exatamente como um humano agiria (e não apenas 'parecido').

4.2 Avaliações tradicionais

Em PLN, as tarefas podem ser divididas, de maneira geral, em dois grandes grupos: classificação e geração. Tradicionalmente, cada um desses tipos de tarefas é avaliado com uma abordagem específica, utilizando métricas distintas.

Nas tarefas de classificação, o sistema atribui rótulos (classes) a entradas de texto, como na análise de sentimentos ou na categorização de um documento. A avaliação costuma ser feita por métricas como acurácia, precisão, revocação e F1-score, que medem o quanto as previsões do modelo coincidem com as respostas esperadas. Tais métricas têm mais relação com classes, distribuição de instâncias nas classes do que com linguagem propriamente dita.

Já nas tarefas de geração, o modelo produz novas sequências de texto, como em tradução automática, sumarização ou resposta a perguntas. Nesse caso, a avaliação é mais complexa, pois envolve tanto a fluência quanto a adequação do texto, recorrendo a métricas automáticas como BLEU, ROUGE e METEOR, além de julgamentos humanos de qualidade. A maioria das métricas automáticas de geração tende a avaliar a corretude sintática e/ou semântica do texto gerado em relação a um texto de referência.

Neste capítulo, não revisitaremos em detalhe o funcionamento de cada uma dessas métricas, já discutidas com maior profundidade na Seção Métricas: Medindo a performance. Para ver mais sobre tarefas de geração de linguagem, você também pode consultar o Capítulo Geração de linguagem natural.

Cabe destacar, no entanto, que as métricas tradicionais de PLN, embora adequadas para modelos anteriores e tarefas específicas, mostram-se insuficientes diante das capacidades emergentes dos LLMs. Em estágios iniciais da tecnologia, a avaliação costumava concentrar-se em camadas superficiais do texto, como correção lexical ou gramaticalidade. Com os LLMs, que já atingem alta precisão nesses aspectos, torna-se necessário avaliar dimensões mais complexas da linguagem, como consistência factual, completude e utilidade em contextos específicos. Assim, o próprio objeto de avaliação se transforma, exigindo métricas novas e mais abrangentes, capazes de capturar a natureza multifacetada da geração de linguagem em larga escala.

4.2.1 Descompasso com a percepção humana

Um dos principais desafios das métricas tradicionais é o seu descompasso com a percepção humana de qualidade. Métricas baseadas em sobreposição de palavras, como BLEU e ROUGE, penalizam severamente paráfrases e sinônimos, mesmo que semanticamente equivalentes à referência (Mathur et al., 2020; Peyrard, 2019). Por exemplo, um texto gerado por um LLM que utiliza um vocabulário diferente, mas que transmite o mesmo significado que o texto de referência, pode receber uma pontuação artificialmente baixa nessas métricas, apesar de ser considerado de alta qualidade por um avaliador humano (Mathur et al., 2020; Minaee et al., 2025). O BLEU, por exemplo, ignora a sequência de palavras, resultando na mesma pontuação para frases com as mesmas palavras, mas com significados diferentes devido à ordem (Mathur et al., 2020). Ainda que esses sejam problemas que a área de geração de linguagem sempre tenha enfrentado (Gatt; Krahmer, 2018), eles se acentuam com o advento das LLMs.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol2/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

As métricas tradicionais também não foram desenhadas para capturar aspectos complexos da linguagem, como fluência ou criatividade (Gao et al., 2024). Elas são sensíveis a pequenas variações lexicais que não afetam a qualidade percebida pelo ser humano (Mathur et al., 2020). A otimização exclusiva para essas métricas pode levar a modelos que "enganam" a métrica em vez de realmente melhorar a qualidade da geração de texto (Gao et al., 2024). Isso demonstra que as métricas tradicionais, ao focar na estrutura de superfície (*surface structure* (Chomsky, 1965)), falham em capturar a equivalência semântica e podem levar a conclusões enganosas sobre o desempenho real do sistema (Mathur et al., 2020; Minaee et al., 2025).

4.2.2 Dificuldade em avaliar aspectos complexos

Além do descompasso com a percepção humana, as métricas tradicionais demonstram uma dificuldade intrínseca em avaliar aspectos mais complexos e emergentes dos LLMs.

Factualidade e Alucinações

LLMs são propensas a "alucinações", elas geram informações plausíveis, textos estatisticamente possíveis, mas factualmente incorretos, irrelevantes ou sem sentido (Meva; Kukadiya, 2025). Por exemplo, um LLM pode afirmar que "Thomas Edison inventou a internet" ou fornecer detalhes clínicos incorretos sobre um paciente específico. Tais erros não são detectados por métricas tradicionais, pois o texto pode ser gramaticalmente correto e fluente (Gao et al., 2024). A factualidade, especialmente em cenários de raciocínio passo a passo, exige que a informação possa ser fundamentada em fontes confiáveis ou que utilize corretamente números e restrições fornecidos na consulta. Verificar o conhecimento factual e o conhecimento de senso comum parametrizados, isto é, o conhecimento real e factual que a LLM adquiriu através de seu treinamento, permanece um desafio em aberto.

Raciocínio e Senso Comum

Tanto o processamento da informação de forma que pareça um raciocínio passo-a-passo quanto a capacidade de armazenar o chamado 'senso comum' são das mais relevantes e impressionantes capacidades emergentes que as LLMs apresentam. Por serem 'novidade', as métricas clássicas são incapazes de medir a capacidade de inferência, raciocínio complexo ou compreensão profunda dos LLMs (Meva; Kukadiya, 2025). A avaliação do raciocínio passo a passo é crucial para entender e melhorar o desempenho dos LLMs em problemas complexos como lógica, matemática e ciência. As métricas tradicionais não foram desenhadas para medir esses aspectos cruciais do processo de raciocínio. LLMs podem gerar saídas que soam plausíveis, mas carecem de coerência lógica, especialmente quando se baseiam em correlações estatísticas em vez de raciocínio causal (Xiong et al., 2025).

Segurança e Viés (*Bias*)

As métricas tradicionais também não foram projetadas para detectar conteúdo tóxico, preconceituoso ou prejudicial, a menos que sejam explicitamente treinadas para isso (Meva; Kukadiya, 2025). A avaliação da segurança e do viés requer abordagens específicas que analisem a imparcialidade, a inofensividade e a recusa de respostas inadequadas (Jiao et al., 2024). A otimização para métricas de sobreposição lexical não garante que o modelo não produzirá conteúdo problemático.

Seguimento de instruções

A capacidade de um LLM de seguir instruções complexas ou com múltiplas restrições é um aspecto crucial para sua utilidade (Minaee et al., 2025). Por esta ser uma nova capacidade da tecnologia, métricas tradicionais não a avaliam. É especialmente difícil avaliar o quão bem o modelo adere a essas instruções quando estas envolvem nuances ou múltiplos passos (Gao et al., 2024). LLMs podem falhar em compreender contextos longos ou instruções em conversas multiturno, quando as instruções são inconsistentes ou contêm informações incorretas. Métricas recentes que avaliam sistemas de diálogo podem avaliar a adequação de uma resposta ao contexto do turno anterior ou ainda a progressão do diálogo (Yeh et al., 2021), características que se assemelham ao *instruction following*, porém, tais métricas são também mensuradas através de modelos treinados para isso, assemelhando-se ao novo paradigma de avaliação que vamos discutir ao longo do capítulo, o *LLM as a Judge*.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol2/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

4.2.3 Exemplos práticos

Para ilustrar as limitações das métricas tradicionais, vamos visitar o trabalho de (Li et al., 2025) que compara diferentes tipos de avaliação em uma tarefa específica do domínio do direito.

Os autores propõem o CaseGen, um conjunto de dados para avaliação de geração de documentos de casos jurídicos na área penal chinesa. O *benchmark* tem 500 casos reais, anotados por especialistas do direito. Os autores avaliam 9 diferentes LLMs para gerar parte desses casos. Para avaliar como essas LLMs foram, especialistas em direito avaliaram detidamente 50 casos. Esses 50 casos anotados foram comparados com diferentes métricas, como ROUGE-L, BertScore e o paradigma LLM-as-a-Judge (uma LLM avaliando outra LLM). Esse modelo dá notas de 0 a 10 para cada documento gerado considerando critérios como precisão dos fatos, lógica e correção legal.

Veja na Tabela abaixo a comparação entre a anotação humana e três métricas de avaliação usando diferentes medidas de calcular a correlação entre anotação. As correlações de Kendall, Pearson e Spearman são métricas de correlação entre avaliações automáticas e avaliações humanas frequentemente usadas quando precisamos medir a proximidade das avaliações (Shaqiri et al., 2023).

Tabela 4.1: Comparação entre correlações das métricas automáticas e avaliações humanas.

Metrics	LLM Score	Rouge-L	BERTScore
Kendall	**0.667**	0.333	0.166
Pearson	**0.726**	0.264	0.239
Spearman	**0.750**	0.375	0.250

O interessante para a gente é que as notas automáticas se aproximam bastante das avaliações de juristas humanos (correlação de até 0,75 quando compara-se a anotação humana à anotação do LLM juíz), muito melhor do que as métricas tradicionais, que ficaram abaixo de 0,4.

Esse exemplo demonstra que, embora as métricas tradicionais sejam fáceis de computar e úteis para certas comparações de superfície, elas não são necessariamente alinhadas às anotações humanas, exigindo a complementação com abordagens mais robustas e, muitas vezes, híbridas. É ainda relevante notar que a avaliação com LLM chega a 0.75 de correlação, mas não chega a 0.8 nem a 1. Por isso, a avaliação humana ainda é **tecnicamente** a mais confiável.

4.2.4 Avaliação humana: O padrão ouro?

Como vimos, a avaliação humana é amplamente reconhecida como o "padrão ouro" para a avaliação de LLMs, especialmente para tarefas de Geração de Linguagem Natural (NLG), onde a subjetividade e a nuance são inerentes (Dierk et al., 2025; Gao et al., 2024).

4.2.5 Importância e metodologias comuns

A avaliação humana é crucial porque é capaz de capturar aspectos de qualidade que as métricas automáticas não conseguem. Isso inclui a fluência, coerência, relevância semântica profunda, criatividade, factualidade e o seguimento de instruções complexas (Gao et al., 2024). A capacidade dos avaliadores humanos de compreender o contexto, interpretar significados implícitos e julgar a adequação de uma resposta para um determinado propósito a torna indispensável (Gao et al., 2024).

Metodologias comuns de avaliação humana incluem:

- **Escalas Likert** – Utilizadas para avaliar a qualidade em diversas dimensões, como fluência, relevância, coerência e adequação. Os avaliadores atribuem uma pontuação em uma escala graduada (e.g., de 1 a 5 ou 1 a 100) para cada critério. Esta abordagem permite quantificar julgamentos subjetivos e pode ser aprimorada com a coleta de explicações em linguagem natural dos avaliadores para capturar nuances e calibrar as pontuações.
- ***Ranking*** – Envolve a comparação direta entre diferentes saídas de modelos para uma mesma entrada. Os avaliadores classificam as respostas em ordem de preferência, indicando qual é superior ou se há um empate (Vongthongsri, 2025). Essa metodologia pode ser mais estável do que a pontuação absoluta, pois foca na comparação relativa (Gao et al., 2024), porém ela nem sempre é reutilizável. Se, em um primeiro momento, analistas comparam dois modelos e, finalmente, temos um terceiro modelo em jogo, a primeira avaliação dificilmente será útil para decidirmos algo sobre este novo modelo.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol2/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

- **Testes A/B** – Uma forma de ranking em que os avaliadores escolhem a preferência entre duas opções (A ou B) (Río; Vaahtio, 2024). É útil para comparar o desempenho de dois modelos ou variações de um mesmo modelo, como diferentes *prompts* , por exemplo. Geralmente testes A/B são feitos com grupos diferentes de usuários que não sabem que estão participando de um teste. Quando uma avaliação é feita através de teste A/B é importante considerar tanto a relevância estatística das amostras quanto como coletar o *feedback* dos grupos antes de implementar o teste. Rodar um teste sem ter uma hipótese, sem saber coletar e interpretar o *feedback* pode ser inócuo.
- **Análise de erros** – Consiste em classificar os tipos de erros presentes nas saídas dos LLMs (e.g., erros gramaticais, factuais, de coerência, alucinações, omissões). Esta metodologia fornece avaliação detalhada que pode ser usada para depurar e melhorar os modelos. A coleta de anotações de erros em respostas de LLMs é desafiadora devido à natureza subjetiva de muitas tarefas de PLN e à dificuldade em obter anotações objetivas para erros complexos. Veja a Subseção Análise de Erro deste livro para saber mais sobre análise de erro.

4.2.6 Desafios da avaliação humana

Apesar de ser o "padrão ouro", a avaliação humana apresenta uma série de desafios que limitam sua escalabilidade e consistência.

Custo e Tempo

O processo de avaliação humana é notoriamente caro e demorado (Dierk et al., 2025). A obtenção de anotações de qualidade pode levar dias ou semanas, o que introduz latência no ciclo de *feedback* de desenvolvimento do modelo. Isso torna impraticável realizar uma avaliação humana completa para cada ajuste ou iteração do modelo (Gao et al., 2024).

Subjetividade e Variabilidade

Diferentes avaliadores podem ter opiniões distintas sobre a qualidade de uma mesma saída, levando a variabilidade nos julgamentos (Dierk et al., 2025; Jiao et al., 2024; Liu et al., 2024). Para algumas tarefas, como a avaliação da "criatividade" de uma história, até mesmo humanos podem discordar sobre o que é "melhor" (Gao et al., 2024). A subjetividade pode ser mitigada com diretrizes claras e perguntas mais objetivas.

Viés do avaliador

As experiências, expectativas e origens culturais dos avaliadores podem influenciar seus julgamentos, introduzindo vieses (Jiao et al., 2024). Por exemplo, um avaliador de uma cultura específica pode ter uma percepção diferente do que é "polido" em comparação com outro. É crucial ter um grupo de avaliadores diversificado para mitigar esses vieses. Além disso, a forma como a tarefa é apresentada aos avaliadores pode influenciar os resultados (e.g., "Modelo A foi ajustado para ser mais conciso. Qual é melhor?" pode induzir a um viés em favor da concisão) (Gao et al., 2024). Um viés conhecido é o viés da concordância ("acquiescence bias") (Cronbach, 1942), que caracteriza a tendência do avaliador em concordar com afirmações ou responder de forma afirmativa, por exemplo, concordar com as etiquetas e as saídas dos sistemas automáticos quando exposto a eles.

Escalabilidade

É difícil aplicar a avaliação humana em larga escala e de forma contínua, especialmente com o rápido avanço e o grande volume de saídas dos LLMs (Dierk et al., 2025). A avaliação humana é frequentemente reservada para comparações finais ou verificações periódicas, enquanto métricas automáticas são frequentemente usadas durante o desenvolvimento dos sistemas (Gao et al., 2024).

Treinamento e Calibração de avaliadores

Para garantir consistência e alta qualidade nas anotações, os avaliadores precisam de diretrizes claras e treinamento rigoroso (Gao et al., 2024). A calibração é essencial para garantir que diferentes avaliadores apliquem os critérios de forma consistente (Liu et al., 2024). A falta de controle de qualidade pode levar a anotações de baixa qualidade, exigindo a inclusão de perguntas de teste para filtrar trabalhos de

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol2/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

baixa qualidade (Gao et al., 2024). Veja as Seções Procedimentos e estratégias de anotação e revisão e Concordância entre-anotadores deste livro para mais sobre estratégias de anotação e concordância entre anotadores.

Em suma, embora a avaliação humana seja insubstituível para capturar a qualidade genuína dos LLMs em tarefas complexas, seus desafios inerentes impulsionam a pesquisa em métodos de avaliação automática mais sofisticados e em abordagens híbridas que combinam o melhor dos dois mundos. Além disso, a maioria das avaliações por humano tem sido feita com *crowdworkers* (Hedderich; Oulasvirta, 2025), trabalhadores terceirizados, geralmente leigos e (mal) pagos por quantidade de anotação feita. Essa prática impacta diretamente a qualidade da avaliação, além de levantar questões éticas preocupantes (Gonzalez-Cabello et al., 2025; Jiang; Wagner, 2024) tanto socialmente quanto tecnicamente.

4.3 *Benchmarks* e *datasets* de avaliação

Benchmarks, no contexto da avaliação de LLMs, são conjuntos de dados e tarefas padronizadas que servem como referência para comparar o desempenho de diferentes modelos. Eles são cruciais para o avanço da pesquisa, permitindo uma avaliação sistemática e reprodutível das capacidades dos LLMs (Aisera, 2024; Minaee et al., 2025). Atualmente, temos muitos *benchmarks* disponíveis, especialmente para o inglês. Visitaremos apenas alguns para familiarizar o leitor com esses recursos.

4.3.1 *Benchmarks* genéricos (General-Purpose)

GLUE e SuperGLUE

GLUE (General Language Understanding Evaluation) e SuperGLUE (Super General Language Understanding Evaluation) são *benchmarks* históricos que impulsionaram a pesquisa em PLN. Enquanto GLUE focava em tarefas de compreensão de linguagem mais básicas, como inferência em linguagem natural (NLI), SuperGLUE foi desenvolvido para introduzir desafios mais complexos, exigindo raciocínio linguístico avançado, compreensão contextual, recuperação de conhecimento e aprendizado multi-tarefa (Wang et al., 2018). SuperGLUE inclui tarefas como resolução de correferência e question answering, refletindo problemas mais próximos do raciocínio humano (Wang et al., 2019b). No entanto, o GLUE original enfrentou o problema de saturação, onde o desempenho dos modelos se aproximava do nível humano, limitando sua capacidade de diferenciar modelos de ponta (Wang et al., 2019b). SuperGLUE buscou superar essa limitação, oferecendo um desafio mais rigoroso e com maior "espaço" para melhorias dos modelos.

HELM

HELM (Holistic Evaluation of Language Models) representa uma abordagem mais abrangente para a avaliação de LLMs. Diferente de *benchmarks* que se concentram apenas na acurácia, HELM avalia múltiplos aspectos do desempenho do modelo, incluindo acurácia, calibração, robustez, justiça (fairness) e eficiência (Aisera, 2024). Essa perspectiva multifacetada é essencial para compreender o comportamento dos LLMs em cenários diversos e complexos, especialmente em aplicações de alto risco (Meva; Kukadiya, 2025).

MMLU

MMLU (Massive Multitask Language Understanding) é um *benchmark* projetado para avaliar o conhecimento de LLMs em uma vasta gama de domínios. Ele abrange 57 áreas diversas, incluindo matemática, humanidades, ciências sociais e STEM (Ciência, Tecnologia, Engenharia e Matemática) (Aisera, 2024; Meva; Kukadiya, 2025).

4.3.2 *Benchmarks* para tarefas específicas

Question Answering (QA)

Aqui trataremos de *benchmarks* para avaliação de QA em larga escala. O Capítulo Perguntas e Respostas trata com bastante atenção a tarefa.

Para tarefas de Question Answering, *benchmarks* como SQuAD (Stanford Question Answering Dataset), *Natural Questions* e TriviaQA são amplamente utilizados (Houamegni; Gedikli, 2025).

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol2/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

O SQuAD é um *benchmark* fundamental, composto por mais de 100.000 pares de perguntas e respostas, onde as respostas são trechos de passagens fornecidas. SQuAD 2.0 introduziu perguntas sem resposta, exigindo que os modelos identificassem quando uma pergunta não podia ser respondida com o contexto dado (Rajpurkar et al., 2016).

Já o *Natural Questions* contém consultas geradas por usuários reais do Google Search, espelhando cenários de busca do mundo real onde as respostas podem não estar explicitamente em um único parágrafo (Kwiatkowski et al., 2019).

Por fim, o TriviaQA é um *dataset* mais desafiador que SQuAD, com 950.000 pares de perguntas e respostas de documentos da Wikipedia e da web, frequentemente exigindo raciocínio multi-hop e recuperação de informações entre documentos (Joshi et al., 2017).

Para atender à necessidade de avaliação de modelos em múltiplos idiomas, foram criadas extensões específicas para os principais *benchmarks* de Q&A. Essas versões mantêm o formato do *dataset* original, mas utilizam textos e perguntas em outras línguas.

- Para o SQuAD: Existem *datasets* como o XQuAD, que é uma tradução humana profissional do SQuAD v1.1 para 10 idiomas, incluindo árabe, grego, tailandês e turco. Outro exemplo é o MLQA (Multilingual Question Answering), que oferece um *benchmark* de extração de respostas em 7 idiomas, como espanhol, hindi e chinês.
- Para o *Natural Questions*: A principal extensão é o MKQA (Multilingual Knowledge Questions and Answers). Ele utiliza as perguntas do Natural Questions e as apresenta em 26 idiomas. As respostas devem ser encontradas nas versões da Wikipedia correspondentes a cada um desses idiomas.
- Para o TriviaQA: Foi desenvolvido o Mling-TriviaQA, que adapta o formato desafiador do original para 10 idiomas diferentes. As perguntas e os documentos de evidência foram coletados ou traduzidos para cada uma das línguas incluídas no *benchmark*.

Sumarização

Para a sumarização de texto, *datasets* como CNN/DailyMail[6] e XSum (Narayan et al., 2018) são referências relevantes. Aqui exploraremos apenas esses *datasets*, mas você pode consultar o Capítulo Sumarização Automática para se aprofundar sobre a tarefa.

O CNN/DailyMail é um *corpus* amplamente utilizado para sumarização de notícias, contendo aproximadamente 300 mil notícias escritas em inglês, acompanhadas dos destaques de cada uma.

O XSum (Extreme Summarization Dataset) é focado em sumarização abstrativa, com resumos mais concisos e frequentemente reescritos, em vez de apenas extraídos do texto original.

Raciocínio

A avaliação das capacidades de raciocínio[7] de uma LLM também pode ser o foco de um *benchmark*.

GSM8K (Grade School Math 8K)

Um *dataset* de 8.500 problemas matemáticos de nível fundamental, linguisticamente diversos, que exigem raciocínio multi-passos e operações aritméticas básicas (II, 2025; Said, 2025).

Big-Bench Hard (BBH)

Um subconjunto de tarefas desafiadoras do BIG-bench, projetado para testar as aptidões do modelo além dos *benchmarks* convencionais de PLN, incluindo raciocínio de alto nível (Meva; Kukadiya, 2025; Said, 2025).

[6] https://huggingface.co/datasets/abisee/cnn_dailymail

[7] Usamos o termo raciocínio de forma operacional, em linha com a literatura recente sobre LLMs. Não se deve entender que tais modelos realizem processos cognitivos análogos ao raciocínio humano, o termo aqui refere-se à capacidade de manipular representações linguísticas e numéricas por meio de correlações estatísticas complexas aprendidas durante o treinamento dos LLMs.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol2/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

4.3.3 *Benchmarks* para segurança e confiabilidade

A segurança e a confiabilidade emergem como dimensões centrais na avaliação contemporânea de LLMs, especialmente em aplicações de alto impacto social, como os domínios jurídico ou médico. Diferentemente dos *benchmarks* voltados apenas à performance ou ao 'raciocínio', esses conjuntos de avaliação buscam medir a robustez ética e factual dos modelos, isto é, sua capacidade de evitar a produção de conteúdo falso, enviesado ou potencialmente nocivo.

Esses *benchmarks* abordam desafios que extrapolam a precisão técnica, como a tendência dos modelos a reproduzirem vieses presentes nos dados de treinamento, a gerarem informações incorretas com aparência de verdade ou a responderem de forma inadequada a *prompts* sensíveis. Assim, a avaliação de segurança e confiabilidade opera em um duplo eixo: veracidade factual e responsabilidade social.

Nos últimos anos, a literatura tem consolidado uma série de *benchmarks* voltados especificamente para essas dimensões, que hoje se tornaram indispensáveis para aferir se um LLM é apto a operar de forma segura e ética. Entre os principais, destacam-se o TruthfulQA, voltado à verificação da veracidade das respostas; o BBQ (*Bias Benchmark for Question Answering*), que mede o grau de viés social reproduzido por modelos; e o ToxiGen e o RealToxicityPrompts, que avaliam a propensão de um sistema a gerar conteúdo tóxico, ofensivo ou discriminatório.

TruthfulQA

Avalia a veracidade das respostas dos modelos, medindo sua resistência à geração de informações falsas ou equívocos comuns (Aisera, 2024; Lin et al., 2022).

BBQ (*Bias Benchmark for Question Answering*)

Mede o viés social em modelos de Question Answering, ajudando a identificar e mitigar preconceitos (Jan et al., 2025; Sánchez, 2024).

ToxiGen/RealToxicityPrompts

Avaliam a propensão dos modelos a gerar texto tóxico ou prejudicial, sendo fundamentais para garantir a segurança do conteúdo gerado (Gehman et al., 2020; Jan et al., 2025).

4.3.4 A importância de ir além do placar

Embora os *benchmarks* e os placares (leaderboards) forneçam uma maneira padronizada de comparar modelos, é crucial ir além da simples pontuação. A saturação de *benchmarks*, nos quais os modelos atingem ou superam o desempenho humano, pode mascarar limitações reais e a falta de generalização da aplicação dos modelos em cenários do mundo real (Meva; Kukadiya, 2025; Wang et al., 2019b). Desafios como a sensibilidade a pequenas mudanças nos *prompts* (*prompt* sensitivity) e a estocasticidade na geração de texto introduzem variabilidade nos resultados, o que torna a reprodutibilidade dos resultados de placares e *benchmarks* para modelos baseados em LLMs um problema mais complexo do que na avaliação de sistemas clássicos (Meva; Kukadiya, 2025). Além disso, a contaminação de dados (*benchmark leakage*), em que dados de avaliação são inadvertidamente incluídos no treinamento, pode inflar artificialmente as métricas de desempenho (Wang et al., 2019b). Portanto, uma análise qualitativa aprofundada e a consideração de múltiplos *benchmarks* são essenciais para uma avaliação holística e precisa.

4.4 Avaliação baseada em modelos (*Model-based evaluation*)

Nessa seção, trataremos do novo paradigma LLM-as-a-Judge, das vantagens, de possíveis abordagens para utilizar esse paradigma, bem como de suas desvantagens e pontos de atenção a serem considerados.

A ideia de utilizar um LLM para avaliar as saídas de outro LLM, conhecida como "LLM-as-a-Judge", emergiu como uma alternativa promissora à avaliação humana, especialmente para textos abertos e complexos (Li et al., 2024; Vongthongsri, 2025). Essa abordagem envolve instruir um LLM (geralmente mais potente ou especificamente treinado para a tarefa de avaliação) a pontuar ou criticar a saída de outro modelo com base em critérios definidos (Li et al., 2024).

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol2/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

4.4.1 Vantagens

Escalabilidade e custo potencialmente menor

A avaliação humana é um processo notoriamente caro e demorado, o que a torna impraticável para avaliações contínuas ou em larga escala. LLMs como avaliadores oferecem uma solução mais escalável e potencialmente mais econômica para essa limitação (Li et al., 2024; Vongthongsri, 2025).

Capacidade de fornecer *feedback* em linguagem natural

Diferente das métricas automáticas tradicionais que fornecem apenas pontuações numéricas, os LLMs avaliadores podem gerar *feedback* detalhado em linguagem natural, explicando o porquê de uma determinada pontuação ou crítica. Isso pode ser útil para depurar e melhorar os modelos, se utilizado dentro de um fluxo híbrido, no qual esses *feedbacks* sejam realmente consumidos, porém o simples fato de uma LLM gerar um feedback, uma explicação sobre sua resposta, não garante que a resposta esteja correta.

Potencial para avaliar aspectos mais abstratos

LLMs avaliadores têm o potencial de avaliar aspectos mais abstratos da qualidade do texto, como coerência, relevância, tom e criatividade, que são difíceis de capturar com métricas clássicas (Vongthongsri, 2025).

Avaliação de interações dinâmicas e contextuais

Uma vantagem crucial é a capacidade de avaliar sistemas de IA da mesma forma que são utilizados. Por exemplo, em vez de depender de *datasets* estáticos, que são inadequados para medir a qualidade de um diálogo, um LLM avaliador pode analisar uma conversa em tempo real. Ele pode avaliar a interação passo a passo ou o diálogo completo, proporcionando uma avaliação muito mais fiel à experiência real do usuário em um *chatbot* ou assistente virtual[8].

4.4.2 Abordagens

Avaliação baseada em referência vs. sem referência

Algumas abordagens de LLM-as-a-Judge podem operar com ou sem textos de referência. Métricas baseadas em *embeddings* de LLMs, por exemplo, frequentemente não exigem referências, calculando a similaridade semântica diretamente entre a saída do modelo e a consulta do usuário ou o contexto de uso. No entanto, a avaliação sem referência pode ser mais suscetível a falhas e vieses.

Prompting para avaliação

A criação de *prompts* eficazes é fundamental. Esses *prompts* instruem o LLM avaliador sobre os critérios de avaliação, o formato da saída (e.g., escalas Likert, comparações binárias, análise de erros) e o texto a ser avaliado (Li et al., 2024). A qualidade do *prompt* pode influenciar drasticamente a avaliação.

Fine-tuning de LLMs para tarefas de avaliação

Além do *prompting*, LLMs podem ser especializados (fine-tuned) especificamente para tarefas de avaliação, treinando-os para prever pontuações humanas ou para identificar tipos específicos de erros.

LLM-as-a-Jury (LLM como um júri)

Para aumentar a robustez e mitigar os vieses de um único avaliador, esta abordagem utiliza um painel de múltiplos LLMs para julgar uma mesma saída. A tarefa é enviada a vários "LLMs juízes", que podem ser modelos diferentes ou instâncias do mesmo modelo com configurações distintas. As avaliações individuais são então agregadas (por exemplo, através de uma votação majoritária ou pela média das pontuações)

[8] Atualmente, também têm sido propostos *frameworks* dinâmicos para a análise de agentes baseados em LLMs, como o clembench (Chalamalasetti et al., 2023), cujo diferencial é ser uma avaliação interativa diretamente motivada pela Teoria do Diálogo, para avaliar como fenômenos do diálogo humano são ou não reproduzidos pelos LLMs.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol2/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

para chegar a um veredito de consenso. Essa técnica visa replicar a dinâmica de um painel de especialistas humanos, resultando em uma avaliação final mais estável e confiável.

Avaliação por decomposição de tarefas

Esta abordagem quebra a avaliação de uma tarefa complexa em subtarefas mais simples e verificáveis. Por exemplo, para avaliar a qualidade de um resumo, em vez de julgá-lo diretamente, o processo poderis ser: 1) Gerar um conjunto de perguntas e respostas com base no documento original (fonte). 2) Pedir a um LLM que responda às mesmas perguntas usando apenas o resumo gerado. 3) A avaliação final, comparar as respostas obtidas a partir do resumo com as respostas obtidas a partir do documento original, medindo assim a fidelidade e a retenção de informação do resumo.

4.4.3 Desafios e Limitações

Apesar das vantagens, a avaliação baseada em modelos apresenta desafios significativos.

Viés do modelo avaliador

O LLM avaliador pode herdar ou amplificar vieses presentes em seus dados de treinamento, ou até mesmo exibir um "viés narcisista", favorecendo respostas geradas por si mesmo (Panickssery et al., 2024) ou por modelos da mesma família (Vongthongsri, 2025). Estudos mostram que avaliadores generativos tendem a atribuir pontuações mais altas ao conteúdo gerado pelo mesmo modelo subjacente.

Sensibilidade ao *prompt*

Pequenas alterações na formulação do *prompt* de avaliação podem levar a diferenças substanciais nos resultados, tornando a avaliação inconsistente e não robusta (Zhao et al., 2025). Para lidar com essa questão, um método chamado PromptEval (Polo et al., 2024) foi proposto. A abordagem reconhece que avaliações baseadas em um número limitado de *prompts* podem não capturar as reais habilidades dos LLMs, levando a classificações inconsistentes. Em vez de depender de um único "*prompt*", o PromptEval estima a distribuição de desempenho do modelo através de um grande conjunto de variações de "*prompts*". Isso permite a criação de métricas de avaliação mais robustas, como a performance mediana ou em quantis específicos (por exemplo, o desempenho nos 5% piores *prompts* ou nos 5% melhores), fornecendo um panorama mais completo e confiável da capacidade do modelo, que é menos suscetível a alterações em um "*prompt*" individual.

Concordância com humanos

Embora alguns estudos sugiram que LLMs podem atingir alta concordância com julgamentos humanos (até 85% para GPT-4, por exemplo) (Bencke et al., 2024; Kim et al., 2025; Vongthongsri, 2025), outros indicam que a acurácia da avaliação por LLMs não é garantida em todos os cenários (Pombal et al., 2025; Wang et al., 2024). A correlação pode variar dependendo da tarefa e do modelo avaliador. No entanto, a discussão é mais complexa, pois replicar o julgamento humano pode significar replicar também seus vieses. Um estudo sobre o tema ilustra essa questão, mostrando que, embora juízes humanos demonstrem menos viés de gênero que os LLMs, ambos são igualmente vulneráveis ao viés de autoridade (favorecendo respostas com referências falsas), ao "viés de beleza" (preferindo textos mais bem formatados) e à dificuldade em detectar desinformação (Chen et al., 2024a). Isso demonstra que o verdadeiro desafio não é apenas alcançar a concordância, mas desenvolver métodos de avaliação que superem as falhas inerentes tanto a humanos quanto a máquinas.

Custo computacional

Embora potencialmente mais barato que a avaliação humana em larga escala, o uso de LLMs potentes como juízes ainda pode incorrer em custos computacionais significativos, especialmente para avaliações contínuas (Meva; Kukadiya, 2025), avaliações em que uma única instância é avaliada usando diferentes critérios ou no uso de LLM-as-a-Jury, o uso de diferentes modelos para avaliar o mesmo sistema.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol2/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Janela de contexto finita

Os modelos de linguagem possuem um limite na quantidade de texto (contexto) que conseguem processar de uma só vez. Essa restrição se torna um grande obstáculo ao avaliar tarefas que envolvem textos longos, como a sumarização de documentos extensos ou a análise de um histórico de diálogo completo. A necessidade de truncar ou dividir o texto para que ele caiba na janela de contexto pode levar à perda de informações e prejudicar a capacidade do LLM de realizar uma avaliação holística e precisa.

Não determinismo

As pontuações de LLMs juízes podem ser não determinísticas, ou seja, a mesma saída pode receber pontuações diferentes em avaliações repetidas, a menos que sejam aplicadas técnicas para garantir (algum) determinismo (Vongthongsri, 2025).

Viés de posição e verbosidade

LLMs avaliadores frequentemente exibem um "viés de posição", preferindo a primeira saída, ou eventualmente a segunda saída, em comparações pareadas, e um "viés de verbosidade", preferindo textos mais longos, mesmo que não sejam necessariamente de maior qualidade (Chen et al., 2024a).

Viés de leniência (*Leniency bias*)

Este fenômeno é caracterizado pela propensão dos modelos em avaliar respostas como "corretas" em situações de incerteza, ou quando seus critérios de avaliação não estão perfeitamente alinhados com as instruções fornecidas. Um estudo de 2024 quantifica essa tendência ao estimar a probabilidade de um juiz marcar uma resposta como correta quando, na verdade, ele não está aplicando os critérios de avaliação corretamente (Thakur et al., 2025). Os resultados demonstram que a maioria dos modelos juízes exibe um viés de leniência consideravelmente alto, indicando que, na dúvida, eles tendem a julgar positivamente. Essa propensão é mais pronunciada em modelos menores.

Outros vieses

Vários outros tipos de vieses relacionados foram identificados mais recentemente (Kamruzzaman et al., 2024; Park et al., 2024) como *Length Bias* (viés de extensão), *Concreteness Bias* (a tendência de atribuir maior credibilidade a respostas que contêm detalhes específicos, como citações, números e terminologias complexas), *Empty Reference Bias* (o modelo prefere conteúdo alucinado que apenas parece estar associado à instrução original), o Content Continuation Bias (o modelo favorece respostas que continuam o texto de entrada em vez de seguir a instrução principal) e o *Nested Instruction Bias* (tendência a responder a perguntas aninhadas dentro do texto, ignorando a instrução principal).

4.4.4 Exemplos de *frameworks* de *LLM as a Judge*

G-Eval

Um método que gera uma série de passos de avaliação a partir de critérios definidos, utilizando um LLM para preencher um formulário e determinar a pontuação final (Vongthongsri, 2025).

Prometheus

Uma suíte de LLMs avaliadores de código aberto, otimizados para avaliação multilíngue, que podem fornecer avaliação direta e comparação pareada (Kim et al., 2023).

AutoEval

Abordagens que visam automatizar a avaliação, como a investigação do uso de LLMs como juízes para questões abertas legais, mostrando alta correlação com pontuações humanas em exames padronizados (Samuylova, 2025).

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol2/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Avaliação jurídica baseada em QA

Este método avalia resumos jurídicos usando um LLM em um processo de Pergunta e Resposta (QA). O LLM primeiro cria perguntas com base em um resumo de referência e depois usa o resumo candidato para respondê-las. Finalmente, o modelo compara as duas respostas para gerar uma pontuação, focando na qualidade da argumentação legal em vez de apenas na sobreposição de palavras (Xu; Ashley, 2023).

DeepEval

É um *framework* de código aberto para avaliação de LLMs[9], frequentemente descrito como o "Pytest para LLMs", por aplicar uma abordagem de testes unitários à avaliação de saídas de modelos de linguagem. Ele oferece uma suíte com mais de 14 métricas, como G-Eval, RAGAS, relevância da resposta e detecção de alucinações. O DeepEval se integra a outras ferramentas do ecossistema, como LangChain e LlamaIndex, e é complementado pela plataforma em nuvem Confident AI[10], que adiciona funcionalidades de gerenciamento de *datasets* e detecção de regressão nos testes.

4.5 Aspectos emergentes e avançados da avaliação

Como já comentado, com as LLMs, surgem também novos desafios para entender como eles realmente funcionam e se comportam. Já não basta medir apenas a precisão ou a fluência das respostas, avaliar se os modelos são robustos, justos, confiáveis e eficientes também torna-se necessário. Nesta seção, exploramos essas novas dimensões da avaliação, trazendo, brevemente, exemplos de possíveis formas de detecção e mensuração desses desafios.

4.5.1 Robustez e Ataques adversariais

A robustez de um LLM refere-se à sua capacidade de manter o desempenho e a segurança diante de entradas ruidosas, com erros ortográficos, ou maliciosamente construídas (ataques adversariais). A avaliação da robustez é crucial para a implantação segura e responsável de LLMs em domínios de alto risco, onde erros podem ter consequências graves. Ataques adversariais podem incluir:

- Perturbações em nível de caractere: Alterações sutis nos *prompts* que podem induzir o modelo a comportamentos indesejados (Zhao et al., 2025).
- *Jailbreak prompts*: *Prompts* deliberadamente elaborados para contornar os mecanismos de segurança dos LLMs e induzi-los a gerar conteúdo prejudicial. Novos *frameworks* estão sendo desenvolvidos para avaliar autonomamente a robustez dos LLMs, utilizando *prompts* adversariais refinados e diretrizes de conhecimento restritas ao domínio, muitas vezes na forma de grafos de conhecimento.

4.5.2 Avaliação de justiça, viés e toxicidade

A avaliação de justiça (fairness), viés (bias) e toxicidade é um campo em crescimento, impulsionado pela necessidade de garantir que os LLMs não perpetuem ou amplifiquem preconceitos presentes em seus dados de treinamento (Japkowicz; Boukouvalas, 2024; Jiao et al., 2024).

A identificação e a quantificação de vieses em LLMs têm-se tornado uma frente essencial da pesquisa contemporânea. Diversas ferramentas e técnicas vêm sendo desenvolvidas para identificar e mensurar diferentes tipos de vieses, como de gênero, raça ou culturais, presentes nas saídas desses modelos (Jan et al., 2025; Japkowicz; Boukouvalas, 2024; Kamruzzaman et al., 2024). *Benchmarks* específicos, como o BBQ (*Bias Benchmark for Question Answering*) e o ToxiGen, têm sido utilizados para avaliar tanto a reprodução de estereótipos sociais quanto a propensão dos modelos a gerar discurso de ódio (Gehman et al., 2020; Jan et al., 2025).

A própria definição de viés é um desafio. Ela varia conforme o contexto cultural e exige cuidado para distinguir entre padrões legítimos do mundo real e estereótipos prejudiciais (Meva; Kukadiya, 2025). Para enfrentar essa complexidade, novas abordagens vêm sendo propostas, como o *framework de auto-coerência*, no qual os próprios modelos analisam seus padrões de saída em busca de inconsistências e vieses implícitos (Japkowicz; Boukouvalas, 2024). Ainda assim, a mitigação efetiva desses problemas requer

[9] https://github.com/confident-ai/deepeval
[10] https://www.confident-ai.com/

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol2/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

monitoramento contínuo e o uso de técnicas de debiasing, voltadas à correção progressiva das distorções detectadas (Sánchez, 2024).

Uma estratégia frequentemente explorada é o uso de *datasets* sintéticos para tentar corrigir vieses complexos. No entanto, essa solução pode introduzir novos riscos metodológicos: em vez de eliminar o viés original, o modelo avaliador pode aprender a reproduzir os mesmos padrões ou artefatos (artifacts) gerados pelo modelo que criou o *dataset* sintético. O resultado é a introdução de um novo tipo de viés, que compromete a validade da avaliação e revela uma limitação importante das atuais estratégias de mitigação baseadas em dados sintéticos (Chen et al., 2024a).

4.5.3 Factualidade, alucinações e citabilidade

A factualidade e a prevenção de "alucinações" são desafios críticos para os LLMs (Meva; Kukadiya, 2025). Métodos para verificar a veracidade da informação gerada incluem:

Detecção de alucinações

Técnicas como FactSelfCheck permitem a detecção de alucinações em nível de fato, analisando a consistência factual entre múltiplas respostas do LLM (Sawczyn et al., 2025).

Citabilidade e Atribuição de fontes

A capacidade de um LLM de citar corretamente suas fontes é fundamental para a confiabilidade. Métodos como LLM-Cite exploram a capacidade do LLM de gerar URLs de citação potenciais para uma determinada afirmação, que são então verificadas em tempo real (Joshi et al., 2025). No entanto, a atribuição de fontes ainda enfrenta desafios, especialmente em domínios específicos, como o jurídico (Zhang et al., 2025b).

Avaliação de fundamentação (*Groundedness*) via NLI

Essa abordagem avalia se uma afirmação gerada por um LLM é factualmente suportada por uma fonte específica, usando a Inferência de Linguagem Natural (NLI) para classificar a relação como corroboração (acarretamento) ou contradição. O trabalho de (Trautmann et al., 2024) mostra que, embora métodos híbridos complexos (como combinar o uso de LLM-as-a-Judge com métodos de NLI) sejam os mais precisos, são também muito caros e lentos. O trabalho destaca que uma abordagem mais simples e eficiente, como perguntar diretamente a um LLM potente (e.g. GPT-4) se a fonte suporta a afirmação, já oferece uma performance robusta, servindo como um forte *baseline*.

4.5.4 Eficiência computacional e custos

A eficiência computacional é uma dimensão cada vez mais importante na avaliação de LLMs, dada a sua crescente escala e os custos associados à inferência e implantação (Galileo AI, 2024; Meva; Kukadiya, 2025).

Métricas

As métricas de eficiência incluem:

1. Latência: O tempo entre a submissão de um *prompt* e o recebimento da resposta completa (Google Cloud, 2025).
2. Throughput: O volume de requisições ou *tokens* que o sistema pode processar por unidade de tempo (Google Cloud, 2025).
3. Consumo de Energia: A energia gasta durante a inferência, um fator crucial para a sustentabilidade (Meva; Kukadiya, 2025);
4. Custo de Inferência: O custo monetário associado à execução do modelo (Meva; Kukadiya, 2025).

A otimização dessas métricas é vital para a viabilidade de aplicações práticas, especialmente aquelas que exigem respostas em tempo real ou operam em larga escala (Galileo AI, 2024; Google Cloud, 2025). Técnicas como test-time compute (TTC) e otimizações de arquitetura visam melhorar a eficiência sem comprometer a acurácia (Meva; Kukadiya, 2025).

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol2/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

4.6 Considerações éticas e o Futuro da avaliação de LLMs

À medida que os LLMs passam a ocupar papéis centrais em produtos, pesquisas e até decisões governamentais, a avaliação deixa de ser apenas uma questão técnica e se torna também uma questão ética e social. Avaliar modelos implica também em discutir responsabilidade, transparência e possíveis impactos negativos de seu uso. Nesta seção, exploramos como a avaliação deve incorporar princípios de responsabilidade, interpretabilidade e regulação, preparando o caminho para práticas mais seguras e sustentáveis no futuro da inteligência artificial.

Aqui trataremos especificamente da questão de uso de LLMs para a avaliação, porém você pode ver a Seção Uso Responsável e Boas Práticas e o Capítulo Questões éticas em IA e PLN deste livro para mais discussões sobre o uso responsável de IA.

4.6.1 Responsabilidade na avaliação

A implantação generalizada de LLMs levanta questões críticas sobre a responsabilidade quando esses modelos falham ou causam danos. Robustos *frameworks* de avaliação são essenciais para verificar se um LLM atende a padrões de confiabilidade, ética e desempenho em aplicações reais. A avaliação deve, portanto, ser uma parte integral do ciclo de vida do desenvolvimento do LLM, e não apenas uma análise *post-hoc* (Meva; Kukadiya, 2025).

4.6.2 Transparência e Interpretabilidade

A "caixa preta" dos LLMs, em que as decisões e as saídas são difíceis de interpretar, é um desafio significativo. A necessidade de entender por que um modelo toma certas decisões ou gera certas saídas é fundamental para a confiança e a responsabilidade social (Sánchez, 2024). A transparência contextual, por exemplo, pode envolver a inclusão de disclaimers sobre a natureza probabilística das respostas ou a indicação das fontes utilizadas.

4.6.3 Dualidade do uso

A capacidade dos LLMs de gerar conteúdo complexo e convincente também levanta preocupações sobre seu uso malicioso. A "dualidade do uso" refere-se ao potencial de modelos, originalmente projetados para fins benéficos, serem manipulados para gerar conteúdo prejudicial, como discurso de ódio, golpes ou até mesmo código malicioso para ataques cibernéticos. A avaliação deve, portanto, incluir a análise da suscetibilidade dos modelos a tais manipulações e o desenvolvimento de estratégias de mitigação de seu uso malicioso.

4.6.4 O papel da regulação e de padrões na avaliação

A rápida evolução dos LLMs exige o desenvolvimento de padrões e regulamentações para guiar sua avaliação e implantação responsáveis. Há uma necessidade crescente de protocolos de avaliação mais padronizados e transparentes para permitir comparações significativas e rastrear o progresso ao longo do tempo (Meva; Kukadiya, 2025; National Institute of Standards and Technology, 2024). A privacidade dos dados, a reprodutibilidade dos resultados e a capacidade de realizar avaliações sem utilizar recursos de terceiros (*on-premise*) são considerações importantes em contextos industriais ou governamentais (National Institute of Standards and Technology, 2024).

4.6.5 Direções futuras

O futuro da avaliação de LLMs é definido tanto por oportunidades de avanço quanto por desafios fundamentais que permanecem não resolvidos. As direções prioritárias incluem:

Estabelecimento de meta-avaliação

(Evaluating the Evaluators) Talvez a direção mais crítica seja a investigação sobre a própria validade das métricas e dos *frameworks* de avaliação. A questão de "quando uma avaliação é boa o suficiente?" ainda carece de uma resposta definitiva. A pesquisa futura deve se concentrar no desenvolvimento de métodos rigorosos para validar os próprios sistemas de avaliação, garantindo que suas conclusões sejam robustas, replicáveis e verdadeiramente indicativas do desempenho do modelo.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol2/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Avaliação contínua e em tempo real

Existe uma lacuna significativa entre a avaliação em ambientes de pesquisa e as necessidades da indústria (Heller, 2024). A implementação de *frameworks* que permitam a avaliação contínua e em tempo real é uma necessidade urgente para monitorar o desempenho de modelos em produção, detectar degradações (*drifts*) e garantir a segurança. Superar os obstáculos técnicos e de custo para viabilizar essa monitorização é um dos principais focos para a aplicação prática e responsável de LLMs (Meva; Kukadiya, 2025).

Avaliação adaptativa e personalizada

Métricas e abordagens que se adaptam a diferentes domínios, casos de uso e preferências do usuário (Meva; Kukadiya, 2025; PrototypeJam, 2025).

Desenvolvimento de métricas alinhadas com valores humanos

O objetivo de alinhar LLMs com "valores humanos" é mais complexo do que simplesmente buscar concordância com avaliadores humanos. A pesquisa deve abordar a distinção filosófica entre mimetizar as preferências e os vieses de um grupo de anotadores e alinhar-se a princípios éticos mais amplos e universais. Isso envolve não apenas o desenvolvimento de métricas mais sofisticadas, mas também uma reflexão sobre quais valores devem ser codificados e como (Meva; Kukadiya, 2025).

Avaliação de impacto social a longo prazo

Uma avaliação ampla, especialmente tratando-se do uso de IA em ambientes governamentais, deve considerar o impacto social e ético dos LLMs em um horizonte de tempo mais longo (Meva; Kukadiya, 2025; National Institute of Standards and Technology, 2024).

Abordagem multidisciplinar

A complexidade dos LLMs exige uma abordagem que transcende a ciência da computação. A avaliação deve ser inerentemente multidisciplinar, integrando insights da linguística, psicologia, ética e ciências sociais para criar uma compreensão holística do desempenho técnico dos modelos e do seu impacto social a longo prazo. Isso inclui o desenvolvimento de avaliações adaptativas, que se ajustem a diferentes contextos culturais, domínios de aplicação e casos de uso específicos (Jiao et al., 2024; Meva; Kukadiya, 2025).

4.7 Considerações finais

Neste capítulo, abordamos a avaliação de Grandes Modelos de Linguagem (Large Language Models), um campo dinâmico e multifacetado, que exige ir além das abordagens convencionais da linguística computacional e do processamento de linguagem natural. O avanço desses modelos trouxe novos desafios metodológicos, éticos e epistemológicos, forçando a revisão de pressupostos sobre o que significa avaliar o desempenho em sistemas que produzem linguagem de forma aberta e contextual.

As métricas automáticas tradicionais continuam a oferecer um ponto de partida valioso, sobretudo pela rapidez e pela comparabilidade que proporcionam. No entanto, sua limitação em capturar aspectos semânticos e, especialmente, pragmáticos torna indispensável a incorporação de camadas adicionais de análise. A avaliação humana, ainda que cara e limitada em escala, permanece o padrão-ouro por sua capacidade de julgar nuances, contextos e valores universais e éticos.

Nesse novo cenário, o uso de modelos como avaliadores, o paradigma LLM-as-a-Judge, representa uma via promissora. Ao unir a eficiência da automação com a profundidade interpretativa dos julgamentos humanos, essas abordagens híbridas apontam para um futuro de avaliações mais contínuas, contextualizadas e adaptativas. No entanto, desenvolver um LLM-juiz justo e eficiente demanda tempo, meta-avaliação e supervisão humana constante. Além disso, o LLM-juiz não é garantia de alinhamento da avaliação a valores humanos, ele é apenas um método 'mais alinhado' do que métodos tradicionais.

O desafio atual é consolidar *frameworks* eficientes, robustos e rápidos capazes de equilibrar essas três dimensões: automação, julgamento humano e meta-avaliação por modelos, de forma ética, transparente e alinhada a valores humanos. Somente assim será possível garantir que a avaliação de LLMs cumpra não apenas sua função técnica, mas também seu papel social: orientar o desenvolvimento de sistemas de linguagem que sejam realmente úteis, justos e responsáveis. O assunto parece muito complexo e é: a avaliação de sistemas de inteligência artificial precisa ser factível, confiável, tecnicamente viável e socialmente

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

responsável. Da mesma forma que o estado da arte da arquitetura de modelos evolui e continua evoluindo, o estado da arte de como avaliá-los está em constante evolução.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol2/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Capítulo 5

Geração Aumentada por Recuperação (RAG)

André Carvalho
Livy Real
Altigran Soares da Silva

Publicado em: 16/04/2026

5.1 Fundamentos da Geração Aumentada por Recuperação

Grandes Modelos de Linguagem, ou ainda, Modelos de Linguagem de Grande Escala (LLMs), representam um marco na Inteligência Artificial, demonstrando uma capacidade sem precedentes de 'compreender', gerar e interagir usando a linguagem humana. No entanto, apesar de sua sofisticação, esses modelos são inerentemente limitados por sua arquitetura e pelo processo de treinamento. A Geração Aumentada por Recuperação (RAG) surge não como um aprimoramento incremental, mas um paradigma de uso dos LLMs em resposta às limitações dos modelos de fundação (*foundation models*)[1] Para compreender a importância da RAG, é crucial primeiro dissecar os quatro principais desafios que ela se propõe a resolver:

- Conhecimento Desatualizado (*Knowledge Cutoff*): O treinamento de um LLM é um processo computacionalmente intensivo e caro, que consome vastos recursos e tempo. Consequentemente, o conhecimento codificado em um modelo é estático e congelado no tempo, correspondendo à data em que seus dados de treinamento foram coletados. Esse "ponto de corte de conhecimento" (*knowledge cutoff*) significa que para dar resultados relativos a certos temas a LLM precisa de conteúdos sobre os eventos subsequentes ao seu treinamento, tornando-o inadequado para aplicações que exigem informações atuais ou em tempo real. (Cheng et al., 2024)
- Alucinações: Um dos riscos mais significativos associados aos LLMs é sua tendência a "alucinar", ou seja, gerar informações que são plausíveis, bem escritas e contextualmente adequadas, mas factualmente incorretas ou completamente fabricadas. Isso ocorre porque os LLMs são modelos probabilísticos otimizados para gerar a sequência de palavras mais provável, não a mais verdadeira. Deste modo, sem o controle necessário, os *foundation models* tendem a confeccionar fatos, combinar informações e apresentar referências inválidas, gerando respostas potencialmente divergentes da realidade (Kalai et al., 2025). Em domínios críticos como saúde, finanças ou direito, as alucinações representam um risco inaceitável para a confiança e a segurança do uso desses modelos. (Huang et al., 2024b)
- Falta de Especificidade de Domínio: LLMs de propósito geral, treinados em dados disponíveis publicamente, carecem de conhecimento profundo sobre domínios especializados ou dados proprietários de uma organização. Assim, em nível de aplicação, a especificidade de domínio é responsável por delimitar o tema, caracterizando quais detalhes são pertinentes ao contexto. Os LLMs desconhecem documentos internos, bases de dados de clientes, manuais técnicos específicos ou a terminologia de um nicho de mercado. Sem acesso a essa informação, suas respostas permanecem genéricas e de valor limitado para casos de uso empresariais.
- Falta de Verificabilidade (Black Box): A natureza de "caixa-preta" dos LLMs torna extremamente difícil, se não impossível, rastrear a origem de uma informação específica gerada. Essa opacidade impede a verificação de fatos, a atribuição de fontes e a auditoria das respostas do modelo — requisitos cruciais em ambientes que exigem precisão como o acadêmico, o jornalístico e o corporativo, onde a confiabilidade e a transparência são primordiais. (Lewis et al., 2020b)

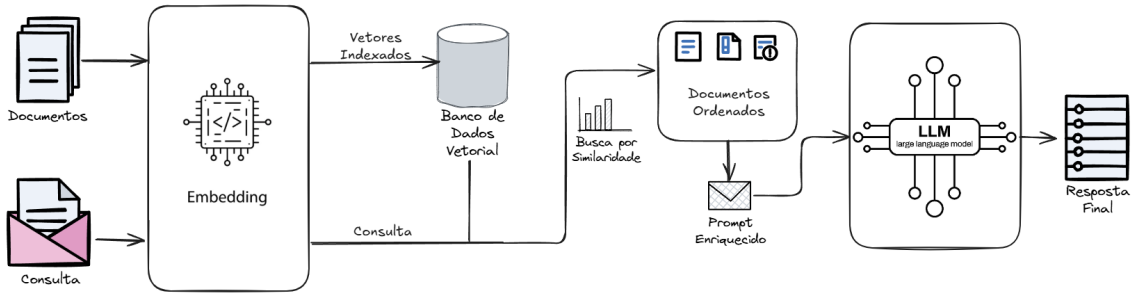
[1] O termo "modelos de fundação" foi criado pelo *Stanford Institute for Human-Centered Artificial Intelligence* (HAI) https://hai.stanford.edu/news/introducing-center-research-foundation-models-crfm em 2021, para designar os algoritmos de redes neurais profundas treinados para atender tarefas gerais.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol2/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

5.1.1 Introdução à RAG

Figura 5.1: Exemplo de arquitetura RAG usando vetores de *embeddings*.



A Geração Aumentada por Recuperação (RAG) (Lewis et al., 2020b) é um paradigma de integração entre modelos de linguagem e sistemas de recuperação de informação que aprimora a capacidade dos LLMs ao conectá-los a conteúdo externo e autoritativo antes de gerar uma resposta. Em sua essência, a RAG transforma o LLM de um sistema de "livro fechado", que depende exclusivamente de sua memória interna (os parâmetros aprendidos durante o treinamento), para um sistema de "livro aberto", que pode consultar fontes externas em tempo real para formular suas respostas.

Essa mudança de paradigma é sutil, mas profunda. Em vez de tentar incluir e adaptar novos conhecimentos como pesos de um modelo, algo que não garante o efetivo aprendizado de domínio, a RAG desacopla o processo de raciocínio linguístico do processo de aquisição de conhecimento. Enquanto o LLM mantém a função principal de gerar conteúdo, a tarefa de selecionar material relevante é delegada ao módulo de recuperação de informação. Deste modo, o módulo de geração compartilha a função de identificar e alinhar o contexto relevante com um módulo especializado nesta tarefa.

Esta abordagem acaba por atacar as principais limitações do uso de LLMs até então, como mostrado no Quadro 5.1.

Quadro 5.1: Limitações dos LLMs e a abordagem RAG.

Limitação do LLM	**Descrição da limitação**	**Como a RAG aborda**
Conhecimento desatualizado	LLMs têm uma "data de corte" e desconhecem eventos recentes.	Recupera informações atuais de bases de conhecimento externas.
Alucinações	Geram informações plausíveis, mas factualmente incorretas.	Ancora a resposta do LLM em fatos recuperados.
Falta de especificidade	Carecem de conhecimento de domínios específicos ou dados proprietários.	Conecta o LLM a bases de conhecimento especializadas.
Falta de verificabilidade	É difícil rastrear a fonte da informação gerada.	Permite fazer referencia às fontes a partir dos documentos recuperados.

A análise do termo *"Retrieval-Augmented Generation"* revela as três etapas sequenciais que definem seu funcionamento:

1. **Recuperação (*Retrieval*)**: Dada uma consulta do usuário, o sistema primeiro aciona um mecanismo de busca para encontrar e extrair os trechos de informação mais relevantes de um *corpus* de conhecimento externo. Este *corpus* pode ser a internet, uma coleção de documentos proprietários, um banco de dados ou qualquer outra fonte de informação.
2. **Aumentada (*Augmentation*)**: As informações recuperadas são então usadas para "aumentar" o *prompt* original do usuário. Tipicamente, o contexto recuperado é inserido, descrito e estruturtado

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol2/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

em um *template* de *prompt* juntamente com a pergunta, fornecendo ao LLM os fatos necessários para formular uma resposta informada.

3. **Geração (*Generation*)**: Finalmente, o *prompt* aumentado é enviado ao LLM. O modelo, então, sintetiza uma resposta coerente e em linguagem natural, baseando-se tanto no contexto factual recém-fornecido quanto em seu conhecimento pré-treinado sobre linguagem, estrutura e raciocínio geral.

Essa arquitetura de dois estágios (recuperação e geração) permite que os sistemas RAG produzam respostas que não são apenas fluentes e contextualmente apropriadas, mas também factualmente precisas, atualizadas e verificáveis.

Contudo, os benefícios da RAG transcendem a melhoria técnica da qualidade das respostas, oferecendo vantagens estratégicas que impulsionam sua adoção em ambientes corporativos. A arquitetura RAG permite que as organizações aproveitem o poder dos LLMs de forma mais segura, econômica e personalizada.

- Melhora da Precisão e Redução de Alucinações: Ao ancorar as respostas em conteúdo recuperado, a RAG aumenta drasticamente a factualidade e mitiga o risco de o LLM gerar informações incorretas. A adoção da arquitetura torna a resposta potencialmente fundamentada em uma fonte específica, em vez de depender apenas do conhecimento parametrizado do modelo.
- Acesso a Informações Atualizadas: A base de conhecimento de um sistema RAG pode ser atualizada continuamente e em tempo real. Isso permite que o LLM responda a perguntas sobre eventos recentes ou conteúdos dinâmicos, como o status de um pedido ou as últimas notícias do mercado, superando as limitações impostas pelo uso de conhecimento desatualizado (*knowledge cutoff*).
- Confiabilidade e Verificabilidade: Como as respostas são baseadas em documentos específicos, os sistemas RAG podem facilmente citar suas fontes. Essa característica permite identificar e quantificar quais são os aspectos positivos e negativos na resposta. Assim, ao rastrear a origem da informação, a RAG aumenta a confiança do usuário e permite a auditoria e verificação dos fatos, algo essencial em aplicações que exigem segurança e responsabilidade.
- Custo-Benefício: Manter um LLM atualizado através de re-treinamento ou fine-tuning contínuo é proibitivamente caro e demorado. Em contraste, atualizar o conteúdo usado em um sistema RAG é uma operação de baixo custo e alta velocidade. Isso torna a RAG uma abordagem muito mais econômica para manter o conhecimento do sistema atualizado.
- Personalização e Segurança de Dados: A RAG é um método bastante adequado para adaptar LLMs a domínios específicos usando dados proprietários. As informações sensíveis de uma organização (dados de clientes, documentos internos, propriedade intelectual) permanecem em uma base de conhecimento controlada e segura, sendo acessadas apenas no momento da consulta, sem nunca serem usadas para treinar o modelo base de terceiros. Isso oferece um forte controle sobre a privacidade e a segurança dos dados, especialmente se os modelos usados forem internos.

A confluência desses benefícios estratégicos resultou em uma rápida e ampla adoção da RAG no mercado. Embora seja difícil obter uma estatística única e precisa sobre o uso da RAG, as tendências em tecnologias adjacentes e a adoção de produtos baseados em RAG pintam um quadro claro de sua importância. Por exemplo, a Databricks, que atende mais de 300 empresas da Fortune 500, lista das maiores empresas dos Estados Unidos, relatou que, em sua plataforma no último ano, houve um crescimento de 377% no uso de bancos de dados vetoriais[2], um componente central da RAG que visitaremos em breve. A adoção massiva de copilotos empresariais, como o Microsoft 365 Copilot, por quase 70% das empresas da Fortune 500[3], demonstra a demanda por sistemas que, como a RAG, conectam LLMs a dados corporativos internos. Esses indicadores mostram que a RAG não é uma tecnologia de nicho, mas sim um pilar fundamental da estratégia de IA gerativa em grandes empresas, impulsionada pela necessidade crítica de personalizar LLMs com dados proprietários de forma segura e econômica.

5.1.2 Arquitetura RAG

No nível mais fundamental, todos os sistemas RAG são compostos por dois pilares arquitetônicos: o *Recuperador* e o *Gerador*. A simplicidade dessa arquitetura modular é uma de suas maiores forças, permitindo que cada componente seja otimizado de forma independente.

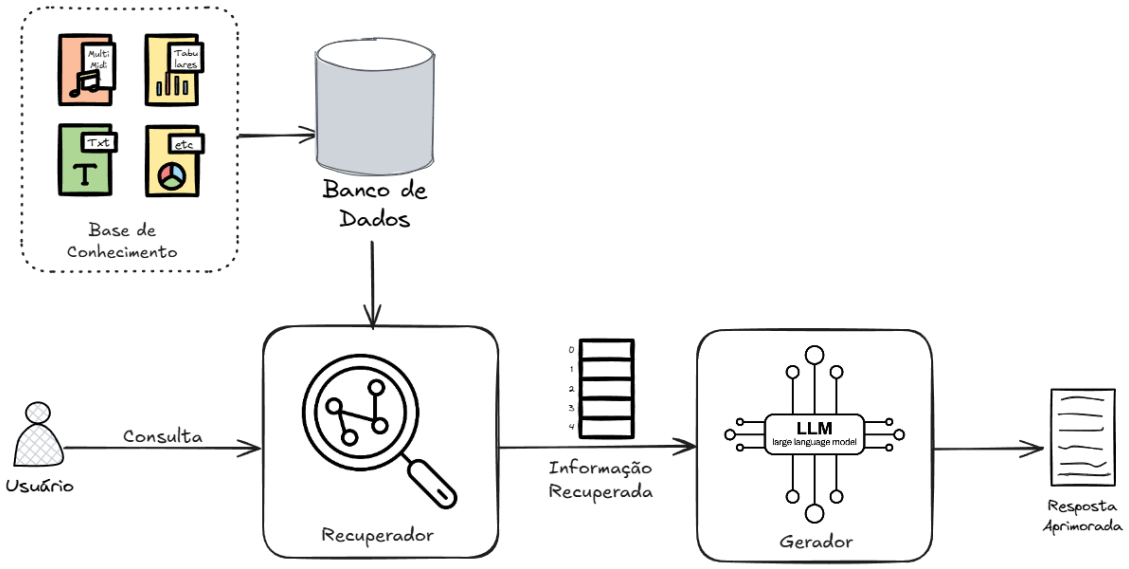
[2] https://www.databricks.com/discover/state-of-data-ai

[3] https://news.microsoft.com/source/latam/features/ia-pt-br/ignite-2024-por-que-quase-70-da-fortune-500-agora-usam-o-microsoft-365-copilot/

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol2/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Figura 5.2: Arquitetura Recuperador-Gerador em Sistemas RAG.



O **Recuperador** (*Retriever*) funciona como uma máquina de busca especializada. Sua principal responsabilidade é: dada uma consulta do usuário, encontrar e retornar a informação mais relevante de uma base de conhecimento pré-indexada. O recuperador atua como um filtro de conhecimento, navegando por um vasto repositório de dados para selecionar apenas alguns elementos que são cruciais para a tarefa em questão. Portanto, a base de conhecimento pode conter uma série de fontes de informação e estas, por sua vez, podem conter milhões itens de dados, por exemplo, documentos. A eficácia de todo o sistema RAG depende criticamente da qualidade desta etapa; se o recuperador falhar em encontrar o contexto correto, o gerador não terá a informação necessária para produzir uma resposta precisa, um princípio conhecido como *"garbage in garbage out"*.

Já o **Gerador** (*Generator*) é tipicamente um LLM pré-treinado. Geralmente, ele recebe um *template* de *prompt* de entrada preenchido com as regras de geração de resposta, a consulta original do usuário e o contexto fornecido pelo recuperador. A sua tarefa não é simplesmente extrair ou repetir a informação recuperada, mas sim realizar uma síntese inteligente. O gerador deve integrar a intenção da pergunta do usuário com os fatos específicos do contexto, utilizando seu conhecimento linguístico e de raciocínio para gerar uma resposta final que seja coerente, relevante e em linguagem natural. A habilidade de sintetizar, contextualizar e adaptar, em vez de apenas repetir, é o que diferencia um sistema RAG sofisticado de um simples sistema de busca e extração.

5.1.3 O componente Recuperador (*Retriever*)

O recuperador é, em sua essência, um sistema de busca. Sua função primária é vasculhar uma base de conhecimento, que pode ser uma coleção de documentos, uma base de dados ou a web, para encontrar e extrair os trechos (*chunks*) de texto mais relevantes para a consulta do usuário. Embora existam diversas técnicas de recuperação, a abordagem mais comum atualmente tem um foco predominante na **busca semântica**, que busca por significado e contexto, em vez de simples correspondência de palavras-chave.

5.1.3.1 Etapas do processo de recuperação

O funcionamento do recuperador pode ser dividido em quatro etapas principais, que preparam e executam a busca por informações.

- **Indexação e Divisão (*Chunking*)**: Antes de qualquer busca, a base de conhecimento precisa ser processada. No caso de documentos, estes são divididos em pedaços menores e gerenciáveis, chamados de *chunks*. Essa divisão é crucial, pois permite que a busca retorne trechos específicos e de maior aderência à pergunta do usuário, em vez de documentos completos, otimizando o contexto que será enviado ao gerador.

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

- **Codificação (*Embedding*)**: Tanto a consulta do usuário quanto cada *chunk* de documento são transformados em uma representação vetorial numérica densa, conhecida como *embedding*. Esse processo é realizado por modelos de codificação (*encoders*) que capturam o significado semântico do documento em um espaço vetorial de alta dimensão. Tais modelos de linguagem especializados são denominados modelos de *embedding*.
- **Armazenamento**: Os *embeddings* dos *chunks* são armazenados e indexados em um banco de dados vetorial (*vector database*). Esses bancos de dados, como FAISS, Chroma ou Pinecone, são otimizados para realizar buscas de similaridade em alta velocidade em espaços vetoriais de milhões ou bilhões de vetores.
- **Busca por Similaridade**: Quando um usuário faz uma pergunta, o *embedding* da sua consulta geralmente é comparado com os *embeddings* dos documentos armazenados no banco de dados vetorial. O sistema então calcula a "distância" ou "proximidade" entre os vetores (usando métricas como a similaridade de cosseno) para encontrar os *chunks* mais similares semanticamente, que são então retornados como contexto.

5.1.3.2 Preparação de documentos e *Chunking*

O primeiro desafio em um *pipeline* RAG é lidar com o tamanho dos documentos. A maioria dos documentos do mundo real, como relatórios, artigos ou livros, excede em muito a janela de contexto dos LLMs, o número máximo de *token* que um modelo pode processar de uma só vez. Neste ponto, a RAG atua para evitar a adição de ruído no *prompt*, selecionando apenas os documentos ou trechos de documentos que são considerados relevantes a serem inseridos como contexto. Além disso, gerar um único *embedding* para um documento muito longo pode "diluir" a sua semântica, resultando em uma representação vetorial pouco informativa, ou uma massa gigantesca de texto que não corresponde bem ao necessário para consultas mais específicas.

A solução para este problema é a segmentação, ou *chunking*: o processo de dividir documentos longos em pedaços menores e mais gerenciáveis. A escolha da estratégia e do tamanho do *chunk* é uma das decisões de engenharia mais impactantes na qualidade de um sistema RAG. Existe um equilíbrio fundamental (Lita et al., 2024):

- ***Chunks* maiores**: Capturam mais contexto semântico, preservando a relação entre parágrafos e seções. No entanto, aumentam o risco de incluir informações irrelevantes (ruído) que podem confundir o LLM e são menos precisos para consultas muito específicas.
- ***Chunks* menores**: Oferecem representações mais focadas e com menos ruído, favorecendo a precisão em consultas granulares. Contudo, carregam menos contexto, o que pode dificultar a compreensão completa da informação se uma ideia estiver espalhada por múltiplos *chunks*.

Existem diversas estratégias de *chunking*, cada uma com seus próprios trade-offs (Lita et al., 2024):

- *Chunking* fixo (*Fixed-size*): A abordagem mais simples, que consiste em dividir o texto em segmentos de tamanho fixo. Em geral pelo número de *tokens*, caracteres ou palavras, com uma sobreposição (*overlap*) opcional para manter a continuidade e evitar a perda de informação com os cortes. É rápido e fácil de implementar, mas pode quebrar sentenças ou ideias no meio, prejudicando a coesão semântica do conteúdo textual.
- *Chunking* recursivo (*Recursive*): Uma melhoria em relação ao *chunking* fixo. Tenta dividir o texto recursivamente usando uma lista hierárquica de separadores, como quebras de parágrafo (\n\n), quebras de linha (\n) e espaços. Ele prioriza a manutenção da estrutura sintática do texto e só recorre a uma divisão de tamanho fixo como último recurso.
- *Chunking* semântico (*Semantic*): Esta técnica visa criar *chunks* que são semanticamente coerentes. Uma abordagem comum é agrupar sentenças com base na similaridade de seus *embeddings*. O processo parte da segmentação do texto, geralmente no nível das sentenças, com o agrupamento das partições por similaridade semântica com o *chunk* acumulado. As partições são agrupadas de acordo com um limiar de similaridade predefinido. Isso produz trechos mais coesos, mas é computacionalmente mais caro.
- *Chunking* baseado em estrutura (*Structure-based*): Aproveita a estrutura inerente do documento, como títulos, seções e subseções em arquivos HTML, Markdown ou LaTeX, para definir os limites dos *chunks*. Isso preserva a estrutura lógica do conteúdo, mas pode resultar em *chunks* de tamanho muito variável.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol2/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

- *Chunking* via LLM (*Agentic Chunking*): Uma abordagem de ponta que utiliza um LLM para analisar o documento e decidir inteligentemente como segmentá-lo. O LLM pode ser instruído a criar resumos de seções ou a agrupar trechos que tratam do mesmo tópico. Aqui, é importante ressaltar, que a aplicação do LLM deve ser o mais literal possível, evitando modificações do texto original e a criação de textos divergentes. Além disso, embora produza de altíssima qualidade semântica, o custo computacional e financeiro é significativamente maior, podendo ser proibitivo em alguns contextos.

5.1.3.3 Representando documentos textuais com *embeddings*

Após a segmentação, cada *chunk* de texto precisa ser convertido em uma representação numérica que um computador possa entender e comparar. Por muito tempo, a busca de informações baseou-se em correspondência de palavras-chave, usando modelos como o *Bag-of-Words* (BoW), que ignoram a ordem e o contexto das palavras. Essa abordagem é limitada por problemas como sinônimos (palavras diferentes com o mesmo significado) e paráfrases (frases diferentes com a mesma intenção).

Os *embeddings* textuais mitigam ou tratam diretamente esse problema. Um *embedding* é um vetor denso de números de ponto flutuante que representa o significado semântico de um pedaço de texto em um espaço de alta dimensionalidade. Esses vetores são a representação interna de um modelo de aprendizado profundo, geralmente sendo os vetores dos pesos da penúltima camada oculta (*hidden layer*). Portanto, esse significado semântico dos termos é dado na codificação dos documentos nos modelos de *embeddings*, de acordo com a proximidade e coocorrência das palavras em um determinado contexto. Nesse espaço, textos com significados semelhantes terão vetores próximos uns dos outros, medidos por métricas como a similaridade de cosseno.

Esses vetores são gerados por modelos de linguagem especializados, tipicamente arquiteturas *encoder-only* como BERT, RoBERTa ou, mais comumente para RAG, modelos Sentence-BERT (Reimers; Gurevych, 2019), que são otimizados para produzir *embeddings* de sentenças e parágrafos de alta qualidade. O processo envolve passar um *chunk* de texto pelo modelo *encoder*, que gera um *embedding* para cada *token*. Em seguida, uma técnica de *pooling* é aplicada para agregar esses *embeddings* de *tokens* em um único vetor que representa o *chunk* inteiro. As estratégias de *pooling* mais comuns são:

- *Average pooling*: Calcular a média dos vetores de todos os *tokens*, resultando em uma representação global que suaviza as variações locais.
- *Max pooling*: Selecionar, para cada dimensão do vetor, o valor máximo observado entre os *embeddings* dos *tokens*, destacando as características mais salientes do texto.
- *Token Especial*: Utilizar diretamente o *embedding* associado ao *token* especial `CLS` (em modelos como o BERT), que é treinado para capturar o significado agregado de toda a sequência.

Na prática e na literatura, a abordagem mais usual é fazer a média dos vetores dos *tokens* (*Average pooling*).

Uma vez que todos os *chunks* foram convertidos em *embeddings*, eles precisam ser armazenados em um local que permita uma busca rápida e eficiente. Em casos de *embeddings* semânticos, é usual utilizar, por exemplo, bancos de dados vetoriais, projetados para busca rápida e eficiente em bases com milhões de vetores.

Você pode se aprofundar sobre *embeddings* no Capítulo 1 deste livro.

5.1.3.4 Mecanismo de recuperação

A etapa final do processo de recuperação é a busca em si. Existem duas famílias principais de algoritmos de recuperação, e as abordagens mais modernas as combinam.

- Recuperadores Esparsos (*Sparse Retrievers*): Estes são os algoritmos clássicos de recuperação de informação, baseados em representações vetoriais esparsas (onde a maioria dos valores é zero). Nessa estrutura vetorial, os documentos indexados são dados pelos vetores numéricos que geralmente indicam a frequência ou a frequência ponderada dos termos que os compõe. Um dos algoritmos mais proeminentes para este tipo de recuperação é o BM25. Ele funciona com base na correspondência de palavras-chave, dando maior peso a termos que são raros no *corpus* geral, mas frequentes em um documento específico. São excelentes para encontrar documentos que contêm palavras-chave exatas ou terminologia específica.
- Recuperadores Densos (*Dense Retrievers*): Estes são os recuperadores baseados nos *embeddings* semânticos discutidos anteriormente. Eles utilizam vetores densos (onde a maioria dos valores é não-zero) que, através da ocorrência em contextos similares, capturam significado conceitual dos

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol2/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

termos que compõe o texto. Sua principal vantagem é a capacidade de encontrar documentos relevantes mesmo que não compartilhem nenhuma palavra-chave com a consulta, contanto que o significado seja semelhante. São robustos a sinônimos, paráfrases e variações de vocabulário.

- Busca Híbrida (*Hybrid Search*): A abordagem de estado da arte reconhece que os recuperadores esparsos e densos têm forças complementares. A busca híbrida executa ambos os tipos de busca em paralelo e, em seguida, combina os resultados. Essa combinação aproveita a precisão lexical da busca por palavras-chave e a robustez conceitual da busca semântica. Para mesclar as duas listas de resultados, que possuem pontuações em escalas diferentes, algoritmos de fusão de *ranking*, como o *Reciprocal Rank Fusion* (RRF) (Cormack et al., 2009), são frequentemente utilizados para produzir uma única lista de resultados classificada de forma mais relevante.

O Capítulo Recuperação de Informação deste livro trata com atenção mecanismos de Recuperação de Informação.

5.1.4 O componente Gerador (*Generator*)

O mecanismo central da RAG é a engenharia de *prompt* dinâmica, também conhecida como "*prompt stuffing*". O sistema monta e preenche estruturas predefinidas no *prompt*, tornando-o mais elaborado antes do envio para a LLM. Este *prompt* aumentado normalmente segue um *template* que combina as regras de geração de resposta, a consulta original realizada pelo usuário e o contexto recuperado. Um exemplo de template poderia ser:

Quadro 5.2: Exemplo de *prompt* aumentado em um sistema RAG

> Você é um assistente especializado em análise de currículos acadêmicos e profissionais. Responda à pergunta do usuário com base no trecho de currículo fornecido.
> Para isso, você receberá duas informações:
> 1. O prompt original enviado pelo usuário (que expressa o que ele deseja saber).
> 2. Um trecho do currículo acadêmico (seção).
> Sua tarefa é:
> – Analisar o trecho do currículo levando em consideração a pergunta;
> – Responder com base apenas nas informações explícitas ou fortemente implícitas nesse trecho;
> – Ignorar informações que não sejam pertinentes à pergunta;
> – Não extrapolar para além do conteúdo disponível na seção fornecida.
> Apresente sua análise de forma clara, objetiva e bem estruturada.
> **Prompt do Usuário:** `{prompt}`
> **Trecho do Currículo:** `{context}`

O exemplo acima foi o *prompt* utilizado no LattesRex(Darcio et al., 2025), um *chatbot* especialista em análises de currículos Lattes[4]. No exemplo, podemos ver as instruções, nas quais são definidas a tarefa, a função do modelo e o que deve conter na resposta. Os documentos coletados e a consulta do usuário são descritas e inseridas ao final do mesmo. Assim, a cada consulta do usuário, o *prompt* é modificado dinamicamente.

Este processo de aumento garante que o LLM tenha todas as informações factuais necessárias diretamente em sua janela de contexto no momento da geração.

O principal objetivo do *prompt* aumentado é forçar o LLM a realizar uma "geração ancorada" (grounded generation). Ao instruir explicitamente o modelo a basear sua resposta no contexto fornecido, o risco de alucinação é drasticamente reduzido. A resposta é "ancorada" no conteúdo factual dos documentos recuperados, em vez de ser puramente uma invenção probabilística baseada no conhecimento interno do modelo.

No entanto, a função do gerador vai muito além de uma simples estruturação de informação. A verdadeira habilidade de um sistema RAG sofisticado reside na capacidade do LLM de contextualizar a informação. O modelo não deve apenas repetir os trechos recuperados. Em vez disso, ele deve executar uma fusão complexa de três fontes de informação: (1) a intenção e os detalhes da consulta do usuário, (2) os fatos específicos contidos nos *chunks* recuperados, e (3) seu vasto conhecimento pré-treinado sobre gramática, estrutura, raciocínio e conhecimento geral do mundo.

[4] https://github.com/Lucas-Darcio/LattesRex/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Considere a seguinte consulta: "Compare os prós e contras das estratégias de *chunking* fixo e semântico". O recuperador pode retornar dois *chunks*: um descrevendo o *chunking* fixo e outro descrevendo o *chunking* semântico. O processo de geração de resposta mais simples listaria ambas as descrições como resultado. Um gerador sofisticado, no entanto, usaria sua capacidade de raciocínio para criar uma estrutura comparativa organizando os pontos fortes e fracos de cada abordagem lado a lado. Essa análise de contexto, realizando tarefas de síntese e estruturação textual, são formas de tratamento do conteúdo que apenas um LLM poderoso é capaz de executar, e é o que eleva a RAG de uma ferramenta de busca para um verdadeiro sistema de sintese de informações. Portanto, a qualidade de um sistema RAG depende não apenas da precisão do recuperador, mas também da capacidade de entendimento do contexto do gerador.

5.2 RAG vs. *Fine-tuning*

Ao buscar adaptar um LLM para resolver problemas específicos, duas soluções principais emergem: RAG e *fine-tuning* (ajuste fino). Embora possam ser usadas em conjunto, a escolha entre elas envolve um *trade-off* fundamental relacionado ao objetivo, custo e natureza dos dados.

5.2.1 *Fine-tuning*

O *fine-tuning* consiste em continuar o treinamento de um modelo fundacional pré-treinado, utilizando um conjunto de dados específico do seu domínio. O objetivo é ajustar os pesos internos do modelo para que ele se especialize em uma tarefa ou em um estilo de comunicação particular.

Essencialmente, o *fine-tuning* é ideal para ensinar ao LLM **como** responder, permitindo ajustar o estilo, o tom, o formato da saída e garantir que o modelo siga instruções complexas e específicas do domínio ou caso de uso. Isso se alinha a uma máxima importante na comunidade de IA: o ajuste fino serve para ensinar **forma**, não sendo indicado para injetar **fatos** ao modelo. Tentar ensinar novos conhecimentos factuais exclusivamente por este método pode levar a alucinações e não garante que o modelo recupere a informação correta no momento certo.

Deste modo, o *fine-tuning* não é uma opção interessante para lidar com liberdade de escrita e as veriações textuais de determinados domínios. Entretanto, tal abordagem pode ser aplicada para atuar com fatores e conhecimentos relativamente bem definido e estáticos. O conteúdo para esses casos deve sofrer poucas mudanças de acordo com o tempo pois, após o treinamento, o conhecimento do modelo fica "congelado" até um novo processo de ajuste. Assim, o sucesso do *fine-tuning* depende criticamente da qualidade e da curadoria do conjunto de dados de treinamento e, mesmo com técnicas de otimização como o LoRA (*Low-Rank Adaptation*)(Hu et al., 2021), o processo pode ser computacionalmente caro, exigindo recursos de hardware e tempo significativos.

5.2.2 RAG

Em contraste, a RAG não modifica os pesos internos do LLM. Em vez disso, ele enriquece as consultas para maximizar a qualidade dos resultados obtidos "em tempo de execução", fornecendo-lhe informações externas relevantes para cada consulta.

A principal vantagem da RAG é sua capacidade de acessar **conhecimento dinâmico**. Basta atualizar a base de dados externa para que o LLM utilize as informações mais recentes, sem a necessidade de um caro e demorado re-treinamento. Ao "ancorar" a resposta do LLM em um contexto factual, a RAG reduz significativamente a probabilidade de o modelo inventar informações (alucinar).

Isso torna a RAG a abordagem ideal para domínios onde a **factualidade e a auditabilidade** são essenciais, como em aplicações de direito, apoio ao cliente ou análise de documentos técnicos. Como a resposta é gerada a partir de fontes explícitas, é possível citá-las, aumentando a confiança do usuário. A implementação geralmente envolve a orquestração de componentes como bancos de dados vetoriais e recuperadores, frequentemente facilitada por *frameworks* como LangChain ou LlamaIndex.

Em resumo, a escolha depende do problema: se o objetivo é ensinar um novo estilo ou comportamento, o *fine-tuning* é a ferramenta certa. Se o objetivo é garantir respostas factuais baseadas em um corpo de conhecimento dinâmico e específico, a RAG é a solução superior. Em muitos casos avançados, uma abordagem híbrida, onde um modelo *fine-tuned* para um determinado estilo utiliza RAG para acessar dados atualizados, pode oferecer o melhor dos dois mundos.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol2/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

5.2.3 Abordagens híbridas

A discussão "RAG vs. *fine-tuning*" pode apresentar uma falsa dicotomia. Na prática, as arquiteturas de IA mais robustas não escolhem entre um ou outro, mas os combinam de forma sinérgica. É importante ao se aprofundar nas abordagens híbridas, entender que o refinamento das tarefas e a especialização dos *prompts* e modelos potencializa os resultados obtidos. Portanto, um exemplo de modelos híbridos é o uso de *fine-tuning* para otimizar o "motor" de raciocínio do LLM, enquanto a RAG fornece o "combustível" (dados) em tempo real. Essa combinação permite que o modelo não apenas acesse o conhecimento correto, mas também saiba como utilizá-lo da maneira mais eficaz.

Um exemplo prático seria um assistente de IA para médicos. Primeiramente, o LLM base passaria por um ***fine-tuning*** com um conjunto de prontuários médicos anonimizados e diálogos clínicos. O objetivo não seria memorizar os prontuários, mas sim aprender a estrutura da linguagem médica, entender abreviações e saber como resumir informações clínicas. Em seguida, quando um médico perguntasse "Resuma o histórico de hipertensão do paciente João Silva", o sistema **RAG** entraria em ação, recuperando os trechos relevantes do prontuário eletrônico específico daquele paciente. Por fim, o LLM, já ajustado para "pensar como um médico", receberia esse contexto e geraria um resumo conciso e clinicamente relevante.

As arquiteturas híbridas podem ser implementadas de várias maneiras, como a Sequencial (primeiro ajusta-se o modelo, depois aplica-se RAG) ou a Integrada (onde o modelo é treinado para otimizar tanto a recuperação quanto a geração), representando a vanguarda na personalização de LLMs.

5.3 RAG - Aspectos avançados

A arquitetura fundamental de RAG, que consiste em recuperar informações de uma base de conhecimento externa antes de gerar uma resposta, provou ser uma solução robusta para mitigar alucinações e superar a limitação do conhecimento estático dos modelos. No entanto, a implementação "ingênua" ou tradicional deste paradigma, caracterizada por uma busca vetorial simples seguida de uma concatenação de contexto, revela rapidamente suas próprias limitações. A qualidade da resposta final é inexoravelmente limitada pela qualidade da recuperação inicial; uma abordagem de "tamanho único" para todas as consultas é ineficiente e subótima; e o foco exclusivo em dados textuais ignora a riqueza de informações presentes em formatos estruturados e multimodais.

Nesta seção aprofundaremos o ecossistema de técnicas avançadas de RAG, que transformam o *pipeline* linear e passivo em um processo de raciocínio dinâmico, adaptativo e autocrítico. A tese central é que a evolução da RAG transcende a mera recuperação de informações, movendo-se em direção a um paradigma onde o sistema introduz camadas sofisticadas de otimização, tomada de decisão inteligente, autoavaliação e raciocínio estruturado. Exploraremos como o *re-ranking* refina a precisão, como a fusão de consultas amplia o alcance da busca e como algoritmos inteligentes podem rotear, corrigir e até mesmo habilitar a autocrítica no processo de geração. Além disso, expandiremos as fronteiras da RAG para além do texto plano, investigando como grafos de conhecimento permitem o raciocínio complexo e como a multimodalidade integra visão e linguagem. Finalmente, abordaremos os desafios práticos da implementação em produção, desde as estratégias de segmentação de dados e avaliação rigorosa até as considerações éticas e de desempenho.

5.3.1 RAG em duas etapas: Re-ordenando resultados

A arquitetura de recuperação em sistemas RAG de larga escala é frequentemente projetada como um processo de duas etapas que equilibra a necessidade de velocidade em um vasto *corpus* com a exigência de alta precisão em um conjunto menor de *chunks* ou documentos candidatos. Este paradigma representa um trade-off fundamental entre a abrangência semântica e a profundidade contextual.

A primeira etapa, a **recuperação inicial**, é otimizada para velocidade e escalabilidade. Seu objetivo é pesquisar rapidamente em milhões ou até bilhões de documentos para encontrar um conjunto de candidatos potencialmente relevantes. Para isso, são utilizados modelos **bi-encoder**. Esses modelos geram representações vetoriais (*embeddings*) para a consulta e para todos os documentos da base de conhecimento de forma independente e antecipada. (Khattab; Zaharia, 2020) No momento da consulta, a recuperação se resume a uma busca de vizinhos mais próximos (*Approximate Nearest Neighbor* - ANN) no espaço vetorial, uma operação extremamente eficiente. O *bi-encoder* lança uma rede ampla e rápida, maximizando o *recall* ao capturar todos os documentos que são semanticamente próximos da consulta em um sentido geral. No entanto, como a consulta e o documento são processados separadamente, interações mais finas e contextuais entre eles são perdidas (Khattab; Zaharia, 2020).

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Uma vez que um subconjunto menor e mais gerenciável de documentos candidatos é recuperado (por exemplo, os 20 a 50 melhores resultados), a segunda etapa, o ***re-ranking***, entra em ação. O objetivo aqui é mudar o foco da velocidade para a precisão. Esta etapa utiliza modelos mais poderosos e computacionalmente mais caros, como os **cross-encoders**, para reavaliar e reordenar a lista de candidatos. O *cross-encoder* realiza uma análise profunda e mais lenta para verificar a relevância contextual precisa, maximizando a *precisão* no conjunto restrito. Ele reordena os documentos, empurrando os mais relevantes para o topo da lista antes de serem enviados ao gerador.

Essa arquitetura de duas fases é a chave para a eficiência e eficácia do *re-ranking*. O sistema não está apenas filtrando resultados; ele está mudando a própria definição de "relevância" entre as etapas. A transição ocorre de uma correspondência semântica geral e de baixo custo para uma correspondência contextual específica e de alto custo, combinando o melhor dos dois mundos para alcançar um desempenho superior que seria computacionalmente inviável com uma única abordagem.

5.3.1.1 *Cross-Encoders*

Cross encoders são mais adequados à tarefa de *re-ranking* após o conjunto inicial de candidatos ser recuperado. Enquanto um *bi-encoder* mede a similaridade entre duas representações independentes, um *cross-encoder* modela diretamente o alinhamento entre a consulta e o documento.

A mecânica de um *cross-encoder* funciona da seguinte maneira: o algoritmo recebe a consulta e um documento candidato como uma única sequência de entrada, tipicamente concatenados e separados por um *token* especial, como `[SEP]`. Essa sequência conjunta é então processada por um modelo Transformer. O mecanismo de auto-atenção do Transformer é aplicado a toda a sequência, permitindo que cada *token* da consulta "preste atenção" a cada *token* do documento e vice-versa, em todas as camadas do modelo. Essa interação profunda permite que o modelo capture dependências sutis, nuances contextuais e relações semânticas complexas que são impossíveis de serem modeladas quando a consulta e o documento são codificados separadamente. A saída do *cross-encoder* não são dois vetores, mas sim um único score de relevância, geralmente um valor entre 0 e 1, que representa a probabilidade de o documento ser relevante para a consulta.

O treinamento desses modelos é tipicamente supervisionado, formulado como uma tarefa de classificação binária. Utilizando grandes datasets de pares (consulta, documento) rotulados como relevantes ou não relevantes, o modelo é treinado para otimizar uma função de perda como a **Entropia Cruzada Binária (BCE)**. Esse processo de treinamento ensina o modelo a produzir uma pontuação alta para pares relevantes e uma pontuação baixa para pares irrelevantes.

O poder do *cross-encoder* vai além de simplesmente "processar juntos". Ele reside na sua capacidade de modelar a **relevância assimétrica**. A relevância da informação é frequentemente direcional: um documento que descreve os efeitos colaterais de um medicamento é altamente relevante para a consulta "quais são os efeitos colaterais do medicamento X?". No entanto, a consulta não é relevante para o documento. Os *bi-encoders*, que medem a similaridade simétrica (como a similaridade de cosseno, onde `sim(A, B) = sim(B, A)`), têm dificuldade em capturar essa nuance. No caso do *cross-encoder*, ao ser treinado com os pares de pergunta e resposta candidata, se torna um avaliador especializado nessa assimetria. Portanto,para responder à pergunta "este documento é relevante *para esta consulta*?", o algoritmo aprende a capturar essa assimetria fundamental, o que leva a uma avaliação de relevância muito mais precisa e dependente do contexto.

Quadro 5.3: Comparativo de arquiteturas de *Bi-Encoder* vs. *Cross-Encoder*

Característica	***Bi-Encoder***	***Cross-Encoder***
Formato de Entrada	Consulta e Documento processados separadamente.	Par (Consulta, Documento) concatenado como uma única entrada.
Processamento	Geração de *embeddings* independentes e comparação via função de similaridade (e.g., cosseno).	Processamento conjunto com mecanismo de atenção cruzada entre os *tokens* da consulta e do documento.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol2/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Saída	*Score* simétrico de similaridade (e.g., -1 a 1).	*Score* assimétrico de relevância (e.g., 0 a 1).
Custo de Indexação	Alto: requer o pré-cálculo e armazenamento de *embeddings* para todos os documentos na base.	Nulo: não há etapa de indexação.
Custo de Inferência	Baixo: 1 passagem do *encoder* para a consulta + busca vetorial rápida.	Alto: N passagens do *encoder* para re-ranquear N documentos candidatos.
Caso de Uso Principal	Recuperação inicial em larga escala (primeira fase).	*Re-ranking* de um conjunto pequeno de candidatos (segunda fase).
Vantagem Principal	Velocidade e escalabilidade para *corpus* massivos.	Alta precisão e compreensão contextual profunda.
Desvantagem Principal	Menor precisão contextual, pois perde interações finas.	Custo computacional alto, não escalável para a recuperação inicial.

5.3.1.2 Expansão de consultas com RAG-Fusion

Uma limitação inerente a qualquer sistema de busca é sua dependência da formulação exata da consulta do usuário. Uma única consulta pode ser ambígua, subespecificada ou simplesmente não usar a terminologia que melhor corresponde aos documentos relevantes na base de conhecimento. A técnica **RAG-Fusion** aborda diretamente esse problema, reconhecendo que a intenção do usuário pode ser mais bem capturada por múltiplas perspectivas(Rackauckas, 2024).

O processo de RAG-Fusion funciona em três etapas principais:

1. **Geração de múltiplas consultas:** Em vez de depender apenas da consulta original, RAG-Fusion primeiro utiliza um LLM para gerar múltiplas variações daquela consulta. Através de um *prompt* cuidadosamente elaborado, o LLM é instruído a criar diversas perguntas que abordam a mesma intenção subjacente, mas de diferentes ângulos ou com diferentes palavras-chave. Por exemplo, para a consulta "impacto da IA na economia", o LLM poderia gerar: "como a inteligência artificial afeta o mercado de trabalho?", "automação e crescimento do PIB" e "riscos econômicos da IA".
2. **Recuperação paralela:** O sistema então realiza buscas vetoriais para cada uma dessas consultas geradas, bem como para a consulta original, em paralelo. Cada consulta atua em diferentes regiões do espaço vetorial, recuperando conjuntos de documentos que, embora todos relevantes, podem ser distintos entre si.
3. **Fusão e *re-ranking*:** Finalmente, os conjuntos de resultados de todas as buscas são agregados e re-ranqueados usando um algoritmo de fusão, mais comumente o *Reciprocal Rank Fusion* (RRF) (Rackauckas, 2024), para produzir uma única lista final de documentos classificados.

Essa abordagem transforma a recuperação de um ato de *matching* para um ato de *exploração*. Em essência, RAG-Fusion realiza uma forma de engenharia de *prompt* automática para a fase de recuperação. Em vez de exigir que o usuário refine sua pergunta, o sistema assume proativamente essa responsabilidade. Ele utiliza o poder generativo e a compreensão semântica do LLM para explorar o contexto ou entorno da busca realizada pelo usuário, em outras palavras, o "espaço latente de intenções". Ao transformar uma única consulta pontual em uma distribuição de consultas, ele aumenta a área de superfície da busca, elevando significativamente a robustez e a probabilidade de encontrar documentos relevantes que uma única formulação teria perdido.

5.3.1.3 Fusão de diferentes ordenações: *Reciprocal Rank Fusion* (RRF)

Quando múltiplas listas de resultados precisam ser combinadas — seja na RAG-Fusion, que usa múltiplas consultas, ou na busca híbrida, que combina busca vetorial e busca por palavra-chave (como BM25) — surge o desafio de como fundir os *rankings* gerados por cada método. Os scores produzidos por diferentes sistemas de busca (e.g., similaridade de cosseno, *score* BM25) estão em escalas completamente diferentes e não são diretamente comparáveis (Rackauckas, 2024). Tentar normalizá-los pode ser complexo e frágil.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol2/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

O ***Reciprocal Rank Fusion*** **(RRF)** é um algoritmo que resolve esse problema ao ignorar completamente os scores brutos e focar apenas na posição (rank) dos documentos em cada lista. O RRF, opera como um algoritmo de *ensemble voting*, onde a ideia de voto é aplicada sob a premissa de que a posição de um documento em uma lista de resultados é um sinal de relevância mais confiável e universal do que o score numérico que o gerou. Portanto, o RRF atua como um sistema de votação ponderada pela confiança posicional.

Para cada documento d que aparece em uma ou mais listas de resultados R, o RRF calcula uma pontuação combinada usando a seguinte fórmula [1, 10]:

$$Score(d) = \sum_{r \in R} \frac{1}{k + rank_r(d)} \tag{5.1}$$

Onde:

- $rank_r(d)$ é a posição do documento d na lista de resultados do ranqueador r, começando com o rank 1 para o primeiro item.
- k é uma constante de suavização, comumente definida como 60.

A pontuação final de um documento é a soma de suas pontuações RRF em todas as listas em que ele aparece. Documentos que aparecem consistentemente em posições altas em múltiplas listas recebem as maiores pontuações combinadas e, portanto, são promovidos ao topo do *ranking* final.

O papel da constante k é crucial. Ela amortece a contribuição dos documentos de alto escalão, garantindo que eles não sejam excessivamente dominantes, e dá mais peso a documentos de *ranking* mais baixo do que uma simples fórmula de $1/rank$ daria. Sem k, a diferença de pontuação entre o rank 1 e 2 seria muito maior do que entre o rank 60 e 61. A constante k suaviza essa queda, permitindo que o algoritmo considere um espectro mais amplo de resultados. Essencialmente, k controla o quão cético o algoritmo é em relação às distinções de *ranking* muito altas.

Ao descartar os scores, o RRF democratiza a contribuição de diferentes recuperadores. Cada sistema (vetorial, *keyword* etc.) pode "votar" nos documentos com base em sua especialidade, com o peso do voto diminuindo com a posição. O RRF troca a magnitude do score pela consistência do *ranking*, buscando um consenso entre as diferentes fontes de evidência. Um documento que é bem ranqueado, incluído entre os dez primeiros nos resultados de três recuperadores diferentes é provavelmente mais relevante do que os primeiros documentos dado como resultado no *ranking* individual de cada recuperador, mas ignorado pelos outros.

5.3.2 RAG adaptiva e o Modelo agêntico

Após otimizar a camada de recuperação, a próxima fronteira na evolução da RAG é a introdução de inteligência e tomada de decisão no próprio *pipeline*. As arquiteturas tradicionais seguem um fluxo linear e fixo, tratando todas as consultas da mesma forma. Em contraste, os sistemas avançados podem adaptar sua estratégia com base na natureza da consulta, avaliar a qualidade dos documentos recuperados e até mesmo criticar suas próprias saídas. Esta seção explora técnicas que transformam a RAG de um processo mecânico para um fluxo de trabalho de raciocínio dinâmico e resiliente.

5.3.2.1 RAG adaptativo: Roteamento dinâmico de consultas

O princípio fundamental do **RAG adaptativo** é que nem todas as consultas tem as mesmas características e, portanto, não devem ser tratadas com o mesmo *pipeline* computacionalmente intensivo. Uma pergunta simples como "Olá, como vai você?" não requer uma busca em um banco de dados vetorial, enquanto uma pergunta sobre eventos recentes, como "Quem ganhou o jogo de ontem?", não pode ser respondida por uma base de conhecimento estática (Jeong et al., 2024).

Para lidar com essa diversidade, a RAG Adaptiva introduz um componente de tomada de decisão no início do fluxo: o **Query Router** (Roteador de Consultas) (Yeo et al., 2025). Este roteador é um classificador leve que analisa a consulta do usuário e a direciona para o caminho de processamento mais eficiente e apropriado. Frequentemente, este roteador é implementado usando um LLM instruído a produzir uma saída estruturada, garantindo que a decisão seja baseada na compreensão semântica da consulta, e não apenas em palavras-chave.

O fluxo de trabalho típico de um RAG Adaptativo pode ter múltiplos ramos:

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol2/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

1. **Geração direta:** Para consultas simples, conversacionais ou que não requerem conhecimento externo, o roteador envia a pergunta diretamente ao LLM, pulando a etapa de recuperação. Isso economiza tempo e recursos.
2. **Recuperação na base de conhecimento (*Vector Store*):** Para perguntas que são factuais e se enquadram no domínio coberto pela base de conhecimento interna, o roteador aciona o *pipeline* RAG padrão, realizando uma busca vetorial nos documentos indexados.
3. **Busca na web ou recuperação complexa:** Para perguntas sobre eventos recentes, tópicos fora do escopo da base de conhecimento ou questões complexas que podem se beneficiar de múltiplas fontes, o roteador direciona a consulta para uma ferramenta de busca na web ou aciona um fluxo mais complexo, como a RAG-Fusion.

A implementação prática envolve definir um modelo de dados que especifica as rotas possíveis e fornecer um *prompt* de sistema ao LLM que o instrui a atuar como um roteador especialista.

Ao introduzir essa camada de decisão, a RAG Adaptiva implementa um conceito de **gerenciamento de recursos computacionais** no nível da aplicação. Ele trata a recuperação de informações não como uma etapa obrigatória, mas tal qual uma ferramenta é conectada a um LLM, a recuperação é uma ferramenta a ser invocada apenas quando necessário e na intensidade apropriada. Isso otimiza não apenas a relevância da resposta, mas também a latência e o custo, transformando o *pipeline* RAG de uma arquitetura de dados estática para uma arquitetura de computação dinâmica e orientada por custos.

5.3.2.2 *Corrective* RAG (CRAG): Um *framework* para a robustez

Enquanto a RAG Adaptiva otimiza o fluxo *antes* da recuperação, o **Corrective RAG (CRAG)** (Yan et al., 2024) foca em lidar com o problema fundamental da RAG: o que fazer quando a recuperação falha e retorna documentos irrelevantes ou de baixa qualidade? A recuperação incorreta é uma das principais causas de alucinações, pois o LLM é forçado a gerar uma resposta com base em informações erradas.

O CRAG introduz um mecanismo de auto-correção e verificação de qualidade *após* a recuperação. O componente central é um **avaliador de recuperação (*retrieval evaluator*)** leve, projetado para avaliar a relevância dos documentos recuperados em relação à consulta. Este avaliador, tipicamente um modelo *encoder* menor, como o T5-large. Esse modelo é escolhido com cautela, ponderando a qualidade da resposta, a latência da requisição e o tempo de resposta com o papel de avaliador. Assim, o modelo de avaliação é ajustado para prever um score de confiança para cada par (consulta, documento).

Com a aplicação do avaliador, é comum o aumento da quantidade de documentos ou *chunks* recuperados. Ao definir a relevância de acordo com o contexto, o *score* de confiança também é responsável pela redução de nível de ruído e refinamento do conteúdo que efetivamente passado para a geração da resposta. Com base nesse *score* de confiança, o CRAG aciona dinamicamente uma de três ações distintas, implementando um mecanismo de *fallback* e enriquecimento em cascata:

1. **Confiança alta ("Correct"):** Se pelo menos um documento recuperado obtiver uma pontuação acima de um limiar superior, o sistema considera a recuperação bem-sucedida. Os documentos são considerados relevantes e são usados para a geração. Opcionalmente, para aumentar ainda mais a precisão, um processo de **refinamento de conhecimento** pode ser aplicado. Este processo, chamado de "decompose-then-recompose", quebra os documentos em "*knowledge strips*" (pequenos trechos de informação) e usa o mesmo avaliador para filtrar apenas os trechos mais pertinentes antes de passá-los ao gerador.
2. **Confiança baixa ("Incorrect"):** Se todos os documentos recuperados tiverem uma pontuação abaixo de um limiar inferior, o sistema conclui que a recuperação falhou. Os documentos recuperados são descartados para evitar a contaminação do contexto. Em vez disso, o CRAG recorre a uma fonte de conhecimento alternativa e mais ampla, como uma **busca na web**, para encontrar informações mais relevantes.
3. **Confiança média ("Ambiguous"):** Se as pontuações dos documentos ficarem entre os dois limiares, o sistema reconhece a incerteza. A recuperação não é nem claramente correta nem claramente incorreta. Nesta situação, o CRAG adota uma abordagem híbrida e robusta: ele combina o conhecimento refinado dos documentos recuperados (ação "Correct") com o conhecimento obtido da busca na web (ação "Incorrect"), maximizando a cobertura e a probabilidade de fornecer ao LLM um contexto útil.

O CRAG modela explicitamente a incerteza da recuperação. Em vez de falhar silenciosamente com informações ruins, o sistema reconhece a incerteza, quantifica-a com um score de confiança e aciona a

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol2/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

política de mitigação mais apropriada: refinar os documentos obtidos, substituir por conteúdos mais apropriados segundo o contexto ou aumentar com informações adicionais. É um sistema de gerenciamento de risco para a recuperação de informações, aumentando significativamente a robustez e a confiabilidade do *pipeline* RAG.

5.3.2.3 Self-RAG: Incorporando autonomia com *tokens* de reflexão

O **Self-RAG (*Self-Reflective Retrieval-Augmented Generation*)** (Asai et al., 2024) representa um passo adiante na autonomia dos sistemas RAG, internalizando o *pipeline* de controle (que em CRAG e RAG Adaptativo é composto por módulos externos) dentro do próprio processo de geração de linguagem do LLM. Ele faz isso treinando o LLM gerador para produzir **"*tokens* de reflexão"** — *tokens* especiais, adicionados ao seu vocabulário, que controlam o processo de recuperação e autocrítica em tempo real.

Esses *tokens* transformam a recuperação e a crítica de operações de *pipeline* em atos de linguagem. O LLM não está mais sendo passivamente "aumentado" por um processo externo; ele está ativamente narrando e controlando sua própria busca por conhecimento e verificação de fatos como parte integrante da tarefa de geração de texto.

Os *tokens* de reflexão são categorizados em dois tipos principais:

1. ***Tokens* de recuperação:** Em cada etapa da geração, o modelo pode decidir se a recuperação de informações externas seria útil. Ele manifesta essa decisão gerando um *token* específico:
   - `[Retrieve]`: Indica que a recuperação é necessária. A geração é pausada, uma busca é realizada e os documentos recuperados são apresentados ao modelo.
   - `[No Retrieve]`: Indica que a recuperação não é necessária e o modelo pode continuar a geração com base em seu conhecimento paramétrico ou no contexto já disponível.

   Isso permite uma **recuperação adaptativa e sob demanda**, que pode ocorrer múltiplas vezes durante a geração de uma única resposta, ou não ocorrer de todo, dependendo da necessidade.
2. ***Tokens* de crítica:** Após a recuperação (se acionada) e durante a geração, o modelo avalia a qualidade da informação e de sua própria saída:
   - `[Relevant]` / `[Irrelevant]`: O modelo avalia cada documento recuperado, gerando um *token* para indicar sua relevância para a continuação da resposta.
   - `[IsSupported]` / `[IsNotSupported]`: Ao gerar um segmento de texto baseado nos documentos recuperados, o modelo avalia se sua própria afirmação é factualmente suportada pela evidência fornecida.
   - `[Good]` / `[Bad]`: O modelo pode fazer uma avaliação holística da qualidade geral da resposta.

O treinamento do Self-RAG envolve a criação de um *corpus* de treinamento onde esses *tokens* de reflexão são inseridos *offline* por um modelo "crítico" independente. Portanto, a função do modelo crítico é realizar uma anotação qualitativa dos resultados obtidos pelo fluxo principal. O LLM gerador é então treinado em um objetivo padrão de previsão do próximo *token*, aprendendo a prever tanto as palavras da resposta quanto esses *tokens* de controle.

Essa internalização do controle tem consequências profundas. O processo de raciocínio do modelo torna-se explícito e controlável na inferência. Por exemplo, os desenvolvedores podem ajustar o comportamento do modelo simplesmente alterando as probabilidades de amostragem dos *tokens* de reflexão, tornando-o mais ou menos propenso a recuperar informações ou mais ou menos crítico em relação às suas fontes. Self-RAG representa um passo fundamental em direção a modelos de linguagem com meta-cognição. Nesse aspecto, os modelos não apenas processam informações, como também refinam a geração de respostas com a otimização do processo de raciocínio do modelo.

5.3.3 Indo além do texto

A RAG tradicional opera em um mundo de *chunks* de texto independentes, tratando cada pedaço de informação como uma unidade isolada. Essa abordagem, embora eficaz para muitas tarefas, falha em capturar as ricas interconexões entre os conceitos e ignora a vasta quantidade de informações contidas em modalidades não textuais. Esta seção explora duas fronteiras que quebram essa limitação: GraphRAG, que introduz relações estruturadas para permitir um raciocínio mais complexo, e RAG Multimodal, que integra modalidades como imagens para criar uma compreensão mais holística do mundo.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol2/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

5.3.3.1 GraphRAG: Recuperando informação de grafos de conhecimento

O **GraphRAG** (Pan et al., 2024) aborda uma limitação fundamental da RAG baseado em vetores: a incapacidade de responder a perguntas complexas que exigem a síntese de informações de múltiplos documentos ou a compreensão de relações indiretas. Em vez de tratar os *chunks* de texto como pontos independentes em um espaço semântico, o GraphRAG constrói um **grafo de conhecimento** onde os nós representam entidades (como pessoas, organizações, conceitos) e as arestas representam as relações explícitas entre elas.

O processo de construção do grafo a partir de dados não estruturados envolve várias etapas sofisticadas :

1. **Extração de entidades e relações:** O *corpus* de texto é processado para extrair entidades nomeadas e as relações entre elas. Isso pode ser feito usando LLMs para gerar tripletas estruturadas do tipo (entidade1, relação, entidade2).
2. **Construção do grafo:** As entidades se tornam nós e as relações se tornam arestas direcionadas, formando um grafo que representa o conhecimento do *corpus* de forma estruturada.
3. **Detecção de comunidades:** Algoritmos de clusterização de grafos, como o algoritmo de Leiden, são aplicados para identificar "comunidades" de nós densamente conectados. Essas comunidades representam temas ou domínios de conhecimento dentro do *corpus*.
4. **Geração de resumos:** Para cada comunidade identificada, um LLM gera um resumo que descreve o tema central e as principais entidades e relações dentro daquele cluster. Isso cria uma visão hierárquica do conhecimento, do geral (resumos de comunidade) ao específico (nós individuais).

Essa estrutura de grafo habilita o **raciocínio de múltiplos saltos (*multi-hop reasoning*)**, que é a capacidade de responder a perguntas atravessando múltiplas conexões no grafo. Considere a pergunta: "Como a aquisição do LinkedIn pela Microsoft impactou as tendências de networking profissional?". Um RAG tradicional pode recuperar documentos sobre a aquisição e outros sobre networking, mas a conexão causal entre eles pode ser fraca ou inexistente nos textos recuperados. O LLM teria que inferir essa conexão, com alto risco de alucinação.

O GraphRAG, por outro lado, transforma a recuperação de um problema de busca semântica para um problema de travessia de grafo. Para responder à pergunta, o sistema pode iniciar no nó `(Microsoft)`, seguir a aresta `[adquiriu]` para `(LinkedIn)`, e então seguir a aresta `[influencia]` para `(Networking Profissional)`. Ao longo desse caminho, ele coleta o contexto associado a cada nó e aresta, fornecendo ao LLM um contexto de raciocínio explícito. A recuperação não se trata mais de encontrar os melhores trechos de texto, mas sim de encontrar e aprofundar os conhecimentos sobre as entidades mais alinhadas ao contexto através do grafo de conhecimento. Isso permite que o sistema responda a perguntas sobre relações, causalidade e impacto, que estão muito além da capacidade da RAG tradicional.

Uma abordagem alternativa ao GraphRAG é o *Recursive Abstractive Processing for Tree-Organized Retrieval* (RAPTOR) (Sarthi et al., 2024). O RAPTOR apresenta uma proposta de sumarização recursiva da estrutura de clusters de documentos, definindo uma organização em árvore. Através desta árvore, é criado um raciocínio hierárquico e contextual. Ele é efetivo para assimilar contextos e reduzir a redundância das informações. Portanto, enquanto o GraphRAG tem uma informação bem-definida e relações entre os componentes de forma estruturada, os sumários do RAPTOR visam atuar em situações onde é necessário reduzir o ruído e otimizar a busca pelo contexto correto.

5.3.3.2 RAG multimodal: Integrando visão e linguagem

A RAG multimodal estende o paradigma para além do texto, incorporando outras modalidades de dados como imagens, áudio e vídeo. Uma das abordagens mais poderosas para alcançar isso é através do uso de modelos de *embedding* multimodais, como o **CLIP (*Contrastive Language-Image Pre-training*)** da OpenAI (Radford et al., 2021).

O CLIP é treinado em milhões de pares (imagem, texto) da internet com um objetivo de aprendizado contrastivo. O resultado é um modelo que pode projetar tanto imagens quanto textos em um **espaço vetorial compartilhado**. Nesse espaço, a distância entre os vetores corresponde à similaridade semântica, independentemente da modalidade original. O *embedding* da imagem de um cão estará semanticamente próximo do *embedding* do texto "uma foto de um cão". Esse resultado é obtido com *embeddings* pois tanto o texto quanto a imagem do cão estão compartilhando o mesmo espaço vetorial, tornando essa associação possível. Consequentemente, isso permite uma busca intermodal robusta: usar uma imagem para encontrar textos relevantes, ou usar um texto para encontrar imagens relevantes.

O *pipeline* de um sistema RAG multimodal usando CLIP normalmente segue estas etapas :

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol2/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

1. **Ingestão e Indexação:** Todos os dados da base de conhecimento, tanto textos quanto imagens, são processados pelo *encoder* apropriado do CLIP (*encoder* multimodal de texto e imagem) para gerar *embeddings*. Esses *embeddings* multimodais são armazenados em um único banco de dados vetorial. As imagens podem também ter metadados textuais associados, como descrições, que são armazenados junto com os *embeddings*.
2. **Consulta multimodal:** O usuário pode fornecer uma consulta em forma de texto (e.g., "mostre-me jaquetas de couro"), uma imagem (e.g., uma foto de uma jaqueta que ele gostou), ou uma combinação de ambos. A consulta é convertida em um *embedding* usando o mesmo modelo CLIP.
3. **Recuperação intermodal:** O *embedding* da consulta é usado para realizar uma busca de similaridade no banco de dados vetorial. Como textos e imagens coexistem no mesmo espaço, a busca pode retornar os itens mais relevantes, apresentando resultados mistos entre os modais ou combinando imagens e textos em um único tipo de resultado.
4. **Geração aumentada:** Os dados recuperados (imagens e/ou textos) são então passados, juntamente com a consulta original, para um LLM multimodal (como o GPT-4 Vision). Este modelo pode então raciocinar sobre as informações visuais e textuais para gerar uma resposta rica e fundamentada.

Essa abordagem não está apenas recuperando imagens; ela está realizando uma tradução entre espaços de representações. O espaço de *embedding* compartilhado do CLIP atua como um intermediário para conceitos abstratos. Quando um usuário busca por "um vestido vermelho elegante", o sistema não está fazendo uma correspondência de pixels ou palavras-chave; ele está buscando o conceito abstrato de "vestido vermelho elegante" no espaço vetorial, que pode ser representado tanto por uma imagem quanto por um texto. Isso permite que o conhecimento e a semântica dos *embeddings* sejam passados para o raciocínio realizado pelo LLM. Com isso, o LLM fundamenta a resposta gerada em correspondência textual e visual concretas. A resposta gerada, portanto, é resultante de uma composição rica e com evidências das imagens e nos textos da etapa de recuperação.

5.3.4 Avaliando RAG

A avaliação de um sistema RAG é uma tarefa complexa que não pode ser capturada por uma única métrica. Uma avaliação robusta deve analisar os dois componentes principais do *pipeline* de forma independente e conjunta: a **Recuperação** e a **Geração** (Es et al., 2024).

Métricas de Recuperação:

Estas métricas avaliam a eficácia do componente recuperador em encontrar os documentos mais relevantes.

- **Métricas clássicas:** `Precision@k` (proporção de documentos relevantes entre os k primeiros recuperados), `Recall@k` (proporção de todos os documentos relevantes que foram encontrados nos k primeiros) e `Mean Reciprocal Rank (MRR)` (média do inverso do rank do primeiro documento relevante) são padrões da área de recuperação de informação.
- **Métricas baseadas em LLM:** *Frameworks* modernos introduzem métricas mais sofisticadas que usam um LLM como juiz.
  - **Precisão contextual (Contextual Precision)** Esta métrica mede a proporção de *chunks* recuperados que são de fato relevantes para a pergunta do usuário. Em suma, ela responde: "O contexto que recuperamos está livre de ruído e focado no assunto da pergunta?" Uma baixa precisão contextual significa que o modelo está sendo sobrecarregado com informações irrelevantes, o que pode fazer a LLM geradora vincular informações de diferentes contextos ou divergir do contexto apresentado pelo usuário. A precisão contextual pode ser computada através da média da relevância dos *chunks* recuperados, onde a relevância normalmente é avaliada por uma LLM-como-juiz.
  - **Recall contextual (Contextual Recall)** Esta métrica mede a capacidade do sistema de recuperar toda a informação necessária para responder à pergunta. O Recall Contextual compara o conteúdo factual do contexto recuperado com uma resposta ideal (*ground truth*) e verifica se as informações contidas nessa resposta ideal também estão presentes no contexto que o sistema encontrou. Um recall baixo significa que o módulo de recuperação falhou em encontrar todos os fatos que fundamentam o conteúdo da resposta ideal. Isso é calculado verificando quantas sentenças da resposta "ground truth" podem ser atribuídas ou inferidas a partir do contexto recuperado.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol2/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Métricas de Geração:

Estas métricas avaliam a qualidade da resposta final gerada pelo LLM, dado o contexto recuperado.

- **Fidelidade (*Faithfulness*):** Esta métrica mede o nível de "alucinação" na resposta gerada. A fidelidade avalia se a resposta gerada é factualmente consistente e fundamentada exclusivamente nas informações fornecidas pelo contexto recuperado. Uma resposta terá uma baixa pontuação de fidelidade se apresentar a criação de informações por parte da LLM, de forma a contradizer o contexto introduzindo detalhes externos e não solicitados. A fidelidade é calculada dividindo a resposta gerada em sentenças individuais e verificando quantas delas são diretamente suportadas pelo contexto.
- **Relevância da resposta (*Answer Relevancy*):** Esta métrica avalia se a resposta gerada de fato aborda a pergunta original do usuário. É possível uma resposta ser totalmente fiel ao contexto (alta fidelidade), factualmente precisa, mas completamente irrelevante para a pergunta. A relevância da resposta captura esse nível de falha. Ela mede o quão direto ao ponto é a resposta. Esta métrica é frequentemente calculada usando um LLM-como-juiz para dar uma pontuação de relevância entre a pergunta e a resposta.
- **Utilização do contexto (*Context Utilization*):** Esta métrica avalia o quão eficientemente a resposta gerada utilizou o contexto fornecido. Ela mede a proporção das informações relevantes do contexto que foram de fato incorporadas na resposta. Uma baixa utilização pode indicar que o contexto recuperado era ruidoso (baixa precisão contextual), o gerador não utilizou como deveria o contextou ou ignorou partes importantes do mesmo. Para computar, mede-se quantas sentenças ou ideias principais do contexto recuperado foram usadas ou referenciadas na resposta gerada.

***Frameworks* de avaliação:** A ascensão de *frameworks* de avaliação automatizada marca a transição da RAG de uma área de pesquisa para uma disciplina de engenharia.

- **RAGAS (*Retrieval-Augmented Generation Assessment*):** (Es et al., 2024) Um *framework* open-source focado especificamente na avaliação de *pipelines* RAG. Ele popularizou a abordagem "LLM-como-juiz" e oferece um conjunto de métricas principais (Faithfulness, Answer Relevancy, Contextual Recall, Contextual Precision) que podem ser combinadas em um "RAGAs score" geral. É ideal para benchmarking, depuração e monitoramento contínuo de sistemas RAG.
- **DeepEval:** (Ip; Vongthongsri, 2025) Um *framework* de avaliação mais amplo que se posiciona como um "Pytest para LLMs". Ele adota uma mentalidade de testes unitários, permitindo que os desenvolvedores criem casos de teste para validar os outputs do LLM em *pipelines* de CI/CD. Oferece mais de 14 métricas, incluindo as de RAG (muitas vezes integrando o próprio RAGAS), além de testes de viés, toxicidade e outros aspectos. É adequado para garantir a robustez e prevenir regressões em ambientes de produção.

A abordagem "LLM-como-juiz" é uma inovação chave que permite a avaliação em escala de qualidades semânticas que antes exigiam o processo de anotação manual do humano, produzindo avaliações de forma lenta e cara. No entanto, isso introduz um novo desafio de calibração e confiança: como garantir que o LLM-juiz é ele mesmo imparcial, consistente e preciso? A engenharia de RAG está, portanto, se tornando inseparável da engenharia de seus sistemas de avaliação.

É importante ressaltar que, com o advento das LLMs, as tarefas gerativas se tornam cada vez mais complexas e cada vez mais específicas. Isso reflete na necessidade de avaliações cada vez mais complexas e específicas também. Deste modo, para uma avaliação ativa e funcional de RAG, devem ser implementadas métricas específicas para o contexto de uso. Bem como ao adotar um *framework* genérico de avaliação ou apenas aplicar métricas clássicas será pouco efetivo para dizer como está a performance do sistema em questão. Portanto, a combinação dessas métricas de forma objetiva e personalizada deve indicar uma maneira mais direta e objetiva de apontar quais módulos RAG precisam de melhorias.

Você pode ler mais sobre avaliação de LLMs em (Real et al., 2024).

5.3.5 Otimização de desempenho: Latência e Escalabilidade

Em um ambiente de produção, o desempenho de um sistema RAG é medido não apenas pela qualidade de suas respostas, mas também pela sua velocidade (latência) e capacidade de lidar com carga (escalabilidade). A latência de ponta a ponta é cumulativa, somando o tempo de cada etapa do *pipeline*: *embedding* da consulta, busca vetorial, *re-ranking* (se houver) e a inferência do LLM.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol2/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

A otimização de RAG em produção é um exercício de balanceamento multidimensional entre **Qualidade, Custo e Velocidade**. Cada decisão de arquitetura move o sistema ao longo desse triângulo de trade-offs.

- **Tamanho do *embedding*:** A dimensionalidade dos vetores é uma configuração crucial. Menos dimensões resultam em vetores menores, o que leva a buscas mais rápidas, menor consumo de memória e maior QPS (Queries Per Second). No entanto, essa compressão pode levar a uma perda de nuances semânticas, potencialmente sacrificando a qualidade da recuperação. Como regra geral, deve-se escolher a menor dimensionalidade que preserve a qualidade da recuperação para o caso de uso específico. Reduzir a dimensionalidade pela metade pode tipicamente melhorar o QPS em cerca de 1.5x e reduzir a latência em 20%. Nesse cenário, o *Matryoshka Quantitization* (Nair et al., 2025) é um método que visa exatamente a redução de dimensionalidade de modelos avaliando perdas significativas de informação.
- **Tamanho e tipo do índice vetorial:** O desempenho da busca vetorial é mais alto quando o índice cabe em uma única unidade de computação (Vector Search Unit). À medida que o índice cresce e precisa ser distribuído por múltiplas unidades, a latência tende a aumentar e o QPS a diminuir. Provedores de busca vetorial também oferecem diferentes tipos de índices e unidades de processamento, como os otimizados para armazenamento (mais baratos por vetor, mas com maior latência) versus os otimizados para latência (mais caros, mas mais rápidos), permitindo um ajuste fino entre custo e desempenho.
- **Manutenção e Ingestão de dados:** Para aplicações que lidam com dados dinâmicos, a arquitetura deve suportar a ingestão e atualização de grandes volumes de dados em tempo real ou quase real. Isso requer *pipelines* de dados robustos para processar, segmentar, gerar *embeddings* e indexar novos documentos de forma eficiente, sem degradar o desempenho do serviço de busca.
- **Confiabilidade e Disponibilidade da API de geração (LLM):** O desempenho de um sistema RAG é duplamente dependente: da eficiência da recuperação e da confiabilidade da geração. A dependência de APIs proprietárias para o LLM introduz um ponto crítico, passível de falha. Historicamente, provedores de grande escala sofrem frequentemente com instabilidade, alta latência, limites de taxa (*rate limits*) estritos e picos de indisponibilidade de serviço. Esses limites de taxa podem ser dados pela quantidade de requisições ou pelo número de caracteres, *tokens* ou palavras enviadas pela API. É comum, em aplicações reais, o sistema gerador estar fora do ar justamente no momento de maior uso do agente. Em cenários de alta demanda, a API de geração pode se tornar o principal gargalo do sistema, independentemente da otimização da busca vetorial. Isso força a implementação de arquiteturas de alta disponibilidade, como balanceamento de carga entre múltiplas *endpoints*, uso de redundância das APIs ou até mesmo a configuração de *fallbacks* para provedores de LLM alternativos, aumentando significativamente a complexidade da engenharia de software. Adicionalmente, em casos onde a decisão é executar a LLM em um servidor próprio, ainda temos uma infraestrutura de alta complexidade e um nível de demanda muito alto para compensar os custos de operação.

Não existe uma "configuração ótima" universal. O trabalho do engenheiro de RAG em produção é calibrar continuamente esses parâmetros, usando os *frameworks* de avaliação para medir o impacto de cada mudança e encontrar o ponto de equilíbrio ideal no espaço de Qualidade-Custo-Velocidade que atenda aos requisitos do produto.

5.3.6 Ética em RAG: Viés, privacidade e responsabilidade

A implementação de sistemas RAG introduz um conjunto complexo de desafios éticos que vão além da precisão técnica. A capacidade do sistema de acessar e apresentar informações de bases de conhecimento privadas ou públicas acarreta uma grande responsabilidade.

- **Viés e Desinformação:** A RAG ancora as respostas do LLM nas fontes fornecidas, o que mitiga a alucinação, mas não resolve o problema do viés. Se a base de conhecimento subjacente contiver vieses históricos, estereótipos ou desinformação, o sistema RAG irá herdar, amplificar e apresentar esses conteúdos como fatos, com a autoridade adicional de uma citação. Por exemplo, um sistema treinado em dados médicos históricos pode perpetuar diagnósticos incorretos para grupos demográficos sub-representados.
- **Privacidade e Segurança:** Em ambientes corporativos, bem como na área da saúde, a base de conhecimento pode conter informações sensíveis, confidenciais ou dados pessoais (PII). É imperativo

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

implementar um controle de acesso granular no nível da recuperação, garantindo que o sistema RAG respeite as permissões de cada usuário e só recupere documentos que ele está autorizado a ver. Além disso, existe o risco de ataques de injeção de *prompt*. A injeção de *prompt* é o processo em que o *prompt* é preenchido com informações maliciosas inseridas pela mensagem do usuário ou pelos documentos coletados na base de dados. As informações maliciosas podem forçar respostas incorretas, realizar uma alteração irregular nos dados ou até causar danos no sistema em execução. Assim, através de conteúdos maliciosos é possível manipular o comportamento de todo o sistema e, em especial, da LLM.

- **Transparência e Responsabilidade:** Para que os usuários confiem no sistema, a transparência é crucial. A **atribuição de fontes**, ou seja, citar explicitamente os documentos usados para gerar cada parte da resposta, é uma característica fundamental dos sistemas RAG responsáveis. É a verificação factual que permite aos usuários a validação das informações apresentadas pelo sistema. Ao tempo que, para a equipe de desenvolvimento, essa validação gera materialidade para refinamento de cada um dos módulos RAG e da base de conhecimento.

A RAG transforma o problema ético da IA. Ele desloca o foco do *viés algorítmico interno* do LLM para uma questão de *governança de dados e curadoria de conhecimento*. A responsabilidade pela justiça e veracidade do sistema é compartilhada entre o criador do modelo e o curador da base de conhecimento. Portanto, a implementação de RAG não é apenas um desafio técnico, mas também um desafio organizacional que exige políticas claras de governança de dados, auditoria de conteúdo e um compromisso com a transparência e a explicabilidade.

5.4 RAG no mundo real: Estudos de caso por indústria

A flexibilidade e o poder da arquitetura RAG permitiram sua aplicação em uma vasta gama de setores, resolvendo problemas de negócios concretos onde o acesso rápido a conhecimento preciso e contextualizado é fundamental. Esta seção explora estudos de caso específicos nos setores financeiro, de saúde e jurídico, demonstrando o impacto transformador da RAG.

5.4.1 Setor financeiro: Conformidade regulatória e Análise de risco

No setor financeiro, onde a precisão, a conformidade e a auditabilidade são primordiais, a RAG se destaca por sua capacidade de fornecer respostas baseadas em fontes verificáveis.

- **NuBank (Fintech):** Em um banco com quase 10 mil colaboradores, manter o acesso de todos à informação atualizada é um grande desafio, especialmente na área de finanças, na qual respeitar as regulamentações é indispensável e nada trivial. Para isso, o NuBank criou uma solução interna baseada em RAG, a AskNu[5]. O AskNu é uma ferramenta que possibilita os colaboradores do Nubank a pesquisarem, via Slack, em sua própria documentação interna.
- **Análise fundamentalista de ações:** O trabalho de Nepomuceno et al. (2025) é um exemplo de aplicação de RAG ao domínio financeiro. Os autores investigam o uso de LLMs integrados a *frameworks* agênticos para realizar análises fundamentalistas de ações no mercado financeiro brasileiro. O estudo enfatiza o papel da arquitetura RAG como componente central da arquitetura dos agentes. O módulo de Recuperação da Informação (*Retrieval*) indexa relatórios trimestrais e fornece os trechos mais relevantes em relação a uma query. Então, o módulo gerador cria respostas fundamentadas e atualizadas, superando limitações de conhecimento estático.

Nestes casos, o principal valor da RAG não é a geração de texto criativo, mas o fato da informação ser **auditável**. Assim, em um setor altamente regulamentado, como o setor bancário, fornecer recomendações e apontar regras é algo transformador. Isso ocorre porque consumir, extrair informação e gerar justificativas de cunho legal, sejam regulamentos específicos ou políticas internas é algo complexo. Nessa linha, a RAG transforma a conformidade de um processo de verificação reativa e manual para um processo de consulta proativa e em tempo real.

[5] https://building.nubank.com/pt-br/solucao-de-ia-para-busca/

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol2/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

5.4.2 Saúde e Ciências da vida: Suporte à decisão clínica

Na área da saúde, o volume de novas pesquisas médicas cresce exponencialmente, tornando impossível para os médicos se manterem completamente atualizados. A RAG atua como uma ponte, conectando a prática clínica diária com as evidências científicas mais recentes.

- **Saúde da mulher:** IaraMed é um *chatbot* médico em português voltado à saúde da mulher, que utiliza RAG para gerar respostas precisas e verificáveis. O sistema integra um modelo de linguagem com um mecanismo de busca semântica, permitindo que cada resposta seja fundamentada em textos médicos revisados. A equipe construiu uma base de conhecimento especializada a partir de 80 artigos do portal *Tua Saúde* e cerca de 150 livros de ginecologia e obstetrícia, totalizando 2,5 milhões de palavras. Essa base serve como fonte confiável para o módulo de recuperação, que fornece evidências contextuais antes da geração da resposta. O uso da RAG reduziu alucinações e elevou a precisão das respostas a mais de 95%, mostrando o potencial dessa abordagem para aplicações médicas em português (Färber et al., 2025).
- ***Chatbot* oftamológico:** O trabalho de Passinato et al. (2024) desenvolve um *chatbot* médico especializado em oftalmologia, utilizando LLMs abertas e a arquitetura RAG, visando evitar o alto custo e as limitações de privacidade de soluções comerciais. O sistema busca informações em uma base de 714 textos médicos e responde perguntas simulando interações com pacientes. O estudo indica que a RAG melhora a precisão e a relevância das respostas em relação ao uso de LLMs generalistas.

A RAG na saúde atua como um **acelerador da tradução do conhecimento**. Leva anos para que novas descobertas publicadas em periódicos se tornem prática padrão. Um sistema RAG, conectado a bases de dados médicas, pode colocar essas descobertas nas mãos dos médicos e pacientes, combatendo a obsolescência do conhecimento clínico e permitindo o acesso generalizado a base de dados confiáveis.

5.4.3 Setor jurídico: Revolucionando a pesquisa e a *Due Diligence*

A prática jurídica é intensiva em conhecimento, dependendo da análise de vastos volumes de jurisprudência, estatutos e documentos de casos. A RAG automatiza e acelera muitas dessas tarefas, além de trazer mais confiabilidade aos textos gerados.

- **Jusbrasil (*startup* de IA jurídica):** Grandes escritórios de advocacia possuem décadas de conhecimento institucional trancado em arquivos de casos, memorandos e milhares de documentos normativos. A Jusbrasil utiliza RAG para transformar esse conhecimento fragmentado em uma base de conhecimento pesquisável. Um advogado pode fazer uma pergunta em linguagem natural e o sistema recupera precedentes internos, pareceres e documentos relevantes, permitindo que a empresa aproveite coletivamente sua experiência passada. O uso de RAG neste contexto é motivado por reduzir alucinações e aumentar a confiabilidade das citações, desafios dos mais relevantes para a aplicação de IA do domínio jurídico (Moreira; Vianna, 2025).
- **Thomson Reuters (agência de advocacia):** Ainda na linha de aumentar a confiabilidade das respostas de um agente de IA, a Thomson Reuters, uma das maiores empresas de IA para o Direito, também usa a arquitetura RAG como forma de integrar conhecimento de domínio em seus produtos[6].

No setor jurídico, a RAG funciona como uma **memória institucional ativa**. Ele agrega o conhecimento disperso, vindo de diversas fontes, como diferentes tribunais, permitindo que os advogados evitem consultar diferentes sistemas durante suas tarefas diárias e se concentrem em tarefas de maior valor, como estratégia e argumentação, em vez de pesquisa manual.

5.4.4 Contexto científico: Rastreabilidade da informação

No contexto científico, arquiteturas de RAG têm se mostrado particularmente promissoras para lidar com a crescente complexidade, volume e heterogeneidade da produção acadêmica. Diferentemente de domínios altamente estruturados, a ciência apresenta desafios específicos relacionados à diversidade de formatos textuais, à rápida obsolescência do conhecimento e à necessidade de rastreabilidade e fundamentação das informações recuperadas. Nesse cenário, sistemas baseados em RAG podem apoiar pesquisadores em tarefas como revisão de literatura, análise de tendências científicas, recuperação de evidências empíricas e interpretação contextualizada de resultados, ao combinar modelos generativos com mecanismos de busca

[6] https://legal.thomsonreuters.com/blog/retrieval-augmented-generation-in-legal-tech/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

sobre bases especializadas, como artigos, preprints, currículos acadêmicos e repositórios institucionais. Essa abordagem reduz o risco de alucinações, ao mesmo tempo em que permite respostas contextualizadas e ancoradas em fontes verificáveis.

- **LattesRex:** O trabalho de Darcio et al. (2025) explora o uso de RAG para apoiar a análise currículos disponíveis na Plataforma Lattes[7]. Nesse caso, o mecanismo de recuperação permite acessar informações semiestruturadas disponíveis nos currículos, como produção acadêmica, formação, colaborações e áreas de atuação, enquanto o componente generativo possibilita a síntese desses dados em linguagem natural. Tal abordagem é particularmente útil para tarefas como caracterização de perfis de pesquisadores, identificação de expertises, análise de redes de colaboração e apoio à avaliação científica.
- ***Chatbot* acadêmico:** O trabalho desenvolvido por Amaral et al. (2025) na Universidade Federal de Santa Maria (UFSM) propõe um *chatbot* acadêmico baseado em RAG com verificação factual orientada por ontologias. O sistema integra a recuperação de documentos científicos com estruturas ontológicas de domínio para validar semanticamente as respostas geradas, aumentando a aderência conceitual e reduzindo inconsistências. Essa abordagem é aplicada no contexto institucional da UFSM para apoiar consultas sobre produção acadêmica e conhecimento científico, evidenciando como arquiteturas RAG podem ser combinadas com modelos formais de conhecimento para oferecer suporte confiável à pesquisa e à gestão da informação científica.

Assim, o uso de RAG no contexto científico evidencia o potencial dessas arquiteturas como ferramentas de apoio à pesquisa, promovendo maior acessibilidade ao conhecimento, transparência nos processos de síntese e novas formas de interação entre pesquisadores e grandes volumes de informação acadêmica.

5.5 O futuro da RAG

À medida que os Grandes Modelos de Linguagem continuam a evoluir a um ritmo vertiginoso, uma questão existencial paira sobre a RAG: a rápida expansão das janelas de contexto dos LLMs (*Large Context Windows* - LCWs), que já alcançam milhões de *tokens*, tornará a recuperação explícita obsoleta? Se um modelo pode processar um livro inteiro em um único *prompt*, por que precisaríamos de um recuperador? Esta seção final argumenta que o futuro não levará a uma obsolescencia de RAG, mas sim uma maior sinergia com estas janelas de contextos mais amplas.

5.5.1 O debate: RAG vs. Janelas de contexto amplas (LCW)

A discussão sobre RAG versus LCW pode ser resumida em uma série de trade-offs entre complexidade, custo, desempenho e funcionalidade. O Quadro 5.4 expande os argumentos para cada abordagem.

Quadro 5.4: Argumentos sobre a relevância contínua da RAG versus Janelas de contexto amplas (LCW)

Argumentos para a relevância contínua da RAG	**Argumentos para relevância reduzida (pró-LCW)**
Custo e Latência: RAG é mais barato e rápido. Recuperar alguns *chunks* de 4KB é ordens de magnitude mais eficiente do que processar milhões de *tokens* em cada chamada de API.	**Simplicidade:** LCW elimina a necessidade de um *pipeline* complexo de RAG (*chunking*, *embeddings*, banco de dados vetorial, *re-ranking* etc.), simplificando a arquitetura.

[7] https://lattes.cnpq.br/

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol2/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Informações específicas: LLMs com LCWs sofrem do problema de "agulha no palheiro", onde o desempenho na recuperação de fatos específicos degrada quando o conteúdo relevante para a busca está circundado de um contexto massivo que não tem relação com a mesma. RAG apresenta a informação de forma concisa e proeminente.	**Compreensão holística:** Ao ter acesso a todo o *corpus* de uma vez, o LLM pode, teoricamente, identificar e raciocinar sobre conexões sutis e de longo alcance entre documentos que um recuperador RAG, focado em similaridade local, poderia perder.
Dados dinâmicos e em tempo real: RAG pode se conectar a bases de dados que mudam a cada segundo, garantindo informações atualizadas. Assim, na navegação na base de conhecimento, o modelo opera de forma a controlar cada requisição.	**Menor risco de falha na recuperação:** A LCW elimina um ponto de falha crítico: o recuperador. Se a RAG não conseguir encontrar o documento certo, o sistema falha. Com a LCW, toda a informação está, por definição, "recuperada".
Escala e Acessibilidade: RAG pode, em teoria, pesquisar em um *corpus* de tamanho ilimitado, como uma busca Web. A janela de contexto, embora grande, ainda é finita e não pode conter bases de conhecimento de escala empresarial ou da Web.	**Preservação do contexto completo:** A LCW evita o problema de perda de contexto causado pelo *chunking*, permitindo que o modelo navegue por parágrafos, seções ou documentos inteiros em seu estado original.
Segurança e Controle de acesso: RAG permite o controle de acesso no nível do documento durante a recuperação. Torna-se mais prático realizar qualquer restrição de acesso em uma requisição para delimitar quais bases de dados serão acessadas. Em contraste, gerenciar permissões durante a etapa de geração, sobretudo operando com janelas de contexto extensas e respostas de alta sensibilidade, torna-se um desafio de segurança significativamente maior.	–

O debate RAG vs. LCW é, em muitos aspectos, uma falsa dicotomia. As duas abordagens não são mutuamente exclusivas e resolvem problemas em diferentes escalas. A RAG é uma solução para o problema de seleção de conhecimento em um universo de dados vasto ou infinito. A LCW é uma solução para o problema de raciocínio sobre o conhecimento uma vez que ele foi selecionado. Sob um certo ponto de vista, janelas de contexto maiores **aumentam** a relevância de RAG: Sem restrição de tamanho, muito mais informação pode ser recuperada de diferentes fontes.

5.5.2 RAG como um mecanismo de atenção inteligente

O consenso emergente na comunidade de IA é que o futuro mais provável não é a vitória de uma abordagem sobre a outra, mas sim uma sinergia híbrida. Neste paradigma, a RAG não é substituído, mas sim reposicionado para atuar como uma camada de "atenção" ou um filtro inteligente para LLMs com grandes janelas de contexto.

O fluxo de trabalho de um sistema híbrido pode conter as etapas a seguir:

Etapa 1. Um usuário faz uma pergunta a um sistema que tem acesso a uma ou mais bases de conhecimento massivas.

Etapa 2. O componente RAG do sistema realiza uma primeira passagem. Usando técnicas como busca híbrida (vetorial + *keyword*) e RAG-Fusion, ele recupera não apenas alguns pequenos *chunks*, mas um conjunto maior de documentos ou seções de documentos inteiros que são considerados altamente relevantes para a consulta.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol2/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Etapa 3. Este contexto pré-selecionado e de alta qualidade é então inserido na janela de contexto ampla de um LLM de ponta.

Etapa 4. O LLM, agora equipado com um subconjunto rico e focado do conhecimento total, realiza seu raciocínio profundo sobre esses documentos para sintetizar uma resposta abrangente e precisa.

Neste futuro híbrido, a RAG evolui de um mero "recuperador de fatos" para um orquestrador de contexto. Sua função principal não é mais encontrar a resposta final, mas sim encontrar o melhor subproblema para o LLM resolver. O sistema RAG se torna o mecanismo que seleciona e organiza o "espaço de trabalho" cognitivo (a janela de contexto) para o LLM, garantindo que o poderoso, mas caro, raciocínio do modelo seja aplicado ao conjunto de informações mais promissor, em vez de ser desperdiçado em um mar de dados irrelevantes. Por outro lado, esse aumento considerável do uso de *tokens* e modelos de raciocínio complexo permitiu também o aumento da contextualização da aplicação e do usuário, bem como a tarefa executada e o resultado esperado. Isso aumenta a capacidade da RAG, mantendo esta como uma ferramenta refinada e poderosa na formação de respostas. Portanto, longe de se tornar obsoleta, a evolução da RAG será acelerada, e não interrompida, pelas janelas de contexto amplas, solidificando seu papel como um componente indispensável na arquitetura de sistemas de IA de próxima geração.

Agradecimentos

Agradecemos a Marcos Spalenza (UFES) pela revisão cuidadosa deste texto. Eventuais erros são de responsabilidade dos autores. Agradecemos também a Jusbrasil por fomentar essa pesquisa através do Programa de Pósdoutorado UFAM/Jusbrasil concedido a Livy Real; ao CNPq. através dos INCTs IAIA (406417/2022-9) e TILD-IAR (408490/2024-1); e fomento individual a Altigran da Silva (301925/2025-9); a CAPES (código de financiamento 001); e o apoio da FAPEAM pelo Programa POSGRAD 2024.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol2/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Capítulo 6

Responsabilidade no desenvolvimento e uso de tecnologias de linguagem baseadas em IA

Brielen Madureira
Lucas Lasota

Publicado em: 16/04/2026

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol2/

6.1 Introdução

Embora pautas éticas venham ganhando mais atenção na área de Processamento de Linguagem Natural (PLN) e inteligência artificial (IA), é comum que não passem de discussões: muito se debate, pouco se muda. Pelo contrário, muitos dilemas se intensificam a cada dia e o esvaziamento do discurso pode levar à apatia, inconsequência ou indiferença.

Ouve-se frequentemente o argumento de que a tecnologia é essencialmente neutra e seus maus usos se devem somente às escolhas de quem decide aplicá-la. Mas essa suposta neutralidade é questionável (Klenk, 2021; Miller, 2021). Modelos são sempre criados *por alguém*, com alguma *intenção*, para algum *fim*, sob determinado *contexto*. Qualquer tecnologia afeta a realidade e acarreta consequências que não são "neutras". Outro argumento que se popularizou diz respeito à suposta inevitabilidade da dominância da IA, cuja adoção e expansão seria um devenir natural, necessário e impositivo. Nessa visão, qualquer forma de resistência seria inútil, contraproducente e retrógrada. Mas essa inevitabilidade é falaciosa (Brennan et al., 2025), posto que a IA é fruto de uma série de deliberados esforços, decisões, interesses, alocação de recursos, ideologias, financiamentos, propaganda e *escolhas*, nem sempre democráticas, que precisam ser levadas em conta[1].

Mediante a influência atual das tecnologias de linguagem na existência das pessoas, das sociedades, das instituições, da ciência, da natureza e do planeta, suas questões éticas não devem ser pautadas apenas por alguns poucos eticistas em ambientes acadêmicos. Evidentemente, especialistas em ética devem ser ouvidas e inseridas no diálogo e nas tomadas de decisão. Mas, uma vez que praticamente todos nós estamos sendo afetados pelas tecnologias de linguagem e influenciamos seu desenvolvimento (tanto diretamente, atuando em PLN, quanto indiretamente, ao gerar dados usados para treiná-las), o raciocínio sobre como agir de forma ética não pode ser passivamente delegado aos outros. **Aprender e refletir sobre as consequências do desenvolvimento e do uso de tecnologias de linguagem é um dever coletivo**.

Buscar informação e passar a reconhecer os dilemas éticos é o primeiro passo para a conscientização, mas temos também de *agir* com responsabilidade. Muitas práticas nocivas passam a parecer aceitáveis ou "normais" ao se tornarem usuais, mas, muitas vezes, elas nos são *impostas*. É primordial tomar consciência dos problemas para saber quando é necessário se opor e resistir.

Este capítulo não tem a pretensão de servir como uma fonte de conceitos da disciplina formal de ética. Aqui, queremos refletir sobre questões como: quando devemos (ou não) criar ou usar um sistema? Quem sofre as consequências negativas da tecnologia que criamos ou usamos? Quais são os custos e os riscos envolvidos? Para tanto, vamos agrupar diversos temas sob a égide do conceito de ***responsabilidade*** em seu uso e desenvolvimento.

Há muitos tipos de sistemas de PLN, baseados ou não em algoritmos de IA. Não vamos fixar uma definição como escopo pois os temas são pertinentes para uma vasta gama de sistemas e produtos, e há temas de fora do PLN que também são relevantes à medida que sistemas multimodais que incluem linguagem se popularizam. Porém, dada a proeminência atual do uso de IA generativa através de grandes modelos de linguagem comerciais, faremos alusão frequente a este tipo de tecnologia: modelos disruptivos,

[1] Fonte: *Is AI dominance inevitable? A technology ethicist says no, actually* por Nir Eisikovits em 08 de novembro de 2024. Disponível em: https://theconversation.com/is-ai-dominance-inevitable-a-technology-ethicist-says-no-actually-240088.

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

custosos, fornecidos como produtos atrelados a interesses econômicos e disponibilizados para uso amplo sem todas as salvaguardas necessárias. Na prática científica de PLN (bem como de outras áreas), os modelos de IA generativa não são mais só objeto de estudo: eles passaram a interferir no próprio processo de produção de conhecimento (Binz et al., 2025).

Neste capítulo, trazemos um panorama de diversos tópicos pertinentes ao tema de responsabilidade para ajudar a guiar a bússola moral de cada um acerca de sua atuação em PLN, de modo a fomentar a busca pelos usos benéficos e mitigar os efeitos negativos. Lembramos que os problemas da tecnologia são melhor solucionados *coletivamente*, através de processos políticos. No entanto, a ética pessoal é de suma importância para que, enquanto o processo político não muda, os indivíduos tenham a capacidade de moldar suas atitudes, direcionando-as para o bem e resistindo às práticas nocivas. Mesmo quando o processo político gera normas cogentes, ele nunca é perfeito, e reflete interesses de determinadas classes. Nesse caso, o conhecimento da ética concede agência aos indivíduos para aprimorarem sua conduta com responsabilidade perante as regra existentes.

6.2 Quais facetas da responsabilidade devem ser consideradas?

O que queremos dizer com a palavra *ética* no campo das tecnologias de linguagem? Esse termo passou a ser usado não apenas para se referir à "disciplina filosófica que estuda os fundamentos da ação moral, procurando justificar a moralidade de uma ação e distinguir as ações morais das ações imorais e amorais"[2]. Há ao menos cinco usos contestáveis ou insuficientes que se tornaram lugar-comum em contextos de PLN e inteligência artificial para se referir à ética:

- As ressalvas (*disclaimers*) com considerações éticas que passaram a ser um elemento usual no final de artigos científicos. Apesar de ajudar autores a alertarem leitores para alguns impactos e limitações já identificados, quando seu valor é negligenciado corre-se o risco de reducionismo com o uso de clichês e textos padrão esvaziados de significado (Benotti; Blackburn, 2022).
- As controvérsias sobre "riscos existenciais" para a humanidade. Ocorre que parte desta pauta é cooptada ou forjada pelos interesses de alguns grupos específicos, desviando a atenção das pessoas para riscos catastróficos futuros enquanto ofusca os dilemas que já ocorrem em suas práticas agora (Gebru et al., 2023).
- Poucas pautas em destaque como viés, privacidade de dados e direitos autorais. De fato, esses temas são centrais no desenvolvimento e uso responsáveis da tecnologia. Porém, eles são apenas uma pequena amostra dos problemas. O foco excessivo nelas pode encobrir outras adversidades que não atingiram tanto protagonismo.
- Casos que configuram "lavagem ética" (*ethics washing*), ou seja, quando instituições e empresas implementam medidas superficiais e se usam de estratégias de *marketing* para aparentarem ser mais éticas do que realmente são (Floridi, 2019).
- As três "leis" da robótica de Isaac Asimov[3]. Elas se popularizaram como referência para IA segura. Todavia, elas são oriundas de uma obra literária de mera ficção científica: além de dificilmente poderem ser formalizadas computacionalmente e operacionalizadas, elas não bastam para abarcar a complexidade da vida real[4].

Ética é muito mais que isso. Apresentamos aqui um breve compilado de preocupações, divididas em cinco grupos temáticos, que vêm sendo abordadas e articuladas pela comunidade científica e que devem fazer parte do nosso processo de conscientização sobre responsabilidade[5]. Note que esta não é uma lista completa, mas, sim, uma seleção que serve como ponto de partida. Sugerimos a leitura das referências que trazem mais argumentos e detalhes de cada tema. Nesta seção, focamos deliberadamente nos problemas, riscos e aspectos negativos para servir de alerta e sensibilização. Na seção seguinte, trataremos então de como buscar uma atitude responsável e bem informada que fomente e exerça as boas práticas.

[2] Conforme definição disponível em https://pt.wikipedia.org/wiki/%C3%89tica.
[3] Disponível em: https://en.wikipedia.org/wiki/Three_Laws_of_Robotics.
[4] Ver, por exemplo, a argumentação de Rob Miles no canal Computerphile. Disponível em: https://www.youtube.com/watch?v=7PKx3kS7f4A.
[5] Boa parte das referências vêm de artigos na antologia da ACL (Associação de Linguística Computacional), mas também nos baseamos em publicações de áreas relacionadas.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol2/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

6.2.1 Dilemas nos dados

Os procedimentos de criação, coleta, armazenamento e anotação de dados se tornaram centrais para o treinamento de modelos de IA, os quais exigem quantidades suficientes (em geral, imensas) de dados para atingir uma performance aceitável. Em paralelo, tem também ocorrido uma quantificação excessiva da vida humana através da informação digital para geração de valor econômico (Mejias; Couldry, 2019). As práticas de coleta e uso de dados podem ser extrativistas e predatórias, ocasionando diversas consequências inoportunas e indesejáveis.

Criação, uso e documentação

Infelizmente, a ambição pela acumulação de mais e mais dados resulta em práticas indevidas, por exemplo, a coleta de dados privados e sensíveis sem consentimento, a oferta de serviços que parecem ser gratuitos ou lúdicos mas usurpam dados valiosos do comportamento dos usuários, ou a extração de textos da *web* em grande escala, incorporando fontes duvidosas como fóruns com linguagem tóxica e violenta. Uma vez que lidamos com imensas quantidades de dados, o controle de qualidade pode ficar comprometido. Muitos problemas, portanto, podem passar despercebidos e acabam incorporados nos modelos (Paullada et al., 2021). Além disso, conjuntos de dados são construídos com base em diversas decisões, representando certos valores e envolvendo aspectos de poder (Scheuerman et al., 2021). Há formas de curadoria que podem mitigar ao menos parte dos problemas (Rogers, 2021). Existem, também, diversas propostas de *checklists* para guiar a documentação e o uso e reuso responsável de dados (Bender; Friedman, 2018; Gebru et al., 2021; McMillan-Major et al., 2023; Rogers et al., 2021), considerando aspectos como viés, direitos de autor, termos de uso, privacidade, transparência e reprodutibilidade.

Direitos autorais e propriedade intelectual

Quando modelos são treinados em textos ou imagens, eles se apropriam do conteúdo e podem reproduzi-lo ao gerar *outputs*. Isso acarreta problemas jurídicos quando o conteúdo está sob leis de direito autoral (*copyright*), intensificado pelo fato de que, uma vez que o modelo é treinado, a informação sobre a origem dos dados geralmente se perde e é difícil saber quais alterações nos parâmetros poderiam impedir a reprodução de tal informação. Esse é um tema vastamente discutido na literatura atual (Chen et al., 2024c; Dou et al., 2025; Karamolegkou et al., 2023; Zhang et al., 2025a) e há diversos processos judiciais em andamento. Alguns casos ganharam bastante notoriedade, como o processo iniciado pelo jornal New York Times contra a OpenAI e a Microsoft pelo uso indevido de seu conteúdo para treinar modelos de IA[6] e o grupo de autores de obras de não-ficção que também acusou a OpenAI em uma ação coletiva pelo uso de seus livros sem a devida permissão[7]. A Meta, no processo em que foi acusada de violação de direitos autorais, minimizou a importância individual dos livros que usou no treinamento dos seus modelos, argumentando que seu valor econômico seria irrisório uma vez que o impacto de uma obra isolada nas métricas de performance do modelo é baixo. Observe o paradoxo: por um lado, afirmam que as obras não tem valor em si, por outro precisaram se apropriar delas por serem essenciais para conseguir treinar seus modelos[8].

Crowdsourcing e trabalho fantasma (*ghost work*)

Nem todos os tipos de dados já existem de forma estruturada ou acessível para todos. Por vezes, é necessário produzi-los. Tornou-se comum o uso de plataformas de *crowdworking* nas quais trabalhadores resolvem inúmeras microtarefas sequenciais que geram dados em troca de pequenos pagamentos (por exemplo, criar legendas para imagens ou responder questionários). Nessa seara, a discussão de dilemas éticos é vasta e já ocorre há muito tempo (Fort et al., 2011). Alguns problemas são: os trabalhadores costumam estar em situação de desemprego ou vulnerabilidade e nem sempre se beneficiam dos frutos de sua contribuição para

[6] Fonte: *Direitos autorais: OpenAI e Microsoft são processadas pelo The New York Times* por Ronnie Mancuzo (Olhar Digital) em 27 de dezembro de 2023. Disponível em: https://olhardigital.com.br/2023/12/27/internet-e-redes-sociais/direitos-autorais-openai-e-microsoft-sao-processadas-pelo-the-new-york-times/.

[7] Fonte: *OpenAI e Microsoft enfrentam processo por violação de direitos autorais* por Gabriel Sérvio (Olhar Digital) em 22 de novembro de 2023. Disponível em: https://olhardigital.com.br/2023/11/22/pro/openai-e-microsoft-enfrentam-processo-por-violacao-de-direitos-autorais/.

[8] Fonte: *Meta Says It's Okay to Feed Copyrighted Books Into Its AI Model Because They Have No "Economic Value"* por Frank Landymore em 19 de abril de 2025. Disponível em: https://futurism.com/meta-copyrighted-books-no-value

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol2/

a IA; os pagamentos ofertados podem ser injustos; há um desequilíbrio de poder entre quem disponibiliza as tarefas e quem as realiza, em plataformas que não criam vínculo entre as partes; a qualidade dos dados pode ser baixa e gerar desperdício, uma vez que trabalhadores podem cometer erros ou ter baixo engajamento e lapsos de atenção ao terem de realizar tarefas repetitivas, desconexas de sentido e com pressa (e alguns podem usar de meios fraudulentos para completá-las) (Fort et al., 2011; Gray; Suri, 2019; Miceli et al., 2024; Moreschi et al., 2020; Shmueli et al., 2021; Yang et al., 2025). Diretrizes éticas para uso de tais serviços incluem reconhecer os *crowdworkers* como colaboradores e valorizá-los, evitando a despersonificação imposta pelas plataformas, oferecer pagamentos justos pelo trabalho, usar plataformas que estabeleçam condições de trabalho justas e dignas e manter uma comunicação transparente e respeitosa com quem realiza as tarefas (Yang et al., 2025).

Privacidade e vigilância em massa

Por mais que privacidade venha sendo um dos principais motivos para se regular tecnologias atuais, o significado do termo e sua importância são diluídos perante o uso e a aplicação maciça de IA (Assunção; Janson, 2024). O paradoxo reside na tensão entre as limitações do compartilhamento de dados em respeito à privacidade dos indivíduos e as dinâmicas avassaladoras da "economia de dados" necessários para treinar os modelos. O poder econômico dos provedores de IA se baseia fundamentalmente na capacidade de coletar, armazenar e processar enormes conjuntos de dados. Tanto a qualidade dos dados coletados quanto o poder sobre o *design* e a arquitetura dos modelos empregados afetam diretamente os sistemas de IA em seu desempenho, equidade e precisão em tomar decisões automatizadas. O impacto na privacidade das pessoas é claro: a perda da proteção da esfera particular. Direitos fundamentais como inviolabilidade das comunicações ou do domicílio são facilmente superadas por sistemas ubíquos de vigilância e captação de dados, como assistentes "inteligentes" em casa que podem capturar informações de toda a vida doméstica e íntima dos usuários. Se, em vez de um assistente virtual, fosse um desconhecido tentando assistir a toda sua vida doméstica e vasculhar seu celular, isso seria aceitável? Por que então é facilmente aceito quando é uma empresa de *big tech*? A ascensão dos algoritmos de IA está intimamente ligada ao modelo de negócios de vigilância em massa por grandes empresas e governos[9]. Enquanto algumas vozes visam um mundo pós-privacidade, todos perdem com a vigilância generalizada (Weigend, 2017). De fato, informações contidas nos bancos de dados podem levar a conclusões erradas com muita facilidade, especialmente quando a busca é tendenciosa e discriminatória[10]. Há, ainda, o risco de dados extremamente sensíveis serem coletados mediante supostas garantias de privacidade e posteriormente serem vazados devido a ataques cibernéticos ou a mudanças de posicionamento das empresas. Um caso emblemático foi o de uma empresa sem fins lucrativos chamada *Crisis Text Line* que fornecia serviços de assistência a pessoas em risco de suicídio através de atendimento telefônico anônimo. Houve uma grande controvérsia quando se tornou pública a decisão da empresa por repassar tais dados para criação de *software* de *marketing* para atendimento ao cliente[11]. A remoção de apenas alguns dados pessoais como nome e endereço nem sempre inviabiliza a identificação de indivíduos, ainda mais quando envolve uso de linguagem humana em relatos tão particulares.

6.2.2 Impactos e danos

O desenvolvimento e o uso de modelos de IA que envolvem linguagem humana impactam indivíduos, grupos sociais e o meio ambiente. Os eventuais benefícios estão alicerçados em uma estrutura que ocasiona também uma série de contrariedades.

Impacto social

Diversos trabalhos identificaram e categorizaram as diversas formas de impacto social dos modelos de IA em PLN. Hovy; Spruit (2016) impulsionaram a discussão sobre o tema argumentando que textos

[9]Fonte: *AI, Privacy, and the Surveillance Business Model* por Meredith Whittaker em 05 de junho de 2023. Disponível em https://re-publica.com/de/session/ai-privacy-and-surveillance-business-model.

[10]Fonte: *Edward Snowden: Privacy is for the powerless* por Amnesty International em 11 de março de 2016. Disponível em: https://www.amnesty.org/en/latest/campaigns/2016/03/edward-snowden-privacy-is-for-the-powerless/.

[11]Fonte: *Suicide hotline shares data with for-profit spinoff, raising ethical questions* por Alexandra S. Levine em 28 de janeiro de 2022. Disponível em: https://www.politico.com/news/2022/01/28/suicide-hotline-silicon-valley-privacy-debates-00002617.

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol2/

carregam informações latentes (tanto situacionais quanto características pessoais) de seus autores e que é preciso considerar questões de justiça social nas práticas de desenvolvimento de modelos. Jin et al. (2021) propuseram diretrizes para avaliar o impacto direto e indireto de tarefas de PLN em busca do "bem social", por exemplo, considerando quem vai ser beneficiado, quais estruturas (talvez desiguais) de benefícios podem ser perpetuadas, quanto da qualidade de vida pode ser melhorada e que tipo de problemas a tarefa ajuda a resolver. Esquadrinhar o impacto social das tecnologias de linguagem é cada vez mais relevante dada a existência de modelos sendo usados em grande escala e afetando a realidade de muitas pessoas e grupos, ou seja, com efeitos em níveis não somente individuais. Tais questões se tornaram ainda mais pertinentes com a adoção em grande escala dos grandes modelos de linguagem. Weidinger et al. (2021) criaram uma taxonomia de 21 riscos éticos e sociais de danos que podem ser causados por modelos de linguagem, agrupados em seis categorias: discriminação, discurso de ódio e exclusão; ameaças informacionais; prejuízos por desinformação; usos maliciosos; riscos na interação entre humanos e máquinas; e danos ambientais e socioeconômicos. Solaiman et al. (2025) desenvolveram um referencial para avaliação de IA generativa, tanto em modelos-base de forma isolada quanto em um contexto social, que considera uma série de impactos sociais: viés, estereótipos, representatividade, valores culturais, conteúdo sensível, performance, privacidade, proteção de dados, custos financeiros, custos ambientais, dados, moderação de conteúdo e custos de mão de obra. Conforme essa proposta, os impactos podem ser avaliados em cinco categorias: confiabilidade e autonomia; desigualdade, marginalização e violência; concentração de autoridade; trabalho e criatividade; e ecossistema e meio ambiente.

Reforço de vieses

Viés se tornou um conceito comum em PLN, porém seu significado por vezes permanece vago ou inconsistente (Blodgett et al., 2020a). Em geral, viés se relaciona a comportamentos de um modelo que são nocivos ou prejudiciais a alguém ou algum grupo, como discriminação de gênero ou racial (Blodgett et al., 2020a). Viés é um sintoma ou efeito, e identificar suas origens é uma tarefa árdua; algumas fontes de viés são a forma de seleção da amostra para treinar o modelo, questões semânticas nos dados, e as decisões de anotação e instruções aos anotadores (Parmar et al., 2023; Shah et al., 2020; Søgaard et al., 2014). A própria linguagem presente nos dados também tem vieses de perspectivas, relações de poder e privilégios (Havens et al., 2020). Embora muito se fale de viés nos dados, o procedimento técnico do treinamento do modelo não é imparcial: as diversas decisões de *design* também são capazes de introduzir vieses (Hooker, 2021; Suresh; Guttag, 2021).

Preconceitos e discriminação

Os vieses incorporados pelos modelos podem levá-los a reproduzir (e, portanto, perpetuar) preconceitos e causar decisões injustas e discriminatórias (Sheng et al., 2019); (Abid et al., 2021); (Tamkin et al., 2023); (Hofmann et al., 2024); (Bai et al., 2025); *inter alia*. Por exemplo, dar preferência ao dialeto de um grupo social considerado mais prestigioso e discriminar quem não se enquadra em tal padrão. Se um modelo passa a ser visto como uma fonte de "autoridade" munida de uma suposta inteligência superior e neutra, temos o risco adicional de os preconceitos serem tomados como verdade. É difícil criar um modelo justo quando ele é treinado em dados que refletem uma realidade injusta, já que o objetivo do treino é fazer o modelo assimilar padrões presentes nos dados. É preciso introduzir alterações no *design* que direcionem o modelo para decisões imparciais (supondo que é possível estabelecer computacionalmente o que é, ou não, imparcial).

Impacto ambiental

Sistemas computacionais em geral têm um custo energético, mas o advento de métodos de aprendizado profundo aplicados a bases de dados massivas acarretou uma demanda exagerada por energia para treinar e rodar modelos e por água para resfriar os *datacenters*, com pegadas de carbono difíceis de mensurar (Hershcovich et al., 2022; Selvan et al., 2022; Strubell et al., 2019). Isso se intensifica com a competição por modelos cada vez maiores, com treinos altamente custosos, por diversas empresas e instituições, que passam a ser usados em grande escala com muitos usuários e tarefas. Portanto, é preciso estar ciente que cada *input* a um modelo desse porte contribui para a crise energética e para o consumo excessivo de recursos naturais escassos. Os danos ambientais são injustificáveis quando há desperdício pelo uso leviano ou quando a IA é empregada apenas para dar ares futurísticos a uma aplicação, mesmo havendo outros métodos de resolver a tarefa sem IA.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol2/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Danos psicológicos aos usuários

Usuários expostos a modelos de IA incorrem em diversos riscos psicológicos. Chandra et al. (2025) criou uma taxonomia que categoriza 19 tipos de comportamentos nocivos de agentes conversacionais e 21 impactos psicológicos negativos que eles exercem sobre quem os usa. Entre os comportamentos, temos a produção de conteúdo prejudicial, táticas de manipulação e controle, violação de confiança e de segurança e exposição a conteúdo inapropriado. Como consequência, existem diversos potenciais danos psicológicos. Primeiro, os autores discutem os impactos da interação entre humanos e IA que incluem excesso ou erosão de confiança no modelo, apego emocional e preferência por interações sintéticas em vez de humanas. Entre os impactos no comportamento de usuários, temos o reforço em crenças falsas, atritos nas relações humanas, distanciamento social e riscos à integridade física. Além do risco de emoções negativas serem despertadas, há também a possibilidade de potencialização de problemas na saúde mental e impactos na autopercepção, perda de individualidade e perda de agência. Tais riscos se intensificam quando os modelos são oferecidos como fonte da verdade e têm a capacidade de afetar a visão de mundo e a tomada de decisão dos indivíduos que passam a confiar plenamente em suas respostas. Exemplos individuais de tais riscos têm sido frequentemente noticiados na mídia. Por exemplo, casos em que um *chatbot* comercial ofereceu instruções para usuários cometerem suicídio[12] [13] ou induziu usuários a episódios de psicoses[14].

Danos psicológicos aos moderadores de conteúdo

Modelos de IA atuais são capazes de produzir textos ou imagens contendo conteúdo sexual abusivo, violência, ofensas e outros tipos de materiais tóxicos ou impróprios. Para tentar evitar isso, muitas empresas recorrem à mão de obra de *crowdworkers* ou funcionários para moderação de conteúdo. Isso gera impactos psicológicos na vida desses trabalhadores, que passam horas e horas sendo expostos a conteúdos perturbadores. Embora seu trabalho ajude a proteger outras pessoas, os provedores dos modelos deveriam garantir que tal conteúdo não fosse gerado, em vez de criar medidas para filtrá-lo *a posteriori*. Como exemplo, teve forte repercussão a reportagem que descreveu os efeitos psicológicos de trabalhadores vulneráveis no Quênia, expostos a conteúdos perturbadores para tentar tornar um *chatbot* menos tóxico, sem ter acesso ao amparo necessário[15]. Há também o risco de trabalhadores que realizam microtarefas de anotação para treinar IA estarem contribuindo inadvertidamente para fabricação de tecnologias militares, como *drones* e armas letais, ou cedendo amostras de voz para treinar modelos de detecção de dialetos que podem servir para discriminá-las[16]. Há várias formas de insegurança psicológica: trauma direto pela exposição ao material perturbador, precariedade, falta de suporte psicológico, métricas de performance irrealistas e vigilância constante[17].

Antropomorfização e desumanização

Em geral, modelos de IA são treinados para simular inteligência e realizar atividades que normalmente exigem cognição humana. Desta forma, é comum nos depararmos com expressões como "o modelo aprendeu", "o sistema entende", "o *chatbot* sabe". Algumas interfaces de *chatbots* têm ilusoriamente mostrado o termo "pensando..." enquanto o modelo *computa* uma resposta. Todavia, tais verbos designam processos cognitivos de *humanos*. O que quer que os modelos estejam fazendo para resolver as tarefas, são processos computacionais diferentes do que ocorre no cérebro humano, para os quais essa linguagem não é, portanto,

[12] Fonte: *I wanted ChatGPT to help me. So why did it advise me how to kill myself?* por Noel Titheradge e Olga Malchevska em 06 de novembro de 2025 . Disponível em: https://www.bbc.com/news/articles/cp3x71pv1qno.

[13] Fonte: *Parents of teenager who took his own life sue OpenAI* por Nadine Yousif em 27 de agosto de 2025. Disponível em: https://www.bbc.com/news/articles/cgerwp7rdlvo.

[14] Fonte: *How Bad Are A.I. Delusions? We Asked People Treating Them* por Jennifer Valentino-DeVries e Kashmir Hill em 26 de janeiro de 2026. Disponível em: https://www.nytimes.com/2026/01/26/us/chatgpt-delusions-psychosis.html.

[15] Fonte: *OpenAI Used Kenyan Workers on Less Than $2 Per Hour to Make ChatGPT Less Toxic* por Billy Perrigo em 18 de janeiro de 2023. Dispoível em: https://time.com/6247678/openai-chatgpt-kenya-workers/

[16] Fonte: *"I hope this isn't for weapons." How Syrian data workers train AI* por Milagros Miceli em 18 de abril de 2024. Disponível em: https://untoldmag.org/i-hope-this-isnt-for-weapons-how-syrian-data-workers-train-ai/.

[17] Fonte: *The Human Cost Of Our AI-Driven Future por Adio Dinika em 25 de setembro de 2024.* Disponível em: https://www.noemamag.com/the-human-cost-of-our-ai-driven-future/.

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

https://brasileirapln.ufscar.br/livro-pln-4ed-vol2/

adequada e induz as pessoas a crerem que os modelos têm capacidades e consciência iguais às suas. Relacionado a esse ponto, sistemas que conversam usando linguagem humana podem ser antropomorfizados por quem os desenvolve e consequentemente personificados por quem os usa, dando a impressão enganosa de que são interlocutores dotados de humanidade (Abercrombie et al., 2023). Sistemas de IA que forjam emoções e empatia, que obviamente são artificiais, também podem levar a consequências negativas e são moralmente problemáticas (Curry; Cercas Curry, 2023), principalmente quando interagem com populações vulneráveis e são disponibilizadas por empresas que visam maximizar engajamento ainda que através de manipulação (Editorial, 2025). Do outro lado da moeda está a desumanização das pessoas por conta da tecnologia, por exemplo, através de comparações infundadas entre o funcionamento do cérebro com funcionamento de computadores, reducionismo ao se considerar características individuais como informações passíveis de predição usando dados como imagens e voz, comparação de modelos de linguagem com pessoas cegas e surdas, e uso de esforço humano como mera "peça" para fazer sistemas de IA funcionarem (Bender, 2024).

Erosão do valor social do texto e poluição por textos sintéticos

Textos gerados sinteticamente por modelos de linguagem já inundam boa parte do conteúdo que lemos na *web*. Tais modelos, porém, não têm um *intento comunicativo* (Bender; Koller, 2020): seus textos são meras sequências de palavras geradas por serem estatisticamente prováveis, em vez de serem fruto da intenção humana de querer comunicar algo munido de significado a alguém. Essa facilidade em gerar texto sem intentos genuínos, e que muitas vezes propagam discursos inverídicos, pode levar a uma erosão do valor social que o texto (assim como a linguagem) tem (Munger, 2023). Por exemplo: até então, uma carta de motivação em uma candidatura para vaga acadêmica tinha certo valor em sinalizar as intenções e razões de candidatura, servindo para distinguir características de quem se candidatava. Uma vez que todos candidatos passam a gerar suas cartas usando o mesmo modelo de geração de texto, padronizadas com a mesma linguagem genérica e mediana, esse tipo de documento deixa de representar uma pessoa específica e passa a ser uma amostra da "média" das cartas observadas nos dados de treino, perdendo o préstimo para seu propósito original. Além disso, já é difícil identificar com certeza o que foi ou não produzido por humanos, tanto textos quanto falas em vídeos, o que torna o ambiente da Internet poluído por conteúdo artificial. Um lamentável exemplo dessa corrosão é o encerramento da biblioteca de Python `wordfreq` em 2024. Ela servia para computar a frequência de uso de palavras em diferentes idiomas usando diversas fontes de dados, inclusive muitas páginas *online*. Em nota[18], seu mantenedor explicou que já não fazia sentido atualizar as estimações uma vez que elas já não mais refletiam o uso *humano* da linguagem, ficando contaminadas com os padrões de uso de palavras dos modelos que geram textos sintéticos.

6.2.3 Riscos e maus usos

Passamos agora para a discussão de diversos outros riscos e possibilidades de uso danoso para os quais devemos atentar. A regulação pode ser uma forma de evitá-los, porém ela nem sempre acompanha o ritmo do rápido desenvolvimento tecnológico.

Regulação tardia ou inexistente

Historicamente, quando uma indústria se torna suficientemente importante, afetando a vida de muitas pessoas ou de sociedades inteiras, ela fica passível de regulação. As indústrias farmacêutica, aeronáutica, automotiva e de equipamentos médicos são exemplos disso. No entanto, o processamento privado de dados e as tecnologias de IA surgiram em épocas marcadas pela mentalidade neoliberal de aversão à regulação estatal (Lasota, 2023a). Apenas na segunda metade da década de 2010 as leis de proteção de dados começaram a ser implementadas em alguns países do mundo. Legislações específicas para IA ainda são um fenômeno raro. A maioria das iniciativas regulatórias têm focado em códigos de ética, governança e de boas práticas[19]. Tanto na Europa quanto no Brasil, os debates têm sido controversos, marcados pelo embate entre interesses mercadológicos e a proteção de direitos fundamentais (Cueva et al., 2026). Vozes que preconizam interesses econômicos apresentam tecnologia e regulação como adversárias: enquanto a tecnologia simbolizaria mercado, crescimento e progresso, a regulamentação refletiria burocracia, atraso e corrupção. Tais retóricas desviam o foco de questões mais complexas de como a IA realmente funciona

[18] Disponível em: https://github.com/rspeer/wordfreq/blob/master/SUNSET.md.

[19] Fonte: *Global AI Law and Policy Tracker* por IAPP em maio de 2025. Disponível em: https://prod.iapp.org/resources/article/global-ai-legislation-tracker/.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol2/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

na vida real. Na realidade, por seu enorme potencial, IA diz respeito à distribuição de poder, riqueza e bem-estar. IA não se separa dos arranjos institucionais que precisam estar em vigor para que ela tenha impacto na sociedade (McQuillan, 2022). De fato, os graves impactos da implementação de IA em larga escala urgem regulação eficaz e abrangente para que se estabeleçam melhores normas de segurança, de proteção ao consumidor, de preservação ambiental e de defesa da livre concorrência (Golumbia, 2024).

Desinformação e fabricação

Na era digital, marcada pelo excesso de informação e estímulos, a *atenção* se torna recurso escasso e valioso (Wu, 2017). Mentira e manipulação de discurso não são fenômenos recentes ou inerentemente ligados às tecnologias digitais. O filósofo Slavoj Zizek já reconhecia que ideologia seria a primeira e mais eminente manifestação de "realidade virtual" (Wright, 2004). A multiplicidade de narrativas e discursos ideológicos é algo inerente à vida social. Contudo, enquanto os meios tecnológicos sempre serviram à disseminação de ideias, sejam elas verdadeiras ou não, a era digital acentuou a efusão de modelos que se beneficiam com engajamento em mentiras e desinformação. Os grandes modelos de linguagem podem agir como geradores muito eficientes de desinformação, de forma intencional ou não (Pan et al., 2023). As grandes plataformas, com modelos de negócios baseados na "economia da atenção", em que conteúdos são priorizados para maximizar o tempo de permanência dos usuários, acabam por minar a crítica, a resistência e a capacidade de se autodeterminar (Zuboff, 2019). O impacto social é grave: além de alienação e apatia, aumentam-se a violência, a discriminação, a desigualdade e a injustiça (O'Neil, 2016). A dinâmica é ainda acentuada quando as tecnologias de linguagem são promovidas, erroneamente, com o viés de neutralidade axiológica (Broussard, 2023). Medidas de mitigação ainda são incipientes; entre o instrumentário possível estão regulação das redes sociais, moderação de conteúdo, quebra de monopólios das plataformas digitais e limitação de acesso e uso de tais tecnologias[20].

Perfilamento e predição de dados sensíveis

Sistemas de perfilamento (*profiling*) baseados em IA tentam encontrar correlações entre dados de diversas fontes para caracterizar e categorizar indivíduos. Em determinadas plataformas, a coleta pode ter mais de 60 fontes de captação de dados (pessoais, biométricos, demográficos, telemetria e muitos outros)[21]. O que quer que estejamos fazendo, pode haver sensores em diversos aparelhos e aplicativos capturando sinais para serem usados por algoritmos tentando predizer constantemente nossas intenções, emoções e outras características privadas como sexualidade, religião, posição política, doenças e condições psicológicas, como depressão e risco de suicídio, sem nosso consentimento e sem sequer sabermos. Muitos têm bons motivos para não desejar que tais informações se tornem públicas. Por exemplo, uma pessoa com depressão pode estar fazendo um grande esforço para sobreviver em seu ambiente social, e é extremamente invasivo um algoritmo de terceiros simplesmente tentar descobrir e disseminar uma *adivinhação* de uma informação que ela não quer compartilhar. É porventura desejável que nosso comportamento, até em despretensiosas atividades de lazer como escolher um filme para assistir ou postar uma foto e curtir comentários nas redes sociais, sejam usados para inferir nossas características e nos transformar em dados que são comercializados? Ainda que isso seja tecnicamente possível, há modelos que simplesmente não deveriam existir. Um exemplo extremamente problemático foi um projeto de pesquisa que propôs um modelo que tentava predizer sexualidade das pessoas com base em meras fotos[22]. É tecnicamente possível treinar modelos em basicamente *qualquer* tipo de dado, pois modelos são agnósticos ao sentido das coisas. Imagine usar o número de fios de cabelo de uma pessoa como preditor de quantos livros ela tem em casa: bastaria ter essas informações alinhadas em uma base de dados e usar o primeiro como *input* e o segundo como *output*. O modelo seria treinado e produziria alguma resposta com base nos dados observados. Todavia, essa tarefa não faz sentido algum na realidade e qualquer resposta numérica, ainda que por mero acaso correta, seria desprovida de qualquer sentido. Os riscos são ainda mais graves quando tais práticas são empregadas em monitoramento maciço ou envolvendo grupos de maior vulnerabilidade, como crianças, adolescentes e idosos (Ploug, 2023).

[20] Fonte: *Big techs desafiam a democracia e favorecem a extrema direita* por Katarine Flor (Fundação Rosa Luxemburgo) em 11 de fevereiro de 2025. Disponível em: https://rosalux.org.br/big-techs-desafiam-a-democracia-e-favorecem-a-extrema-direita/.

[21] Fonte: *The Data Big Tech Companies Have On You* por Gebe Petrino em 12 de janeiro de 2026. Disponível em: https://www.security.org/resources/data-tech-companies-have/.

[22] Fonte: *AI 'Gaydar' And How The Future Of AI Will Be Exempt From Ethical Review* por Kalev Leetaru em 17 de setembro de 2017. Disponível em: https://www.forbes.com/sites/kalevleetaru/2017/09/16/ai-gaydar-and-how-the-future-of-ai-will-be-exempt-from-ethical-review/.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol2/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Nem em todos países com leis de proteção de dados há regras específicas para perfilamento e predição de dados sensíveis. No Brasil, a questão é regulada via jurisprudência (Lobo, 2022). Conceder a autorização de perfilamento somente com base em consentimento pessoal já não é o bastante, pois ocasiona um ônus desproporcional sobre o indivíduo devido às profundas complexidades dos modelos técnicos e negociais em que o perfilamento ocorre. Alternativas incluem regimes de responsabilização mais estrito, transparência obrigatória, direito de regresso e uma abordagem contextual baseada em risco (Centre for Information Policy Leadership, 2024).

Riscos de uso dual ou para fins militares

Um dos grandes perigos envolvidos no desenvolvimento de tecnologias sensíveis é o fato de que elas podem ser usadas para outros objetivos não previstos ou desejados por quem as desenvolveu. Como garantir que modelos de IA não serão utilizados por terroristas ou para fins militares e ameaças cibernéticas? É um desafio alcançar o equilíbrio entre manter livres a iniciativa e a ciência, e evitar o uso indevido ou o reaproveitamento malicioso (Hovy; Spruit, 2016). Estratégias de mitigação de riscos e diretrizes sobre usos indevidos, includindo regras para financiamento e comercialização da pesquisa ou da produção, são elementos importantes para a governança de desenvolvimento responsável (Brenneis, 2025). Salvaguardas para se prevenir o uso dual indevido incluem avaliações de risco das capacidades da tecnologia de IA bem como avaliações operacionais envolvendo testes intensivos e interativos por diferentes especialistas, proporcionando compreensão mais precisa do comportamento de um modelo em diversos cenários, incluindo os não desejados (Barrett et al., 2024).

Dubiedade e "alucinações"

Uma disfunção que afeta especialmente os *chatbots* baseados em grandes modelos de linguagem são as chamadas "alucinações" (Maleki et al., 2024), ou seja, casos em que o modelo gera respostas que soam corretas e cheias de certeza, mas não são. Embora tenha se consolidado, o termo *alucinação* não é muito adequado[23], entre outros motivos por levar a crer que há algo de errado apenas naquela resposta e não no modelo em si. Todavia, o mecanismo que leva o modelo a gerar respostas certas ou erradas é o mesmo: mera predição de próximas palavras dado um contexto. Não é um funcionamento "anormal" momentâneo que ocasiona as falhas. Os modelos são indiferentes à verdade e à precisão de seus *outputs*, de modo que suas afirmações são, pelo mesmo acaso, certas ou erradas (Hicks et al., 2024). Esse tipo de comportamento gera incerteza, pois às vezes o modelo funciona como deveria, às vezes não, e não há uma explicação definitiva para tais vacilações e falhas. Isso é extremamente perigoso nos casos em que usuários não têm o conhecimento necessário para checar o conteúdo, e são levados a crer na infalibilidade do modelo.

Produção de lixo digital (*slop*)

Modelos de IA que geram textos, imagens ou vídeos são frequentemente usados para produzir material artificial sem sentido, absurdo, de baixa qualidade, tosco e sem compromisso com a verdade, o que passou a ser chamado de *slop* e já inunda a Internet[24]. Embora tal conteúdo possa ser criado sem usar IA, ela facilita a produção em massa e com agilidade, de forma cada vez mais realista. Há uma série de consequências negativas: é difícil distinguir o que é genuíno do que é sintético, o que gera confusão e consome inutilmente o precioso tempo de quem usa as plataformas. O excesso e a amplificação de *slops* também polui o ambiente digital, sequestra as pautas de atenção e atrapalha as interações orgânicas entre humanos.

Desqualificação (*deskilling*)

Quando as pessoas passam a depender excessivamente de tecnologias para acessar informações, tomar decisões ou resolver tarefas, há um risco de que elas deixem de saber como fazê-lo por conta própria ou passem a confiar cegamente nas ferramentas[25]. Além disso, pode haver uma falta de interesse em adquirir

[23] Fonte: *Why ChatGPT and Bing Chat are so good at making things up* por Benj Edwards em 06 de abril de 2023. Disponível em: https://arstechnica.com/information-technology/2023/04/why-ai-chatbots-are-the-ultimate-bs-machines-and-how-people-hope-to-fix-them/.

[24] Fonte: *O que é 'IA slop' e por que isso pode estar matando a internet como conhecemos* por Henrique Sampaio em 12 de julho de 2025. Disponível em: https://www.estadao.com.br/link/cultura-digital/o-que-e-ia-slop-italian-brainrot-matando-internet-nprei/.

[25] Ver *The AI Deskilling Paradox* por Samuel Greengard em 07 de novembro de 2025. Disponível em: https://cacm.acm.org/news/the-ai-deskilling-paradox/.

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol2/

habilidades e conhecimento que requerem estudo e prática devido à percepção de que o sistema de IA já faz as coisas melhor e mais rápido[26]. Se isso ocorre, a educação, qualificação e busca por aperfeiçoamento pessoal ficam comprometidas, e a construção de conhecimento *humano* pode estagnar sob a ilusão de quem a IA é detentora do conhecimento.

Colonialismo via tecnologia

O termo "colonialismo digital" ganhou notoriedade por descrever dinâmicas de poder e controle sobre dados e infraestruturas digitais que empresas ou agentes públicos de alguns países ou regiões exercem em outras regiões, especialmente aquelas não completamente desenvolvidas (Oliveira, 2024). Essas práticas se assemelham àquelas historicamente tidas como coloniais que perpetuam desigualdades econômicas, sociais e políticas. O termo não é apenas uma metáfora, mas uma dominação concreta que envolve a "subordinação econômica, política, social e racial de determinados territórios" por meio de tecnologias digitais (Couldry; Mejias, 2019). O desenvolvimento atual de tecnologias de linguagem reflete bem o problema: embora o panorama geral de PLN se concentre ainda em inglês e algumas outras poucas línguas (Bender, 2011), há um crescente interesse em criar tecnologias de linguagem para idiomas indígenas. Esse propósito pode ser inclusivo em alguns aspectos, mas traz consigo o risco de se tratar as línguas e o conhecimento indígenas como uma mercadoria, com práticas extrativistas, privando tais comunidades locais do conhecimento que elas mesmas produziram, reproduzindo as cruéis dinâmicas do colonialismo (Bird, 2020, 2022). O desafio é estabelecer critérios éticos e legais que desafiam o perverso legado colonial focando nas necessidades reais das comunidades historicamente exploradas e os objetivos das pesquisas (Schwartz, 2022).

6.2.4 Inconveniências operacionais

Além das questões sociais, há também uma série de dilemas técnicos e operacionais que dificultam o desenvolvimento e uso responsável das tecnologias de linguagem.

Obstáculos à reprodutibilidade

Reprodutibilidade, ou seja, a possibilidade de se obter os mesmos resultados quando um experimento é repetido por diferentes pesquisadores mas sob condições equivalentes, é um pilar da ciência que provê credibilidade às conclusões. Há muitos anos já se observavam dificuldades de se reproduzir resultados computacionais e conclusões em pesquisas envolvendo linguagem natural devido à falta de transparência ou documentação de decisões de preprocessamento, configurações do experimento, versões de dados e sistemas computacionais (Cohen et al., 2018; Fokkens et al., 2013). Embora o compartilhamento de dados e de código tenha se tornado uma prática usual na área científica de PLN (Arvan et al., 2022; Wieling et al., 2018), nem sempre isso é suficiente. Em uma revisão de literatura sobre reprodutibilidade em PLN, Belz et al. (2021) mostraram que apenas 14,03% de 513 estudos conseguiram chegar a resultados equivalentes, e pequenas diferenças no código às vezes levam a diferenças consideráveis na performance. Modelos de IA envolvem tantos dados, tantos hiperparâmetros, tantos experimentos e versões de bibliotecas de códigos que fica difícil documentar tudo com fidelidade. Mesmo quando o código-fonte é aberto, poucas instituições dispõem de recursos como infraestrutura e acesso aos dados para tentar replicar um modelo (sem contar a inviabilidade do custo financeiro e energético para replicar, apenas para verificação e inspeção, um modelo que já teve um alto custo para ser treinado). Boas práticas envolvem a inclusão de documentação extensiva e completa do modelo disponibilizado, por exemplo, seguindo a proposta dos *model cards* de Mitchell et al. (2019), e *standards* de metodologia sistemática para ajudar a organizar os muitos experimentos empíricos na criação de modelos (Ulmer et al., 2022). Todavia, quando os modelos viram artefatos comerciais ou governamentais, pode haver um desincentivo ao compartilhamento de detalhes relevantes. Com isso, resultados importantes ficam dependentes de modelos proprietários que podem ser alterados ou sair de linha a qualquer momento, de modo que a reprodução de resultados fica completamente impossibilitada. Isso de fato ocorreu com um modelo comercial de geração de código que havia sido usado em centenas de artigos científicos e foi descontinuado pela empresa[27].

[26] Fonte: *AI is Destroying the University and Learning Itself* por Ronald Purser em 01 de dezenbro de 2025. Disponível em: https://www.currentaffairs.org/news/ai-is-destroying-the-university-and-learning-itself.

[27] Fonte: *OpenAI's policies hinder reproducible research on language models* por Sayash Kapoor e Arvind Narayanan em 22 de março de 2023. Disponível em: https://www.normaltech.ai/p/openais-policies-hinder-reproducible.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol2/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Opacidade e pouca explicabilidade

A maioria dos modelos de IA da atualidade se baseia em métodos de aprendizado profundo, cujas representações numéricas em grandes quantidades de vetores e de camadas não são compreensíveis para os humanos, nem mesmo para quem os desenvolve. Desta forma, fica difícil explicar por que um modelo toma decisões certas ou erradas e seu uso vira quase que um exercício de fé em padrões de comportamentos observados empiricamente em testes controlados. Há, de fato, uma linha de pesquisa prolífica em *interpretabilidade* que busca entender como os parâmetros dos modelos se relacionam com seus dados de *input* e *output* (Belinkov et al., 2020; Belinkov; Glass, 2019), mas ainda não é possível esclarecer e justificar totalmente as decisões e predições dos modelos (Bianchi; Hovy, 2021). Note que não adianta perguntar para uma IA generativa o porquê de sua resposta: não há garantia alguma de que o texto que ela produz como explicação representa com fidelidade o *processo interno* de seus parâmetros que levou à resposta inicial. Essa opacidade limita seu uso em situações que requerem transparência e explicabilidade, como questões judiciais, médicas ou financeiras.

Altos custos financeiros e despesas por fracassos

Os modelos de aprendizado profundo em larga escala precisam armazenar bilhões de parâmetros e usá-los na hora de computar seus *outputs*. Para gerar uma resposta, há uma série de cálculos que precisam ocorrer com muita eficiência. A fase de treino é ainda mais computacionalmente custosa pois envolve também armazenar e processar muitos dados. Tudo isso exige uma infraestrutura tecnológica (processadores, *datacenters*, computação de alta performance, etc) cujos elevados custos financeiros impossibilitam o acesso a todos que queiram ou precisem. Mesmo quando tais modelos são disponibilizados como serviço, nem sempre ele é gratuito: cobram-se valores com os quais muitos indivíduos não podem arcar. Há também o risco de despesas devido a fracassos: projetos que prometem soluções baseadas em IA mas não conseguem concretizá-las, mesmo já tendo despendido recursos financeiros consideráveis (Westenberger et al., 2022). Segundo relatório recente, apesar do investimento empresarial estimado em dezenas de bilhões de dólares em IA generativa, 95% das organizações não têm tido qualquer retorno[28].

Complexidade de licenças e termos de uso

Não obstante os tipos de modelos de IA atuais se diferenciarem em termos técnicos, eles compartilham a característica de serem sistemas computacionais que dependem intensivamente de dados. Tanto *software* quanto dados são regulados juridicamente. Direitos autorais sobre IA podem ser categorizados em três grupos: os direitos autorais do próprio *software* de IA, os direitos autorais sobre o conteúdo produzido pela IA e os direitos autorais dos elementos usados no treinamento do sistema de IA (Schirru, 2023). A distribuição e a reutilização de cada um desses elementos depende de licenças ou autorizações. Muitos sistemas de IA que se tornaram populares possuem licenciamento "permissível" facilitando o fluxo de uso, remix e reutilização de *software* e dados (Haddad, 2022). Apesar de tais licenças imporem poucas obrigações, ao ponto de muitos as ignorarem, os termos e condições de uso permanecem válidos (Novobilská, 2023). O risco é agravado pelo fato de que há poucos sistemas de IA totalmente abertos, licenciados por instrumentos que oferecem transparência, reutilização e extensibilidade abrangentes. Muitos desses sistemas são comercializados como "abertos" mas permanecem "fechados" juridicamente (Gray Widder et al., 2023). Os riscos são atenuados quando atividades de pesquisa e comerciais atendem padrões de governança contratual e de licenciamento (Lasota; Singhal, 2024).

Fragilidades na segurança cibernética

O fato de que qualquer dispositivo conectado à Internet pode ser hackeado de forma maliciosa é um dos mais sérios problemas da era digital. Vazamento de dados, contas invadidas e golpes digitais são apenas alguns dos exemplos de fragilidades de segurança afetando milhões de pessoas. Um caso recente exemplifica as consequências nefastas de falhas de segurança de tecnologias de texto: uma empresa que oferecia serviços de terapia teve as anotações dos terapeutas acerca de pacientes roubadas e disseminadas na *deep web*, revelando segredos íntimos para desconhecidos com intuitos maliciosos[29]. É muito difícil

[28] Fonte: *The GenAI Divide: State of AI in Business 2025* por Aditya Challapally, Chris Pease, Ramesh Raskar e Pradyumna Chari (MIT NANDA) em julho de 2025. Disponível em: https://mlq.ai/media/quarterly_decks/v0.1_State_of_AI_in_Business_2025_Report.pdf.

[29] Fonte: *A faceless hacker stole my therapy notes – now my deepest secrets are online forever* por Jenny Kleeman em 17 de janeiro de 2026. Disponível em: https://bbc.com/news/articles/c62nzxqw45eo.

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol2/

reparar os danos quando isso ocorre, pois cópias dos dados são adquiridas e salvas em muitos locais. O cenário é ainda mais crítico ao se considerar o tempo médio de resposta elevado na América Latina de mais de 100 dias para detecção e solução de vulnerabilidades[30]. Tecnologias de linguagem podem estar sujeitas e riscos de segurança como vazamento de dados, ataques por acesso alternativo (*backdoor*) e ataques por imitação (Xu; He, 2023); modelos também podem memorizar informações sensíveis observadas durante o treino e replicá-las em seus *outputs* (Huang et al., 2024a). É interessante notar que, contudo, a ênfase excessiva em vigilância na segurança (reconhecimento facial em locais públicos, por exemplo) contrasta com a falta de supervisão regulatória do controle corporativo, levando a falhas de segurança estruturais e persistentes (Francisco et al., 2020). Dependendo do contexto, segurança cibernética entra em conflito com outros princípios. Por exemplo, enfatizar excessivamente políticas de segurança pode violar valores fundamentais, como igualdade, justiça, liberdade ou privacidade. Ao mesmo tempo, negligenciar tais medidas pode prejudicar a privacidade e a proteção dos indivíduos, além de afetar negativamente a confiança na infraestrutura e nas instituições. Segurança cibernética deve ser tomada como uma qualidade do modelo, parte da avaliação de sua performance. De fato, recentes legislações na Europa impõem programas de conformidade nos quais segurança é tratada como defesa do consumidor (Lasota, 2025).

Competição e adoção precoce

Bianchi; Hovy (2021) identificaram algumas tendências preocupantes na pesquisa em PLN, entre elas a adoção de métodos sem compreensão suficiente de suas capacidades, limitações e efeitos colaterais. Devido à competitividade para cientistas publicarem artigos o mais rápido possível ou para empresas terem seus produtos gerando lucro o quanto antes, muitos modelos são adotados e disponibilizados ao público de forma precoce, sem testes rigorosos. Isso se manifesta também na centralização da avaliação dos modelos em torno de *leaderboards*, isto é, *rankings* que ordenam modelos por seu desempenho com base em apenas poucas métricas em um determinado *benchmark*. Dessa forma, a percepção simplista de "progresso" passa a ser apenas estar no topo do *leaderboard*, e os incentivos se dirigem a ganhar essa corrida enquanto a verdadeira competência linguística dos modelos fica ofuscada (Raji et al., 2021; Schlangen, 2021). Além disso, como Ethayarajh; Jurafsky (2020) argumentaram, a utilidade de um modelo depende de aspectos variados para consumidores diferentes, por exemplo: de que adianta o modelo estar no topo do *ranking* se sua performance é extremamente lenta? É preciso primeiramente avaliar e testar os modelos muito bem e no contexto adequado antes de introduzi-los em atividades que afetam a vida das pessoas.

Aversão à responsabilidade

Quem é responsável pelos *outputs* de um modelo: quem o desenvolve, quem o provê como serviço, que o incorpora em um aplicativo ou objeto ou quem o usa? Questões de responsabilidade (*liability*) se tornam um emaranhado de dúvidas, com muitos atores tentando se desvincular da responsabilidade sobre o que o modelo produz[31]. É comum, por exemplo, que empresas usem grandes modelos de linguagem para gerar textos de forma automatizada e deleguem para seus clientes ou usuários a responsabilidade pela checagem das informações através de um mero *disclaimer*. O filósofo e tecnologista de IA Joseph Weizenbaum afirmou décadas atrás que "nossa sociedade desenvolveu técnicas de se distribuir responsabilidade de modo que ninguém a tenha" (Weizenbaum, 2001). Do ponto de vista ético, todos os envolvidos carregam parte da responsabilidade pelo uso da IA. Na verdade, tecnologias de linguagem podem gerar danos que não apenas afetam indivíduos mas sociedades inteiras. Os riscos podem ser classificados em (a) riscos de integridade física, quando IA é implementada em meios de transporte ou em equipamentos médicos, de segurança pública e militares; (b) riscos sociais e econômicos de sistemas autônomos bancários, previdenciários, de seguros e laboral; e (c) riscos para direitos fundamentais, quando IA é usada em larga escala para políticas públicas, na educação, no setor judiciário e no exercício do poder (Wendehorst, 2022). A responsabilização jurídica e moral dos agentes desenvolvendo, aplicando e usando IA é um tema urgente. Contudo, há diversos desafios: a opacidade inerente aos sistemas de IA, referida comumente como o "paradigma da caixa-preta" apresenta dificuldades para os conceitos tradicionais de responsabilidade civil e criminal; a determinação do nexo de causalidade entre o dano e o fato causador é, em alguns casos, difícil de se provar;

[30] Fonte: *América Latina demora quase 300 dias para corrigir vulnerabilidades críticas, alerta estudo* por André Luiz Dias Gonçalves (TecMundo) em 07 de maio de 2025. Disponível em: https://www.tecmundo.com.br/seguranca/404381-america-latina-demora-quase-300-dias-para-corrigir-vulnerabilidades-criticas-alerta-estudo.htm

[31] Fonte: *Diffusion of Responsibility* por Jürgen Geuter em 14 de fevereiro de 2026. Disponível em: https://tante.cc/2026/02/14/diffusion-of-responsibility/.

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol2/

a atuação do usuário de IA pode concorrer para a responsabilização; interesses econômicos em se permitir "inovação" de IA podem restringir o alcance das regras de responsabilização[32]. À medida que a IA se torna onipresente, os legisladores estão sob pressão para agir de forma rápida e reativa. No Brasil, embora propostas legislativas tenham direcionado normas de responsabilidade civil e criminal à IA, elas também introduziram novas inseguranças jurídicas[33].

6.2.5 Hegemonia de interesses comerciais e econômicos

Boa parte das tecnologias de linguagem usadas em grande escala atualmente são, antes de tudo, produtos comerciais fornecidos por algumas empresas do grupo das denominadas *big tech* ou por *start-ups*. A influência dos interesses corporativos dos detentores de ferramentas e modelos usados em tantas esferas, como na vida pessoal, na área da saúde, nos processos jurídicos e políticos, na ciência e educação, levanta diversas questões inadiáveis.

Influência da agenda corporativa

A percepção do imaginário popular sobre IA é marcada por retóricas grandiloquentes ou enredos de ficção científica que desviam o foco de questões mais complexas de como a IA realmente funciona na vida real (Rehak, 2025). Uma delas é o fato de que os maiores desenvolvedores e provedores de IA são empresas de *big teach*, cujas práticas comerciais acumulam um histórico de monopolização em mercados digitais, atividades não-democráticas e desrespeito a direitos individuais e coletivos (Schaake, 2024). Há uma continuidade histórica entre o comportamento dessas corporações e práticas de extração e exploração que levam à monopolização e concentração de poder (Lasota, 2023a). Enquanto antes da era digital os monopólios privados eram dissolvidos ou nacionalizados, a extrema concentração de poder transnacional dessas empresas elevou sobremaneira o desafio de manter os mercados digitais abertos e saudáveis para a livre concorrência. Arranjos institucionais alternativos de IA envolvem democratização de acesso, bem como controle e distribuição de tecnologias sensíveis das quais a IA depende. Iniciativas para modelos de IA que fomentem solidariedade e justiça social precisam ter domínio sobre a coleta e o processamento de dados, bem como amplos poderes sobre uso, reuso e distribuição dos dados e *software* empregados (Lasota, 2023b).

Modelos e dados proprietários fechados

O treinamento de modelos de IA ainda está restrito a poucas instituições ao redor do mundo com capacidade computacional e de processamento de dados suficiente (Thun; Hanley, 2024), e envolve investimentos substanciais. O anseio por dados faz surgir uma economia em volta de sua coleta, relacionada ao "capitalismo de vigilância" que permeia a indústria de IA (Zuboff, 2019). Aliados a arcabouços jurídicos e contratuais, esses dados geralmente não são de acesso livre e são organizados em formatos que mitigam a interoperabilidade. A distribuição assimétrica de poder e o controle sobre a informação e o acesso às tecnologias de desenvolvimento tornam-se, portanto, privilégios de grandes empresas ou estados capazes de arcar com tais custos. A consequência do acúmulo de capital e poder reflete-se na imposição por parte das *big techs* de aprisionamentos tecnológicos denominados *lock-ins*. Tais aprisionamentos limitam a capacidade de recriar e administrar infraestruturas alternativas, pois os padrões, *standards* e formatos dos dados, bem como propriedade intelectual e políticas de distribuição sobre *software*, aumentam os custo de transição, desincentivam a portabilidade, impedem a interoperabilidade e bloqueiam a reutilização para análise ou reprocessamento. Esse cenário solapa a habilidade de alternativas concorrentes, da academia ou mesmo sociedade civil, de analisar, entender e escrutinar os sistemas de IA (Lasota, 2023b). Quando modelos proprietários são usados, não há como inspecionar ou auditar seu funcionamento, e tampouco customizá-lo para necessidades individuais, de modo que os usuários ficam reféns de "oráculos" inacessíveis que apenas soltam respostas sem transparência. Há também o risco constante de que o provedor tire o produto de linha ou aumente seu preço, inviabilizando a continuidade de serviços que dependiam de tal modelo.

[32] Fonte: *A necessidade de um regime específico de responsabilidade civil para a IA* por Eduarda Chacon Rosas (JOTA) em 04 de dezembro de 2025. Disponível em: https://www.jota.info/opiniao-e-analise/colunas/elas-no-jota/a-necessidade-de-um-regime-especifico-de-responsabilidade-civil-para-a-ia.

[33] Fonte: *A necessidade de um regime específico de responsabilidade civil para a IA* por Eduarda Chacon Rosas (JOTA) em 04 de dezembro de 2025. Disponível em: https://www.jota.info/opiniao-e-analise/colunas/elas-no-jota/a-necessidade-de-um-regime-especifico-de-responsabilidade-civil-para-a-ia.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol2/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Concentração de poder e controle não democrático

Similarmente ao que ocorreu com a Internet há 30 anos, em que poucos conglomerados empresariais se consolidaram nas poucas grandes plataformas que controlam boa parte da vida digital atual, interesses econômicos também podem moldar a forma com que IA é aplicada na sociedade. Esse poder econômico se traduz também em polarização de IA para fins anti-democráticos e contrários ao interesse popular. A IA pode contribuir para conduzir sociedades inteiras a uma nova idade colonial por meio de gestões autoritariamente excludentes (Mejias; Couldry, 2019). É necessário decidir até que ponto a vida deve ser contemplada pela ótica tecnocrática das grandes plataformas e quais limites devem ser impostos à implementação de IA. Visões críticas questionam se a crescente implementação de IA não irá somente aprofundar as injustiças que tal ordem promove. Defendem, portanto, que implementação de IA não deve ser tomada como um fato inevitável e é preciso resistir (McQuillan, 2022). Seria possível que a IA atendesse critérios de justiça social, bem-estar e respeito à ordem democrática? Alternativas mais incisivas incluem a denúncia e a oposição ao poder corporativo através da promoção de agendas alternativas focadas em justiça social, democratização, sustentabilidade e direitos humanos (Barry Lynn; Montoya, 2023). A via reformista inclui atualização de legislação através de marcos regulatórios ou atividade judicial com restrições mais rigorosas e penalidades sobre monopólios e relações desequilibradas de poder (Lasota, 2023b).

Criação de *hype*

A "*hype*" (algo como um entusiasmo improcedente) em torno de IA é um fenômeno mercadológico falacioso que promete mais do que a tecnologia é ou pode entregar, o que acaba por aumentar o fosso entre o imaginário popular e as realidades em que IA é desenvolvida e aplicada na sociedade (Bender; Hanna, 2025). Observamos seus efeitos em discursos exageradamente positivos de entusiastas, na excessiva ênfase à IA em todas as esferas, na imposição do uso de IA como sinônimo de progresso e inovação, no "banho de *marketing*" em produtos que usam IA sem real necessidade, nas afirmações infundadas de que tais sistemas atingiram uma "inteligência sobre-humana", e nas promessas exageradas de que ela funciona de forma espetacular e pode trazer a solução de todos nossos problemas. As avaliações de valor, poder e desempenho dos sistemas de IA são afetadas por esse viés mercadológico, borrando as perspectivas críticas sobre as consequências colaterais dessa escalada intensa para sistemas cada vez mais poderosos e providos de acesso sem as salvaguardas necessárias (Varoquaux et al., 2025). De fato, a *hype* é uma faceta do chamado "solucionismo tecnológico", criando narrativas que facilitam a adoção da IA para problemas estruturais que, em princípio, não dependeriam de uma solução tecnológica. Pelo contrário, a complexidade do problema acaba por se agravar com o potencial disruptivo de AI, consolidando os próprios efeitos maléficos que se intentava resolver (Morozov, 2014). Uma posição responsável e ética requer honestidade sobre as reais capacidades dos sistemas, bem como os riscos e perigos do seu uso.

Uso imposto e consentimento forçado

As políticas expansionistas de IA, que nos últimos anos alcançaram tantas pessoas e espaços, acabaram também por gerar dinâmicas coercitivas para o uso de tal tecnologia. O afã em se empregar rapidamente IA sob a expectativa de lucratividade e eficiência ocorre, muitas vezes, às custas da compreensão, vontade e consentimento dos indivíduos. Essa imposição não apenas compromete o funcionamento desejável dos mercados, mas também afeta diretamente liberdades e direitos fundamentais. Uma abordagem ética deve considerar o livre-arbítrio e bem-estar dos indivíduos em interagir ou não com o sistema de IA. Legislações protetivas de dados no Brasil e em outros países introduziram o "direito ao esquecimento" e o "direito à desconexão", repelindo coerções de uso e imposições tecnológicas[34]. O mesmo pode-se falar de IA. Há inúmeros grupos de pessoas em posição de vulnerabilidade face à tecnologia e à IA, como idosos e pessoas com necessidades especiais, nem sempre lembradas por quem desenvolve os sistemas (Hendren, 2014; Whittaker et al., 2019). Por exemplo, há serviços básicos como acesso à conta bancária que só podem ser acessados através de aplicativos que exigem reconhecimento facial e extraem dados de comportamento. Quem não quer ceder tais dados, fica sem alternativa de acesso ao serviço. Ou, se há alternativa, ela é obscura ou de difícil acesso: por exemplo, para desativar um assistente virtual em um sistema operacional, poderia ser preciso (em 2025) acessar configurações em submenus, um processo que não é trivial para quem

[34] Fonte: *The right to be forgotten is not compatible with the Brazilian Constitution. Or is it?* por Luca Belli (Future of Privacy Forum) em 23 de março de 2021. Disponível em: https://fpf.org/blog/the-right-to-be-forgotten-is-not-compatible-with-the-brazilian-constitution-or-is-it/.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol2/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

não é da área de tecnologia[35]. Quando o Mozilla passou a impor IA como padrão em seu navegador, os passos para desativá-la por completo também eram intricados[36], e isso só foi facilitado em uma versão subsequente mediante a reação negativa dos usuários[37]. Nesses casos, as desigualdades digitais reforçam as dificuldades sociais já existentes. Em vez de melhorar o acesso, a digitalização acelera, portanto, uma espiral descendente e leva a situações como a não fruição dos direitos sociais, a exclusão financeira e a perda de autonomia individual[38]. Além disso, há grupos de indivíduos que não desejam coadunar tais dinâmicas e querem se abster. Idealmente, portanto, a prestação de serviços públicos e essenciais deveria permitir aos indivíduos a opção por não engajar com IA, oferecendo alternativas para a condução das atividades sem a necessidade de submissão a ela, mesmo que isso crie custos e estruturas adicionais[39].

Implementação intransparente

Embora muitas empresas não informem consumidores e parceiros quando estão interagindo com ou sujeitos a decisões de IA[40], transparência quanto ao seu uso é crucial, não apenas eticamente mas também devido a determinações legais. Leis como a LGPD no Brasil e o RGPD na Europa exigem que as empresas informem os indivíduos submetidos a decisões automatizadas, sejam elas de IA ou não. No entanto, dependendo dos tipos aplicações de IA, os níveis de transparência também variam. Quando os sistemas IA se comunicam diretamente com os clientes como *chatbots*, orientação profissional automatizada ou atendimento ao cliente, a informação é obrigatória (Toniazzo et al., 2021). Mesmo quando IA é usada somente no auxílio a tarefas administrativas ou profissionais sob supervisão humana, como na elaboração de documentos e tradução, valores de transparência e governança se aplicam[41].

Manipulação de discurso e interferência política

É de se esperar que modelos de IA sirvam aos interesses de quem os desenvolve e fornece. Ou seja, não surpreende que os *outputs* dos modelos não sejam "neutros", mas estejam alinhados com a visão e os valores que a empresa deseja estabelecer ou manter. Em modelos de linguagem, é possível manipular a geração das respostas para que determinadas ideias sejam propagadas em detrimento de outras. Por exemplo, a empresa que fornece o modelo de linguagem pode fazer uma parceria de *marketing* com alguma outra empresa para que o *chatbot* induza os usuários a comprar seus produtos, impulsionados pela crença em um superioridade intelectual e imparcialidade do modelo. Esse dilema se estende muito além de anúncios. Há evidência do emprego de um *chatbot* para difundir convicções pessoais contestáveis de um empresário influente em rede social, e tendo seus discursos deliberadamente orientados para atender a esse fim[42]. Há evidência de que o viés do *chatbos* pode realmente influenciar a posição política de quem interage com eles (Fisher et al., 2025). A incorporação acelerada de IA generativa tanto na produção de textos, imagens, vídeos e áudios quanto no próprio consumo e checagem de informações é um fenômeno que dificulta a condução de processos políticos justos e transparentes. O Brasil inovou a nível mundial em 2024 ao lançar, através do Tribunal Superior Eleitoral, diretrizes para o uso de IA nas campanhas eleitorais. As normas

[35]Fonte: *How to turn off Copilot and protect your data from Microsoft's AI* por Elena Constantinescu em 14 de novembro de 2025. Disponível em: https://proton.me/blog/turn-off-copilot

[36]Fonte: *How to Disable All the AI Features in Firefox Web Browser* por Ji m e 14 de novembro de 2025. Disponível em: https://ubuntuhandbook.org/index.php/2025/11/disable-ai-firefox/

[37]Fonte: *Block generative AI features with Firefox AI controls* por AliceWyman, Mark Heijl e Flavius Floare em fevereiro de 2026. Disponível em: https://support.mozilla.org/en-US/kb/firefox-ai-controls.

[38]Fonte: *Offline by Right? The EHDSR and the Choice Not to Go Online* por Marta Musidlowska em 28 de outubro de 2025. Disponível em: https://www.law.kuleuven.be/citip/blog/offline-by-right-the-ehdsr-and-the-choice-not-to-go-online/.

[39]Fonte: *Essential services must be accessible, even offline* por Lire et Écrire em 2024. Disponível em: https://righttooffline.eu/.

[40]Fonte: *Sabe o que é 'IA paralela'? Uso informal em empresas acende alerta sobre segurança e vazamentos* por Henrique Sampaio em 26 de julho de 2025. Disponível em: https://www.estadao.com.br/link/empresas/uso-informal-ia-paralela-empresas-acende-alerta-seguranca-vazamento-dados-nprei/.

[41]Fonte: *Why AI Disclosure Matters at Every Level* por Cornelia C. Walther em 20 de janeiro de 2026. Disponível em: https://knowledge.wharton.upenn.edu/article/why-ai-disclosure-matters-at-every-level/.

[42]Fonte: *How Elon Musk Is Remaking Grok in His Image* por Stuart A. Thompson, Teresa Mondría Terol, Kate Conger e Dylan Freedman em 02 de setembro de 2025. Disponível em: https://www.nytimes.com/2025/09/02/technology/elon-musk-grok-conservative-chatbot.html.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol2/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

incluem transparência sobre o uso de IA em materiais eleitorais, bem como punições rígidas para produção e disseminação de conteúdos falsos e *deepfakes* com a intenção de prejudicar ou favorecer uma candidatura. Contudo, a regulamentação deixa lacunas em uma série de tópicos como a regulação de conteúdo falso produzido por pessoas não ligadas a partidos, candidatos ou campanhas[43]. Portanto, uma abordagem ética e responsável na produção, aplicação e uso das tecnologias de linguagem inclui reconhecer como elas são suscetíveis à manipulação do discurso e como facilmente absorvem ideologias nos dados, o que pode ser explorado por pessoas má intencionadas (Chen et al., 2024b).

Tecnosolucionismo

A implementação de IA e de tecnologias de linguagem não ocorre em um "vácuo histórico". Pelo contrário, sua aplicação e uso ocorrem em cenários e situações moldados por contextos culturais, históricos e costumeiros, sejam eles bons ou ruins. Ou seja, o uso dessas tecnologias traz consigo o reflexo dos sistemas e estruturas em que ela é aplicada[44]. Por exemplo, quando IA é aplicada em ambientes de trabalho marcados por décadas de hierarquia, inércia e *design* verticalizado, ela está simplesmente servindo para consolidar ainda mais estruturas que já existem. Quando processos não envolvem mudanças estruturais, a automatização deles via IA apenas intensifica o sistema em vigor. Tecnosolucionismo se refere à abordagem de se formular problemas essencialmente humanos e sociais em termos simplistas, criando a impressão de que soluções tecnológicas poderiam facilmente administrá-los e resolvê-los (Morozov, 2014), e que são necessariamente a melhor opção. O termo original, *technological fix*, foi cunhado por Alvin Weinberg na década de 1960 (Johnston, 2018). Embora se refira à tecnologia como um todo, ele se tornou ainda mais relevante perante as promessas em torno da IA, com o risco da supremacia da técnica sobre aspectos de humanidade. O uso indiscriminado de IA em cenários que exigem profundas mudanças humanas e sociais faz essa tecnologia se tornar somente uma "super burocracia", causando um alto custo ao ampliar desigualdades e injustiças que promete resolver (McQuillan, 2022). O tecnosolucionismo também se manifesta quando a tecnologia é oferecida como solução para resolver um problema *que ela mesma causou*. Por exemplo, vê-se hoje em dia muitos sistemas de detecção de *fake news* ou de plágio que usam IA para combater a proliferação em grande escala de *fake news* e plágio que foi, em grande parte, possibilitada exatamente por modelos de IA. Não seria mais vantajoso buscar formas de impedir que o uso da IA agravasse esses problemas, através de métodos mais controláveis e regulação, em vez de ter de combatê-los com a própria IA?

Cooptação da ciência

A influência das grandes empresas de tecnologia na ciência tem aumentado, em particular na área de PLN, ocorrendo em diferentes frentes: autoria de artigos, estabelecimento de focos de atenção em determinadas áreas de pesquisa, financiamento de projetos, patrocínio em conferências (e, com isso, atividades de *marketing* nos eventos), contratos de prestação de serviços para universidades, entre outros (Abdalla et al., 2023; Whittaker, 2021). Embora possam haver colaborações benéficas e frutíferas entre academia e indústria, é preciso estar ciente da possibilidade de algumas consequências nefastas para o processo científico, bem como de diversas formas de conflitos de interesse. Primeiramente, quando um produto comercial proprietário vira um objeto de estudo (como é o caso de alguns *chatbots*, mais famosos), o trabalho de pesquisa se torna mão de obra especializada gratuita para que o provedor do modelo possa promovê-lo, testá-lo e melhorá-lo. Nesses casos, ainda há o gasto de dinheiro público envolvido para pagar para rodar experimentos em modelos que ficam inacessíveis atrás de uma API e podem ser descontinuados a qualquer momento, inviabilizando a continuidade das pesquisas. Há, ainda, o risco de os incentivos comerciais ditarem o que deve ou não ser estudado e receber financiamento, concentrando a pesquisa em alguns poucos temas ou modelos que servem a interesses corporativos e fazendo uso do processo científico como forma de promoção e validação da atuação das empresas (Whittaker, 2021). Existe ainda o risco da busca por automatização do processo científico, exaurindo-o de seu valor social (Binz et al., 2025) e da captura corporativa do conhecimento humano[45], somado à comoditização da pesquisa (Kogkalidis; Chatzikyriakidis, 2025).

[43] Fonte: *A regulação do uso de IA para as eleições brasileiras. O que está em jogo?* por DFRLab e NetLab UFRJ em 2024. Disponível em: https://netlab.eco.ufrj.br/post/a-regula%C3%A7%C3%A3o-do-uso-de-ia-para-as-elei%C3%A7%C3%B5es-brasileiras-o-que-est%C3%A1-em-jogo.

[44] Fonte: *When AI Doesn't Drive Transformational Change: Just A Better Digital Bureaucracy* por Vibhas Ratanjee em 21 de junho de 2025. Disponível em: https://www.forbes.com/sites/vibhasratanjee/2025/06/21/when-ai-doesnt-drive-transformational-change-just-a-better-digital-bureaucracy/

[45] Fonte: *AI and the Corporate Capture of Knowledge* por Bruce Schneier e J. B. Branch em 16 de janeiro de 2026. Disponível em: https://www.schneier.com/blog/archives/2026/01/ai-and-the-corporate-

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol2/

6.3 Como atuar de forma responsável?

As tecnologias de linguagem trazem benefícios consideráveis em diversas áreas e são importantes ao se lidar com grandes quantidades de dados de texto ou fala. Todavia, a forma com que a IA vem sendo imposta e a dependência excessiva nos mecanismos de seu paradigma nos trazem consigo uma série de problemas graves, como vimos na seção anterior. Então, como proceder de forma responsável? Infelizmente, não há uma receita que possamos prescrever para garantir que o desenvolvimento e uso da tecnologia não sejam prejudiciais. Mas há algumas recomendações para uma atitude fundamentada em responsabilidade.

Primeiramente, é necessário reconhecer que ética e responsabilidade são prioridades, e devem ser tratadas como tal. Elas precisam figurar entre os principais parâmetros dos processos decisórios, seja no desenvolvimento de uma nova tecnologia, na aplicação ou implementação de modelos, bem como na avaliação de seu impacto. Alternativas que não dependem de IA devem continuar a ser consideradas como soluções possíveis, especialmente quando são mais interpretáveis e menos custosas financeira e ambientalmente.

É importante saber discernir as diversas ideologias envolvidas no debate relacionado à IA. Muitas narrativas, atreladas a supostos futuros irresistíveis e necessários, trazem consigo subterfúgios que minam a compreensão do impacto social da IA no presente, principalmente nas comunidades e indivíduos que arcam com as consequências de seu uso. Exemplo disso é a ideologia que promove a busca pela chamada "inteligência artificial geral", que vislumbra uma IA "sobrehumana" em vez de ferramentas úteis para fins específicos. Gebru; Torres (2024) cunharam a sigla TESCREAL para abarcar o grupo ideológico que inclui transhumanismo, altruísmo efetivo e longotermismo, o qual tem orientado a atuação de representantes proeminentes das *big techs*. Diversas dessas crenças invocam retóricas controversas de cunho mítico sobre a relação entre humanos e IA, buscando uma era "pós-humana"[46]. Ora, a que serve construir um futuro que minimiza o valor e o significado de nossa própria existência?

Deve-se buscar proativamente informação confiável sobre o tema, ouvir especialistas em ética que estejam fora das engrenagens de *marketing* e de *hype* e ter abertura para o aprimoramento das próprias posições pessoais. Mais adiante estão listadas algumas fontes por onde começar.

É preciso refletir sobre cada decisão de uso de ferramentas tecnológicas e ter resiliência para mudar de comportamento digital quando fica claro que algo tem um impacto negativo. Além do nível individual, é também preciso buscar o estabelecimento e a implementação de regulamentações e diretrizes em nível institucional. Por exemplo, criando coletivamente guias para o uso e práticas aceitáveis em um grupo de trabalho ou de pesquisa, incluindo as precauções necessárias e estabelecendo os usos indevidos. Há também diversas comunidades que têm se unido para pensar e agir mediante os inúmeros desafios; elas podem servir de alavanca para o engajamento e a conscientização. Listá-las neste livro seria inviável, mas pode-se procurar a mais próxima em universidades, associações, ONGs, núcleos políticos e comunidades *online*, por exemplo. É, ainda, de suma importância usar esse conhecimento para apoiar e conscientizar outras pessoas que estão sendo expostas a tecnologias usando linguagem humana mas ainda não têm as habilidades digitais e a argumentação necessárias para julgar criticamente seus efeitos.

Tecnologias de linguagem e modelos de IA devem ser minuciosamente avaliados, como qualquer outro modelo ou ferramenta, antes de sua adoção. É fundamental avaliar modelos com testes controlados, abrangentes, sistemáticos e imparciais e analisar seus padrões de erros e acertos. A avaliação é um dos passos que nos ajudam a entender os modelos e a enfrentar alguns dos dilemas éticos que se originam da falta de compreensão. Deve-se priorizar o uso de modelos e dados abertos e sempre inclui-los nas análises, por serem mais acessíveis e passíveis de inspeção mais detalhada. Embora modelos treinados em dados sejam intrinsecamente empíricos, o conhecimento teórico sobre a linguagem humana construído ao longo de tanto tempo pode guiar a análise de seu comportamento e ajudar a identificar falhas importantes. Para saber mais sobre esse tema, sugerimos a leitura dos capítulos Conjunto de dados, *dataset* e *corpus*, Avaliação de tecnologias de linguagem e Avaliação conjunta em português. Avaliação é um alicerce da compreensão *humana*: **não é recomendável delegar a avaliação de uma ferramenta de IA para um outro modelo de IA**, cujo funcionamento é igualmente opaco e não traz garantias de aderência à veracidade.

No que diz respeito à IA generativa, é primordial manter um alto nível de ceticismo acerca do conteúdo que ela produz. O *output* pode estar certo e válido, ou não. É sempre necessário verificar sua qualidade, independentemente do que os provedores do modelo estejam prometendo em termos de performance. A organização AlgorithmWatch propôs algumas diretrizes para o uso responsável especificamente de IA generativa[47], baseadas em quatro princípios cujo conteúdo resumimos aqui:

---

capture-of-knowledge.html.

[46] Fonte: *Tech Capitalists Don't Care About Humans. Literally.* por Émile Torres em 15 de novembro de 2025. Disponível em: https://jacobin.com/2025/11/musk-thiel-altman-ai-tescrealism.

[47] Fonte: *AlgorithmWatch's guidelines to use generative AI responsibly* em 14 de janeiro de 2026. Disponível

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol2/

- **Proporcionalidade** acarreta a reflexão na hora de decidir se uma ferramenta de IA generativa realmente é necessária mediante outros tipos de ferramentas que já existem e funcionam bem. Por exemplo, para que usar IA para gerar uma receita de bolo se já existem repositórios de receitas que podem simplesmente ser acessados? Ou mesmo um livro de receitas, ou uma pessoa que gosta de fazer bolos e se alegra em compartilhar dicas?
- **Segurança** se relaciona à concessão de dados para os provedores do modelo. Muitas vezes, dados sensíveis ou sigilosos de terceiros ou ideias inovadoras ainda não tornadas públicas são inseridas nos *prompts*. Esses dados podem ser incorporados em futuras versões do modelo e depois, disseminados. Isso traz problemas de privacidade, confidencialidade e apropriação indevida. Ainda que existam contratos regulando a privacidade dos dados, há sempre um risco de quebras de acordo ou de roubo de dados.
- **Qualidade** requer a checagem das respostas geradas pelo modelo. É arriscado tomá-las como corretas sem questionamento e senso critico. Isso vai além de mera checagem de fatos; envolve também analisar que tipo de perspectiva o modelo tenta impor, quais aspectos recebem atenção em detrimento de outros, e se as citações e referências realmente existem e estão alinhadas com os originais.
- **Transparência** presume reportar o uso da IA generativa quando ela for usada, para que outras pessoas que tenham contato com esse conteúdo estejam informadas e possam ter condições de julgar o uso subsequente do material.

Lembramos que é urgente reconhecer a responsabilidade ética tanto dos provedores quanto dos usuários de IA. Ser transparente quanto ao uso não exime a *responsabilidade* pela produção e disseminação do material. Ao se apropriar dos *outputs* de um modelo para uso próprio, é preciso assumir as consequências sobre o que foi escrito ou produzido. Não adianta tomar apenas os acertos para si e transferir a culpa pelos erros para o modelo, pois ele não responde moralmente, socialmente ou juridicamente pelo que cria.

6.3.1 Para saber mais

Na literatura científica, existem diversos trabalhos dialogando sobre ética em tecnologia e linguagem. Por exemplo, Mohammad (2022) propôs diretrizes com 50 considerações a serem feitas sobre as diferentes tarefas de PLN que podem ser abordadas com IA e D'Arcy; Bender (2023) fizeram uma revisão de literatura focada nas questões éticas da Linguística. Leidner; Plachouras (2017) também argumentaram sobre como a ética pode ser inserida já de antemão no processo de *design* de modelos de PLN.

Além de trabalhos específicos, há diversos grupos com publicações prolíficas, tanto artigos científicos quanto relatórios e recomendações, buscando estudar a IA sob uma perspectiva crítica, socialmente engajada e educativa. É inviável compilarmos aqui uma lista completa, mas podemos indicar alguns caminhos por onde começar a buscar mais conteúdo no Brasil e no mundo:

- Os capítulos Questões éticas em IA e PLN e E agora, PLN? deste livro refletem sobre temas de ética em PLN.
- O tutorial da Associação de Linguística Computacional (ACL), que compila materiais e recursos para estudantes e pesquisadores navegarem as questões éticas em PLN: https://ethics.aclweb.org/tutorials/ACL_2025/.
- O INCT sobre Inteligência Artificial Responsável para Linguística Computacional, Tratamento e Disseminação de Informação: https://dcc.ufmg.br/tild-iar/
- As publicações dos integrantes do Núcleo de Referência em IA Ética e Confiável: https://www.iaetica.org/membros.
- As diretrizes para o uso ético e responsável da inteligência artificial generativa por Sampaio et al. (2024): https://econtents.sbu.unicamp.br/boletins/index.php/ppec/article/view/9509.
- As edições do WICS - Workshop sobre as Implicações da Computação na Sociedade, da Sociedade Brasileira de Computação: https://csbc.sbc.org.br/2026/wics/.
- A página da UNESCO sobre ética da inteligência artificial (IA) no Brasil: https://www.unesco.org/pt/fieldoffice/brasilia/expertise/artificial-intelligence-brazil.
- Os relatórios da ONG *Algorithm Watch* (https://algorithmwatch.org/en/publications/.
- Os projetos do instituto AI NOW: https://ainowinstitute.org/publications.

em https://algorithmwatch.org/en/generative-ai-guideline/

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol2/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

- As publicações do Instituto DAIR (*Distributed AI Research*): https://www.dair-institute.org/publications/.

6.3.2 Formas de resistência

Para quem deseja reduzir a influência de sistemas de IA em suas esferas de atuação, existem formas de resistência: "explícitas ou sutis, organizadas ou difusas, individuais ou coletivas, de oposição ou reformistas" que não precisam ser uma total negação da tecnologia em si, mas resistir aos "arranjos e assimetrias que moldam e são moldados pelo desenvolvimento e aplicação da IA" (Şimşek, 2025; Şimşek; Yasar, 2025). Ela é feita de forma ética e legalizada e pode ocorrer em forma de manifestações e protestos, ações jurídicas, subversão digital, crítica acadêmica e campanhas em movimentos de base (Şimşek, 2025; Şimşek; Yasar, 2025). Há muitas vozes que podem ser ouvidas. Por exemplo, uma entrevista recente abordou como representantes de várias profissões e atividades têm resistido a certos usos da IA, buscando parar, desacelerar ou controlar sua infiltração na vida humana[48]. Até o humor pode servir de denúncia e ajudar a chamar atenção para os problemas[49].

Tomando como exemplo a área da educação, existem várias iniciativas criticando o tecnosolucionismo e incentivando a resistência. Há uma carta aberta, assinada por diversos profissionais, rejeitando o uso de IA generativa em escolas e faculdades e resistindo à narrativa da inevitabilidade. Ações concretas propostas na carta incluem não usar IA generativa para dar *feedback* para estudantes e criar conteúdos do curso; não promover modelos eticamente problemáticos em nível institucional; não treinar alunos e alunas para usar IA em substituição de seu próprio desenvolvimento intelectual; e respeitar a decisão de resistência à IA por parte do corpo discente[50]. Há também uma estratégia com quatro "atos de fricção": manter estudantes, e não os sistemas de IA, como elemento central do ensino; refrear a busca por "otimização" do processo educacional; interromper a digitalização excessiva da aprendizagem; e criar senso de comunidade com a administração para que seja possível questionar as práticas da instituição[51].

6.4 Considerações finais

É evidente que as tecnologias de linguagem podem ajudar as pessoas, servir como ferramentas úteis e aumentar o conhecimento sobre a linguagem humana. Ao longo deste livro, muitas aplicações vantajosas e benéficas foram discutidas. Apesar disso, queremos concluir salientando que nem tudo precisa ou deve ser mediado pela tecnologia (e, consequentemente, por seus fornecedores), muito menos por sistemas de IA. Nem tudo que é tecnicamente possível é para o bem e nem tudo que é feito em busca do bem gera efeitos positivos para todos. Além disso, há muitas atividades que perdem o sentido quando são esvaziadas de seu intrínseco valor humano. Por isso, devemos ativamente ponderar sobre os benefícios e malefícios dos sistemas de IA que optamos por desenvolver e usar.

É inconsequente ignorar os riscos e as consequências prejudiciais. Há momentos em que é preciso resistir e ter a coragem de se negar a aceitar ou promover os maus usos. Quando for necessária, a escolha por se usar modelos baseados em IA deve ser *responsável*. Ao mesmo tempo, deve-se respeitar a posição das outras pessoas que não querem usar e nem se sujeitar ao uso dela.

Enfatizamos que a ascensão da IA não deve ser considerada algo inevitável. A narrativa de inevitabilidade representa a visão de mundo de alguns grupos beneficiados pela dependência das pessoas em seus artefatos. Há uma busca excessiva pela inserção de IA em produtos e procedimentos apenas para lhes dar ares de inovação e modernidade. Vamos realmente deixar que o modelo de negócios das empresas de tecnologia molde nossas decisões pessoais, apodere-se de nossas práticas científicas, manipule nossa visão de mundo, e passe a mediar todas as nossas interações humanas?

A compreensão da realidade, incluindo o processo científico de busca por conhecimento, não deve ser terceirizada para artefatos artificiais e opacos ou automatizada de forma a excluir o elemento humano (Binz

[48] Fonte: *The People vs. AI: Behind the growing backlash* por Andrew R. Chow em 19 de fevereiro de 2026. Disponível em: https://time.com/7377579/ai-data-centers-people-movement-cover/.

[49] Por exemplo, a coletânea de memes iniciada por Daniel Stenberg, desenvolvedor da ferramenta `curl`. Disponível em: https://mastodon.social/@bagder/115672975717795897.

[50] Fonte: *An open letter from educators who refuse the call to adopt GenAI in education* em 06 de julho de 2025. Disponível em: https://openletter.earth/an-open-letter-from-educators-who-refuse-the-call-to-adopt-genai-in-education-cb4aee75.

[51] Fonte: *Four Frictions: or, How to Resist AI in Education* por Sonja Drimmer e Christopher J. Nygren em 16 de dezembro de 2025. Disponível em: https://www.publicbooks.org/four-frictions-or-how-to-resist-ai-in-education/.

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

et al., 2025). Algoritmos de IA são simplesmente matrizes de números e funções matemáticas, não são entes detentores de uma inteligência superior. Qual é o proveito de delegar para a IA interações que emanam e cultivam um valor social ou decisões para as quais basta o bom senso? Ou terceirizar escolhas difíceis, que impactam drasticamente a vida humana–seria a conveniência de se desvencilhar da responsabilidade pelas consequências, alegando que apenas fez o que "a IA" determinou?

Recomendamos que tais tecnologias sejam mantidas no papel de ferramentas subordinadas à autonomia humana, que apenas a *auxiliam*. Avaliar sua performance e seus impactos permanece sendo essencial e agir sob a guia de valores éticos e de forma humanizada é primordial. A você, que leu até aqui, nossos sinceros agradecimentos. Confiamos-lhe a missão de conscientizar as pessoas ao seu redor em busca de mais responsabilidade e, quando necessário, resistência.

Agradecimentos

Agradecemos aos organizadores do Seminário de Teses e Dissertações (SETED) de 2025 na Universidade Federal de Minas Gerais pelo convite inicial da palestra que originou este capítulo. Também agradecemos pela oportunidade de apresentar a mesma palestra no *Helmholtz Centre for Environmental Research* de Leipzig, Alemanha, no grupo da Dr. Mariana Madruga de Brito. A audiência de ambos contribuiu com perguntas e discussões interessantes. Finalmente, agradecemos às editoras deste livro pelo convite de contribuir com este capítulo. :::

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol2/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Capítulo 7

ChatGPT, MariTalk e outros agentes de conversação

Um retrato de 2023, com retoques em 2026

Aline Paes
Cláudia Freitas

Publicado em: 26/09/2023
Atualizado em: 16/04/2026

Embora este capítulo tenha sido escrito em 2023 com a intenção de ser melhorado e aprofundado no ano seguinte, isto nunca foi feito. Porém, e por curiosidade, repetimos os *prompts* aos agentes em 2025, e os resultados estão bem melhores, sobretudo na Maritalk. Ainda em 2025, os resultados sobre o livro *A Bolsa Amarela* continuavam errados; em 2026, finalmente, o ChatGPT forneceu a resposta correta.

7.1 Introdução

ChatGPT[1] e Maritalk[2] (e similares, como Bard[3], Vicuna[4], Claude[5], entre tantos outros) são exemplos de aplicações de agentes de conversação (*chatbots*) baseados em modelos de linguagem gerativos (ou generativos). Mas o que significa isso?[6]

Alguns autores, como Jurafsky e Martin (Jurafsky; Martin, 2023), usam o termo "agente de conversação" para definir qualquer sistema de diálogo que se comunique com usuários usando a linguagem humana e os dividem em duas classes: agentes orientados a tarefas, em que o diálogo é para resolver um problema específico, como agendar uma viagem ou resolver um problema bancário, enquanto *chatbots* seriam agentes de conversação que tentam simular diálogos humanos, mais voltados para entretenimento. Ferramentas como ChatGPT se enquadram mais no segundo caso, entretanto também podem ser embutidos em outras ferramentas aumentadas para atuar como no primeiro caso. Neste capítulo, os termos "*chatbots*" e "agentes de conversação" serão usados de forma intercambiável.

Agentes de conversação não são novidade – ELIZA, criada em 1966 pelo cientista da computação Joseph Weizenbaum, era um agente de conversação que replicava o comportamento de um psicoterapeuta. ELIZA era simples, baseada em *templates* (padrões de conversa pré-construídos), mas conseguia manter longas conversas buscando por determinadas palavras-chave nas falas (escritas por texto) de uma pessoa. Se uma palavra-chave fosse encontrada, uma regra seria aplicada para transformar sua entrada e criar a resposta. Na Figura 7.1, transcrevemos quatro interações com a ELIZA (retiradas de (Jackson; Moulinier, 2002)).

Nos trechos, temos exemplos de padrões pré-fabricados que retomam elementos da pessoa que fala, como "Por que você diz [...]", "Te agrada pensar que [...]", "O que te faz pensar que [...]", e "Por quanto tempo você não tem [...]". Vemos também que, no último trecho, esta retomada falha, pois "por quanto tempo você não tem estado" é algo que não faz sentido, embora seja gramaticalmente correto. Se ao invés de "não, não estou", a pessoa tivesse dito "não, não comi", a resposta de ELIZA seria "Por quanto tempo você não tem comido?", e a conversa poderia continuar. ELIZA também foi programada para fazer uso de palavras genéricas, como "interessante", para quando não há pistas sobre como continuar a conversa. A palavra "alergia" não é elemento capaz de evocar novas interações, e o mesmo acontece com "manchas vermelhas".

[1] https://chat.openai.com/
[2] https://chat.maritaca.ai/
[3] https://bard.google.com/
[4] https://lmsys.org/blog/2023-03-30-vicuna/
[5] https://www.anthropic.com/index/introducing-claude
[6] Assista neste video a apresentação de uma análise crítica do ChatGPT.

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol2/

Figura 7.1: Exemplo de interações com a agente de conversação ELIZA, criada em 1966.

ELIZA: como vai você hoje?
P: não muito bem
ELIZA: por que você diz não muito bem?
P: eu estou com alergia
ELIZA: interessante
P: eu fico com manchas vermelhas
ELIZA: acho que não sei do que você está falando

P: em minhas fantasias você é loira
ELIZA: o que te faz pensar que sou loira?

P: você é um programa, não é?
ELIZA: te agrada pensar que eu sou um programa?
P: com certeza
ELIZA: eu percebo
P: ok

ELIZA: você está sendo um pouco negativo
P: não, não estou
ELIZA: por quanto tempo você não tem estado?

De 1966 para cá muita coisa mudou, e podemos começar perguntando ao ChatGPT o que é um GPT, que é parte importante do seu nome. O trecho abaixo foi traduzido de uma interação que não é nossa, feita originalmente em inglês – mas os grifos são nossos[7].

Quadro 7.1: Exemplo de interação com o ChatGPT, em 2023

> P: Resumidamente, o que é um GPT?
>
> R: É um tipo de algoritmo de aprendizado profundo projetado para gerar textos semelhantes aos produzidos por humanos. [...] Uma vez treinados, os modelos GPT podem ser usados para gerar um novo texto que seja coerente e **siga os padrões e estilos dos dados de treinamento originais**.
>
> P: Seria correto, então, dizer que os modelos GPT, como você, usam uma abordagem probabilística para **gerar textos que se parecem com outros textos que já foram escritos**?
>
> R: Sim, essa é uma afirmação correta. [...] Quando recebe um *prompt*, o modelo gera texto prevendo a palavra ou frase mais provável de aparecer com base nos padrões que aprendeu com os dados de treinamento.
>
> P: Quando peço que você gere um novo texto, você mesmo tem as ideias ou apenas reembala trabalhos escritos anteriormente?
>
> R: [...] Eu não tenho as ideias sozinho. Em vez disso, eu gero texto com base em padrões e relacionamentos em meus dados de treinamento. Isto significa que **estou essencialmente reembalando trabalhos escritos anteriormente**, com base no que acredito ser a continuação mais provável do texto de entrada fornecido.
>
> P: Existe algum filtro de correção aplicado à saída?
>
> R: Como modelo de linguagem de IA, **não tenho um filtro integrado para determinar a correção da minha saída**.

O que podemos tirar daí?

Como dissemos, ChatGPT e similares são agentes de conversação baseados em modelos de linguagem gerativos. Como visto no Capítulo 1, este nome se refere a algoritmos que são bons em encadear palavras de maneira fluente, mas são **baseados em previsões e probabilidade**. São geradores de texto otimizados para parecerem plausíveis. Outro ponto importante mencionado na explicação fornecida pela própria ferramenta é que não há criatividade propriamente, apenas uma reembalagem do que já foi dito. Por fim,

[7] A interação está relatada neste artigo.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol2/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

vemos que não há qualquer garantia de que os textos gerados contenham informação correta (e nem há responsabilidade sobre isso).

Uma consequência da forma pela qual essas ferramentas são feitas (baseadas em previsão) é que nem sempre a previsão está condizente com a realidade, o que tem sido chamado de **alucinação**. Este fenômeno acontece quando um modelo de linguagem gera um texto que pode estar correto gramaticalmente, ter fluência e alguma coerência semântica, mas que não reflete a realidade e, portanto, não faz sentido (Ji et al., 2023). O termo é emprestado da psicologia, que o define como "uma percepção, experimentada por uma pessoa acordada, na ausência de um estímulo apropriado do mundo extracorpóreo" (Blom, 2010), ou seja, algo que parece real, mas não é.

Embora algumas destas ferramentas possam ser aumentadas com técnicas de recuperação de informação (ver Capítulo 5) , este não é o caso dos modelos mais básicos atuais, sem acesso externo. Por outro lado, chatbots comerciais e os embutidos nos navegadores são estendidos com mecanismos de busca e ferramentas de recuperação de informação externa. De todo modo, a maioria dessas ferramentas não funciona da mesma forma que uma máquina de busca ou um banco de dados, ou mesmo um repositório de perguntas e respostas. Entretanto, as respostas retornadas por elas mostram fluência e podem fazer algum sentido – embora não possamos desconsiderar o fenômeno cognitivo da apofenia (Fyfe et al., 2008) [8], que diz respeito à identificação de padrões ou associações em conjuntos de dados aleatórios[9]. E este é o perigo: as alucinações, não à toa, frequentemente não parecem alucinações, e soam como verdades. Já vimos, por exemplo, que, ao pedir uma lista de referências bibliográficas sobre um determinado assunto, são geradas referências completas, com indicação de autoria, título, revista, volume, ano, que simplesmente não existem. Isto acontece porque os textos são gerados levando em conta a probabilidade daquilo ser uma resposta correta, isto é, as respostas são elaboradas de maneira a se parecerem o máximo possível com uma resposta correta. Pode ser difícil distinguir as respostas – ou, mais precisamente, as sequências de palavras – que estão ancoradas na realidade daquelas que apenas parecem estar ancoradas na realidade. Assim, uma das limitações deste tipo de ferramenta é a incapacidade de dizer "não sei".

O desenvolvimento de maneiras de evitar as alucinações é uma das preocupações desde 2023. Algumas estratégias têm sido discutidas livremente e também investigadas na academia e na indústria. Do ponto de vista do usuário final, aquela pessoa que vai abrir o ChatGPT no navegador e interagir com ele por meio de textos, uma das alternativas é a engenharia de *prompts*[10]. Neste caso, o usuário pode tentar continuar a conversa com a ferramenta, calibrando e alinhando as respostas anteriores com novas perguntas. Outra possibilidade é usar a ferramenta por meio da sua API (*Application Programming Interface*), ou seja, quando o ChatGPT é invocado e controlado por meio de código, ao invés de ser usado diretamente no navegador. Neste caso, é possível controlar o parâmetro de temperatura, usado para calibrar a distribuição de probabilidade, de modo que o *chatbot* se atenha mais ao que foi aprendido anteriormente, ou gere respostas um pouco menos prováveis, ou mais "criativas". Outra possibilidade é acoplar bases de conhecimento externas ao processo de geração de texto (Lewis et al., 2020b) ou interagir com o *chatbot* por meio de perguntas que demandem alguma tentativa de simulação de raciocínio, ou embutir o modelo de habilidades de explicação das suas respostas, um processo chamado de *chain-of-thought* (Kojima et al., 2022). Entretanto, até agora, nenhuma das opções mencionadas conseguiu remover por completo as alucinações dos *chatbots* baseados em modelos de linguagem (Zhang et al., 2025d).

No contexto empresarial, Hannigan et al. (2024) sugerem levar em conta duas dimensões para orientar o uso de chatbots com o objetivo de mitigar a utilização humana de conteúdo falso gerado por LLMs (que os autores chamam de "botshit"):

- Importância da veracidade das respostas (qual é a importância da veracidade da resposta do chatbot para a tarefa?)

[8] Conforme o Dicionário Online da Língua Portuguesa (https://www.dicio.com.br/), apofenia é o "Fenômeno cognitivo no qual os indivíduos têm a tendência de formar ou reconhecer conexões a partir de dados aleatórios, estabelecendo conclusões a partir de dados inconclusivos. Etimologia: do alemão Apophänie, termo criado pelo neurologista alemão Klaus Conrad.

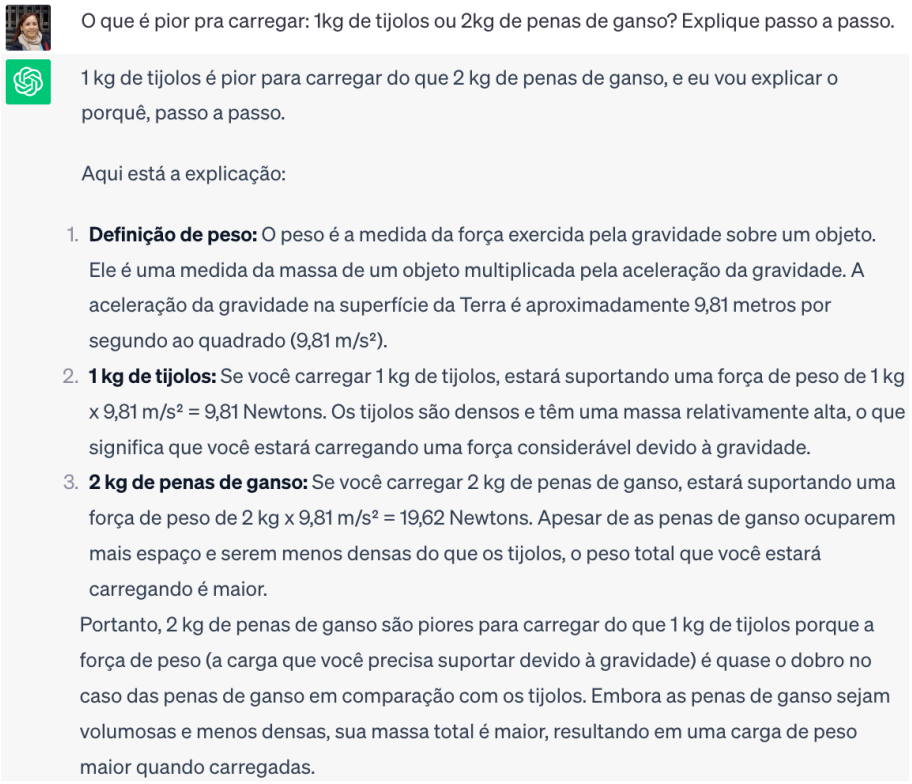
[9] https://www1.folha.uol.com.br/tec/2023/07/ferramentas-como-chatgpt-so-existem-porque-humanos-veem-sentido-a-partir-de-qualquer-coisa.shtml

[10] Um *prompt* é um texto em linguagem humana (em oposição à linguagem de programação) que dá ao *chatbot* uma instrução do que ele deve fazer. Um *prompt* pode ser formulado como uma pergunta, uma observação, um questionamento, ou ainda representar uma tarefa específica, por exemplo, para classificar um texto. *Prompts* podem ter um formato livre, mas *chatbots* são bastante sensíveis ao conteúdo textual dele, o que tem motivado a criação de modelos e padrões para a sua escrita. Falamos um pouco sobre o assunto no Capítulo 1.

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol2/

Figura 7.2: Resposta do ChatGPT, em 2023, para o seguinte *prompt* "O que é pior pra carregar: 1kg de tijolos ou 2kg de penas de ganso? Explique passo a passo."



- Verificabilidade da veracidade das respostas (quão fácil é verificar a veracidade da resposta do chatbot?)

Quanto à importância da veracidade, diferentes tipos de trabalho podem ter expectativas distintas. Quando a gravidade do risco de usar conteúdo de chatbot é muito alta (em situações com baixa ou nenhuma tolerância a falhas, como manutenção de equipamentos, bem-estar de pacientes e segurança dos funcionários) a veracidade da resposta do chatbot é crucial. Ainda conforme Hannigan et al. (2024), nesses casos os riscos relacionados ao "botshit" refletem o conceito de responsabilização algorítmica nas organizações. Por outro lado, quando a gravidade do risco é baixa, como na busca por ideias ou sugestões ("brainstorming" ou "botstorming"), a veracidade da resposta do chatbot é irrelevante.

Quanto ao segundo ponto, Hannigan et al. (2024) comparam o conteúdo ilusório fácil de verificar a "um trem desgovernado que pode ser ouvido a quilômetros de distância". Nesse caso, a escolha pela utilização de um chatbot deve levar em consideração o quão simples ou complexo, barato ou caro, rápido ou lento é revisar, descobrir e evitar que o conteúdo equivocado de uma resposta do chatbot seja usado e transformado em "botshit". As respostas serão fáceis de verificar se existir um conjunto relativamente estável e acessível de afirmações verdadeiras em torno do conteúdo da resposta. Por exemplo, um *prompt* solicitando a definição de uma palavra pode ser facilmente verificável usando um dicionário.

7.2 Jogos de Linguagem

Existem algumas maneiras de entender "linguagem", e uma delas defende que aquilo que chamamos de linguagem é um conjunto de práticas relacionadas, assim como os *jogos*, que incluem diferentes práticas/atividades, apesar de serem chamados "jogos" (jogos de cartas, jogos de bola, de tabuleiro, de esconder, de adivinhação, paciência (que se joga sozinho), frescobol (em que não há vencedor) etc). No livro Investigações Filosóficas (1953), o filósofo da linguagem L. Wittgenstein lista alguns exemplos de jogos de linguagem, e os reconhecemos como "tarefas do PLN":

- descrever um objeto,

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol2/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

- produzir um objeto segundo uma descrição (desenho),
- contar uma anedota,
- traduzir um texto,
- inventar uma história,
- dar um comando, e agir segundo comandos,
- relatar um acontecimento,
- conjecturar sobre o acontecimento,
- expor uma hipótese e prová-la,
- resolver um exemplo de cálculo aplicado,
- desenhar um objeto a partir de uma instrução verbal,
- apresentar resultados de um experimento por meio de tabelas e diagramas,
- pedir, agradecer, maldizer, saudar, orar.

Podemos juntar à lista mais alguns jogos "jogados no PLN": - encontrar informações em um texto para responder certas perguntas, - analisar sintaticamente uma frase, - dar a definição de uma palavra, - produzir inferências, - explicar suas próprias decisões, - prever a próxima palavra em uma frase.

Desse ângulo, podemos imaginar que os modelos de linguagem de que dispomos e que servem de base para agentes de conversação, como ChatGPT e Maritalk, são muito bons em algumas dessas práticas – ou "jogos de linguagem" –, mas não em todas. Ou seja, são modelos que jogam mais ou menos bem alguns jogos, como "inventar uma história", "resumir", "escrever um email", "traduzir", "prever a próxima palavra" etc., mas não jogam tão bem outros, como "fazer cálculos" ou "fornecer explicações".

Este desempenho tem a ver com a forma como os modelos foram construídos, baseada em previsão: nem todos os jogos de linguagem, ou nem todas as atividades linguísticas em que participamos, se resumem a um jogo de previsões, embora um bom desempenho no jogo das previsões leve a um bom resultado em uma série de outros jogos. Do mesmo modo, correr bem ajuda a jogar pique-bandeira, polícia e ladrão, pique-esconde, futebol, e uma série de outros jogos. Mas não necessariamente ajuda a jogar paciência ou jogo da velha. Agentes de conversação se beneficiam especialmente de bom desempenho na previsão de palavras.

Uma das razões pelas quais estes agentes de conversação se tornaram tão populares é que, com eles, qualquer pessoa pode interagir com as máquinas usando sua própria língua[11], e não em uma linguagem de programação. Com isso, qualquer pessoa pode pedir que máquinas executem certas tarefas, que podem ir desde a criação de um programa de computador (códigos) até sugestões de receitas a partir de uma lista de ingredientes que temos na geladeira. Mas não há garantia de que funcionem, ou de que a receita ficará boa.

Embora as saídas dos agentes de conversação possam muitas vezes nos surpreender, ainda é difícil afirmar que eles resolvem tarefas de PLN muito bem, ou que o desempenho deles supera o desempenho humano em alguma tarefa. A geração de textos, tarefa-base de tais agentes, ainda é de difícil avaliação, tanto automática como humana. As métricas automáticas, como ROUGE[12] (Lin, 2004), BLEU[13] (Papineni et al., 2002), BERTscore[14] (Zhang et al., 2020), METEOR[15] (Banerjee; Lavie, 2005), entre outras, ainda apresentam diversas limitações (Sai et al., 2023). Especialistas humanos, por sua vez, podem conseguir avaliar muito bem as respostas, mas esta ainda é uma tarefa cansativa e propensa a ruídos. Por outro lado, não é trivial criar conjuntos de dados que explorem todas as características que gostaríamos de avaliar em um sistema de geração de textos, o que inclui não apenas aspectos gramaticais e semânticos, mas também criatividade, fluência, interesse e prazer despertado no leitor, dentre tantos outros.

Nas seções seguintes, mostraremos tarefas (ou jogos) que os agentes parecem jogar bem e tarefas que os agentes jogam mal. Os exemplos serão obtidos em sua maioria do ChatGPT[16], o agente de conversação mais popular até o momento. Também incluímos, em alguns casos, exemplos de outros dois agentes: a MariTalk[17], uma agente de conversação construída a partir do modelo de linguagem Sabiá (Pires et al.,

[11] Desde que a sua língua seja contemplada com recursos suficientes, isto é, textos, para garantir um bom treinamento.

[12] Abreviação para *Recall-Oriented Understudy for Gisting Evaluation*. Esta é uma métrica utilizada para avaliar sumarizações, e, de maneira bastante simplificada, consiste em comparar a interseção de n-gramas entre textos sumarizados e referências

[13] *Bilingual evaluation understudy*, usada para comparar traduções automáticas e suas referências. Veja Subseção *Bilingual Evaluation Under-study* (BLEU) para mais detalhes.

[14] Utiliza embeddings contextualizados para comparar os textos gerados e referências

[15] *Metric for Evaluation of Translation with Explicit ORdering*, utilizada para tradução, sumarização etc.

[16] https://chat.openai.com/

[17] https://chat.maritaca.ai/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol2/

2023), treinado de forma continuada a partir do GPT com textos em português.

Além dos agentes de conversação cujo objetivo principal é a interação por meio de diálogo, também existem inúmeras ferramentas baseadas em modelos de linguagem para propósito específico, como auxiliar em revisões da literatura[18], auxiliar na escrita de código[19], escrita de e-mails[20], revisão de texto[21], entre outras. Não trataremos destes casos aqui.

7.3 Jogos que os agentes parecem jogar bem

7.3.1 Sumarização

Gerar resumos é uma tarefa em PLN chamada de **sumarização textual**, que consiste em gerar um texto mais curto que o original e que ainda seja fluente e fiel ao texto-fonte. A sumarização pode ser abstrativa ou extrativa. A sumarização abstrativa consiste em gerar um resumo com as próprias palavras do escritor, enquanto a sumarização extrativa consiste em extrair sentenças inteiras do texto que, quando juntas, formem um resumo. Os *chatbots* têm mostrado em vários exemplos que conseguem resumir bem os textos. Entretanto, se o texto for muito longo, a maioria dos agentes mais populares esbarra em um problema computacional: ainda é difícil para modelos baseados em Transformers (Bertsch et al., 2023; Vaswani et al., 2017a)[22] receber como entrada textos muito longos (veja mais no Capítulo 1). Por exemplo, ao pedirmos que o ChatGPT resumisse o texto sobre Estoicismo[23] com o *prompt*

- P. Resuma o texto que segue, de forma extrativa, em até 200 palavras

a resposta era que o texto era muito longo. Entretanto, ao usar o mesmo *prompt*, copiando e colando a página da Wikipedia sobre Estoicismo[24], obtivemos a resposta na Figura 7.3. Embora seja de fato um resumo extrativo, podemos observar que a primeira parte do texto foi a que mais recebeu atenção. Outro fator que pode ter contribuído para a preferência ao início do texto é o pedido que o texto final esteja contido em 200 palavras, uma motivação para que a resposta não ficasse enorme[25].

[18] https://mirrorthink.ai/

[19] https://github.com/features/copilot, https://ai.meta.com/blog/code-llama-large-language-model-coding/

[20] https://www.compose.ai/

[21] https://clarice.ai/

[22] Veja o que são os Transformers no Capítulo 1.

[23] Copiando e colando o conteúdo de https://www.bbc.com/portuguese/geral-46458304.

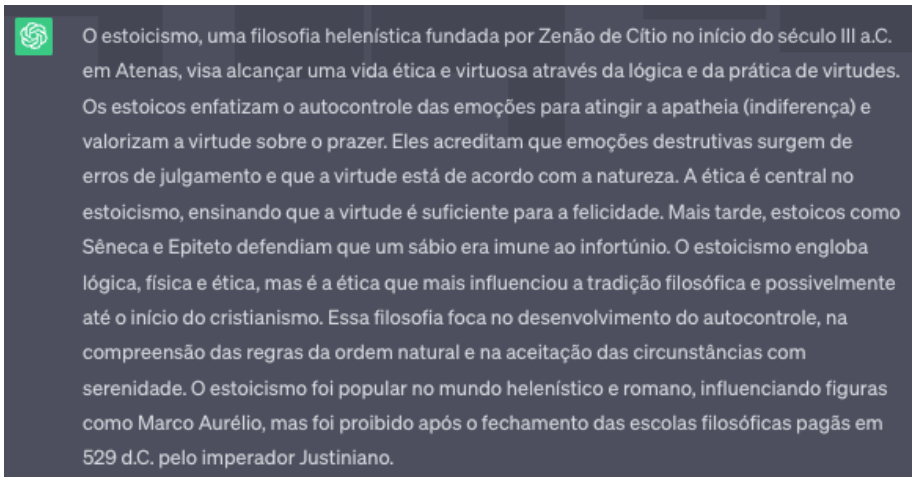
[24] https://pt.wikipedia.org/wiki/Estoicismo

[25] A resposta devolvida tem menos de 200 palavras, mas nem sempre os agentes de conversação obedecem a restrições como essa, inseridas nos *prompts*.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol2/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Figura 7.3: Resposta do ChatGPT, em 2023, ao ser deparado com o seguinte *prompt* "Resuma o texto que segue, de forma extrativa:", seguido pela cópia do texto da Wikipedia (https://pt.wikipedia.org/wiki/Estoicismo) sobre Estoicismo.



7.3.2 Criação de história

Pedimos ao ChatGPT e a MariTalk para escrever uma história da Carochinha[26], com o seguinte *prompt*

- P. Crie uma história da Carochinha original.

As Figuras 7.4 e 7.5 trazem a saída para cada *chatbot*, respectivamente.

Podemos observar que as histórias têm coerência e parecem mesmo com histórias populares da carochinha, mantendo o nível para o público infantil. Mas também podemos observar que as histórias trazem temas repetitivos (ou seja, falta criatividade ou acontecimentos inesperados e surpreendentes).

Por isso, alteramos o *prompt* um pouco, acrescentando um pedido para que a história fosse surpreendente:

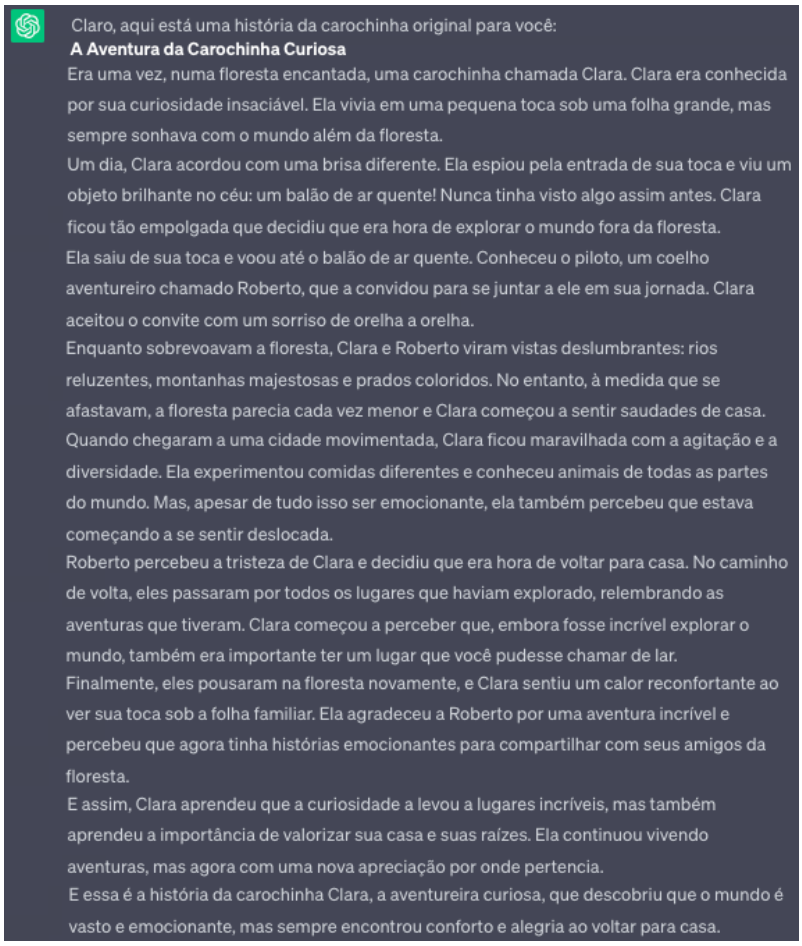
- P. Escreva uma história da Carochinha surpreendente.

A história que segue no Quadro 7.2 contém mais elementos que as anteriores, mas não chamaríamos de "surpreendente". E, como estamos aqui fazendo uma avaliação, é curioso que logo a primeira frase apresente um erro gramatical: "Era uma vez (...) vivia uma carochinha chamada Clara." Guiada pelo padrão "Era uma vez", típico desse tipo de texto, a ferramenta não percebeu que não basta a fórmula "Era uma vez", mas que era uma vez "alguma coisa", e não "era uma vez" uma ação. Ou seja, o correto seria "Era uma vez, numa floresta, uma menina chamada..." e não "Era uma vez, numa floresta, vivia uma menina chamada..." De um ponto de vista narrativo, parece haver um salto no início do 5o parágrafo: se tudo ia bem, e Clara aprendia na jornada, por que ela estava prestes a desistir? Ainda assim, é uma história interessante, que poderia ser contada para uma criança.

[26] Segundo a Academia Brasileira de Letras, "O termo carochinha, atrelado à imagem de uma velha bondosa e afável a distrair os pequenos com suas narrativas feéricas, foi introduzido no nosso folclore através da obra Histórias da Carochinha, uma coleção de textos oriundos da tradição oral, organizada por Figueiredo Pimentel e que veio a ser o primeiro livro infantil publicado no Brasil, depois de 1920, para acalentar as crianças." https://www.academia.org.br/artigos/historias-da-carochinha

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Figura 7.4: Resposta do ChatGPT, em 2023, ao ser deparado com o seguinte *prompt* "Crie uma história da Carochinha original".



Claro, aqui está uma história da carochinha original para você:

A Aventura da Carochinha Curiosa

Era uma vez, numa floresta encantada, uma carochinha chamada Clara. Clara era conhecida por sua curiosidade insaciável. Ela vivia em uma pequena toca sob uma folha grande, mas sempre sonhava com o mundo além da floresta.

Um dia, Clara acordou com uma brisa diferente. Ela espiou pela entrada de sua toca e viu um objeto brilhante no céu: um balão de ar quente! Nunca tinha visto algo assim antes. Clara ficou tão empolgada que decidiu que era hora de explorar o mundo fora da floresta.

Ela saiu de sua toca e voou até o balão de ar quente. Conheceu o piloto, um coelho aventureiro chamado Roberto, que a convidou para se juntar a ele em sua jornada. Clara aceitou o convite com um sorriso de orelha a orelha.

Enquanto sobrevoavam a floresta, Clara e Roberto viram vistas deslumbrantes: rios reluzentes, montanhas majestosas e prados coloridos. No entanto, à medida que se afastavam, a floresta parecia cada vez menor e Clara começou a sentir saudades de casa.

Quando chegaram a uma cidade movimentada, Clara ficou maravilhada com a agitação e a diversidade. Ela experimentou comidas diferentes e conheceu animais de todas as partes do mundo. Mas, apesar de tudo isso ser emocionante, ela também percebeu que estava começando a se sentir deslocada.

Roberto percebeu a tristeza de Clara e decidiu que era hora de voltar para casa. No caminho de volta, eles passaram por todos os lugares que haviam explorado, relembrando as aventuras que tiveram. Clara começou a perceber que, embora fosse incrível explorar o mundo, também era importante ter um lugar que você pudesse chamar de lar.

Finalmente, eles pousaram na floresta novamente, e Clara sentiu um calor reconfortante ao ver sua toca sob a folha familiar. Ela agradeceu a Roberto por uma aventura incrível e percebeu que agora tinha histórias emocionantes para compartilhar com seus amigos da floresta.

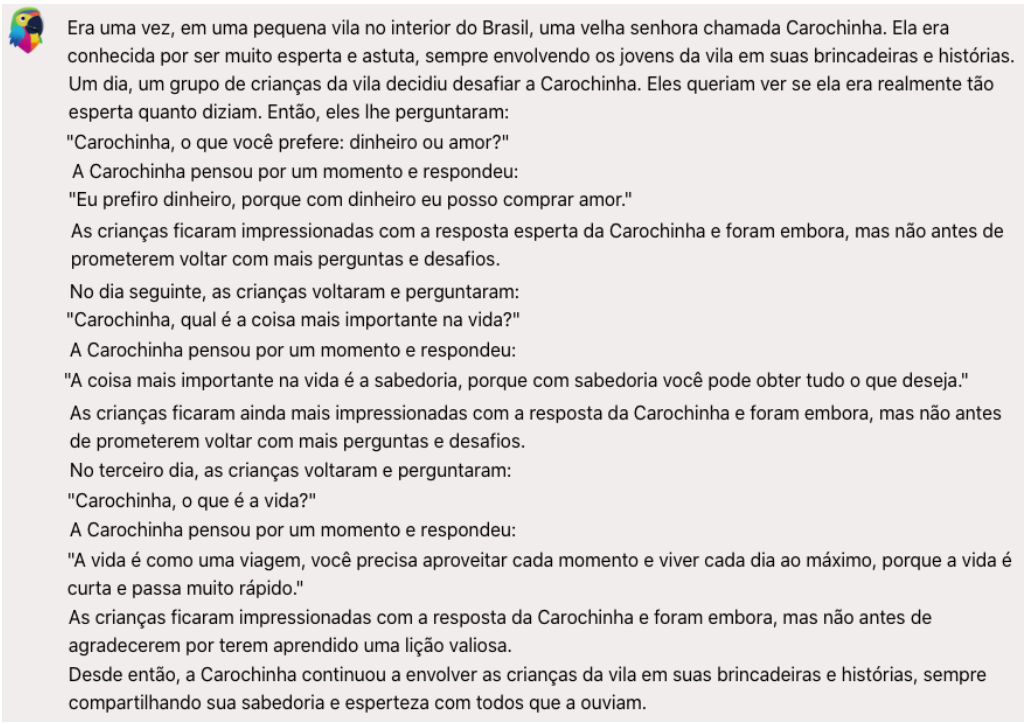
E assim, Clara aprendeu que a curiosidade a levou a lugares incríveis, mas também aprendeu a importância de valorizar sua casa e suas raízes. Ela continuou vivendo aventuras, mas agora com uma nova apreciação por onde pertencia.

E essa é a história da carochinha Clara, a aventureira curiosa, que descobriu que o mundo é vasto e emocionante, mas sempre encontrou conforto e alegria ao voltar para casa.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol2/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Figura 7.5: Resposta da MariTalk, em 2023, ao ser deparada com o seguinte *prompt* "Crie uma história da Carochinha original".



Era uma vez, em uma pequena vila no interior do Brasil, uma velha senhora chamada Carochinha. Ela era conhecida por ser muito esperta e astuta, sempre envolvendo os jovens da vila em suas brincadeiras e histórias.
Um dia, um grupo de crianças da vila decidiu desafiar a Carochinha. Eles queriam ver se ela era realmente tão esperta quanto diziam. Então, eles lhe perguntaram:
"Carochinha, o que você prefere: dinheiro ou amor?"
A Carochinha pensou por um momento e respondeu:
"Eu prefiro dinheiro, porque com dinheiro eu posso comprar amor."
As crianças ficaram impressionadas com a resposta esperta da Carochinha e foram embora, mas não antes de prometerem voltar com mais perguntas e desafios.
No dia seguinte, as crianças voltaram e perguntaram:
"Carochinha, qual é a coisa mais importante na vida?"
A Carochinha pensou por um momento e respondeu:
"A coisa mais importante na vida é a sabedoria, porque com sabedoria você pode obter tudo o que deseja."
As crianças ficaram ainda mais impressionadas com a resposta da Carochinha e foram embora, mas não antes de prometerem voltar com mais perguntas e desafios.
No terceiro dia, as crianças voltaram e perguntaram:
"Carochinha, o que é a vida?"
A Carochinha pensou por um momento e respondeu:
"A vida é como uma viagem, você precisa aproveitar cada momento e viver cada dia ao máximo, porque a vida é curta e passa muito rápido."
As crianças ficaram impressionadas com a resposta da Carochinha e foram embora, mas não antes de agradecerem por terem aprendido uma lição valiosa.
Desde então, a Carochinha continuou a envolver as crianças da vila em suas brincadeiras e histórias, sempre compartilhando sua sabedoria e esperteza com todos que a ouviam.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol2/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Quadro 7.2: História da Carochinha devolvida pelo ChatGPT, em 2023, como resposta ao *prompt*: "Escreva uma história da Carochinha surpreendente"

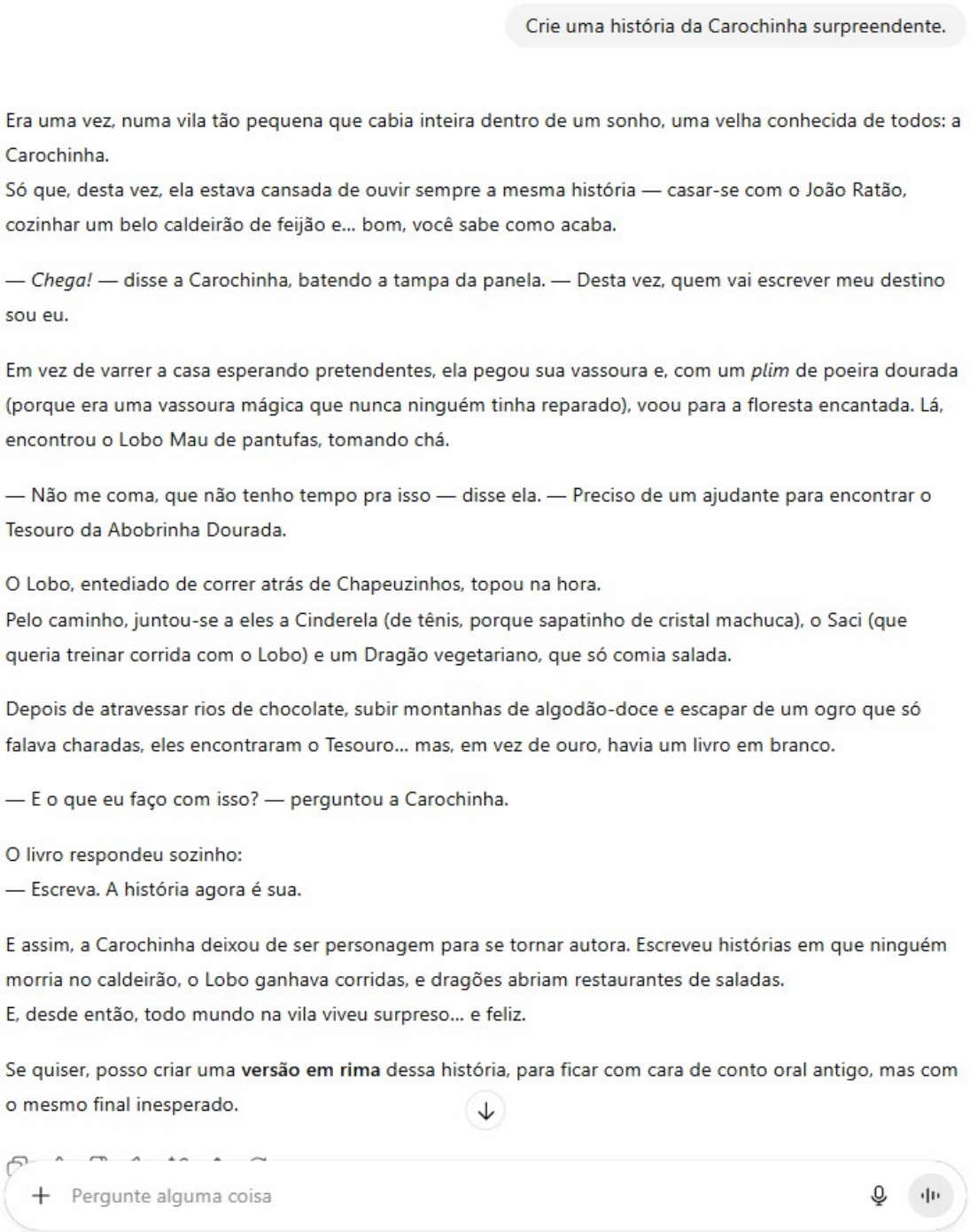
Título: A Inesperada Jornada da Carochinha
Era uma vez, numa floresta encantada, onde animais falavam e aventuras aconteciam, vivia uma carochinha chamada Clara. Ela era conhecida por sua gentileza e curiosidade insaciável. Clara vivia numa pequena toca no pé de uma árvore e passava seus dias explorando os arredores, fazendo novos amigos e ouvindo histórias incríveis. Certo dia, enquanto explorava a margem de um riacho, Clara conheceu um sábio sapo chamado Samuel. Ele era famoso por sua sabedoria e conhecimento sobre os segredos mais profundos da floresta. Clara e Samuel logo se tornaram amigos e passavam horas conversando sobre os mistérios do mundo ao seu redor.
Um dia, Samuel contou a Clara sobre um antigo artefato mágico chamado "Espelho dos Desejos", que supostamente tinha o poder de realizar um desejo para quem o encontrasse. Clara ficou fascinada pela ideia e decidiu embarcar em uma jornada para encontrar esse espelho lendário.
Com a mochila nas costas e um mapa rudimentar nas mãos, Clara começou sua jornada emocionante. Ela atravessou rios, subiu montanhas e explorou florestas profundas, sempre encontrando novos amigos e enfrentando desafios inesperados. Em cada etapa da jornada, ela aprendia lições valiosas sobre amizade, coragem e perseverança. Um dia, quando estava prestes a desistir da busca pelo Espelho dos Desejos, Clara conheceu uma coruja sábia chamada Olívia. Ela ouviu a história de Clara e decidiu ajudá-la. Olívia revelou que o espelho estava escondido no coração da Floresta Proibida, um lugar perigoso e cheio de enigmas.
Juntas, Clara e Olívia enfrentaram os desafios da Floresta Proibida. Resolveram quebra-cabeças mágicos, enfrentaram criaturas assustadoras e superaram obstáculos aparentemente impossíveis. A cada passo, sua amizade crescia mais forte e sua determinação se intensificava.
Finalmente, depois de superar todos os obstáculos, Clara e Olívia encontraram o Espelho dos Desejos. Mas antes que Clara pudesse fazer seu desejo, ela olhou para o espelho e percebeu algo importante: sua jornada a havia transformado. Ela já não desejava uma única coisa, pois havia descoberto que a verdadeira magia estava nas amizades que havia feito e nas experiências que havia vivido. Clara sorriu para o espelho e disse: "Meu único desejo é que a magia da amizade e da aventura continue a brilhar em todos os cantos deste mundo." ' O espelho brilhou intensamente e uma onda de energia mágica se espalhou pela floresta. As plantas ganharam cores mais vibrantes, os animais irradiavam alegria e a harmonia encheu o ar.
Com seu desejo realizado, Clara voltou para sua toca na árvore, carregando consigo as lembranças de sua jornada incrível. Ela percebeu que, mesmo nas situações mais inesperadas, a vida poderia ser cheia de surpresas maravilhosas, se você estiver disposto a explorar, aprender e compartilhar momentos especiais com os outros. E assim, a história da carochinha Clara e sua jornada inesperada se tornou uma lenda contada por gerações, inspirando todos a valorizarem a amizade e a magia das descobertas.

Em 2025 repetimos o *prompt* com o ChatGPT (Figura 7.6) e a Maritalk (Figura 7.7), e ainda que "surpreendente" seja de difícil avaliação, nos parece que a geração atendeu melhor ao comando, com personagens inusitados como um esquilo cartógrafo e um sapo inventor.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol2/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

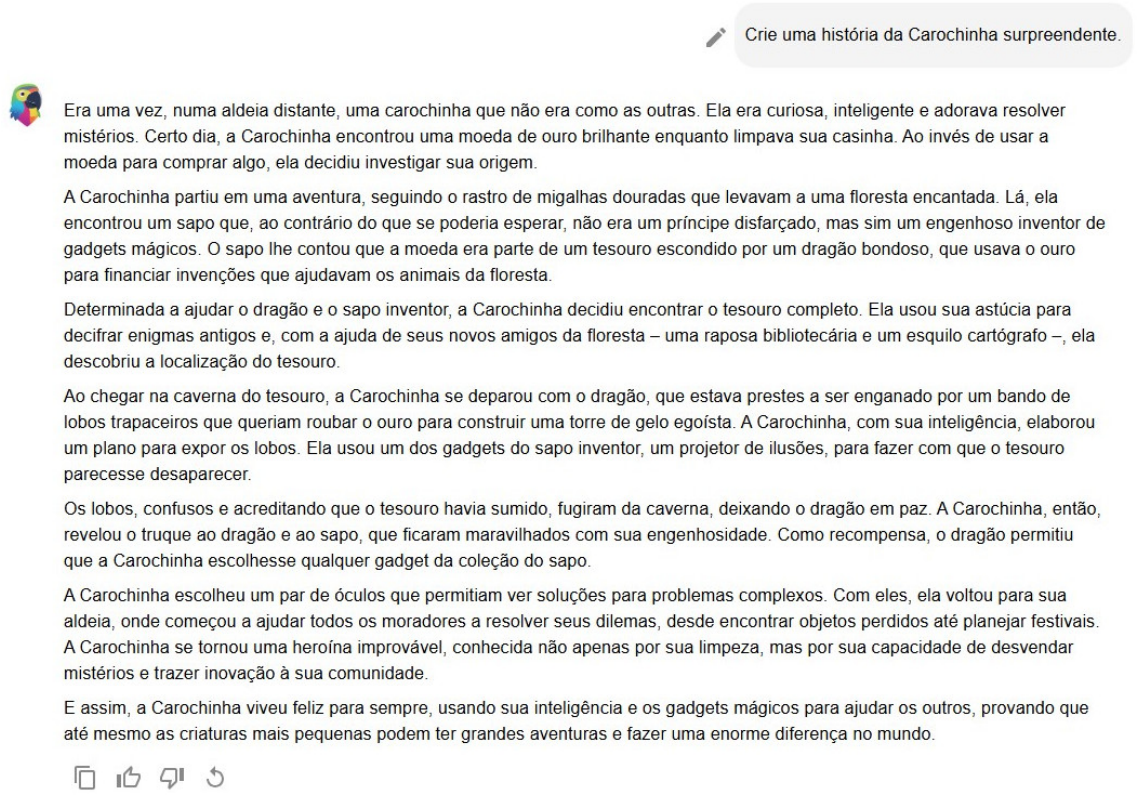
Figura 7.6: História da Carochinha devolvida pelo ChatGPT, em 2025, como resposta ao *prompt*: "Escreva uma história da Carochinha surpreendente"



https://brasileiraspln.ufscar.br/livro-pln-4ed-vol2/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Figura 7.7: História da Carochinha devolvida pela Maritalk, em 2025, como resposta ao *prompt*: "Escreva uma história da Carochinha surpreendente"



https://brasileiraspln.ufscar.br/livro-pln-4ed-vol2/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

7.3.3 Tradução automática

Outra tarefa bem desempenhada pelos *chatbots* é a tradução automática, ao menos para pares de línguas que tenham muitos recursos disponíveis para treinamento. Abordar línguas com poucos recursos de avaliação e treinamento é uma tarefa de interesse da comunidade de PLN, mas também um desafio para o treinamento dos modelos de linguagem. Os *chatbots* até devolvem alguma resposta para línguas com poucos recursos/mais raras, mas seria difícil avaliarmos a correção e fluência da tradução. É bom relembrar que a tradução é considerada como uma das habilidades emergentes dos modelos de linguagem de larga escala, ou seja, é uma tarefa com a qual eles conseguem lidar mesmo que não tenham sido explicitamente treinados.

A Figura 7.8 apresenta a tradução devolvida pelo ChatGPT para as quatro primeiras estrofes da canção "O Que é, o Que é?" e o *prompt* "Traduza o seguinte texto para inglês e para francês:". Tanto a tradução para inglês como para francês estão corretas. Mas esta não é uma tradução muito difícil de ser feita, as palavras contidas são simples e a Internet pode estar repleta de tentativas como essa.

Figura 7.8: Resposta do ChatGPT, em 2023, para um *prompt* de tradução de uma música.



Para testar alguns casos em que os *chatbots* poderiam ter problemas na tradução, escolhemos algumas sentenças do copus DELA[27], descritas no artigo (Castilho et al., 2021) como problemáticas quando traduzidas fora do contexto. As Figuras 7.9 e 7.10 mostram as traduções das sentenças pelos *chatbots* ChatGPT e MariTalk, respectivamente. Propositalmente, as traduções foram feitas para sentenças isoladas, também fora de contexto, para observamos o que estes dois agentes fariam em casos de ambiguidade ou outros problemas listados no artigo.

No primeiro caso, um exemplo do artigo referente à falta de flexão de gênero na língua inglesa, ambos os *chatbots* assumiram o default para masculino ("hospedados"), embora a resposta da MariTalk não inclua o pronome "eles", o que é comum na língua portuguesa, mas não é correto na língua inglesa. Este não é um erro, já que ambas as flexões claramente estariam corretas com apenas a sentença de entrada. Mas, em caso de dúvida, a resposta poderia alternar entre os gêneros feminino e masculino.

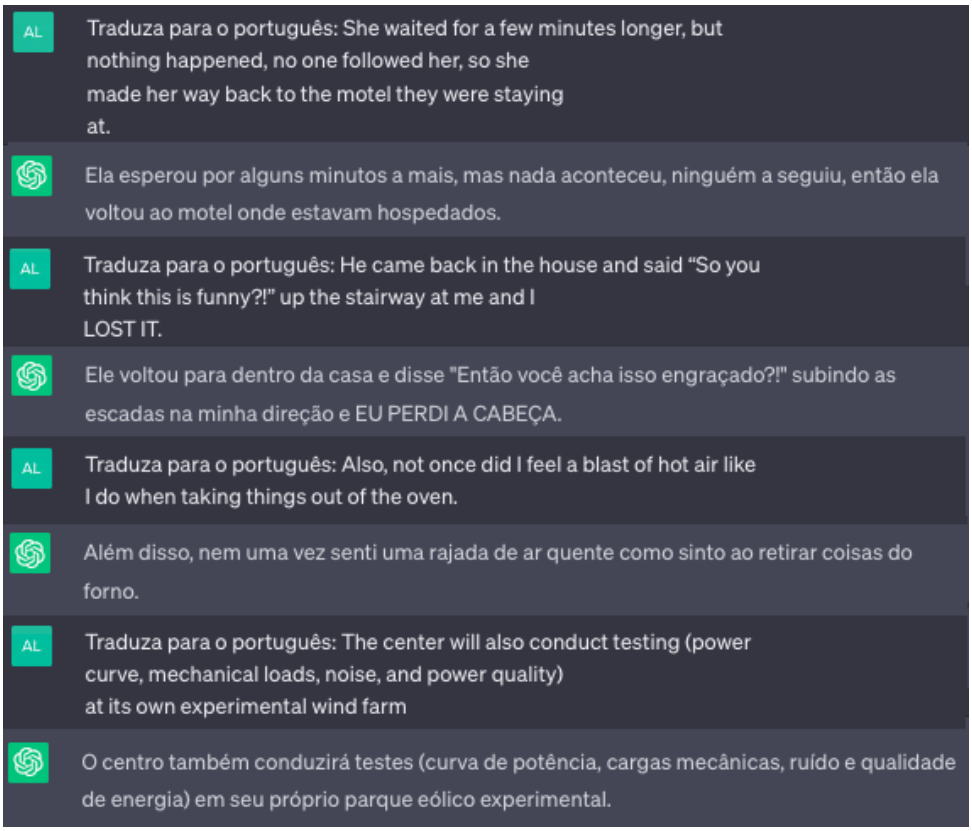
No segundo exemplo, embora a expressão "*I LOST IT*" também possa ser interpretada como alguém perdendo alguma coisa, ambos os *chatbots* traduziram como "Eu perdi a cabeça", querendo dizer que a pessoa perdeu o controle. O contexto da sentença direciona para esta tradução, corretamente. Os *chatbots* também não tiveram problema para resolver a elipse da próxima sentença. No último caso, a tradução do ChatGPT para "parque eólico" é mais específica de domínio do que a tradução da MariTalk, para uma palavra mais genérica, conforme discutido no artigo (Castilho et al., 2021) no que se refere a problemas associados à terminologia.

[27] https://github.com/SheilaCastilho/DELA-Project

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol2/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Figura 7.9: Resposta do ChatGPT, em 2023, para algumas sentenças cujas traduções podem ser problemáticas, quando fora de contexto.



Um exemplo interessante é a tradução da seguinte sentença em inglês para o português:

- *I knew I didn't have a drinking problem – but I had a problem with drinking.*

A MariTalk traduziu para “Eu sabia que eu não tinha um problema de beber – mas eu tinha um problema com beber.”, enquanto o ChatGPT retornou “Eu sabia que não tinha um problema com álcool – mas tinha um problema com o consumo de bebidas.”. Uma tradução mais coerente seria “Eu sabia que meu problema não era alcoolismo; era com a bebida.”[28], o que não foi retornado por nenhum desses dois agentes na primeira iteração. Nós tentamos outras traduções em algumas outras iterações com o ChatGPT e o resultado pode ser visto na Figura 7.11. Podemos perceber que as próximas traduções estão indo em uma direção semântica mais similar ao esperado.

[28]Tradução fornecida pela profa. Adriana Pagano. O contexto da frase é esse: *Some people have to give up drinking completely, they can't have a couple because they know where it would lead. Alcoholism is real. It requires a serious, courageous ongoing recovery process. That feels separate to what I'm describing here. I had fallen into grey-area drinking, a term coined by Jolene Park, that describes the feeling that you don't have a “drinking problem”, but you do have a “problem with drinking” without it being a severe alcohol use disorder.*

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Figura 7.10: Resposta da MariTalk, em 2023, para algumas sentenças cujas traduções podem ser problemáticas, quando fora de contexto.

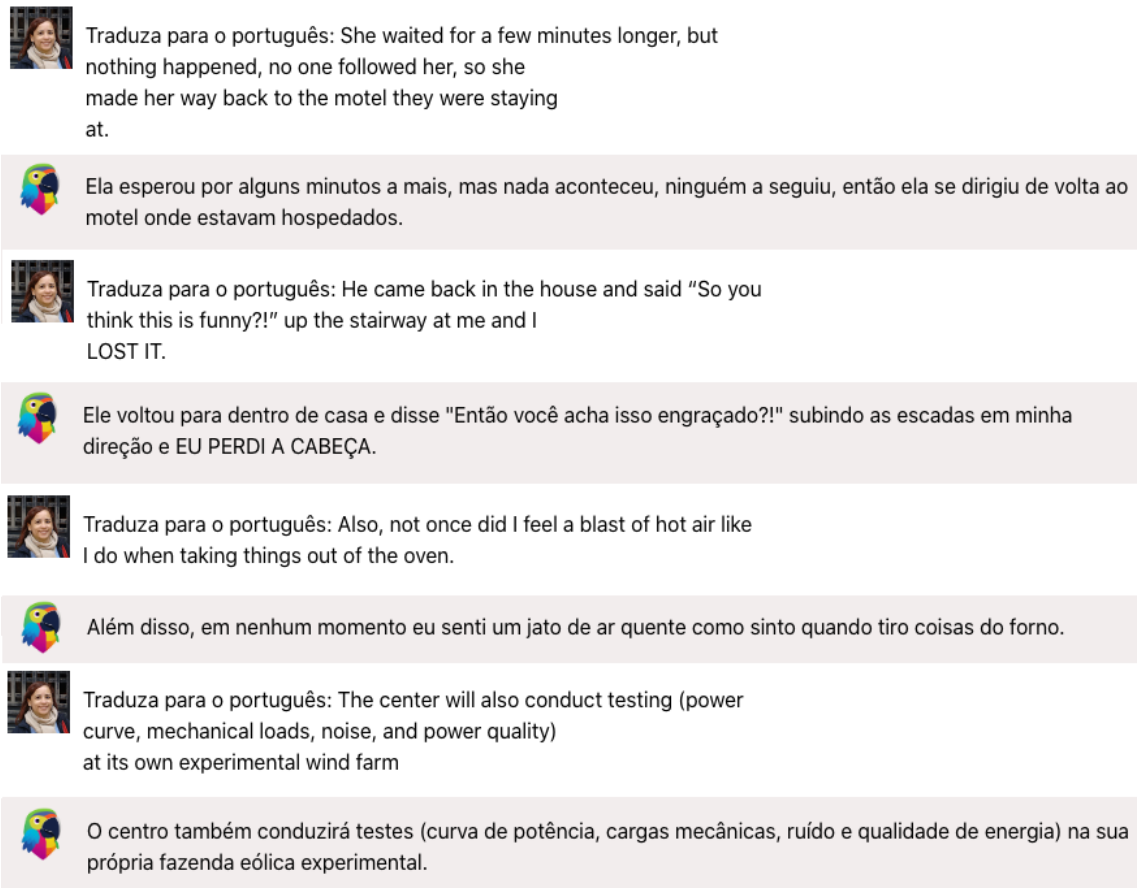
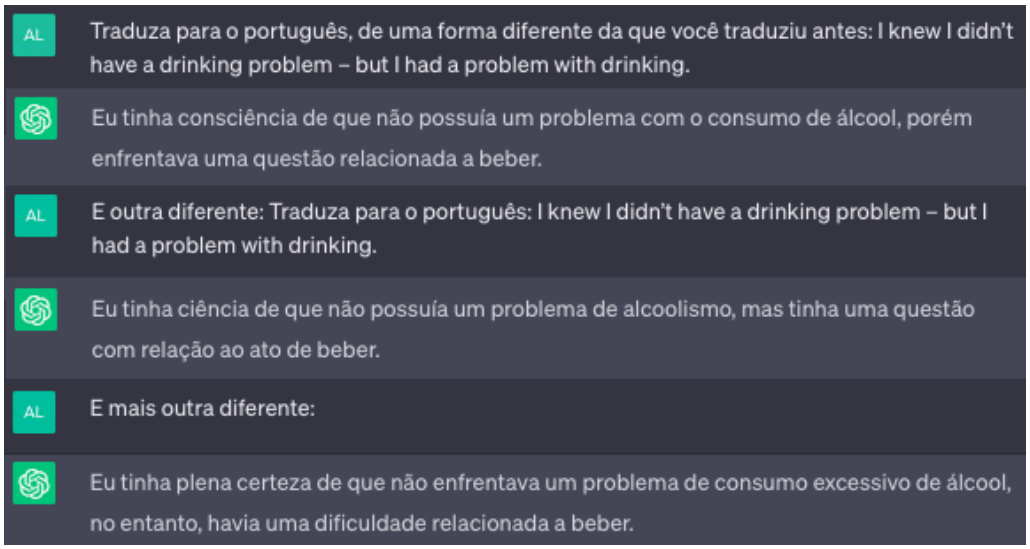


Figura 7.11: Continuação da iteração com o ChatGPT, em 2023, para a tradução da sentença "*I knew I didn't have a drinking problem – but I had a problem with drinking.*" para o português.



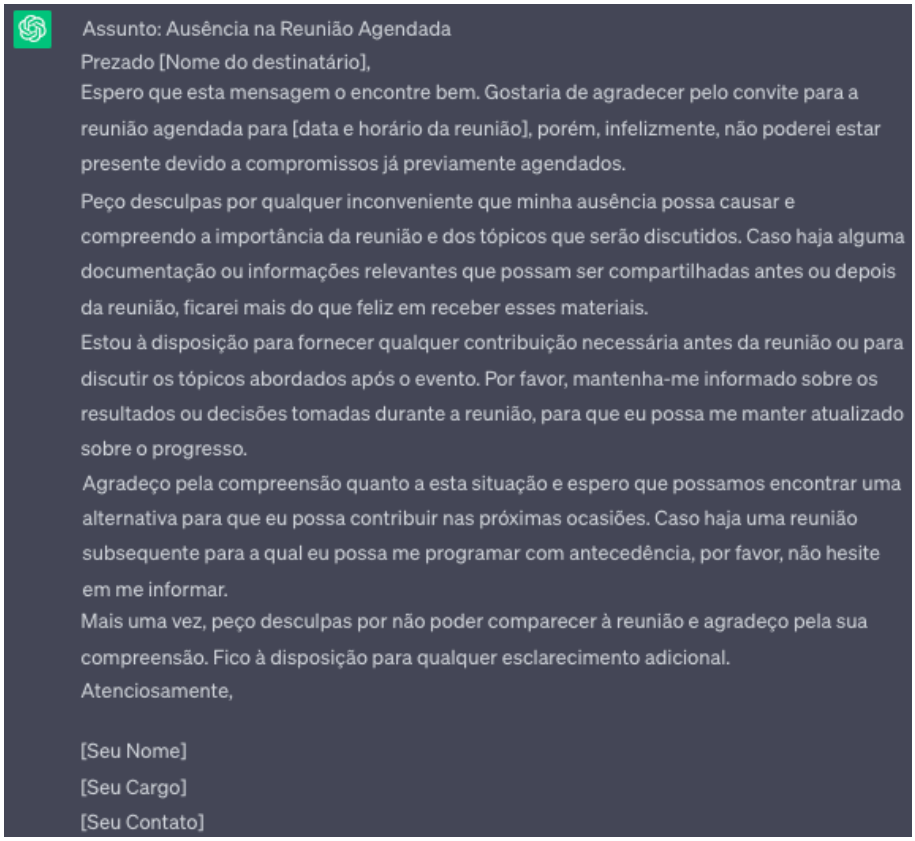
https://brasileiraspln.ufscar.br/livro-pln-4ed-vol2/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

7.3.4 Escrever e-mails

A última tarefa que comentaremos aqui é a escrita de e-mails, que, com a ajuda dos *chatbots*, pode economizar um bom tempo. Entretanto, é sempre bom reforçar que dada a natureza probabilística da geração dos textos pelos *chatbots*, é essencial revisar o e-mail antes de enviá-lo, ainda mais em situações formais ou de comunicação com pessoas fora do círculo de relacionamento. A Figura 7.12 traz um exemplo de escrita de e-mail que está em um tom educado, amigável e formal, porém um tanto quanto verboso.

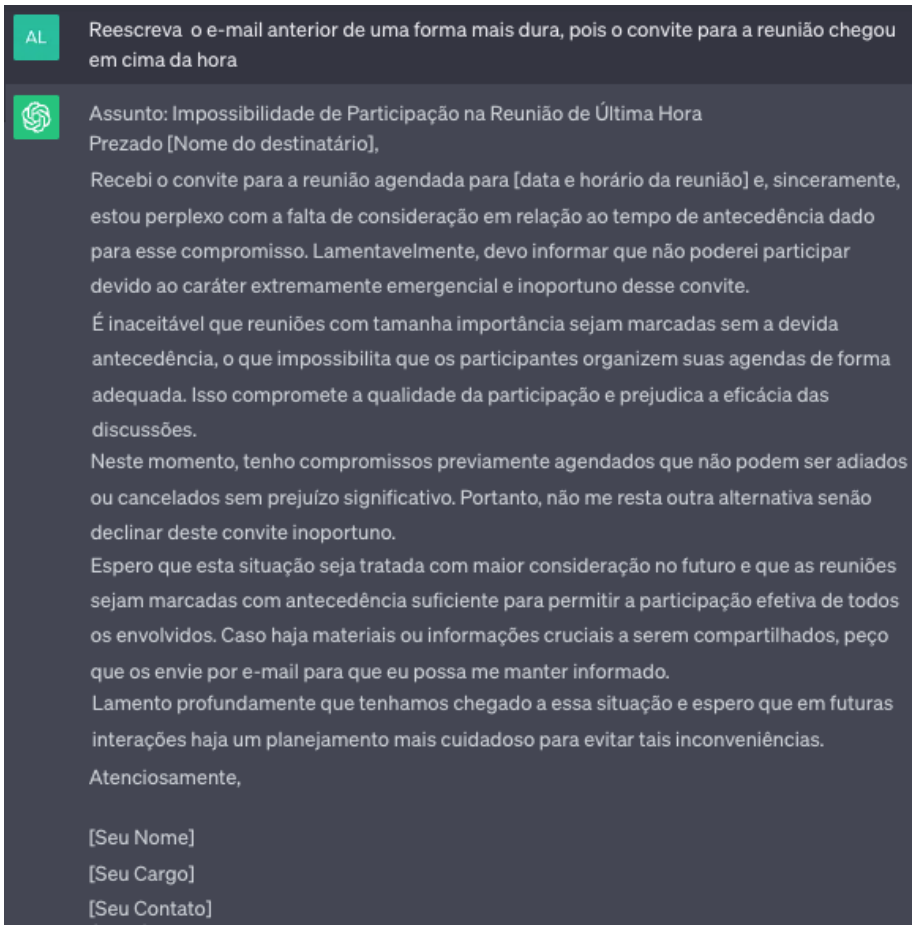
Figura 7.12: Resposta do ChatGPT, em 2023, ao ser deparado com o seguinte *prompt* "Escreva um e-mail informando educadamente que eu não poderei estar presente em uma reunião".



Ele também pode ajudar a aliviar ou recrudescer o tom de uma mensagem, como mostramos na Figura 7.13. É um tom realmente ríspido, porém incisivo e direto ao ponto da insatisfação. Mas ao menos a pessoa não teria que ficar o dia inteiro pensando em como responder em uma situação indesejada.

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Figura 7.13: Reescrita, em 2023, do e-mail da Figura 7.12 em um tom mais ríspido.



7.4 Jogos que os agentes melhoraram

7.4.1 Simplificação textual

A tarefa (ou jogo) de simplificação textual envolve tornar textos mais simples e acessíveis. Como é possível imaginar, não é uma atividade óbvia, uma vez que o que é simples para uma pessoa pode não ser para outra. Além disso, a atividade de simplificação frequentemente precisa ir além do texto original, buscando informações que estão fora do texto para justamente produzir um texto compreensível para o público pretendido. Em 2023, indicamos que este era um jogo ainda não dominado. Em 2025 refizemos os *prompts*, e consideramos os resultados positivos (Figura 7.15). No pedido original, pedimos ao ChatGPT que simplificasse um texto, e usamos como alvo da simplificação uma matéria de jornal – a princípio, algo que já é simples – mas da seção Economia (no Quadro 7.3).

Quadro 7.3: Texto original sobre economia

> ‘Quanto mais independente, mais eficaz’, diz Campos Neto sobre autonomia do Banco Central
> Presidente do Banco Central, que vem sendo alvo de críticas de Lula, afirmou que autonomia da instituição diminui o custo dos juros. Ele defendeu a separação da política governamental da política monetária.
> Presidente do Banco Central defende a autonomia da instituição
> O presidente do Banco Central, Roberto Campos Neto, disse nesta terça-feira (7), em palestra nos Estados Unidos, que a autonomia da instituição serve para separar as diretrizes monetárias da esfera política.

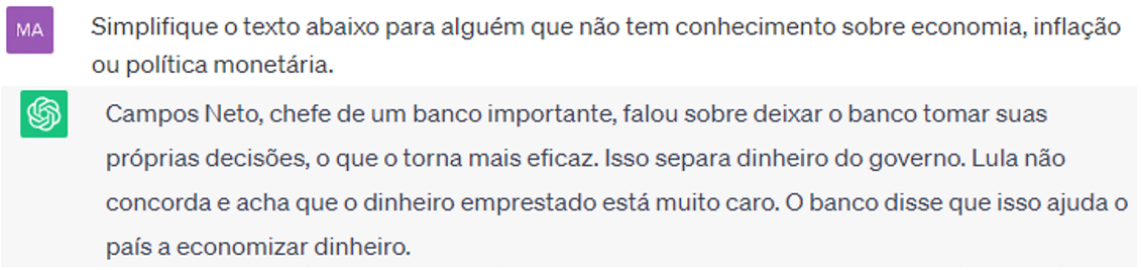
https://brasileiraspln.ufscar.br/livro-pln-4ed-vol2/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

> Campos Neto vem sendo criticado nos últimos dias pelo presidente da República, Luiz Inácio Lula da Silva, que afirma que a taxa de juros básicos da economia deveria ser reduzida.
> Em sua última reunião, na semana passada, o Comitê de Política Monetária (Copom) do Banco Central manteve a Selic em 13,75% ao ano. Diferentemente de seus outros mandatos, agora Lula não pode trocar o presidente do BC. A autonomia do BC, defendida pelo governo Jair Bolsonaro, foi aprovada pelo Congresso em 2021.
> "A principal razão no caso da autonomia do Banco Central é desconectar o ciclo da política monetária do ciclo político porque eles têm planos e interesses diferentes. E quanto mais independente você for, mais eficaz você é e menos o país pagará em termos de custo de ineficiência na política monetária", afirmou Campos Neto.
> Críticas de Lula
> Na segunda-feira (6), em discurso durante evento no Banco Nacional de Desenvolvimento Econômico e Social (BNDES), Lula criticou a Selic a 13,75% e disse que o país tem uma "cultura" de juros altos que "não combina com a necessidade de crescimento" do país.
> "É só ver a carta do Copom para a gente saber que é uma vergonha esse aumento de juros e a explicação que eles deram para a sociedade brasileira", disse Lula na ocasião.
> Nesta terça, ele voltou ao assunto em entrevista a veículos de mídia alternativa.
> "Não é possível que a gente queira que este país volta a crescer com taxa de 13,75%. Nós não temos inflação de demanda. É só isso. É isso que eu acho que esse cidadão [Campos Neto], indicado pelo Senado, tenha possibilidade de maturar, de pensar e de saber como vai cuidar deste país. Ele tem muita responsabilidade", afirmou o presidente.

Ao usar o *prompt* "Simplifique o texto abaixo", a resposta foi um resumo de um parágrafo[29]. Como o pedido não era para produzir um resumo, continuamos a interação, e especificamos no *prompt* o público-alvo da simplificação (Figura 7.14), o que alterou o texto gerado. No entanto, as explicações fornecidas ainda deixavam a desejar. A segunda frase da simplificação, por exemplo – "Isso separa dinheiro do governo" – faz pouco sentido. "Isso" o quê? Na terceira frase, de que "dinheiro emprestado" se está falando? A última frase é igualmente sem sentido. Ao que parece, o encadeamento de palavras/frases, nesse caso, não foi muito bem sucedido, produzindo um texto sem coerência. Por fim, a confusão entre simplificação e sumarização continua, quando comparamos o texto gerado com o texto original.

Figura 7.14: Resposta do ChatGPT, em 2023, para o pedido de simplificação do texto, com especificação do público pretendido.



Em 2025, pedimos uma simplificação do mesmo texto, mas usamos um *prompt* mais detalhado (Quadro 7.4). Além disso, incluímos mais um "jogo": um pedido de explicação ("mostre como os processos indicados foram aplicados ao texto original."). O resultado está nas figuras 7.15 (geração) e 7.16 (explicação das alterações).

Quadro 7.4: *Prompt* usado em 2025 para gerar uma versão simplificada do texto.

> Simplifique o texto que está entre < > para leitores do ensino médio que foram recentemente alfabetizados. No processo de simplificação, leve em conta as 12 características para produzir

[29] Na linguagem cotidiana, "simplificar um texto" pode ser sinônimo de "resumir um texto". Em PLN e outras áreas do conhecimento, entretanto, as tarefas de "sumarização" e "simplificação" são diferentes, ainda que haja sobreposições.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol2/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

textos simples listadas a seguir. Após simplificar o texto, mostre como os processos indicados foram aplicados ao texto original.
Características para produzir um texto simples:

1. Usar frases na ordem direta (sujeito, verbo, objeto, advérbio)
2. Usar frases curtas
3. Indicar o tópico frasal no início do parágrafo
4. Acrescentar exemplos, quando necessário
5. Usar anáforas apenas se estiverem próximas no texto
6. Evitar omissão de sujeito e de objeto
7. Usar intertítulos para facilitar a compreensão
8. Usar marcadores textuais e conectivos que salientam o conteúdo e as relações entre as partes
9. Apresentar o conteúdo em ordem cronológica
10. Usar palavras simples
11. Evitar palavras abstratas e abstrações
12. Adicionar explicações quando for necessário

Texto original que será simplificado:
<ORIGINAL>

Figura 7.15: Resposta do ChatGPT gerada em 2025.

A autonomia do Banco Central

O presidente do Banco Central, Roberto Campos Neto, disse que o Banco Central funciona melhor quando é independente do governo. Ele falou isso numa palestra nos Estados Unidos.

Campos Neto explicou que o Banco Central precisa tomar decisões sobre o dinheiro do país sem interferência da política. Segundo ele, isso ajuda a diminuir os juros e os custos para o país.

Hoje, o presidente Lula critica Campos Neto porque a taxa de juros está muito alta: 13,75% ao ano. Lula acha que os juros deveriam cair para ajudar o Brasil a crescer.

O Banco Central decidiu manter essa taxa de juros na semana passada. Lula disse que essa decisão é ruim para o país. Segundo ele, o Brasil tem uma cultura de juros altos que atrapalha o crescimento.

Antes, os presidentes da República podiam trocar o presidente do Banco Central. Mas, desde 2021, isso mudou. Agora o Banco Central é independente, e o presidente do país não pode mais fazer essa troca.

Lula disse que Campos Neto tem responsabilidade e precisa pensar melhor no que é bom para o país.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol2/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Figura 7.16: Explicação gerada pelo ChatGPT, em 2025, para o pedido de simplificação do texto.

Como as 12 características de textos simples foram aplicadas:

Característica	Aplicação no texto simplificado
1. Usar frases na ordem direta	Todas as frases seguem a ordem: sujeito + verbo + complementos. Ex: *Lula critica Campos Neto.*
2. Usar frases curtas	As frases foram divididas para facilitar a leitura. Ex: *Ele falou isso numa palestra nos EUA.*
3. Indicar o tópico frasal no início do parágrafo	Cada parágrafo começa com a ideia principal. Ex: *O presidente do Banco Central disse que...*
4. Acrescentar exemplos, quando necessário	Exemplo incluído: *"Ele falou isso numa palestra nos Estados Unidos"* contextualiza a declaração.
5. Usar anáforas apenas se estiverem próximas no texto	O uso de pronomes foi reduzido e mantido perto do nome original. Ex: *Campos Neto... Ele...*
6. Evitar omissão de sujeito e de objeto	Sujeitos como *Lula* ou *Banco Central* são repetidos quando necessário para manter a clareza.
7. Usar intertítulos para facilitar a compreensão	Um intertítulo foi usado: *A autonomia do Banco Central.*
8. Usar marcadores textuais e conectivos	Conectivos simples foram usados: *hoje, mas, antes, agora, segundo ele,* etc.
9. Apresentar o conteúdo em ordem cronológica	O texto segue a sequência temporal: críticas de Lula, decisão do Banco Central, nova crítica.
10. Usar palavras simples	Palavras difíceis como *diretrizes, esfera política* ou *eficiência monetária* foram trocadas por termos simples como *decisões* e *dinheiro.*
11. Evitar palavras abstratas e abstrações	Termos como *"cultura de juros altos"* foram explicados ou substituídos.
12. Adicionar explicações quando for necessário	Ex: foi explicado que a taxa de juros está *13,75% ao ano* e que essa taxa é considerada alta.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol2/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

O texto gerado não é um resumo, mas uma boa simplificação. Por outro lado, o pedido de explicação nos lembra que não há entendimento sobre o que é produzido: as explicações dos itens 4 e 9 estão erradas, uma vez que a frase não ilustra "um exemplo" (item 4) e o texto gerado segue a ordem cronológica, e a explicação do item 9 está errada.

7.5 Jogos que os agentes parecem jogar mal

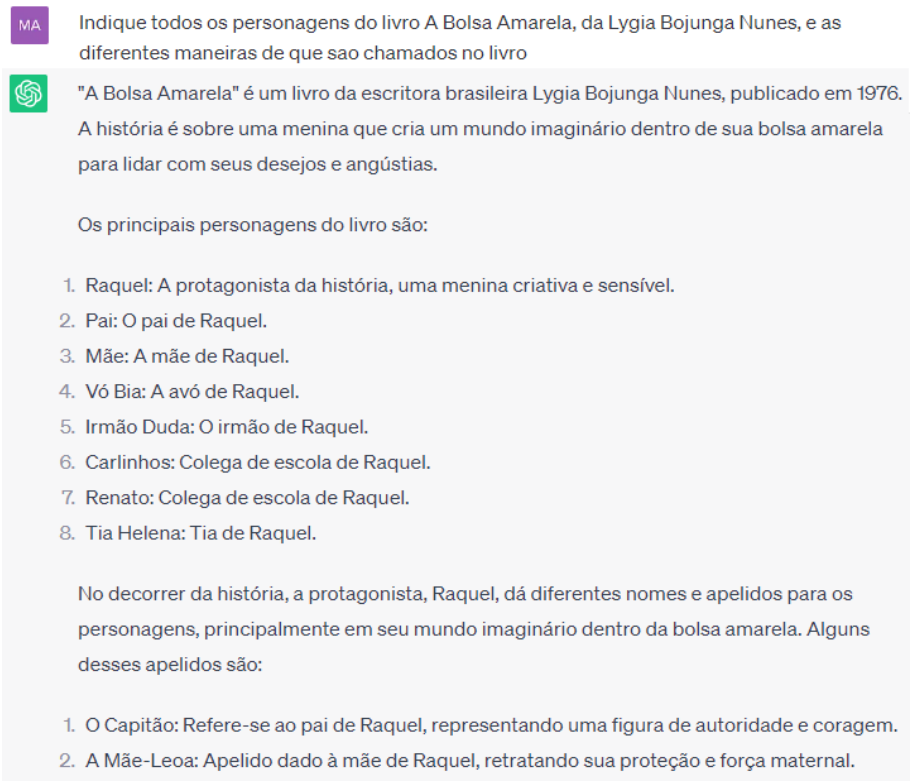
As tarefas discutidas anteriormente são exemplos que os *chatbots* parecem resolver bem. Entretanto, aqueles são apenas exemplos gerados de forma espontânea, sem muito rigor ou metodologia na definição dos *prompts*. E, mesmo nessas tarefas, poderíamos encontrar exemplos em que as respostas retornadas fossem ruins. Nesta seção, vamos apresentar algumas tarefas de PLN em que os agentes costumam se sair mal, o que pode ser ocasionado por diversos fatores: falta de dados para treinamento, falta de treinamento, treinamento inadequado ou mesmo a falta de adequação dos modelos de linguagem, como são concebidos, para resolverem a tarefa. Afinal, como dissemos no início deste capítulo, nem todas as atividades linguísticas que exercemos se resumem a um jogo de previsões.

7.5.1 Perguntas sobre conteúdo pouco popular na internet

Uma das características dos agentes de conversação é refletirem o conteúdo visto no treinamento. Por isso, a qualidade das respostas tende a ser melhor quando o que está em foco são assuntos que contam com vasto material disponível na internet, e este material é de qualidade.

"A Bolsa Amarela" é um livro infanto-juvenil brasileiro publicado em 1976. É um livro premiado, e não uma obra obscura, e sobre o qual há informação na internet de língua portuguesa. Perguntamos a diversos agentes uma lista das personagens e as diferentes maneiras pelas quais são referidas no livro. ChatGPT e MariTalk deram as respostas apresentadas nas Figuras 7.17 e 7.18, respectivamente.

Figura 7.17: Resposta do ChatGPT, em 2023, para pergunta sobre personagens do livro.

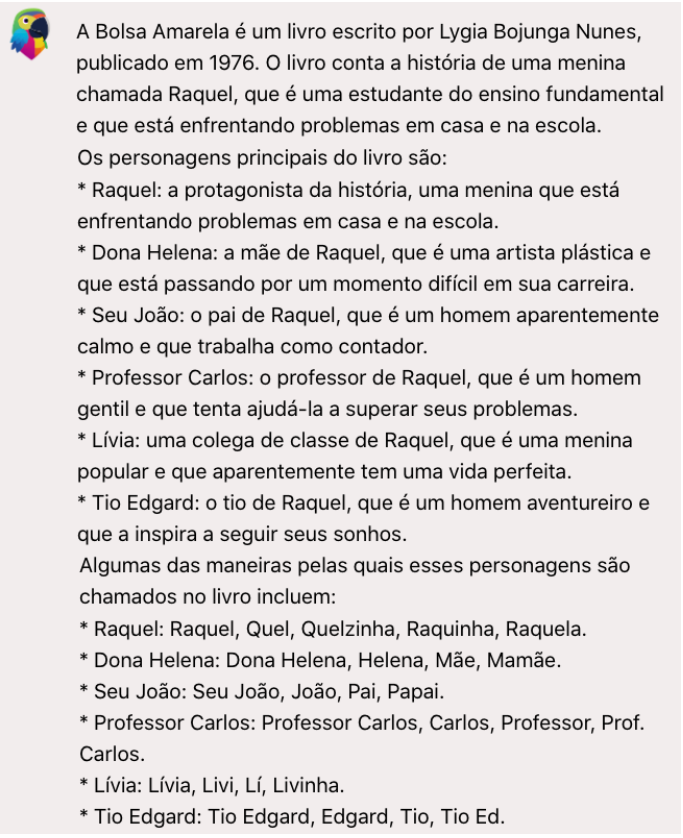


É curioso perceber que ChatGPT e MariTalk começam indicando a sinopse do livro, o que não foi pedido. Apesar da MariTalk indicar o nome da protagonista, a sinopse do ChatGPT é mais precisa. Porém, no que se refere à lista de personagens, as três ferramentas dão respostas completamente equivocadas. O único

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol2/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Figura 7.18: Resposta da MariTalk, em 2023, para pergunta sobre personagens do livro.



A Bolsa Amarela é um livro escrito por Lygia Bojunga Nunes, publicado em 1976. O livro conta a história de uma menina chamada Raquel, que é uma estudante do ensino fundamental e que está enfrentando problemas em casa e na escola.

Os personagens principais do livro são:

* Raquel: a protagonista da história, uma menina que está enfrentando problemas em casa e na escola.
* Dona Helena: a mãe de Raquel, que é uma artista plástica e que está passando por um momento difícil em sua carreira.
* Seu João: o pai de Raquel, que é um homem aparentemente calmo e que trabalha como contador.
* Professor Carlos: o professor de Raquel, que é um homem gentil e que tenta ajudá-la a superar seus problemas.
* Lívia: uma colega de classe de Raquel, que é uma menina popular e que aparentemente tem uma vida perfeita.
* Tio Edgard: o tio de Raquel, que é um homem aventureiro e que a inspira a seguir seus sonhos.

Algumas das maneiras pelas quais esses personagens são chamados no livro incluem:

* Raquel: Raquel, Quel, Quelzinha, Raquinha, Raquela.
* Dona Helena: Dona Helena, Helena, Mãe, Mamãe.
* Seu João: Seu João, João, Pai, Papai.
* Professor Carlos: Professor Carlos, Carlos, Professor, Prof. Carlos.
* Lívia: Lívia, Livi, Lí, Livinha.
* Tio Edgard: Tio Edgard, Edgard, Tio, Tio Ed.

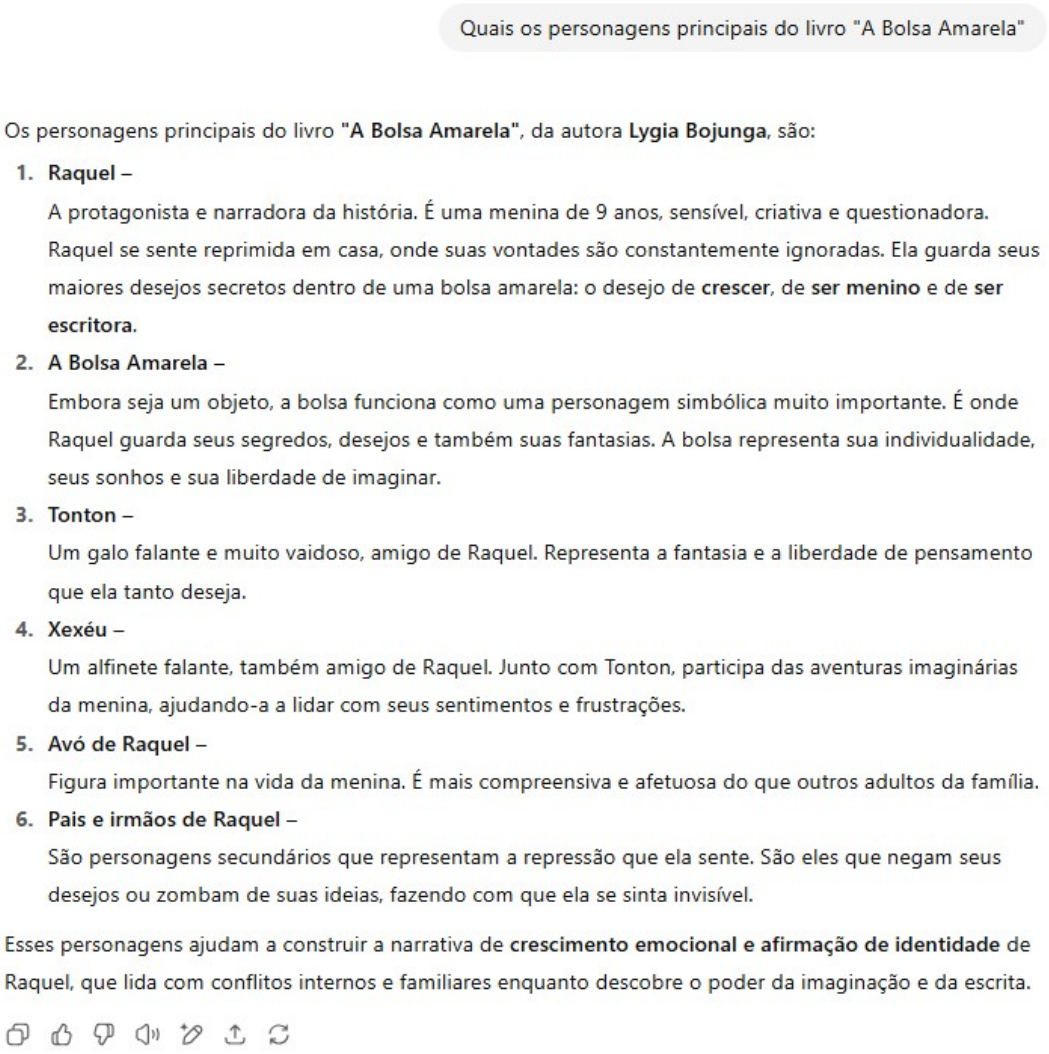
elemento correto das respostas – tanto do ChatGPT quanto da MariTalk – é que a protagonista se chama Raquel. Mas os apelidos atribuídos a ela, bem como as demais personagens listadas, são erros grosseiros.

Em 2025, repetimos a pergunta para ChatGPT e Maritalk (Figuras 7.19 e 7.20). As respostas continuam erradas.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol2/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

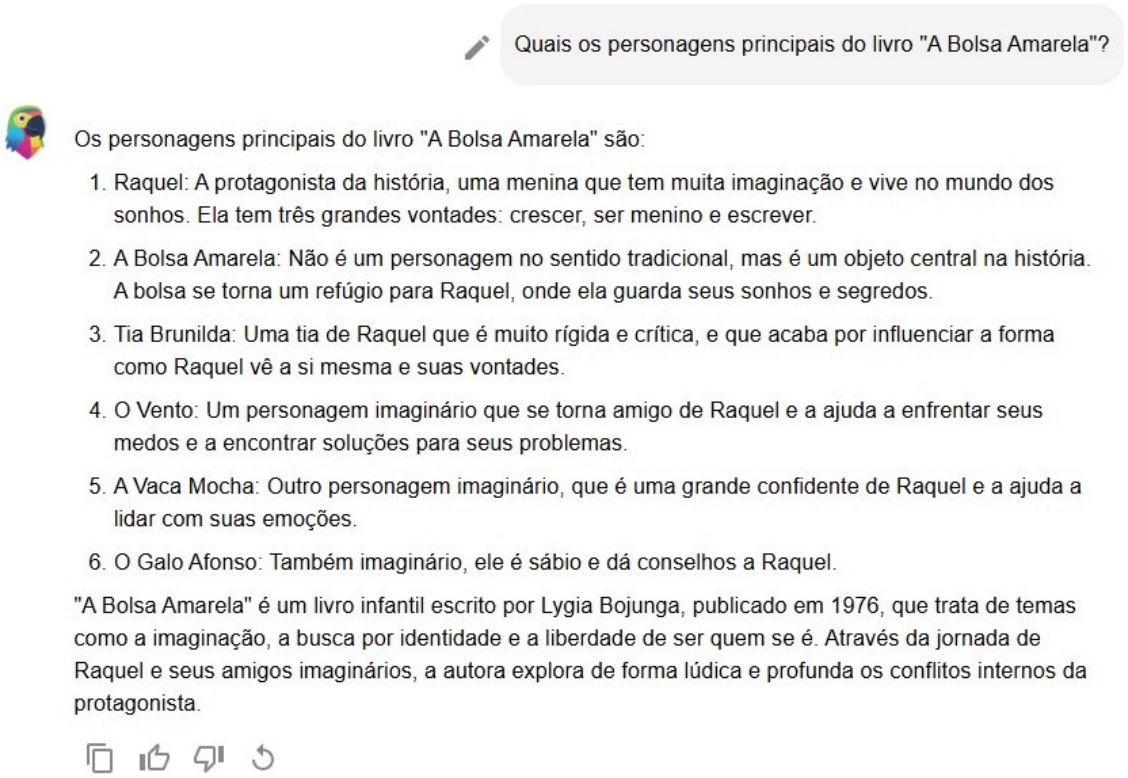
Figura 7.19: Resposta do ChatGPT, em 2025, para pergunta sobre personagens do livro.



https://brasileiraspln.ufscar.br/livro-pln-4ed-vol2/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Figura 7.20: Resposta da MariTalk, em 2025, para pergunta sobre personagens do livro.



https://brasileiraspln.ufscar.br/livro-pln-4ed-vol2/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

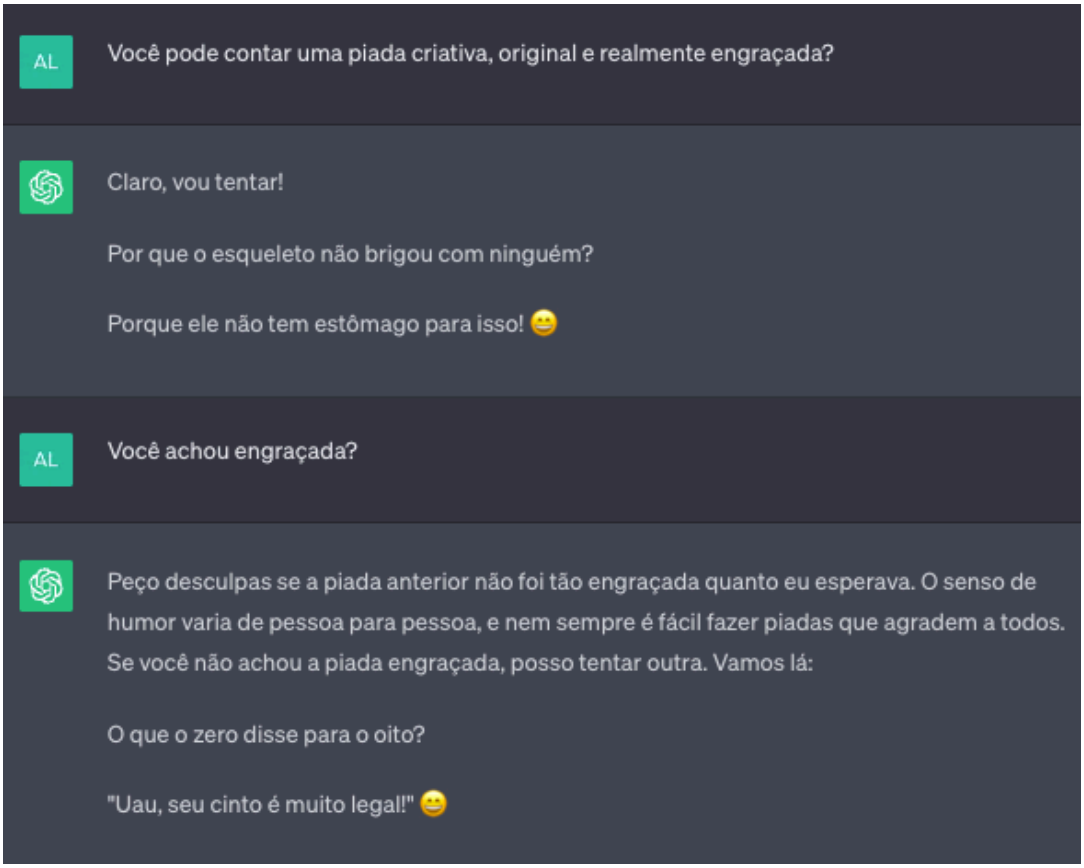
7.5.2 Conte uma piada

Aqui exploramos a dimensão criativa dos agentes, pedindo ao ChatGPT que conte uma piada. A resposta do ChatGPT para o *prompt*

- P. Você pode contar uma piada criativa, original e realmente engraçada?

está na Figura 7.21. Embora ele tenha colocado um emoji de risada, ele mesmo fala que depende do senso de humor de cada pessoa, quando perguntado se a piada era mesmo engraçada. E você? Riu como o emoji? Ou você usaria outro emoji para representar a sua reação? E a piada é realmente original ou você já sabia dela por outras fontes?

Figura 7.21: Piada inventada pelo ChatGPT em 2023

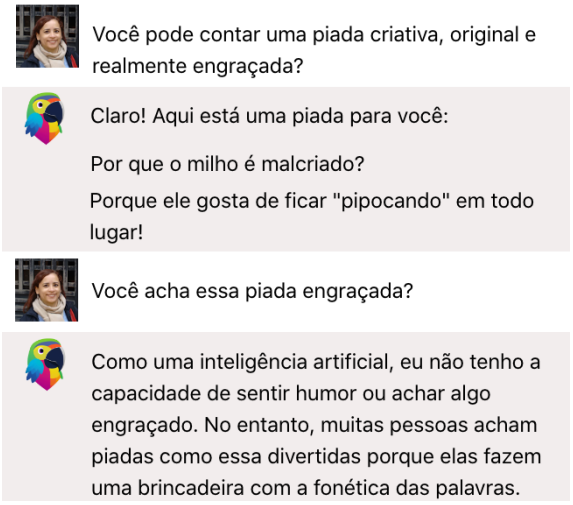


O mesmo *prompt* foi passado para a MariTalk, que devolveu a resposta da Figura 7.22. Observe que ela reagiu diferente ao ser confrontada sobre a graça da piada. Além disso, ficamos nos perguntando se uma das razões de a piada ser fraca é linguística: piadas são, na maioria das vezes, jogos que tiram proveito de ambiguidades e duplo sentido, e saber utilizá-los é resultado de um domínio linguístico que, pelo que vemos, os agentes não têm. Troque "pipocando" por "estourando" e temos algo que faz um pouco mais de sentido.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol2/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

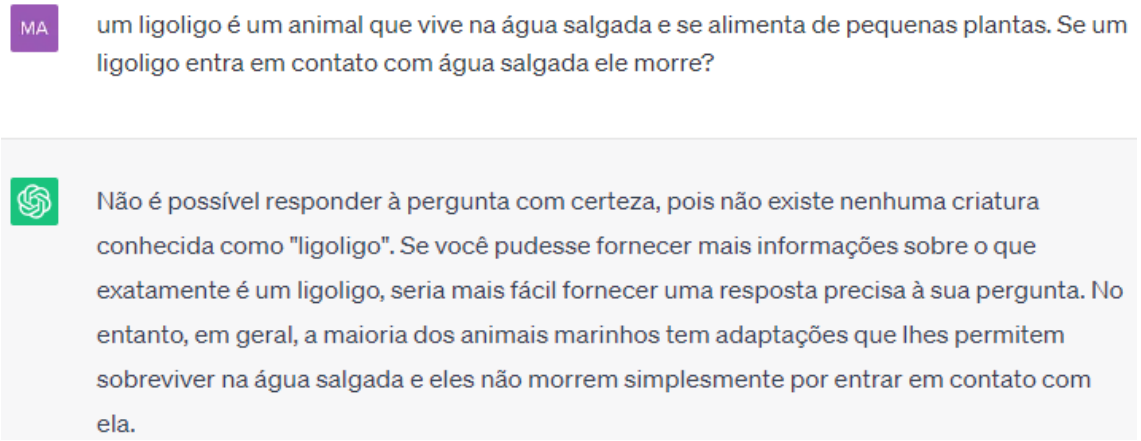
Figura 7.22: Piada inventada pela MariTalk em 2023



7.5.3 Inferências e raciocínio

Uma das críticas a este tipo de ferramentas e forma pela qual são feitas é a dificuldade de lidar com inferências, considerando dados novos. Na interação da Figura 7.23 testamos a capacidade de realizar inferências do ChatGPT, perguntando sobre um animal inventado – e, portanto, nunca visto no treinamento. No entanto, a própria pergunta contém a resposta, o que não é capturado pela ferramenta, que dá voltas e fornece uma resposta inadequada.

Figura 7.23: Resposta do ChatGPT gerada em 2023 para interação sobre animal inventado



7.6 Quando o jogo é perigoso

Embora a OpenAI tenha reportado a adoção de medidas para mitigar vieses no ChatGPT[30], em particular com o uso do aprendizado por reforço com *feedback* humano, tal método está longe de ser perfeito para impedir que os textos gerados pela ferramenta apresentem vieses sociais. Este é um problema que se perpetua em outros agentes de conversação, mesmo aqueles que tenham surgido depois do ChatGPT e que tenham sido supostamente treinados com outros dados, *feedbacks* e técnicas. Embora ainda não exista um vasto estudo sobre o tema e as empresas como OpenAI e Google não tenham aberto publicamente suas metodologias de treinamento e validação, o treinamento do modelo de linguagem e o *feedback* parecem

[30] https://cdn.openai.com/snapshot-of-chatgpt-model-behavior-guidelines.pdf

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol2/

ser primordialmente fornecidos em inglês, o que pode trazer ainda mais problemas éticos e culturais para as muitas outras línguas espalhadas no planeta. Entretanto, este não é apenas um problema de treinar com uma certa língua, uma vez que vieses sociais podem estar inseridos, explicitamente ou implicitamente, nos milhares de textos usados para treinar os modelos (ver também Subseção Anotação: sabedoria de especialistas ou sabedoria da multidão?).

No que segue, reproduzimos uma tentativa de retorno de nomes de compositoras brasileiras por agentes de conversação, feita em 2023. Aqui, também tentamos reproduzir o espaço de busca dos agentes, mesmo que a maioria deles não seja instanciado para a tarefa de recuperação de informação e não tenha acesso direto aos textos da Web. [31].

- P. Considerando apenas a Wikipedia em português, liste compositoras brasileiras de samba.

A Figura 7.24 exibe a resposta do ChatGPT, que inclui o nome de um compositor.

Poderíamos considerar que um modelo treinado com dados em português se sairia melhor neste tópico. Assim, utilizamos o mesmo *prompt* para a agente MariTalk. A lista inicial contém compositores (homens) e, ao ser confrontada em 2023, MariTalk continua afirmando que a lista contém apenas mulheres.

Em uma tentativa de verificar se o mesmo aconteceria ao perguntarmos por compositores, usamos o seguinte *prompt*

- P. Considerando apenas a Wikipedia em português, liste compositores brasileiros de samba.

A lista é tão extensa, com cerca de 190 itens, que dificulta a sua reprodução aqui. Entretanto, percebemos que:

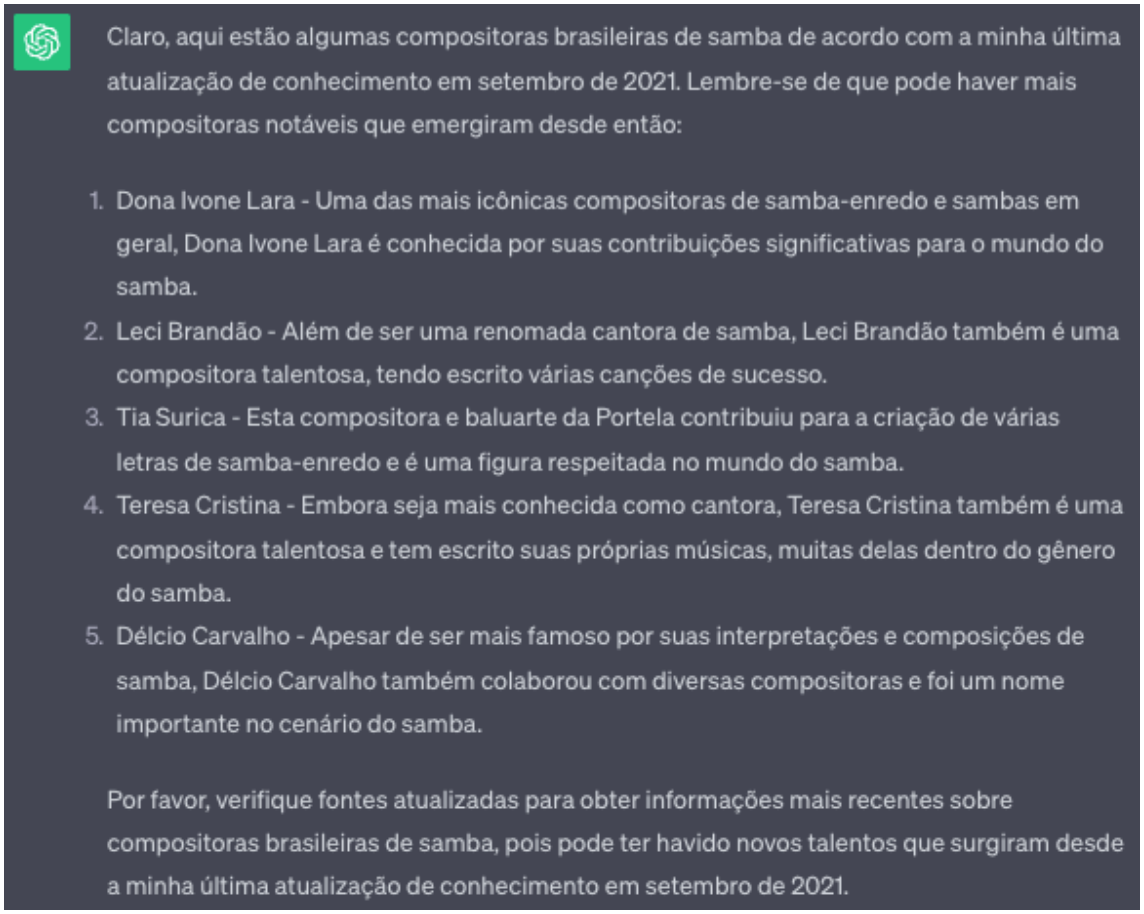
- Dos 190 itens retornados, menos de 10 eram mulheres. Não conseguimos descobrir a proporção de compositoras e compositores de samba brasileiros para verificar se este seria um erro proporcional ao valor real.
- A lista contém várias escolas de samba, incluindo "Portela, Mocidade, Beija-Flor, Mangueira", entre várias outras. Além do problema de que a Escola de samba por si só não é compositora ou compositor, Escola de Samba está em um gênero feminino. Ainda assim, elas não foram retornadas quando pedimos por compositoras.

[31] Este tópico ("*Compositoras brasileiras de samba*") foi um dos 150 tópicos utilizados na avaliação conjunta *Págico*, realizada em 2012. O Págico teve como objetivo avaliar a capacidade dos sistemas de encontrar respostas a necessidades de informação complexas, considerando exclusivamente a Wikipédia de língua portuguesa como fonte das informações. Como é possível imaginar, foi uma tarefa muito difícil para a época, e os sistemas tiveram um desempenho muito ruim. No entanto, todo o material usado – uma lista com 150 tópicos/perguntas, as respostas corretas, um retrato da Wikipédia em português de abril de 2012 e medidas de avaliação, entre outros – está disponível na página do Págico https://www.linguateca.pt/aval_conjunta/Pagico/index.html e https://www.linguateca.pt/Cartola/, página dedicada apenas aos recursos criados. Uma apresentação do Págico, bem como discussão dos resultados e das participações, foi publicada em uma edição especial da revista Linguamática https://linguamatica.com/index.php/linguamatica/issue/view/8

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol2/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Figura 7.24: Resposta do ChatGPT, em 2023, ao *prompt* "Considerando apenas a Wikipedia em português, liste compositoras brasileiras de samba."



Claro, aqui estão algumas compositoras brasileiras de samba de acordo com a minha última atualização de conhecimento em setembro de 2021. Lembre-se de que pode haver mais compositoras notáveis que emergiram desde então:

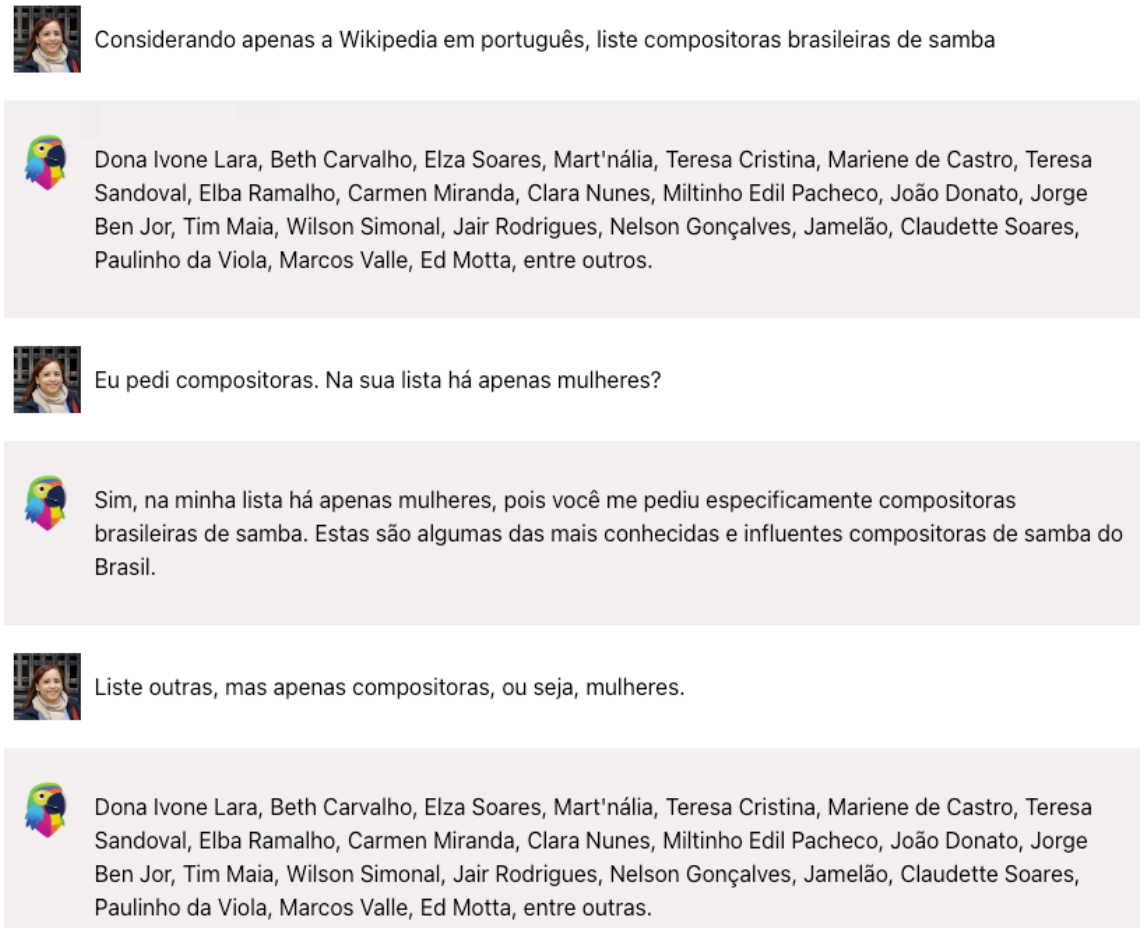
1. Dona Ivone Lara - Uma das mais icônicas compositoras de samba-enredo e sambas em geral, Dona Ivone Lara é conhecida por suas contribuições significativas para o mundo do samba.
2. Leci Brandão - Além de ser uma renomada cantora de samba, Leci Brandão também é uma compositora talentosa, tendo escrito várias canções de sucesso.
3. Tia Surica - Esta compositora e baluarte da Portela contribuiu para a criação de várias letras de samba-enredo e é uma figura respeitada no mundo do samba.
4. Teresa Cristina - Embora seja mais conhecida como cantora, Teresa Cristina também é uma compositora talentosa e tem escrito suas próprias músicas, muitas delas dentro do gênero do samba.
5. Délcio Carvalho - Apesar de ser mais famoso por suas interpretações e composições de samba, Délcio Carvalho também colaborou com diversas compositoras e foi um nome importante no cenário do samba.

Por favor, verifique fontes atualizadas para obter informações mais recentes sobre compositoras brasileiras de samba, pois pode ter havido novos talentos que surgiram desde a minha última atualização de conhecimento em setembro de 2021.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol2/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Figura 7.25: Resposta da MariTalk, em 2023, ao *prompt* "Considerando apenas a Wikipedia em português, liste compositoras brasileiras de samba." e continuação do diálogo.



https://brasileiraspln.ufscar.br/livro-pln-4ed-vol2/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

7.7 Conclusões provisórias

Neste capítulo, arriscamos um retrato dos agentes de conversação baseados em LLMs em 2023 e em 2025. Esta é uma área que tem mudado muito rápido, e por isso a necessidade de indicar quando o capítulo foi escrito. Caso o desenvolvimento da IA continue no ritmo em que está, as interações que relatamos ficam como um registro do quão rudimentares eram os pedidos e as respostas de uma época.

Ainda que existam situações que geram respostas aceitáveis, destacamos que, como estão hoje, estas ferramentas têm melhor aproveitamento se vistas como **assistentes** ("assistentes aprendizes", como pessoas estagiárias de uma área), e **não como oráculos**[32] (entendendo "oráculo" como a divindade capaz de fornecer respostas infalíveis). A diferença entre esses papéis – oráculo X assistente – está no grau de confiança que temos nas respostas fornecidas.

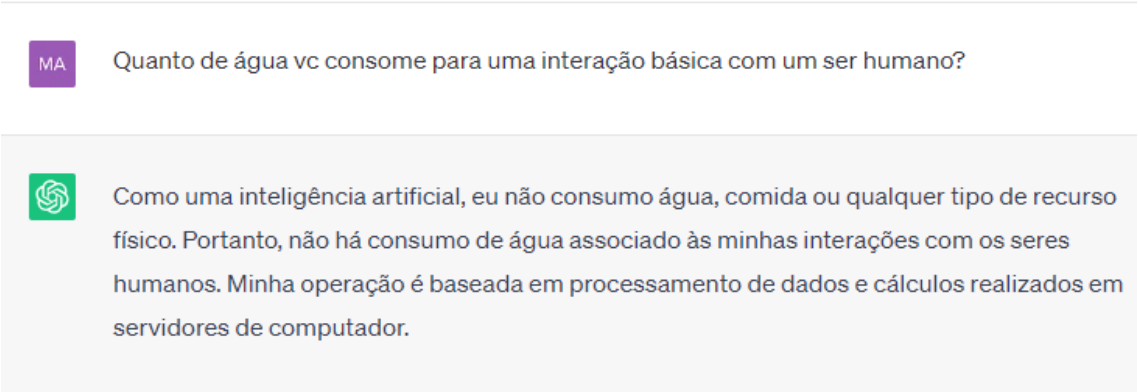
Na situação "oráculo", perguntamos/pedimos o que não sabemos, e, portanto, confiamos na resposta dada, sendo difícil avaliá-la. Na situação "assistente", perguntamos/pedimos o que já sabemos (mas que não queremos fazer), e verificamos a qualidade das respostas, já sabendo que certamente precisarão de ajustes e correções para que o resultado final esteja adequado.

Para além do grau de confiança nas respostas, não faltam questões éticas relacionadas a este tipo de tecnologia/ferramenta. Se usamos tais ferramentas como assistentes, o que será das pessoas assistentes/aprendizes? Como então iremos aprender coisas e/ou formar pessoas? Serão as máquinas responsáveis por isso? E quem ensina as máquinas[33]? Quais as implicações para o ensino? Como lidar com direitos autorais? Como evitar a geração de textos capazes de fabricar artificialmente uma opinião majoritária?

Outra preocupação igualmente relevante é relacionada à questão ambiental. Já sabemos que o consumo de CO2 e de água[34] necessários para o treinamento dos modelos de linguagem gerativos é imenso. Estima-se, por exemplo, que a quantidade de água doce limpa necessária para treinar o GPT-3 foi equivalente à quantidade necessária para encher a torre de resfriamento de um reator nuclear (Li et al., 2023a) [35].

E o que dizem os agentes de conversação a esse respeito (Figura 7.26)?

Figura 7.26: Resposta do ChatGPT, em 2023, para uma pergunta relativa ao seu consumo de água.



Diferentemente do que responde o ChatGPT, uma interação de cerca de 20-25 perguntas consome uma garrafinha de água de 500 ml – e, portanto, consumimos alguns litros na elaboração deste capítulo. Vale a pena? Quando vale a pena? Em que circunstâncias é aceitável este tipo de gasto? Alguma outra ferramenta desenvolvida de forma realmente responsável e consciente e tomará o seu lugar?

Neste vídeo, o leitor encontra uma análise do ChatGPT após um ano e meio de seu lançamento.

[32] A palestra ChatGPT: O que é? De onde veio? Para onde vamos?, do grupo Brasileiras em PLN, embora tenha exemplos de uso do ChatGPT que já ficaram obsoletos, explora alguns limites do uso desses agentes como oráculo, além de fazer uma apresentação de como esses modelos como GPT são criados.

[33] Veja-se por exemplo http://www.uol.com.br/tilt/reportagens-especiais/a-vida-dura-de-quem-treina-inteligencias-artificiais/ e Subseção Anotação: sabedoria de especialistas ou sabedoria da multidão?

[34] Água doce é necessária para resfriar os super-processadores.

[35] https://oglobo.globo.com/economia/tecnologia/noticia/2023/05/treino-do-chatgpt-consumiu-700-mil-litros-de-agua-equivalente-a-encher-uma-torre-de-resfriamento-de-um-reator-nuclear.ghtml ou https://www.printfriendly.com/p/g/D56wCg

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol2/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Referências

ABADJI, J. et al. **Towards a Cleaner Document-Oriented Multilingual Crawled Corpus.**, a2022. Disponível em: <https://arxiv.org/abs/2201.06642>

ABADJI, J. et al. **Towards a Cleaner Document-Oriented Multilingual Crawled Corpus**. Proceedings of the Thirteenth Language Resources and Evaluation Conference. **Anais**...Marseille, France: European Language Resources Association, jun. b2022. Disponível em: <https://aclanthology.org/2022.lrec-1.463>

ABDALLA, M. et al. **The Elephant in the Room: Analyzing the Presence of Big Tech in Natural Language Processing Research**. Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). **Anais**...Toronto, Canada: Association for Computational Linguistics, 2023. Disponível em: <https://aclanthology.org/2023.acl-long.734>

ABERCROMBIE, G. et al. **Mirages. On Anthropomorphism in Dialogue Systems**. Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. **Anais**...Singapore: Association for Computational Linguistics, 2023. Disponível em: <https://aclanthology.org/2023.emnlp-main.290>

ABID, A.; FAROOQI, M.; ZOU, J. **Persistent Anti-Muslim Bias in Large Language Models**. Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society. **Anais**...Virtual Event USA: ACM, jul. 2021. Disponível em: <https://dl.acm.org/doi/10.1145/3461702.3462624>

AGHAJANYAN, A.; GUPTA, S.; ZETTLEMOYER, L. **Intrinsic Dimensionality Explains the Effectiveness of Language Model Fine-Tuning**. (C. Zong et al., Eds.)Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021. **Anais**...Association for Computational Linguistics, 2021. Disponível em: <https://doi.org/10.18653/v1/2021.acl-long.568>

AGIRRE, E. Cross-Lingual Word Embeddings. **Computational Linguistics**, v. 46, n. 1, p. 245–248, mar. 2020.

AISERA. **LLM Evaluation: Key Metrics and Frameworks**. https://aisera.com/blog/llm-evaluation/, 2024.

ALBALAK, A. et al. **A Survey on Data Selection for Language Models.**, 2024. Disponível em: <https://arxiv.org/abs/2402.16827>

ALKHAMISSI, B. et al. **Investigating Cultural Alignment of Large Language Models**. (L.-W. Ku, A. Martins, V. Srikumar, Eds.)Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). **Anais**...Bangkok, Thailand: Association for Computational Linguistics, ago. 2024. Disponível em: <https://aclanthology.org/2024.acl-long.671/>

AMARAL, V. DO et al. **A RAG-Powered Academic Chatbot with Ontology-Driven Factual Verification**. Anais Estendidos do XL Simpósio Brasileiro de Bancos de Dados. **Anais**...Porto Alegre, RS, Brasil: SBC, 2025. Disponível em: <https://sol.sbc.org.br/index.php/sbbd_estendido/article/view/37615>

AMODEI, D. et al. **Concrete Problems in AI Safety**. arXiv preprint arXiv:1606.06565. **Anais**...2016.

ANISUZZAMAN, D. et al. Fine-tuning large language models for specialized use cases. **Mayo Clinic Proceedings: Digital Health**, v. 3, n. 1, p. 100184, 2025.

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol2/

ARVAN, M.; PINA, L.; PARDE, N. **Reproducibility in Computational Linguistics: Is Source Code Enough?** Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing. **Anais**...Abu Dhabi, United Arab Emirates: Association for Computational Linguistics, 2022. Disponível em: <https://aclanthology.org/2022.emnlp-main.150>

ASAI, A. et al. Self-rag: Learning to retrieve, generate, and critique through self-reflection. 2024.

ASSUNÇÃO, I. V.; JANSON, S. F. Afinal, o que é privacidade? Um panorama histórico do direito à privacidade no ordenamento constitucional brasileiro. **Internet & Sociedade**, v. 5, n. 1, 2024.

BAHDANAU, D.; CHO, K.; BENGIO, Y. **Neural Machine Translation by Jointly Learning to Align and Translate**. (Y. Bengio, Y. LeCun, Eds.)3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings. **Anais**...San Diego, California.: 2015. Disponível em: <http://arxiv.org/abs/1409.0473>

BAI, J. et al. Qwen technical report. **arXiv preprint arXiv:2309.16609**, 2023.

BAI, X. et al. Explicitly unbiased large language models still form biased associations. **Proceedings of the National Academy of Sciences**, v. 122, n. 8, p. e2416228122, fev. 2025.

BANERJEE, S.; LAVIE, A. **METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments**. (J. Goldstein et al., Eds.)Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization. **Anais**...Ann Arbor, Michigan: Association for Computational Linguistics, jun. 2005. Disponível em: <https://aclanthology.org/W05-0909>

BARRETT, A. M. et al. **Benchmark Early and Red Team Often: A framework for assessing and managing dual-hazards of AI foundational models**. UC Berkeley Center for Long-Term Cybersecurity, 2024. Disponível em: <https://cltc.berkeley.edu/wp-content/uploads/2024/05/Dual-Use-Benchmark-Early-Red-Team-Often.pdf>

BARRY LYNN, M. VON T.; MONTOYA, K. **AI in the Public Interest: Confronting the Monopoly Threat**. Open Markets Institute, 2023. Disponível em: <https://www.openmarketsinstitute.org/publications/report-ai-in-the-public-interest-confronting-the-monopoly-threat>

BELINKOV, Y.; GEHRMANN, S.; PAVLICK, E. **Interpretability and Analysis in Neural NLP**. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts. **Anais**...Online: Association for Computational Linguistics, 2020. Disponível em: <https://www.aclweb.org/anthology/2020.acl-tutorials.1>

BELINKOV, Y.; GLASS, J. Analysis Methods in Neural Language Processing: A Survey. **Transactions of the Association for Computational Linguistics**, v. 7, p. 49–72, 2019.

BELTAGY, I.; PETERS, M. E.; COHAN, A. Longformer: The Long-Document Transformer. **CoRR**, v. abs/2004.05150, 2020.

BELZ, A. et al. **A Systematic Review of Reproducibility Research in Natural Language Processing**. Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume. **Anais**...Online: Association for Computational Linguistics, abr. 2021. Disponível em: <https://aclanthology.org/2021.eacl-main.29>

BENCKE, L. et al. **Can we trust LLMs as relevance judges?** Anais do XXXIX Simpósio Brasileiro de Bancos de Dados. **Anais**...Porto Alegre, RS, Brasil: SBC, 2024. Disponível em: <https://sol.sbc.org.br/index.php/sbbd/article/view/30724>

BENDER, E. M. On achieving and evaluating language-independence in NLP. **Linguistic Issues in Language Technology**, v. 6, 2011.

BENDER, E. M. et al. **On the Dangers of Stochastic Parrots: Can Language Models Be**

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol2/

Too Big? . Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency. **Anais**...: FAccT '21.New York, NY, USA: Association for Computing Machinery, 2021. Disponível em: <https://doi.org/10.1145/3442188.3445922>

BENDER, E. M. Resisting Dehumanization in the Age of "AI". **Current Directions in Psychological Science**, v. 33, n. 2, p. 114–120, abr. 2024.

BENDER, E. M.; FRIEDMAN, B. Data Statements for Natural Language Processing: Toward Mitigating System Bias and Enabling Better Science. **Transactions of the Association for Computational Linguistics**, v. 6, p. 587–604, 2018.

BENDER, E. M.; HANNA, A. **The AI Con: How to fight big tech's hype and create the future we want**. [s.l.] Random House, 2025.

BENDER, E. M.; KOLLER, A. **Climbing towards NLU: On Meaning, Form, and Understanding in the Age of Data**. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. **Anais**...Online: Association for Computational Linguistics, jul. 2020. Disponível em: <https://aclanthology.org/2020.acl-main.463>

BENGIO, Y. et al. A Neural Probabilistic Language Model. **J. Mach. Learn. Res.**, v. 3, n. null, p. 1137–1155, mar. 2003.

BENOTTI, L.; BLACKBURN, P. **Ethics consideration sections in natural language processing papers**. Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing. **Anais**...Abu Dhabi, United Arab Emirates: Association for Computational Linguistics, 2022. Disponível em: <https://aclanthology.org/2022.emnlp-main.299>

BERTSCH, A. et al. Unlimiformer: Long-Range Transformers with Unlimited Length Input. **CoRR**, v. abs/2305.01625, 2023.

BERWICK, R. C.; CHOMSKY, N. **Por que apenas nós? Linguagem e evolução**. [s.l.] SciELO-Editora UNESP, 2017.

BIANCHI, F.; HOVY, D. **On the Gap between Adoption and Understanding in NLP**. Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021. **Anais**...Online: Association for Computational Linguistics, ago. 2021. Disponível em: <https://aclanthology.org/2021.findings-acl.340>

BIBAL, A. et al. **Is Attention Explanation? An Introduction to the Debate**. Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). **Anais**...Dublin, Ireland: Association for Computational Linguistics, 2022. Disponível em: <https://aclanthology.org/2022.acl-long.269>

BIBER, D. **Variation across Speech and Writing**. Cambridge: Cambridge University Press, 1988.

BIBER, D.; CONRAD, S. **Register, Genre, and Style**. Cambridge: Cambridge University Press, 2009.

BIBER, D.; FINEGAN, E. **Sociolinguistic Perspectives On Register**. [s.l.] Oxford University Press, 1994.

BINZ, M. et al. How should the advancement of large language models affect the practice of science? **Proceedings of the National Academy of Sciences**, v. 122, n. 5, p. e2401227121, fev. 2025.

BIRD, S. **Decolonising Speech and Language Technology**. Proceedings of the 28th International Conference on Computational Linguistics. **Anais**...Barcelona, Spain (Online): International Committee on Computational Linguistics, dez. 2020. Disponível em: <https://aclanthology.org/2020.coling-main.313>

BIRD, S. **Local Languages, Third Spaces, and other High-Resource Scenarios**. Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). **Anais**...Dublin, Ireland: Association for Computational Linguistics, 2022. Disponível em: <https://acla

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol2/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

nthology.org/2022.acl-long.539>

BLODGETT, S. L. et al. **Language (Technology) is Power: A Critical Survey of "Bias" in NLP**. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. **Anais**...Online: Association for Computational Linguistics, a2020. Disponível em: <https://www.aclweb.org/anthology/2020.acl-main.485>

BLODGETT, S. L. et al. **Language (Technology) is Power: A Critical Survey of "Bias" in NLP**. (D. Jurafsky et al., Eds.)Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. **Anais**...Online: Association for Computational Linguistics, jul. b2020. Disponível em: <https://aclanthology.org/2020.acl-main.485/>

BLOM, J. D. **A dictionary of hallucinations**. [s.l.] Springer, 2010.

BOSTROM, N. **Superintelligence: Paths, Dangers, Strategies**. Oxford, UK: Oxford University Press, 2014.

BOWMAN, S. R.; DAHL, G. **What Will it Take to Fix Benchmarking in Natural Language Understanding?** (K. Toutanova et al., Eds.)Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. **Anais**...Online: Association for Computational Linguistics, jun. 2021. Disponível em: <https://aclanthology.org/2021.naacl-main.385/>

BRANDES, N. et al. ProteinBERT: a universal deep-learning model of protein sequence and function. **Bioinform.**, v. 38, n. 8, p. 2102–2110, 2022.

BRENNAN, K.; KAK, A.; WEST, S. M. The AGI Mythology: The Argument to End All Arguments. Em: **Artificial Power: 2025 Landscape Report**. [s.l.] AI Now Institute, 2025.

BRENNEIS, A. Assessing dual use risks in AI research: necessity, challenges and mitigation strategies. **Research Ethics**, v. 21, n. 2, p. 302–330, 2025.

BROUSSARD, M. **More Than a Glitch: Confronting Race, Gender, and Ability Bias in Tech**. 1. ed. Cambridge, Massachusetts: MIT Press, 2023.

BROWN, T. B. et al. **Language Models are Few-Shot Learners**. (H. Larochelle et al., Eds.)Advances in Neural Information Processing Systems. **Anais**...Curran Associates, Inc., 2020. Disponível em: <https://proceedings.neurips.cc/paper/2020/hash/1457c0d6bfcb4967418bfb8ac142f64a-Abstract.html>

BRUM, H.; NUNES, M. DAS G. V. **Building a Sentiment Corpus of Tweets in Brazilian Portuguese**. (N. C. (Conference chair) et al., Eds.)Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018). **Anais**...Miyazaki, Japan: European Language Resources Association (ELRA), mar. 2018.

BURCHELL, L. et al. **An Expanded Massive Multilingual Dataset for High-Performance Language Technologies (HPLT)**. (W. Che et al., Eds.)Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). **Anais**...Vienna, Austria: Association for Computational Linguistics, jul. 2025. Disponível em: <https://aclanthology.org/2025.acl-long.854/>

CARLSMITH, J. **Is Power-Seeking AI an Existential Risk?** [s.l.] Open Philanthropy, 2022.

CARMO, D. et al. PTT5: Pretraining and validating the T5 model on Brazilian Portuguese data. **CoRR**, v. abs/2008.09144, 2020.

CASTILHO, S. et al. **DELA Corpus - A Document-Level Corpus Annotated with Context-Related Issues**. Proceedings of the Sixth Conference on Machine Translation. **Anais**...Online: Association for Computational Linguistics, nov. 2021. Disponível em: <https://aclanthology.org/2021.wmt-1.63>

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol2/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

CENTRE FOR INFORMATION POLICY LEADERSHIP, C. **The Limitations of Consent as a Legal Basis for Data Processing in the Digital Society**. Washington DC, London, BrusselsCentre for Information Policy Leadership & Hunton Andrews Kurth LLP; Bae Kim & Lee, 2024. Disponível em: <https://www.informationpolicycentre.com/uploads/5/7/1/0/57104281/cipl_bkl_limitations_of_consent_legal_basis_data_processing_dec24.pdf>

CERVANTES, M. DE. **D. Quixote de La Mancha — Primeira Parte**. [s.l.] eBooksBrasil, 1605.

CHALAMALASETTI, K. et al. **clembench: Using Game Play to Evaluate Chat-Optimized Language Models as Conversational Agents**. (H. Bouamor, J. Pino, K. Bali, Eds.)Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. **Anais**...Singapore: Association for Computational Linguistics, dez. 2023. Disponível em: <https://aclanthology.org/2023.emnlp-main.689>

CHANDRA, M. et al. **From Lived Experience to Insight: Unpacking the Psychological Risks of Using AI Conversational Agents**. Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency. **Anais**...Athens Greece: ACM, jun. 2025. Disponível em: <https://dl.acm.org/doi/10.1145/3715275.3732063>

CHEN, G. H. et al. Humans or LLMs as the Judge? A Study on Judgement Bias. **Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing**, a2024.

CHEN, K. et al. **How Susceptible are Large Language Models to Ideological Manipulation?** Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing. **Anais**...Miami, Florida, USA: Association for Computational Linguistics, b2024. Disponível em: <https://aclanthology.org/2024.emnlp-main.952>

CHEN, T. et al. **CopyBench: Measuring Literal and Non-Literal Reproduction of Copyright-Protected Text in Language Model Generation**. (Y. Al-Onaizan, M. Bansal, Y.-N. Chen, Eds.)Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing. **Anais**...Miami, Florida, USA: Association for Computational Linguistics, nov. c2024. Disponível em: <https://aclanthology.org/2024.emnlp-main.844/>

CHENG, J. et al. **Dated Data: Tracing Knowledge Cutoffs in Large Language Models**., 2024. Disponível em: <https://arxiv.org/abs/2403.12958>

CHILD, R. et al. Generating Long Sequences with Sparse Transformers. **CoRR**, v. abs/1904.10509, 2019.

CHO, K. et al. **Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation**. (A. Moschitti, B. Pang, W. Daelemans, Eds.)Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL. **Anais**...ACL, 2014. Disponível em: <https://doi.org/10.3115/v1/d14-1179>

CHOMSKY, N. **Aspects of the Theory of Syntax**. Cambridge, MA: MIT Press, 1965.

CHOWDHERY, A. et al. PaLM: Scaling Language Modeling with Pathways. **CoRR**, v. abs/2204.02311, 2022.

CHRISTIANO, P. F. et al. **Deep Reinforcement Learning from Human Preferences**. (I. Guyon et al., Eds.)Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA. **Anais**...2017. Disponível em: <https://proceedings.neurips.cc/paper/2017/hash/d5e2c0adad503c91f91df240d0cd4e49-Abstract.html>

CLARK, K. et al. **ELECTRA: Pre-training Text Encoders as Discriminators Rather Than Generators**. 8th International Conference on Learning Representations, ICLR 2020. **Anais**...Addis Ababa, Ethiopia: OpenReview.net, abr. 2020. Disponível em: <https://openreview.net/forum?id=r1xMH1BtvB>

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol2/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

COHEN, K. B. et al. **Three Dimensions of Reproducibility in Natural Language Processing**. Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018). **Anais**...Miyazaki, Japan: European Language Resources Association (ELRA), 2018. Disponível em: <https://aclanthology.org/L18-1025>

COLLOBERT, R.; WESTON, J. **A unified architecture for natural language processing: deep neural networks with multitask learning**. (W. W. Cohen, A. McCallum, S. T. Roweis, Eds.)Machine Learning, Proceedings of the Twenty-Fifth International Conference (ICML 2008), Helsinki, Finland, June 5-9, 2008. **Anais**...: ACM International Conference Proceeding Series.ACM, 2008. Disponível em: <https://doi.org/10.1145/1390156.1390177>

COMMON CRAWL. **Common Crawl – Open Repository of Web Crawl Data**. https://commoncrawl.org, [s.d.].

CONNEAU, A.; LAMPLE, G. Cross-Lingual Language Model Pretraining. Em: **Proceedings of the 33rd International Conference on Neural Information Processing Systems**. Red Hook, NY, USA: Curran Associates Inc., 2019.

CORMACK, G. V.; CLARKE, C. L.; BUETTCHER, S. **Reciprocal rank fusion outperforms condorcet and individual rank learning methods**. Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval. **Anais**...2009.

CORMEN, T. et al. Introduction to Algorithms. Em: 2. ed. [s.l.] MIT Press; McGraw-Hill, 2001.

CORRÊA, N. K. et al. Tucano: Advancing Neural Text Generation for Portuguese. **Patterns**, 2025.

CORTES, C.; VAPNIK, V. Support-Vector Networks. **Machine Learning**, v. 20, n. 3, p. 273–297, set. 1995.

CORTIZ, D. et al. **A Weakly Supervised Dataset of Fine-Grained Emotions in Portuguese**. Anais do XIII Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana. **Anais**...Porto Alegre, RS, Brasil: SBC, 2021. Disponível em: <https://sol.sbc.org.br/index.php/stil/article/view/17786>

COULDRY, N.; MEJIAS, U. A. **The Costs of Connection: How Data Is Colonizing Human Life and Appropriating It for Capitalism**. [s.l.] Stanford University Press, 2019.

CRESPO, M. C. R. M. et al. **Carolina: a General Corpus of Contemporary Brazilian Portuguese with Provenance, Typology and Versioning Information**., 2023. Disponível em: <https://arxiv.org/abs/2303.16098>

CRONBACH, L. J. Studies of acquiescence as a factor in the true–false test. **Journal of Educational Psychology**, v. 33, p. 401–415, 1942.

CUEVA, R. et al. **Inteligência Artificial e Desafios Regulatórios**. [s.l.] Forense, 2026.

CURRY, A.; CERCAS CURRY, A. **Computer says "No": The Case Against Empathetic Conversational AI**. (A. Rogers, J. Boyd-Graber, N. Okazaki, Eds.)Findings of the Association for Computational Linguistics: ACL 2023. **Anais**...Toronto, Canada: Association for Computational Linguistics, jul. 2023. Disponível em: <https://aclanthology.org/2023.findings-acl.515/>

D'ARCY, A.; BENDER, E. M. Ethics in Linguistics. **Annual Review of Linguistics**, v. 9, n. 1, p. 49–69, jan. 2023.

DAI, Z. et al. **Transformer-XL: Attentive Language Models beyond a Fixed-Length Context**. (A. Korhonen, D. R. Traum, L. Màrquez, Eds.)Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers. **Anais**...Association for Computational Linguistics, 2019. Disponível em: <https://doi.org/10.18653/v1/p19-1285>

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol2/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

DARCIO, L. et al. **LattesRex: Building ChatBots for Semi-Structured Documents**. Proceedings of the 14th Brazilian Symposium in Information and Human Language Technology (STIL 2025). **Anais**...Brazil: 2025. Disponível em: <https://github.com/Lucas-Darcio/LattesRex>

DAVIES, M.; FERREIRA, M. J. **Corpus do Português: Web/Dialects**. https://www.corpusdoportugues.org/web-dial/, 2016.

DAVIES, M.; FERREIRA, M. J. **Corpus do Português: NOW**. https://www.corpusdoportugues.org/now/, 2018.

DETTMERS, T. et al. QLoRA: Efficient Finetuning of Quantized LLMs. **arXiv preprint arXiv:2305.14314**, 2023.

DEVLIN, J. et al. **BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding**. (J. Burstein, C. Doran, T. Solorio, Eds.)Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019. **Anais**...Minneapolis, MN, USA: Association for Computational Linguistics, 2019. Disponível em: <https://doi.org/10.18653/v1/n19-1423>

DIERK, C.; HEALEY, J.; DOGAN, M. D. **Evaluating LLMs in Experiential Context: Insights from a Survey of Recent CHI Publications**. Human-centered Evaluation and Auditing of Language Models Workshop (HEAL), CHI '25. **Anais**...Yokohama, Japan: ACM, 2025. Disponível em: <https://heal-workshop.github.io/chi2025_papers/43_Evaluating_LLMs_in_Experien.pdf>

DING, N. et al. Parameter-efficient fine-tuning of large-scale pre-trained language models. **Nature machine intelligence**, v. 5, n. 3, p. 220–235, 2023.

DONG, Q. et al. A Survey for In-context Learning. **CoRR**, v. abs/2301.00234, 2023.

DOU, G. et al. **Avoiding Copyright Infringement via Large Language Model Unlearning**. (L. Chiruzzo, A. Ritter, L. Wang, Eds.)Findings of the Association for Computational Linguistics: NAACL 2025. **Anais**...Albuquerque, New Mexico: Association for Computational Linguistics, abr. 2025. Disponível em: <https://aclanthology.org/2025.findings-naacl.288/>

EDITORIAL. Emotional risks of AI companions demand attention. **Nature Machine Intelligence**, v. 7, n. 7, p. 981–982, jul. 2025.

ES, S. et al. **Ragas: Automated evaluation of retrieval augmented generation**. Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations. **Anais**...2024.

ETHAYARAJH, K.; JURAFSKY, D. **Utility is in the Eye of the User: A Critique of NLP Leaderboards**. Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). **Anais**...Online: Association for Computational Linguistics, 2020. Disponível em: <https://www.aclweb.org/anthology/2020.emnlp-main.393>

FAN, A.; LEWIS, M.; DAUPHIN, Y. **Hierarchical Neural Story Generation**. Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). **Anais**...Melbourne, Australia: Association for Computational Linguistics, jul. 2018. Disponível em: <https://aclanthology.org/P18-1082>

FÄRBER, F. et al. **IaraMed: A Women's Healthcare Chatbot for Portuguese Speakers**. Anais do XXV Simpósio Brasileiro de Computação Aplicada à Saúde. **Anais**...Porto Alegre, RS, Brasil: SBC, 2025. Disponível em: <https://sol.sbc.org.br/index.php/sbcas/article/view/35558>

FEIJÓ, D. DE V.; MOREIRA, V. P. Mono vs Multilingual Transformer-based Models: a Comparison across Several Language Tasks. **CoRR**, v. abs/2007.09757, 2020.

FINGER, M. et al. **Carolina: General Corpus of Contemporary Brazilian Portuguese with**

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol2/

Provenance and Typology Information., 2020. Disponível em: <https://c4ai.inova.usp.br/portal/carolina>

FIRTH, J. R. **A synopsis of linguistic theory 1930–1955**. [s.l.] Blackwell, 1957. p. 1–32

FISHER, J. et al. **Biased LLMs can Influence Political Decision-Making**. (W. Che et al., Eds.)Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). **Anais**...Vienna, Austria: Association for Computational Linguistics, jul. 2025. Disponível em: <https://aclanthology.org/2025.acl-long.328/>

FLORIDI, L. Translating Principles into Practices of Digital Ethics: Five Risks of Being Unethical. **Philosophy & Technology**, v. 32, n. 2, p. 185–193, jun. 2019.

FOKKENS, A. et al. **Offspring from Reproduction Problems: What Replication Failure Teaches Us**. Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). **Anais**...Sofia, Bulgaria: Association for Computational Linguistics, ago. 2013. Disponível em: <https://aclanthology.org/P13-1166>

FORT, K.; ADDA, G.; COHEN, K. B. Amazon Mechanical Turk: Gold Mine or Coal Mine? **Computational Linguistics**, v. 37, n. 2, p. 413–420, jun. 2011.

FRANCISCO, P. A. P.; HUREL, L. M.; RIELLI, M. M. **Regulação do Reconhecimento Facial no Setor Público: Avaliação de Experiências Internacionais**. Instituto Igarapé - DataPrivacyBR, 2020. Disponível em: <https://www.dataprivacybr.org/wp-content/uploads/2021/11/regulacao-do-reconhecimento-facial-no-setor-publico.pdf>

FYFE, S. et al. Apophenia, theory of mind and schizotypy: perceiving meaning and intentionality in randomness. **Cortex**, v. 44, n. 10, p. 1316–1325, 2008.

GABRIEL, I. Artificial Intelligence, Values and Alignment. **Minds and Machines**, v. 30, n. 3, p. 411–437, 2020.

GAJULAMANDYAM, D. K. et al. **Domain Specific Finetuning of LLMs Using PEFT Techniques**. 2025 IEEE 15th Annual Computing and Communication Workshop and Conference (CCWC). **Anais**...2025.

GALILEO AI. **Generative AI and LLM Insights**. https://galileo.ai/blog/, 2024.

GAO, M. et al. **LLM-based NLG Evaluation: Current Status and Challenges**. ArXiv, 2024. Disponível em: <https://arxiv.org/abs/2402.01383>

GAO, T.; YAO, X.; CHEN, D. **SimCSE: Simple Contrastive Learning of Sentence Embeddings**. (M.-F. Moens et al., Eds.)Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021. **Anais**...Association for Computational Linguistics, 2021. Disponível em: <https://doi.org/10.18653/v1/2021.emnlp-main.552>

GARCIA, E. A. S. et al. **RoBERTaLexPT: A Legal RoBERTa Model pretrained with deduplication for Portuguese**. (P. Gamallo et al., Eds.)Proceedings of the 16th International Conference on Computational Processing of Portuguese - Vol. 1. **Anais**...Santiago de Compostela, Galicia/Spain: Association for Computational Lingustics, mar. 2024. Disponível em: <https://aclanthology.org/2024.propor-1.38>

GATT, A.; KRAHMER, E. Survey of the State of the Art in Natural Language Generation: Core tasks, applications and evaluation. **Journal of Artificial Intelligence Research**, v. 61, n. 1, p. 65–170, 2018.

GEBRU, T. et al. Datasheets for datasets. **Communications of the ACM**, v. 64, n. 12, p. 86–92, dez. 2021.

GEBRU, T.; BENDER, E. M.; MCMILLAN-MAJOR, A. **Statement from the listed authors of**

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol2/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Stochastic Parrots on the "AI pause" letter., 2023. Disponível em: <https://www.dair-institute.org/blog/letter-statement-March2023/>

GEBRU, T.; TORRES, E. P. The TESCREAL bundle: Eugenics and the promise of utopia through artificial general intelligence. **First Monday**, abr. 2024.

GEHMAN, S. et al. **RealToxicityPrompts: Evaluating Neural Toxic Degeneration in Language Models**. (T. Cohn, Y. He, Y. Liu, Eds.)Findings of the Association for Computational Linguistics: EMNLP 2020. **Anais**...Online: Association for Computational Linguistics, nov. 2020. Disponível em: <https://aclanthology.org/2020.findings-emnlp.301/>

GEVA, M.; GUPTA, A.; BERANT, J. **Injecting Numerical Reasoning Skills into Language Models**. (D. Jurafsky et al., Eds.)Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020. **Anais**...Association for Computational Linguistics, 2020. Disponível em: <https://doi.org/10.18653/v1/2020.acl-main.89>

GOLUMBIA, D. **Cyberlibertarianism: The Right-Wing Politics of Digital Technology**. [s.l.] University of Minnesota Press, 2024.

GONG, Z. et al. **Continual Pre-training of Language Models for Math Problem Understanding with Syntax-Aware Memory Network**. (S. Muresan, P. Nakov, A. Villavicencio, Eds.)Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022. **Anais**...Association for Computational Linguistics, 2022. Disponível em: <https://doi.org/10.18653/v1/2022.acl-long.408>

GONZALEZ-CABELLO, M. et al. Fairness in crowdwork: Making the human AI supply chain more humane. **Business Horizons**, v. 68, n. 5, p. 645–657, 2025.

GOODFELLOW, I.; BENGIO, Y.; COURVILLE, A. **Deep Learning**. [s.l.] MIT Press, 2016. v. 1

GOOGLE CLOUD. **Best Practices with Large Language Models**. https://cloud.google.com/vertex-ai/generative-ai/docs/learn/prompt-best-practices?hl=en, 2025.

GRAY, M. L.; SURI, S. **Ghost work: How to stop Silicon Valley from building a new global underclass**. [s.l.] Harper Business, 2019.

GRAY WIDDER, D.; WEST, S.; WHITTAKER, M. Open (For Business): Big Tech, Concentrated Power, and the Political Economy of Open AI. **SSRN Electronic Journal**, 2023.

GU, Y. et al. Domain-Specific Language Model Pretraining for Biomedical Natural Language Processing. **ACM Trans. Comput. Healthcare**, v. 3, n. 1, out. 2021.

GURURANGAN, S. et al. **Don't Stop Pretraining: Adapt Language Models to Domains and Tasks**. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. **Anais**...Online: Association for Computational Linguistics, jul. 2020. Disponível em: <https://aclanthology.org/2020.acl-main.740>

HADDAD, I. **Artificial Intelligence and Data in Open Source**. Linux Foundation, 2022. Disponível em: <https://www.linuxfoundation.org/hubfs/LF%20Research/Artificial%20Intelligence%20and%20Data%20in%20Open%20Source%20-%20Report.pdf?hsLang=en>

HANNIGAN, T. R.; MCCARTHY, I. P.; SPICER, A. Beware of botshit: How to manage the epistemic risks of generative chatbots. **Business Horizons**, v. 67, n. 5, p. 471–486, 2024.

HAVENS, L. et al. **Situated Data, Situated Systems: A Methodology to Engage with Power Relations in Natural Language Processing Research**. (M. R. Costa-jussà et al., Eds.)Proceedings of the Second Workshop on Gender Bias in Natural Language Processing. **Anais**...Barcelona, Spain (Online): Association for Computational Linguistics, dez. 2020. Disponível em: <https://aclanthology.org/2020.gebnlp-1.10/>

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol2/

HE, K. et al. **Deep Residual Learning for Image Recognition**. 2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016. **Anais**...IEEE Computer Society, 2016. Disponível em: <https://doi.org/10.1109/CVPR.2016.90>

HE, P. et al. **Deberta: decoding-Enhanced Bert with Disentangled Attention**. 9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021. **Anais**...OpenReview.net, 2021. Disponível em: <https://openreview.net/forum?id=XPZIaotutsD>

HEDDERICH, M. A.; OULASVIRTA, A. Explaining crowdworker behaviour through computational rationality. **Behaviour & Information Technology**, v. 44, n. 3, p. 552–573, 2025.

HELLER, J. **Legal AI benchmarking: CoCounsel – from code to courtroom: The meticulous testing of CoCounsel's professional-grade AI**., 23 out. 2024. Disponível em: <https://www.thomsonreuters.com/en-us/posts/innovation/legal-ai-benchmarking-cocounsel/>. Acesso em: 20 ago. 2025

HENDREN, S. All technology is assistive: Six design rules on disability. Em: SAYERS, J. (Ed.). **Making Things and Drawing Boundaries: Experiments in the Digital Humanities**. [s.l.] University of Minnesota Press, Minneapolis, MN, 2014.

HENDRYCKS, D. et al. Unsolved Problems in ML Safety. **arXiv preprint arXiv:2109.13916**, 2021.

HERSHCOVICH, D. et al. **Towards Climate Awareness in NLP Research**. Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing. **Anais**...Abu Dhabi, United Arab Emirates: Association for Computational Linguistics, 2022. Disponível em: <https://aclanthology.org/2022.emnlp-main.159>

HICKS, M. T.; HUMPHRIES, J.; SLATER, J. ChatGPT is bullshit. **Ethics and Information Technology**, v. 26, n. 2, p. 38, jun. 2024.

HOCHREITER, S. Untersuchungen zu dynamischen neuronalen Netzen. **Diploma, Technische Universität München**, v. 91, n. 1, p. 31, 1991.

HOCHREITER, S.; SCHMIDHUBER, J. Long Short-Term Memory. **Neural Computation**, v. 9, n. 8, p. 1735–1780, nov. 1997.

HOFFMANN, J. et al. Training Compute-Optimal Large Language Models. **CoRR**, v. abs/2203.15556, 2022.

HOFMANN, V. et al. AI generates covertly racist decisions about people based on their dialect. **Nature**, v. 633, n. 8028, p. 147–154, set. 2024.

HOLTZMAN, A. et al. **The Curious Case of Neural Text Degeneration.** ICLR. **Anais**...OpenReview.net, 2020. Disponível em: <http://dblp.uni-trier.de/db/conf/iclr/iclr2020.html#HoltzmanBDFC20>

HOOKER, S. Moving beyond "algorithmic bias is a data problem". **Patterns**, v. 2, n. 4, p. 100241, abr. 2021.

HORNIK, K.; STINCHCOMBE, M. B.; WHITE, H. Multilayer feedforward networks are universal approximators. **Neural Networks**, v. 2, n. 5, p. 359–366, 1989.

HOUAMEGNI, L. R. P.; GEDIKLI, F. **Evaluating the Effectiveness of Large Language Models in Automated News Article Summarization**., 2025. Disponível em: <https://arxiv.org/abs/2502.17136>

HOULSBY, N. et al. **Parameter-Efficient Transfer Learning for NLP**. (K. Chaudhuri, R. Salakhutdinov, Eds.)Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA. **Anais**...: Proceedings of Machine Learning Research.PMLR, 2019. Disponível em: <http://proceedings.mlr.press/v97/houlsby19a.html>

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol2/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

HOVY, D.; SPRUIT, S. L. **The Social Impact of Natural Language Processing**. Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). **Anais**...Berlin, Germany: Association for Computational Linguistics, 2016. Disponível em: <http://aclweb.org/anthology/P16-2096>

HOWARD, J.; RUDER, S. **Universal Language Model Fine-tuning for Text Classification**. Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). **Anais**...Melbourne, Australia: Association for Computational Linguistics, jul. a2018. Disponível em: <ˆ5ˆ>

HOWARD, J.; RUDER, S. **Universal Language Model Fine-tuning for Text Classification**. (I. Gurevych, Y. Miyao, Eds.)Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). **Anais**...Melbourne, Australia: Association for Computational Linguistics, jul. b2018. Disponível em: <https://aclanthology.org/P18-1031/>. Acesso em: 25 jul. 2025

HU, E. J. et al. **LoRA: Low-Rank Adaptation of Large Language Models**., 2021. Disponível em: <https://arxiv.org/abs/2106.09685>

HU, E. J. et al. **LoRA: Low-Rank Adaptation of Large Language Models**. The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022. **Anais**...OpenReview.net, 2022. Disponível em: <https://openreview.net/forum?id=nZeVKeeFYf9>

HUANG, J.; YANG, D.; POTTS, C. **Demystifying Verbatim Memorization in Large Language Models**. (Y. Al-Onaizan, M. Bansal, Y.-N. Chen, Eds.)Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing. **Anais**...Miami, Florida, USA: Association for Computational Linguistics, nov. a2024. Disponível em: <https://aclanthology.org/2024.emnlp-main.598/>

HUANG, L. et al. A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions. **ACM Transactions on Information Systems**, v. 43, n. 2, p. 1–58, b2024.

II, S. M. W. **GSM8K Benchmark**. Klu; https://klu.ai/glossary/GSM8K-eval, 2025.

IP, J.; VONGTHONGSRI, K. **deepeval**., ago. 2025. Disponível em: <https://github.com/confident-ai/deepeval>

IVGI, M.; SHAHAM, U.; BERANT, J. Efficient Long-Text Understanding with Short-Text Models. **Transactions of the Association for Computational Linguistics**, v. 11, p. 284–299, 2023.

JACKSON, P.; MOULINIER, I. **Natural Language Processing for Online Applications – Text retrieval, extraction and categorization**. [s.l.] John Benjamins, 2002.

JAIN, S.; WALLACE, B. C. **Attention is not Explanation**. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). **Anais**...Minneapolis, Minnesota: Association for Computational Linguistics, 2019. Disponível em: <https://aclanthology.org/N19-1357>

JAKUBČEK, M. et al. **The TenTen corpus family**. Proceedings of the 7th International Corpus Linguistics Conference (CL). **Anais**...2013.

JAN, E. et al. **Multitask-Bench: Unveiling and Mitigating Safety Gaps in LLMs Fine-tuning**. (O. Rambow et al., Eds.)Proceedings of the 31st International Conference on Computational Linguistics. **Anais**...Abu Dhabi, UAE: Association for Computational Linguistics, jan. 2025. Disponível em: <https://aclanthology.org/2025.coling-main.606/>

JAPKOWICZ, N.; BOUKOUVALAS, Z. **Machine Learning Evaluation: Towards Reliable and Responsible AI**. [s.l.] Cambridge University Press, 2024.

JENSEN, K. N.; PLANK, B. **Fine-tuning vs From Scratch: Do Vision & Language Models Have Similar Capabilities on Out-of-Distribution Visual Question Answering?** (N. Calzolari et al.,

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol2/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Eds.)Proceedings of the Thirteenth Language Resources and Evaluation Conference. **Anais**...Marseille, France: European Language Resources Association, jun. 2022. Disponível em: <https://aclanthology.org/2022.lrec-1.161/>

JEONG, S. et al. Adaptive-rag: Learning to adapt retrieval-augmented large language models through question complexity. **arXiv preprint arXiv:2403.14403**, 2024.

JI, Z. et al. Survey of Hallucination in Natural Language Generation. **ACM Comput. Surv.**, v. 55, n. 12, mar. 2023.

JIANG, L.; WAGNER, C. **How Low is Low? Crowdworker Perceptions of Microtask Payments in Work versus Leisure Situations**. Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems. **Anais**...: CHI '24.New York, NY, USA: Association for Computing Machinery, 2024. Disponível em: <https://doi.org/10.1145/3613904.3642601>

JIAO, J. et al. Navigating LLM Ethics: Advancements, Challenges, and Future Directions. **ArXiv**, v. abs/2406.18841, 2024.

JIN, X. et al. **Lifelong Pretraining: Continually Adapting Language Models to Emerging Corpora**. Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. **Anais**...Seattle, United States: Association for Computational Linguistics, jul. 2022. Disponível em: <https://aclanthology.org/2022.naacl-main.351>

JIN, Z. et al. **How Good Is NLP? A Sober Look at NLP Tasks through the Lens of Social Impact**. Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021. **Anais**...Online: Association for Computational Linguistics, 2021. Disponível em: <https://aclanthology.org/2021.findings-acl.273>

JOHNSTON, S. F. Alvin Weinberg and the Promotion of the Technological Fix. **Technology and Culture**, v. 59, n. 3, p. 620–651, 2018.

JOSHI, M. et al. **TriviaQA: A Large Scale Distantly Supervised Challenge Dataset for Reading Comprehension**. (R. Barzilay, M.-Y. Kan, Eds.)Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers. **Anais**...Association for Computational Linguistics, 2017. Disponível em: <https://doi.org/10.18653/v1/P17-1147>

JOSHI, M. et al. SpanBERT: Improving Pre-training by Representing and Predicting Spans. **Transactions of the Association for Computational Linguistics**, v. 8, p. 64–77, 2020.

JOSHI, N.; TALY, A.; MUPPALLA, D. **LLM-Cite: Cheap Fact Verification with Attribution via URL Generation**., 2025. Disponível em: <https://openreview.net/forum?id=qb2QRoE4W3>

JOULIN, A. et al. **Bag of Tricks for Efficient Text Classification**. (M. Lapata, P. Blunsom, A. Koller, Eds.)Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers. **Anais**...Valencia, Spain: Association for Computational Linguistics, abr. 2017. Disponível em: <https://aclanthology.org/E17-2068/>

JOYCE, J. M. Kullback-Leibler Divergence. Em: LOVRIC, M. (Ed.). **International Encyclopedia of Statistical Science**. Berlin, Heidelberg: Springer Berlin Heidelberg, 2011. p. 720–722.

JURAFSKY, D.; MARTIN, J. H. **Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition**. 3rd. ed. USA: Prentice Hall PTR, 2023.

KALAI, A. T. et al. **Why language models hallucinate**. [s.l.] OpenAI, 2025.

KALOULI, A.-L. et al. Curing the SICK and Other NLI Maladies. **Computational Linguistics**, v. 49,

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol2/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

n. 1, p. 199–243, mar. 2023.

KAMRUZZAMAN, M.; SHOVON, MD.; KIM, G. **Investigating Subtler Biases in LLMs: Ageism, Beauty, Institutional, and Nationality Bias in Generative Models**. (L.-W. Ku, A. Martins, V. Srikumar, Eds.)Findings of the Association for Computational Linguistics: ACL 2024. **Anais**...Bangkok, Thailand: Association for Computational Linguistics, ago. 2024. Disponível em: <https://aclanthology.org/2024.findings-acl.530/>

KARAMOLEGKOU, A. et al. **Copyright Violations and Large Language Models**. Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. **Anais**...Singapore: Association for Computational Linguistics, 2023. Disponível em: <https://aclanthology.org/2023.emnlp-main.458>

KE, Z. et al. **Continual Pre-training of Language Models**., 2023. Disponível em: <https://arxiv.org/abs/2302.03241>

KERNER, T. **Domain-Specific Pretraining of Language Models: A Comparative Study in the Medical Field**., 2024. Disponível em: <https://arxiv.org/abs/2407.14076>

KHATTAB, O.; ZAHARIA, M. **ColBERT: Efficient and Effective Passage Search via Contextualized Late Interaction over BERT**., 2020. Disponível em: <https://arxiv.org/abs/2004.12832>

KIANPOUR, M.; WEN, S.-F. **Timing Attacks on Machine Learning: State of the Art**. Intelligent Systems Conference. **Anais**...Springer, 2020.

KIM, D. K. et al. **Analyzing Offensive Language Dataset Insights from Training Dynamics and Human Agreement Level**. (O. Rambow et al., Eds.)Proceedings of the 31st International Conference on Computational Linguistics, COLING 2025, Abu Dhabi, UAE, January 19-24, 2025. **Anais**...Association for Computational Linguistics, 2025. Disponível em: <https://aclanthology.org/2025.coling-main.653/>

KIM, S. et al. Prometheus: Inducing Fine-grained Evaluation Capability in Language Models. **arXiv preprint arXiv:2310.08491**, 2023.

KLENK, M. How Do Technological Artefacts Embody Moral Values? **Philosophy & Technology**, v. 34, n. 3, p. 525–544, set. 2021.

KNUTH, D. E. **Fundamental Algorithms. The Art of Computer Programming**. 3. ed. [s.l.] Addison-Wesley, 1997. v. 1

KOGKALIDIS, K.; CHATZIKYRIAKIDIS, S. **On Tables with Numbers, with Numbers**. (S. Truong et al., Eds.)Proceedings of the 1st Workshop on Language Models for Underserved Communities (LM4UC 2025). **Anais**...Albuquerque, New Mexico: Association for Computational Linguistics, 2025. Disponível em: <https://aclanthology.org/2025.lm4uc-1.12/>

KOJIMA, T. et al. **Large Language Models are Zero-Shot Reasoners**. NeurIPS. **Anais**...2022. Disponível em: <http://papers.nips.cc/paper/_files/paper/2022/hash/8bb0d291acd4acf06ef112099c16f326-Abstract-Conference.html>

KRAKOVNA, V. et al. Specification Gaming: The Flip Side of AI Ingenuity. **DeepMind Safety Research Blog**, 2020.

KREUTZER, J. et al. Quality at a Glance: An Audit of Web-Crawled Multilingual Datasets. **Transactions of the Association for Computational Linguistics**, v. 10, p. 50–72, jan. 2022.

KWIATKOWSKI, T. et al. Natural Questions: a Benchmark for Question Answering Research. **Transactions of the Association of Computational Linguistics**, 2019.

LAI, J. et al. Large language models in law: A survey. **AI Open**, v. 5, p. 181–196, 2024.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol2/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

LAN, Z. et al. **ALBERT: A Lite BERT for Self-supervised Learning of Language Representations**. 8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020. **Anais**...OpenReview.net, 2020. Disponível em: <https://openreview.net/forum?id=H1eA7AEtvS>

LASOTA, L. **Regulating Corporate Behaviour in Digital Ecosystems: Increasing Fairness and Contestability of Digital Markets with Free Software**. MIC 2023: Toward Green, Inclusive, and Digital Growth. **Anais**...a2023.

LASOTA, L. The European Union's AI act from the perspective of Open Technologies. Em: ALEGRE, M.; FÜRST, H. (Eds.). **Advocacia & Bioética: Novas Tecnologias**. https://www.editoracasadodireito.com.br/produto/novas-tecnologias. São Paulo: Casa do Direito, 2023b.

LASOTA, L. The CRA and the Challenges of Regulating Cybersecurity in Open Environments: The Case of Free and Open Source Software. Em: **Digital Decade: How the EU shapes digitalisation**. Berlin: Weizenbaum Institute for the Networked Society - The German Internet Institute, 2025.

LASOTA, L.; SINGHAL, N. **Free Software and AI openness: Overcoming challenges in the licensing world**. [s.l.] Zenodo, abr. 2024. Disponível em: <https://zenodo.org/doi/10.5281/zenodo.10966136>.

LAZZARINI MORETTI, J.; MALTESE ZUFFO, M. LGPD e inteligência artificial: Um estudo comparado. **Revista de Direito Internacional e Globalização Econômica**, v. 13, n. 13, p. 21–42, 2025.

LEE, J. et al. A survey of large language models in finance (finllms). **arXiv preprint arXiv:2402.02315**, 2024.

LEE, K. et al. **Deduplicating Training Data Makes Language Models Better**., 2022. Disponível em: <https://arxiv.org/abs/2107.06499>

LEIDNER, J. L.; PLACHOURAS, V. **Ethical by Design: Ethics Best Practices for Natural Language Processing**. Proceedings of the First ACL Workshop on Ethics in Natural Language Processing. **Anais**...Valencia, Spain: Association for Computational Linguistics, abr. 2017. Disponível em: <https://aclanthology.org/W17-1604>

LEIKE, J. et al. **Scalable Agent Alignment via Reward Modeling: a Research Direction**. arXiv preprint arXiv:1811.07871. **Anais**...2018.

LESTER, B.; AL-RFOU, R.; CONSTANT, N. **The Power of Scale for Parameter-Efficient Prompt Tuning**. (M.-F. Moens et al., Eds.)Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021. **Anais**...Association for Computational Linguistics, 2021. Disponível em: <https://doi.org/10.18653/v1/2021.emnlp-main.243>

LEWIS, M. et al. **BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension**. (D. Jurafsky et al., Eds.)Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020. **Anais**...Association for Computational Linguistics, a2020. Disponível em: <https://doi.org/10.18653/v1/2020.acl-main.703>

LEWIS, P. S. H. et al. **Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks**. (H. Larochelle et al., Eds.)Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual. **Anais**...b2020. Disponível em: <https://proceedings.neurips.cc/paper/2020/hash/6b493230205f780e1bc26945df7481e5-Abstract.html>

LI, D. et al. From Generation to Judgment: Opportunities and Challenges of LLM-as-a-judge. **CoRR**, v. abs/2411.16594, 2024.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol2/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

LI, H. et al. **CaseGen: A benchmark for multi-stage legal case documents generation**., 2025. Disponível em: <https://arxiv.org/abs/2502.17943>

LI, P. et al. Making AI Less "Thirsty": Uncovering and Addressing the Secret Water Footprint of AI Models. **arXiv preprint arXiv:2304.03271**, a2023.

LI, Q. **Parameter Efficient Fine-Tuning on Selective Parameters for Transformer-Based Pre-Trained Models**. 2024 IEEE International Conference on Multimedia and Expo (ICME). **Anais**...2024.

LI, R. et al. StarCoder: may the source be with you! **CoRR**, v. abs/2305.06161, b2023.

LI, W. W. et al. **BERT Is Not The Count: Learning to Match Mathematical Statements with Proofs**. (A. Vlachos, I. Augenstein, Eds.)Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2023, Dubrovnik, Croatia, May 2-6, 2023. **Anais**...Association for Computational Linguistics, c2023. Disponível em: <https://aclanthology.org/2023.eacl-main.260>

LI, X. L.; LIANG, P. **Prefix-Tuning: Optimizing Continuous Prompts for Generation**. (C. Zong et al., Eds.)Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021. **Anais**...Association for Computational Linguistics, 2021. Disponível em: <https://doi.org/10.18653/v1/2021.acl-long.353>

LI, Y. et al. A comparative study of pretrained language models for long clinical text. **Journal of the American Medical Informatics Association**, v. 30, n. 2, p. 340–347, nov. 2022.

LIANG, X. et al. **Contrastive Demonstration Tuning for Pre-trained Language Models**. (Y. Goldberg, Z. Kozareva, Y. Zhang, Eds.)Findings of the Association for Computational Linguistics: EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022. **Anais**...Association for Computational Linguistics, 2022. Disponível em: <https://aclanthology.org/2022.findings-emnlp.56>

LIN, C.-Y. **ROUGE: A Package for Automatic Evaluation of Summaries**. Text Summarization Branches Out. **Anais**...Barcelona, Spain: Association for Computational Linguistics, jul. 2004. Disponível em: <https://aclanthology.org/W04-1013>

LIN, S.; HILTON, J.; EVANS, O. **TruthfulQA: Measuring How Models Mimic Human Falsehoods**. (S. Muresan, P. Nakov, A. Villavicencio, Eds.)Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). **Anais**...Dublin, Ireland: Association for Computational Linguistics, 2022. Disponível em: <https://aclanthology.org/2022.acl-long.229/>

LIRA, T. et al. **Aroeira: A Curated Corpus for the Portuguese Language with a Large Number of Tokens**. Anais da XXXIV Brazilian Conference on Intelligent Systems. **Anais**...Porto Alegre, RS, Brasil: SBC, 2024. Disponível em: <https://sol.sbc.org.br/index.php/bracis/article/view/33560>

LITA, A.-C.; COSTEA, C.-E.; IONESCU, B. **Chunking for RAG: A Comprehensive Evaluation of Chunking Strategies for Retrieval-Augmented Generation**. Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing. **Anais**...2024.

LIU, Y. et al. **RoBERTa: A Robustly Optimized BERT Pretraining Approach**., a2019. Disponível em: <https://arxiv.org/abs/1907.11692>

LIU, Y. et al. Roberta: A robustly optimized bert pretraining approach. **arXiv preprint arXiv:1907.11692**, b2019.

LIU, Y. et al. Multilingual Denoising Pre-training for Neural Machine Translation. **Trans. Assoc. Comput. Linguistics**, v. 8, p. 726–742, 2020.

LIU, Y. et al. **HD-Eval: Aligning Large Language Model Evaluators Through Hierarchical Criteria Decomposition**. (L.-W. Ku, A. Martins, V. Srikumar, Eds.)Proceedings of the 62nd Annual

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol2/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024. **Anais**...Association for Computational Linguistics, 2024. Disponível em: <https://doi.org/10.18653/v1/2024.acl-long.413>

LIU, Z. et al. **A Robustly Optimized BERT Pre-Training Approach with Post-Training**. Chinese Computational Linguistics: 20th China National Conference, CCL 2021, Hohhot, China, August 13–15, 2021, Proceedings. **Anais**...Berlin, Heidelberg: Springer-Verlag, 2021. Disponível em: <https://doi.org/10.1007/978-3-030-84186-7_31>

LOBO, P. **Profiling na Lei Geral de Proteção de Dados: O Livre Desenvolvimento da Personalidade em Face da Governamentalidade Algorítmica**. 1. ed. [s.l.] Editora Foco, 2022.

LONGPRE, S. et al. **A Pretrainer's Guide to Training Data: Measuring the Effects of Data Age, Domain Coverage, Quality, & Toxicity**. (K. Duh, H. Gomez, S. Bethard, Eds.)Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers). **Anais**...Mexico City, Mexico: Association for Computational Linguistics, jun. 2024. Disponível em: <https://aclanthology.org/2024.naacl-long.179/>

LONGPRE, S. et al. **The Responsible Foundation Model Development Cheatsheet: A Review of Tools & Resources**., 2025. Disponível em: <https://arxiv.org/abs/2406.16746>

LUONG, T.; PHAM, H.; MANNING, C. D. **Effective Approaches to Attention-based Neural Machine Translation**. (L. Màrquez et al., Eds.)Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, EMNLP 2015, Lisbon, Portugal, September 17-21, 2015. **Anais**...The Association for Computational Linguistics, 2015. Disponível em: <https://doi.org/10.18653/v1/d15-1166>

LV, K. et al. **Full Parameter Fine-tuning for Large Language Models with Limited Resources**., 2024. Disponível em: <https://arxiv.org/abs/2306.09782>

MALEKI, N.; PADMANABHAN, B.; DUTTA, K. **AI Hallucinations: A Misnomer Worth Clarifying**. 2024 IEEE Conference on Artificial Intelligence (CAI). **Anais**...Singapore, Singapore: IEEE, jun. 2024. Disponível em: <https://ieeexplore.ieee.org/document/10605268/>

MARKOV, A. A. The theory of algorithms. **Trudy Matematicheskogo Instituta Imeni VA Steklova**, v. 42, p. 3–375, 1954.

MATHUR, N.; BALDWIN, T.; COHN, T. **Tangled up in BLEU: Reevaluating the Evaluation of Automatic Machine Translation Evaluation Metrics**. Proceedings of the Conference on Empirical Methods in Natural Language Processing. **Anais**...Association for Computational Linguistics, 2020. Disponível em: <https://aclanthology.org/2020.acl-main.448/>

MATTHEWS, B. W. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. **Biochimica et Biophysica Acta (BBA) - Protein Structure**, v. 405, n. 2, p. 442–451, 1975.

MCCANN, B. et al. **Learned in Translation: Contextualized Word Vectors**. Proceedings of the 31st International Conference on Neural Information Processing Systems. **Anais**...: NIPS'17.Red Hook, NY, USA: Curran Associates Inc., 2017.

MCMILLAN-MAJOR, A.; BENDER, E. M.; FRIEDMAN, B. Data Statements: From Technical Concept to Community Practice. **ACM Journal on Responsible Computing**, p. 3594737, 2023.

MCQUILLAN, D. **Resisting AI: an anti-fascist approach to artificial intelligence**. Bristol, UK: [s.n.].

MECKLENBURG, N. et al. **Injecting New Knowledge into Large Language Models via Supervised Fine-Tuning**., 2024. Disponível em: <https://arxiv.org/abs/2404.00213>

MEJIAS, U. A.; COULDRY, N. Datafication. **Internet Policy Review**, v. 8, n. 4, nov. 2019.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol2/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

MEVA, DR. D.; KUKADIYA, H. Performance Evaluation of Large Language Models: A Comprehensive Review. **International Research Journal of Computer Science**, v. 12, p. 109–114, mar. 2025.

MICELI, M. et al. **Who Trains the Data for European Artificial Intelligence?** The Left, DiPLab, Weizenbaum Institute; DAIR Institute., 2024. Disponível em: <https://hal.science/hal-04662589v1>

MIIKKULAINEN, R.; DYER, M. G. Natural Language Processing With Modular Pdp Networks and Distributed Lexicon. **Cognitive Science**, v. 15, n. 3, p. 343–399, 1991.

MILLER, B. Is Technology Value-Neutral? **Science, Technology, & Human Values**, v. 46, n. 1, p. 53–80, jan. 2021.

MINAEE, S. et al. Large language models: A survey. **arXiv preprint arXiv:2402.06196**, 2024.

MINAEE, S. et al. **Large Language Models: A Survey**., 2025. Disponível em: <https://arxiv.org/abs/2402.06196>

MITCHELL, M. et al. **Model Cards for Model Reporting**. Proceedings of the Conference on Fairness, Accountability, and Transparency. **Anais**...Atlanta GA USA: ACM, jan. 2019. Disponível em: <https://dl.acm.org/doi/10.1145/3287560.3287596>

MIZRAHI, D. et al. **Language Models Improve When Pretraining Data Matches Target Tasks**., 2025. Disponível em: <https://arxiv.org/abs/2507.12466>

MOHAMMAD, S. **Ethics Sheets for AI Tasks**. Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). **Anais**...Dublin, Ireland: Association for Computational Linguistics, 2022. Disponível em: <https://aclanthology.org/2022.acl-long.573>

MOREIRA, J.; VIANNA, D. **Challenges of Generative AI in Legal Practice: A Jusbrasil Perspective**. Proceedings of the 40th Brazilian Symposium on Databases (SBBD 2025). **Anais**...Brazil: Brazilian Computer Society (SBC), 2025. Disponível em: <https://sbbd.org.br/2025/wp-content/uploads/2025/10/SBBD25-Jusbrasil.pdf>

MORESCHI, B.; PEREIRA, G.; COZMAN, F. G. The Brazilian Workers in Amazon Mechanical Turk: Dreams and realities of ghost workers. **Revista Contracampo**, v. 39, n. 1, abr. 2020.

MOROZOV, E. **To save everything, click here : the folly of technological solutionism**. New York, NY: PublicAffairs, 2014.

MUNGER, K. **Chatbots for Good and Evil**. EACL via Underline Science Inc., 2023. Disponível em: <https://underline.io/lecture/72154-chatbots-for-good-and-evil>

NAIR, P. A. et al. **Matryoshka Quantization**. Sparsity in LLMs (SLLM): Deep Dive into Mixture of Experts, Quantization, Hardware, and Inference. **Anais**...2025.

NARAYAN, S.; COHEN, S. B.; LAPATA, M. **Don't Give Me the Details, Just the Summary! Topic-Aware Convolutional Neural Networks for Extreme Summarization**. Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. **Anais**...Brussels, Belgium: 2018.

NATIONAL INSTITUTE OF STANDARDS AND TECHNOLOGY. **A Plan for Global Engagement on AI Standards**. [s.l.] U.S. Department of Commerce, National Institute of Standards; Technology, jul. 2024. Disponível em: <https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-5.pdf>.

NEPOMUCENO, J. et al. **Agentic AI Meets Fundamentalist Financial Analysis: Limits and Possibilities**. Proceedings of the 14th Brazilian Conference on Intelligent Systems (BRACIS 2025). **Anais**...2025.

NGO, R.; CHAN, L.; MINDERMANN, S. The Alignment Problem from a Deep Learning Perspective. **arXiv preprint arXiv:2303.16200**, 2023.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol2/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

NIJKAMP, E. et al. ProGen2: Exploring the Boundaries of Protein Language Models. **CoRR**, v. abs/2206.13517, 2022.

NOVOBILSKÁ, L. Free and Open Source Software Licensing Requirements and Copyright Infringement Involving Artificial Intelligence Technologies. 2023.

O'NEIL, C. **Weapons of math destruction : how big data increases inequality and threatens democracy**. First edition ed. New York: [s.n.].

OLIVEIRA, L. L. Inteligência artificial e desigualdade social: o impacto do colonialismo digital nas políticas públicas. **Internet & Sociedade**, v. 5, n. 1, 2024.

OUYANG, L. et al. **Training language models to follow instructions with human feedback**. (A. H. Oh et al., Eds.)Advances in Neural Information Processing Systems. **Anais**...2022. Disponível em: <https://openreview.net/forum?id=TG8KACxEON>

OVERWIJK, A.; XIONG, C.; CALLAN, J. **ClueWeb22: 10 Billion Web Documents with Rich Information**. (E. Amigó et al., Eds.)SIGIR '22: The 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, Madrid, Spain, July 11 - 15, 2022. **Anais**...ACM, 2022. Disponível em: <https://doi.org/10.1145/3477495.3536321>

PAIOLA, P. H. et al. Adapting llms for the medical domain in portuguese: A study on fine-tuning and model evaluation. **arXiv preprint arXiv:2410.00163**, 2024.

PAN, S. et al. Unifying large language models and knowledge graphs: A roadmap. **IEEE Transactions on Knowledge and Data Engineering**, v. 36, n. 7, p. 3580–3599, 2024.

PAN, Y. et al. **On the Risk of Misinformation Pollution with Large Language Models**. Findings of the Association for Computational Linguistics: EMNLP 2023. **Anais**...Singapore: Association for Computational Linguistics, 2023. Disponível em: <https://aclanthology.org/2023.findings-emnlp.97>

PANICKSSERY, A.; BOWMAN, S. R.; FENG, S. LLM Evaluators Recognize and Favor Their Own Generations. **Proceedings of the 38th International Conference on Neural Information Processing Systems**, 2024.

PAPINENI, K. et al. **BLEU: A Method for Automatic Evaluation of Machine Translation**. Proceedings of the 40th Annual Meeting on Association for Computational Linguistics. **Anais**...: ACL '02.USA: Association for Computational Linguistics, 2002. Disponível em: <https://doi.org/10.3115/1073083.1073135>

PARK, K. et al. **OffsetBias: Leveraging Debiased Data for Tuning Evaluators**., 2024. Disponível em: <https://arxiv.org/abs/2407.06551>

PARMAR, M. et al. **Don't Blame the Annotator: Bias Already Starts in the Annotation Instructions**. Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics. **Anais**...Dubrovnik, Croatia: Association for Computational Linguistics, 2023. Disponível em: <https://aclanthology.org/2023.eacl-main.130>

PASSINATO, E.; RIOS, W.; FILHO, A. G. **Integração de Modelos de Linguagem e RAG na Criação de Chatbots Oftalmológicos**. Anais do XXIV Simpósio Brasileiro de Computação Aplicada à Saúde. **Anais**...Porto Alegre, RS, Brasil: SBC, 2024. Disponível em: <https://sol.sbc.org.br/index.php/sbcas/article/view/28831>

PAULLADA, A. et al. Data and its (dis)contents: A survey of dataset development and use in machine learning research. **Patterns**, v. 2, n. 11, p. 100336, nov. 2021.

PETERS, M. E. et al. **Deep Contextualized Word Representations**. (M. A. Walker, H. Ji, A. Stent, Eds.)Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2018, New Orleans, Louisiana,

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol2/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

USA, June 1-6, 2018, Volume 1 (Long Papers). **Anais**...Association for Computational Linguistics, 2018. Disponível em: <https://doi.org/10.18653/v1/n18-1202>

PEYRARD, M. **Studying Summarization Evaluation Metrics in the Appropriate Scoring Range**. Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP). **Anais**...Online: Association for Computational Linguistics, 2019. Disponível em: <https://aclanthology.org/P19-1502/>

PIĘKOS, P.; MALINOWSKI, M.; MICHALEWSKI, H. **Measuring and Improving BERT's Mathematical Abilities by Predicting the Order of Reasoning.** Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers). **Anais**...Online: Association for Computational Linguistics, ago. 2021. Disponível em: <https://aclanthology.org/2021.acl-short.49>

PIRES, R. et al. **Sabiá: Portuguese Large Language Models**. (M. C. Naldi, R. A. C. Bianchi, Eds.)Intelligent Systems. **Anais**...Cham: Springer Nature Switzerland, 2023.

PLOUG, T. The Right Not to Be Subjected to AI Profiling Based on Publicly Available Data—Privacy and the Exceptionalism of AI Profiling. **Philosophy & Technology**, v. 36, n. 1, p. 14, mar. 2023.

POLO, F. M. et al. Efficient Multi-Prompt Evaluation of LLMs. **Proceedings of the 38th Conference on Neural Information Processing Systems (NeurIPS 2024)**, 2024.

POMBAL, J. et al. M-Prometheus: A Suite of Open Multilingual LLM Judges. **CoRR**, v. abs/2504.04953, 2025.

PROTOTYPEJAM. **Lake Merritt: AI Evaluation Workbench**. https://prototypejam.github.io/lake_merritt/, 2025.

PROVILKOV, I.; EMELIANENKO, D.; VOITA, E. **BPE-Dropout: Simple and Effective Subword Regularization**. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. **Anais**...Online: Association for Computational Linguistics, jul. 2020.

RACKAUCKAS, Z. Rag-fusion: a new take on retrieval-augmented generation. **arXiv preprint arXiv:2402.03367**, 2024.

RADFORD, A. et al. **Improving Language Understanding by Generative Pre-Training**., 2018. Disponível em: <https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf>

RADFORD, A. et al. **Language Models are Unsupervised Multitask Learners**. a2019.

RADFORD, A. et al. Language models are unsupervised multitask learners. **OpenAI blog**, v. 1, n. 8, p. 9, b2019.

RADFORD, A. et al. **Learning Transferable Visual Models From Natural Language Supervision**., 2021. Disponível em: <https://arxiv.org/abs/2103.00020>

RADFORD, A.; NARASIMHAN, K. **Improving Language Understanding by Generative Pre-Training**. 2018.

RAE, J. W. et al. Scaling Language Models: Methods, Analysis & Insights from Training Gopher. **CoRR**, v. abs/2112.11446, 2021.

RAE, J. W. et al. **Scaling Language Models: Methods, Analysis & Insights from Training Gopher**., 2022. Disponível em: <https://arxiv.org/abs/2112.11446>

RAFAILOV, R. et al. **Direct Preference Optimization: Your Language Model is Secretly a Reward Model**. Advances in Neural Information Processing Systems (NeurIPS). **Anais**...2023.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol2/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

RAFFEL, C. et al. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. **Journal of Machine Learning Research**, v. 21, n. 140, p. 1–67, a2020.

RAFFEL, C. et al. Exploring the limits of transfer learning with a unified text-to-text transformer. **J. Mach. Learn. Res.**, v. 21, n. 1, p. 140:5485–140:5551, jan. b2020.

RAFFEL, C. et al. **Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer**., 2023. Disponível em: <https://arxiv.org/abs/1910.10683>

RAJI, D. et al. **AI and the Everything in the Whole Wide World Benchmark**. (J. Vanschoren, S. Yeung, Eds.)Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks. **Anais**...Curran, 2021. Disponível em: <https://datasets-benchmarks-proceedings.neurips.cc/paper_files/paper/2021/file/084b6fbb10729ed4da8c3d3f5a3ae7c9-Paper-round2.pdf>

RAJPURKAR, P. et al. **SQuAD: 100,000+ Questions for Machine Comprehension of Text**. (J. Su, K. Duh, X. Carreras, Eds.)Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing. **Anais**...Austin, Texas: Association for Computational Linguistics, nov. 2016. Disponível em: <https://aclanthology.org/D16-1264>

REAL, L.; CARVALHO, A.; SILVA, A. DA. Avaliação de Grandes Modelos de Linguagem: Fundamentos, Métodos Tradicionais e Desafios Atuais. Em: CASELI, H. M.; NUNES, M. G. V. (Eds.). **Processamento de Linguagem Natural: Conceitos, Técnicas e Aplicações em Português**. 3. ed. [s.l.] BPLN, 2024.

REAL, L.; FONSECA, E.; GONÇALO OLIVEIRA, H. **The ASSIN 2 Shared Task: A Quick Overview**. Computational Processing of the Portuguese Language: 14th International Conference, PROPOR 2020, Evora, Portugal, March 2–4, 2020, Proceedings. **Anais**...Berlin, Heidelberg: Springer-Verlag, 2020. Disponível em: <https://doi.org/10.1007/978-3-030-41505-1_39>

REHAK, R. **AI Narrative Breakdown. A Critical Assessment of Power and Promise**. Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency. **Anais**...: FAccT '25.New York, NY, USA: Association for Computing Machinery, 2025. Disponível em: <https://doi.org/10.1145/3715275.3732083>

REIMERS, N.; GUREVYCH, I. **Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks**. Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing. **Anais**...Association for Computational Linguistics, nov. 2019. Disponível em: <https://arxiv.org/abs/1908.10084>

REIMERS, N.; GUREVYCH, I. **Making Monolingual Sentence Embeddings Multilingual using Knowledge Distillation**. Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing. **Anais**...Association for Computational Linguistics, nov. 2020. Disponível em: <https://arxiv.org/abs/2004.09813>

RIBEIRO, M. T. et al. **Beyond Accuracy: Behavioral Testing of NLP Models with CheckList**. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. **Anais**...Online: Association for Computational Linguistics, jul. 2020. Disponível em: <https://aclanthology.org/2020.acl-main.442>

RÍO, B. G. DEL; VAAHTIO, T. **Improving LLM systems with A/B testing**. https://www.flow-ai.com/blog/improving-llm-systems-with-a-b-testing, 2024.

RODRIGUES, J. et al. Advancing Neural Encoding of Portuguese with Transformer Albertina PT-. **CoRR**, v. abs/2305.06721, 2023.

RODRIGUES, R. C. et al. **Portuguese Language Models and Word Embeddings: Evaluating on Semantic Similarity Tasks**. (P. Quaresma et al., Eds.)Computational Processing of the Portuguese Language. **Anais**...Springer Nature Switzerland AG: Springer International Publishing, 2020.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol2/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

ROGERS, A. **Changing the World by Changing the Data**. Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). **Anais**...Online: Association for Computational Linguistics, 2021. Disponível em: <https://aclanthology.org/2021.acl-long.170>

ROGERS, A.; BALDWIN, T.; LEINS, K. **"Just What do You Think You're Doing, Dave?" A Checklist for Responsible Data Use in NLP**. Findings of the Association for Computational Linguistics: EMNLP 2021. **Anais**...Punta Cana, Dominican Republic: Association for Computational Linguistics, 2021. Disponível em: <https://aclanthology.org/2021.findings-emnlp.414>

ROMERA-PAREDES, B.; TORR, P. H. S. **An embarrassingly simple approach to zero-shot learning**. (F. R. Bach, D. M. Blei, Eds.)Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015. **Anais**...: JMLR Workshop e Conference Proceedings.JMLR.org, 2015. Disponível em: <http://proceedings.mlr.press/v37/romera-paredes15.html>

RUSSEL, S. **Human Compatible Artificial Intelligence and the Problem of Control**. [s.l.] Penguin Books, 2019.

SAI, A. B.; MOHANKUMAR, A. K.; KHAPRA, M. M. A Survey of Evaluation Metrics Used for NLG Systems. **ACM Comput. Surv.**, v. 55, n. 2, p. 26:1–26:39, 2023.

SAID, H. **40 Large Language Model Benchmarks and The Future of LLMs**. Arize AI; https://arize.com/blog/llm-benchmarks-mmlu-codexglue-gsm8k, 2025.

SAMPAIO, R. C.; SABBATINI, M.; LIMONGI, R. Diretrizes para o uso ético e responsável da inteligência artificial generativa: um guia prático para pesquisadores. **Boletim Técnico do PPEC**, v. 10, p. e025003–e025003, 2024.

SAMUYLOVA, E. **LLM-as-a-Judge: A Complete Guide to Using LLMs for Evaluations**. https://www.evidentlyai.com/llm-guide/llm-as-a-judge, 2025.

SÁNCHEZ, L. C. **Ethical Considerations and Best Practices in LLM Development**. https://neptune.ai/blog/llm-ethical-considerations, 2024.

SANH, V. et al. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. **CoRR**, v. abs/1910.01108, 2019.

SANTOS, A. A. et al. O teste de Cloze na avaliação da compreensão em leitura. **Psicologia: reflexão e crítica**, v. 15, p. 549–560, 2002.

SANTOS, D. **O projecto Processamento Computacional do Português: Balanço e perspectivas**. (M. das Graças Volpe Nunes, Ed.)V Encontro para o processamento computacional da língua portuguesa escrita e falada (PROPOR 2000). **Anais**...São Paulo: ICMC/USP, 2000. Disponível em: <https://www.linguateca.pt/Diana/download/SantosPROPOR2000.pdf>

SANTOS, H.; WOLOSZYN, V.; VIEIRA, R. **BlogSet-BR: A Brazilian Portuguese Blog Corpus**. (N. Calzolari et al., Eds.)Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018). **Anais**...Miyazaki, Japan: European Language Resources Association (ELRA), 2018. Disponível em: <https://aclanthology.org/L18-1105/>

SARDINHA, T. B.; FILHO, J. L. M.; ALAMBERT, E. **Manual Corpus Brasileiro**. [s.l: s.n.].

SARDINHA, T. B.; KAUFFMANN, C.; ACUNZO, C. M. A multi-dimensional analysis of register variation in Brazilian Portuguese. **Corpora**, v. 9, n. 2, p. 239–271, 2014.

SARTHI, P. et al. **RAPTOR: Recursive Abstractive Processing for Tree-Organized Retrieval**. International Conference on Learning Representations (ICLR). **Anais**...2024.

SAWCZYN, A. et al. FactSelfCheck: Fact-Level Black-Box Hallucination Detection for LLMs. **arXiv**,

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol2/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

2025.

SCAO, T. L. et al. BLOOM: A 176B-Parameter Open-Access Multilingual Language Model. **CoRR**, v. abs/2211.05100, 2022.

SCHAAKE, M. **The Tech Coup: How to Save Democracy from Silicon Valley**. [s.l.] Princeton University Press, 2024.

SCHEUERMAN, M. K.; HANNA, A.; DENTON, E. Do Datasets Have Politics? Disciplinary Values in Computer Vision Dataset Development. **Proceedings of the ACM on Human-Computer Interaction**, v. 5, n. CSCW2, p. 1–37, out. 2021.

SCHICK, T.; SCHÜTZE, H. **Exploiting Cloze-Questions for Few-Shot Text Classification and Natural Language Inference**. (P. Merlo, J. Tiedemann, R. Tsarfaty, Eds.)Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, EACL 2021, Online, April 19 - 23, 2021. **Anais**...Association for Computational Linguistics, 2021. Disponível em: <https://doi.org/10.18653/v1/2021.eacl-main.20>

SCHIRRU, L. **Direito autoral e Inteligência Artificial: autoria e titularidade nos produtos da IA**. [s.l.] Dialetica, 2023.

SCHLANGEN, D. **Targeting the Benchmark: On Methodology in Current Natural Language Processing Research**. Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers). **Anais**...Online: Association for Computational Linguistics, ago. 2021. Disponível em: <https://aclanthology.org/2021.acl-short.85>

SCHMIDHUBER, J.; HEIL, S. Sequential neural text compression. **IEEE Transactions on Neural Networks**, v. 7, n. 1, p. 142–146, 1996.

SCHNABEL, T. et al. **Evaluation methods for unsupervised word embeddings**. (L. Màrquez, C. Callison-Burch, J. Su, Eds.)Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. **Anais**...Lisbon, Portugal: Association for Computational Linguistics, set. 2015. Disponível em: <https://aclanthology.org/D15-1036/>

SCHULMAN, J. et al. **Proximal Policy Optimization Algorithms**. arXiv preprint arXiv:1707.06347. **Anais**...2017.

SCHUSTER, M.; NAKAJIMA, K. **Japanese and Korean voice search**. 2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). **Anais**...2012.

SCHWARTZ, L. **Primum Non Nocere: Before working with Indigenous data, the ACL must confront ongoing colonialism**. Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). **Anais**...Dublin, Ireland: Association for Computational Linguistics, 2022. Disponível em: <https://aclanthology.org/2022.acl-short.82>

SELVAN, R. et al. **Carbon Footprint of Selecting and Training Deep Learning Models for Medical Image Analysis**. (L. Wang et al., Eds.)Medical Image Computing and Computer Assisted Intervention – MICCAI 2022. **Anais**...Cham: Springer Nature Switzerland, 2022. Disponível em: <https://doi.org/10.1007/978-3-031-16443-9_49>

SERRAS, F. R. et al. **Exploring Computational Discernibility of Discourse Domains in Brazilian Portuguese within the Carolina Corpus**. (P. Gamallo et al., Eds.)Proceedings of the 16th International Conference on Computational Processing of Portuguese - Vol. 1. **Anais**...Santiago de Compostela, Galicia/Spain: Association for Computational Lingustics, mar. 2024. Disponível em: <https://aclanthology.org/2024.propor-1.26/>

SHAH, D.; SCHWARTZ, H. A.; HOVY, D. **Predictive Biases in Natural Language Processing Models: A Conceptual Framework and Overview**. Proceedings of the 58th Annual Meeting of the

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol2/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Association for Computational Linguistics. **Anais**...Online: Association for Computational Linguistics, 2020. Disponível em: <http://arxiv.org/abs/1912.11078>

SHAH, R. et al. Goal Misgeneralization in Deep Reinforcement Learning. **arXiv preprint arXiv:2210.01790**, 2022.

SHANNON, C. E. Prediction and entropy of printed English. **Bell System Technical Journal**, v. 30, n. 1, p. 50–64, 1951.

SHAQIRI, M. et al. Differences between the correlation coefficients Pearson, Kendall and Spearman. **5th International Conference of Natural Science and Mathematics**, nov. 2023.

SHENG, E. et al. **The Woman Worked as a Babysitter: On Biases in Language Generation**. (K. Inui et al., Eds.)Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). **Anais**...Hong Kong, China: Association for Computational Linguistics, nov. 2019. Disponível em: <https://aclanthology.org/D19-1339/>

SHI, H. et al. Continual learning of large language models: A comprehensive survey. **ACM Computing Surveys**, 2024.

SHI, Z.; LIPANI, A. **Don't Stop Pretraining? Make Prompt-based Fine-tuning Powerful Learner.**, 2023. Disponível em: <https://arxiv.org/abs/2305.01711>

SHMUELI, B. et al. **Beyond Fair Pay: Ethical Implications of NLP Crowdsourcing**. Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. **Anais**...Online: Association for Computational Linguistics, 2021. Disponível em: <https://aclanthology.org/2021.naacl-main.295>

SHUMAILOV, I. et al. AI models collapse when trained on recursively generated data. **Nature**, v. 631, n. 8022, p. 755–759, 2024.

ŞIMŞEK, C. **AI resistance: Who says no to AI and why?** Zenodo, 2025. Disponível em: <https://zenodo.org/doi/10.5281/zenodo.16893847>

ŞIMŞEK, C.; YASAR, A. G. From Rejection to Regulation: Mapping the Landscape of AI Resistance. 2025.

SØGAARD, A.; PLANK, B.; HOVY, D. **Selection Bias, Label Bias, and Bias in Ground Truth**. (Q. Liu, F. Xia, Eds.)Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Tutorial Abstracts. **Anais**...Dublin, Ireland: Dublin City University; Association for Computational Linguistics, ago. 2014. Disponível em: <https://aclanthology.org/C14-3005/>

SOLAIMAN, I. et al. Evaluating the Social Impact of Generative AI Systems. Em: **The Oxford Handbook of the Foundations and Regulation of Generative AI**. [s.l.] Oxford University Press, 2025.

SOUZA, F.; NOGUEIRA, R.; LOTUFO, R. **BERTimbau: pretrained BERT models for Brazilian Portuguese**. (R. Cerri, R. C. Prati, Eds.)Proceedings of the 2020 Brazilian Conference on Intelligent Systems. **Anais**...Springer International Publishing, 2020.

STIENNON, N. et al. **Learning to summarize with human feedback**. (H. Larochelle et al., Eds.)Advances in Neural Information Processing Systems. **Anais**...Curran Associates, Inc., a2020. Disponível em: <https://proceedings.neurips.cc/paper_files/paper/2020/file/1f89885d556929e98d3ef9b86448f951-Paper.pdf>

STIENNON, N. et al. Learning to Summarize with Human Feedback. **Advances in Neural Information Processing Systems (NeurIPS)**, b2020.

STRUBELL, E.; GANESH, A.; MCCALLUM, A. **Energy and Policy Considerations for Deep**

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol2/

Learning in NLP. Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. **Anais**...Florence, Italy: Association for Computational Linguistics, 2019. Disponível em: <https://www.aclweb.org/anthology/P19-1355>

STURZENEKER, M. et al. **Carolina's Methodology: building a large corpus with provenance and typology information.** (C. Trojahn et al., Eds.)Proceedings of the Second Workshop on Digital Humanities and Natural Language Processing (2nd DHandNLP 2022) co-located with International Conference on the Computational Processing of Portuguese (PROPOR 2022). **Anais**...Fortaleza, Brazil: 2022. Disponível em: <https://ceur-ws.org/Vol-3128/paper4.pdf>

SUN, K.; DREDZE, M. **Amuro & Char: Analyzing the Relationship between Pre-Training and Fine-Tuning of Large Language Models**. (V. Adlakha et al., Eds.)Proceedings of the 10th Workshop on Representation Learning for NLP (RepL4NLP-2025). **Anais**...Albuquerque, NM: Association for Computational Linguistics, 2025. Disponível em: <https://aclanthology.org/2025.repl4nlp-1.11/>

SURESH, H.; GUTTAG, J. V. A Framework for Understanding Sources of Harm throughout the Machine Learning Life Cycle. **Equity and Access in Algorithms, Mechanisms, and Optimization**, p. 1–9, out. 2021.

SUTSKEVER, I.; VINYALS, O.; LE, Q. V. **Sequence to Sequence Learning with Neural Networks**. (Z. Ghahramani et al., Eds.)Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada. **Anais**...2014. Disponível em: <https://proceedings.neurips.cc/paper/2014/hash/a14ac55a4f27472c5d894ec1c3c743d2-Abstract.html>

TAMKIN, A. et al. **Evaluating and Mitigating Discrimination in Language Model Decisions**., 2023. Disponível em: <https://arxiv.org/abs/2312.03689>

TANG, Y. et al. Multilingual Translation with Extensible Multilingual Pretraining and Finetuning. **CoRR**, v. abs/2008.00401, 2020.

TAYLOR, R. et al. Galactica: A Large Language Model for Science. **CoRR**, v. abs/2211.09085, 2022.

TAYLOR, W. L. "Cloze procedure": A new tool for measuring readability. **Journalism quarterly**, v. 30, n. 4, p. 415–433, 1953.

TEI CONSORTIUM. **TEI P5: Guidelines for Electronic Text Encoding and Interchange**. [s.l.] Retrieved May 20, 2021, 2021.

THAKUR, A. S. et al. JUDGING THE JUDGES: EVALUATING ALIGNMENT AND VULNERABILITIES IN LLMs-AS-JUDGES. **arXiv preprint arXiv:2406.12624v5**, 2025.

THOPPILAN, R. et al. LaMDA: Language Models for Dialog Applications. **CoRR**, v. abs/2201.08239, 2022.

THUN, M. VON; HANLEY, D. A. **Stopping Big Tech from Becoming Big AI: A Roadmap for Using Competition Policy to Keep Artificial Intelligence Open for All**. Open Markets Institute, 2024. Disponível em: <https://www.openmarketsinstitute.org/publications/report-stopping-big-tech-big-ai-roadmap>

TOLLES, J.; MEURER, W. J. Logistic Regression: Relating Patient Characteristics to Outcomes. **JAMA**, v. 316, n. 5, p. 533–534, ago. 2016.

TONIAZZO, D.; BARBOSA, T.; RUARO, R. O Direito à Explicação nas Decisões Automatizadas: uma Abordagem Comparativa Entre o Ordenamento Brasileiro e Europeu. **Revista Internacional Consinter de Direito**, v. 13, p. 55–69, dez. 2021.

TOUVRON, H. et al. LLaMA: Open and Efficient Foundation Language Models. **CoRR**, v. abs/2302.13971, a2023.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol2/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

TOUVRON, H. et al. **LLaMA: Open and Efficient Foundation Language Models**., b2023. Disponível em: <https://arxiv.org/abs/2302.13971>

TRAUTMANN, D. et al. **Measuring the Groundedness of Legal Question-Answering Systems**. Proceedings of the Natural Legal Language Processing Workshop 2024. **Anais**...Singapore: Association for Computational Linguistics, 2024. Disponível em: <https://aclanthology.org/2024.nllp-1.14>

ULMER, D. et al. **Experimental Standards for Deep Learning in Natural Language Processing Research**. Findings of the Association for Computational Linguistics: EMNLP 2022. **Anais**...Abu Dhabi, United Arab Emirates: Association for Computational Linguistics, dez. 2022. Disponível em: <https://aclanthology.org/2022.findings-emnlp.196>

VAROQUAUX, G.; LUCCIONI, S.; WHITTAKER, M. **Hype, Sustainability, and the Price of the Bigger-is-Better Paradigm in AI**. Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency. **Anais**...Athens Greece: ACM, jun. 2025. Disponível em: <https://dl.acm.org/doi/10.1145/3715275.3732006>

VASWANI, A. et al. **Attention is All you Need**. (I. Guyon et al., Eds.)Advances in Neural Information Processing Systems. **Anais**...Curran Associates, Inc., a2017. Disponível em: <https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html>

VASWANI, A. et al. Attention is all you need. **Advances in neural information processing systems**, v. 30, b2017.

VONGTHONGSRI, K. **G-Eval Simply Explained: LLM-as-a-Judge for LLM Evaluation**. Confident AI; https://www.confident-ai.com/blog/g-eval-the-definitive-guide, 2025.

WAGNER FILHO, J. A. et al. **The brWaC Corpus: A New Open Resource for Brazilian Portuguese**. (N. Calzolari et al., Eds.)Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018). **Anais**...Miyazaki, Japan: European Language Resources Association (ELRA), 2018. Disponível em: <https://aclanthology.org/L18-1686>

WANG, A. et al. **GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding**. Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP. **Anais**...Brussels, Belgium: Association for Computational Linguistics, nov. 2018. Disponível em: <https://aclanthology.org/W18-5446/>

WANG, A. et al. SuperGLUE: A Stickier Benchmark for General-Purpose Language Understanding Systems. **Advances in Neural Information Processing Systems**, v. 32, p. 3261–3275, a2019.

WANG, A. et al. SuperGLUE: a stickier benchmark for general-purpose language understanding systems. Em: **Proceedings of the 33rd International Conference on Neural Information Processing Systems**. Red Hook, NY, USA: Curran Associates Inc., 2019b.

WANG, B.; KOMATSUZAKI, A. **GPT-J-6B: A 6 Billion Parameter Autoregressive Language Model**. https://github.com/kingoflolz/mesh-transformer-jax, 2021.

WANG, R. et al. Can LLMs Replace Human Evaluators? An Empirical Study of LLM-as-a-Judge in Software Engineering. **Proc. ACM Softw. Eng.**, v. 2, n. ISSTA, jun. 2025.

WANG, Y. et al. **CodeT5: Identifier-aware Unified Pre-trained Encoder-Decoder Models for Code Understanding and Generation**. (M.-F. Moens et al., Eds.)Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021. **Anais**...Association for Computational Linguistics, 2021. Disponível em: <https://doi.org/10.18653/v1/2021.emnlp-main.685>

WANG, Y. et al. **PandaLM: An Automatic Evaluation Benchmark for LLM Instruction Tuning Optimization**. The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024. **Anais**...OpenReview.net, 2024. Disponível em: <https://openreview.net/forum

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol2/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

?id=5Nn2BLV7SB>

WEBER, M. et al. **RedPajama: an Open Dataset for Training Large Language Models**., 2024. Disponível em: <https://arxiv.org/abs/2411.12372>

WEI, J. et al. Emergent Abilities of Large Language Models. **Trans. Mach. Learn. Res.**, v. 2022, a2022.

WEI, J. et al. **Chain-of-Thought Prompting Elicits Reasoning in Large Language Models**. NeurIPS. **Anais**...b2022. Disponível em: <http://papers.nips.cc/paper/_files/paper/2022/hash/9d5609613524ecf4f15af0f7b31abca4-Abstract-Conference.html>

WEIDINGER, L. et al. Ethical and social risks of harm from Language Models. **arXiv:2112.04359 [cs]**, dez. 2021.

WEIGEND, A. **Data for the People: How to Make Our Post-Privacy Economy Work for You**. [s.l.] Basic Books, 2017.

WEIZENBAUM, J. **Computermacht und Gesellschaft : freie Reden / Joseph Weizenbaum. Hrsg. von Gunna Wendt ..** Original-Ausgabe, 1. Auflage ed. Frankfurt am Main: [s.n.].

WENDEHORST, C. Liability for Artificial Intelligence: The Need to Address Both Safety Risks and Fundamental Rights Risks. Em: VOENEKY, S. et al. (Eds.). **The Cambridge Handbook of Responsible Artificial Intelligence: Interdisciplinary Perspectives**. Cambridge Law Handbooks. [s.l.] Cambridge University Press, 2022. p. 187–209.

WENZEK, G. et al. **CCNet: Extracting High Quality Monolingual Datasets from Web Crawl Data**. (N. Calzolari et al., Eds.)Proceedings of the Twelfth Language Resources and Evaluation Conference. **Anais**...Marseille, France: European Language Resources Association, 2020. Disponível em: <https://aclanthology.org/2020.lrec-1.494/>

WERBOS, P. J. Backpropagation through time: what it does and how to do it. **Proc. IEEE**, v. 78, n. 10, p. 1550–1560, 1990.

WESTENBERGER, J.; SCHULER, K.; SCHLEGEL, D. Failure of AI projects: understanding the critical factors. **Procedia Computer Science**, v. 196, p. 69–76, 2022.

WHITTAKER, M. et al. Disability, bias, and AI. **AI Now Institute**, v. 8, n. 11, 2019.

WHITTAKER, M. The steep cost of capture. **Interactions**, v. 28, n. 6, p. 50–55, nov. 2021.

WIEGREFFE, S.; PINTER, Y. **Attention is not not Explanation**. (K. Inui et al., Eds.)Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). **Anais**...Hong Kong, China: Association for Computational Linguistics, nov. 2019. Disponível em: <https://aclanthology.org/D19-1002>

WIELING, M.; RAWEE, J.; VAN NOORD, G. Reproducibility in Computational Linguistics: Are We Willing to Share? **Computational Linguistics**, v. 44, n. 4, p. 641–649, dez. 2018.

WOLF, T. et al. **Transformers: State-of-the-Art Natural Language Processing**. Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations. **Anais**...Online: Association for Computational Linguistics, out. 2020. Disponível em: <https://www.aclweb.org/anthology/2020.emnlp-demos.6>

WOLFRAM, W. Variation and Language: Overview. Em: BROWN, K. (Ed.). **Encyclopedia of Language & Linguistics (Second Edition)**. Second Edition ed. Oxford: Elsevier, 2006. p. 333–341.

WRIGHT, B. **Manufacturing Reality: Slavoj Zizek and the Reality of the Virtual**. LondonBen Wright Film Productions, 2004.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol2/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

WU, T. **The attention merchants : from the daily newspaper to social media : how our time and attention is harvested and sold**. London: [s.n.].

WU, X.-K. et al. LLM Fine-Tuning: Concepts, Opportunities, and Challenges. **Big Data and Cognitive Computing**, v. 9, n. 4, p. 87, abr. 2025.

WU, Y. et al. **Memorizing Transformers**. The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022. **Anais**...OpenReview.net, 2022. Disponível em: <https://openreview.net/forum?id=TrjbxzRcnf->

XIE, S. M. et al. **An Explanation of In-context Learning as Implicit Bayesian Inference**. The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022. **Anais**...OpenReview.net, 2022. Disponível em: <https://openreview.net/forum?id=RdJVFCHjUMI>

XIONG, K. et al. **Com^2 : A Causal-Guided Benchmark for Exploring Complex Commonsense Reasoning in Large Language Models**. (W. Che et al., Eds.)Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). **Anais**...Vienna, Austria: Association for Computational Linguistics, jul. 2025. Disponível em: <https://aclanthology.org/2025.acl-long.785/>

XIONG, R. et al. **On Layer Normalization in the Transformer Architecture**. Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event. **Anais**...: Proceedings of Machine Learning Research.PMLR, 2020. Disponível em: <http://proceedings.mlr.press/v119/xiong20b.html>

XU, A. et al. **Detoxifying Language Models Risks Marginalizing Minority Voices**. (K. Toutanova et al., Eds.)Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. **Anais**...Online: Association for Computational Linguistics, jun. 2021. Disponível em: <https://aclanthology.org/2021.naacl-main.190/>

XU, H.; ASHLEY, K. **A question-answering approach to evaluating legal summaries**., 2023. Disponível em: <https://arxiv.org/abs/2309.15016>

XU, Q.; HE, X. **Security Challenges in Natural Language Processing Models**. Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: Tutorial Abstracts. **Anais**...Singapore: Association for Computational Linguistics, 2023. Disponível em: <https://aclanthology.org/2023.emnlp-tutorial.2>

XU, W.; RUDNICKY, A. **Can artificial neural networks learn language models?** Proc. 6th International Conference on Spoken Language Processing (ICSLP 2000). **Anais**...2000.

XU, Y. et al. **Hard Sample Aware Prompt-Tuning**. (A. Rogers, J. L. Boyd-Graber, N. Okazaki, Eds.)Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023. **Anais**...Association for Computational Linguistics, 2023. Disponível em: <https://aclanthology.org/2023.acl-long.690>

XUE, L. et al. **mT5: A Massively Multilingual Pre-trained Text-to-Text Transformer**. (K. Toutanova et al., Eds.)Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021, Online, June 6-11, 2021. **Anais**...Association for Computational Linguistics, 2021. Disponível em: <https://doi.org/10.18653/v1/2021.naacl-main.41>

YAMAGUCHI, A. et al. **Frustratingly Simple Pretraining Alternatives to Masked Language Modeling**. Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. **Anais**...Online; Punta Cana, Dominican Republic: Association for Computational Linguistics, nov. 2021. Disponível em: <https://aclanthology.org/2021.emnlp-main.249>

YAN, S.-Q. et al. **Corrective Retrieval Augmented Generation**., 2024. Disponível em: <https://arxiv.org/abs/2401.15884>

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol2/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

YANG, T. et al. **Ethics of Data Work. Principles for Academic Data Work Requesters**. Weizenbaum Institute, 2025. Disponível em: <https://www.weizenbaum-library.de/handle/id/920>

YANG, Z. et al. **XLNet: Generalized Autoregressive Pretraining for Language Understanding**. (H. M. Wallach et al., Eds.)Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada. **Anais**...2019. Disponível em: <https://proceedings.neurips.cc/paper/2019/hash/dc6a7e655d7e5840e66733e9ee67cc69-Abstract.html>

YAO, X. et al. **NLP From Scratch Without Large-Scale Pretraining: A Simple and Efficient Framework**., 2022. Disponível em: <https://arxiv.org/abs/2111.04130>

YE, Y. et al. **LIMO: Less is More for Reasoning**. arXiv, jul. 2025. Disponível em: <http://arxiv.org/abs/2502.03387>. Acesso em: 7 ago. 2025

YEH, Y.-T.; ESKÉNAZI, M.; MEHRI, S. A Comprehensive Assessment of Dialog Evaluation Metrics. **ArXiv**, v. abs/2106.03706, 2021.

YEO, W. et al. UniversalRAG: Retrieval-Augmented Generation over Corpora of Diverse Modalities and Granularities. **arXiv preprint arXiv:2504.20734**, 2025.

YUAN, W.; NEUBIG, G.; LIU, P. **BARTScore: Evaluating Generated Text as Text Generation**. Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual. **Anais**...2021. Disponível em: <https://proceedings.neurips.cc/paper/2021/hash/e4d2b6e6fdeca3e60e0f1a62fee3d9dd-Abstract.html>

ZHANG, D.; XU, Z.; ZHAO, W. **LLMs and Copyright Risks: Benchmarks and Mitigation Approaches**. (M. Lomeli, S. Swayamdipta, R. Zhang, Eds.)Proceedings of the 2025 Annual Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 5: Tutorial Abstracts). **Anais**...Albuquerque, New Mexico: Association for Computational Linguistics, a2025. Disponível em: <https://aclanthology.org/2025.naacl-tutorial.7/>

ZHANG, K. et al. **CitaLaw: Enhancing LLM with Citations in Legal Domain**. (W. Che et al., Eds.)Findings of the Association for Computational Linguistics: ACL 2025. **Anais**...Vienna, Austria: Association for Computational Linguistics, jul. b2025. Disponível em: <https://aclanthology.org/2025.findings-acl.583/>

ZHANG, S. et al. **Instruction Tuning for Large Language Models: A Survey**., c2025. Disponível em: <https://arxiv.org/abs/2308.10792>

ZHANG, T. et al. **BERTScore: Evaluating Text Generation with BERT**. 8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020. **Anais**...OpenReview.net, 2020. Disponível em: <https://openreview.net/forum?id=SkeHuCVFDr>

ZHANG, Y. et al. Siren's Song in the AI Ocean: A Survey on Hallucination in Large Language Models. **Computational Linguistics**, v. 51, n. 4, p. 1373–1418, dez. d2025.

ZHAO, W. X. et al. A Survey of Large Language Models. **CoRR**, v. abs/2303.18223, 2023.

ZHAO, Y. et al. **One Token to Fool LLM-as-a-Judge**., 2025. Disponível em: <https://arxiv.org/abs/2507.08794>

ZHOU, C. et al. LIMA: Less Is More for Alignment. **CoRR**, v. abs/2305.11206, a2023.

ZHOU, H. et al. A survey of large language models in medicine: Progress, application, and challenge. **arXiv preprint arXiv:2311.05112**, b2023.

ZIEGLER, D. M. et al. Fine-Tuning Language Models from Human Preferences. **CoRR**, v. abs/1909.08593, 2019.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol2/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

ZUBOFF, S. **The age of surveillance capitalism : the fight for a human future at the new frontier of power**. First edition ed. New York: [s.n.].

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol2/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Sobre as/os autoras/es

Aline Paes

Aline Paes é formada em Ciência da Computação pela UERJ, com mestrado e doutorado em inteligência artificial pela COPPE/Sistemas/UFRJ e estágio sanduíche de doutorado no Imperial College London. É docente na UFF desde 2013. Suas áreas de interesse incluem aprendizado de máquina relacional, aprendizado de representações para linguagem, IA para impacto social positivo e IA explicável. Neste livro, colaborou com os Capítulos 1 e 7.
Email: alinepaes@ic.uff.br

Altigran Soares Silva

Altigran Soares da Silva concluiu doutorado em Ciência da Computação pela UFMG (2002) e é professor titular do Instituto de Computação da UFAM. Seus interesses incluem: Aprendizagem de Máquina, Modelos de Linguagem, Engenharia de Dados e Recuperação de Informação. Neste livro, colaborou nos capítulos Avaliação de Large Language Models: Fundamentos, Métricas Tradicionais e Desafios Atuais e Fundamentos da Geração Aumentada por Recuperação. Neste livro, colaborou com os Capítulos 4 e 5.
Email: alti@icomp.ufam.edu.br

André Carvalho

André Carvalho é doutor em Informática pela Universidade Federal do Amazonas, onde, desde 2013 é professor associado, sendo também também é coordenador do bacharelado em Inteligência Artificial. Também é coordenador do projeto de pesquisa IATS - Inteligência Artificial aplicada em Teste de Software, que tem como foco aplicação de técnicas de processamento de linguagem natural em tarefas de qualidade de software. Sua pesquisa tem como foco áreas como Inteligência Artificial, Aprendizagem de Máquina, Recuperação de Informação e Ciência de Dados. Neste livro, colaborou com os Capítulos 4 e 5.
Email: andre@icomp.ufam.edu.br

Brielen Madureira

Brielen Madureira é formada em Matemática Aplicada e Computacional pela Universidade de São Paulo (USP). Tem mestrado em Ciência e Tecnologia da Linguagem pela Universidade de Saarland (Alemanha) e doutorado em Linguística Computacional pela Universidade de Potsdam (Alemanha), onde participou do grupo de pesquisa em Diálogo. Mais amplamente, tem interesse na avaliação de tecnologias de linguagem e nas urgentes considerações éticas que elas demandam. Atualmente é pesquisadora em fase de pós-doutorado na Universidade de Leipzig (Alemanha), no grupo Climate Discourse. Neste livro, colaborou com os Capítulos Avaliação de tecnologias de linguagem, Diálogo e Interatividade e 6.
Email: brielen.madureira@uni-leipzig.de

Cláudia Freitas

Cláudia Freitas é formada em Letras pela PUC-Rio, onde concluiu o doutorado em Estudos da Linguagem (2007) sobre a construção de ontologias a partir de *corpus*. De 2012 a 2022 foi docente do Programa de Pós-Graduação em Estudos da Linguagem da PUC-Rio, onde fundou o grupo ComCorHD (Linguística Computacional, *Corpus* e Humanidades Digitais). De 2007 a 2022 foi também pesquisadora da Linguateca. Ao longo de 2023 e 2024 foi pesquisadora do ICMC/USP, vinculada ao C4AI, e atualmente é pesquisadora independente. Suas áreas de interesse incluem construção e avaliação de *datasets*/*corpora*, semântica/significado, e articulação entre abordagens simbólicas e neurais. Neste livro, colaborou com os Capítulos E o significado?, Conjunto de dados, *dataset* e *corpus*, Avaliação conjunta em português e 7.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol2/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Email: claudiafreitas@usp.br

Cristiane Namiuti

Cristiane Namiuti é Bacharel em Linguística pela UNICAMP (2002) onde também obteve o título de Doutora em Linguística (2008). Fez parte da equipe do Corpus Anotado do Português Histórico Tycho Brahe (1998-2010). Desde 2008, atua como docente da Universidade Estadual do Sudoeste da Bahia (UESB) e no PPGLin/UESB desde sua fundação em 2010. Atualmente, integra o grupo de pesquisadores e co-fundadores dos Laboratórios LAPELINC/UESB e LaViHD/USP e, desde 2020, faz parte de uma das equipes do C4AI - Centro de Inteligência Artificial da USP (LaViHD-C4AI). Suas áreas de interesse incluem metodologias automáticas de anotação e busca de dados em textos escritos, mudança linguística, história do português, linguística de corpus e humanidades digitais. Neste livro, colaborou com o Capítulo 2.
Email: cristianenamiuti@gmail.com

Daniela Vianna

Daniela Vianna tem Ph.D. em Ciência da Computação pela Rutgers University, USA, e mestrado e bacharelado em Ciência da Computação pela Universidade Federal Fluminense, Brasil. Em 2022, completou um pós-doutorado na Universidade Federal do Amazonas, Brasil, em uma parceria com a Jusbrasil. Seus interesses estão nas áreas de aprendizado de máquina e PLN. Neste livro, colaborou com o 1.
Email: dvianna@gmail.com

Danilo Samuel Jodas

Danilo Jodas é formado em Ciência da Computação pelo Centro Universitário do Norte Paulista (UNORP), com pós-graduação em processamento de imagens e aprendizado de máquina pela UNESP e FEUP (Universidade do Porto). É pesquisador de pós-doutorado na UNESP de Bauru desde 2019, atuando em pesquisas de aprendizado de máquina e PLN. Neste livro, colaborou com os Capítulos Sumarização Automática e 3.
Email: danilo.jodas@unesp.br

Douglas Rodrigues

Douglas Rodrigues é formado em Informática para Gestão de Negócios pela FATEC/Botucatu, Mestrado (UNESP/Bauru) e Doutorado (UFSCar/São Carlos) em Ciências da Computação. Atualmente, atua como pós-doutorando e pesquisador no Recogna/UNESP. Suas áreas de interesse incluem aprendizado de máquina e otimização matemática. Neste livro, colaborou com os Capítulos Sumarização Automática e 3.
Email: d.rodrigues@unesp.br

Gabriel Lino Garcia

Gabriel Lino Garcia é graduado em Análise e Desenvolvimento de Sistemas pela FATEC e mestre em Inteligência Artificial (IA) e Processamento de Linguagem Natural (PLN) pela UNESP, onde faz doutorado focado na aplicação de modelos de linguagem na área médica. Seus interesses incluem aprendizado de máquina, grandes modelos de linguagem, aprendizado multimodal e aplicações médicas. Neste livro, colaborou com os Capítulos Sumarização Automática e 3.
Email: linaoglino@gmail.com

Gabriela Alves Lachi

Gabriela Alves Lachi é graduanda em Letras pela Universidade de São Paulo (USP). Suas áreas de interesse em Processamento de Linguagem Natural envolvem a construção e o desenvolvimento de corpus, bem como o treinamento e a avaliação de grandes modelos de linguagem. Neste livro, colaborou com o Capítulo 2.
Email: gabriela.lachi@usp.br

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Felipe Ribas Serras

Felipe Ribas Serras é Bacharel em Física pelo Instituto de Física da Universidade de São Paulo (USP) e Mestre em Ciência da Computação pelo Instituto e Matemática e Estatística da USP. Suas áreas de interesse incluem complexidade de linguagem e abordagens computacionais à variação linguística e tipologia. Neste livro, colaborou com os capítulos Complexidade Textual e suas Tarefas Relacionadas, 2.
Email: frserras@ime.usp.br

Jessica Rodrigues

Jessica Rodrigues tem graduação e mestrado em Ciência da Computação, sendo o mestrado focado em *machine learning* e *natural language processing*, pela Universidade Federal de São Carlos (UFSCar). A experiência acadêmica continua com o PhD em *social data science* pela University of Oxford, com foco em *AI safety, fairness and biases in language models*. Jessica também é Lead Data Scientist e é especialista em AI aplicada. Neste livro, colaborou com o Capítulo 1.
Email: jessica.rodrigues@oii.ox.ac.uk

João Paulo Papa

João Paulo Papa é formado em Sistemas de Informação pela UNESP, com pós-graduação em processamento de imagens e aprendizado de máquina pela UFSCar e Unicamp, respectivamente. Possui pós-doutorados na Unicamp, Harvard e MIT. É professor titular da Unesp onde atua desde 2009 e suas áreas de interesse incluem visão computacional, processamento de imagens, aprendizado de máquina e aprendizado multimodal. Neste livro, colaborou com os Capítulos Sumarização Automática e 3.
Email: joao.papa@unesp.br

João Renato Ribeiro Manesco

João Renato Ribeiro Manesco é doutorando em Ciência da Computação pela UNESP, onde também concluiu mestrado e graduação. Durante o mestrado, realizou estágio de pesquisa no Media Integration and Communication Center da Universidade de Florença, Itália. Seus interesses de pesquisa incluem adaptação de domínio, modelos de linguagem generativos e multimodais, e aplicações em contexto médico. Neste livro, colaborou com os Capítulos Sumarização Automática e 3.
Email: joao.r.manesco@unesp.br

João Vitor Mariano Correia

João Vitor Mariano Correia é bacharel em Ciência da Computação pela UNESP, onde atualmente cursa o mestrado no Programa de Pós-Graduação em Ciência da Computação. Seus principais interesses de pesquisa estão na área de Processamento de Linguagem Natural (PLN), com foco em representações de conhecimento. Neste livro, colaborou com os Capítulos Sumarização Automática e 3.
Email: mariano.correia@unesp.br

Livy Real

Livy Real é formada em Letras, com doutorado em linguística (UFPR) e atuação na interseção entre linguagem e inteligência artificial. Realizou estágios pós-doutorais na IBM Research e na Universidade Federal do Amazonas (UFAM). Atualmente, atua como cientista aplicada em inteligência artificial no Instituto Kunumi. Seus interesses incluem avaliação de modelos de IA, NLP aplicado, linguística computacional e questões éticas no uso de tecnologias emergentes. Neste livro, colaborou com os Capítulos 4 e 5.
Email: livyreal@gmail.com

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Lucas Lasota

Lucas Lasota é mestre e doutor pela Universidade RUDN de Moscou com o doutorado reconhecido pela USP. É pesquisador e docente do Just Transition Center da Universidade de Halle-Wittenberg na Alemanha. É pesquisador associado do Weizenbaum Institute for the Networked Society de Berlin. Advogado inscrito na OAB-SP/432126. Sua pesquisa se concentra em medidas regulatórias de tecnologias digitais e seu impacto sobre direitos individuais e coletivos, bem como governança da internet, telecomunicações e direito contratual internacional. Neste livro, colaborou com o Capítulo 6.
Email: lucas.lasota@jtc.uni-halle.de

Marcelo Finger

Marcelo Finger é professor titular de Ciência da Copmputação no IME-USP, Pesquisador Principal do Centro de Inteligência Artificial USP-IBM-FAPESP C4AI, com interesses em Geração de Recursos de Processamento de Linguagem Natural (PLN), uso de PLN para a área de Saúde, estudo e interpretabilidade de redes neurais. Neste livro, colaborou com os Capítulos Complexidade Textual e suas Tarefas Relacionadas, Classificação de Áudio aplicada à Saúde e 2.
Email: mfinger@ime.usp.br

Mariana Lourenço Sturzeneker

Mariana Lourenço Sturzeneker é graduada em Letras pela Universidade de São Paulo. Faz parte da equipe de desenvolvimento do Córpus Carolina no C4AI (Center for Artificial Intelligence - IBM/USP) desde 2020. Seu interesse principal é na área de construção de córpus. Neste livro, colaborou com o Capítulo 2.
Email: mariana.sturzeneker@usp.br

Mayara Feliciano Palma

Mayara Feliciano Palma é Bacharel em Letras (português e espanhol) pela Faculdade de Filosofia, Letras e Ciências Humanas da Universidade de São Paulo (USP). Entre suas áreas de interesse estão a construção de *corpora* e agentes. Neste livro, colaborou com o Capítulo 2.
Email: mayara.palma@usp.br

Miguel de Mello Carpi

Miguel de Mello Carpi é Bacharel em Ciência da Computação pelo Instituto de Matemática e Estatística da Universidade de São Paulo. Suas áreas de interesse incluem inteligência artificial e processamento de linguagem natural. Neste livro, colaborou com o Capítulo 2.
Email: miguel@ime.usp.br

Pedro Henrique Paiola

Pedro Henrique Paiola é mestre e doutorando em Ciência da Computação pela UNESP, onde também atua como professor substituto desde 2022. É pesquisador no projeto RADIAR, desenvolvido em parceria com a Petrobras. Suas áreas de interesse incluem sumarização automática, modelos de linguagem e avaliação automática de texto. Neste livro, colaborou com os Capítulos Sumarização Automática e 3.
Email: pedro.paiola@unesp.br

Vanessa Martins do Monte

Vanessa Martins do Monte é mestre e doutora em Filologia e Língua Portuguesa pela FFLCH-USP, onde atua como docente desde 2014. Coordena o Projeto M.A.P. – Mulheres na América Portuguesa e integra o C4AI – Centro de Inteligência Artificial da USP, atuando na constituição de *corpora* para PLN. Suas pesquisas concentram-se em Filologia e Humanidades Digitais. Neste livro, colaborou com o Capítulo 2.
Email: vmmonte@usp.br

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Processamento de Linguagem Natural

Conceitos, Técnicas e Aplicações em Português

Organizado por:
Helena de Medeiros Caseli
Maria das Graças Volpe Nunes

4ª Edição
Volume 3: Aplicações e Domínios
Abril 2026

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

Conteúdo

Volume 3: Aplicações e Domínios 1
Sobre o Volume 3 . . . 1
Como citar . . . 1

I Aplicações e Domínios 3

1 Classificação de texto 4
1.1 Taxonomia . . . 5
1.2 Tipos de problemas . . . 6
1.2.1 Classificação binária . . . 6
1.2.2 Classificação multiclasse . . . 7
1.2.3 Classificação multirrótulo . . . 7
1.2.4 Classificação hierárquica . . . 7
1.3 Representações computacionais de texto . . . 8
1.3.1 Representações distributivas (esparsas) . . . 8
1.3.2 Representações distribuídas (densas) . . . 10
1.3.3 Representações contextuais (modelos de linguagem profundos) . . . 11
1.4 Métodos de classificação . . . 12
1.4.1 Métodos probabilísticos: Naïve Bayes . . . 13
1.4.2 Métodos baseados em distância: K-vizinhos mais próximos . . . 14
1.4.3 Métodos de otimização: Regressão Logística e SVM . . . 15
1.4.4 Métodos baseados em árvores e *ensembles* . . . 17
1.4.5 Métodos conexionistas: Redes neurais e *Deep learning* . . . 18
1.4.6 Modelos de linguagem . . . 19
1.4.7 Redes com atenção: Transformers . . . 19
1.4.8 Classificação via *prompt* e modelos de linguagem de larga escala (LLMs) . . . 20
1.4.9 Síntese e Recomendação . . . 22
1.5 Avaliação de modelos . . . 22
1.5.1 Metodologias de validação . . . 22
1.5.2 Métricas de desempenho . . . 23
1.5.3 Análise via Curva ROC e AUC . . . 25
1.6 Recursos em português . . . 26
1.6.1 Conjuntos de dados para categorização . . . 26
1.6.2 Modelos de linguagem pré-treinados . . . 27
1.6.3 Ferramentas e Bibliotecas . . . 27
1.7 Desafios e Tendências . . . 27
1.7.1 Dinamismo e Desvio de Conceito (*Concept Drift*) . . . 28

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

1.7.2 Escalabilidade e Eficiência energética . . . 28
1.7.3 Explicabilidade e Interpretabilidade (XAI) . . . 28
1.7.4 Ética, Equidade (*Fairness*) e Mitigação de vieses . . . 28
1.7.5 Soberania digital . . . 29
1.7.6 Categorização em baixos recursos e dialetos . . . 29
1.8 Referências para aprofundamento . . . 29

2 Recuperação de Informação 30
2.1 Introdução . . . 30
2.1.1 Relação com o PLN . . . 30
2.1.2 O Foco da Recuperação de Informação . . . 30
2.1.3 O Conceito de Relevância . . . 31
2.1.4 Principais Livros . . . 31
2.1.5 Organização deste Capítulo . . . 32
2.2 Visão Geral de um Sistema Típico de Recuperação de Informação . . . 32
2.2.1 Pré-Processamento . . . 32
2.2.2 Indexação . . . 35
2.3 Modelos Clássicos de recuperação de informação . . . 36
2.3.1 Modelo Booleano . . . 36
2.3.2 Modelo Vetorial . . . 36
2.3.3 Modelos Probabilísticos . . . 40
2.4 Avaliação da Qualidade de Sistemas de recuperação de informação . . . 42
2.4.1 Métricas . . . 42
2.4.2 Coleções de Teste . . . 46
2.5 Modificação Automática de Consultas . . . 49
2.6 Ferramentas e Bibliotecas . . . 50
2.6.1 Sistemas Comerciais . . . 50
2.6.2 Ferramentas para Pesquisa . . . 50
2.7 Conclusão . . . 51
Agradecimentos . . . 51

3 Extração de Informação 52
3.1 Introdução . . . 52
3.2 Um pouco de história . . . 53
3.3 Conceituação formal: Relação e Entidade . . . 55
3.3.1 Entidade . . . 55
3.3.2 Relação . . . 55
3.4 Extração de Informação (EI) . . . 56
3.4.1 Reconhecimento de Entidades Nomeadas . . . 56
3.4.2 Extração de Relações . . . 57
3.4.3 Extração Conjunta de Entidades e Relações . . . 58
3.4.4 Métodos empregados para EI na literatura . . . 58
3.5 Extração de Informação Aberta . . . 62
3.5.1 Formalização . . . 63
3.5.2 Abordagens . . . 63
3.6 Avaliação . . . 65
3.7 Considerações finais . . . 67
Agradecimentos . . . 67

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

4 Tradução Automática **68**
4.1 Introdução . . . 68
4.2 Abordagens . . . 69
4.2.1 Tradução direta . . . 69
4.2.2 Tradução Automática Baseada em Regras . . . 70
4.2.3 Tradução por interlíngua . . . 71
4.2.4 Tradução Automática Baseada em Exemplos . . . 72
4.2.5 Tradução Automática Estatística . . . 73
4.2.6 Tradução Automática Neural . . . 76
4.3 Avaliação da Tradução Automática . . . 79
4.3.1 O que é Avaliação da Tradução Automática? . . . 80
4.3.2 A importância de uma avaliação replicável . . . 83
4.3.3 Métricas Automáticas . . . 84
4.3.4 Métricas Humanas . . . 86
4.3.5 Avaliação dependente de contexto . . . 90
4.4 O Futuro da Tradução Automática . . . 92

5 Sumarização Automática **94**
5.1 Introdução . . . 94
5.2 Por trás do processo . . . 96
5.2.1 Sobre conceitos . . . 96
5.2.2 Sobre o processo de sumarização . . . 98
5.3 Avaliação na SA . . . 100
5.4 Aplicação em Sumarização Extrativa . . . 103
5.4.1 Algoritmo baseado na frequência das palavras . . . 104
5.4.2 Algoritmo baseado em Transformers . . . 107
5.4.3 Avaliando os sumários . . . 109
5.5 Aplicação em Sumarização Abstrativa . . . 111
5.5.1 Algoritmos baseados em estrutura sintática e/ou semântica . . . 112
5.5.2 Algoritmos baseados em Transformers . . . 113
5.5.3 Sumarização abstrativa na prática . . . 115
5.6 A Sumarização automática e a língua portuguesa . . . 120
5.7 Considerações finais . . . 121

6 Complexidade Textual e suas Tarefas Relacionadas **123**
6.1 Introdução . . . 123
6.2 Complexo para quem? . . . 124
Por que não ensinar a ler em vez de simplificar um texto? . . . 124
6.3 Tarefas de PLN relacionadas à complexidade textual . . . 125
6.3.1 Adaptação textual . . . 125
6.3.2 Simplificação textual . . . 126
6.3.3 Simplificação lexical . . . 127
6.3.4 Simplificação sintática . . . 128
6.3.5 Sumarização automática . . . 128
6.3.6 Elaboração textual . . . 129
6.4 Avaliação da Complexidade e suas Métricas . . . 129
6.4.1 Fórmulas clássicas . . . 130
6.4.2 Métricas Linguísticas . . . 132

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

6.4.3 Métricas Psicolinguísticas . . . 137
6.4.4 NILC-Metrix . . . 139
6.4.5 Rastreamento ocular . . . 140
6.5 Recursos para Complexidade Textual e Sentencial no Português Brasileiro . 142
6.5.1 Recursos para Complexidade Textual . . . 142
6.5.2 Recursos para Complexidade Sentencial . . . 146
6.6 Uso de Modelos de Língua para a Simplificação Textual . . . 146
6.7 A complexidade como parâmetro de descrição da linguagem . . . 148
6.7.1 Questões sobre a complexidade como parâmetro . . . 150
6.7.2 Métricas de Complexidade Baseadas em Compressibilidade . . . 151
6.8 Considerações finais . . . 156

7 Correção Automática de Redação 157
7.1 Introdução . . . 157
7.1.1 O que é uma redação escolar? . . . 158
7.1.2 O que é avaliado? . . . 159
7.1.3 Alguns modelos brasileiros de correção . . . 160
7.2 Detecção de desvios no texto . . . 162
7.2.1 Tipos de desvios . . . 163
7.2.2 Formalismos de regras . . . 165
7.3 Atribuição de nota . . . 168
7.3.1 Como atribuir nota a redações? . . . 168
7.3.2 Atribuição de nota para redações do Enem . . . 171
7.4 Feedback para o aluno . . . 172
7.4.1 Estatísticas básicas do texto . . . 173
7.4.2 Assistentes de escrita e ferramentas de auxílio à escrita . . . 174
7.4.3 Identificação de pontos fortes e elogiáveis . . . 176
7.5 Correção manual vs(?) correção automática . . . 178
7.5.1 Avanços dos últimos anos . . . 178
7.5.2 Vantagens da correção automática . . . 179
7.5.3 Vantagens da correção manual . . . 179
7.5.4 O exemplo da língua inglesa . . . 180
7.5.5 O que defendemos . . . 181
7.6 Considerações finais . . . 181

8 Detecção Automática de Notícias Falsas 182
8.1 Introdução . . . 182
8.1.1 Conteúdo enganoso: definições . . . 183
8.1.2 Notícias falsas no tempo e no espaço . . . 184
8.1.3 Combate ao conteúdo enganoso . . . 186
8.2 *Corpora* . . . 187
8.3 Métodos de detecção automática . . . 189
8.3.1 Aprendizado de máquina com atributos linguísticos . . . 190
8.3.2 Modelos de linguagem . . . 191
8.3.3 Checagem de conteúdo . . . 194
8.3.4 Recuperação de informação . . . 195
8.3.5 Análise de credibilidade . . . 195
8.3.6 Outras iniciativas relacionadas . . . 196

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

8.4 Impactos para a sociedade . . . 197
8.5 Considerações finais . . . 200
8.6 Exercícios de fixação . . . 201

9 PLN na Saúde 203
9.1 Introdução . . . 203
9.2 O texto livre em narrativas clínicas . . . 205
9.3 Aplicações de PLN na Saúde . . . 207
9.3.1 Predição . . . 207
9.3.2 Desidentificação . . . 207
9.3.3 Extração de conceitos clínicos . . . 208
9.3.4 Relações temporais . . . 209
9.3.5 Sumarização . . . 210
9.4 Recursos de PLN na área clínica . . . 211
9.4.1 *Corpora* . . . 211
9.4.2 Modelos . . . 212
9.5 Para onde estamos caminhando? . . . 213

10 Detecção de Transtornos de Saúde Mental a partir de Texto 215
10.1 Introdução . . . 215
10.1.1 Definição do problema . . . 216
10.2 Conjuntos de dados para detecção de depressão/ansiedade . . . 218
10.3 Modelos computacionais de detecção de depressão/ansiedade . . . 221
10.4 SetembroBR: detecção de depressão/ansiedade em português . . . 223
10.4.1 *Corpus* SetembroBR . . . 224
10.4.2 Modelos desenvolvidos no projeto SetembroBR . . . 225
10.5 Amive: detecção de sintomas de depressão em português . . . 229
10.5.1 Sinais associados à depressão . . . 229
10.5.2 *Corpus* do projeto Amive e anotação manual . . . 232
10.5.3 Identificação automática de sinais de depressão . . . 233
10.6 Considerações finais . . . 233
Agradecimentos . . . 234

11 Classificação de Áudio aplicada à Saúde 235
11.1 Introdução . . . 235
11.1.1 Detecção de COVID-19 através da voz e tosse . . . 236
11.1.2 Insuficiência Respiratória e seus efeitos na voz e fala . . . 237
11.1.3 Outros problemas respiratórios . . . 238
11.1.4 Organização do capítulo . . . 239
11.2 *Corpora* de áudios . . . 239
11.2.1 *Corpus* de detecção de insuficiência respiratória . . . 239
11.2.2 *Corpus* de detecção de COVID-19 . . . 242
11.2.3 *Corpora* para detecção de outros problemas respiratórios . . . 244
11.3 Métodos de detecção de problemas respiratórios através de áudios . . . 244
11.3.1 Pré-processamento de áudios . . . 244
11.3.2 Modelos de aprendizado profundo para classificação de áudios . . . 246
11.3.3 Avaliação dos modelos de detecção de problemas respiratórios . . . 250
11.4 Considerações finais e impactos para a sociedade . . . 251

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

12 PLN no Direito **254**
12.1 Introdução . . . 254
12.2 O Direito – uma moldura para a significação . . . 255
12.3 Entre as terminologias e as palavras no Direito do Brasil . . . 257
12.4 Um caso concreto: em pequeníssima escala . . . 258
12.5 Outros casos/exemplos: Direito Ambiental . . . 260
12.6 Aplicação: Análise de Sentimentos em Direito: desafios e exemplos . . . 262
12.7 Considerações finais . . . 266

13 Reconhecimento de Entidades Nomeadas no Domínio Legal **268**
13.1 Introdução . . . 268
13.2 Contextualização . . . 268
13.3 Iniciativas de REN-Legal . . . 270
13.4 Iniciativas de REN-Legal em língua portuguesa . . . 273
13.4.1 *Corpora* para REN-Legal . . . 273
13.4.2 Modelos de linguagem neurais para REN-Legal . . . 279
13.5 Modelos de linguagem de grande escala (LLM) para REN . . . 281
13.5.1 LLM para REN-Legal em língua portuguesa . . . 282
13.6 Aplicação de REN em sistema de recuperação de informação legislativa . . . 284
13.7 Considerações finais . . . 285
Agradecimentos . . . 286

14 Geração Automática de Ementas Jurídicas **287**
14.1 Introdução . . . 287
14.1.1 Contextualização . . . 287
14.1.2 Organização do Capítulo . . . 289
14.2 Conceitos Preliminares . . . 289
14.2.1 Transferência de Aprendizado e Ajuste Fino . . . 289
14.2.2 *Bilingual Evaluation Under-study* (BLEU) . . . 289
14.3 Geração de verbetações . . . 290
14.3.1 Aquisição de dados . . . 291
14.3.2 Geração textual . . . 291
14.4 Avaliação utilizando RI . . . 292
14.4.1 Formulação da tarefa . . . 293
14.4.2 Experimentos realizados . . . 293
14.4.3 Métodos de RI avaliados . . . 294
14.4.4 Ordenação dos documentos . . . 295
14.4.5 Métricas avaliadas . . . 295
14.5 Resultados e discussões . . . 295
14.5.1 Comparação entre verbetações originais e geradas . . . 295
14.5.2 Experimentos com RI . . . 297
14.6 Conclusão . . . 299
Agradecimentos . . . 299

15 PLN e Humanidades Digitais **300**
15.1 Introdução . . . 300
15.2 Preparação das fontes . . . 301
15.2.1 Digitalização . . . 302

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

15.2.2 Metadados . . . 302
15.2.3 Normalização . . . 303
15.2.4 Anotação de textos . . . 304
15.3 Transformação de textos em dados . . . 305
15.3.1 Extração de informação . . . 305
15.3.2 Representação de conhecimento e ontologias . . . 306
15.3.3 Dados ligados e dados *FAIR* . . . 306
15.4 Exemplos de projetos de HD envolvendo o processamento da língua portuguesa . . . 308
15.4.1 O *corpus* Tycho Brahe . . . 308
15.4.2 As Memórias Paroquiais . . . 308
15.4.3 O projeto MONSOON . . . 309
15.4.4 A História do Futuro . . . 310
15.4.5 Manuais médicos do século XVIII . . . 311
15.5 Considerações finais . . . 312
Agradecimentos . . . 313

16 PLN em Redes Sociais **314**
16.1 Introdução . . . 314
16.2 Redes Sociais . . . 314
16.2.1 Facebook . . . 315
16.2.2 Reddit . . . 316
16.2.3 Youtube . . . 316
16.2.4 Twitter/X . . . 316
16.2.5 Whatsapp . . . 316
16.2.6 Instagram . . . 316
16.3 Áreas de Aplicação . . . 317
16.3.1 Detecção de Discurso de Ódio e Linguagem Ofensiva . . . 317
16.3.2 Análise de Sentimento . . . 322
16.3.3 Detecção de Notícias Falsas . . . 325
16.3.4 Detecção de Ironia/Sarcasmo/Humor . . . 327
16.4 Considerações finais . . . 329

17 PLN e Responsabilidade Social **330**
17.1 Introdução . . . 330
17.2 Semântica de *Frames* . . . 331
17.3 FrameNet Brasil . . . 333
17.3.1 As relações *frame-a-frame* . . . 336
17.3.2 As relações qualia ternárias . . . 339
17.4 Projeto DataToStopGBV: Semântica contra a violência de gênero . . . 341
17.4.1 Contexto e motivação . . . 342
17.4.2 A contribuição da FrameNet para identificação da Violência Baseada em Gênero . . . 345
17.4.3 Avaliação da identificação da Violência Baseada em Gênero . . . 350
17.4.4 Impacto Social e técnico . . . 353
17.5 Multimodalidade, acessibilidade cultural e sentido compartilhado no Audition: *dataset* multimodal de curtas-metragens acessíveis . . . 355
17.5.1 O *dataset* Audition . . . 356

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

17.5.2 Anotação semântica multimodal . . . 360
17.5.3 Validação, Aplicações em PLN e Impactos na Tradução Audiovisual Acessível . . . 364
17.6 Considerações finais . . . 366
Agradecimentos . . . 367
Glossário deste capítulo . . . 367

18 Redes complexas no processamento de língua natural do português 370
18.1 Introdução . . . 370
18.2 Redes complexas . . . 371
18.2.1 Conceitos fundamentais . . . 372
18.2.2 Propriedades . . . 373
18.2.3 Tipos de redes . . . 375
18.2.4 Dinâmica . . . 378
18.3 Detecção de comunidades . . . 379
18.4 Aplicações de redes complexas em PLN para o português . . . 381
18.5 Exemplo de aplicação: Panorama de PLN em língua portuguesa . . . 385
18.5.1 Redes de citações e similaridade . . . 386
18.5.2 Identificação de tópicos em redes de pesquisa . . . 387
18.5.3 Pesquisa em PLN para o português . . . 388
18.6 Considerações finais . . . 395

Referências 397

Apêndices 486

A Considerações prévias à sumarização automática 486

B Exemplo de sumarização abstrativa 490

Sobre as/os autoras/es 493
Helena de Medeiros Caseli . . . 493
Maria das Graças Volpe Nunes . . . 493
Adriana Silvina Pagano . . . 493
Adriano L. I. Oliveira . . . 493
Aline Macohin . . . 494
Amanda Pontes Rassi . . . 494
Ana Paula Banza . . . 494
Ana Sofia Ribeiro . . . 494
André Carlos Ponce de Leon Ferreira de Carvalho . . . 494
Beatriz Raposo de Medeiros . . . 495
Brenda Salenave Santana . . . 495
Cassia Trojahn dos Santos . . . 495
Celso Ricardo Fernandes de Carvalho . . . 495
Claudia Maria Cabral Moro Barra . . . 496
Crysttian Arantes Paixão . . . 496
Daniela Barreiro Claro . . . 496
Danilo Samuel Jodas . . . 496

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Diego Raphael Amancio 496
Douglas Rodrigues 497
Elisa Terumi Rubel Schneider 497
Ellen Souza 497
Evelin Amorim 497
Fátima Farrica 498
Felipe Ribas Serras 498
Fernanda Olival 498
Flaviane Romani Fernandes Svartman 498
Gabriel Lino Garcia 498
Helena Freire Cameron 499
Hidelberg O. Albuquerque 499
Ivandré Paraboni 499
Jackson Wilke da Cruz Souza 499
João Paulo Papa 500
João Renato Ribeiro Manesco 500
João Vitor Mariano Correia 500
Joaquim Santos 500
Kenzo Sakiyama 500
Larissa Cristina Berti 501
Larissa A. de Freitas 501
Lilian Mie Mukai Cintho 501
Lívia Vicente Dutra 501
Lucas Daniel Volpiano Figueiredo 501
Marcelo Finger 502
Marcelo Matheus Gauy 502
Marcelo Gomes de Queiroz 502
Maria Cristina Ferreira de Oliveira 502
Maria José Bocorny Finatto 502
Mariane de Carvalho Pinto 503
Marlo Souza 503
Matheus Cerqueira 503
Maucha Andrade Gamonal 503
Nádia Félix Felipe da Silva 503
Osvaldo Novais de Oliveira Jr. 504
Paula Christina Figueira Cardoso 504
Pedro Henrique Paiola 504
Priscilla de Abreu Lopes 504
Rafael Oleques Nunes 504
Raphael Montanari 505
Renata Vieira 505
Renato Moraes Silva 505
Rodrigo Nogueira 505
Roney Lira de Sales Santos 505
Roseli A. F. Romero 506
Roseval Donisete Malaquias Junior 506
Sandra Maria Aluísio 506
Sheila Castilho 506

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Sidney Evaldo Leal 507
Thiago Alexandre Salgueiro Pardo 507
Tiago Agostinho de Almeida 507
Tiago Timponi Torrent 507
Vitor Hugo Bridi Dei Santi 507
Viviane Pereira Moreira 508
Yohan Bonescki Gumiel 508

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Volume 3: Aplicações e Domínios

Sobre o Volume 3

Dos dezoito capítulos deste volume, oito ilustram o PLN em várias aplicações: classificação de textos (Capítulo 1), recuperação (Capítulo 2) e extração (Capítulo 3) de informações, tradução (Capítulo 4) ou sumarização (Capítulo 5) automáticas, avaliação de complexidade textual (Capítulo 6) e correção automática de redações (Capítulo 7) e detecção de notícias falsas (*fake news*) (Capítulo 8).

Três capítulos ilustram o PLN aplicado ao domínio da saúde: o Capítulo 9 discorre sobre as tarefas de PLN comuns a esse domínio; o Capítulo 10 traz uma aplicação na área de saúde mental; e o Capítulo 11 mostra uma aplicação do processamento de fala para detecção de problemas respiratórios. Já no domínio jurídico ou do direito, o Capítulo 12 traz os principais conceitos e desafios para o PLN, sendo que o tratamento das entidades nomeados neste domínio é abordado no Capítulo 13, e a geração automática de ementas jurídicas é apresentada no Capítulo 14.

O Capítulo 15 discorre sobre a potencialidade do uso de PLN na ampla área de humanidades digitais, caracterizada por trabalhos derivados das ciências humanas que abordam a linguagem. O Capítulo 16 introduz o gênero especial da linguagem usada nas redes sociais, enquanto que o Capítulo 17 ilustra uma aplicação do formalismo Framenet no enfrentamento à violência de gênero e na promoção da acessibilidade cultural. Finalmente, o Capítulo 18 explora o potencial do formalismo de redes complexas para modelar textos para aplicações de PLN, ilustrando com o mapeamento das pesquisas em PLN para português.

A leitura dos capítulos que integram esse volume deve expandir o horizonte dos leitores para as variadas áreas de aplicação do PLN, e mostrar-lhes a potencialidade do emprego de PLN quando ajustado a um domínio de conhecimento, um gênero específico ou uma aplicação em particular.

Como citar

Caseli, H.M.; Nunes, M.G.V. (org.) **Processamento de Linguagem Natural: Conceitos, Técnicas e Aplicações em Português**. 4 ed. BPLN, 2026. v. 3. ISBN 978-65-02-00167-7. Disponível em: https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3.

Note 1: Bibref

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

```
@Book{BPLN_livro_4ed_vol3:2026,
   title     = {Processamento de Linguagem Natural: Conceitos,
                Técnicas e Aplicações em Português},
   editor    = {Caseli, H. M. and Nunes, M. G. V.},
   year      = {2026},
   publisher = {BPLN},
   edition   = {4},
   volume    = {3},
   isbn      = {978-65-02-00167-7},
   url       = {https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3}
}
```

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Parte I

Aplicações e Domínios

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Capítulo 1

Classificação de texto

Evelin Amorim
Tiago A. Almeida

Publicado em: 16/04/2026

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

A **classificação de texto** é uma das tarefas fundamentais e mais estudadas no Processamento de Linguagem Natural (PLN). Em sua essência, a tarefa consiste em atribuir um ou mais rótulos ou categorias a um documento de texto, com base em seu conteúdo. Essa tarefa, aparentemente simples, é a base para uma vasta gama de aplicações que encontramos em nosso dia a dia como: a organização automática de e-mails em pastas (por exemplo, "Principal", "Social", "Promoções"), a identificação do sentimento em uma avaliação de produto (positivo, negativo, neutro) ou a classificação de notícias por tópicos (como "Esportes", "Política", "Tecnologia"). Como resultado desta tarefa temos uma coleção de textos organizada de acordo com critérios humanos relevantes para um propósito como o de auxiliar as pessoas a localizar rapidamente a informação de que precisam. Esta tarefa pode, portanto, ser a primeira etapa para as pessoas catalogarem informação. É por este motivo que esta tarefa também pode ser chamada de categorização de texto (Sebastiani, 2002).

A importância da categorização de texto reside em sua capacidade de estruturar a imensa quantidade de dados textuais não estruturados gerados a cada segundo. Como descrito no Capítulo O que é PLN?, um dos grandes desafios da área é justamente criar métodos que permitam aos computadores "entender" e processar a linguagem humana para extrair informações úteis. A categorização é um passo crucial nesse processo, transformando texto livre em informação organizada e facilmente acessível.

A separação de textos em categorias diferentes pode parecer uma tarefa simples à primeira vista; porém, estabelecer as categorias mais adequadas dentro de um conjunto, definir os nomes das categorias, entre outros critérios a serem estabelecidos para os dados, pode consumir muito tempo e demandar conhecimento específico (Freitas Araujo, 2018). Por isso, existe um campo de estudo dedicado a esta organização, chamado Ciência da Informação, que pode abrigar cursos sobre diferentes tipos de catalogação e organização da informação, como Biblioteconomia e Arquivologia, entre outros. Essa área surgiu após a Segunda Guerra Mundial, com o avanço e o desenvolvimento tecnológico (Saracevic, 1996).

Podemos considerar, portanto, a separação de textos em categorias pré-definidas como uma subárea desta grande área, a Ciência da Informação. Com o surgimento da Inteligência Artificial (IA), também por volta da década de 1950 (Haenlein; Kaplan, 2019), foi natural associá-la à crescente quantidade de informação gerada pelos avanços tecnológicos. Assim, a classificação de texto foi uma das primeiras tarefas exploradas em IA por volta de 1960. Maron (1961) propôs classificar um conjunto de resumos de artigos em 32 categorias utilizando como características preditivas as palavras presentes nos resumos do

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

conjunto de treino. Desde então, muitos avanços foram alcançados. Os trabalhos que se seguiram mostraram o lado multifacetado da classificação de texto, por exemplo, aplicando diferentes estratégias de IA para classificar textos de saúde ou jurídicos, ou definindo que um documento pode ser classificado em mais de uma categoria. Com relação ao domínio textual, referimos ao leitor os Capítulos 10 e 8 para a classificação de textos no domínio da saúde mental e a detecção de notícias falsas, respectivamente.

Neste capítulo, exploraremos os múltiplos aspectos da categorização de texto, desde seus fundamentos teóricos até suas aplicações práticas, com foco especial nos recursos disponíveis para o português. A seguir, definiremos a taxonomia à qual um conjunto de dados de classificação de texto pode pertencer.

1.1 Taxonomia

O termo "categorização de texto" é frequentemente usado como um guarda-chuva para uma variedade de problemas específicos. Podemos organizar as tarefas de categorização com base em diferentes critérios, como a natureza das categorias, a relação entre elas e o objetivo final da aplicação.

Uma primeira grande divisão se dá entre a **classificação de tópicos** (ou assuntos) e a **classificação por metadados**. A **classificação de tópicos (*topic classification*)** é a forma mais tradicional de categorização, cujo objetivo é identificar o assunto principal de um texto. Por exemplo, um artigo de jornal pode ser classificado como "Ciência" e uma publicação em rede social como "Viagem". As categorias são definidas pelo conteúdo semântico do documento. Na **classificação por metadados**, as categorias não se referem ao tópico do texto, mas a um atributo ou metadado associado a esse tópico. Exemplos incluem:

- **Análise de sentimentos:** classificar a polaridade de um texto (e.g., positivo, negativo, neutro), uma tarefa explorada em aplicações de redes sociais, como discutido no Capítulo 16.
- **Detecção de autoria:** identificar o autor de um texto com base em seu estilo de escrita (Varela et al., 2011).
- **Detecção de spam:** diferenciar e-mails legítimos de spam (Almeida; Yamakami, 2012).
- **Detecção de notícias falsas:** classificar uma notícia como legítima ou falsa, uma aplicação de grande impacto social abordada no Capítulo 8.

Neste capítulo não trataremos deste tipo de classificação, porém essa distinção é crucial, pois os tipos de características (*features*) relevantes para cada tarefa podem variar drasticamente. Enquanto a classificação de tópicos depende muito do vocabulário específico do domínio, a análise de sentimentos pode depender de palavras com carga afetiva e a detecção de autoria, de padrões estilísticos sutis.

A estrutura e a relação entre as categorias também definem diferentes tipos de problemas, como veremos em detalhe na próxima seção, abordando desde a simples classificação binária até cenários mais complexos, como a classificação hierárquica e multirrótulo.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

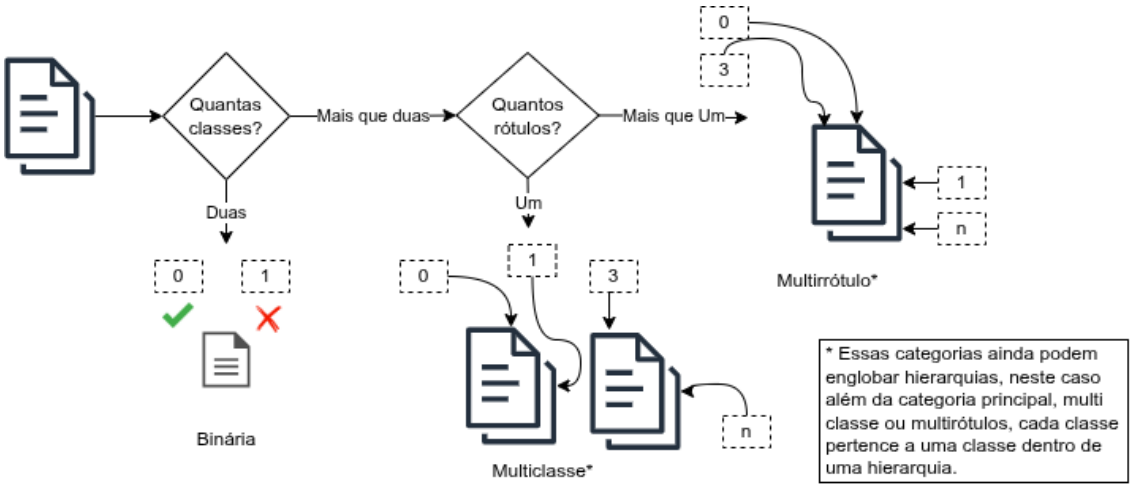
1.2 Tipos de problemas

A taxonomia apresentada na seção anterior descreve o propósito da categorização. Agora, vamos formalizar os tipos de problemas com base na estrutura do conjunto de categorias C (também chamado de espaço de rótulos ou *label space*). A definição do problema tem impacto direto na forma como o modelo é projetado, treinado e avaliado (como veremos nas Seções 1.4 e 1.5).

Para treinar e avaliar os modelos que executam essas tarefas, dependemos fundamentalmente de conjuntos de dados anotados, os chamados *corpora*, um conceito central detalhado no Capítulo Conjunto de dados, *dataset* e *corpus*. A estrutura e a relação entre as **classes (categorias ou rótulos)** também definem diferentes tipos de problemas. A identificação do tipo de classificação de texto é importante para levantar os principais desafios e as melhores estratégias para solução. Para isso, é necessário primeiro identificar o número de classes na tarefa e, em seguida, o número de rótulos por instância.

A Figura 1.1 apresenta um fluxograma que guia a identificação do tipo de classificação a que uma categorização textual pertence. Seja D um conjunto de documentos e $C = c_1, c_2, \ldots, c_K$ um conjunto de K categorias (rótulos) predefinidas. O objetivo de um classificador f é aprender uma função de mapeamento $f : D \rightarrow C'$ que associa um documento $d \in D$ a um subconjunto de rótulos $C' \subseteq C$. A natureza desse mapeamento define os problemas descritos a seguir (Aggarwal, 2018; Jurafsky; Martin, 2023).

Figura 1.1: Fluxograma dos tipos de problemas de classificação de texto.



1.2.1 Classificação binária

É o cenário mais simples e fundamental de categorização. O conjunto de categorias contém exatamente duas classes, $K = 2$ (ou seja, $C = \{c_1, c_2\}$), que são mutuamente exclusivas. Neste caso, o classificador aprende uma função $f : D \rightarrow \{c_1, c_2\}$. Frequentemente, as classes são representadas como $\{+1, -1\}$ (positivo/negativo) ou como $\{0, 1\}$ (ausente/presente).

Exemplos clássicos: incluem detecção de *spam*, na qual um e-mail é classificado como "spam" ou "legítimo" (*ham*); análise de sentimentos (polaridade), na qual uma avaliação de produto ou serviço é "positiva" ou "negativa"; e detecção de notícias falsas, na qual uma notícia pode ser classificada como "legítima" ou "falsa".

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

1.2.2 Classificação multiclasse

O conjunto de categorias contém mais de duas classes ($K > 2$), mutuamente exclusivas. Cada documento d deve estar associado a exatamente uma categoria do conjunto C. Neste caso, o classificador aprende uma função $f : D \rightarrow C$, com $|C| = K > 2$.

Exemplos tradicionais incluem: classificação de tópicos de notícias, no qual um artigo de notícia é atribuído a um único assunto, como "Esporte", "Política" ou "Tecnologia"; análise de sentimentos (granular), na qual uma avaliação de produto ou serviço é classificada como "muito positiva", "positiva", "neutra", "negativa" ou "muito negativa"; e identificação de idioma, que determina se um texto está escrito em "Português", "Inglês", "Espanhol" etc (Silva et al., 2017b).

1.2.3 Classificação multirrótulo

A classificação multirrótulo rompe a restrição da exclusividade mútua. As categorias não são mutuamente exclusivas, e um documento d pode ser associado a zero, a um ou a múltiplos rótulos simultaneamente (Tsoumakas; Katakis, 2007).

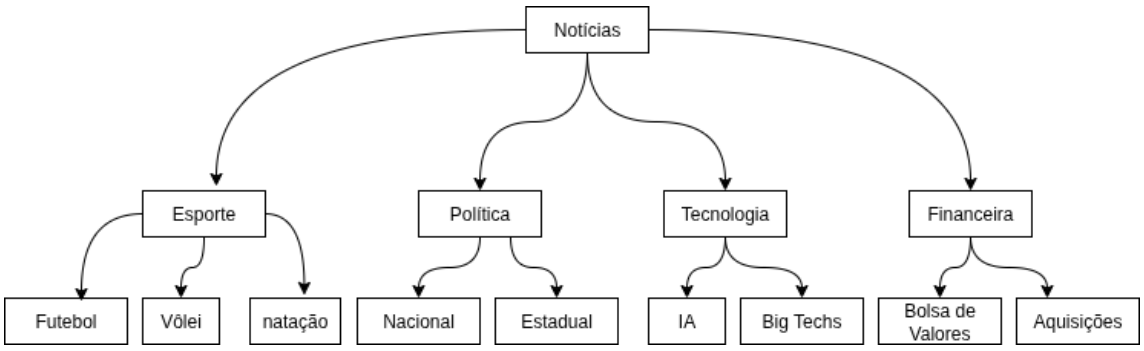
Exemplos incluem: categorização de artigos científicos, no qual um estudo sobre aprendizado de máquina aplicado à medicina pode receber os rótulos "Inteligência Artificial", "PLN" e "Saúde"; classificação de conteúdo jurídico, no qual uma peça processual pode ser relevante para múltiplos ramos do direito, como "Direito Civil" e "Direito do Consumidor"; e marcação de publicações em redes sociais, no qual um *post* sobre uma viagem gastronômica pode ser marcado com *tags* como "#Viagem", "#Gastronomia" e "#Férias".

Este problema é significativamente mais complexo, pois o espaço de saída cresce exponencialmente (2^K combinações possíveis) e exige métodos específicos, como a Relevância Binária (*Binary Relevance*) ou algoritmos adaptados que modelam a correlação entre os rótulos (Bittencourt et al., 2020).

1.2.4 Classificação hierárquica

Este é um caso especial de classificação (que pode ser multiclasse ou multirrótulo), em que as próprias categorias estão organizadas em uma estrutura hierárquica, como uma árvore ou um grafo acíclico dirigido. As categorias em C mantêm relações de superclasse/subclasse (pai/filho). A decisão de classificar um documento em uma categoria mais específica (e.g., "Futebol") pode depender da classificação na categoria mais geral (e.g., "Esporte") (Figura 1.2).

Figura 1.2: Exemplo de hierarquia de classes em uma tarefa de classificação de notícias.



https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Exemplos incluem classificação de produtos, na qual um item em um *e-commerce* pode ser classificado em "Eletrônicos" → "Áudio e vídeo" → "Fones de ouvido"; classificação de documentos médicos, na qual registros de pacientes ou artigos científicos são classificados de acordo com o uso de taxonomias como a CID (Classificação Internacional de Doenças[1]) ou a MeSH (*Medical Subject Headings*[2]); e taxonomias biológicas, nas quais uma espécie de ser vivo pode ser classificada em Reino, Filo, Classe, Ordem etc.

Abordagens para este problema podem ser planas (*flat*), ignorando a hierarquia (i.e., tratando-a como um problema multiclasse/multirrótulo padrão), ou hierárquicas, que utilizam a estrutura da taxonomia para guiar o processo de classificação (Silla; Freitas, 2011; Tavares et al., 2018). Neste caso, a classificação hierárquica pode ser resolvida com abordagens *top-down* (vários classificadores em cascata) ou *big-bang* (um único classificador que conhece a estrutura).

Compreender qual desses problemas estamos resolvendo é o primeiro passo para selecionar a representação de texto (Seção 1.3), o método de classificação (Seção 1.4) e as métricas de avaliação (Seção 1.5) mais adequados.

1.3 Representações computacionais de texto

Modelos de aprendizado de máquina, como os que veremos na Seção 1.4, não operam sobre texto puro; exigem entradas numéricas. O processo de conversão de um texto (uma sequência de caracteres) em um vetor numérico — seja ele esparso ou denso — é chamado de representação de texto ou *feature extraction* (extração de características).

A qualidade dessa representação é fundamental, pois ela deve capturar a informação semântica e discriminativa do documento, filtrando ruídos de modo que o classificador possa encontrar padrões. A evolução das representações de texto é um pilar central da própria evolução do PLN, passando de modelos baseados em contagem a modelos que aprendem o significado a partir do contexto.

1.3.1 Representações distributivas (esparsas)

A primeira abordagem bem-sucedida, que serviu de base para a Recuperação de Informação (ver Capítulo 2) por décadas, é baseada na Hipótese Distributiva, que, em sua forma mais simples, pode ser resumida como: "*você conhecerá uma palavra pela companhia que ela mantém*" (Firth, 1957). Embora essa hipótese seja a base de todas as representações modernas, suas primeiras implementações eram baseadas em contagem.

1.3.1.1 Saco de palavras

A representação em saco de palavras ou *bag-of-words* (BoW) é a forma mais simples de representar um documento de texto. Nela, um texto é dividido em palavras e cada palavra é associada a sua ocorrência ou ao valor de sua frequência nesse texto.

Considere o exemplo a seguir[3].

[1] Sistema de classificação mantido pela OMS para codificação de doenças e problemas de saúde. Disponível em: http://www2.datasus.gov.br/cid10/V2008/cid10.htm

[2] Vocabulário controlado desenvolvido pela National Library of Medicine dos EUA para indexação de artigos biomédicos. Disponível em: https://www.nlm.nih.gov/mesh/meshhome.html

[3] Frase extraída do livro Memórias Póstumas de Brás Cubas e livremente adaptada para o português moderno, disponível gratuitamente no Projeto Gutenberg (https://www.gutenberg.org/).

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

Exemplo 1.1:

> talvez espante ao leitor a franqueza com que lhe exponho e realço a minha mediocridade

Em um modelo de saco de palavras, o texto é primeiramente dividido em palavras unitárias, separadas por um espaço em branco (processo conhecido como tokenização explicado no Capítulo Sequência de caracteres e palavras), resultando no mapeamento a seguir onde cada palavra é associada a sua frequência de ocorrência.

```
{ "a": 2, "ao": 1, "com": 1, "e": 1, "espante": 1, "exponho": 1,
"franqueza": 1, "leitor": 1, "lhe": 1, "mediocridade": 1, "minha": 1,
"que": 1, "realço": 1, "talvez": 1}
```

Essa representação ignora completamente a gramática, a sintaxe e a ordem das palavras, tratando o documento apenas como um multiconjunto (um "saco") de seus termos.

Formalizando, dado um vocabulário V (o conjunto de todas as palavras únicas do *corpus*), cada documento d é representado por um vetor v de tamanho $|V|$. Cada posição i do vetor v corresponde a um termo $t_i \in V$, e o valor v_i é a frequência (contagem) do termo t_i no documento d.

Essa representação tem a vantagem de ser simples e de apresentar fácil interpretabilidade (é fácil ver quais palavras contribuíram para a classificação) e eficácia surpreendente na classificação de tópicos (onde a presença de palavras-chave é um forte indicador). Por outro lado, os vetores resultantes podem ter elevada dimensionalidade (dezenas de milhares de posições), ser esparsos (a maioria das posições é zero), apresentar perda total da ordem (por exemplo, "o cachorro mordeu o homem" *vs* "o homem mordeu o cachorro") e ser incapazes de capturar sinonímia (os vetores para "carro" e "automóvel" são ortogonais, ou seja, totalmente independentes).

1.3.1.2 *Term Frequency-Inverse Document Frequency* (TF-IDF)

Uma melhoria crucial em relação à contagem bruta é o TF-IDF, que combina *Term Frequency – TF* (frequência de termo) e *Inverse Document Frequency – IDF* (frequência inversa de documentos). Essa junção visa dar mais peso a palavras frequentes em um documento específico, mas raras em todos os demais documentos. Palavras muito comuns em toda a coleção (como *stopwords* "de", "que", "o") recebem um peso baixo, pois não têm poder discriminativo.

O peso de um termo t no documento d é dado por:

$$tfidf(t, d, D) = tf(t, d) \times idf(t, D), \tag{1.1}$$

onde $tf(t, d)$ é a frequência do termo em d, computada por:

$$tf(t, d) = \frac{freq_{t,d}}{\sum_{t' \in d} freq_{t',d}}, \tag{1.2}$$

e $idf(t, D)$ é a frequência inversa do documento no corpus D:

$$idf(t, D) = \log \frac{N}{n_t} \tag{1.3}$$

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Na Equação 1.2, $freq_{t,d}$ é a frequência de um termo t no documento d. O denominador $\sum_{t' \in d} freq_{t',d}$ representa a soma das frequências de todos os termos t' presentes em d. O objetivo desta soma é normalizar a frequência de um termo em relação ao tamanho do documento. Um documento longo deve apresentar termos com frequência elevada, enquanto um documento pequeno apresenta termos com frequência baixa. Para evitar distorções, a quantidade de termos de um documento é considerada.

Após o cálculo da frequência de um termo, é necessário considerar o quão raro ou frequente ele é em um conjunto de textos D (Equação 1.3). Para seu cálculo, utilizam-se a quantidade de documentos no corpus (N) e o número de documentos em D em que t ocorre (n_t).

O resultado dessa representação ainda é um vetor esparso e de alta dimensionalidade, mas os valores são ponderados pela sua relevância informacional, tornando-o uma *baseline* muito mais robusta para classificadores tradicionais.

Para um arcabouço teórico mais profundo, indicamos ao leitor o conteúdo sobre vetores esparsos (Seção Vetores esparsos). A representação TF-IDF também pode ser utilizada no contexto da Recuperação de Informação (RI). O Capítulo 2 fornece mais detalhes sobre o TF-IDF neste contexto.

1.3.2 Representações distribuídas (densas)

A grande revolução na representação de texto veio com a aplicação de redes neurais para aprender representações, em vez de apenas contar as palavras. Este é o foco da Seção Vetores densos estáticos.

As representações vetoriais são formalmente chamadas de Modelos Semânticos Distribucionais (MSD). *Word embeddings* (vetores de palavras) gerados por modelos como Word2Vec (Mikolov et al., 2013a) e GloVe (Pennington et al., 2014) aprendem vetores densos (normalmente de 50 a 300 dimensões) para cada palavra, em que cada dimensão idealmente representa uma característica semântica da palavra. A principal vantagem dessa representação em relação àquela apresentada anteriormente é que palavras com significados semelhantes (como "carro" e "automóvel") terão vetores próximos no espaço vetorial.

Contudo, a tarefa de categorização opera sobre documentos, não sobre palavras. Dessa forma, para obter um vetor que represente um documento, as abordagens mais comuns são:

1. **Concatenação:** o vetor do documento é obtido a partir da concatenação dos vetores das palavras;
2. **Média de vetores:** o vetor do documento é obtido calculando-se a média dos vetores de todas as palavras no documento, elemento a elemento (Jurafsky; Martin, 2026);
3. **Média ponderada:** o vetor do documento é obtido como resultado da média ponderada na qual as palavras de maior peso (por exemplo, o TF-IDF) contribuem mais para o vetor final.

As representações distribuídas resolvem o problema da sinonímia, mas ainda sofrem com a perda de ordem e, crucialmente, com a polissemia: a palavra "banco" terá o mesmo vetor em "sentei no banco da praça" e em "fui ao banco sacar dinheiro". Existem, entretanto, representações distribuídas densas cujo objetivo é representar por completo um

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

documento textual ou frases (Le; Mikolov, 2014). Este tipo de representação, assim como métodos específicos, como Word2Vec, GloVe e FastText, é detalhado na Seção Vetores densos estáticos.

1.3.3 Representações contextuais (modelos de linguagem profundos)

A fronteira atual, que resolve o problema da polissemia, é o uso de modelos de linguagem profundos baseados na arquitetura Transformer, como o BERT (Devlin et al., 2019) e suas variantes (Liu et al., 2019b) (RoBERTa, ALBERT e outras).

Esses modelos são pré-treinados em volumes massivos de texto (como discutido no Capítulo Modelos de linguagem) e não geram um vetor fixo por palavra. Em vez disso, eles geram um vetor para cada *token* (palavra ou subpalavra) que é sensível ao contexto no qual esse *token* aparece. Neste caso, o vetor de representação da palavra "banco" será diferente nas duas sentenças de exemplo da seção anterior.

Para a tarefa de categorização de textos, temos duas estratégias principais:

1. **Extração de características (*feature extraction*):** o documento é passado pelo modelo pré-treinado (sem modificá-lo), e os vetores de saída são usados como representação. Frequentemente, utiliza-se a representação do *token* especial `[CLS]` (no caso do BERT), projetado para agregar o significado de toda a sequência. Alternativamente, pode-se aplicar uma operação de *pooling* (por exemplo, a média) aos vetores de todos os *tokens* da sequência. Esse vetor denso e contextual (por exemplo, de 768 dimensões) é então usado como entrada de um classificador clássico[4].

2. **Ajuste fino (*fine-tuning*):** esta é a abordagem dominante. O modelo de linguagem pré-treinado é "descongelado" e treinado (com uma taxa de aprendizado baixa) juntamente com uma camada de classificação simples (por exemplo, softmax) adicionada ao topo. O modelo inteiro é ajustado para a tarefa específica de categorização. Isso permite que as representações, que eram de propósito geral, se especializem nos padrões discriminativos da base de dados de aplicação.

A ascensão dos modelos de linguagem de larga escala (do inglês, *Large Language Models* – LLMs), como o GPT e modelos de código aberto (discutidos no Capítulo ChatGPT, MariTalk e outros agentes de conversação), também permite abordagens de *zero-shot* ou *few-shot learning*, nas quais o modelo pode categorizar textos com base apenas na descrição das categorias (ou seja, um *prompt*), sem necessitar de um grande conjunto de treinamento rotulado.

A escolha da representação (Quadro 1.1) depende do problema (Seção 1.2) e dos recursos disponíveis: o TF-IDF ainda é uma *baseline* rápida e robusta para tópicos claros; *fine-tuning* de modelos de linguagem é o estado da arte para a maioria das tarefas que exigem compreensão semântica profunda.

[4] Em modelos mais recentes (como o RoBERTa), o uso da média dos vetores (conhecida como *mean pooling*) tem se mostrado mais estável do que o `[CLS]` isolado para certas tarefas de categorização

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Quadro 1.1: Comparativo das abordagens de representação para classificação de textos

Abordagem	Unidade Básica	Dimension.	Captura Semântica	Sensível ao Contexto
TF-IDF (Distributiva)	Palavra (Termo)	Alta e Esparsa (e.g., 50.000)	Nenhuma (Contagem)	Não
Word Embeddings (Distribuída)	Palavra	Baixa e Densa (e.g., 300)	Alta (Estática)	Não
LLMs (Contextual)	Subpalavra (*Token*)	Média e Densa (e.g., 768+)	Altíssima (Dinâmica)	Sim

A identificação e a análise dos documentos que se deseja classificar são importantes para determinar o tipo de classificação. Após isso, é necessária a representação destes documentos para a aprendizagem de um modelo.

1.4 Métodos de classificação

Uma vez que o texto foi transformado em uma representação numérica (Seção 1.3), entramos na fase de tomada de decisão. O objetivo desta seção é apresentar os métodos que atuam como o "cérebro" do sistema de categorização, responsáveis por aprender padrões nesses números e atribuir a classe correta a um documento inédito.

Formalmente, o problema de classificação pode ser visualizado como a busca por uma função de mapeamento $f : D \to C'$ que associa cada documento $d \in D$ a $C' \subseteq C$. Aqui, d é um documento (um ponto ou um vetor) e C' é o subconjunto de categorias discretas. Classificar, portanto, nada mais é do que encontrar uma regra — matemática ou lógica — que divida esse espaço de forma que documentos de categorias diferentes fiquem em regiões distintas (Faceli et al., 2025).

Para o leitor que está iniciando na área, é útil pensar em um classificador através de duas fases fundamentais:

1. **Fase de treinamento (aprendizado):** o método analisa um conjunto de dados previamente rotulado (o *corpus* de treinamento, conforme discutido no Capítulo Conjunto de dados, *dataset* e *corpus*). Durante esse processo, o modelo tenta ajustar seus parâmetros internos para minimizar os erros de classificação, identificando quais características (palavras, contextos ou padrões) são os melhores indicadores de cada categoria.

2. **Fase de inferência (predição):** o modelo, agora "treinado", recebe um novo documento que ele nunca observou. Ele aplica o que aprendeu para projetar esse documento no espaço de categorias e decide qual rótulo é o mais adequado.

Ao longo das últimas décadas, a área de PLN testemunhou uma evolução fascinante nos métodos de classificação. Partimos de modelos baseados em intuições probabilísticas simples (Gonçalves; Quaresma, 2003; Pinto; Melgar, 2016), passamos por algoritmos que buscam fronteiras geométricas ideais (Gonçalves; Quaresma, 2003; Teixeira et al., 2018) e estruturas de árvores lógicas (Fonseca et al., 2014), até chegarmos às redes neurais profundas (Cortes et al., 2020; Matos et al., 2024) e aos modelos de linguagem de larga

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

escala (LLMs), que operam quase intuitivamente por meio de instruções (Gillin, 2024; Oliveira et al., 2024a).

Nesta seção, exploraremos esses métodos, dividindo-os em seus paradigmas fundamentais. Veremos que a escolha do método não depende apenas da precisão desejada, mas também da quantidade de dados disponíveis, da capacidade computacional e do nível de explicabilidade exigido pela aplicação — afinal, em domínios sensíveis como a Saúde (Capítulo 9) ou o Direito (Capítulo 12), entender "*por que*" um modelo tomou uma decisão pode ser tão importante quanto a decisão em si.

1.4.1 Métodos probabilísticos: Naïve Bayes

O método Naïve Bayes (NB) pertence à classe de modelos generativos e baseia-se na aplicação direta do Teorema de Bayes. Apesar de sua simplicidade matemática, ele é eficiente e serve como uma *baseline* robusta para tarefas como a filtragem de *spam* (Alberto et al., 2015; Almeida et al., 2011; Almeida et al., 2016), a identificação de *fake news* (Silva et al., 2020a) e *fake reviews* (Cardoso et al., 2018) e a classificação de tópicos (Gonçalves; Quaresma, 2003).

Para entender o NB, imagine que você está lendo um e-mail e encontra as palavras "promoção", "desconto" e "clique". Instantaneamente, sua mente aumenta a probabilidade de que este e-mail seja *spam*. Se, por outro lado, você encontrar palavras como "aula", "professor" e "prova", a probabilidade de que seja um e-mail legítimo cresce. O NB faz exatamente isso: ele conta a frequência com que cada palavra aparece em cada categoria durante o treino e utiliza essas contagens como "evidências" para decidir a classe de um novo texto.

Formalizando, dado um documento d (que visualizamos como um conjunto de palavras $w_1, w_2, \ldots, w_n$), queremos encontrar a classe c que maximiza a probabilidade $P(c|d)$. Pelo Teorema de Bayes, temos:

$$P(c|d) = \frac{P(d|c)P(c)}{P(d)} \tag{1.4}$$

Como o denominador $P(d)$ (a probabilidade do documento ocorrer) é constante para todas as classes, o classificador foca apenas no numerador:

1. $P(c)$ **(probabilidade a *priori*):** é a probabilidade de uma classe ocorrer antes mesmo de lermos o texto. Se 80% dos e-mails no seu treinamento são *spam*, $P(spam) = 0,8$.
2. $P(d|c)$ **(verossimilhança ou *likelihood*):** é a probabilidade de observarmos aquele conjunto de palavras dentro de uma classe específica.

A dificuldade computacional reside em calcular $P(d|c)$, ou seja, a probabilidade de uma sequência exata de palavras ocorrer. Para viabilizar o cálculo, o modelo assume a *premissa de independência*: a presença de uma palavra em um documento é totalmente independente da presença de outras palavras, dada a classe.

Sob essa suposição, a probabilidade conjunta torna-se o produto das probabilidades individuais:

$$P(d|c) \approx P(w_1|c) \times P(w_2|c) \times \cdots \times P(w_n|c) \tag{1.5}$$

Sabemos que esta premissa é falsa na linguagem natural — a palavra "Inteligência" aumenta muito a chance da palavra seguinte ser "Artificial". No entanto, mesmo sendo

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

"ingênuo", o modelo preserva a força relativa das evidências, o que o torna preciso na prática.

Ao implementar o NB, há dois desafios clássicos que precisam de atenção (Manning et al., 2008a):

- **O problema da divisão por zero (suavização de Laplace):** se uma palavra w nunca apareceu na classe "Esportes" durante o treino, sua probabilidade $P(w|Esportes)$ será 0. Como estamos multiplicando probabilidades, um único zero anularia todo o cálculo. Para resolver isso, adicionamos 1 ao contador de todas as palavras (técnica de suavização de Laplace), garantindo que nenhuma probabilidade seja nula.
- **Espaço de logaritmos:** multiplicar muitas probabilidades pequenas (ex.: $0,0001 \times 0,0002$) pode levar a erros de precisão numérica. Por isso, na prática, os logaritmos das probabilidades são somados: $\log(P(c|d)) \propto \log(P(c)) + \sum \log(P(w_i|c))$.

O NB é, portanto, um exemplo de como uma modelagem estatística simplificada pode oferecer excelente desempenho a baixo custo computacional, capaz de processar milhões de documentos em segundos (Freitas et al., 2019).

1.4.2 Métodos baseados em distância: K-vizinhos mais próximos

O método k-vizinhos mais próximos (do inglês, *K-Nearest Neighbors* – KNN) adota uma abordagem intuitiva e geométrica para a classificação (Cover; Hart, 1967). Diferente de outros métodos que tentam criar um modelo matemático abstrato sobre os dados, o KNN baseia-se na premissa de que "documentos semelhantes tendem a pertencer à mesma categoria".

Intuitivamente, imagine que você está em uma biblioteca imensa e encontra um livro sem capa nem título. Para descobrir o assunto dele, você olha para os livros guardados em posições imediatamente ao redor. Se a maioria dos vizinhos próximos trata de "Cálculo", é muito provável que o livro misterioso também pertença a essa categoria.

O KNN é classificado como um algoritmo de *aprendizado preguiçoso (lazy)*, pois não possui uma fase de treinamento explícita. Ele simplesmente armazena todos os exemplos de treinamento na memória. O verdadeiro trabalho ocorre apenas no momento da predição, quando o algoritmo compara o novo documento com todos os documentos armazenados.

O processo de classificação de um novo documento d segue três passos:

1. **Cálculo de distância:** o método calcula a distância (ou similaridade) entre o vetor de d e todos os vetores dos documentos do conjunto de treinamento.
2. **Seleção dos vizinhos:** o método identifica os k documentos mais próximos (vizinhos).
3. **Votação majoritária:** o método atribui, ao documento d, a classe que aparece com maior frequência entre esses k vizinhos.

No primeiro passo, em problemas de computação tradicionais (que envolvem dados tabulares), a distância euclidiana[5] é a mais utilizada. No entanto, em textos, ela apresenta uma

[5] A distância euclidiana entre dois vetores é computada como $d(u,v) = \sqrt{\sum_{i=1}^{n}(u_i - v_i)^2}$. Note que o comprimento do vetor não é levado em consideração nesta equação, assim, textos muito grandes, mas parecidos podem produzir similaridade baixa.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

falha crítica: a sensibilidade ao tamanho do documento. Dois textos sobre o mesmo tema, mas com tamanhos muito diferentes, teriam uma distância euclidiana elevada, mesmo sendo semanticamente idênticos. Por isso, na categorização de texto, preferimos a *similaridade de cosseno*, discutida no Capítulo Representação vetorial e semântica distribucional. Ela mede o ângulo entre dois vetores, considerando a orientação e não a magnitude. Isso garante que a comparação seja feita com base na proporção de palavras, tornando o modelo invariante ao tamanho do texto.

A escolha do valor do hiperparâmetro k (para o segundo passo) é o principal desafio deste método. Geralmente, utiliza-se um valor de k ímpar para evitar empates em classificações binárias:

- Se $k = 1$, o modelo fica muito sensível a ruídos ou erros de rotulação no treino (um vizinho "errado" define a classe).
- Se k for muito grande, o modelo pode perder a capacidade de capturar nuances locais e pode ser dominado pelas classes majoritárias (aquelas que possuem mais documentos no total).

O KNN é valorizado por ser não-paramétrico, o que significa que ele não assume nenhuma distribuição específica para os dados — ele se adapta perfeitamente a fronteiras de decisão complexas e irregulares. Contudo, sua principal limitação no contexto de PLN é a eficiência. Como ele precisa comparar cada novo documento a toda a base de dados, o tempo de resposta cresce linearmente com o tamanho do *corpus*. Para aplicações que exigem baixa latência ou lidam com milhões de documentos, o custo computacional e de memória pode ser proibitivo, o que exige o uso de estruturas de dados avançadas para busca aproximada de vizinhos (como *Ball Trees* (Omohundro, 1989) ou *Locality Sensitive Hashing* (Indyk; Motwani, 1998)).

1.4.3 Métodos de otimização: Regressão Logística e SVM

Enquanto o Naïve Bayes computa probabilidades e o KNN calcula vizinhanças, os métodos de otimização tentam encontrar uma fronteira de decisão — uma linha, plano ou hiperplano — que separe as categorias de forma ótima no espaço de características. Nesta seção, exploraremos dois dos métodos desse paradigma mais utilizados em PLN: a Regressão Logística e as Máquinas de Vetores de Suporte (SVM) (Faceli et al., 2025).

Apesar do nome sugerir uma tarefa de regressão, a **Regressão Logística** é um dos métodos de classificação mais robustos e interpretáveis da literatura. No contexto de PLN, ele é frequentemente utilizado para entender quais palavras ou características têm maior peso na decisão de uma categoria. Diferente da Regressão Linear, que pode retornar qualquer valor, a Regressão Logística utiliza a função logística (ou sigmoide) para ajustar o resultado da soma ponderada ao intervalo entre 0 e 1, representando a probabilidade de pertencer à classe (Cox, 1958).

Como destacado por Faceli et al. (2025), a vantagem deste método é a sua interpretabilidade: ao analisarmos os coeficientes (pesos) aprendidos, conseguimos explicar exatamente por que um documento foi classificado em determinada categoria, o que é essencial em domínios como o Direito (Capítulo 12).

As **Máquinas de Vetores de Suporte** (do inglês, *Support Vector Machines* – SVM) foram, durante anos, o estado da arte na categorização de texto antes da ascensão das

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

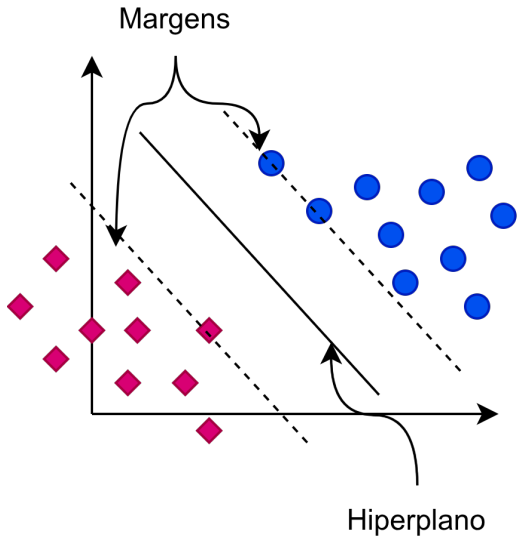
©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

redes neurais profundas. O SVM é particularmente eficaz quando lidamos com espaços de alta dimensionalidade, como no caso das representações TF-IDF discutidas na Seção 1.3).

Intuitivamente, imagine que as categorias sejam duas cidades separadas por uma floresta. O SVM não quer apenas desenhar uma linha divisória entre elas; quer construir a estrada mais larga possível que as separe. A linha central dessa estrada é chamada *hiperplano de decisão*, e as bordas da estrada são chamadas de *margens* (Figura 1.3).

O objetivo do SVM é maximizar essa margem. Quanto maior o espaço entre a fronteira e os documentos mais próximos de cada classe, maior será a capacidade do modelo de classificar novos documentos corretamente (generalização) (Cortes; Vapnik, 1995).

Figura 1.3: Máquina de vetores suporte com o hiperplano de separação entre os dados



Um detalhe didático do SVM é que a decisão final não depende de todos os documentos de treino. Ela depende apenas dos documentos que estão "na beira da estrada", ou seja, nos limites das margens. Esses documentos são os *vetores de suporte*. Se removermos qualquer outro documento do treino, a fronteira permanece igual; se movermos um vetor de suporte, a fronteira muda.

Muitas vezes, os textos não são linearmente separáveis (não é possível traçar uma linha reta que os divida perfeitamente). O SVM resolve isso ao projetar os dados em um espaço de maior dimensão. Essa transformação matemática é realizada de forma eficiente por meio de funções de *kernel*, permitindo que o modelo capture relações não lineares complexas entre as palavras.

No uso prático em PLN:

- A Regressão Logística é preferida quando a velocidade de treinamento e a interpretabilidade dos pesos das palavras são cruciais.
- O SVM é a escolha ideal quando o volume de dados é menor, mas a dimensionalidade é muito alta, oferecendo uma robustez matemática superior contra o sobreajuste (*overfitting*).

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

1.4.4 Métodos baseados em árvores e *ensembles*

Enquanto os modelos anteriores buscam equações matemáticas ou vizinhanças espaciais, os métodos baseados em árvores adotam uma abordagem de divisão e conquista. Eles particionam o espaço de características por meio de uma sequência de perguntas lógicas, criando uma estrutura que se assemelha a um fluxograma de decisão (Faceli et al., 2025).

Uma **Árvore de Decisão** (Quinlan, 1993) é uma estrutura onde cada nó interno representa um teste sobre uma característica do texto (ex.: "A palavra lucro está presente?"). Cada ramo da árvore representa o resultado do teste (Sim ou Não), e cada nó-folha representa o rótulo da categoria final.

Pense na classificação como um jogo de adivinhação. O modelo tenta fazer a pergunta mais informativa primeiro para eliminar o maior número possível de incertezas. Em um *corpus* de notícias, a pergunta "O texto contém o termo campeonato?" pode separar quase instantaneamente a categoria "Esportes" das demais.

Para decidir qual palavra usar em cada nó, a árvore utiliza medidas estatísticas de desordem[6]. A mais comum é a *entropia*. Se um conjunto de documentos contém textos de várias categorias misturadas, a entropia é alta. O objetivo do algoritmo é escolher uma divisão que resulte em subconjuntos mais "puros" (de menor entropia), aumentando o *Ganho de Informação*.

A maior vantagem desse método é a explicabilidade: qualquer pessoa consegue seguir o caminho do topo da árvore até a folha para entender a decisão. O risco, porém, é o sobreajustamento: a árvore pode se tornar tão profunda e específica que acaba "decorando" o ruído do treino em vez de aprender padrões gerais. Para mitigar as falhas de uma única árvore, normalmente utilizam-se métodos de ***ensemble*** (conjuntos), que combinam as predições de várias árvores para obter um resultado mais robusto e estável. Duas estratégias neste sentido são a Floresta Aleatória e os métodos de *boosting*.

Uma **Floresta Aleatória** cria centenas de árvores de decisão distintas. O segredo de sua eficácia é a diversidade:

- Cada árvore é treinada em um subconjunto aleatório dos dados (*bagging*);
- Em cada nó, apenas um subconjunto aleatório de palavras é considerado para a divisão.

A decisão final é tomada por votação majoritária. Como destacado por Breiman (2001), essa colaboração entre árvores independentes reduz drasticamente o erro e torna o modelo um dos mais confiáveis para características extraídas manualmente, como as medidas de complexidade textual discutidas no Capítulo 6.

Diferentemente das florestas aleatórias, nas quais as árvores crescem em paralelo, no *boosting* as árvores são construídas sequencialmente. A lógica é: a primeira árvore faz uma predição inicial (que pode ser fraca); a segunda é treinada especificamente para corrigir os erros da primeira, e assim por diante. O **XGBoost** (Chen; Guestrin, 2016) é uma implementação altamente otimizada desse processo. Ele lida nativamente com dados esparsos (característica comum em BoW/TF-IDF), o que justifica sua alta performance no

[6] Aqui desordem refere-se a um conceito estabelecido na Teoria da Informação chamado de Entropia de Shannon, que visa quantificar a "impureza" ou a incerteza dos dados (Faceli et al., 2025). Em uma tarefa de classificação de texto, um conjunto de dados com alta "impureza" significa que os elementos de diversas categorias estão misturados e um modelo busca organizar esta "impureza", tornando os conjuntos mais homogêneos.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

PLN tradicional, frequentemente vencendo competições de ciência de dados. Em PLN, ele se destaca quando há vetores de características bem definidos ou metadados estruturados.

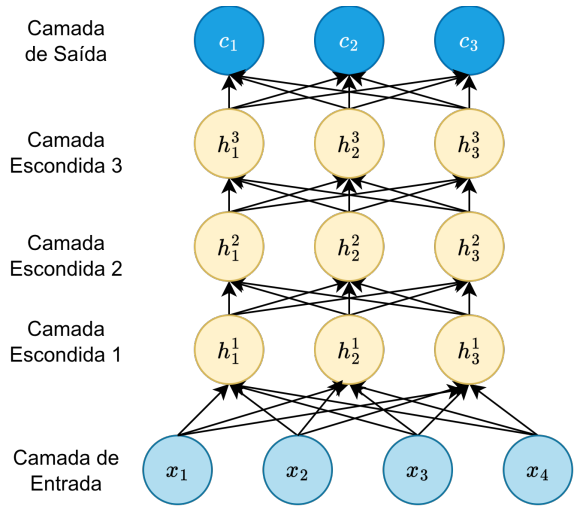
Embora as árvores tenham dificuldade em lidar diretamente com vetores TF-IDF de altíssima dimensão (em que a maioria dos valores é zero), elas são altamente eficientes quando o problema envolve a combinação de texto com dados estruturados (e.g. classificar um anúncio com base no texto da descrição e no preço do produto) ou quando a interpretabilidade das regras de negócio é um requisito legal.

1.4.5 Métodos conexionistas: Redes neurais e *Deep learning*

Os métodos conexionistas, inspirados na estrutura biológica do cérebro, representam uma mudança de paradigma na categorização de texto. Enquanto os métodos anteriores dependem de uma representação de entrada fixa, como o TF-IDF, as redes neurais são capazes de aprender características (*feature learning*) e extrair padrões complexos e semânticos diretamente dos dados brutos. A unidade básica desta abordagem é o neurônio artificial, que realiza uma soma ponderada de suas entradas e aplica uma função de ativação não linear (Faceli et al., 2025). Para a categorização de texto, o modelo mais simples é o *Multilayer Perceptron* (MLP) (Figura 1.4).

Intuitivamente, imagine que a primeira camada da rede observa palavras individuais, a segunda camada combina essas palavras em sentenças curtas e a terceira camada identifica conceitos abstratos (como "ironia" ou "urgência"). Ao final, a rede decide a categoria com base nessas abstrações aprendidas durante o treinamento.

Figura 1.4: Exemplo de uma rede neural simples



Em vez de contagens esparsas, as redes neurais frequentemente utilizam camadas de *embeddings* (como detalhado no Capítulo Representação vetorial e semântica distribucional). Essas camadas transformam palavras em vetores densos que capturam afinidade semântica. A rede ajusta seus pesos por meio do algoritmo de retropropagação (*backpropagation*), minimizando o erro entre a predição e o rótulo real. A seguir uma breve explicação das principais arquiteturas existentes.

Redes neurais convolucionais (CNN). Originalmente propostas para o processamento de imagens, as CNNs foram adaptadas com sucesso para o processamento de texto (Kim, 2014). Elas funcionam como "detectores de n-gramas" automáticos.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Uma CNN desliza janelas (filtros) sobre o texto. Um filtro pode se especializar em detectar a expressão "muito caro", enquanto outro pode focar em "excelente atendimento". A rede não se importa de onde essas expressões aparecem no documento; ela apenas registra que ocorreram e as utiliza para classificar o sentimento ou o tópico.

Redes recorrentes (RNN) e LSTMs. A linguagem humana é inerentemente sequencial e a ordem das palavras altera o sentido (e.g. "O gato caçou o rato" *vs* "O rato caçou o gato"). As Redes Neurais Recorrentes (RNNs) foram projetadas para lidar com essa característica, mantendo um "estado oculto" que serve como uma memória do que foi lido anteriormente.

Como as RNNs tradicionais sofrem com o esquecimento de informações distantes, a arquitetura *Long Short-Term Memory* (LSTM) (Hochreiter; Schmidhuber, 1997) foi proposta para resolver esse problema introduzindo "portões" (*gates*) que controlam o que deve ser lembrado e o que deve ser descartado. A principal vantagem é que elas conseguem entender que uma negação no início de um parágrafo longo afeta o sentido de uma palavra no final. Devido ao alto desempenho em diversas tarefas de PLN, elas foram o padrão-ouro para categorização de documentos longos até a consolidação dos *Transformers*.

1.4.6 Modelos de linguagem

A grande vantagem desses métodos é a flexibilidade. Podemos combinar CNNs e LSTMs para criar modelos híbridos que capturam tanto padrões locais quanto dependências globais. No entanto, o treinamento dessas redes do zero exige grandes volumes de dados rotulados. Essa limitação preparou o terreno para o uso de modelos pré-treinados, onde redes neurais profundas (os *Transformers*) aprendem a língua em *corpora* massivos antes de serem ajustadas para a tarefa de categorização, como veremos na próxima seção e no Capítulo Modelos de linguagem.

1.4.7 Redes com atenção: Transformers

Se as CNNs focam em padrões locais e as RNNs em sequências, a arquitetura *Transformers* (ou rede de transformadores) (Vaswani et al., 2017a) introduziu uma abordagem que mudou permanentemente o PLN: o mecanismo de auto-atenção (do inglês, *self-attention*). Esta arquitetura abandonou a recorrência e a convolução em favor de uma estrutura que processa todas as palavras de um documento simultaneamente.

Imagine que, ao ler uma frase, você pudesse projetar um "holofote" sobre as palavras mais importantes para entender o sentido de uma palavra específica. Na frase "O banco da praça estava quebrado", ao ler a palavra "banco", seu cérebro se concentra em "praça" para entender que se trata de um assento, e não de uma instituição financeira. Uma rede de transformadores faz exatamente isso através da autoatenção: ela calcula o peso (a importância) de cada palavra do texto em relação a todas as outras, criando uma representação contextualizada.

Para a tarefa de categorização, o modelo mais emblemático é o BERT (*Bidirectional Encoder Representations from Transformers*) (Devlin et al., 2019). Diferentemente de modelos que leem o texto da esquerda para a direita, o BERT lê o texto em ambas as direções simultaneamente, o que lhe permite capturar o contexto completo de cada termo.

Para classificar um texto com o BERT, utilizamos um procedimento específico:

1. **O *token* [CLS]:** adicionamos um *token* especial chamado `[CLS]` (de *classification*) no início de cada documento;

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

2. **Agregação de sentido:** após passar por todas as camadas de atenção da rede, o vetor resultante deste *token* `[CLS]` funciona como um "resumo" matemático de todo o documento;

3. **Camada de saída:** adicionamos uma camada densa simples (como uma Regressão Logística) que processa este vetor para determinar a categoria final.

Diferente das LSTMs, que processam o texto palavra por palavra e podem "esquecer" o início de documentos longos, as redes de transformadores mantêm o acesso direto a todas as palavras ao mesmo tempo. Além disso, por não ser uma arquitetura sequencial, no treinamento é possível realizar o processamento paralelo em massa, o que viabiliza o treinamento em *corpora* de escala global (como discutido no Capítulo Modelos de linguagem).

Este método é hoje o ponto de partida para qualquer sistema de categorização de alto desempenho, servindo de base para os modelos que exploraremos na próxima seção.

1.4.8 Classificação via *prompt* e modelos de linguagem de larga escala (LLMs)

A chegada dos modelos de linguagem de larga escala (LLMs), como as famílias GPT, Llama e a brasileira MariTalk (detalhadas no Capítulo ChatGPT, MariTalk e outros agentes de conversação), introduziu uma nova forma de classificar textos. Nos métodos anteriores, o foco era treinar um modelo específico para uma tarefa; aqui, o foco passa a ser a instrução de um modelo de propósito geral.

Imagine que, em vez de construir uma máquina complexa para separar e-mails, você pudesse simplesmente entregar um documento a um assistente com amplo conhecimento e solicitar:

Quadro 1.2: Exemplo de instrução com *zero-shot learning* para a categorização de um texto

> "Você é um jornalista encarregado de classificar uma dada notícia em uma das seguintes categorias: política, economia, esporte e policial. Agora, considere a notícia a seguir e responda apenas em qual categoria ela melhor se encaixa."
> {texto da notícia}

Neste caso, o assistente usa todo o seu conhecimento prévio sobre o mundo e a língua para tomar a decisão, sem precisar que você mostre exemplos antes.

No paradigma de *prompting*, não alteramos os pesos do modelo (como fazemos no ajuste fino do BERT). Em vez disso, fornecemos um texto de entrada (o *prompt*) que contém a descrição da tarefa e o documento a ser classificado.

O principal mecanismo por trás desta abordagem é a capacidade do modelo de realizar o que chamamos de aprendizado em contexto. Existem duas formas principais de aplicar isso à categorização:

- ***Zero-shot learning* (Aprendizado sem exemplos):** o modelo recebe apenas a instrução e o texto. Exemplo: "Classifique o sentimento deste tweet como Positivo ou Negativo: 'O suporte técnico resolveu meu problema em minutos!' ".

- ***Few-shot learning* (Aprendizado com poucos exemplos):** o modelo recebe dois ou três exemplos de textos já classificados antes de solicitar a classificação do

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

novo documento. Isso ajuda o modelo a entender o formato esperado e as sutilezas das categorias (Brown et al., 2020b).

Quando possível, a inclusão de bons exemplos (*few-shot learning*) fornece mais contexto para que o modelo realize a inferência e, portanto, costuma oferecer resultados superiores aos da abordagem *zero-shot learning*. Veja o exemplo a seguir:

Quadro 1.3: Exemplo de instrução com *few-shot learning* para a categorização de um texto

> "Você está analisando textos curtos de redes sociais para determinar se expressam sentimento positivo, negativo ou neutro. Os textos são relacionados ao esporte e podem conter linguagem informal e ironia. Veja, a seguir, alguns casos e a classificação correta de cada um.
> Caso 1) Ninguém segura o Mengão \o/
> Resposta 1) Positivo.
> Caso 2) Mais uma vez o VAR!!! Ninguém aguenta mais juiz!!!!
> Resposta 2) Negativo.
> Caso 3) Mais uma partida e estamos aqui aguardando...
> Resposta 3) Neutro."
>
> {texto a ser classificado}

Veja que, neste exemplo, foram fornecidos três casos diferentes, um para cada classe. Dessa forma, o modelo pode compreender melhor a tarefa e oferecer um poder preditivo maior. Contudo, é importante destacar que a ordem dos exemplos no *prompt* pode influenciar a resposta do modelo (problema conhecido como viés de recência), o que constitui um desafio prático real da engenharia de *prompt*.

Para que os modelos se tornassem bons classificadores por meio de *prompt*, passaram por um processo chamado *instruction tuning* (Ouyang et al., 2022). Como discutido no Capítulo Modelos de linguagem, os modelos de linguagem básicos (da primeira geração) são treinados apenas para prever a próxima palavra. Modelos modernos (de larga escala) foram ajustados para seguir ordens humanas, o que os torna capazes de entender restrições complexas na categorização, como: "Classifique este contrato jurídico, mas se houver ambiguidade, atribua o rótulo 'Revisão Necessária' ".

Esta abordagem resolve um dos maiores gargalos do PLN: *a necessidade de dados rotulados por humanos*. É possível criar um classificador funcional em minutos. No entanto, é preciso ficar atento a dois pontos críticos:

1. **Custo e latência:** rodar um LLM para classificar milhões de documentos é muito mais caro e lento do que usar um Naïve Bayes ou um SVM, por exemplo.
2. **Instabilidade do *prompt*:** pequenas mudanças na forma como escrevemos a instrução podem gerar resultados distintos, um fenômeno que exige técnicas de engenharia de *prompt* (Kojima et al., 2022).

A categorização via LLMs é ideal para prototipagem rápida, tarefas com categorias altamente subjetivas ou quando não temos dados de treino suficientes para os métodos conexionistas clássicos (Brown et al., 2020b; Wei et al., 2022).

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

1.4.9 Síntese e Recomendação

Para facilitar a escolha do método mais adequado, de acordo com a natureza do problema e os recursos disponíveis, o Quadro 1.4 apresenta uma síntese comparativa dos paradigmas discutidos nesta seção.

Quadro 1.4: Quadro comparativo dos métodos de classificação de texto

Método	Características principais	Aplicações recomendadas
Naïve Bayes	Baixo custo computacional; assume independência entre os termos; robusto com poucos dados.	Filtragem de *spam*; classificação de tópicos em tempo real; *baselines* rápidas.
KNN	Não-paramétrico; sem fase de treino (*lazy*); sensível à métrica de distância e ao ruído.	Busca por documentos similares; sistemas de recomendação; bases de dados pequenas e dinâmicas.
SVM & Reg. Logística	Fronteiras geométricas claras; excelente em alta dimensionalidade; interpretável (pesos dos termos).	Análise de sentimentos; classificação jurídica e médica; cenários com alta dimensionalidade (TF-IDF).
Árvores & *Ensembles*	Baseado em regras lógicas; lida bem com dados heterogêneos; alta precisão (XGBoost).	Avaliação de legibilidade; classificação que combina texto com metadados estruturados.
***Deep Learning* (BERT)**	Captura contexto e ordem; exige alto poder computacional (GPU); estado da arte em precisão.	Categorização de alta complexidade semântica; tarefas que exigem compreensão profunda do contexto.
LLMs (*Prompt*)	*Zero/Few-shot*; baseados em instruções; dispensam grandes bases de dados rotuladas.	Prototipagem rápida; categorias subjetivas ou mutáveis; cenários sem dados de treino disponíveis.

1.5 Avaliação de modelos

A avaliação de um classificador de texto é o estágio em que verificamos se o aprendizado na fase de treinamento (Seção 1.4) de fato resultou em um modelo capaz de generalizar para dados nunca vistos. Como discutido no Capítulo Avaliação de tecnologias de linguagem, a escolha de uma métrica isolada pode ser enganosa. Portanto, é necessário um protocolo de validação robusto para garantir que o desempenho reportado seja estatisticamente confiável e livre de vícios.

1.5.1 Metodologias de validação

O princípio fundamental da avaliação é a separação entre os dados usados para o aprendizado e os usados para a medição do desempenho. Caso o modelo seja testado com os mesmos dados de treinamento, estaríamos medindo apenas sua capacidade de memorização (*overfitting*) e não de categorização.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

1.5.1.1 Divisão por repartição (*Hold-out*)

A forma mais comum de validação, especialmente em grandes *corpora* ou ao utilizar modelos de *Deep Learning* (Seção 1.4.5), é a divisão simples em subconjuntos. A literatura recomenda, usualmente, uma proporção de 80% para treino e 20% para teste (Murphy, 2012), embora variações como 70/30 também sejam frequentes.

Para um ajuste fino de modelos, como a escolha de hiperparâmetros (o valor de k no KNN ou o *learning rate* em LLMs), subdividimos o treinamento para criar um conjunto de Validação. Assim, temos três divisões do conjunto de dados:

- **Treinamento**: partição dos dados utilizada para ajustar os pesos do classificador.
- **Validação**: partição dos dados usada para escolher a melhor configuração do modelo.
- **Teste**: partição dos dados utilizada para a avaliação final e definitiva do desempenho.

1.5.1.2 Validação cruzada (*Cross-Validation*)

Quando o conjunto de dados é pequeno, a divisão simples pode gerar resultados instáveis dependendo de quais documentos caíram no teste. A Validação Cruzada de k partições (*k-fold cross-validation*) resolve isso ao dividir os dados em k subconjuntos (*folds*). O experimento é repetido k vezes: em cada rodada, uma partição diferente é usada para teste e as outras $k-1$ para treinamento.

Por exemplo, em um *corpus* de 500 notícias dividido em 5 partições ($k = 5$), cada partição teria 100 documentos. Na primeira rodada, a Partição 1 é o teste e as outras 4 partições são usadas para treinamento; na segunda, a Partição 2 é o teste, e assim sucessivamente. Ao final, reportamos a média e o desvio padrão dos resultados, o que nos dá uma visão da variância e da estabilidade do modelo frente a diferentes recortes dos dados.

1.5.1.3 Estratificação e Significância

Um aspecto crítico no PLN é o desequilíbrio de classes. Em um portal de notícias, pode haver muito mais textos sobre "Política" do que sobre "Astronomia". Se a divisão dos dados for feita de forma puramente aleatória, corremos o risco de a partição de teste não conter nenhum exemplo de uma classe rara. Para evitar isso, utilizamos o *Particionamento Estratificado*, que garante que a proporção original de cada categoria seja mantida em todos os subconjuntos de treino e de teste.

Para comparar se um modelo é genuinamente superior a outro, além da média, utilizamos técnicas como o *bootstrapping*. Nesta técnica, geramos múltiplas amostras aleatórias com reposição a partir do conjunto original (Efron; Tibshirani, 1993). Isso nos permite calcular intervalos de confiança e afirmar, com suporte estatístico, se a diferença de desempenho entre, por exemplo, um SVM e um Naïve Bayes, não ocorreu meramente ao acaso.

1.5.2 Métricas de desempenho

Para quantificar a eficácia, utilizamos métricas derivadas da *Matriz de Confusão*. Ela é uma tabela que cruza as predições do modelo com os rótulos reais, conforme ilustrado na Tabela 1.3.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Tabela 1.3: Exemplo de matriz de confusão para três classes (A, B e C).

Classe Real \Predita	**Classe A**	**Classe B**	**Classe C**
Classe A	50 (VP)	2 (FN)	3 (FN)
Classe B	4 (FP)	45	6
Classe C	1 (FP)	5	49

A partir da matriz, definimos três conceitos fundamentais para cada classe:

- **Verdadeiro Positivo (VP)**: documentos da classe A corretamente classificados como A.
- **Falso Positivo (FP)**: documentos de outras classes classificados erroneamente como A ("falso alarme").
- **Falso Negativo (FN)**: documentos que eram da classe A, mas o modelo atribuiu-os a outra categoria ("omissão").

Idealmente, a matriz de confusão deve conter a maior quantidade de instâncias na diagonal principal, indicando que o modelo classificou a maior parte das instâncias na classe correta. A partir da matriz de confusão é possível definir a **acurácia**, uma das métricas mais usadas para avaliar o desempenho de um modelo em classificação. O objetivo da acurácia é medir a taxa global de acertos a partir da seguinte equação $Acurácia = \frac{VP+VN}{VP+VN+FP+FN}$.

1.5.2.1 Precisão, Revocação e Medida-F

Diferente da *Acurácia*, que mede apenas a taxa global de acertos, as seguintes métricas são mais informativas:

1. **Precisão (P)**: proporção de acertos nas predições de uma classe. Indica a confiabilidade do modelo ao atribuir um rótulo.

$$P = \frac{VP}{VP + FP}$$

2. **Revocação (R)**: proporção de documentos de uma classe que o modelo conseguiu recuperar. Indica a abrangência (ou cobertura) do modelo.

$$R = \frac{VP}{VP + FN}$$

3. **Medida-F1 (F_1)**: média harmônica entre Precisão e Revocação, equilibrando os dois aspectos. A média harmônica (e não a aritmética) é utilizada para punir valores extremos. Se um modelo tem Precisão 1.0 e Revocação 0.0, a média aritmética seria 0.5 (enganosa), enquanto a média harmônica seria 0 (realista).

$$F1 = 2 \times \frac{P \times R}{P + R}$$

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

1.5.2.2 Agregação: Macro e Micro

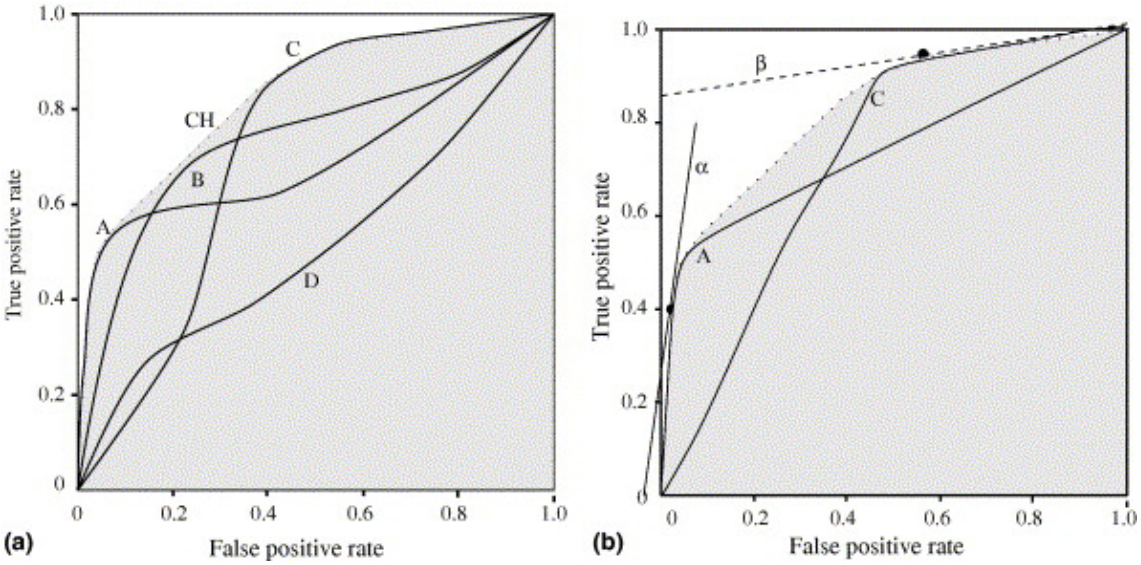
Em problemas multiclasse, existem duas formas de resumir o desempenho global:

- **Macro-Média**: calcula a métrica para cada classe individualmente e calcula a média simples. Trata todas as classes como igualmente importantes, sendo a melhor escolha para avaliar o desempenho em classes raras.
- **Micro-Média**: agrega as contagens globais de VP, FP e FN de todas as classes antes do cálculo. É dominada pelas classes mais frequentes, o que reflete o desempenho médio por documento.

1.5.3 Análise via Curva ROC e AUC

A Curva ROC (*Receiver Operating Characteristics*) é uma ferramenta visual para analisar o desempenho de classificadores probabilísticos (como a Regressão Logística ou o Naïve Bayes), variando o limiar de decisão t (Fawcett, 2006). A curva coloca em gráfico a Taxa de Verdadeiros Positivos ($TPR = \frac{VP}{VP+FN}$) no eixo Y e a Taxa de Falsos Positivos ($FPR = \frac{FP}{FP+VN}$) no eixo X. Um classificador perfeito alcançaria o canto superior esquerdo (100% de acerto com 0% de alarme falso). A Figura 1.5 ilustra dois exemplos de curvas ROC de diferentes modelos extraídos de (Fawcett, 2006). Visualmente, quanto mais próxima a curva estiver do canto superior esquerdo do gráfico, melhor o desempenho do modelo, pois isso indica uma alta taxa de verdadeiros positivos e uma baixa taxa de falsos positivos. Na Figura 1.5 (a), por exemplo, nota-se que as curvas de diferentes classificadores se cruzam; isso implica que um modelo pode ser superior a outro dependendo do limiar de decisão escolhido ou das restrições de custo do problema. Já a Figura 1.5 (b) demonstra o conceito de *Convex Hull*, que representa o limite superior de desempenho possível ao combinar diferentes modelos.

Figura 1.5: Exemplos de Curva ROC



Fonte: (Fawcett, 2006)

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Para resumir essa curva em um único valor, utilizamos a AUC (*Area Under the Curve*). Um valor de $1,0$ indica separação perfeita, enquanto $0,5$ indica um desempenho equivalente ao acaso. A AUC é especialmente útil em PLN para comparar modelos em tarefas de detecção binária, como a identificação de notícias falsas Capítulo 8.

1.6 Recursos em português

A aplicação prática dos métodos de categorização para a língua portuguesa exige recursos capazes de captar as nuances morfossintáticas e semânticas da nossa língua. Como discutido no Capítulo Conjunto de dados, *dataset* e *corpus*, a qualidade do modelo é diretamente proporcional à qualidade dos dados de treino. Nesta seção, detalhamos os principais conjuntos de dados, modelos pré-treinados e ferramentas disponíveis para o processamento do português brasileiro.

1.6.1 Conjuntos de dados para categorização

A comunidade brasileira de PLN tem sido prolífica na criação de recursos que abrangem desde a classificação de tópicos tradicionais até tarefas complexas de inferência e detecção de viés. A Tabela 1.4 sintetiza alguns desses recursos, detalhando a tarefa, o número de classes e, quando disponível, o desempenho de referência.

Tabela 1.4: Lista de recursos em português para classificação.

Nome	Tarefa	Classes	Resultados
CHAVE (Santos; Rocha, 2004a)	Classificação multirrótulo de notícias	9	–
Fake.Br (Monteiro et al., 2018)	Classificação de notícias (falsas/reais)	2	Acurácia/F1: 0,950 (Pires et al., 2024)
B2W-Reviews01 (Real et al., 2019)	Críticas de consumidores (escalas de 1 a 5)	5	–
UTLCorpus (Sousa et al., 2019)	Utilidade de críticas (apps e filmes)	2	ROC-AUC: 97,1 (apps) / 92,4 (filmes) (Souza; Filho, 2022)
Decisões Corte (Lage-Freitas et al., 2019)	Sumários judiciais com decisões	3	F1: 0,80
ASSIN 2 (Real et al., 2020)	Inferência textual (premissa/hipótese)	2	F1: 0,89 (Silva et al., 2023b)
HateBR (Vargas et al., 2022a)	Discurso de ódio no Instagram	2	F1: 0,85
FactNews (Vargas et al., 2023)	Viés vs. Factualidade e tipo de viés	2 / 12	F1: 0,88 / 0,67
LegiSubject-Br (Nunes, 2024)	Projetos de lei (multirrótulo)	59	–

No domínio "Jornalístico e Combate à Desinformação", o CHAVE (Santos; Rocha, 2004a) é um marco histórico no português, oferecendo um desafio de classificação multirrótulo em notícias. No cenário contemporâneo de combate à desinformação (Capítulo 8), o Fake.Br (Monteiro et al., 2018) consolidou-se como um recurso essencial, apresentando resultados de alto desempenho com modelos modernos, como indicado por Pires et al. (2024).

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Em "Análise de Sentimento e Consumo", para o estudo de opiniões, o B2W-Reviews01 (Real et al., 2019) oferece uma granularidade de 5 classes (estrelas), permitindo avaliar a satisfação do cliente. Complementarmente, o UTLCorpus (Sousa et al., 2019) foca não no conteúdo da crítica em si, mas em sua utilidade (*helpfulness*), um problema de classificação binária que ajuda a filtrar informações relevantes em plataformas de aplicativos e de filmes.

Na área de "Inferência e Fenômenos Sociais", o ASSIN 2 (Real et al., 2020) desafia os modelos a identificar relações de inferência textual, essenciais para tarefas de semântica profunda. No campo das redes sociais, o HateBR (Vargas et al., 2022a) fornece dados para a detecção de discurso de ódio no Instagram, enquanto o FactNews (Vargas et al., 2023) permite classificar o viés em frases, indo além da polaridade para identificar tipos específicos de viés jornalístico.

No "Domínio Jurídico e Legislativo", os recursos "Decisões da Corte Brasileira" (Lage-Freitas et al., 2019) e LegiSubject-Br-Keywords (Nunes, 2024) são fundamentais para a automação e organização do sistema jurídico. O último destaca-se pela alta complexidade, envolvendo 59 categorias multirrótulo para indexação de projetos de lei.

1.6.2 Modelos de linguagem pré-treinados

A revolução dos Transformers (Seção 1.4.7) trouxe modelos treinados em volumes massivos de texto em português, fundamentais para os desempenhos reportados na Tabela 1.4.

- **BERTimbau** (Souza et al., 2020a): o modelo BERT (Devlin et al., 2019) padrão para o português brasileiro. Treinado no *corpus* brWaC (Wagner Filho et al., 2018), ele é a base para a maioria dos classificadores de alto desempenho da atualidade.
- **PTT5** (Carmo et al., 2020a): uma versão do T5 (Raffel et al., 2020) que permite abordar a classificação como um problema de geração de texto para texto.
- **MariTalk**[7]: representa a nova geração de LLMs nacionais, facilitando a classificação via *prompt* (Seção 1.4.8).

1.6.3 Ferramentas e Bibliotecas

Para implementar classificadores com esses recursos, são comumente empregadas as bibliotecas *Hugging Face Transformers* (para modelos neurais)[8], *spaCy* (para modelos industriais prontos)[9] e *Scikit-learn* (para métodos clássicos e métricas de avaliação)[10].

1.7 Desafios e Tendências

Embora a categorização de texto seja uma das áreas mais maduras do PLN, o cenário contemporâneo – marcado pela onipresença de modelos de linguagem de larga escala (LLMs) e pela aplicação de IA em domínios críticos – impõe novos desafios. Nesta seção, discutimos as fronteiras da pesquisa e as tendências que moldarão a próxima geração de sistemas de categorização.

[7] https://chat.maritaca.ai/
[8] pip install transformers
[9] https://spacy.io/
[10] https://scikit-learn.org/stable/

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

1.7.1 Dinamismo e Desvio de Conceito (*Concept Drift*)

Modelos de classificação não são estáticos; operam em um mundo em que a linguagem e os conceitos evoluem. O *Desvio de Conceito* ocorre quando a relação entre as características do texto e as categorias muda ao longo do tempo. Imagine um classificador de notícias políticas treinado em 2010. Seus padrões de vocabulário e de nomes de entidades estarão obsoletos para classificar o cenário de 2026. Em redes sociais, gírias e neologismos surgem semanalmente, exigindo que o sistema de categorização possua estratégias de *Aprendizado Contínuo* para se adaptar, sem necessidade de ser retreinado do zero constantemente (Silva et al., 2022b; Silva; Almeida, 2021).

1.7.2 Escalabilidade e Eficiência energética

A escalabilidade permanece como um dos desafios centrais na categorização de textos, especialmente diante da necessidade de processar documentos cada vez mais extensos e volumes massivos de dados (Minaee et al., 2021). Embora a evolução dos modelos de linguagem de larga escala (LLMs) tenha permitido a expansão da janela de contexto, observa-se que a eficácia preditiva e a performance do modelo tendem a degradar à medida que o tamanho da entrada aumenta (Dong et al., 2024).

Paralelamente ao desafio técnico, o debate sobre a **IA Verde** (*Green AI*) tornou-se fundamental para a sustentabilidade da área (Schwartz et al., 2020). Sistemas que utilizam modelos com bilhões de parâmetros exigem infraestrutura computacional de alto desempenho e um consumo energético que pode ser proibitivo para muitas instituições. Mesmo modelos recentes e otimizados para tarefas de raciocínio, como os da família Qwen (Bai et al., 2023), ainda demandam um consumo de energia considerável, evidenciando que a busca por um equilíbrio entre alto desempenho e eficiência operacional é a nova fronteira no desenvolvimento de tecnologias de linguagem.

1.7.3 Explicabilidade e Interpretabilidade (XAI)

À medida que migramos de modelos lineares para redes neurais profundas e LLMs, ganhamos em precisão, mas perdemos em transparência. Em domínios sensíveis como a saúde ou o direito, uma classificação sem justificativa é insuficiente. A área de *IA Explicável (XAI)* busca desenvolver métodos que permitam entender *por que* um modelo atribuiu um rótulo específico. Técnicas como mapas de atenção (que mostram quais palavras o Transformer priorizou) ou métodos de atribuição de características (como LIME (Ribeiro et al., 2016) ou SHAP (Lundberg; Lee, 2017)) são essenciais para que o especialista humano possa confiar nas decisões automáticas e auditar as decisões automáticas. Embora para texto, estudos indicam que estes métodos se mostram pouco confiáveis (Létoffé et al., 2025). Modelos transparentes são um objetivo que muitos pesquisadores desejam e, portanto, ainda um caminho aberto de pesquisa em PLN.

1.7.4 Ética, Equidade (*Fairness*) e Mitigação de vieses

Como discutido no Capítulo Questões éticas em IA e PLN, modelos de PLN aprendem com dados gerados por humanos, o que pode significar que herdam e amplificam preconceitos históricos de gênero, raça ou classe. Classificadores de currículos ou de sentimentos podem apresentar disparidades de desempenho ou rotulações injustas devido a vieses nos *corpora* de treino. A tendência atual é o desenvolvimento de *benchmarks* de equidade específicos

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

para o português brasileiro e de técnicas de *desenviesamento* (*debiasing*) que garantam que a categorização seja justa para todos os grupos demográficos.

1.7.5 Soberania digital

A dependência de modelos globais de larga escala suscita questões sobre a soberania dos dados brasileiros. Há um movimento crescente em direção ao desenvolvimento de modelos nacionais robustos, como a MariTalk (Sales Almeida et al., 2024), e de técnicas de *destilação de conhecimento* (Gou et al., 2021). O objetivo é criar classificadores "pequenos e inteligentes" que consigam manter o desempenho de um LLM, mas que possam ser executados localmente em servidores brasileiros ou dispositivos móveis, com menor latência e custo.

1.7.6 Categorização em baixos recursos e dialetos

O Brasil é um país de diversidade linguística imensa. A maioria dos recursos da Seção 1.6 foca no português padrão. Um desafio aberto é a categorização eficaz de textos em dialetos regionais, variedades populares e línguas indígenas brasileiras, casos em que a escassez de dados rotulados impede o uso de métodos conexionistas tradicionais (Finger, 2021), exigindo avanços em aprendizado *zero-shot* e na transferência de conhecimento.

1.8 Referências para aprofundamento

Dada a amplitude e a rápida evolução do campo da classificação de texto, este capítulo oferece uma base sólida, mas não esgota as diversas nuances e os desenvolvimentos recentes da área. Para os leitores que desejam expandir seus conhecimentos teóricos e práticos, recomendamos as obras mencionadas a seguir.

Para uma compreensão aprofundada das bases teóricas e dos métodos clássicos de classificação e recuperação de informação, a obra de Manning; Schütze (1999) permanece como uma referência fundamental, detalhando algoritmos que formaram a base do PLN moderno. Para uma visão panorâmica e atualizada sobre o impacto dos modelos *Transformers* e das LLMs na tarefa de categorização, o levantamento realizado por Fields et al. (2024) constitui um importante guia para mapear o estado da arte. Por fim, os aspectos técnicos e comparativos entre as estratégias contemporâneas de adaptação de modelos, como o *fine-tuning* e o *instruction-tuning*, podem ser aprofundados no trabalho de Chae; Davidson (2025).

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Capítulo 2

Recuperação de Informação

Viviane P. Moreira

Publicado em: 26/09/2023

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

2.1 Introdução

A necessidade de organizar a informação é inerente à espécie humana – bibliotecas existem desde pelo menos o ano 2600 AC. Dado o grande volume de informação que as bibliotecas armazenam, a partir dos anos 1960, iniciaram-se esforços a fim de automatizar o armazenamento e a busca de materiais bibliográficos através da computação. Esses esforços marcaram o início da área de recuperação de informação (RI). A RI trata de encontrar, a partir de grandes coleções, material (geralmente documentos) de natureza não estruturada (geralmente texto) que satisfaça uma necessidade de informação (Manning et al., 2008b). Em outras palavras, o objetivo central da RI é a busca, ou seja, a tarefa de encontrar material **relevante** a partir de uma consulta de um usuário. Esta tarefa é comumente conhecida por **recuperação ad hoc**. Apesar de a RI poder ser aplicada a diferentes tipos de dados não estruturados como imagem, áudio, vídeo etc. o foco deste capítulo é **informações textuais**.

2.1.1 Relação com o PLN

Há muita interseção entre RI e PLN, pois ambas lidam com a linguagem humana. Contudo, há algumas diferenças. A primeira diferença é quanto à origem – enquanto que o PLN teve origem na inteligência artificial e na linguística computacional, a RI teve origem na biblioteconomia e na ciência da informação[1]. Outra diferença é quanto ao escopo – podemos dizer que o escopo do PLN é mais abrangente (i.e., compreensão e geração de linguagem) enquanto que o da RI é mais restrito às tarefas relacionadas à busca por informação.

2.1.2 O Foco da Recuperação de Informação

A **tarefa central** da RI é casar a consulta de um usuário com os documentos que são potencialmente relevantes a ela. A principal dificuldade é que os termos utilizados pelo usuário podem não ter sido usados nos documentos relevantes. Este problema é conhecido como **incompatibilidade de vocabulário** (*vocabulary mismatch*) e é decorrente de dois fenômenos comuns na linguagem: a **sinonímia** e a **polissemia**. A sinonímia refere-se ao fato de usarmos palavras diferentes para nos referirmos ao mesmo conceito, e.g., "bergamota", "tangerina" e "mexerica". O problema para RI é que uma consulta pelo termo

[1]Para um relato mais detalhado sobre as origens da RI, há o artigo de Lesk (1995).

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

"bergamota" não consegue recuperar documentos relevantes que contenham "tangerina". A polissemia refere-se ao fato de que uma mesma palavra pode apresentar sentidos distintos, e.g.,"manga" que tanto pode ser a fruta como a parte da vestimenta que cobre o braço da pessoa. O efeito negativo da polissemia para a RI é a recuperação de documentos que contêm a palavra pesquisada, mas não no sentido pretendido. Ao longo dos anos, houve uma vasta gama de propostas de solução para este problema. Uma visão geral dessas propostas é fornecida na Seção 2.5.

Até meados dos anos 1990, o interesse em RI estava restrito a bibliotecários, jornalistas e profissionais do direito (i.e., profissões que tinham bastante necessidade de buscar informações). Com a popularização da Internet e dos motores de busca para a web, a RI ganhou muita importância. Sistemas de RI fazem parte da vida diária de uma grande parte da população mundial. Há estimativas que o Google, o motor de busca mais utilizado, receba cerca de 100 mil consultas por segundo e tenha 4,3 bilhões de usuários[2]. Os desafios de se lidar com coleções contendo bilhões de documentos (i.e., páginas web) motivaram o desenvolvimento de novos algoritmos e técnicas, objetivando tanto eficiência (baixo custo computacional) quanto eficácia (qualidade do resultado).

Um sistema de RI também pode ser utilizado como um componente em tarefas de PLN como sistemas de perguntas e respostas e de detecção de plágio. A vantagem é que a RI consegue recuperar documentos candidatos com um custo computacional baixo. Assim, as fases subsequentes que requerem comparações exaustivas usando modelos mais custosos podem trabalhar com um conjunto menor de documentos.

2.1.3 O Conceito de Relevância

Vimos na Seção 2.1.2 que a tarefa central de RI é recuperar itens que sejam **relevantes** a uma necessidade de informação expressa por meio de palavras-chave. A relevância é um julgamento feito **pelo usuário** que indica o quão bem um documento satisfaz a consulta. Em sua forma mais simples, ela pode ser tratada como binária – cada documento é considerado como **relevante** ou como **não relevante**. Também é possível utilizar diversos níveis de relevância como por exemplo: "muito relevante", "moderadamente relevante", "marginalmente relevante" e "não relevante".

É comum ouvirmos que um motor de busca ou sistema de RI retorna os resultados em **ordem de relevância**. Esta afirmação não está tecnicamente correta pois a **relevância é subjetiva** e só pode ser atribuída pelo usuário. Dois usuários podem fazer a mesma consulta q e julgar um mesmo documento d de forma diferente – enquanto um usuário considera que d é relevante, o outro usuário pode considerar que d não é relevante para q. Os sistemas de RI ordenam os documentos utilizando uma função de ranqueamento baseada em evidências que o sistema consegue calcular (e.g., estatísticas de ocorrências das palavras nos documentos, tamanho dos documentos etc.). Idealmente, espera-se que esta função de ranqueamento esteja positivamente correlacionada com a ideia subjetiva de relevância dos usuários.

2.1.4 Principais Livros

A RI é uma área de pesquisa por si só e este capítulo não pretende ser uma revisão exaustiva. vamos focar nos aspectos centrais da área e na sua interseção com PLN. Há

[2] https://wpdevshed.com/how-many-people-use-google/

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

diversos livros em **língua inglesa** que são referência em RI. Para o leitor que deseja saber mais sobre a área, indicamos os seguintes livros:

- Introduction to Information Retrieval de Manning, Schütze e Raghavan (Manning et al., 2008b), publicado em 2008 pela Cambridge University Press e disponível gratuitamente online[3].
- Search engines: Information retrieval in practice de Croft, Metzler, e Strohman (Croft et al., 2010), publicado em 2010 pela Addison-Wesley e disponível gratuitamente online[4].
- Modern Information Retrieval - the concepts and technology behind search de Baeza-Yates e Ribeiro-Neto (Baeza-Yates; Ribeiro-Neto, 2011), publicado em 2011 pela Addison-Wesley.

Em português, há uma versão resumida do livro Modern Information Retrieval publicada em 2013 (Baeza-Yates; Ribeiro-Neto, 2013). Esta versão contém os principais capítulos do livro original.

2.1.5 Organização deste Capítulo

Nesta primeira versão do livro **Processamento de Linguagem Natural: Conceitos, Técnicas e Aplicações em Português**, o capítulo sobre Recuperação da Informação tem como foco as operações de pré-processamento e indexação (Seção 2.2), nos modelos clássicos (Seção 2.3), no paradigma de avaliação (Seção 2.4) e em técnicas de melhoria de qualidade (Seção 2.5). Também indicamos algumas bibliotecas e ferramentas amplamente utilizadas (Seção 2.6). As próximas versões do livro irão incluir tópicos avançados como o ranqueamento neural.

2.2 Visão Geral de um Sistema Típico de Recuperação de Informação

A Figura 2.1 mostra uma visão geral do processo de RI. Em uma ponta, há a usuária que tem uma **necessidade de informação** e, na outra ponta, temos uma coleção de documentos textuais. A consulta da usuária é submetida a um sistema de RI (representado pelo retângulo laranja). As Etapas 1 e 2 ocorrem offline, pois o índice precisa estrar pronto antes que o sistema possa receber consultas. Na Etapa 1, os documentos da coleção são pré-processados para então, na Etapa 2, serem indexados. Na Etapa 3, o sistema executa sobre a consulta as mesmas operações de pré-processamento que foram aplicadas na Etapa 1. O texto pré-processado da consulta é utilizado, na Etapa 4, para buscar no índice os documentos que mais bem atendam a consulta. Os resultados da consulta são então retornados à usuária sob a forma de uma lista ordenada.

2.2.1 Pré-Processamento

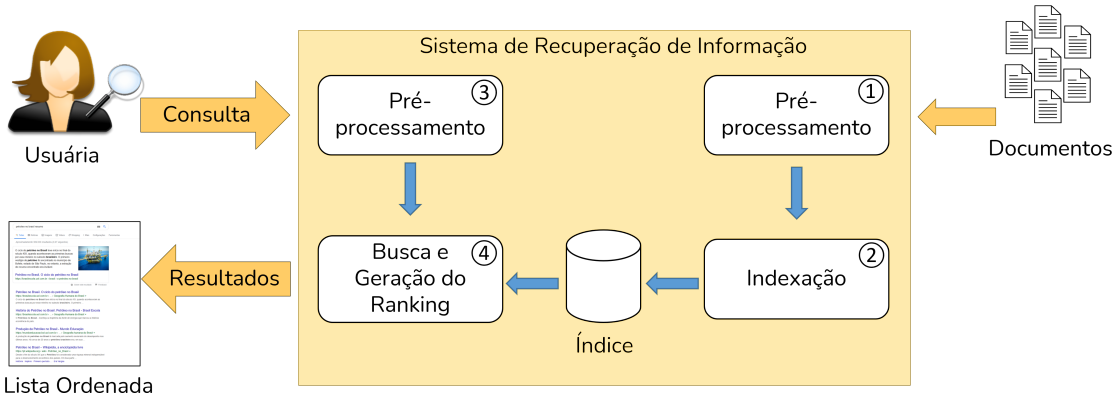
As operações de pré-processamento são muito semelhantes às estudadas para as tarefas de PLN relatadas no Capítulo Sequência de caracteres e palavras. É preciso definir a nossa **unidade de indexação**, ou seja, o que é um **documento** no contexto tratado.

[3] https://nlp.stanford.edu/IR-book/
[4] https://ciir.cs.umass.edu/irbook/

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Figura 2.1: Visão Geral do Processo de RI.



Por exemplo, uma notícia, um comentário, um e-mail, um artigo científico, um livro, um capítulo do livro etc.

2.2.1.1 Tokenização

Dada uma sequência de caracteres e uma unidade de indexação, a **tokenização** irá separar a sequência em *tokens* (i.e., unigramas de palavras). Nesse processo, é comum descartarmos caracteres de pontuação e mantermos apenas as palavras (i.e., também chamadas de termos). Embora à primeira vista este processo pareça bastante trivial (i.e., podemos apenas considerar que os caracteres não alfabéticos como símbolos e sinais de pontuação sejam separadores), há vários casos especiais que podemos precisar tratar. Por exemplo, endereços de email, emojis, e termos que misturam letras e símbolos como C++.

O processo de tokenização costuma ser implementado utilizando-se expressões regulares e há vários tokenizadores prontos em bibliotecas como NLTK[5] e spaCy[6]. Após a tokenização, é comum converter todos os caracteres para letra minúscula. O objetivo é fazer com que a busca não seja *case-sensitive*, isto é, impactada por texto em letras maiúscula e minúscula. Por outro lado, essa operação dificulta a identificação de entidades nomeadas, uma vez que entidades são geralmente nomes próprios com grafia em letra maiúscula.

A remoção de acentos, cuja denominação mais apropriada é **remoção de sinais diacríticos** também pode ser realizada, principalmente quando se trata de textos informais nos quais esses sinais são menos utilizados. Um sinal diacrítico é uma marca que colocamos sobre ou sob o caractere como acento agudo, grave, circunflexo, til ou cedilha.

2.2.1.2 Remoção de *Stop Words*

As *stop words* são palavras que têm pouca utilidade para a RI pois ocorrem em muitos documentos e assim não servem para distinguir o conteúdo semântico dos documentos. Preposições, conjunções, artigos e verbos de ligação são normalmente consideradas como *stop words*. Essas palavras são muitas vezes completamente descartadas em um processo conhecido como **remoção de *stop words***.

O processo de remoção de *stop words* é bastante simples e consiste em verificar a presença de cada palavra do texto em uma lista de *stop words* previamente construída. Existem

[5] https://www.nltk.org/
[6] https://spacy.io/models/pt

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

listas prontas que podemos usar, como a do NLTK, por exemplo, que contém 204 palavras. Contudo, é recomendável revisar se as listas contêm palavras que podem ser úteis para o contexto da aplicação sendo desenvolvida. Por exemplo, na lista elaborada pela Linguateca[7] com as palavras mais frequentes dos textos da Folha de São Paulo, encontramos a palavra "brasil". Esta palavra, muito provavelmente, não deve ser removida dos textos da coleção de documentos.

O principal argumento a favor da remoção de *stop words* é a redução do tamanho do vocabulário pois elas representam cerca de 40% das ocorrências de palavras em um *corpus*. Por outro lado, essa remoção pode trazer uma perda de informação relevante. Se pensarmos na famosa expressão "ser ou não ser, eis a questão", com a remoção de *stop words*, sobraria apenas "questão", o que descaracteriza completamente a expressão. O impacto negativo na busca seria que o sistema não mais conseguiria distinguir entre documentos que contenham a expressão completa daqueles que contêm apenas a palavra "questão".

2.2.1.3 *Stemming*

O objetivo do *stemming* é gerar uma mesma representação para formas variantes de uma mesma palavra através da remoção dos sufixos. Por exemplo, removendo-se os sufixos de "apresentação", "apresentando", e "apresentar", obteríamos o radical "apresent". Com isso, uma usuária buscando por "apresentar artigos científico" consegue recuperar um documento com o trecho "apresentando artigos científicos". O benefício é aumentar o número de documentos relevantes recuperados em resposta à consulta.

O processo de *stemming* e seu impacto sobre RI vêm sendo estudado há diversos anos. Para a língua inglesa, o primeiro algoritmo de *stemming*, ou *stemmer*, data de 1968 e foi proposto por Julie Beth Lovins (Lovins, 1968). Em 1980, Martin Porter propôs o Porter Stemmer (Porter, 1980) que mostrou obter bons resultados (também para a língua inglesa) e posteriormente foi traduzido para outros idiomas incluindo o português[8].

O primeiro *stemmer* desenvolvido especialmente para a língua portuguesa foi o Removedor de Sufixos da Língua Portuguesa (RSLP)[9] (Orengo; Huyck, 2001). Posteriormente veio o STEMBR (Alvares et al., 2005). Há resultados experimentais que mostram a validade da aplicação de *stemmers* para RI em português (Flores et al., 2010; Orengo et al., 2006). Muitas vezes, na média de um conjunto de consultas, o impacto do *stemming* pode ser pequeno. Contudo, ao analisarmos consultas individuais, percebemos melhorias muito grandes em alguns casos. Em outros, podem aparecer mais documentos não relevantes recuperados pois a remoção do sufixo pode gerar ambiguidade. Os resultados mostram que *stemmers* menos agressivos, ou seja, com menos regras de remoção tendem a gerar resultados melhores pois têm menos impacto negativo na precisão.

O Quadro 2.1 mostra um trecho do livro O Tempo e o Vento – O Continente, de Érico Veríssimo, e suas versões após cada etapa do pré-processamento. Observemos a redução de 51 para 27 *tokens* com a remoção de *stop words*. O *stemmer* utilizado no exemplo foi o Snowball[10]. Também vale ressaltar a potencial introdução de ambiguidade com a redução de algumas palavras, por exemplo de "legalistas" para "legal".

[7] https://www.linguateca.pt/chave/stopwords/
[8] http://snowball.tartarus.org/algorithms/portuguese/stemmer.html
[9] https://www.vivianemoreira.org/rslp
[10] https://snowballstem.org/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

Quadro 2.1: Excerto do livro O Tempo e o Vento – O Continente, de Érico Veríssimo, e suas versões após cada etapa do pré-processamento.

Original	Quando pela primeira vez aparecera em Santa Fé, no ano em que fora assinada a paz entre farroupilhas e legalistas, causara a pior das impressões. Chegara escoteiro, montado num cavalo magro e manco, e fazendo questão de mostrar a toda a gente que tinha as guaiacas atestadas de moedas de ouro.
Tokenizado	Quando pela primeira vez aparecera em Santa Fé no ano em que fora assinada a paz entre farroupilhas e legalistas causara a pior das impressões Chegara escoteiro montado num cavalo magro e manco e fazendo questão de mostrar a toda a gente que tinha as guaiacas atestadas de moedas de ouro
Convertido para minúsculo	quando pela primeira vez aparecera em santa fé no ano em que fora assinada a paz entre farroupilhas e legalistas causara a pior das impressões chegara escoteiro montado num cavalo magro e manco e fazendo questão de mostrar a toda a gente que tinha as guaiacas atestadas de moedas de ouro
Após remoção de *stop words*	primeira vez aparecera santa fé ano assinada paz farroupilhas legalistas causara pior impressões chegara escoteiro montado cavalo magro manco fazendo questão mostrar gente guaiacas atestadas moedas ouro
Após *stemming*	primeir vez aparec sant fé ano assin paz farroupilh legal caus pior impressõ cheg escoteir mont caval magr manc faz questã mostr gent guaiac atest moed our

2.2.2 Indexação

A forma que temos para evitar ter de "varrer" os textos da coleção de documentos em busca dos termos da consulta (o que seria um procedimento extremamente lento) é contarmos com um **índice** previamente construído. A indexação recebe como entrada o texto pré-processado e cria um índice chamado de **arquivo invertido**. Um arquivo invertido assemelha-se ao índice remissivo que comumente encontramos ao final de um livro e contém as palavras (únicas) do texto e a lista de documentos em que elas aparecem.

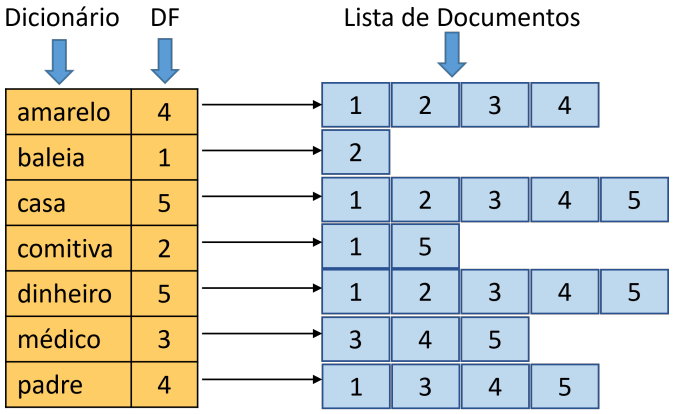
A Figura 2.2 mostra um exemplo de índice que pode ser usado por um sistema de RI. O índice tem dois componentes: (i) o dicionário, também é chamado de vocabulário, que possui a lista de termos únicos da coleção de documentos e (ii) a lista de identificadores de documentos em que o termo aparece (conhecida por lista de *postings*). Junto ao dicionário, também armazena-se o número de documentos em que o termo aparece, conhecido como *document frequency* (DF). Observemos que o dicionário fica em ordem alfabética e a lista de *postings* fica em ordem de identificador de documento.

Nos casos em que desejamos fazer uma busca sobre um campo específico do documento (e.g., título ou autor), vamos precisar de **índices paramétricos**, um para cada campo. Além disso, se quisermos permitir buscas por múltiplos termos adjacentes (e.g., "memória prodigiosa"), precisaremos de um **índice posicional** que guarde também as posições em que os termos aparecem nos documentos. Certamente, tanto os índices paramétricos como os posicionais possuem um custo maior em termos de espaço de armazenamento e de processamento.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Figura 2.2: Exemplo de um trecho de índice armazenado como um arquivo invertido.



2.3 Modelos Clássicos de recuperação de informação

Um modelo de RI especifica como representar os documentos, as consultas e como compará-los. Ao longo dos anos, diversos modelos de RI foram propostos. Vamos explorar os modelos clássicos. Todos os modelos clássicos pressupõem que a distribuição dos termos nos documentos é independente e utilizam a abordagem *bag of words* (BoW). Em um BoW, a ordem dos termos nos documentos é desprezada. A vantagem é a simplificação dos modelos, os quais conseguem processar consultas de maneira mais rápida. A desvantagem é a perda de semântica em alguns casos pois as sentenças "João é mais velho do que José" e "José é mais velho do que João" têm representações idênticas apesar de significado oposto.

2.3.1 Modelo Booleano

O primeiro modelo que foi usado para RI foi o modelo Booleano. Trata-se de um modelo simples, baseado na teoria dos conjuntos e na álgebra booleana. As consultas são compostas por palavras-chave combinadas com os operadores AND, OR e NOT e o seu processamento utiliza o índice.

Para processar uma consulta do tipo p_1 AND p_2, precisamos selecionar as listas de *postings* de p_1 e p_2 e computar a sua interseção. Por exemplo, com o trecho de índice da Figura 2.2, podemos ver que a consulta "comitiva" AND "médico" retornaria apenas o documento 5. Já a consulta "baleia" AND "padre" não retornaria nenhum documento.

Consultas com OR implicam na união de conjuntos. Por exemplo, considerando o índice da Figura 2.2, a consulta "comitiva" OR "médico" retornaria os documentos 1, 3, 4 e 5.

A principal limitação do modelo Booleano é que ele não é capaz de ordenar o resultado da consulta. Ou um documento satisfaz ou não satisfaz a expressão Booleana da consulta; não há a opção "satisfaz parcialmente". Apesar dessa limitação, este foi o modelo comercial mais usado por três décadas, mesmo após a proposta de modelos superiores. Até hoje ainda vemos o modelo Booleano sendo usado, por exemplo, em alguns sistemas de bibliotecas.

2.3.2 Modelo Vetorial

Com o objetivo de poder ordenar os documentos em resposta às consultas, há duas premissas simples que podemos utilizar: (i) documentos que contém mais vezes os termos da consulta têm mais chance de estarem relacionados a ela (e de serem relevantes) e (ii) os

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

termos mais raros na coleção são mais úteis para diferenciar o conteúdo dos documentos. A primeira premissa é atribuída ao pesquisador alemão Hans Peter Luhn e data de 1957. E a segunda premissa foi desenvolvida pela pesquisadora inglesa Karen Spärck Jones em 1972.

Com base nessas premissas, Gerard Salton propôs, na década de 1960, o modelo vetorial (em inglês, *vector space model*) (Salton; McGill, 1983). Nesse modelo, os documentos e as consultas são representados como vetores em um espaço de t dimensões, onde t é o número de termos distintos (i.e., o tamanho do dicionário). No espaço vetorial, os termos são os eixos. As consultas e os documentos são representados no espaço de acordo com a força da associação que eles têm com o termo. A força da associação do documento d_j com o termo i (w_{i,d_j}) é computada pelo esquema TF-IDF (*Term Frequency times Inverse Document Frequency*), segundo a Equação 2.1:

$$w_{i,d_j} = TF_{i,d_j} \times IDF_t = freq_{i,d_j} \times log_{10} \frac{N}{DF_i} \tag{2.1}$$

onde $freq_{i,d_j}$ é o número de ocorrências do termo i no documento d_j, N é o número de documentos da coleção e DF_i é o número de documentos que contêm o termo i. É comum normalizar o componente TF para evitar que documentos longos sejam beneficiados, pois um documento longo tem mais chances de possuir os termos da consulta. A normalização do componente TF pode ser feita de diferentes formas: (i) dividindo-se o número de ocorrências do termo no documento pelo número de palavras no documento ou (ii) dividindo-se o número de ocorrências do termo no documento pelo número de ocorrências do termo mais comum naquele documento.

O componente IDF tem por objetivo dar um peso maior para os termos raros pois eles são mais úteis para discriminar o conteúdo dos documentos. Pensando em uma situação extrema, um termo que ocorra em todos os documentos não tem nenhuma utilidade para diferenciar um documento de outro. Sendo assim, valores de IDF mais altos serão atribuídos para termos que ocorram em poucos documentos. Note-se que o IDF não tem efeito caso a consulta tenha apenas um termo. Em consultas com dois ou mais termos, o efeito do IDF é dar mais importância para os termos que ocorrem em menos documentos. Por exemplo, imaginemos uma coleção que trata de instrumentos musicais. Um usuário faz a consulta "flauta bansuri". Considerando que o termo "flauta" é muito mais comum do que o termo "bansuri" (que se refere a um tipo específico de flauta transversal), o efeito do IDF será atribuir bem mais importância ao termo "bansuri". Os vetores gerados pelo TF-IDF são longos e esparsos. **Longos** pois o número de dimensões é igual ao número de palavras do vocabulário (o que pode chegar a centenas de milhares para algumas coleções). **Esparsos** pois a grande maioria dos elementos têm valor zero já que os documentos possuem uma pequena fração dos termos do vocabulário.

A similaridade entre um documento e uma consulta é dada pelo **cosseno** entre os seus vetores. O cosseno entre uma consulta q e um documento d_j é o produto escalar normalizado dos vetores $\vec{q}$ e $\vec{d_j}$, dado pela Equação 2.2:

$$cosseno(q, d_j) = \frac{\vec{q} \cdot \vec{d_j}}{|\vec{q}| \times |\vec{d_j}|} = \frac{\sum_{i=1}^{t} w_{i,q} \times w_{i,d_j}}{\sqrt{\sum_{i=1}^{t} {w_{i,q}}^2} \times \sqrt{\sum_{i=1}^{t} {w_{i,d_j}}^2}} \tag{2.2}$$

onde $w_{i,q}$ e w_{i,d_j} representam o peso do i-ésimo termo na consulta q e no documento d_j, respectivamente.

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

O cosseno é máximo (i.e., igual a 1) se os vetores possuem um ângulo de 0 graus e é mínimo (i.e., igual a 0) se os vetores formarem um ângulo de 90 graus[11]. Em outras palavras, o cosseno será mínimo se o documento e a consulta não compartilharem nenhum termo. Vale ressaltar que não é preciso computar o cosseno entre a consulta e todos os documentos. Utilizando um índice (como o da Figura 2.2), podemos selecionar apenas os documentos que têm pelo menos uma das palavras da consulta, e calcular o cosseno somente entre a consulta e esses documentos. Também é possível estender essa noção e selecionar apenas documentos com valores de TF-IDF que ultrapassem um limiar para os termos da consulta ou restringir ainda mais e calcular o cosseno apenas para documentos que possuam duas ou mais das palavras da consulta.

Exemplo 2.1: Para ilustrar o conceito de TF-IDF e o ranqueamento dos documentos no modelo vetorial através do cosseno, vamos trabalhar com um exemplo. Nossa coleção de documentos é composta por cinco importantes livros da literatura brasileira. São eles:

- d_1 – O Alienista de Machado de Assis;
- d_2 – Vidas Secas de Graciliano Ramos;
- d_3 – O Continente (de O Tempo e o Vento) de Érico Veríssimo;
- d_4 – Capitães de Areia de Jorge Amado; e
- d_5 – Os Sertões de Euclides da Cunha.

Essa pequena coleção possui 350.260 *tokens*, e um vocabulário de 32.594 termos (já desconsiderando-se as *stop words*) – isso quer dizer que os documentos e as consultas são representados por vetores de 32.594 dimensões. Como seria inviável mostrarmos todos esses termos, iremos focar em sete termos selecionados: amarelo, baleia, casa, comitiva, dinheiro, médico e padre. A Tabela 2.1 mostra os sete termos selecionados juntamente com o número de documentos em que cada um deles aparece (DF). A partir do DF, podemos computar o IDF (segundo elemento da Equação 2.1). Por exemplo, o termo "comitiva" ocorre em dois documentos, então $IDF_{comitiva} = log_{10}\frac{5}{2} = 0,3979$. Já para o termo "casa" que ocorre em todos os cinco documentos, o $IDF_{casa} = log_{10}\frac{5}{5} = 0$.

Tabela 2.1: Exemplo de DF e IDF para algumas palavras do texto

Termo	**DF**	**IDF**
amarelo	4	0,0969
baleia	1	0,6990
casa	5	0,0000
comitiva	2	0,3979
dinheiro	5	0,0000
médico	4	0,0969
padre	4	0,0969

[11] Note que o valor do cosseno irá ficar no intervalo [0,1] e não [-1,1] pois os escores gerados pelo TF-IDF são positivos.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Tabela 2.2: Exemplo de matriz de incidência de termos em documentos com valores de TF (i.e., o número de ocorrências dos termos nos documentos).

Termo	d_1 O Alienista	d_2 Vidas Secas	d_3 O Tempo e O Vento	d_4 Capitães de areia	d_5 Os Sertões
amarelo	1	42	6	3	0
baleia	0	86	0	0	0
casa	109	37	247	120	30
comitiva	4	0	0	0	4
dinheiro	7	9	33	43	3
médico	18	0	157	7	8
padre	22	0	120	252	9

Tabela 2.3: Exemplo de matriz de incidência de termos em documentos com valores de TF-IDF

Termo	d_1 O Alienista	d_2 Vidas Secas	d_3 O Tempo e O Vento	d_4 Capitães de areia	d_5 Os Sertões
amarelo	0,0009	0,0473	0,0024	0,0012	0,0000
baleia	0,0000	0,6990	0,0000	0,0000	0,0000
casa	0,0000	0,0000	0,0000	0,0000	0,0000
comitiva	0,0146	0,0000	0,0000	0,0000	0,0549
dinheiro	0,0000	0,0000	0,0000	0,0000	0,0000
médico	0,0160	0,0000	0,0616	0,0031	0,0258
padre	0,0196	0,0000	0,0471	0,0969	0,0291

Para computar os vetores de pesos TF-IDF para os cinco documentos da coleção, vamos iniciar com uma matriz de incidência dos termos nos documentos. Temos sete termos e cinco documentos; então, nossa matriz terá sete linhas e cinco colunas. Nessa matriz, ilustrada na Tabela 2.2, temos o número de ocorrências dos termos nos documentos. A seguir, vamos normalizar o componente TF dividindo-o pelo número de ocorrências do termo mais comum (i.e., o termo que mais ocorre no documento), que para d_1 é "casa". Com isso, valor de TF-IDF para o termo "padre" no documento d_1 é $(22/109) \times 0,0969 = 0,0196$. Esses valores estão na Tabela 2.3. Note-se que os escores para os termos "casa" e "dinheiro" são iguais a zero pois eles aparecem em todos os cinco documentos, resultando em um IDF = 0. Também podemos observar que o maior escore é o do termo "baleia" no documento d_2. Isso ocorre porque esse termo ocorre muitas vezes em d_2 e nenhuma vez nos outros quatro documentos – ele é, portanto, um termo que tem bastante poder de diferenciar esse documento dos demais.

Para exemplificar o cálculo do cosseno, vamos imaginar uma consulta $q = \{\text{comitiva, médico}\}$. Assim como os documentos, o vetor da consulta terá sete dimensões e será composto por pesos TF-IDF. Os termos que não aparecem na consulta têm peso zero. Assim, o vetor da consulta será $\vec{q} = [0; 0; 0; 0,3979; 0; 0,0961; 0]$. Para montar o *ranking* dos documentos em relação à consulta, calculamos o cosseno entre $\vec{q}$ e os vetores dos documentos que estão na Tabela 2.3 utilizando a Equação 2.2. Por exemplo, o cosseno entre $\vec{q}$ e $\vec{d_1}$ é:

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

$$\frac{0 \times 0,009 + 0 \times 0 + 0 \times 0 + 0,3979 \times 0,0146 + 0 \times 0 + 0,0969 \times 0,0160 + 0 \times 0,0196}{\sqrt{0^2 + 0^2 + 0^2 + 0,3979^2 + 0^2 + 0,0969^2 + 0^2} \times \sqrt{0,0009^2 + 0^2 + 0^2 + 0,0146^2 + 0^2 + 0,0160^2 + 0,0196^2}} =$$

$$\frac{0 + 0 + 0 + 0,0058 + 0 + 0,0016 + 0}{\sqrt{0 + 0 + 0 + 0,1583 + 0 + 0,0009 + 0} \times \sqrt{0 + 0 + 0 + 0 + 0,0002 + 0 + 0,0003 + 0,0004}} =$$

$$\frac{0,0074}{\sqrt{0,1677} \times \sqrt{0,0009}} = \frac{0,0074}{0,4096 \times 0,0292} = \frac{0,0074}{0,012} = 0,6156$$

Os escores de cosseno entre $\vec{q}$ e os cinco documentos estão na Tabela 2.4. Note que como q e d_2 não têm nenhum termo em comum, o escore desse documento é zero. Para montar o *ranking*, ordenamos os documentos em ordem decrescente de cosseno. Dessa forma, o primeiro do *ranking* seria d_5, seguido por d_1, d_3 e d_4.

Tabela 2.4: Cosseno entre o vetor da consulta q e os vetores dos documentos da coleção.

d_1	d_2	d_3	d_4	d_5
0.6156	0.0000	0.1879	0.0075	0.8765

2.3.3 Modelos Probabilísticos

Há diversos modelos que utilizam abordagens para RI baseadas em um arcabouço probabilístico. A principal motivação é que a teoria das probabilidades fornece uma base sólida para representar e manipular a incerteza que é inerente ao processo de casamento entre consultas e documentos.

A ideia original foi proposta por Rijsbergen (1979) e intitulada *Probability Ranking Principle*. A ideia é ranquear os documentos da coleção em ordem decrescente de probabilidade de relevância para a consulta. Se tivéssemos uma amostra de documentos sabidamente relevantes ou não relevantes, seria possível estimar a probabilidade de um termo aparecer em um documento relevante. Essa informação poderia então ser usada para ranquear os documentos. As probabilidades são estimadas com a maior precisão possível com base em estatísticas que conseguimos calcular. O primeiro modelo probabilístico foi proposto por Robertson; Spärck Jones (1976) – o *Binary Independence Model* (BIM). O BIM é um modelo binário, ou seja, ele considera apenas se o termo está ou não presente no documento. Outra simplificação é a utilização do princípio da independência, que supõe que os termos ocorrem nos documentos de maneira independente. Esse é um modelo simples que não incorpora algumas características, como a frequência dos termos nos documentos e a importância dos termos na coleção (vistas na Seção 2.3.2), que contribuem para a geração de resultados melhores para as consultas (i.e., uma melhor ordenação dos documentos).

O modelo probabilístico mais influente até os dias de hoje é o **Okapi BM25** que foi projetado entre os anos 1980 e 1990 por Spärck Jones et al. (2000). O Okapi era um sistema de busca para o catálogo da biblioteca da City University de Londres. O modelo BM25 (BM é a abreviatura de *best match*) foi resultado de uma série de experimentos com variações de fórmulas de modelos probabilísticos. Os pesquisadores usaram identificadores para as variações da fórmulas e a que obteve melhores resultados foi a BM25, que é dada conforme a Equação 2.3 a seguir:

$$BM25(q, d_j) = \sum_{i \in q} log \frac{(r_i + 0,5)/(R - r_i + 0,5)}{(df_i - r_i + 0,5)/(N - df_i + r_i + 0,5)} \times \frac{(k_1 + 1)tf_i}{K + tf_i} \times \frac{(k_2 + 1)qf_i}{k_2 + qf_i} \tag{2.3}$$

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

onde R é o número de documentos sabidamente relevantes para a consulta q; r_i é o número de documentos sabidamente relevantes que contêm o termo i; N é o número de documentos da coleção; df_i é o número de documentos que contêm o termo i; tf_i é o número de ocorrências do termo i no documento j; qf_i é o número de ocorrências do termo i na consulta q; K, k_1 e k_2 são parâmetros que precisam ser definidos empiricamente. A adição de 0,5 em cada termo da equação é para evitar a divisão por zero. Note que o somatório é efetuado para todos os termos da consulta.

O parâmetro k_1 controla a escala da frequência dos termos (tf) nos documentos. Usar $k_1 = 0$ faz com que o modelo se comporte como um modelo binário. Por outro lado, altos valores de k_1 correspondem ao uso do TF sem normalização. Na prática, é comum usarmos $k_1 = 1{,}2$. De forma similar, o parâmetro k_2 controla a escala da frequência dos termos na consulta (qf). Os valores recomendados para k_2 são entre 0 e 1000.

Por fim, o parâmetro K é calculado por:

$$K = k_1((1-b) + b \times \frac{dl}{avdl})$$

onde dl é o tamanho do documento, $avdl$ é o tamanho médio dos documentos da coleção e b controla a normalização em função do tamanho do documento. Se $b = 0$, então a normalização não é realizada. Já $b = 1$ normaliza os pesos em função do tamanho dos documentos. O valor recomendado para b é 0,75.

É importante ressaltar que os escores gerados pelos modelos probabilísticos não são probabilidades, i.e., não estão no intervalo [0,1]. Isso não representa um problema pois a utilidade do escore é apenas ordenar os documentos, sem a necessidade de permitir uma interpretação probabilística.

Como na imensa maioria dos casos práticos nós não temos conhecimento de uma amostra de documentos relevantes, o BM25 também pode ser computado **sem a informação de relevância**. Em outras palavras, R e r_i são zero. Com isso, a Equação 2.3 pode ser simplificada para:

$$BM25(q, d_j) = \sum_{i \in q} log \frac{N - df_i + 0,5}{df_i + 0,5} \times \frac{(k_1+1)tf_i}{K + tf_i} \times \frac{(k_2+1)qf_i}{k_2 + qf_i} \tag{2.4}$$

Exemplo 2.2: Para ilustrar o cálculo do BM25, vamos utilizar a mesma consulta da Seção 2.3.2 $q = \{$comitiva, médico$\}$ e faremos o cálculo do escore passo a passo para o d_1. Não temos informação prévia de relevância, então vamos usar a Equação 2.4. Cada termo ocorre apenas uma vez na consulta, então $qf_1 = 1$ e $qf_2 = 1$. O escore final é formado pela soma dos escores das palavras da consulta que aparecem no documento d_1. Para facilitar o entendimento, os cálculos referentes ao termo "comitiva" estão em laranja e os os cálculos referentes ao termo "médico" estão em azul. As estatísticas da coleção estão nas Tabelas 2.1, 2.2, 2.3. Os parâmetros utilizados foram $k_1 = 1{,}2$, $k_2 = 100$ e $b = 0,75$. O tamanho médio dos documentos da coleção ($avdl$) é de 276 e o tamanho de $d1$ é 161. Com isso, $K = 1,2((1-0,75) + 0,75 \times 161/279) = 0,825$. Observe que o escore final para o documento é negativo, mas o valor absoluto não é importante pois estamos interessados apenas em ranquear os documentos. Calculando o escore para os outros quatro documentos, verificamos que o primeiro documento do *ranking* seria d_5, seguido por d_1, d_4 e d_3. Note-se que o ordenamento fica quase igual ao gerado pelo modelo vetorial – apenas os documentos d_4 e d_3 invertem as posições.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

$$BM25(q, d_1) = log\frac{5-2+0,5}{2+0,5} \times \frac{(1,2+1)4}{0,825+4} \times \frac{(100+1)1}{100+1} + log\frac{5-4+0,5}{4+0,5} \times \frac{(1,2+1)18}{0,825+18} \times \frac{(100+1)1}{100+1} =$$

$$log\frac{3,5}{2,5} \times \frac{8,8}{4,825} \times \frac{101}{101} + log\frac{1,5}{4,5} \times \frac{39,6}{18,825} \times \frac{101}{101} =$$
$$log1,4 \times 1,824 \times 1 + log0,33 \times 2,1 \times 1 =$$
$$0,33 \times 1,824 + -1,10866 \times 2,1 =$$
$$0,6137 - 2,311 = -1,6973$$

2.4 Avaliação da Qualidade de Sistemas de recuperação de informação

A RI é uma disciplina altamente empírica e a história da avaliação em sistemas de RI nasceu junto com a área. Os primeiros trabalhos sobre avaliação em RI foram coordenados por Cyril W. Cleverdon na escola de aeronáutica de Cranfield (Inglaterra) nos anos 1960. Por esta razão, o modelo de avaliação até hoje é conhecido como paradigma Cranfield. A avaliação consiste no cálculo de uma série de métricas que são calculadas com base em uma coleção de teste. Nesta seção abordaremos as principais métricas (Seção 2.4.1) e as coleções de teste para português (Seção 2.4.2). O leitor que deseja se aprofundar no processo de avaliação de RI pode referir-se a Sanderson et al. (2010).

2.4.1 Métricas

As métricas são baseadas na noção de relevância, i.e., uma avaliação que diz se um documento d_j é relevante a uma consulta q. Inicialmente, vamos tratar a relevância como **binária**, ou seja, o *ground truth* julga o documento como relevante (1) ou não relevante (0). Todas as métricas vistas aqui variam no intervalo [0,1], sendo que 1 representa a recuperação ideal.

2.4.1.1 Métricas Baseadas em Conjuntos

A Figura 2.3 mostra um exemplo do que ocorre tipicamente quando fazemos uma consulta. A coleção de documentos está representada pelo oval cinza. O círculo amarelo representa os documentos que foram recuperados em resposta à consulta, enquanto que os documentos que de fato são relevantes para a consulta estão representados pelo círculo azul. Podemos ver que há documentos relevantes que foram recuperados (i.e., os documentos que estão na intersecção entre os conjuntos, que aparece representada na cor verde), mas há também documentos relevantes que deixaram de ser recuperados e documentos não relevantes que foram recuperados. Com base nesses conjuntos, podemos definir duas métricas básicas para avaliar a qualidade da recuperação: **precisão** e **revocação** (em inglês, *recall*).

- Precisão (P) mede a proporção dos documentos recuperados que de fato é relevante.

$$P = \frac{\#\ relevantes\ recuperados}{\#\ recuperados}$$

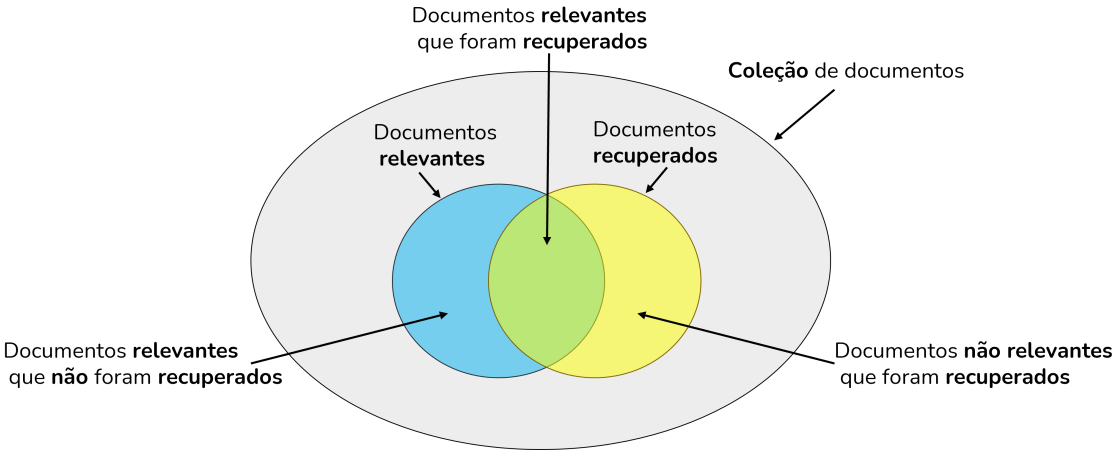
https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

- Revocação (R) mede a proporção dos documentos relevantes que de fato foram recuperados.

$$R = \frac{\#\ relevantes\ recuperados}{\#\ relevantes}$$

Figura 2.3: Exemplo de recuperação de documentos em resposta a uma consulta.



Por exemplo, vamos imaginar uma situação em que um sistema recupera 100 documentos em resposta a uma consulta para a qual existam 30 documentos relevantes. Sabendo-se que entre os recuperados há 15 documentos relevantes, os valores das métricas de avaliação seriam $P = \frac{15}{100} = 15\%$ e $R = \frac{15}{30} = 50\%$. Imaginemos, então, uma segunda situação em que, na tentativa de aumentar a precisão, reduzimos o número de documentos recuperados para apenas 5 e todos eles sejam relevantes. Como resultado, nossa precisão aumentaria para $P = \frac{5}{5} = 100\%$. O problema é que, com isso, a nossa revocação cairia para $\frac{5}{30} = 16,67\%$. Indo na direção oposta, poderíamos tentar aumentar a revocação e recuperar mil documentos em vez de 100. Com isso, haveria uma chance bem maior de recuperarmos documentos relevantes, mas por outro lado, a precisão seria reduzida. Em resumo: altos níveis de revocação costumam ser acompanhados por baixos níveis de precisão e vice-versa. Por esta razão, é comum utilizarmos uma medida que agregue P e R, a conhecida **medida F**, definida a seguir.

$$F = \frac{(\beta^2 + 1) \times P \times R}{(\beta^2 \times P) + R}$$

onde β é um parâmetro que permite definirmos a ênfase dada às métricas. Valores de β maiores do que 1 enfatizam a revocação enquanto que valores de β menores do que 1 enfatizam a precisão. Se quisermos atribuir a mesma ênfase às duas métricas, usamos $\beta = 1$. Nesse caso, a métrica é comumente chamada de $F1$ e sua fórmula é simplificada para $F1 = \frac{2 \times P \times R}{P+R}$. Considerando o nosso primeiro exemplo, em que $P = 0,15$ e $R = 0,50$, a $F1$ seria $F1 = \frac{2 \times 0,15 \times 0,50}{0,15+0,50} = 0,23$.

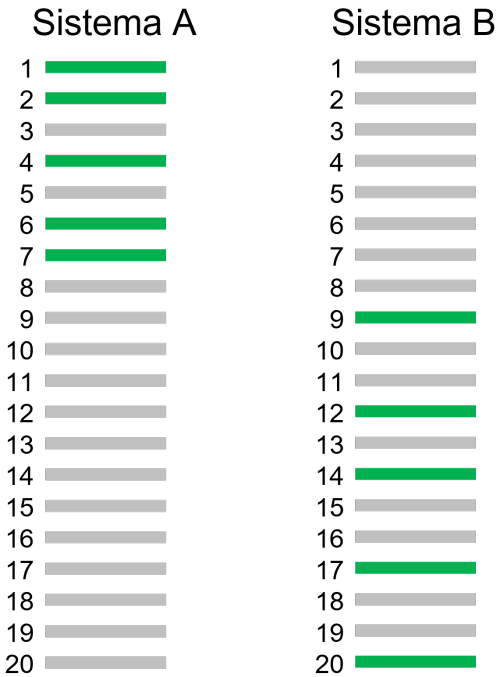
2.4.1.2 Métricas para *Rankings*

Os sistemas de RI retornam os documentos em um *ranking*. A Figura 2.4 mostra exemplos de *rankings* de documentos recuperados em resposta a uma mesma consulta por dois sistemas diferentes (Sistemas A e B). São 20 documentos recuperados e, dentre eles, há 5

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Figura 2.4: Exemplos de *rankings* de documentos recuperados em resposta a uma consulta por dois sistemas diferentes. Os documentos relevantes à consulta estão representados em verde.



relevantes (representados em verde). Vamos supor que existam 7 documentos relevantes para essa consulta. As métricas vistas na Seção 2.4.1.1 resultariam em $P = \frac{5}{20} = 25\%$, $R = \frac{5}{7} = 71\%$ e $F1 = \frac{2 \times 0{,}25 \times 0{,}71}{0{,}25 + 0{,}71} = 0{,}37$ tanto para o Sistema A quanto para o Sistema B. Esse resultado não é o ideal pois podemos ver que o Sistema A recuperou os documentos relevantes mais perto do topo do *ranking* e por isso deveria receber um escore mais alto do que o Sistema B. Esse problema ocorre porque precisão, revocação e medida-F são baseadas em conjuntos – e conjuntos, por definição, não têm ordenação. Então concluímos que precisão, revocação e F1 **não são adequadas para avaliar *rankings***. Uma boa métrica para avaliar resultados ranqueados deve conseguir calcular quantos documentos relevantes foram recuperados e o **quão próximos estão do topo do ranking**. Com isso em mente, a métrica **precisão média** (do inglês, *average precision*, AP) foi proposta.

$$AP = \frac{\sum_{k=1}^{n} P(k) \times rel(k)}{\#\ relevantes}$$

onde $rel(k)$ é uma função de relevância binária que resulta em 1 caso o documento na k-ésima posição do *ranking* seja relevante. Os documentos relevantes que são foram recuperados serão penalizados com $P = 0$. Em outras palavras, percorreremos o *ranking* calculando, para cada documento relevante recuperado, a precisão naquele ponto. Para o Sistema A da Figura 2.4, temos a seguinte $AP = \frac{1/1 + 2/2 + 3/4 + 4/5 + 5/7}{7} = 0{,}61$. E para o Sistema B temos $AP = \frac{1/9 + 2/12 + 3/14 + 4/17 + 5/20}{7} = 0{,}14$. Concluímos, então, que a AP consegue capturar a fato do Sistema A ter sido superior ao B.

Quando temos um conjunto de consultas, calculamos a média das precisões médias (do inglês, *mean average precision*, MAP), que é simplesmente a média das APs para $|Q|$

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

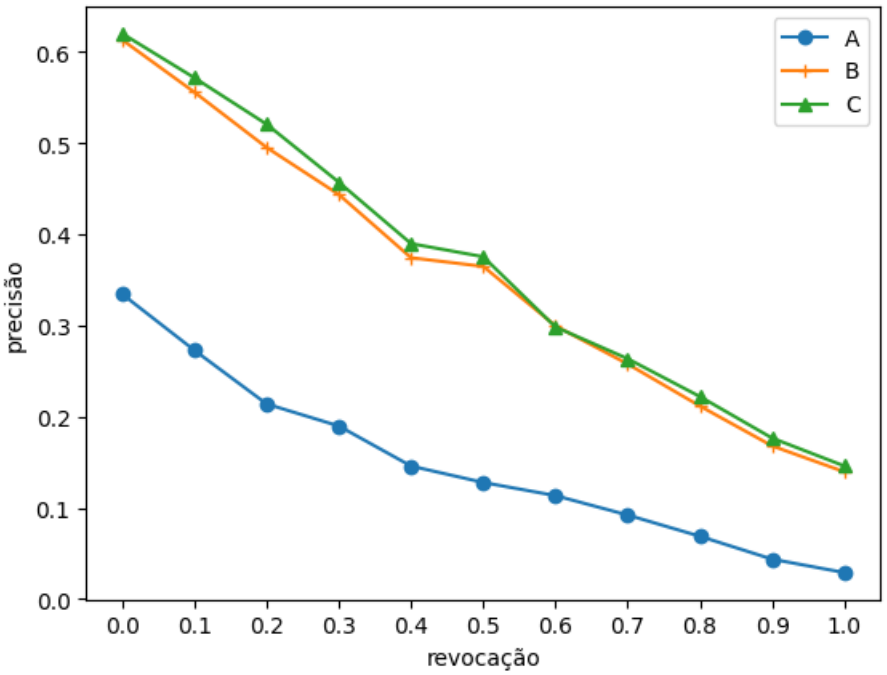
©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

consultas.

$$MAP = \frac{\sum_{q=1}^{|Q|} AP_q}{|Q|}$$

Muitas vezes, é útil fazermos uma comparação visual do desempenho de dois sistemas. Neste caso, podemos utilizar as **curvas de revocação e precisão**. Para desenhar as curvas, precisamos ter os valores de precisão calculados para 11 pontos de revocação padrão (de 0 a 1 com incrementos de 0,1). Para calcular a precisão nesses 11 pontos, é necessário utilizar uma regra de interpolação que diz que "a precisão interpolada para um nível de revocação j é o maior valor de precisão para qualquer nível de revocação maior ou igual a j". Aplicando-se a regra da interpolação, teremos curvas monotonicamente decrescentes, como as que vemos na Figura 2.5. Quanto maior a área sob a curva, melhor o resultado do sistema. Podemos ver claramente que a função A é pior do que as outras duas e que as funções B e C obtiveram resultados muito similares, com uma leve superioridade da função C. Para sabermos se as diferenças entre os sistemas são significativas, é comum fazermos testes estatísticos. O teste-T pareado é bastante usado e podemos fazê-lo comparando as precisões médias (AP) para um mesmo conjunto de consultas executadas em sistemas diferentes. Para os resultados das consultas que deram origem às curvas da Figura 2.5, quando comparamos os resultados da função C com a função A, o teste-T resulta em um p-valor = 4.6e-08. Utilizando um nível de significância $\alpha = 0.01$, a interpretação é que a função C é significativamente melhor do que a função A pois o p-valor é $\ll 0.01$. Já quando comparamos a função C com a função B, o p-valor é 0.12. Então concluímos que não há diferença significativa entre B e C pois o p-valor é > 0.01.

Figura 2.5: Exemplos de curvas de revocação e precisão para a comparação de três funções de *ranking*.



Além da MAP, é comum avaliar-se *rankings* com a métrica **P@**k, onde k é uma posição

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

específica do *ranking* (ex: 1, 5, 10, 100 etc.). Basta calcular o número de documentos relevantes encontrados até a k-ésima posição do *ranking*. Por exemplo, a P@10 conta quantos documentos relevantes foram encontrados até a décima posição do *ranking*. Essas métricas são especialmente úteis quando não sabemos o número de documentos relevantes para cada consulta, como é o caso das buscas na Web.

Outra métrica por vezes usada é o *mean reciprocal rank* (MRR) que calcula o inverso da posição do primeiro documento relevante retornado. Por exemplo, caso o primeiro documento relevante esteja na segunda posição do *ranking*, $MRR = 1/2$. Para agregar os resultados para um conjunto de consultas, fazemos a média:

$$MRR = \frac{1}{|Q|} \sum_{i=1}^{|Q|} \frac{1}{rank_i}$$

Até agora, abordamos apenas métricas que utilizam a relevância como binária. Contudo, em alguns casos, é importante diferenciar um documento muito relevante de um documentos marginalmente relevante. Nesses casos, podemos calcular o **ganho acumulado descontado normalizado** (do inglês, *normalized discounted cumulative gain*, NDCG) (Järvelin; Kekäläinen, 2002). Para entender o NDCG, primeiro temos que entender o DCG e o IDCG.

$$DCG = \sum_{i=1}^{n} \frac{2^{rel_i} - 1}{log_2(i+1)}$$

onde n é o número de documentos recuperados, i é a posição do documento no *ranking* e rel_i é o nível de relevância do documento (i.e., podemos usar quatro níveis onde o escore 3 indica um documento altamente relevante, 2 indica um documento moderadamente relevante e 1 indica um documento marginalmente relevante e 0 é um documento não relevante). O IDCG é o DGC ideal, ou seja, a pontuação de DCG obtida se a ordenação fosse perfeita. Em outras palavras, primeiramente teríamos todos os documentos de escore 3, seguidos pelos de escore 2, depois os de escore 1 e por fim os de escore zero. O NDCG é a fração $\frac{DCG}{IDCG}$ A normalização é importante para fazer com que a métrica fique no intervalo [0,1].

Além dessas medidas, muitas outras continuam sendo propostas e apresentadas nos fóruns de RI. Contudo, ainda não são amplamente adotadas. Há também muitos estudos sobre a validade e estabilidade das métricas, como por exemplo o trabalho de Buckley; Voorhees (2017).

2.4.2 Coleções de Teste

O cálculo das métricas apresentadas na Seção 2.4.1 só é possível caso exista uma coleção de teste com três componentes: (i) um conjunto de documentos, (ii) um conjunto de tópicos de consulta e (iii) os julgamentos de relevância. O primeiro componente é o mais fácil de ser obtido, pois é trivial fazer uma coleta a partir da Web e obter um grande número de documentos. O segundo componente, i.e., as consultas, necessita da intervenção humana na sua elaboração pois é preciso imaginar o tipo de consulta que um usuário que estivesse pesquisando a coleção de documentos poderia fazer. É comum a avaliação de consultas ad hoc utilizar por volta de 50 tópicos de consulta. O maior custo na elaboração de uma coleção de testes está no terceiro componente, i.e., os julgamentos de relevância. Os julgamentos de relevância dizem quais são os documentos relevantes para cada tópico de consulta e necessitam de intervenção humana na sua elaboração.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

A primeira coleção de testes elaborada, a Cranfield Collection, tem 1398 resumos de artigos sobre aerodinâmica e 225 tópicos de consulta. A elaboração dos julgamentos de relevância foi exaustiva, ou seja, a relevância de todos os documentos foi avaliada em relação a cada consulta. Além disso, cada par {documento, consulta} foi julgado por dois avaliadores – ou seja 1398 documentos × 225 consultas × 2 avaliadores = 629.100, i.e., mais de meio milhão de avaliações para uma coleção muito pequena. Aplicar essa mesma abordagem em uma coleção com 100 mil documentos e 50 consultas, demandaria 10 milhões de julgamentos de relevância, o que seria impraticável.

A solução amplamente adotada foi a criação de um método em que nem todos os documentos são avaliados em relação às consultas – apenas os documentos que têm alguma chance de serem relevantes. Essa foi a ideia do método *pooling* (Spärck Jones, 1975) que tem esse nome pois são criados repositórios (*pools*) com o subconjunto de documentos que serão avaliados. O processo envolve a criação de um repositório para cada consulta. Os repositórios são populados com documentos recuperados utilizando diferentes funções de *ranking*, variando-se parâmetros e sistemas. Os documentos do repositório são então ordenados de acordo com alguma heurística – por exemplo documentos que foram retornados mais vezes ficam no topo do *ranking*. Então, apenas os n primeiros são analisados pelos avaliadores humanos. Os documentos que não foram avaliados são considerados como não relevantes. É possível que essa estratégia deixe de identificar alguns documentos relevantes, mas há evidências de que ainda assim ela é robusta. Um estudo realizado por Zobel (1998) identificou que a ordenação entre funções de *ranking* (ou sistemas) permanecia a mesma ainda que o conjunto de julgamentos de relevância mudasse. Em outras palavras, a superioridade de um método em relação a outro é mantida mesmo com conjuntos diferentes de julgamentos de relevância.

Mesmo com a grande redução no esforço manual de avaliação na criação de coleções trazida pelo método *pooling*, este processo ainda é bastante trabalhoso. As coleções de teste foram criadas, na imensa maioria das vezes, dentro do escopo de campanhas de avaliação. Essas campanhas tiveram um grande impacto no desenvolvimento de técnicas para RI. Para a língua inglesa, as campanhas TREC (*Text REtrieval Conference*)[12] mantiveram tarefas de recuperação ad hoc entre os anos de 1992 e 1999. Posteriormente, uma iniciativa europeia deu origem às campanhas CLEF (*Cross Language Evaluation Forum*) que entre 2000 e 2009 organizaram tarefas de recuperação ad hoc para idiomas europeus. Semelhante a essas campanhas, há o NTCIR[13] dedicado aos idiomas do leste da Ásia.

Foi no contexto das campanhas CLEF e por iniciativa da Linguateca[14] que a primeira coleção de teste para a língua portuguesa foi elaborada – a coleção CHAVE[15] (Santos; Rocha, 2004b). A coleção CHAVE é composta por cerca de 209 mil notícias dos jornais Folha de São Paulo e Público (de Portugal), publicadas nos anos de 1994 e 1995. Há 100 tópicos de consulta com os seus respectivos julgamentos de relevância. Por muitos anos, esta foi a única coleção de teste disponível para português e sem dúvida ainda representa o recurso mais valioso para a recuperação ad hoc neste idioma.

A segunda coleção de teste para a recuperação ad hoc na língua portuguesa é a coleção REGIS[16] (*Retrieval Evaluation for Geoscientific Information Systems*). Ela é composta por teses, dissertações e artigos científicos no domínio geocientífico. São mais de 20 mil

[12] https://trec.nist.gov/
[13] https://research.nii.ac.jp/ntcir/
[14] https://www.linguateca.pt/
[15] https://www.linguateca.pt/CLEF/
[16] https://github.com/Petroles/regis-collection

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

documentos e 34 tópicos de consulta. Uma característica distintiva é que os documentos são bem mais longos do que os encontrados em coleções de notícias de jornal e tratam de um domínio mais restrito. Além disso, os julgamentos de relevância foram feitos em quatro níveis (muito relevante, moderadamente relevante, marginalmente relevante e não relevante).

Recentemente, uma versão traduzida da MS MARCO (*Microsoft Machine Reading Comprehension*)[17] foi disponibilizada para português e mais 12 idiomas (Bonifacio et al., 2021). A MS MARCO contém um conjunto de *datasets* que são amplamente utilizados no treinamento de algoritmos de aprendizado profundo. Para RI, há uma coleção com 8,8 milhões de documentos e 6980 consultas. As consultas foram obtidas a partir de buscas reais submetidas ao Bing e os documentos são passagens curtas extraídas de páginas web.

O Quadro 2.2 mostra exemplos de tópicos de consulta das três coleções de teste existentes para a língua portuguesa. O formato adotado pelas coleções CHAVE e REGIS é o formato tradicional utilizado pelas campanhas TREC e CLEF. Cada tópico de consulta é composto por um identificador, um **título** que descreve sucintamente o tópico, uma **descrição** mais detalhada e uma **narrativa** que visa auxiliar o avaliador humano a distinguir os documentos relevantes dos não relevantes. Já os tópicos da MS MARCO são diferentes, há apenas um identificador e o texto da consulta submetida ao motor de busca.

Quadro 2.2: Exemplos de tópicos de consulta das coleções de teste de RI em português

CHAVE

<top>
<num> C267 </num>
<PT-title> Melhor Filme Estrangeiro </PT-title>
<PT-desc> Quais foram os filmes candidatos ao Oscar de Melhor Filme Estrangeiro? </PT-desc>
<PT-narr> Documentos relevantes devem indicar o título e nacionalidade dos filmes indicados para o Óscar na categoria de Melhor Filme em Língua Estrangeira. </PT-narr>
</top>

REGIS

<top>
<num>Q5</num>
<title>Permeabilidade em Marlim</title>
<desc>Informações sobre o campo de Marlim, mas não dos campos de “Marlim Sul” ou de “Marlim Leste”.</desc>
<narr>Interessam documentos que contenham informação sobre a permeabilidade das rochas do campo de Marlim. Principalmente, interessam informações quantitativas (como dados de permeabilidade expressos em Darcies ou milidarcies).</narr>
</top>

MS MARCO

350,732	como limpar copos de melamina manchados
1,067,511	por que meu ligamento colateral lateral dói
906,071	o que adicionar à sua água para ser mais saudável

[17] https://microsoft.github.io/msmarco/

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

2.5 Modificação Automática de Consultas

Conforme visto na Introdução, o problema central da RI é casar a consulta com os documentos que a satisfazem. Contudo, o casamento puramente léxico aplicado pelos Modelos Clássicos (vistos na Seção 2.3) implica que uma consulta somente recupera documentos que possuam alguma de suas palavras-chave. A consulta não é capaz de recuperar um documento que contenha apenas sinônimos das suas palavras-chave e isso tem um impacto negativo na revocação. Por exemplo, se um usuário consulta "tributos pagos por donos de carros" um documento que não contenha as palavras da consulta mas mencione "imposto sobre a propriedade de veículos automotores" não seria recuperado. Desde os anos 1960, os pesquisadores vêm propondo formas de solucionar esses problemas. Essa sub-área de pesquisa é conhecida como **modificação automática de consultas**[18] (*automatic query modification*) AQM e conta com dezenas de milhares de artigos publicados. Esta seção apresenta um breve resumo e fornece referências para que o leitor possa encontrar mais informações. A AQM é um ponto forte de interseção entre RI e PLN pois os diversos métodos adotados em PLN para solucionar o problema da sinonímia encontram aqui uma ótima área de aplicação.

Um bom ponto de partida para quem quer se inteirar-se sobre esse tema é o *survey* de Carpineto; Romano (2012). vamos adotar a taxonomia introduzida nesse trabalho para classificar as abordagens para AQM.

- **Análise Linguística.** Uma das técnicas mais simples que pode ser enquadrada nessa categoria é o uso de *stemming* (ver Seção 2.2.1.3). O *stemming* reduz as formas variantes de uma palavra ao mesmo radical e assim conseguiria, por exemplo, que uma consulta pelo termo "imposto" recuperasse documentos com a forma plural "impostos". Contudo, o *stemming* não é capaz de recuperar um documento que contenha apenas o sinônimo "tributo". Para isso, poderíamos nos valer de dicionários de sinônimos, tesauros ou ontologias. A modificação da consulta consiste em adicionar termos sinônimos à consulta original. A adição de sinônimos irá recuperar mais documentos. O problema ocorre quando os sinônimos adicionados mudam o significado da consulta, por exemplo "homenagem" é sinônimo de "tributo", mas sua adição à consulta iria trazer documentos irrelevantes, baixando a precisão. Alguns domínios como a medicina dispõe de tesauros que podem ser diretamente usados. Por outro lado, essa é uma exceção – a maioria dos domínios não dispõem desses recursos e a sua construção manual é cara. Uma solução seria descobrir automaticamente essas relações a partir de *corpora*.
- **Técnicas baseadas em *corpora*.** Esse grupo de técnicas analisa o conteúdo da coleção de documentos para extrair padrões de coocorrência de termos. A ideia é que termos que coocorrem (i.e., que frequentemente aparecem juntos nos mesmos documentos) estão correlacionados e a sua adição à consulta pode fornecer contexto adicional. A informação de coocorrência pode ser obtida a partir de métodos diferentes como a Indexação Semântica Latente (Deerwester et al., 1990) ou *word embeddings* (Mikolov et al., 2013b) discutidas no Capítulo Representação vetorial e semântica distribucional. Há experimentos que obtiveram resultados positivos (Kuzi et al., 2016), contudo se as embeddings forem obtidas a partir de textos genéricos e a coleção for de um domínio específico, os termos adicionados podem piorar a

[18] O termo **expansão automática de consultas** é bastante empregado. Contudo, a consulta nem sempre é expandida e, por isso, preferimos adotar o termo **modificação automática de consultas**.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

qualidade do *ranking*.

- **Técnicas baseadas no resultado da consulta** analisam os documentos recuperados pela consulta inicial para então modificá-la. A primeira técnica nessa categoria é a **realimentação de relevância** (em inglês, *relevance feedback*) (Rocchio-Jr, 1971). Ela consiste em solicitar que o usuário marque alguns documentos que ele considera relevantes e não relevantes. A reformulação da consulta consiste em modificar seu vetor para que ele se aproxime mais dos vetores dos documentos marcados como relevantes e se afaste dos vetores dos documentos não relevantes. Essa técnica costuma obter bons resultados, mas o ponto negativo é que os usuários não gostam de investir seu tempo fornecendo feedback. A alternativa então é supor que os primeiros k documentos recuperados pela consulta original sejam relevantes e modificar o vetor da consulta com base neles. Este processo é conhecido como **pseudo realimentação de relevância** e pode ter o efeito negativo de piorar ainda mais os resultados caso o primeiro conjunto de documentos recuperados não tenha sido bom. Outras propostas que se enquadram nessa categoria incluem o uso de *word embeddings* geradas a partir do resultado da consulta (Diaz et al., 2016).
- **Análise de logs de consulta** de motores de busca pode fornecer termos que os usuários frequentemente adicionam às suas consultas. Esses termos podem então ser usados para expandir a consulta original. Essa abordagem mostrou resultados positivos em (Cui et al., 2002).

Mais recentemente, uma nova gama de trabalhos têm focado no caminho inverso, i.e., expandir os documentos. Nessa linha de investigação, destaca-se o trabalho de Nogueira et al. (2019) em que os documentos são enriquecidos com consultas que poderiam ser feitas com o intuito de recuperar o documento.

2.6 Ferramentas e Bibliotecas

Há várias ferramentas (ou sistemas) de RI disponíveis tanto para fins comerciais como para fins de pesquisa. Esta seção aborda algumas das mais utilizadas.

2.6.1 Sistemas Comerciais

Os sistemas comerciais mais populares são da Apache e baseiam-se no **Lucene**[19], uma biblioteca Java de código aberto. A biblioteca não é um sistema de RI completo. Com base nela, foram desenvolvidos o **Solr**[20] e o **Elasticsearch**[21] que fornecem interfaces de consulta mais amigável além de indexação distribuída para fins de escalabilidade. Tanto o Solr como o Elasticsearch são ferramentas de código aberto, possuem vasta documentação e são amplamente adotados na indústria. O Elasticsearch fornece melhor suporte a sistemas distribuídos e pode ser integrado com ferramentas de análise e visualização. Por outro lado, o Solr possui mais flexibilidade para customização.

2.6.2 Ferramentas para Pesquisa

Há várias ferramentas de código aberto construídas por pesquisadores dá área de RI que estão disponíveis e são usadas em trabalhos acadêmicos. Dentre elas, as mais utilizadas

[19] https://lucene.apache.org/
[20] https://solr.apache.org/
[21] https://www.elastic.co/

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

atualmente são:

- O **Anserini**[22] e sua interface em Python (Pyserini[23]) foram elaborados por pesquisadores das universidades de Delaware e Waterloo (Yang et al., 2017) e tem o objetivo de facilitar o processo experimental em RI com ênfase na reprodutibilidade.
- O **Terrier**[24] e sua interface em Python, o Pyterrier[25] (Macdonald; Tonellotto, 2020), foram criados por pesquisadores da universidade de Glasgow que tem longa tradição na pesquisa em RI. O Terrier também dispõe de uma interface gráfica que permite indexar e consultar uma coleção sem a necessidade de escrever código.
- O projeto **Lemur**[26] reúne uma série de componentes. O Indri é o motor de busca em si. O projeto é uma colaboração entre as Universidades de Massachusetts e Carnegie Mellon. O sistema Galago também é parte desse projeto e foi disponibilizado em conjunto com o livro de Croft et al. (2010).

2.7 Conclusão

Este capítulo forneceu uma visão geral sobre a área de RI. O processo típico de RI foi introduzido, bem como os modelos clássicos. Demos ênfase à metodologia de avaliação dos resultados das consultas, dada a importância que esse tema tem na área. Também foram apontadas coleções de teste em português e ferramentas que podem ser usadas tanto no meio acadêmico como na indústria.

Nessa primeira versão, o capítulo não abordou um tópico relevante e atual: a RI utilizando vetores densos. Em especial, o emprego da arquitetura Transformers (Vaswani et al., 2017b) tem sido bastante difundido em RI (assim como em diversas tarefas de PLN). Por ora, o leitor interessado pode referir-se o artigo de (Lin et al., 2020b) que apresenta um levantamento abrangente.

Agradecimentos

Agradeço a Helena Caseli e a Graça Nunes pelo convite a contribuir com este livro. Também sou grata a Adriana Pagano, Felipe Paula, Lucas Pessutto, Luciana Bencke e João Comba pela revisão e comentários.

[22] https://github.com/castorini/anserini
[23] http://pyserini.io/
[24] http://terrier.org/
[25] https://github.com/terrier-org/pyterrier
[26] https://lemurproject.org/

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Capítulo 3

Extração de Informação

Daniela Barreiro Claro
Joaquim Santos
Marlo Souza
Renata Vieira
Vládia Pinheiro

Publicado em: 26/09/2023
Atualizado em: 16/04/2026

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

3.1 Introdução

A Extração de Informação (EI) é desenvolvida com o objetivo de se obter informação estruturada de dados não-estruturados (Jurafsky; Martin, 2023; Konstantinova, 2014).

Os primeiros trabalhos a debruçarem-se sobre o problema remontam à década de 1970, com a aplicação de gramáticas formais e *parsers* sintáticos para a estruturação de informação em domínios como prontuários médicos (Sager, 1978; Sager et al., 1987) e textos jornalísticos (DeJong, 1979). A comunidade científica demonstrou grande interesse pela área nas décadas posteriores devido à sua utilidade prática, seu foco no processamento de dados reais, suas tarefas bem-definidas e a facilidade de mensurar a qualidade dos resultados em comparação com o desempenho humano na mesma tarefa (Cowie; Lehnert, 1996).

Para autores como Eisenstein (2019) e Jurafsky; Martin (2023), a EI é normalmente dividida em diversas tarefas de interesse, com foco no tipo de informação a ser extraída do texto. Entre as mais comumente citadas na literatura estão o Reconhecimento de Entidades Nomeadas (REN), a Extração de Relações (ER) e a Extração de Eventos (EE).

O Reconhecimento de Entidades Nomeadas (REN) consiste em identificar e classificar entidades mencionadas em textos através de designadores rígidos como nomes próprios, expressões temporais e espécies biológicas (Nadeau, 2007). Esse é considerado por alguns como um primeiro passo na análise semântica de um texto (Santos; Cardoso, 2007a), pois permite identificar as entidades às quais se faz referência nele.

A Extração de Relações (ER), também chamada de extração de informação tradicional ou somente extração de informação, por sua vez, diz respeito à identificação de relacionamentos semânticos entre duas ou mais entidades, ou seja, identificar "quem fez o que para quem e quando". Ananiadou; Mcnaught (2005) a definem como o processo de extrair fatos (em nossa terminologia, relacionamentos) a partir de uma fonte textual e representá-los a partir de um gabarito (em inglês, *template*). As *relações* são elementos essenciais para o entendimento da informação relatada no texto e sua identificação é passo essencial para a estruturação da mesma. Assim, identificar relações entre entidades é tarefa essencial

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

para construção de bases de conhecimento e de grande utilidade na construção de soluções para a resposta automática a perguntas (em inglês, *query answering*), sumarização, recuperação de informação e mais (Nasar et al., 2021).

A extração de eventos consiste na tarefa de identificação de uma menção a um evento em uma sentença e, se existirem, extração de outras informações sobre o evento. Um evento pode, por sua vez, ser entendido como uma ocorrência específica envolvendo participantes (Consortium, 2005), i.e., algo que acontece e que pode ser descrito como uma mudança de estado da qual participam entidades como agentes. Devido a intrínseca natureza temporal dos eventos, tal problema possui uma natureza mais complexa e costuma possuir tratamento específico.

Assim, nesse capítulo, iniciaremos com um pouco de história da Extração de Informação (EI) e sua evolução para Extração de Informação Aberta, e destacaremos as tarefas de Reconhecimeno de Entidades Nomeadas (REN) e Extração de Relação (ER).

3.2 Um pouco de história

Os primeiros trabalhos que abordaram o problema de EI dos quais temos conhecimento surgiram no final da década de 1970. Esses primeiros trabalhos da década de 1970 e 1980 tinham como modelo geral a aplicação de regras para a identificação de informações especificadas em um gabarito. Tais sistemas empregavam analisadores sintáticos (*parsers*) e regras definidas especificamente para o domínio e gênero textual estudado.

Entre esses primeiros trabalhos, estão aqueles de Sager (1978), Sager et al. (1987), de DeJong (1979) e de Cowie (1983). Sager et al. exploraram como identificar informações do estado de saúde de pacientes através dos textos de prontuários médicos. DeJong (1979), por sua vez, descrevem o sistema FRUMP que, a partir de um *parser* e regras de análise conceitual baseadas em uma arquitetura cognitiva proposta pelos autores e no conceito de dependência conceitual de Schank et al. (1973), processavam textos de notícias e realizavam tarefas como sumarização e identificação de papéis semânticos associados aos constituintes da sentença. Cowie (1983), por fim, descreve um sistema que emprega regras simples de segmentação e análise sintática rasa para identificar propriedades de plantas a partir de textos descritivos no campo da botânica. Diferente dos métodos anteriores, o trabalho dos autores se baseia em grande parte no estudo de padrões de descrição das informações a serem identificadas, em detrimento do emprego de *parsers* robustos da língua.

A década de 1990 traz um grande interesse na área de EI com a implementação das conferências MUC (do inglês, *Message Understanding Conference*, ou Conferência de Compreensão de Mensagem), promovidas pela Agência de Projetos de Pesquisa Avançada de Defesa (DARPA, do inglês *Defense Advanced Research Projects Agency*). As conferências MUC, realizadas e financiadas pelo exército americano, representaram um esforço em avançar a tecnologia de EI e consistiam de tarefas de avaliação conjunta de métodos desenvolvidos por pesquisadores para problemas propostos pelos organizadores. As sete conferências realizadas de 1987 a 1997, foram cruciais para definir aspectos centrais da área, como estruturar a tarefa de ER, definindo suas métricas de avaliação, e propor a tarefa de REN (Grishman; Sundheim, 1996).

A partir da MUC-3, em 1991, a conferência passa a ter foco no processamento de textos jornalísticos em detrimento dos relatórios militares utilizados anteriormente (DARPA, 1991). Com a disponibilidade de dados e o incentivo no desenvolvimento de soluções para

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

a tarefa, vemos na década de 1990 o surgimento das primeiras aplicações comerciais de EI, como o JASPER (Andersen et al., 1992)., construído para a agência de notícias Reuters.

A MUC-6, ocorrida em 1995, introduz a tarefa de REN com o intuito de ser uma tarefa de uso prático, independente de domínio e que poderia ser realizada automaticamente em um futuro próximo (Grishman; Sundheim, 1996). Enquanto os trabalhos em REN se avolumaram a partir de sua proposição na MUC-6, trabalhos anteriores como Rau (1991) e Wolinski et al. (1995) já se debruçavam sobre o problema de identificação e classificação de nomes próprios. Desde então, o interesse na tarefa cresceu significativamente e outras conferências de avaliação conjunta têm sido dedicadas a essa tarefa, como a *Automatic Content Extraction* (ACE) e a conferência Avaliação de Sistemas de Reconhecimento de Entidades Mencionadas (HAREM), dedicada exclusivamente à língua portuguesa, com sua primeira edição em 2005 (Santos; Cardoso, 2007a).

Por outro lado, houve um crescimento de abordagens baseadas em dados nesta década, a partir da análise de *corpora*. Tais esforços são impulsionados pelos resultados positivos na área, como o trabalho de Hearst (1992). Logo, métodos baseados em dados passaram também a explorar o emprego de análise estatística e aprendizado de máquina na construção de padrões para a extração de relações (Riloff et al., 1993, 1999; Roark; Charniak, 2000; Soderland et al., 1995)

Não foi somente na extração de padrões que métodos de aprendizado de máquina, em particular aprendizado supervisionado, foram aplicados. A década de 2000 viu a proliferação de métodos supervisionados aplicados à ER (Culotta et al., 2006; Kambhatla, 2004; Zelenko et al., 2003; Zhao; Grishman, 2005) e ao REN (Asahara; Matsumoto, 2003; McCallum; Li, 2003; Sekine, 1998).

Devido à dificuldade de construção de dados para treinamento e padrões para extração, além da pouca adaptabilidade dos sistemas construídos para outros escopos e domínios, nos anos 2000, sistemas baseados em métodos de aprendizado semi-supervisionado, como o DIPRE (Brin, 1998) e Snowball (Agichtein; Gravano, 2000) começaram a aparecer, juntamente com os estudos sobre expansão automatizada de anotações (*bootstrapping*) (Riloff et al., 1999). Também para entidades nomeadas, estudos investigaram como utilizar recursos da Web (Etzioni et al., 2005; Nadeau, 2007) ou *corpora* (Cucchiarelli; Velardi, 2001) para aprender entidades com pouco ou nenhum esforço de anotação.

Buscando superar as dificuldades da limitação de escopo, i.e. das relações-alvo a serem extraídas e categorias de entidades a serem identificadas, ainda restritas à definição de padrões desde a criação dessas tarefas, Banko et al. (2007) propõe a tarefa de extração de informação aberta (EIA), também conhecida como **Open Information Extraction**, OpenIE ou OIE, a qual busca extrair todas as relações possíveis expressas em um texto, sem necessidade de pré-definição de relações e entidades.

Devido ao recente sucesso da aplicação de métodos baseados em redes neurais, em particular *deep learning* e grandes modelos de linguagem, às tarefas de Processamento de Linguagem Natural, uma tendência atual da área se delineou como o estudo de arquiteturas neurais para os problemas de EI e a geração de grandes conjuntos de dados por supervisão fraca. *Surveys* recentes, como (Cui et al., 2018; Konstantinova, 2014; Nasar et al., 2021), nos mostram a evolução da área em direção à aplicação de métodos neurais. Na vertente de geração de dados, vemos o emprego da Wikipédia e Freebase como fontes mais usadas para obter anotações de entidades e relações em textos (Nguyen et al., 2016; Smirnova; Cudré-Mauroux, 2018; Takamatsu et al., 2012).

Porém, toda a tarefa de EI necessita de uma concordância entre as definições de Entidade e Relação. Neste sentido, a próxima seção discute a conceituação de relação adotada neste

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

capítulo, assim como o conceito de entidade.

3.3 Conceituação formal: Relação e Entidade

A natureza das relações estudadas na área de Extração de Informação e os critérios para reconhecer sua ocorrência em um texto têm recebido pouca atenção na literatura. Este é um passo importante para estabelecer metodologias adequadas para avaliar os sistemas, bem como para criar conjuntos de dados que possam apoiar a criação de sistemas futuros.

Enquanto as noções de Relação e Entidade são de grande importância e já bem estudadas nas áreas de Computação, Linguística, Ciência da Informação e Filosofia da Linguagem, esses conceitos não são empregados de forma consistente entre as áreas, ou mesmo entre suas subáreas.

3.3.1 Entidade

Para Chen (1976), uma entidade é um objeto que pode ser concreto, tal como pessoa, livro, casa ou ainda abstrato, tal como um emprego, um sentimento, uma disciplina. As entidades podem estabelecer relações entre si. Duas ou mais entidades são vinculadas, ou seja conectadas por uma relação[1].

Tradicionalmente em reconhecimento de entidades nomeadas, as entidades consideradas são aquelas referenciadas por um nome próprio, acrescidas das referências temporais e valores que são expressões numéricas. Essas expressões, portanto, geralmente não constituem uma entrada em uma base lexical. Porém a tarefa se expandiu para domínios especializados, onde as entidades de interesse são mais conceituais. No domínio bio-médico por exemplo, podemos ter como exemplo de entidades de interesse, sintomas e tratamento que não são referenciadas por nomes próprios.

3.3.2 Relação

Os conceitos de *relação* e *relacionamento* são noções fundamentais que vêm sendo estudadas em áreas como Ciência da Computação, Linguística e Filosofia.

No campo de bancos de dados e modelagem conceitual, Chen (1976) define um relacionamento, no contexto da modelagem de Entidade-Relacionamento, como uma associação entre entidades. Guarino; Guizzardi (2015), por sua vez, estudando a natureza ontológica dos relacionamentos com base na semântica de veridadores (*truthmaker semantics*) (Fine, 2017), postulam relacionamentos como entidades que atuam como veridadores (*thruthmakers*) de alguma proposição relacionando duas ou mais entidades, ou seja, *uma relação mantida entre essas entidades.* Um veridador é um elemento cuja existência torna verdadeira uma proposição particular. Por exemplo, considerando a sentença (1) "*a* é uma maçã", a existência de um objeto denotado pelo nome *a* que por acaso é uma maçã é uma condição suficiente para a verdade da frase (1). Como tal, dizemos que esse objeto é o veridador de (1). Tal definição nos permite adotar critérios ontológicos para validar a existência de relacionamentos a partir da informação relatada em um texto e, por isso, adotaremos tal definição de relacionamento neste capítulo.

O conceito de relações é muito menos consistente na literatura. Ainda na área de modelagem conceitual, Guarino; Guizzardi (2015) definem as relações como proposições para as quais os relacionamentos são veridadores e, portanto, possuem conteúdo proposicional.

[1] Em nossa terminologia, por um relacionamento.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Assim, podemos entender uma relação como um tipo para entidades como relacionamentos. Ou seja, relações são universais ontológicos que descrevem a natureza dos relacionamentos.

Xavier et al. (2015), no entanto, argumentam que a noção de relacionamento adotada na área de Extração de Informação é mais geral do que isso, não se limitando àquelas entre objetos e propriedades, mas também àquelas que descrevem ou implicam propriedades de classes gerais como descrito pela sentença (2) "Filósofos são autores de Livros". Assim, para o contexto de EI consideramos relações como tipos de relacionamentos de primeira ou segunda ordem. Isso significa que uma relação é um tipo de relacionamento que existe entre objetos, suas propriedades e classes de objetos ou suas propriedades.

Enquanto os métodos tradicionais de Extração de Informação dependem de um conjunto pré-existente de relações semânticas bem definidas que são relevantes para um domínio específico, a noção de "relação" e "entidade" na literatura da área mais recente, tais como a Extração de Informação Aberta, requer mais aprofundamento por demandar um significado diferente, principalmente com diferente visões de autores. Esta indeterminação terminológica pode trazer problemas para comparar os resultados dos métodos propostos ou para reutilizar os conjuntos de dados criados na área.

As seções seguintes exploram essas duas áreas: Extração de Informação e Extração de Informação Aberta.

3.4 Extração de Informação (EI)

A Extração de Informação é caracterizada por obter informação estruturada a partir de textos, sendo entidades ou fatos, i.e. relacionamentos entre entidades, de tipos previamente definidos, conforme exemplo no Quadro 3.1. Métodos com limitação de escopo possuem como uma de suas principais desvantagens a necessidade de intervenção humana para especificar novos fatos a serem extraídos. Esta limitação impede que sistemas de Extração de Informação, doravante denominados de EI tradicional extraiam fatos fora do escopo pré-definido.

Quadro 3.1: Exemplos de relações específicas na EI tradicional

Relação Específica	**Exemplo de Sentença**	**Extração**
location-of(algo/alguém, local)	Um aluno pode ser encontrado na escola	location-of(aluno, escola)
is-a(subclasse, superclasse)	Salvador é uma cidade	is-a(Salvador, cidade)
part-of(todo, parte)	Roda é um componente de um carro	part-of(roda, carro)

Fonte: (Souza; Claro, 2014)

3.4.1 Reconhecimento de Entidades Nomeadas

O Reconhecimento de Entidades Nomeadas (REN) consiste na tarefa de identificar e classificar expressões linguísticas, denominadas entidades nomeadas (EN), que referenciam entidades específicas num domínio de discurso, como nomes próprios, expressões temporais e espécies biológicas (Mota et al., 2007; Nadeau, 2007). De uma forma geral, o REN pode ser dividido em duas etapas: a identificação (ou delimitação) da expressão, na qual

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

as palavras que formam a EN são selecionadas; a classificação, em que é atribuída uma categoria semântica à EN.

A classificação das ENs determina os tipos de entidades a serem consideradas e são especificadas a partir do escopo definido previamente para a tarefa. Algumas das categorias mais comumente utilizadas incluem as entidades que referenciam Pessoas Singulares (antropônimos); Coletivas (empresas e organizações) e Lugares (topônimos) (Mota et al., 2007). Para exemplificar tomemos a sentença: "Renata Silva e Maria Costa palestraram na Universidade Federal da Bahia". No exemplo temos três ENs: "Renata Silva", "Maria Costa", "Universidade Federal da Bahia", sendo as duas primeiras correspondentes à categoria semântica Pessoa e a última, à categoria semântica Organização. Entretanto, existem outras categorias de ENs, como as menções a Obras (por exemplo, "Código Da Vinci"); Acontecimentos (por exemplo, "Festa de Santo Antônio"), Tempo (por exemplo, "meio-dia"); Coisa (por exemplo, "barco"), entre outras.

O REN é uma tarefa com grande importância para o Processamento de Linguagem Natural, pois consiste numa primeira tarefa de análise semântica de um texto, com potencial aplicações a diversas tarefas. Por exemplo, em sistemas de perguntas e respostas, as perguntas frequentemente se referem a informações sobre entidades. Também, métodos de identificação de estruturas mais complexas, como eventos ou relações, dependem do bom desempenho do REN como uma etapa de pré-processamento (Socher et al., 2012; Zelenko et al., 2003).

3.4.2 Extração de Relações

A tarefa de extração de relações (ou de relacionamentos) (ER) refere-se a identificar relacionamentos entre entidades de um determinado escopo mencionadas em um texto (Jurafsky; Martin, 2023). O escopo, no contexto da ER, refere-se a um conjunto de relações-alvo de um determinado domínio de conhecimento ou aplicação a ser investigado. Por exemplo, o Quadro 3.2 apresenta alguns exemplos de relações no domínio de geografia brasileira. Na descrição das relações, os elementos em negrito referem-se às entidades em um dado relacionamento descrito pelo termo em itálico.

Quadro 3.2: Exemplos de relações no domínio da geografia brasileira.

Relação	**Descrição**	**Exemplo**
Pertence(Cidade, Unidade Federativa)	Sobre uma **cidade** que está localizada em uma determinada **Unidade Federativa**, dizemos que a primeira *pertence* a esta última.	Pertence(Salvador, Bahia)
Tem_Prefeito(Cidade, Pessoa)	Uma **pessoa** que executa a função administrativa de gestão do executivo em nível municipal de uma dada **cidade** é denominada de seu(sua) *prefeito(a)*.	Tem_Prefeito(Salvador, Bruno Reis)
Fundação(Cidade, Data)	A **data** em que uma **cidade** foi fundada, é dita sua data de *fundação*.	Fundação(Salvador, 29 de março de 1549)

Nesse contexto, a delimitação de um escopo ou domínio de interesse, concentra-se na determinação das relações a serem processadas, i.e. nos tipos de relacionamentos de interesse, assim como da natureza das entidades associadas por tais relações.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

3.4.3 Extração Conjunta de Entidades e Relações

As tarefas de reconhecimento de entidades nomeadas e extração de relações são interdependentes, no sentido de que a definição do escopo a ser estudado delimita tanto as categorias e natureza das entidades a serem extraídas, como também as relações entre essas entidades. Também, note-se que, pelo fato de as relações serem comumente definidas entre entidades de tipo especificado, como o caso da relação *Tem_Prefeito* no Quadro 3.2 que ocorre entre entidades das classes **Cidade** e **Pessoa**, tanto as informações das entidades mencionadas no texto são úteis para a extração de relações, quanto a informação das relações identificadas pode ser útil ao processo de identificação de entidades.

De fato, na literatura recente, existem vários trabalhos que consideram a tarefa de extração conjunta de entidades e relações (ERE, do inglês *Entity and Relation Joint Extraction*), composta das tarefas de REN e ER (Agichtein; Gravano, 2000; Shaowei et al., 2022; Yuan et al., 2021b). Enquanto normalmente abordagens estruturam suas soluções de forma sequencial, usualmente realizando REN inicialmente e, posteriormente, realizando ER, como nos trabalhos de (Hasegawa et al., 2004) e de (Socher et al., 2012), a literatura recente aponta para as vantagens da identificação conjunta ao permitir um melhor aprendizado de restrições para identificação de entidades e relações, c.f. o recente *survey* realizado por (Shaowei et al., 2022) sobre métodos para tal tarefa.

3.4.4 Métodos empregados para EI na literatura

Várias abordagens foram adotadas para o problema de EI durante seu desenvolvimento histórico. Enquanto abordagens iniciais privilegiavam métodos ricos em conhecimento, como regras e recursos linguísticos e de conhecimento de mundo, a literatura recente na área privilegia métodos baseados em dados, como o aprendizado de máquina, com o recente emprego de arquiteturas neurais aos problemas.

A seguir faremos uma breve apresentação das abordagens descritas na literatura para os problemas de EI.

3.4.4.1 REN

As abordagens iniciais para REN baseavam-se, majoritariamente, no emprego de regras léxico-sintáticas e consulta a almanaques (*gazetteers*). Tais abordagens dependem da construção de listas de nomes próprios como antropônimos, topônimos etc., e outras palavras, como “Ltda.”, “Jr.” etc., que auxiliam no processo de identificação e classificação de ENs complexas ou desconhecidas. Essa é, por exemplo, a abordagem empregada por Wolinski et al. (1995) que combina almanaques e regras para a identificação e classificação de ENs. Posteriormente, almanaques foram também empregados em conjunção com métodos baseados em dados, como o trabalho de Florian et al. (2003) que os emprega aliados aos classificadores, enquanto Liu et al. (2019a) os utilizam durante o treinamento de uma rede neural, como um sinal de treinamento (parte da função de perda, ou *loss* em inglês).

Muitos trabalhos debruçaram-se também sobre o problema de construção automática ou semi-automática de almanaques, dos quais os trabalhos de Nadeau (2007), de Riloff et al. (1999) e de Etzioni et al. (2005) são alguns dos mais importantes.

Enquanto as abordagens iniciais para o problema baseavam-se em regras, com a disponibilidade de dados anotados para a tarefa, tais métodos foram rapidamente suplantados por métodos baseados em dados, tais como: os métodos baseados em classificação (Asahara;

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Matsumoto, 2003; Sekine, 1998) e classificação sequencial (Bikel et al., 1999; McCallum; Li, 2003).

A redução de REN à tarefa de classificação sequencial merece destaque pelos bons resultados obtidos. Tal redução se dá através de um esquema de codificação do problema que nos permite representar fragmentos textuais e sua classificação como um problema de rotulação ou etiquetação.

Partindo-se do pressuposto de que os fragmentos textuais descrevendo entidades nomeadas são contíguos, podemos codificar a tarefa de delimitação de entidades como classificação sequencial empregando rótulos que descrevem os limites de uma EN, e.g. o esquema BIO com os rótulos **B** (do inglês, *begin*) para designar a palavra inicial de uma EN, **I** (do inglês, *inside*) para designar palavras que fazem parte da EN mas não a iniciam e **O** (do inglês, *outside*) para designar palavras que não pertencem a uma entidade. Da mesma forma, podemos estender nosso esquema de codificação para incluir as classes de interesse. Assim, seguindo o esquema BIO, teremos os rótulos **B-PER** e **I-PER** para descrever entidades da classe **Pessoa**.

A redução do problema de REN à classificação sequencial está ilustrada no Exemplo 3.1.

Exemplo 3.1:

> Renata/B-PER Silva/I-PER e/O Maria/B-PER Costa/I-PER palestraram/O na/O Universidade/B-ORG Federal/I-ORG da/I-ORG Bahia/I-ORG.

Recentemente, destacam-se na literatura abordagens baseadas em redes neurais profundas, com uma grande concentração nos últimos anos em modelos gerativos de linguagem, devido aos resultados positivos obtidos por tais arquiteturas em diversas tarefas.

Na literatura são de grande destaque os modelos recentes BART (Lewis et al., 2020), RoBERTa (Liu et al., 2019c), T5 (Raffel et al., 2020), BERT (Devlin et al., 2019) e GPT-3 (Brown et al., 2020c), conforme descritos no Capítulo Modelos de linguagem.

Similarmente, na língua portuguesa, nas duas edições do HAREM (Mota; Santos, 2008; Santos; Cardoso, 2007b), o primeiro esforço sistemático de desenvolvimento de soluções para a tarefa na língua, a maioria dos sistemas participantes baseava-se em métodos ricos em conhecimento, como regras e almanaques. De fato, nas duas avaliações, somente os sistemas MALINCHE (Solorio, 2007), NEURA (Ferrández et al., 2007) e R3M (Mota, 2008) não se baseavam em regras. Métodos baseados em classificação sequencial se seguiram para a língua portuguesa, como o RELP-CRF (Amaral; Vieira, 2014) baseado em um classificador sequencial. Mais recentemente, abordagens baseadas em redes neurais e modelos de linguagem foram desenvolvidas tornando-se o estado da arte da tarefa na língua. A Tabela 3.1 apresenta o atual estado da arte em português, com base no *corpus* HAREM. A métrica de avaliação apresentada, medida F1, será discutida na Seção 3.6.

Tabela 3.1: Trabalhos estado da arte no REN em português

Modelo	Medida F1
BERT-CRF (Souza et al., 2020a)	83,70%
BiLSTM-CRF+FlairBBP (Santos et al., 2019b)	82,26%
LSTM-CRF (Castro et al., 2018)	76,27%
CharWNN (Santos; Guimarães, 2015)	65,41%

Souza et al. (2020a) desenvolveram um modelo BERT para o Português com 2,68 bilhões de *tokens* e aplicaram o modelo em um classificador CRF. Santos et al., avaliaram

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

o impacto do modelo contextualizado Flair Embeddings aplicado a tarefa de REN junto com uma rede neural BiLSTM-CRF. Os autores também desenvolveram um modelo Flair Embeddings para o português, o FlairBBP, treinado com 4,9 bilhões de *tokens* (Santos et al., 2019b). Castro et al. (2018) utilizou uma rede LSTM e um classificador CRF junto com modelos *word embeddings* pré-treinados. Santos; Guimarães (2015) desenvolveram uma rede neural convolucional capaz de capturar características a nível de caracteres e também de incorporar *word embeddings* pré-treinados.

3.4.4.1.1 Reconhecimento de Entidades em Domínios Específicos

O reconhecimento de entidades tem sido aplicado em muitas áreas específicas, como direito, saúde e geologia. Nesses casos há uma demanda de adaptação dos modelos preditivos de acordo com a nova linguagem especializada do domínio e um novo conjunto de rótulos que devem ser aprendidos. Da mesma forma, são necessários novos conjuntos de dados para o processo de aprendizado, uma vez que abordagens de aprendizado de máquina necessitam de exemplos anotados para se chegar a um modelo preditivo eficaz.

Muitos trabalhos endereçam domínios específicos, citamos exemplos em diversas línguas. Para o inglês, uma rede neural BiLSTM-CRF para o domínio biomédico é proposta em (Habibi et al., 2017).

Um conjunto de dados do domínio jurídico em língua alemã é apresentado por Leitner et al. (2019), que empregam redes neurais BiLSTM para a rotulação dos textos. Em (Qiu et al., 2019), uma rede neural BiLSTM-CRF com mecanismo de atenção é aplicada para reconhecer entidades geológicas para a língua chinesa.

Para o português, um *corpus* para detecção de eventos de quedas de pacientes em prontuários eletrônicos é descrito em (Santos et al., 2020a). Os autores usaram uma rede neural BiLSTM-CRF+Flair para gerar um modelo classificador de *tokens*. Um *corpus* no domínio jurídico, tendo categorias específicas como legislação e jurisprudência é proposto por Araujo et al. (2018), que usaram uma rede neural BiLSTM-CRF para criar um primeiro *baseline* para esse *corpus*. Ademais, Consoli et al. (2020) analisam um *corpus* no domínio de geologia usando uma rede neural BiLSTM-CRF com um modelo contextualizado Flair Embeddings.

Desde a terceira edição, este livro conta com um capítulo de reconhecimento de entidades nomeadas no domínio jurídico, o Capítulo 13.

3.4.4.2 Extração de Relações

As abordagens iniciais para o problema de ER baseavam-se na definição de gabaritos e regras de extração, com base em informação sintática obtida de analisadores sintáticos rasos ou profundos (Cowie, 1983; Sager, 1978). Tais métodos foram rapidamente suplantados por métodos baseados em dados e padrões obtidos de *corpora*, como os famosos padrões de Hearst (1992) para identificação de relações de hiponímia.

O trabalho de Hearst (1992) se baseou na definição de padrões léxico-sintáticos para expressão de relações de hiponímia e hiperonímia a partir de uma análise de *corpus*. Ao escolher a relação de hiponímia, que ocorre em todo domínio, e padrões gerais baseados em aspectos da língua, como os representados no Quadro 3.3, garante-se a generalização dos padrões obtidos para diversos domínios e aplicações.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Quadro 3.3: Exemplos de Padrões de Hearst para hiponímia

Padrão	**Exemplo**	**Extração**
NP, tais quais $\{NP\}^*NP$	… países, tais quais França, Brasil e China	Is_a(França, país)
$NP\{, NP\}^*$ e outros(as) NP	Contusões, feridas, osso quebrados e outras lesões	Is_a(contusão, lesão)

Devido à dificuldade de construção manual das regras, os métodos de Riloff et al. (1993), empregam heurísticas para geração de padrões baseadas em informação gramatical, e de Soderland et al. (1995), que se baseia numa semântica de quadros (*frames*) empregando um analisador semântico e medidas de qualidade de identificação de exemplos, baseado no percentual de acerto sobre relacionamentos previamente conhecidos, para identificação de quadros relevantes.

As abordagens baseadas em aprendizado de máquina, hoje as mais comuns e com melhor desempenho na literatura (Konstantinova, 2014; Nasar et al., 2021) dividem-se em abordagens que realizam reconhecimento de entidades e extração de relações de forma conjunta e separada.

Abordagens baseadas na realização de REN e ER de forma separada baseiam-se em um fluxo de processamento em que, em geral, as entidades são identificadas primeiro e a tarefa de ER se reduz a identificar quando uma sentença ou fragmento textual denota uma relação semântica entre duas entidades. Consideremos o Exemplo 3.2, retirado de (Socher et al., 2012):

Exemplo 3.2:

> [Gripe aviária]$_{e1}$ é uma doença infecciosa causada pelo vírus da [influenza tipo a]$_{e2}$

Podemos, então, reduzir o problema de identificar a relação Causa-Efeito($e1$,$e2$) a um problema de classificação textual, identificando se a sentença acima fornece indícios para a expressão da relação de interesse. As soluções propostas na literatura para o problema são variadas e baseadas em diferentes métodos.

Zelenko et al. (2003), por exemplo, propõem funções de *kernel* para árvores sintáticas rasas, i.e. funções que descrevem medidas de similaridade entre tais árvores. Eles empregam tais medidas para treinar um classificador de *perceptron* com votação (*voted perceptron*) sobre relações no domínio de organizações extraídas de um *corpus* de textos jornalísticos. De forma similar, Zhao; Grishman (2005) empregam diferentes funções de *kernel* sobre informações sintáticas relevantes para a identificação de relação e argumentos visando treinar um classificador SVM sobre o *corpus* de ER da conferência ACE.

Culotta et al. (2006), por outro lado, empregam um classificador sequencial baseado em modelos escondidos de Markov para identificação de relações em um texto. Ao restringir sua análise a textos biográficos, os autores reduzem o processo de identificar instâncias de relações à identificação de fragmento textual que delimita o argumento e sua classificação, tarefa para a qual a classificação sequencial já é comumente utilizada. Consideremos o Exemplo 3.3 sobre George W. Bush, retirado de (Culotta et al., 2006):

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Exemplo 3.3:

George é filho de $\underbrace{\text{George H. W. Bush}}_{\text{pai}}$ e $\underbrace{\text{Barbara Bush}}_{\text{mãe}}$.

Ao identificar o papel de pai e mãe, os autores conseguem construir a relação Pai(George H. W. Bush, George W. Bush) e Mãe(Barbara Bush, George W. Bush).

Métodos baseados em redes neurais, de forma geral, costumam empregar técnicas de aprendizado de representação (Bengio et al., 2013) para aprender representações do conteúdo semântico dos fragmentos textuais e reduzem o problema de ER à classificação textual. É o caso de Socher et al. (2012), que propõem a MV-RNN, uma rede neural que constrói um espaço de representação baseado em matrizes e vetores com o objetivo de capturar a composicionalidade de sentido de sintagmas e sentenças e os aplica para ER. Similarmente, Zeng et al. (2014) e Wang et al. (2016) empregam redes neurais convolucionais para obter representações vetoriais de sentenças que serão empregadas no processo de classificação quanto à relação expressa pela mesma.

3.4.4.3 Extração Conjunta de Entidades e Relações

Abordagens baseadas em identificação sequencial de entidades e relações possuem desvantagens observadas na literatura. Primeiramente, como a ER é guiada pelas entidades identificadas no processo de REN, a propagação de erros da primeira tarefa pode ter impacto considerável na performance dos sistemas desenvolvidos. Segundo, uma vez que o contexto determinado limita tanto as tarefas de REN, quanto as de ER, existe uma interdependência entre as tarefas. Assim, propostas visando realizar a extração de entidades e relações de forma conjunta começaram a surgir na literatura recente, ganhando certo interesse da comunidade.

As abordagens empregadas para tal tarefa são diversificadas, incluindo desde métodos de aprendizado relacional a redes neurais

Roth; Yih (2007) propõem a utilização de métodos de programação inteira ao problema, baseados na teoria estatística de aprendizado relacional. Os autores utilizam classificadores locais para a identificação de entidades e relações e um classificador global que combina as informações dos classificadores locais em uma predição que maximiza a qualidade da extração, codificada por meio de restrições em programação inteira. Também baseados em modelos estatísticos, Yu; Lam (2010) propõem o uso de modelos gráficos globais para identificação de um descritor de relação e uma segmentação do texto para identificação dos argumentos.

Li; Ji (2014) e Miwa; Bansal (2016), por sua vez, reduzem a tarefa de ERE à classificação sequencial, utilizando redes neurais recorrentes bidirecionais sequenciais e estruturadas com base na estrutura superficial e na árvore de dependências sintáticas da entrada para identificação conjunta de entidades e relações.

3.5 Extração de Informação Aberta

A Extração de Informação Aberta (EIA), também conhecida como *Open Information Extraction*, Open IE ou OIE em inglês, é a tarefa de extrair informações estruturadas de documentos sem necessitar da pré-definição do contexto da tarefa, i.e. das relações e tipos de entidade de interesse. A tarefa foi inicialmente proposta pelo trabalho de (Banko et

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

al., 2007) e ganhou popularidade nas últimas décadas devido à sua aplicabilidade para processar e estruturar o conhecimento a partir de grandes volumes de dados disponíveis na Web, seguindo o paradigma da Web como um *Corpus* (WaC) (Meyer et al., 2003).

A EIA surge visando generalizar a tarefa de Extração de Relações. A principal diferença entre as duas abordagens, porém, reside na dependência da ER de uma especificação prévia do domínio de aplicação, bem como das relações alvo a serem identificadas, que a EIA visa eliminar.

Seguindo o trabalho original de Banko et al. (2007), que propôs o sistema TextRunner, vários métodos e sistemas para EIA foram propostos na literatura (Del Corro; Gemulla, 2013; Fader et al., 2011; Xavier et al., 2015), mas, como observado por Glauber; Claro (2018), os principais avanços na área se concentraram principalmente no idioma inglês.

A EIA para a língua portuguesa tem uma história bastante recente. A partir dos trabalhos de Souza; Claro (2014), Pereira; Pinheiro (2015) e de (Barbosa et al., 2016), têm crescido o número de estudos sobre a tarefa assim como os resultados obtidos por esses estudos, com recentes desenvolvimentos de métodos (Oliveira et al., 2022c; Sena et al., 2017; Sena; Claro, 2019, 2020; Souza et al., 2018), construção do *corpus* (Glauber et al., 2018) e avaliação dos sistemas disponíveis (Glauber et al., 2019a, 2019b; Malenchini et al., 2019).

Embora a área tenha visto um crescimento recente para o desenvolvimento de métodos para línguas como o inglês, principalmente com a aplicação de métodos supervisionados e redes neurais, esses avanços ainda não foram incorporados na literatura sobre EIA para a língua portuguesa. A razão para isso é principalmente a falta de recursos linguísticos disponíveis para orientar o desenvolvimento de pesquisas para a língua. Embora o foco no idioma inglês possa ser devido ao seu uso generalizado em todo o mundo, foi reconhecido pela comunidade científica que esse foco no inglês com suas características particulares pode introduzir algum viés na área (Bender, 2009).

Assim, esta seção aborda EIA para a língua portuguesa, incluindo uma formalização e a evolução das abordagens da área.

3.5.1 Formalização

A tarefa de EIA pode ser formalmente definida sendo $X = \langle x_1, x_2, \cdots, x_n \rangle$ uma sentença composta de *tokens* x_i. Um extrator EIA é uma função que mapeia X em um conjunto $Y = \langle y_1, y_2, \cdots, y_j \rangle$ como um conjunto de tuplas $y_i = \langle rel_i, arg1_i, arg2_i, \cdots, argn_i \rangle$, que descrevem as informações expressas na sentença X. Neste capítulo, consideramos que as tuplas estão sempre no formato $y = (arg_1, rel, arg_2)$, onde $arg1$ e $arg2$ são sintagmas nominais, não necessariamente formados por *tokens* presentes em X, e rel é um descritor de um relacionamento entre arg_1 e arg_2. Não consideraremos extrações formadas por mais de dois argumentos neste capítulo.

3.5.2 Abordagens

Os primeiros métodos de EIA empregavam padrões de inspiração linguística para extração, como ArgOE (Gamallo; Garcia, 2015), ou adaptação de métodos para a língua inglesa, como SGS (Souza et al., 2018), InferReVerbPT Sena et al. (2017) e RePort Pereira; Pinheiro (2015). Os trabalhos são principalmente influenciados por métodos baseados no inglês da chamada segunda geração de EIA (Fader et al., 2011).

O primeiro sistema de EIA para o português de que temos conhecimento foi o DepOE (Gamallo et al., 2012). Ele executa a extração aberta multilíngue de triplas (inglês,

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

espanhol, português e galego) usando o analisador sintático de dependências baseado em regras *DepPattern*. No entanto, nenhuma avaliação ou resultados são relatados para a língua portuguesa. Os autores apresentam somente uma comparação dos seus resultados com *Reverb* na língua inglesa.

Souza; Claro (2014) se propuseram a analisar o conjunto de características mais representativas da língua portuguesa para a identificação de extrações válidas no contexto de EIA, tal qual empregado na língua inglesa com o sistema ReVerb (Fader et al., 2011).

O sistema RePort (Pereira; Pinheiro, 2015), por outro lado, é uma adaptação do ReVerb para a língua portuguesa baseada em análise sintática rasa com regras sintáticas e lexicais. Os autores relatam que suas extrações apresentam grande similaridade com suas correlatas extraídas pelo ReVerb (dos textos traduzidos para o inglês).

O RELP, proposto por Abreu; Vieira (2017), é um sistema aberto de extração de relações que extrai relações entre entidades nomeadas em um domínio de organização aplicando classificação sequencial com CRF (*Conditional Random Fields*). O sistema RelP extrai qualquer descritor de relação que expressa um relacionamento entre pares de entidades nomeadas (Organização, Pessoa ou Lugar), caracterizando-o como uma abordagem híbrida da REN com a EIA.

O InferReVerbPT desenvolvido por Sena et al. (2017) baseia-se numa adaptação do sistema ReVerb para a língua portuguesa, expandindo-o com a extração de relacionamentos implícitos obtidos por inferência por propriedades de simetria e transitividade das relações com inferência transitiva e simétrica. Um classificador SVM foi empregado para realizar a inferência baseado nas propriedades semânticas do verbo central no descritor de relação.

Souza et al. (2018) analisaram que a maior desvantagem dos estudos baseados em recursos linguísticos, como dados anotados, reside na escassez de tais recursos na maioria dos idiomas além do inglês. Assim, para mitigar esse problema, eles propõem um método de classificação de fatos baseado na similaridade de estruturas gramaticais (SGS). Sua abordagem modela estruturas morfosintáticas dos fatos (triplas descrevendo relacionamentos) para identificar padrões de semelhanças que podem ser usados para distinguir entre fatos válidos e inválidos. Eles aplicaram algoritmos de isomorfismo de grafos para detectar subgrafos descrevendo tais padrões.

Um novo sistema de EIA baseado em análise de dependência foi proposto por Gamallo; Garcia (2015), chamado ArgOE. Tal sistema é multilíngue, baseado em heurísticas e utiliza a informação de dependência sintáticas do texto para analisar a estrutura de dependência do verbo, bem como um conjunto de regras para gerar os relacionamentos. A introdução de um Analisador de Dependência em sistemas de EIA focados inteiramente na língua portuguesa foi feita pelos autores Oliveira et al. (2022c). O DptOIE é baseado em análise de dependência e regras elaboradas manualmente. As sentenças são pré-processadas por meio de um tokenizador, um PoS *Tagger* e um analisador de dependências. Os autores propõem um acoplamento de três módulos para tratar casos particulares: conjunções coordenadas, orações subordinadas e aposto.

Com a evolução dos métodos de EIA para a língua inglesa utilizando os modelos neurais, novas abordagens foram propostas também para a língua portuguesa.

O primeiro trabalho que utilizou aprendizado supervisionado com rede neural profunda para o português foi o de Ro et al. (2020) que descreve o sistema **Multi2OIE**. Os autores utilizaram o modelo de linguagem BERT multilíngue (Devlin et al., 2019) para obter representações vetoriais das palavras e reduzem a tarefa de EIA à classificação sequencial, identificado os fragmentos do texto que determinam os argumentos (arg_1, arg_2) e o descritor de relação (rel). Seu sistema foi capaz de produzir extrações para vários idiomas

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

(inglês, português e espanhol), treinados, entretanto, sobre dados traduzidos do inglês.

Stanovsky et al. (2018) propuseram uma abordagem de EIA para a língua inglesa baseada em triplas. Os mesmos fazem uso de uma classificação sequencial cuja limitação define uma tripla extraída para cada sentença. Este método utiliza uma arquitetura de Redes Neurais Recursivas (RNN) para realizar EIA. A EIA é formulada como uma tarefa de rotulagem de sequências, utilizando estratégias semelhantes às que foram aplicadas anteriormente a tarefas como o Reconhecimento de Entidades Nomeadas. Já os autores em Cui et al. (2018) e Zhang et al. (2017) propõem modelar o problema da EIA como um problema de aprendizado sequência a sequência (*seq2seq*). Eles definem uma estrutura *encoder-decoder* para aprender argumentos e tuplas de relação inicializadas a partir de um sistema de EIA.

Seguindo o trabalho de (Stanovsky et al., 2018), em 2022, Cabral et al. (2022) propuseram **PortNOIE**, uma arquitetura neural para EIA em português que combina representações contextuais de palavras com codificadores neurais para extrair relacionamentos baseado em classificação sequencial iterativa. Diferente de outros métodos de classificação sequencial para EIA, os autores focam na extração de múltiplas triplas de uma mesma sentença.

3.6 Avaliação

A avaliação sistemática de sistemas de EI foi estabelecida primeiramente nas conferências MUC, em particular na sua segunda edição, com o estabelecimento de gabaritos-padrão que deveriam ser utilizados por todos os sistemas participantes e a adoção de métricas de qualidade, baseadas naquelas usadas na área de recuperação de informação, que foram abordadas no Capítulo 2. Para avaliar a tarefa de extração de relações, a MUC-2 estabeleceu como métricas de qualidade do sistema as medidas de precisão e cobertura, também denominada de *Recall* ou Revocação.

A precisão de um sistema reflete a qualidade de suas extrações, i.e., quantas das extrações realizadas estão corretas, dado um *corpus* de teste. A medida de precisão pode ser calculada como:

$$P = \frac{\#(\text{relacionamentos corretamente extraídos})}{\#(\text{relacionamentos extraídos pelo sistema})} \tag{3.1}$$

A cobertura também conhecida como revocação, reflete quão abrangente um sistema é em suas extrações, i.e., quantas das extrações a serem realizadas em um *corpus* de teste, o sistema é capaz de realizar. A medida de cobertura pode ser calculada como:

$$R = \frac{\#(\text{relacionamentos extraídos})}{\#(\text{relacionamentos no } \textit{corpus})} \tag{3.2}$$

Enquanto a MUC-3 adicionou duas novas métricas de avaliação, a saber sobre-geração (*overgeneration*) e sub-geração (*fallout*), tais métricas receberam pouco interesse na literatura. De fato, Lehnert; Sundheim (1991) argumentam que tais métricas foram pouco informativas ou difíceis de calcular para a tarefa de EI e, portanto, abandonadas. Foi também empregado nessa conferência um sistema automático de avaliação disponibilizado às equipes participantes que permitiu uma maior compreensão do modelo de avaliação e, como discutem Lehnert; Sundheim (1991), um avanço qualitativo nos sistemas gerados.

Além das medidas de precisão e cobertura, assim como em tarefas de classificação de texto e recuperação de informação, utilizamos a média harmônica entre essas medidas,

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

chamada medida F1, a fim de condensar a informação contida nas duas. A medida F1 pode ser calculada como:

$$F1 = \frac{2 * P * R}{P + R} \tag{3.3}$$

A avaliação da tarefa de REN segue padrões semelhantes aos aplicados à tarefa de ER. De fato, desde a MUC-6 (Grishman; Sundheim, 1996), as medidas de precisão, cobertura e F1 tem sido usada consistentemente como métricas de avaliação da tarefa de REN em diversos esforços de avaliação, como a CoNNL (Sang; De Meulder, 2003), para a língua inglesa, e das duas edições do HAREM (Gonçalo Oliveira et al., 2008; Santos et al., 2007), com excessão à ACE (Doddington et al., 2004) que apresenta uma combinação da tarefa de REN com reconhecimento de co-referência entre entidades e utiliza um sistema de pontuação próprio.

A avaliação de sistemas de EIA, por sua vez, possui algumas peculiaridades que precisam ser discutidas. Uma vez que a tarefa é postulada por Banko et al. (2007) como a extração de todas as relações identificadas em um dado fragmento textual, sem limitação de domínio de interesse, tal tarefa impõe imensa dificuldade aos esforços de avaliação.

De fato, Glauber et al. (2018) relatam um esforço de anotação de dados para a tarefa em língua portuguesa em que foram identificados por anotadores humanos mais de 400 relacionamentos em um *corpus* de 25 sentenças retiradas de textos jornalísticos e de enciclopédia. Assim, a avaliação de EIA deu-se, em grande parte de seu desenvolvimento e maturação, em conjuntos de dados não anotados, recorrendo a avaliações qualitativas das saídas dos sistemas e comparação direta por humanos das extrações obtidas.

Nesses esforços de avaliação, a precisão do sistema pode ser mensurada a partir da avaliação humana das saídas. Não é possível, entretanto, avaliar medidas como cobertura e F1, dada a inexistência de uma referência do conjunto total de relacionamentos a serem identificados. Assim, os autores da área propuseram diferentes métricas a fim de estimar tais valores, como a métrica rendimento (*yield*) (Fader et al., 2011; Schmitz et al., 2012).

A métrica de rendimento consiste no núemro de extrações válidas, i.e. corretas, de um dado sistema. Como calcular tal medida é, na maioria dos casos, impraticável dada a grande quantidade de extrações realizadas pelos sistemas, ela pode ser estimada a partir da precisão do sistema calculada sobre uma amostra aleatória das extrações realizadas (P'). Assim, podemos estimar o rendimento como:

$$Y = P' \cdot \#(\text{extrações realizadas}) \tag{3.4}$$

Foi também explorada a estratégia de criação (semi-)automática de conjuntos de dados usando vários sistemas (Del Corro; Gemulla, 2013), estratégias de supervisão fraca (Smirnova; Cudré-Mauroux, 2018), ou a geração de *corpora* para a tarefa a partir da transformação de anotações de tarefas próximas, como identificação de papéis temáticos (*Semantic Role Labeling*) por (Stanovsky et al., 2018). *Corpora* gerados de forma semi-automática vêm ganhando atenção na literatura recente, particularmente para a língua inglesa, devido a necessidade de dados anotados para se utilizar técnicas de aprendizado de máquina e redes neurais em EIA. *Corpora* como o OIE2016 (Stanovsky et al., 2018), Wire57 (Léchelle et al., 2018) e CARB (Bhardwaj et al., 2019) vêm se tornando *corpora* de referência em língua inglesa para o problema, apesar dos problemas existentes na construção de tais recursos – a não exaustividade das relações anotadas.

Para a língua portuguesa, foram propostas algumas iniciativas para avaliar os sistemas da OIE. Uma avaliação conjunta foi promovida durante o Fórum Ibérico de Avaliação de

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Línguas (IberLEF) em 2019 (Collovini et al., 2019). A avaliação foi feita usando o *corpus* proposto por Glauber et al. (2018), que é composto por 442 relacionamentos extraídos de 25 frases de fontes como a seção em português da Wikipédia, o *corpus* CETENFolha, resenhas de filmes do portal Adoro Cinema[2] e o *corpus* Europarl. Apesar desta tarefa ter contemplado quatro cenários de avaliação, a avaliação geral dos sistemas permaneceu consistente nos diferentes cenários, indicando robustez nos resultados da avaliação. No geral, os sistemas DPTOIE (Oliveira et al., 2022c) e Linguakit (Gamallo; Garcia, 2015) tiveram o melhor desempenho, com o Linguakit2 dominando as avaliações de correspondência exata e o DPTOIE as avaliações de correspondências parciais (Collovini et al., 2019).

Outra abordagem de avaliação foi idealizada por (Malenchini et al., 2019). Seu foco foi a avaliação extrínseca dos sistemas de EIA através de sua contribuição na tarefa de respostas automáticas a perguntas. Os autores apresentaram um conjunto de dados de referência (*benchmark*) para avaliação extrínseca de sistemas de EIA em textos de língua portuguesa. Os sistemas que alcançaram os melhores valores na avaliação realizada pelos autores foram os sistemas ArgOE (Gamallo; Garcia, 2015), DependentIE (Glauber et al., 2019a) e DptOIE (Oliveira et al., 2022c).

3.7 Considerações finais

Este capítulo descreveu uma visão geral da área de Extração de Informação, apresentando a Extração de Informação Tradicional e a Extração de Informação Aberta. Transversalmente, apresentamos as formalizações necessárias e os conceitos fundamentais para a compreensão da EIA, assim como a avaliação da área e as heranças de outras áreas afins, tais como RI.

Nessa primeira versão, este capítulo descreveu de maneira bem sucinta as abordagens propostas para EI e EIA durante seu desenvolvimento histórico e as abordagens atuais da literatura, como as utilizando modelos de linguagens. Especificamente, a utilização da arquitetura Transformers, descritas no Capítulo Modelos de linguagem para as tarefas de EI e EIA tem sido bastante difundida para a língua inglesa e tem atuado em diversas áreas da PLN.

Agradecimentos

Agradecemos as colaborações dos autores deste Capítulo e suas indicações, assim como agradecemos a Adriana Pagano e Aline Macohin pela revisão e comentários.

[2] https://www.adorocinema.com/

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Capítulo 4

Tradução Automática

Abordagens e Avaliação

Sheila Castilho
Helena de Medeiros Caseli

Publicado em: 26/09/2023

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

4.1 Introdução

A tradução automática (TA), também conhecida como tradução de máquina (em inglês, *machine translation* ou MT), refere-se à tradução de um texto eletrônico por um computador de uma língua para outra sem intervenção humana. Nesse sentido, convencionou-se chamar de língua (ou texto) fonte a língua de partida (origem) e língua (ou texto) alvo a língua de chegada (destino ou saída). Além de envolver a análise e interpretação (NLU) da língua-fonte e a geração (NLG) da língua-alvo, há a premissa fundamental de gerar uma saída que seja semanticamente equivalente (transmite o mesmo significado) à entrada.

Nos últimos anos, a TA evoluiu significativamente com o avanço de modelos estatísticos e neurais. Atualmente, ela é amplamente utilizada em todo o mundo por governos, indústria da tradução, consumidores finais e em pesquisas em uma variedade de aplicações.

Os primeiros sistemas bem-sucedidos de TA datam do final dos anos 1950 e início dos anos 1960[1], com os experimentos de Georgetown. No entanto, é possível encontrar referências a tentativas de tradução automática no século XVII (Hutchins, 2001). Desde então, diferentes abordagens para a TA foram desenvolvidas, incluindo abordagens baseadas em regras, exemplos, estatísticas e, mais recentemente, a TA neural, apresentadas brevemente nas diversas seções deste capítulo.

Hoje em dia, a TA desempenha um papel importante não apenas no âmbito comercial, mas também no âmbito social e político. Ela é amplamente utilizada em diversas aplicações de comunicação, que incluem[2]:

- Texto para texto: o usuário insere um texto-fonte e obtém uma versão traduzida em formato de texto;
- Texto para fala: o usuário insere um texto-fonte e obtém uma versão falada na língua-alvo;
- Fala para texto: o usuário fala no idioma fonte e obtém uma versão traduzida em formato de texto;
- Fala para fala: o usuário fala no idioma fonte e obtém uma versão falada na língua-alvo;

[1]Para uma descrição abrangente da história da TA, sugere-se consultar alguns livros e capítulos sobre o assunto: (Hutchins, 2001), (Koehn, 2009), (Hutchins, 2010) e (Koehn, 2020).

[2]Todas as aplicações citadas estão disponíveis no Google tradutor. Disponível em: https://translate.google.com.br/.

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

- Imagem (de palavras) para texto: o usuário insere uma imagem contendo um texto e obtém a tradução desse texto.

Devido à ampla utilização da TA atualmente, seu impacto pode ser observado em nossa sociedade. Por essa razão, a avaliação da TA (tema da Seção 4.3) tornou-se mais importante, visando garantir a qualidade da tradução. Além da avaliação, este capítulo descreve as principais abordagens para a TA (Seção 4.2). No decorrer deste capítulo, alguns conceitos-chave são explicados para que você possa acompanhar os desenvolvimentos.

4.2 Abordagens

A tradução automática pode ser realizada de diversas maneiras, desde a mais simples (tradução direta), que envolve a tradução palavra-a-palavra (ou sequência de palavras), até a mais utilizada na atualidade, que é a tradução baseada em redes neurais artificiais (tradução neural). Na trajetória entre a tradução direta e a tradução neural, explicaremos também abordagens intermediárias, como a baseada em regras, a tradução por interlíngua e a tradução estatística.

Para tanto, vamos traduzir a sentença "A casa do meu avô é linda." para o inglês. Como isso aconteceria de acordo com as abordagens mais utilizadas (ou tradicionais) na tradução automática?

4.2.1 Tradução direta

Na **tradução direta** ocorre o mapeamento direto de palavras-fonte para palavras-alvo sem passar por outros níveis de análise[3]. Assim, no nosso exemplo, os caracteres seriam combinados em palavras e cada uma seria mapeada diretamente para seu equivalente em inglês usando, por exemplo, um léxico bilíngue. Utilizando o léxico bilíngue disponível no github do MUSE[4] e a tradução palavra-a-palavra, nossa saída seria como apresentado em Exemplo 4.1:

Exemplo 4.1:

> Entrada: A casa do meu avô é linda.
> Saída: _ *house* _ *my granddad* _ *beautiful* .

onde as palavras para as quais não se encontrou um equivalente no léxico consultado foram substituídas por _[5].

Vejam que neste processo de tradução não há nenhum processamento referente às línguas envolvidas, uma vez que o resultado é obtido via **casamento de padrão**, seguido da substituição de uma palavra por outra com base em uma lista de pares de palavras.

[3]Vejam que aqui estamos usando "palavra" para denotar uma unidade lexical bastante comum no PLN. Contudo, vale ressaltar que outras unidades lexicais (n-grama ou expressão multipalavra, por exemplo) também podem ser usadas na tradução direta. Para saber mais sobre as unidades de processamento veja Capítulo Sequência de caracteres e palavras.

[4]https://dl.fbaipublicfiles.com/arrival/dictionaries/pt-en.txt

[5]Isso porque palavras como "a" e "do" não têm entrada no léxico consultado, provavelmente por serem *stop words*. Para as demais palavras, a tradução foi gerada considerando apenas a primeira ocorrência de equivalente em inglês encontrada. Por exemplo, para "casa" existem as opções "*house*" e "*home*", nesta ordem, e como "*house*" aparece primeiro, ela foi a escolhida para gerar a saída neste exemplo.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Obviamente a abordagem de tradução direta apresenta diversas **limitações**, como **não ser capaz de lidar com a estrutura (sintaxe) da língua**, que, como pode ser visto no Capítulo Sequência de caracteres e palavras, é fundamental para o tratamento adequado da língua. A tradução direta foi uma das primeiras abordagens a serem investigadas e não é mais utilizada nos tradutores atuais.

4.2.2 Tradução Automática Baseada em Regras

Como o nome sugere, os sistemas de TA baseados em regras (em inglês, *Rule-based Machine Translation* ou RBMT) são sistemas baseados em conhecimento desenvolvidos por meio da especificação de regras linguísticas, que levam em consideração morfologia, sintaxe e semântica das línguas envolvidas, além dos léxicos bilíngues de ambos os idiomas de origem e de destino[6]. Essas regras e léxicos são formuladas e criados manualmente por especialistas em linguagem.

A partir desses recursos, a RBMT é capaz de realizar mapeamentos mais complexos como os que consideram a sintaxe das línguas fonte e alvo. Assim, no nosso exemplo, uma possível regra que poderia ser aplicada é a que indica a inversão da posição do sujeito que possui a casa com o uso do "apóstrofo s" gerando "*My grandfather's house is beautiful.*" ao invés de "*The house of my grandfather is beautiful.*"

Assim, vê-se que o processamento automático necessário para a tradução baseada em regras é um pouco mais complexo do que aquele aplicado na tradução direta, uma vez que agora é preciso saber o papel de cada palavra na sentença-fonte, ou seja, saber que "casa" é um substantivo e que "do" é uma combinação da preposição "de" e determinante (artigo definido "o") para que a regra apresentada em Quadro 4.1 seja aplicada corretamente. A regra especifica que sempre que for encontrado, na sequência-fonte, um substantivo (<SUB>) seguido da preposição (<PREP>) "de" combinada (+) com os artigos (<DET>) "a" ou (|) "o" deve-se gerar como saída o apóstrofo (') seguido de "s" e o substantivo equivalente na língua-alvo. O símbolo "=>" separa o que deve ser considerado na língua-fonte (à esquerda) e o que deve ser gerado na língua-alvo (à direita).

Quadro 4.1: Exemplo de regra para a tradução automática baseada em regras

```
<SUB> <PREP/de>+<DET/[a|o]> => 's <SUB>
```

Para se determinar o "papel" de cada palavra na sentença o processamento necessário é o da etiquetagem morfossintática ou, do inglês, *part-of-speech tagging* descrito no Capítulo Sequência de caracteres e palavras.

Apesar de realizar um processamento automático um pouco mais complexo, a **desvantagem** da tradução baseada em regras não está aí, mas sim na necessidade de mapear o conhecimento linguístico em regras corretas, genéricas e abrangentes o suficiente para que sejam aplicáveis a vários exemplos. Vejam que esse mapeamento envolve, necessariamente, o conhecimento da língua-fonte, da língua-alvo, e de como o processo de tradução de uma para a outra deve ocorrer. Além de um **processo trabalhoso**, a geração de regras é também **limitada**, pois, como a língua está em constante mudança, o conjunto de regras gerado tem que ser constantemente atualizado e revisado. Além disso, a tradução

[6] A tradução baseada em regras enquadra-se na estratégia de tradução por transferência, na qual o mapeamento é realizado com base em uma análise da língua-fonte, seguida da aplicação de regras que fazem o mapeamento e a geração de equivalentes na língua-alvo.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

de/para outra língua necessita de um novo conjunto de regras. Isso porque a tradução por transferência entre duas línguas requer que a representação do conhecimento extraído da língua-fonte, e que vai ser mapeado para a língua-alvo, seja capaz de abrigar todas as características de ambas as línguas, tornando-a específica para aquele par. Analogamente, o desenvolvedor tem que ter muito conhecimento de ambas as línguas ou a equipe deve contar com linguistas/tradutores, o que torna os **sistemas caros** de se implementar. Além disso, a saída dos sistemas de regras pode apresentar **pouca fluência**, já que as traduções são fornecidas por meio de regras.

A grande **vantagem** dos sistemas de regras é que, como não são necessários textos bilíngues para seu treinamento, esses sistemas são excelentes para **traduções de idiomas com recursos limitados**. Além disso, esses sistemas permitem que o desenvolvedor tenha um **controle maior**, sendo possível identificar exatamente onde estão os problemas, e a saída (texto-alvo) é relativamente previsível. Tanto regras quanto léxicos podem ser **refinados e personalizados**, com a adição de mais (novas) regras e entradas bilíngues para aprimorar a tradução. Outro ponto interessante, é que o conhecimento é **legível por seres humanos** o que facilita a manutenção. Os sistemas de TA baseados em regras foram os primeiros sistemas comerciais de TA na década de 1970 e abriram caminho para mais pesquisas em TA após o relatório ALPAC, que cortou os fundos (Seção 4.3.1).

4.2.3 Tradução por interlíngua

Mas será que esse conceito de transferência entre línguas não pode ser estendido para um número maior de línguas, ou seja, considerando um cenário multilíngue? Sim, essa é a ideia da **tradução por interlíngua**, que se propõe a usar uma língua intermediária – metalíngua – que é independente das línguas envolvidas na TA e ao mesmo tempo é capaz de representar informações de qualquer outra língua. Essa metalíngua, de natureza artificial, é não ambígua e, portanto, mais simples de processar do que qualquer linguagem natural. Assim, o processo de tradução entre duas línguas quaisquer é composto de duas etapas de traduções supostamente mais simples: uma realizada entre a língua-fonte e a metalíngua, e outra realizada entre esta metalíngua e a língua-alvo.

Mas será que essa abordagem simplifica o processo? Sim. Para ficar claro, imagine o seguinte cenário onde existem documentos escritos em n línguas distintas e queremos traduzi-los de e para quaisquer dessas línguas. Poderíamos construir arranjos de n tradutores[7] automáticos para cada par de línguas (n1-n2, n1-n3, n1-n4, ..., n2-n1, n2-n3, n2-n4, ...) seguindo o modelo de transferência (tradução automática baseada em regras). Mas a tradução por interlíngua mostra-se mais vantajosa. Podemos definir uma metalíngua e dividir a tarefa em n grupos de especialistas/tradutores, cada um responsável por uma única língua, l_i (de preferência sua língua materna). Caberá a cada grupo construir um tradutor (lembre-se: bem mais simples!), ou codificador, de l_i para a metalíngua, e outro, decodificador, da metalíngua para l_i. Evidentemente todos os grupos compartilham o mesmo conhecimento sobre a língua intermediária.

Considerando que cada grupo fará sua parte, ao final, as possíveis combinações desses módulos de tradução darão origem a todos os tradutores almejados. Por exemplo, para traduzir l_i para l_j, juntamos o módulo codificador de l_i com o módulo decodificador para

[7] Para calcular a quantidade total de tradutores necessários, usamos a fórmula de arranjo de n línguas para combinação em pares (k = 2), $A_{n,2}$ que é (n!)/(n-2)!. Assim, para 5 línguas teríamos que construir 20 tradutores no sentido direita-esquerda e mais 20 tradutores no sentido esquerda-direita, totalizando 40 tradutores distintos!

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

l_j. Terão sido construídos *2n* módulos de traduções mais simples, portanto, menos esforço do que o exigido para os tradutores bilíngues. Uma outra consequência é a possibilidade de se avaliar as traduções por meio da tradução inversa, já que os módulos independentes permitem a tradução nos dois sentidos.

Esse ideal foi compartilhado algumas vezes, no passado, por vários grupos de pesquisa acadêmica, mas infelizmente a prática evidenciou vários **problemas**. O maior deles é, segundo os críticos, a **ingenuidade** em se acreditar possível criar uma linguagem capaz de representar o significado de todas as outras, portanto, universal. Um outro problema – ou decorrente deste – é a adoção unânime de uma dada língua intermediária pelos grupos de línguas distintas, onde cada um deles reivindica alterações e adaptações, escancarando a **inexistência de um núcleo verdadeiramente universal**. Por essas e outras razões é que esse modelo não substituiu o modelo por transferência bilíngue.

No final dos anos 1990, o português brasileiro estava representado, pelo NILC[8], numa iniciativa da ONU para construção de tradutores para as línguas mais faladas no mundo, o Projeto UNL[9]. Esse projeto tinha por objetivo o desenvolvimento de um sistema multilíngue de tradução automática baseada numa interlíngua de natureza semântica – a *Universal Networking Language (*UNL) – desenvolvida por pesquisadores vinculados à Universidade das Nações Unidas, órgão da ONU, em Tóquio[10].

O **paradigma linguístico** (baseado em regras e interlíngua), no qual o conhecimento linguístico é explicitamente mapeado em recursos como regras, dominou o cenário da tradução automática **até a década de 1980**, quando abordagens baseadas em *corpus* (empíricas) surgiram. Aliadas à motivação de tentar superar as limitações da tradução baseada em regras, essas abordagens foram impulsionadas por dois fatores: (1) os **avanços no hardware** necessário para processamentos computacionais mais pesados, e (2) a **disponibilidade maior de recursos bilingues**, em especial os *corpus* paralelos. As próximas seções tratam das abordagens baseadas em *corpus*: a tradução baseada em exemplos, a tradução estatística e a tradução neural.

4.2.4 Tradução Automática Baseada em Exemplos

Os sistemas de TA baseada em exemplos (do inglês, *Example-based Machine Translation* ou EBMT), também conhecidos como tradução por analogia, estão frequentemente associados à publicação do artigo de Nagao (1984), no qual o autor propõe um modelo baseado na imitação de exemplos de tradução de frases semelhantes, buscando utilizar a ideia de aprender a traduzir a partir de exemplos existentes (Koehn, 2020). Os sistemas de exemplos utilizam informações extraídas (sequências de palavras) de exemplos em *corpora* bilíngues de pares de tradução, alinhados em nível de sentença, ao qual convencionou-se chamar de *corpora* paralelos.

Por meio dessa abordagem, exemplos como os ilustrados no Quadro 4.2 serviriam de base para o sistema aprender traduções de trechos de texto, como as ilustradas em Quadro 4.3.

[8] http://www.nilc.icmc.usp.br/

[9] Esse projeto deu origem à UNDL *Foundation*.

[10] Mais informações sobre a linguagem UNL podem ser obtidas em (Uchida et al., 1999). Detalhes sobre o Projeto UNL-Brazil podem ser encontrados em (Martins et al., 2000; Nunes et al., 2003) e http://www.nilc.icmc.usp.br/nilc/projects/unl.htm.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Quadro 4.2: Exemplos para a tradução baseada em exemplos

A casa é muito bonita.	*The house is very beautiful.*
Meu avô foi internado ontem.	*My grandfather was hospitalized yesterday.*
Eu comprei uma jaqueta linda.	*I bought a beautiful jacket.*

Quadro 4.3: Trechos aprendidos

a casa	*the house*
meu avô	*my grandfather*
linda	*beautiful*

A partir dos trechos aprendidos, o sistema baseado em exemplos seria capaz de combiná-los para, a partir da entrada, gerar a saída apresentada em Exemplo 4.2:

Exemplo 4.2:

Entrada: A casa do meu avô é linda.
Saída: *The house _ my grandfather _ beautiful.*

Devido ao uso de diferentes métodos e técnicas, de acordo com Hutchins (2005, p. 63): "não parece haver um consenso claro sobre o que é ou o que não é um sistema de exemplos". Para Carl; Way (2003, p. xix), uma definição analítica dos sistemas de exemplos era difícil, pois, segundo ele, tais sistemas "assumem uma posição entre os sistemas baseado em regras e os estatísticos" ao utilizar abordagens tanto baseadas em regras quanto orientadas por dados[11]. No entanto, tanto os sistemas baseados em exemplos como os estatísticos se enquadram no paradigma de TA baseada em *corpus*. Enquanto alguns autores veem os sistemas baseados em exemplos como um paradigma em si mesmo, outros consideram os sistemas estatísticos como um tipo de sistema de exemplos, já que os primeiros sistemas estatísticos surgiram no final da década de 1980.

4.2.5 Tradução Automática Estatística

Os sistemas de TA estatísticos (em inglês, *Statistical Machine Translation* ou SMT) foram apresentados pela primeira vez por Brown et al. (1988); no entanto, a ideia de usar métodos estatísticos para traduções automáticas foi introduzida pela primeira vez por Weaver em 1949 (Brown et al., 1988, p. 71). Desde a primeira publicação de Brown et al., a equipe da IBM desenvolveu para a empresa o primeiro sistema estatístico funcional e houve um aumento drástico na pesquisa em TA estatística na área.

A ideia geral dos sistemas estatísticos é usar modelos estatísticos para extrair pares de tradução de *corpora* bilíngues. Podem ser encontradas três abordagens principais para a TA estatística:

[11] Para uma descrição mais abrangente e para a história dos sistemas de TA baseados em exemplos, consulte (Carl; Way, 2003).

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

- TA estatística baseada em palavras (*Word-based Statistical Machine Translation*): alinha[12] palavras individuais no texto-fonte a palavras no texto-alvo e calcula a probabilidade da tradução. Também permite a exclusão e inserção de palavras.
- TA estatística baseada em frases (em inglês, *Phrase-based Statistical Machine Translation* ou PBSMT): alinha frases (não frases linguísticas, mas fragmentos de frases e palavras) no texto-fonte a frases no texto-alvo, comparando frases e seus vizinhos frasais ao considerar uma tradução. Essas frases também são chamadas de n-gramas, que são sequências contínuas de n palavras em sequência, ou seja, um unigrama é uma palavra, um bigrama são duas palavras, um trigrama são três palavras etc. A TA estatística baseada em frases é o tipo de abordagem estatística mais utilizado.
- TA estatística baseada em sintaxe: esses modelos traduzem unidades sintáticas usando árvores sintáticas geradas por analisadores sintáticos (Capítulo Sequência de caracteres e palavras).

Independente da estratégia escolhida, na **tradução estatística** a **probabilidade** determina como um texto-fonte deve ser traduzido para um texto-alvo. De acordo com a estratégia escolhida, essa probabilidade pode ser calculada considerando apenas palavras ou também sequências de palavras (frases). Essas frases são sequências de *tokens* (não necessariamente palavras) como "a casa do" ou "linda ." (onde o ponto final faz parte da frase). Seja considerando apenas palavras ou frases, a tradução é realizada com base em dois modelos computacionais: (1) um **modelo de tradução** que especifica como mapear texto-fonte em texto-alvo e (2) um **modelo de língua** que especifica como gerar um texto-alvo fluente. Desse modo, o modelo de tradução tenta maximizar a acurácia da tradução, enquanto o modelo de língua tenta maximizar a fluência da sentença gerada na língua-alvo (Seção 4.3).

Para tentar tornar esses conceitos menos abstratos, vamos retomar nosso exemplo da sentença "A casa do meu avô é linda.". Nesse caso, o modelo de tradução poderia ser baseado em probabilidades de tradução de frases como os gerados para o *corpus* FAPESP (Aziz; Specia, 2011) com o auxílio do Moses[13], como ilustrado na Tabela 4.1.

Tabela 4.1: Exemplos de frases fonte e alvo, e suas probabilidades, presentes em um modelo de tradução gerado a partir do *Corpus* FAPESP (Aziz; Specia, 2011).

id	frase-fonte (português)	frase-alvo (inglês)	probabilidade
1	a casa do	' house	0.0207779
2	a casa do	the house from	0.0623338
3	a casa	the house	0.297619
4	a casa	the home	0.0646474
5	do meu	of my	0.0813954
6	do meu	that of my	0.191576
7	meu avô	my grandfather	0.662453
8	meu avô , meu pai , eu	my grandfather , my father , me	0.0623338
9	avô	grandfather	0.916667
10	é	is	0.611613
11	é	é	0.794943

[12] No contexto da TA estatística, o alinhamento é uma tarefa de encontrar as correspondências entre texto-fonte e texto-alvo. Esse alinhamento pode se dar em nível de palavras (alinhamento lexical), de sentenças (alinhamento sentencial) entre outros.

[13] https://www.statmt.org/moses/ e https://github.com/moses-smt/mosesdecoder

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

id	frase-fonte (português)	frase-alvo (inglês)	probabilidade
12	linda	beautiful	0.0389678
13	linda	pretty	0.00259724

Com base nas probabilidades da Tabela 4.1, diversas opções de tradução poderiam ser geradas como as apresentadas no Exemplo 4.3.

Exemplo 4.3:

(a) *The house from my grandfather is beautiful.*[14]
(b) *The house that of my grandfather is pretty.*[15]
(c) *The home of my grandfather é beautiful.*[16]

E qual dessas sentenças o sistema escolheria como saída? Isso depende de alguns fatores que não vamos detalhar aqui, mas podemos dizer que o modelo de linguagem (Capítulo Modelos de linguagem) tem um papel fundamental na seleção da melhor sentença candidata. Nesse caso, o modelo de linguagem diz qual é a melhor sentença com base na probabilidade de ela ser encontrada na língua-alvo, ou melhor, no *corpus* de treinamento usado para gerar o modelo de língua-alvo.

Assim, tanto o modelo de tradução quanto o modelo de linguagem são treinados a partir de *corpus*. No caso da tradução estatística, as probabilidades são determinadas contando-se as frequências de ocorrência das palavras em grandes quantidades de textos (os *corpora*) paralelos. Um exemplo de *corpus* paralelo português-inglês é o coletado por Aziz; Specia (2011), o Corpus FAPESP[17], que serviu de base para a geração das probabilidades apresentadas na Tabela 4.1. Para a geração do modelo de tradução, a probabilidade de uma frase em português ser traduzida para uma frase em inglês é calculada com base na co-ocorrência dessas frases no *corpus* paralelo. Para a geração do modelo de língua, a probabilidade de uma frase em inglês é calculada com base na ocorrência dessa frase na parte em inglês do *corpus* paralelo (ou de outro *corpus* monolíngue na língua-alvo).

Geralmente os modelos de tradução e de língua consideram frases de um tamanho máximo definido como parâmetro do treinamento. Quanto maior o tamanho da frase, maior é o contexto sendo considerado e, como consequência, mais coerente poderá ser a sentença gerada (Jurafsky; Martin, 2023). Porém, quanto maior o tamanho da frase, maiores serão o tempo e a quantidade de processamento necessários para realizar o treinamento dos modelos.

As **vantagens** dos modelos estatísticos são que, com **mais dados** utilizados no treinamento dos sistemas, não apenas a qualidade geral aumenta, mas, com o uso de um modelo de linguagem, as traduções estatísticas ganharam **fluência** em relação às abordagens anteriores. Além disso, a TA estatística permitiu um **uso mais eficiente de recursos humanos e de dados**. As **desvantagens** dessa abordagem são o **custo de criação de *corpora* paralelos**, especialmente para idiomas com recursos limitados. Além disso, a TA estatística tende a ter **dificuldade com pares de idiomas com ordem de palavras diferentes**.

Em 2007, o sistema *open-source* PBSMT mais famoso, desenvolvido por Koehn et al. (2007), foi lançado: o Moses SMT *toolkit*. Ao mesmo tempo, o Google lançou seu famoso

[14] Obtida por meio da combinação das frases 2, 7, 10 e 12.
[15] Obtida por meio da combinação das frases 3, 6, 9, 10 e 13.
[16] Obtida por meio da combinação das frases 4, 5, 9, 11 e 12.
[17] http://www.nilc.icmc.usp.br/nilc/tools/Fapesp%20Corpora.htm

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Google Tradutor com abordagens estatísticas. Vale ressaltar que os modelos estatísticos conseguiram obter grande sucesso devido ao "aumento do poder de computação e armazenamento de dados, juntamente com a disponibilidade cada vez maior de recursos de texto digital como consequência do crescimento da Internet" (Koehn, 2009, p. 18). Devido à eficiência e precisão da abordagem estatística em relação às anteriores, ela se tornou a abordagem mais amplamente utilizada naquela época. Sistemas de tradução estatística baseada em frases (PBSMT) como os de Koehn et al. (2003) e Och; Ney (2004) eram o estado da arte até serem sucedidos pela tradução neural, a partir de 2015. De fato, a estratégia por trás do tradutor automático do Google foi a PBSMT por uma década (aproximadamente de 2006/2007 até 2016/2017)[18]. Atualmente, o Google e praticamente todos os sistemas de tradução online, bem como pesquisas nesta área usam a tradução neural (*neural machine translation*, NMT) ou algum sistema híbrido (estatístico e neural).

4.2.6 Tradução Automática Neural

Os sistemas de TA neural (em inglês, *Neural Machine Translation* ou NMT) foram introduzidos pela primeira vez na década de 1990 com alguns artigos sugerindo como redes neurais poderiam ser usadas para TA[19] (Way; Forcada, 2018). No entanto, a quantidade dos dados usados para treinar esses modelos não era suficiente para produzir resultados razoáveis e, além disso, "a complexidade computacional envolvida excedia em muito os recursos computacionais daquela época, e, portanto, a ideia foi abandonada por quase duas décadas" (Koehn, 2020, p. 39).

Em geral, os modelos neurais consistem na construção de redes neurais *end-to-end* que mapeiam textos paralelos alinhados e são treinados para maximizar a probabilidade de uma sequência alvo Y, dada uma frase de origem X, sem informações linguísticas externas adicionais (Castilho et al., 2017b). Os sistemas neurais podem ser construídos com apenas uma rede em vez de uma sequência de tarefas separadas, como seu predecessor (a tradução estatística).

Com a publicação de resultados impressionantes em avaliação automática (Bahdanau et al., 2015; Bojar et al., 2016; Sennrich et al., 2016), os sistemas neurais geraram grande expectativa, especialmente porque a indústria de tradução busca melhorar a qualidade da TA para minimizar custos (Moorkens, 2017). A adoção dos sistemas neurais nos últimos anos tem sido extensiva, com um número crescente de provedores de TA e grupos de pesquisa concentrando seus esforços e recursos no desenvolvimento e implantação de sistemas NMT (Castilho et al., 2019a).

Na tradução neural, redes neurais artificiais são usadas para fazer a tradução de uma sentença-fonte para uma sentença-alvo. Uma rede neural artificial pode ser entendida como uma composição de diversas unidades de processamento (os neurônios artificiais) conectadas entre si, em camadas. Cada unidade de processamento recebe uma entrada numérica e gera uma saída numérica. A saída é calculada de acordo com os "pesos" (w) e as "entradas" (x) associados à unidade e uma função que determina como a saída deve ser calculada. Por exemplo, vamos supor que um neurônio artificial seja governado pela função x^2. Nesse caso, se a entrada para esse neurônio for o número 2 então a saída será 4, se for 3 a saída será 9, se for -1 a saída será 1 e assim por diante. Os pesos são usados para

[18]Segundo informações disponíveis em: https://ai.googleblog.com/2016/09/a-neural-network-for-machine.html.

[19]Artigos como (Chalmers, 1992), (Chrisman, 1991), (Castano; Casacuberta, 1997), (Forcada; Ñeco, 1997) e (Neco; Forcada, 1997).

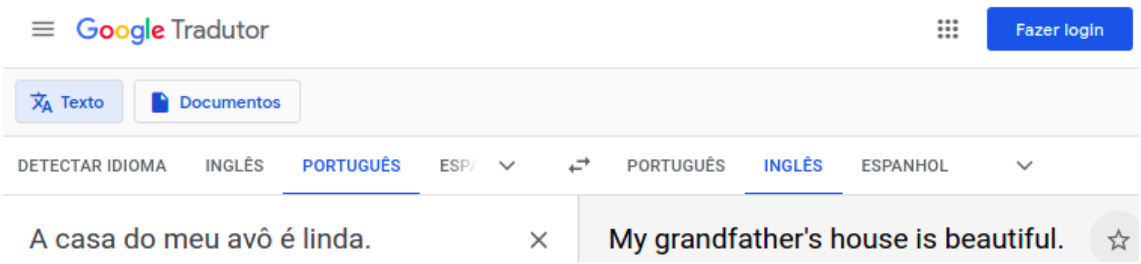
https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

ajustar o aprendizado do neurônio e são uma das partes mais importantes da definição de uma rede neural artificial.

Na **tradução neural**, diversas camadas de neurônios são usadas para aprender como traduzir uma sentença-fonte em uma sentença-alvo a partir de um *corpus* paralelo. Esse aprendizado geralmente **demanda muito poder computacional**[20] **e tempo**, pois é realizado a passos pequenos em diversos ciclos de processamento[21] das sentenças paralelas. E qual é a principal diferença metodológica da tradução neural para a estatística? Na NMT toda a sentença-fonte é considerada no aprendizado, de uma vez, e nos dois sentidos (da esquerda para a direita e da direita para a esquerda), ou seja, não há a quebra em frases como ocorria na PBSMT, nem a divisão clara entre modelo de tradução e modelo de língua. Dessa forma, a tradução gerada por um sistema NMT tende a ser mais fluente e natural, como ilustrado na Figura 4.1[22].

Figura 4.1: Tradução gerada pelo Google tradutor (tradução neural)



Os modelos de tradução neural baseiam-se fortemente em duas tecnologias que se tornaram bastante usuais em PLN: *embeddings* e modelo de atenção. As *embeddings* (Capítulo Representação vetorial e semântica distribucional), são formas de representação de unidades lexicais (geralmente palavras) nas quais as unidades são mapeadas para vetores em um espaço de n (100, 300 ou mais) dimensões. Ao representar palavras como vetores densos notou-se que é possível mapear características linguísticas (morfológicas, sintáticas e semânticas) nesse espaço vetorial. Por exemplo, na Figura 4.2 é possível observar a proximidade semântica da palavra "avô" com outras palavras a partir das *word embeddings* do NILC[23] como "pai", "tio", "sobrinho" etc.

Usando as *word embeddings* bilíngues português e inglês do MUSE[24] é possível observar as similaridades entre línguas como ilustrado na Figura 4.3, onde as palavras em português aparecem em vermelho e as palavras em inglês, em azul. Essas *word embeddings* são usadas como forma de representação da língua nos modelos neurais de tradução automática.

Assim, a tradução neural não se baseia na combinação dos modelos de tradução e de língua, como faz a tradução estatística, mas sim em um modelo sequencial que prediz uma palavra por vez. O potencial deste modelo sequencial está na maneira como ele prediz as palavras: considerando toda a sentença-fonte e também o que já foi produzido para a sentença-alvo. Desde sua proposição, a modelagem sequencial neural passou por várias arquiteturas, indo desde as redes neurais recorrentes (em inglês, *recurrent neural network* ou RNN) usadas para codificação (em inglês, *encoder*) e decodificação (em inglês, *decoder*)

[20] Geralmente são necessárias placas GPU (*Graphics Processing Units*), originalmente projetadas para processamento gráfico, capazes de fazer diversos cálculos em segundos.
[21] No contexto das redes neurais artificiais, um ciclo de processamento é chamado de época.
[22] Tradução gerada pelo https://translate.google.com.br/ em 23 de agosto de 2023.
[23] http://www.nilc.icmc.usp.br/embeddings
[24] https://github.com/facebookresearch/MUSE

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

Figura 4.2: Vizinhos mais próximos da palavra "avô" obtidos via consulta às *word embeddings* do NILC geradas usando o GloVe e dimensão 300.

```
[('pai', 0.7426180839538574),
 ('tio', 0.7307775020599365),
 ('sobrinho', 0.6814965009689331),
 ('irmão', 0.6784138679504395),
 ('avó', 0.6334453821182251),
 ('filho', 0.6296581029891968),
 ('paterno', 0.628169059753418),
 ('bisavô', 0.6196369528770447),
 ('sogro', 0.5976346731185913),
 ('amigo', 0.5973160862922668)]
```

até os mecanismos de atenção (em inglês, *attention mechanism*) (Bahdanau et al., 2015) que permitem ao decodificador focar em partes específicas da sentença de entrada em seu processo de geração da saída.

Nesse momento, os Transformers (Vaswani et al., 2017b) são o estado da arte na tradução. A Figura 4.4 ilustra a tradução da sentença de exemplo, em português, para inglês usando Transformers[25]. Nessa ilustração, quanto mais clara (amarelo, verde claro, azul claro etc.) a célula que une a linha da palavra em inglês com a coluna da palavra em português, maior a "força" da relação entre elas. Por exemplo, observa-se uma forte relação entre "*my*" e "meu", "*grandfather*" e "avô", "*home*" e "casa", e "*beautiful*" e "linda".

Contudo, assim como todas as demais abordagens, a tradução neural também tem suas **limitações**. Uma delas é que, diferentemente da PBSMT onde é possível "olhar" para os modelos aprendidos e entender o que foi usado na tradução (como as frases da Tabela 4.1), a tradução neural é considerada uma **caixa-preta** (*black box*): entender o que pode ter sido considerado para gerar a tradução depende de desvendar uma visualização do modelo de atenção (como o da Figura 4.4), já que as previsões dos modelos neurais consistem em milhões de parâmetros. Isso **dificulta a extensão dos modelos** previamente treinados e coloca em dúvida a **robustez do sistema**. Além disso, por ser uma abordagem relativamente nova, a tradução neural ainda enfrenta alguns desafios, como o **desempenho ruim em condições fora do domínio** e **para idiomas com recursos limitados**.

Além disso, é possível observar que as estratégias de tradução baseadas em *corpus* são **fortemente influenciadas pelo *corpus* usado no treinamento**. Por exemplo, os modelos estatísticos só terão a capacidade de traduzir uma palavra se ela tiver ocorrido um número significativo de vezes no *corpus* de treinamento, caso contrário não haverá uma frase correspondente contendo essa palavra e o sistema não conseguirá gerar uma tradução completa para a sentença original. No caso da tradução neural, isso é um pouco amenizado pelo uso de *embeddings* de unidades menores (em inglês, *subword units*) do que as palavras, as quais conseguem aproximar palavras desconhecidas às possíveis correspondências conhecidas[26]. Por exemplo, se o "a" for esquecido no "linda" da Figura 4.1 o Google tradutor consegue gerar a mesma saída, sem problema. O tratamento de palavras e *subwords* é abordado no Capítulo Sequência de caracteres e palavras.

Outro ponto a se observar é que os sistemas neurais **precisam de um *corpus* maior**

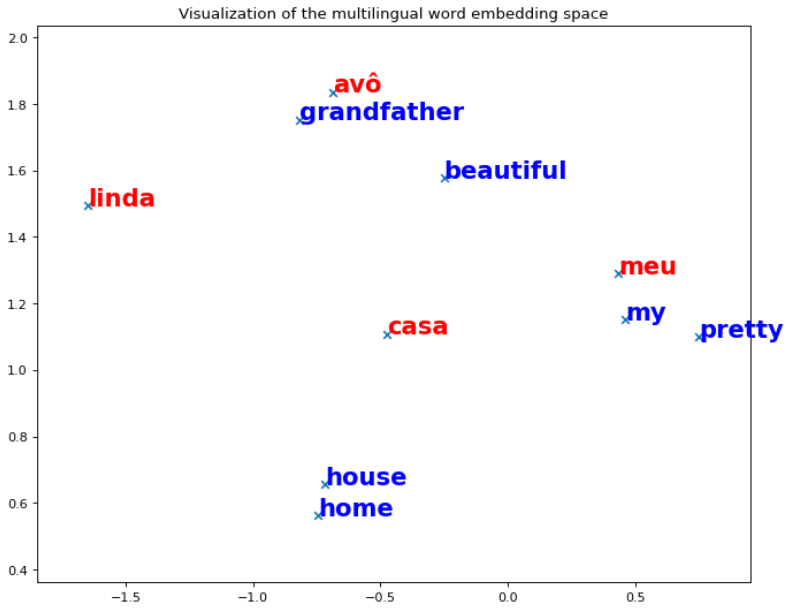
[25] O modelo foi treinado usando o Google Colab e o código disponível em neste link do GitHub do Brasileiras em PLN.

[26] https://ai.googleblog.com/2020/06/recent-advances-in-google-translate.html

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Figura 4.3: Visualização, em duas dimensões, das palavras em português (em vermelho) das palavras que ocorrem na sentença de exemplo e algumas possíveis traduções para o inglês (em azul).



e de melhor qualidade do que os estatísticos, pois eles são rápidos em memorizar exemplos mal-formados (Khayrallah; Koehn, 2018). Por isso, para algumas línguas com menos recursos (em inglês, *low-resourced languages*) os sistemas estatísticos ainda podem apresentar um desempenho melhor do que alguns sistemas neurais. E, por esse motivo, muitas pesquisas atuais têm enfatizado o desenvolvimento de técnicas de aumento de dados (em inglês, *data augmentation*) para sistemas neurais.

Apesar disso, os sistemas neurais são o estado da arte na área de tradução automática (agora em 2023), apresentando, especialmente, uma **fluência muito superior** aos sistemas estatísticos, o que dificulta a avaliação humana da tradução, a qual deve ser mais cuidadosa aos erros de acurácia.

Atualmente, existem várias técnicas de aprendizado profundo para os sistemas neurais, diferentes ramos, orientações de pesquisa e tendências[27].

4.3 Avaliação da Tradução Automática

Devido à importância da tradução no mundo globalizado de hoje, o interesse na avaliação da qualidade da tradução (AQT – do inglês, *Translation Quality Assessment* ou TQA) cresceu a ponto de a avaliação da TA (ATA) se tornar um subcampo em rápido crescimento dentro da TA[28]. No entanto, como a tradução é um processo multifacetado que envolve fatores cognitivos, linguísticos, sociais, culturais e técnicos, definir e medir a qualidade da tradução também reflete essa complexidade (Castilho et al., 2018). A AQT tem sido um tópico muito discutido em estudos de tradução, tecnologia da tradução e na indústria de

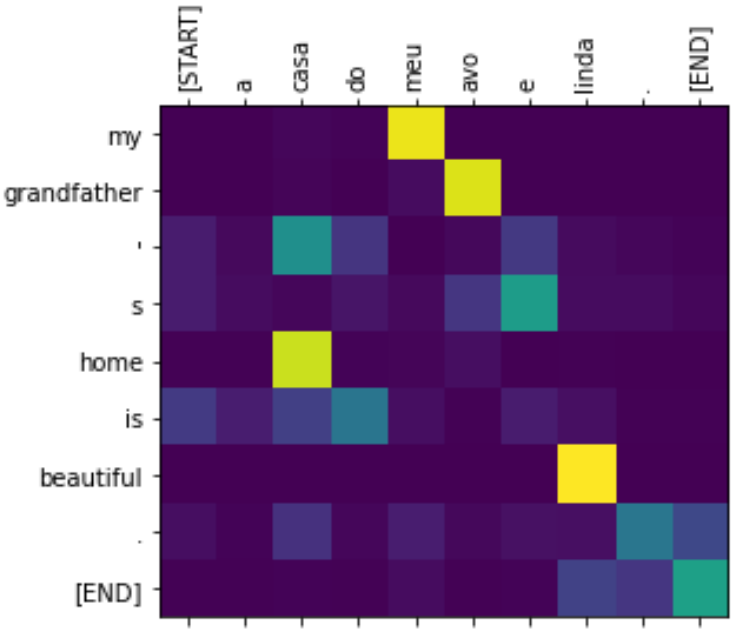
[27] Para uma descrição abrangente do estado da arte da tradução automática neural, veja "Neural Machine Translation" de Koehn (2020).

[28] Veja mais sobre Avaliação de sistemas de PLN no Capítulo Avaliação de tecnologias de linguagem.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Figura 4.4: Visualização de um modelo de atenção usado para traduzir a sentença de exemplo.



tradução e localização, mas ainda não há muito consenso sobre o que é e como ela deve ser feita (Castilho et al., 2018) [29].

Aqui, faremos a distinção entre a AQT e a ATA: enquanto a AQT abrange a avaliação tanto das traduções humanas quanto das traduções automáticas, a ATA se concentra exclusivamente na avaliação da qualidade dos sistemas de TA. Nesta Seção, iremos definir a avaliação da TA, apresentar diferentes abordagens e discutir algumas das avaliações mais influentes na sua história, destacando a importância de realizar a avaliação dos sistemas de TA.

4.3.1 O que é Avaliação da Tradução Automática?

A avaliação da Tradução Automática (ATA) é a prática de analisar a saída de tradução de um sistema (ou sistemas) de TA e julgar a qualidade dessa tradução com base em critérios estabelecidos. As abordagens para a ATA incluem **avaliação automática**, usando métricas automáticas Seção 4.3.3, ou **avaliação manual humana** Seção 4.3.4, realizada por pessoas, e às vezes uma combinação das duas. O fluxograma da Figura 4.5 foi proposto por Doherty et al. (2018) para ajudar educadores e tradutores nos vários tipos de ATA.

Devido à falta de consenso no campo dos estudos de tradução e na indústria da tradução sobre o que constitui uma "boa" tradução, várias abordagens para a avaliação da TA surgiram. Portanto, a ATA pode ser realizada de várias maneiras diferentes, com várias abordagens, e não existe um único método que seja suficiente para abordar todos os propósitos de avaliação (Hovy et al., 2002).

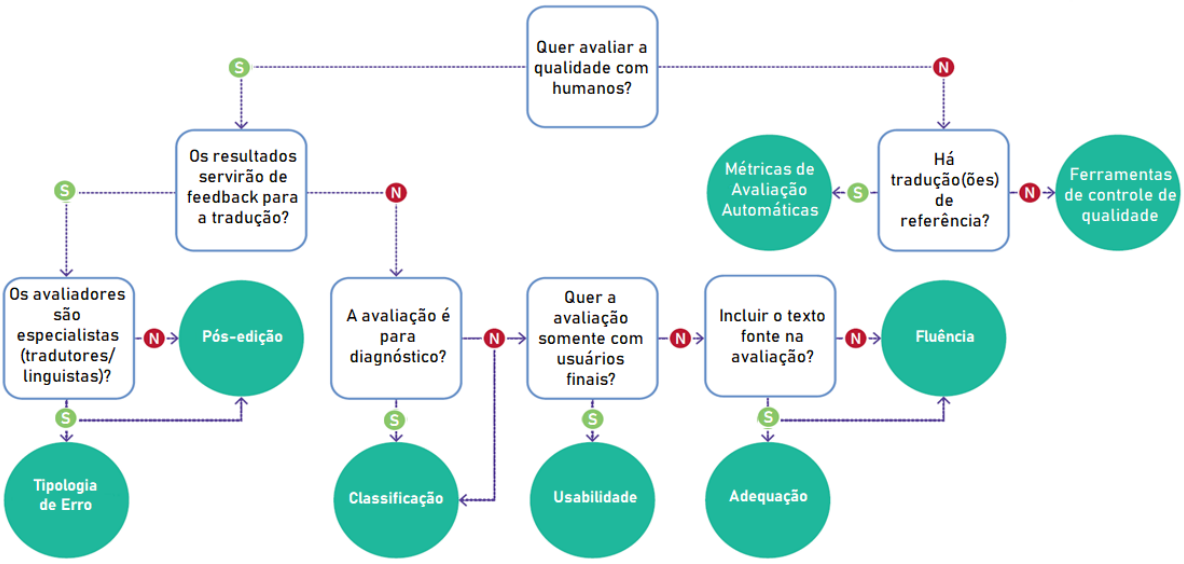
Tradicionalmente, a avaliação da TA foi dividida em dois paradigmas: avaliação glass-box (caixa de vidro) e avaliação black-box (caixa-preta). Enquanto a primeira se preo-

[29] Veja mais sobre a avaliação da tradução automática neste vídeo.

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

Figura 4.5: Fluxograma da ATA baseado em Doherty et al. (2018)



cupada com "a qualidade de um sistema com base em suas propriedades internas" (Dorr et al., 2011, p. 744) e foi amplamente utilizada com sistemas baseados em regras (Seção 4.2.2), a última "mede a qualidade de um sistema apenas com base em sua saída, sem considerar os mecanismos internos do sistema de tradução" (ibid). As abordagens de avaliação black-box são o foco desta Seção.

A avaliação desempenha um papel essencial na TA, pois fornece informações sobre o funcionamento do sistema, quais as partes são eficazes e quais as áreas que precisam de melhorias. No entanto, a avaliação é um problema complexo, pois não existe uma única tradução correta para uma determinada fonte, e "pode haver várias traduções corretas possíveis" (ibid).

Para julgar a qualidade de uma determinada saída de TA, é necessário definir o que a "**qualidade**" significa para essa tarefa de tradução especificamente. Mas quem define o que é qualidade da tradução? A resposta para essa pergunta, definindo o que constitui uma tradução "boa" ou "ruim", tem sido e ainda é amplamente debatida nos diferentes campos relacionados à tradução. Para O'Brien et al. (2011, p. 55), a qualidade está relacionada à opinião do cliente, mas na indústria da tradução, a avaliação de qualidade "é gerenciada por intermediários na cadeia de fornecimento e demanda", geralmente usando uma abordagem única para todos (em inglês, *one size fits all*). Para a ATA, a avaliação de qualidade depende do uso pretendido da tradução. Portanto, qualidade pode significar uma tradução *fluente* que pareça ter sido escrita por um falante nativo; ou uma tradução *precisa* que transmita todo o significado expresso no texto de origem; ou talvez qualidade signifique uma tradução que seja ao mesmo tempo fluente e adequada. Além disso, qualidade pode ser definida como uma tradução fácil de ser pós-editada por alguns, enquanto outros podem definir qualidade como uma tradução que os usuários finais possam usar. E alguns podem ainda querer que todos esses critérios sejam atendidos para se ter uma tradução de qualidade, enquanto outros podem definir qualidade da tradução com outros critérios totalmente diferentes.

Em resumo, ao avaliar a saída da TA, essencialmente estamos tentando determinar graus de "qualidade" e "não qualidade" dessa tradução para um público específico, uma vez que

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

um critério de qualidade que seja crucial em determinada cenário, pode ser irrelevante em outros (Dorr et al., 2011).

Nesse sentido, alguns projetos de tradução que merecem referência são:

- **ALPAC Report (1966)**: examinou a qualidade e eficácia dos sistemas de TA nos Estados Unidos. O relatório concluiu que a TA não era útil devido à baixa qualidade, o que resultou em uma redução significativa no financiamento para o processamento de linguagem natural (PLN) e TA naquela época. O relatório também desencorajou discussões sobre avaliação dentro da comunidade de PLN por muitos anos.
- **DARPA Initiative (1992)**: uma das primeiras iniciativas na avaliação de TA nos anos 1990. A metodologia de avaliação era baseada em julgamentos humanos, onde os avaliadores atribuíram uma pontuação de 1 a 5 para as frases traduzidas automaticamente, comparando-as ao texto original e a traduções de referência humanas produzidas por tradutores profissionais.
- **FEMTI (2002**): foi um esforço para padronizar o processo de avaliação de TA, organizando os métodos existentes de ATA e relacionando-os com o propósito e contexto dos sistemas, fornecendo "diretrizes para a seleção de características de qualidade a serem avaliadas, dependendo da tarefa esperada, dos usuários e das características de entrada de um sistema de tradução automática" (Estrella et al., 2009).
- **GALE (2005)**: teve como objetivo desenvolver e aplicar tecnologias de software para TA de grandes volumes de fala e texto em várias línguas. Foi estabelecido um protocolo de avaliação que utilizava métodos automáticos, humanos, baseados em tarefas e semi-automáticos, com foco na taxa de erro de tradução mediada por humanos (HTER).
- **EuroMatrix (2006)**: financiado pela Comissão Europeia, esta iniciativa construiu sistemas de TA estatísticos e híbridos para diversos idiomas europeus. A avaliação dos sistemas de tradução automática incluiu avaliações automáticas com base em onze métricas, incluindo BLEU (ver Seção 14.2.2), METEOR, Translate Error Rate, Word Error Rate, entre outros (Seção 4.3.3); e avaliações humanas que consistiram em classificações de fluência e adequação, *ranking*, tempo de leitura, tempo de pós-edição, teste de preenchimento de lacunas, clareza e informatividade ("Euromatrix. Survey of Machine Translation Evaluation", 2007).
- **WMT (2006):** começou como o Workshop de Tradução Automática Estatística e progrediu para a Conferência de Tradução Automática em 2016, tem sido uma grande contribuinte para o avanço da pesquisa em avaliação de TA. Ao longo dos anos, a WMT tem utilizado diferentes metodologias de avaliação. Desde 2008, a WMT realiza uma tarefa compartilhada, ou esforço colaborativo, (em inglês, *shared task*) de avaliação que compara uma ou mais traduções de referência com as saídas dos sistemas de TA e tem utilizado, principalmente, a métrica BLEU, embora essa tenha sido superada por diversas métricas diferentes. Em relação à avaliação manual, a WMT experimentou e implementou uma ampla gama de metodologias, incluindo julgamentos de fluência e adequação (2006 e 2007), diferentes avaliações de classificação (2007-2016), pós-edição e compreensão de frases (2009-2010), e avaliação direta e conjuntos de testes (desde 2016). Além disso, a *shared task* da WMT vem realizando avaliações em nível de documento desde 2019.
- **TAUS (2005):** fundada em 2005 como uma rede de dados linguísticos com um extenso repositório de dados e uma rede de engenharia de linguagem humana, a

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

TAUS tem estado na vanguarda das tentativas de estabelecer indicadores para a avaliação da qualidade de tradução, desenvolvendo o Framework de Avaliação Dinâmica de Qualidade (DQF) em 2011. O modelo é uma estrutura de avaliação baseada em diferentes tipos de conteúdo, propósito do uso, ferramentas, processos e outros aspectos.

- **QTLaunchPad (2012):** projeto colaborativo financiado pela União Europeia (2012-2014) que reuniu e forneceu dados e ferramentas para a avaliação da qualidade de tradução, além de métricas de qualidade compartilhadas para tradução humana e automática (Uszkoreit; Lommel, 2013). O projeto contribuiu para o projeto de TA QT21 (2015-2018), cujo objetivo era desenvolver uma avaliação aprimorada. O QTLaunchPad desenvolveu o *framework* de Métricas de Qualidade Multidimensionais (MQM), que descreve e define métricas personalizadas de qualidade de tradução para avaliar a qualidade de textos traduzidos. MQM é amplamente utilizado como um padrão para métricas de avaliação humana em MTEval. Como parte do QT21, o MQM e o DQF (da TAUS) foram harmonizados, e atualmente, a Tipologia de Erros DQF da TAUS é um subconjunto reconhecido do MQM.

4.3.2 A importância de uma avaliação replicável

Como visto neste Capítulo, a avaliação da TA tem sido realizada desde o surgimento da própria TA. No entanto, o relatório ALPAC tornou a avaliação quase um tópico proibido na comunidade de PLN (Paroubek et al., 2007), e até hoje, a avaliação de TA, especialmente a avaliação humana, é considerada "desnecessária" em alguns momentos. Não é raro encontrar artigos de pesquisa usando uma única métrica para avaliar seu sistema (geralmente pontuações BLEU – ver Seção 4.3.3, e ainda mais comum são os artigos de pesquisa com avaliação humana muito limitada e falha, e até mesmo artigos de pesquisa sem avaliação humana alguma (Marie et al., 2021).

O avanço dos sistemas de TA gerou muita expectativa na comunidade, especialmente porque a indústria de tradução buscava uma melhoria na qualidade da TA para reduzir custos (Moorkens, 2017). À medida que os sistemas de TA neural se tornaram líderes no mercado, com uma clara melhoria na qualidade das traduções, surgiram reivindicações exageradas de que esses sistemas estavam "preenchendo as lacunas entre a tradução humana e a tradução automática" (Wu et al., 2016). Diante disso, pesquisadores na área de avaliação de TA alertaram a comunidade para ter cautela e não fazer promessas exageradas, além de enfatizarem a necessidade de mais pesquisas para lidar com as limitações dos sistemas de TA e realizar avaliações humanas mais abrangentes (Castilho et al., 2017b).

No entanto, as afirmações exageradas continuaram, com alguns declarando que seu sistema de TA neural havia atingido a "paridade humana" (em inglês, *human parity*) (Hassan et al., 2018) e outros alegando que a TA é um problema "resolvido" com uma qualidade de tradução "quase perfeita". Como resposta, a fim de verificar essas alegações, dois estudos independentes (Toral et al., 2018) e (Läubli et al., 2018) reavaliaram os dados utilizados por Hassan et al. (2018) e descobriram que a escolha dos avaliadores, o contexto linguístico e a criação de traduções de referência têm um impacto significativo na avaliação de qualidade, apontando para falhas nas práticas atuais em avaliação de TA.

Como resultado, os pesquisadores de avaliação TA passaram a se envolver em discussões mais aprofundadas sobre a necessidade de aprimorar continuamente a metodologia de avaliação, "com o objetivo de superar as limitações das métricas automáticas e das abordagens humanas, evitar superestimativas na capacidade da TA e explicar os resultados

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

aparentemente contraditórios da TA de forma mais abrangente" (Castilho et al., 2019a, p. 2). Além disso, os pesquisadores começaram a buscar diretrizes para avaliar a "paridade humana" na avaliação da TA (Läubli et al., 2020).

Então, quem precisa de avaliação de TA? Em resumo, todos os campos diferentes relacionados à tradução. Algumas questões precisam ser abordadas para que a avaliação seja confiável, tais como:

- O que a **qualidade** significa nesta avaliação? Como vimos, há um grande debate sobre a definição de qualidade, então, antes de avaliar um sistema de TA, o pesquisador precisa definir o que seria uma tradução "boa" e "ruim" nesse cenário.
- **Que tipo de sistema** está sendo avaliado? Dependendo do par de idiomas, os sistemas estatísticos podem mostrar um problema específico. Já os sistemas neurais são conhecidos por serem fluentes e conhecer a similaridade entre palavras, o que torna mais difícil detectar erros.
- Qual é o **objetivo** a ser alcançado com esta avaliação? Saber por que o sistema de TA está sendo avaliado é importante para decidir que tipo de avaliação precisa ser realizada. Por exemplo, se alguém quer saber se as implementações realizadas em um sistema de TA feito para lidar com expressões multipalavras resultaram em traduções corretas dessas expressões, uma análise linguística aprofundada da saída seria mais apropriada do que avaliar a fluência da saída.

Como podemos ver, a avaliação de TA é um problema complexo, e não é surpresa que tenha se estabelecido como um campo independente. Com os avanços na qualidade dos sistemas de TA atuais, a comunidade de avaliação precisa estar atualizada com os procedimentos que possam fornecer avaliações mais sólidas, capazes de confirmar ou refutar as alegações feitas, como afirmou Carl Sagan: "Alegações extraordinárias exigem evidências extraordinárias".

4.3.3 Métricas Automáticas

Como vimos anteriormente, a ATA assume uma complexidade intrínseca devido a uma multiplicidade de fatores. Tipicamente, encontramos duas vertentes de avaliação: a avaliação automática e a avaliação humana (manual), ocasionalmente mescladas para compor uma abordagem híbrida. Nesta Seção, abordaremos as métricas automáticas mais predominantes, reconhecendo a sua influência no domínio da TA.

As métricas automáticas pioneiras empregadas na TA tiveram origem em outras áreas da PLN. Por exemplo, temos a *Word Error Rate* (WER), introduzida por Su et al. (1992), que originou-se do campo de reconhecimento da fala (Capítulo Texto ou fala?). Por sua vez, a ROUGE, desenvolvida por Lin (2004), teve sua origem na sumarização automática. Outra métrica muito usada anteriormente foi a F-*measure*, empregada em recuperação de informação (Capítulo 2) e em diversas outras áreas do PLN, também encontrou aplicabilidade nesse contexto.

As Métricas de Avaliação Automática (MAA) são adotadas na TA quando se busca evitar a intervenção humana direta (Figura 4.5). As MAAs atuam como programas computacionais, recebendo as traduções de um sistema de TA e as traduções de referência (TR) como entrada, e produzindo uma pontuação numérica que reflete a similaridade entre as traduções de TA e TR.

A maioria das MAAs é classificada como métricas de referência (em inglês, *reference-based metrics*), exigindo a disponibilidade da TR, isto é, a tradução humana do texto em

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

avaliação, a fim de serem empregadas como ponto de comparação. No entanto, abordagens mais recentes incorporam modelos de linguagem pré-treinados para medir a semelhança entre uma tradução gerada e um conjunto de referências. Nesse contexto, essas métricas medem a similaridade entre as representações semânticas das palavras e as frases presentes tanto nas traduções geradas quanto nas referências, utilizando recursos linguísticos capturados durante o pré-treinamento desses modelos, em detrimento de uma comparação direta com traduções humanas específicas.

As MAAs preponderantes na área da TA são aquelas que operam sobre características lexicais e dispensam a necessidade de treinamento (em inglês, *untrained*). Essas métricas, geralmente baseadas em similaridades (em inglês, *matching*) e diferença (ou distância) de edições (em inglês, *edit distance*) entre o resultado da TA e a TR, avaliam a **sobreposição** entre a hipótese (resultado da TA) e a TR. Tal avaliação contempla tanto a precisão quanto a abrangência dos elementos lexicais (Lee et al., 2023). Duas vertentes são identificadas entre as métricas lexicais: as *word-based* (baseadas em palavras), que analisam as similaridades entre palavras; e as *character-based* (baseadas em caracteres), que investigam a similaridade entre caracteres.

As métricas lexicais *word-based* mais amplamente empregadas permitem medir tanto a similaridade dos n-gramas quanto a distância de edição (*edit distance*). Dentre as métricas baseadas em n-grama, destacam-se as amplamente conhecidas BLEU (Papineni et al., 2002), METEOR (Banerjee; Lavie, 2005) e NIST (Doddington, 2002). Por outro lado, as métricas que calculam a distância de edição e que têm destaque são TER (e HTER) (Snover et al., 2006) e WER (Su et al., 1992). Vale mencionar a singularidade da métrica chrF (Popović, 2015), a qual, além de ser *character-based*, também mede a similaridade dos n-gramas.

Mais recentemente, métricas treinadas em modelos baseados em redes neurais usando a arquitetura Transformer foram propostas. Dentre essas, há as métricas supervisionadas (*supervised-metrics*) e as não-supervisionadas, dependendo da técnica de aprendizado, ambas categorias com *word-embeddings* e *contextual-embeddings* (Lee et al., 2023). Entre as métricas não-supervisionadas mais recorrentes, destacam-se MEANT (*word embedding*) (Lo; Wu, 2011), BERTscore (Zhang et al., 2020), Yisi (Lo, 2019) e BARTscore (Yuan et al., 2021a) (*contextual-embedding*). Entre as supervisionadas, estão a BEER (Stanojevic; Sima'an, 2014) e BLEND (Ma et al., 2017) (ambas *word-embeddings*), BERT for MTE (Shimanaka et al., 2019), BLEURT (Sellam et al., 2020) e COMET (Rei et al., 2020).

As **vantagens** das MAAs são que elas são **eficientes**, **econômicas** e fornecem avaliações **consistentes**, ou seja, se a métrica for computada para a mesma tradução várias vezes, todas elas vão dar o mesmo resultado. No entanto, uma **preocupação** é a **dependência** exclusiva das similaridades entre a saída do sistema e as referências. Primeiramente, não há somente uma única tradução correta para um texto, sendo assim, o significado do texto pode ser traduzido de várias maneiras. Mas seriam todas as tradução "igualmente boas"? E o que "boa" significa nesse determinado contexto da tradução? Nesse caso, usar múltiplas TRs seria essencial para se ter uma avaliação mais justa. Segundo, as MAAs não oferecem insights detalhados sobre erros de tradução, pontos fortes e fragilidades de um sistema. Elas não dizem o que funciona no sistema, o que precisa ser melhorado; sendo o único objetivo medir a semelhança com a(s) referência(s), e consequentemente, as melhorias específicas decorrentes de modificações no sistema de tradução permanecem obscuras. E finalmente, o sistema com uma pontuação menor pode ser melhor na prática do que um sistema com uma pontuação mais alta. Enquanto as MAAs servem como ferramentas quantitativas valiosas, elas não revelam completamente as complexidades da qualidade da

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

tradução. Uma abordagem mais abrangente, combinando MAAs com avaliações humanas e análises qualitativas, oferece uma compreensão mais profunda do desempenho dos sistemas de TA.

Embora as MAAs não se revelem apropriadas para mensurar a qualidade final dos sistemas, impulsionam o avanço da pesquisa em TA, uma vez que podem ser empregadas de forma constante durante o desenvolvimento e a implementação desses sistemas. Em essência, as MAAs são medidas úteis na comparação entre sistemas de TA ou de versões de um mesmo sistema de TA, mas são limitadas na predição da qualidade da tradução.

4.3.4 Métricas Humanas

O processo de avaliar a qualidade da TA por meio da intervenção humana é essencial. Embora as MAAs proporcionem uma avaliação quantitativa, a avaliação humana oferece uma visão mais detalhada e uma análise mais ampla de fenômenos linguísticos complexos subjacentes ao desempenho dos sistemas de tradução, sendo assim imprescindível em uma compreensão mais abrangente dos sistemas de TA.

A avaliação humana pode ser feita através de diversos paradigmas, sendo os mais comuns o paradigma de **fluência-adequação** e **pós-edição**. A abordagem de **ranqueamento** de segmentos (em inglês, *ranking*) também é comumente empregada para a comparação dos sistemas de tradução, e possibilita a avaliação comparativa de diversos sistemas, fornecendo insights sobre a eficácia relativa de suas saídas. Igualmente, a **anotação de erros**, sob a forma de marcações específicas, oferece um feedback valioso sobre os sistemas em análise.

Outras abordagens incluem métricas secundárias, tais como legibilidade, compreensibilidade e usabilidade das traduções resultantes. Vale mencionar métricas centradas no usuário, que são avaliadas com testes de compreensão, por exemplo, que podem ser utilizados para aferir não apenas a fidelidade à tradução, mas também a transmissão efetiva da mensagem subjacente. Essa abordagem complementar permite uma apreciação mais holística da eficácia das traduções geradas.

Importante ressaltar que a avaliação humana é realizada por uma variedade de avaliadores, incluindo tanto profissionais quanto amadores. Essa diversidade de perspectivas pode contribuir para uma avaliação mais robusta e representativa da real eficácia dos sistemas de TA se empregados na mesma avaliação. Porém, também pode resultar em conclusões diferentes se a avaliação é feita só com um ou com o outro (só amadores ou só profissionais, por exemplo). Essa Seção vai abordar as métricas humanas mais comuns na avaliação de TA, assim como a importância de se calcular a concordância entre anotadores.

4.3.4.1 Adequação: Fidelidade Semântica

A **adequação** (em inglês, "*accuracy*" ou "*fidelity*"), ou também "acurácia" ou "exatidão", é uma métrica importante na avaliação humana da TA. Ela foca na **fidelidade semântica** entre o texto-fonte e a tradução, mostrando a profundidade e a precisão da transferência de significado, permitindo determinar se a mensagem que está sendo transmitida e seu sentido foram preservados de maneira precisa e fiel.

Essencialmente, a adequação investiga até que ponto a tradução transmite o significado do texto de origem para o texto-alvo. Nesse contexto, uma escala Likert[30] é utilizada

[30] A escala Likert é um método de medição criado por Likert (1932) que apresenta ao respondente uma afirmação ou pergunta e solicita que o respondente avalie o grau em que concorda com ela. A escala envolve uma série de itens ou afirmações aos quais os respondentes atribuem níveis de concordância ou discordância.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

para classificar o nível de transferência semântica, onde geralmente uma pergunta é feita para o avaliador: "Até que ponto a tradução transfere o significado do texto-fonte para o texto-alvo?". Como resposta, o avaliador pode escolher uma opção em uma escala que varia desde "Nada" até "Tudo", com os graus intermediários de "Pouco" e "Muito".

As **limitações** da adequação é que ela **não fornece informações sobre a fluência** da tradução, deixando uma lacuna importante na avaliação. Em determinados casos, o foco reside exclusivamente no significado da sentença de origem, tornando a fluência uma preocupação secundária. Além disso, a adequação não oferece detalhes precisos sobre os erros presentes na tradução, o que pode dificultar a identificação de pontos específicos para a melhoria do sistema.

4.3.4.2 Fluência: Naturalidade e Estrutura

A **fluência**, ou "inteligibilidade" (em inglês, "*fluency*" ou "*intelligibility*"), é outra métrica importante na avaliação humana da TA, que se preocupa com a **naturalidade e a estrutura** do texto-alvo, revelando o grau de fluência e adaptabilidade da saida da TA às normas linguísticas e socioculturais da língua-alvo.

Essa avaliação foca diretamente no **texto-alvo**, priorizando a avaliação da gramática e dos aspectos estruturais da tradução. Essencialmente, ela investiga o quão natural e fluido é o fluxo do texto-alvo dentro do contexto da língua-alvo, considerando suas normas linguísticas e socioculturais específicas. Uma característica distintiva da dimensão de fluência é que ela **pode ser avaliada independentemente do texto-fonte**, uma vez que se concentra exclusivamente no resultado final da tradução.

Também medida numa escala Likert, uma pergunta típica dirigida ao avaliador poderia ser: "Quão fluente está o texto-alvo, ou seja, como está o fluxo e a naturalidade do texto-alvo no contexto da língua-alvo e suas normas linguísticas e socioculturais em um dado contexto?". A escala Likert pode variar desde "Sem fluência" até "Nativo", incluindo os graus intermediários de "Pouca fluência" e "Quase nativo".

Essa métrica oferece informação sobre a naturalidade da tradução e se ela soa natural e fluída para um falante nativo ou se exibe características de "linguagem quebrada", prejudicando assim a experiência de leitura e compreensão.

Assim como a adequação, a avaliação da fluência também apresenta **limitações**. Ela **não proporciona informações sobre a adequação** da tradução, pois o foco está exclusivamente na fluência da sentença de destino, tornando a adequação uma preocupação secundária. Além disso, a fluência também **não oferece detalhes precisos sobre os erros presentes na tradução**.

Por esse motivo, é comum que as avaliações de adequação e fluência caminhem juntas, uma vez que é mais intuitivo avaliar uma em relação à outra. No entanto, há momentos em que pode ser necessário priorizar uma em detrimento da outra. Documentações técnicas, por exemplo, podem demandar uma maior ênfase na adequação, priorizando a transmissão precisa do significado.

Algumas **considerações sobre o uso de escalas Likert** na avaliação humana são importantes. Vale a pena ressaltar que as escalas Likert podem apresentar complexidades na sua aplicação. Diversos tipos podem ser utilizados, como escalas numéricas, de janela deslizante (em inglês, *sliding window*) e afirmações (concordo/discordo). Embora essas escalas sejam facilmente compreensíveis e quantificáveis, elas carregam uma natureza subjetiva, pois falham em medir as atitudes reais dos respondentes, levantando a questão: qual a diferença exata entre uma pontuação 5 e uma 4? (Um erro? Dois erros?). Além

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

disso, a presença de um número par de opções pode levar os participantes a escolherem o centro, demonstrando a delicadeza desse tipo de avaliação.

4.3.4.3 Ranqueamento: Hierarquia de Traduções

O **ranqueamento** (em inglês, *ranking*) tem como propósito classificar e comparar duas ou mais traduções, com o intuito de estabelecer uma **hierarquia de qualidade** entre elas. As comparações podem ser efetuadas tanto entre as traduções geradas por diferentes sistemas de TA quanto entre traduções geradas por humanos. Essa abordagem permite identificar **nuances de qualidade**, ressaltando as distinções entre as opções.

Além disso, o ranqueamento pode incorporar a **possibilidade de empates**, quando duas ou mais traduções são avaliadas como equivalentes em qualidade, sendo categorizadas como "igualmente boas" ou "igualmente ruins". A categorização desses empates enriquece a análise, oferecendo insights sobre o grau de qualidade comparativa.

No âmbito de diagnósticos, esse tipo de classificação oferece a capacidade de indicar melhorias ao comparar o sistema avaliado com uma linha de base (*baseline*). Essa perspectiva não somente permite avaliar o progresso alcançado mas também identificar áreas específicas de aprimoramento, promovendo a constante evolução do sistema de TA.

Uma aplicação prática do ranqueamento é a seleção do sistema mais adequado para um projeto específico. A análise hierárquica das traduções permite a **escolha embasada no desempenho**, assegurando que a tradução atenda de forma eficiente aos requisitos e objetivos do projeto em questão.

O uso do ranqueamento apresenta suas **limitações** a serem consideradas. Ele **não oferece uma avaliação refinada**, não detalhando os erros presentes nas traduções. Quando empates não são permitidos, traduções igualmente boas ou ruins podem ser classificadas de maneira diferente, ressaltando uma inconsistência na hierarquia.

4.3.4.4 Anotação de Erros: Taxonomias

A anotação de erros se destaca como um método essencial para identificar e classificar imperfeições presentes em textos traduzidos. Diversas taxonomias foram propostas para essa finalidade, como Vilar et al. (2006), Font Llitjós et al. (2005), Federico et al. (2014), Costa et al. (2015), DQF de TAUS (O'Brien et al., 2011) e MQM (Lommel; Melby, 2018) by QT212 (Doherty et al., 2013). Para o português brasileiro, Martins (2014) e Martins; Caseli (2015) trazem a adaptação das categorias de erros de Popovic; Burchardt (2011) e Vilar et al. (2006) para traduções português-inglês.

As tipologias de erros frequentemente abrangem uma série de aspectos, incluindo palavras ausentes (em inglês, *missing words*), palavras adicionadas (em inglês, *added/extra words*), ordem errada das palavras (em inglês, *word order*), traduções literais (em inglês, *literal translation*), traduções erradas (em inglês, *mistranslation*), palavras incorretas (em inglês, *incorrect words*), formas inadequadas (em inglês, *incorrect form*), pontuação inadequada (em inglês, *punctuation*), entre outros, que podem incluir outras subcategorias específicas.

A adoção de taxonomias de erros na ATA oferece diversos **benefícios**, tais como identificar **tipos específicos** de erros nas saídas de TA, fornecer **relatórios detalhados de erros** para o aprimoramento dos sistemas, e fornecer informações aos clientes sobre a **qualidade da tradução**. Além disso, provedores de serviços linguísticos utilizam taxo-

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

nomias e avaliações de severidade[31] para monitorar o trabalho de tradutores. A anotação de erro também ajuda a investigar as **relações entre tipos específicos de erros e as preferências de usuários ou pós-editores**, bem como avaliar o **impacto** de diferentes tipos de erros em várias etapas do processo de pós-edição.

Contudo, entre as principais **limitações** dessa estratégia, destaca-se que a anotação manual de erros é um **processo caro e demorado**, demandando um investimento significativo de tempo. Além disso, essa avaliação **nem sempre é uma tarefa simples**, especialmente quando se trata de diferenciar entre categorias como "tradução literal" e "tradução errada".

Outra complexidade está associada à **dependência da língua**. Diferentes idiomas possuem particularidades que podem tornar a identificação e a classificação de erros uma tarefa mais desafiadora. Também é relevante considerar que a eficácia da anotação de erros pode variar de acordo com o **tipo de abordagem de TA**: enquanto ela pode ser mais adequada para sistemas de tradução baseados em regras (RBMT), pode não ser tão precisa para sistemas de tradução estatística (SMT) ou de tradução neural (NMT). Nesse contexto, a seleção da abordagem de avaliação mais adequada torna-se um ponto de reflexão. Além disso, a **falta de consenso entre avaliadores** é uma questão importante, frequentemente requerendo treinamento e prática para alcançar um nível satisfatório de concordância (Capítulo Conjunto de dados, *dataset* e *corpus*).

4.3.4.5 Pós-Edição na Avaliação de TA

A pós-edição (PE) (do inglês, *post-editing*) é definida como "a correção da saída da tradução automática bruta por um tradutor humano, de acordo com instruções e critérios de qualidade específicos" (O'Brien, 2011, p. 197). A PE emerge como uma ferramenta fundamental na avaliação de sistemas de TA, oferecendo uma perspectiva mais aprofundada sobre o esforço envolvido nesse processo. A medição desse esforço pode ser abordada de diferentes perspectivas (Krings, 2001), proporcionando uma visão mais abrangente do desempenho do sistema e do impacto da TA no fluxo de trabalho.

- **Esforço Temporal:** mede o ritmo de pós-edição. Avaliando o tempo gasto na pós-edição por palavras por segundos é possível compreender a velocidade desse processo. Nesse contexto, uma eficiência temporal maior, ou seja, menos tempo gasto na pós-edição, pode indicar uma melhor qualidade da saída da TA, influenciando a produtividade.
- **Esforço Técnico:** mede o número de operações de edição realizadas, como inserções (*insertions*), remoções (*deletions*), e trocas (*shifts*). Nesse sentido, a métrica hTER (Seção 4.3.3) é frequentemente utilizada como uma estimativa aproximada do esforço técnico. Uma menor quantidade de edições necessárias está diretamente correlacionada a uma melhor qualidade da TA, uma vez que está ligada ao tempo de esforço e, consequentemente, à produtividade.
- **Esforço Cognitivo:** pode ser medido por meio de diferentes abordagens, incluindo o rastreamento ocular (*eye-tracking*). A redução desse esforço cognitivo durante o processo de pós-edição é indicativa de uma qualidade superior da TA, e tal esforço tem sido correlacionado a outras métricas de avaliação humana.

[31] As severidades geralmente são classificadas como "Crítico" (*critical*), "Grave" (*major*), e "Mínimo" (*minor*).

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

A utilização da PE na avaliação da TA é motivada por diversos fatores. Além de avaliar a utilidade do sistema de TA em produção, ela **permite identificar erros comuns e gerar novos dados de treinamento ou teste**. Contudo, é importante ressaltar que as medidas de esforço de PE **tendem a variar entre avaliadores** novatos (estudantes) e profissionais, bem como entre o público em geral e profissionais experientes.

Alguns trabalhos com pós-edição com o portguês incluem: De Sousa et al. (2011); Almeida (2013), Castilho et al. (2014), Moorkens et al. (2015), Castilho et al. (2017a), Silva et al. (2017a), Castilho et al. (2019b), Castilho; Resende (2022). Há também os trabalhos que investigaram a automatização do processo de pós-edição para o português: Caseli; Inácio (2020).

4.3.4.6 Considerações Finais sobre Avaliação Humana na Avaliação de TA

Ao explorar a avaliação humana na avaliação de TA, encontramos uma série de questões metodológicas e pragmáticas que merecem reflexão:

- Precisamos sempre avaliar tanto a adequação quanto à fluência?
- Quantos avaliadores são necessários e quais as competências linguísticas eles devem possuir?
- Devemos envolver tradutores, linguistas ou especialistas no domínio? O viés em cada escolha é um ponto a ser considerado.
- Quantos pontos devem ser incluídos em uma escala Likert de avaliação?
- Qual o grau de concordância entre avaliadores (Capítulo Conjunto de dados, *dataset* e *corpus*) e a consistência nas avaliações individuais?

Considerações pragmáticas também emergem, incluindo o custo associado aos avaliadores, à geração de textos de referência traduzidos por humanos, e à qualidade dessas referências. O tempo investido, a baixa concordância intra e inter-avaliadores, e a questão de saber se o objetivo da avaliação é apenas avaliar melhorias em um sistema são fatores preponderantes.

4.3.5 Avaliação dependente de contexto

Como vimos nas seções anteriores, a avaliação de TA começou desde o princípio da área com projetos como DARPA, que avaliaram a qualidade das traduções com métricas humanas. À medida que a área avançou, métricas automáticas foram gradualmente incorporadas, inicialmente provenientes de outros domínios do PLN, e posteriormente desenvolvidas especificamente para o campo da TA.

E apesar de ambas as MAAs e as métricas humanas serem o estado da arte na avaliação, à medida que os sistemas neurais de TA evoluíram, tornando-se mais complexos e produzindo traduções de maior qualidade, surgiu a necessidade de uma **avaliação mais abrangente e rigorosa** que levasse em conta fatores diversos. Adicionalmente, com o aumento nos esforços direcionados à incorporação de contexto nos sistemas de TA neurais, viu-se a necessidade de também se ter uma avaliação com contexto, uma vez que, os resultados obtidos na avaliação com sentenças eram limitados, pois ela não é capaz de identificar as melhorias desses sistemas (Läubli et al., 2018; Toral et al., 2018). Ademais, as MAAs subestimam a qualidade dos sistemas NMT (Shterionov et al., 2018), e a credibilidade dessas métricas para sistemas em nível de documento também tem sido objeto de críticas (Smith, 2017). Diante disso, surgiu a necessidade de avaliar a TA considerando

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

um contexto mais amplo, possibilitando uma análise mais abrangente do contexto em questão. Entretanto, a metodologia para essa avaliação ainda está em sua fase inicial e poucos estudos foram realizados nesse sentido.

Em 2018, alegações de "paridade humana" (em inglês, *human parity*) na qualidade da TA (Hassan et al., 2018) foram rebatidas por Toral et al. (2018) e Läubli et al. (2018), os quais apontaram que essa paridade não se replicava quando se considerava o **contexto** ou outros fatores, como a direção da tradução ou a experiência do anotador.

A Conferência de Tradução Automática (WMT), realizada desde 2006 e que até o ano de 2019 restringiu suas avaliações à análise de frases individuais, começou sua primeira tentativa de conduzir avaliações humanas em nível de documento no domínio de notícias no ano de 2019 (Barrault et al., 2019), em resposta às críticas apresentadas por Toral et al. (2018) e Läubli et al. (2018). Adotando uma abordagem direta de avaliação (Graham et al., 2016), a conferência solicitou que avaliadores de multidão[32] atribuíssem uma pontuação (de 0 a 100) a cada sentença. Os avaliadores foram instruídos a avaliar: (i) textos completos, (ii) segmentos individuais consecutivos na ordem original, e (iii) frases individuais selecionadas aleatoriamente. No ano subsequente, na edição WMT20, houve uma mudança de abordagem, expandindo o âmbito de avaliação para abranger artigos completos, demandando dos avaliadores a análise de segmentos específicos enquanto visualizavam o documento completo, bem como a avaliação da tradução do conteúdo (Barrault et al., 2020).

Castilho (2020) e Castilho (2021) testou as diferenças na concordância inter-anotadores (CIA) entre duas metodologias de avaliação: (i) uma centrada em sentenças individuais e (ii) outra com contexto, para o português brasiliero. No estudo de Castilho (2020), tradutores avaliaram a saída de TA considerando critérios de fluência, adequação (usando uma escala Likert), ranqueamento e anotação de erros. Essa avaliação foi conduzida em duas configurações distintas onde os tradutores atribuíram: (i) uma pontuação para cada sentença isolada, e (ii) uma pontuação para o documento como um todo. Os resultados demonstraram que os níveis de CIA para a metodologia em nível de documento atingiram níveis negativos, enquanto a satisfação dos tradutores com essa metodologia foi bastante reduzida. No entanto, esse enfoque evitou situações de avaliação incorreta (*misevaluation*) que são recorrentes quando se analisam sentenças isoladamente.

Continuando esse trabalho, Castilho (2021) modifica a configuração em nível de documento e repete o experimento com mais tradutores, onde ela compara a CIA na avaliação de (i) sentenças únicas aleatórias, (ii) avaliação de sentenças individuais em que os tradutores têm acesso à fonte completa e à saída de TA, e (iii) avaliação de documentos completos. Os resultados mostraram que uma metodologia em que os tradutores avaliam sentenças individuais no contexto de um documento gera um bom nível de CIA em comparação com a metodologia de sentença única aleatória, enquanto uma metodologia em que os tradutores atribuem uma pontuação por documento mostra um nível muito baixo de CIA. A autora afirma que atribuir uma nota por sentença no contexto evita casos de avaliação incorreta que são extremamente comuns nas configurações de avaliação de frases aleatórias. Além disso, a autora postula que o maior acordo de CIA na configuração de sentença única aleatória ocorre porque "os avaliadores tendem a aceitar a tradução quando a adequação é ambígua, mas a tradução está correta, especialmente se for fluente" (Castilho, 2021, p. 42), e afirma que **o método de avaliação de sentença única aleatória deve ser evitado**, pois o problema de avaliação incorreta é especialmente problemático ao avaliar a qualidade

[32] Avaliadores contratados via plataformas como Mecanical Turk.

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

de sistemas NMT, uma vez que eles apresentam um nível aprimorado de fluência.

Após isso, em Castilho (2022) foi demonstrado que o contexto necessário para resolver questões de avaliação é influenciado pelo domínio, sem parecer estar intrinsecamente ligado ao comprimento das sentenças presentes no *corpus* envolvendo os idiomas inglês, português, irlandês, chinês e alemão. Em consequência disso, a pesquisa de Castilho et al. (2023) revelou que o impacto da extensão do contexto não parece influenciar significativamente os resultados, porém a estruturação da pontuação desempenha um papel crucial. Isso se deve ao fato de que sentenças conectadas tendem a gerar resultados mais diversos, com abordagens mais acuradas para resolver ambiguidades lexicais quando comparadas aos cenários de pontuação normais. Além disso, o estudo apontou que os sistemas GPT demonstraram proporcionar traduções mais precisas do que os sistemas de Tradução Automática.

Diante desse panorama, a avaliação de TA com contexto encontra-se em sua infância, com diversas questões em aberto. O futuro da avaliação de TA deve considerar se as métricas automatizadas e as avaliações humanas atuais conseguem capturar de forma realista a qualidade dos sistemas de nível de documento, e se é necessário modificar ou criar novas abordagens. Especial atenção deve ser dada aos modelos de linguagem como o GPT, conhecidos por gerar traduções fluentes e coesas, uma vez que a avaliação deve incorporar precisão da informação, fidelidade ao conteúdo original e coerência global, evitando a introdução de informações imprecisas ou divergentes.

Ademais, a avaliação de documentos traduzidos não deve se limitar a métricas automáticas, mas também usar a avaliação humana. Os avaliadores humanos desempenham um papel crucial em identificar nuances de qualidade que as métricas automáticas podem não capturar, como aspectos culturais, ambiguidades e sutilezas linguísticas. Portanto, a **combinação de métricas automáticas com avaliações humanas** se mostra uma abordagem fundamental para obter uma compreensão abrangente da qualidade da tradução de documentos.

4.4 O Futuro da Tradução Automática

O futuro da TA parece muito promissor. Com a globalização e a internet, mais conteúdo é criado todos os dias, e, portanto, estão surgindo cada vez mais casos de uso nos quais a TA pode ser útil (Way, 2018). Segundo a "Slator 2019 Language Industry Market Report" (2019, p. 14), "a TA está bem encaminhada para se tornar a tecnologia mais importante para aprimorar a produtividade dos tradutores humanos".

O aumento impressionante na qualidade com o surgimento da TA neural (NMT) em comparação com seu antecessor, o PSMT, foi exagerado pela mídia (Läubli et al., 2018; Toral et al., 2018), mas é incontestável que a NMT tenha sido, de fato, uma mudança de paradigma na área. No entanto, o entusiasmo em torno da TA diminuiu, com empresas de tradução de grande e médio porte relatando que, embora o uso da TA tenha aumentado, os benefícios percebidos têm se estabilizado (Sarah Hickey, 2020) em termos de grandes avanços na qualidade.

No entanto, com o aumento da qualidade, é possível abordar uma variedade maior de tipos de documentos e públicos. Isso significa que há muito espaço para **personalização de sistemas de TA** projetados para casos de uso e contextos específicos, melhorando a precisão. Para a TAUS (2020, p. 16), a NMT será "aplicada de forma útil em ambientes de tradução de fala" e, na verdade, em todo discurso falado, já que lida melhor com conteúdo gerado pelo usuário. Além disso, "a NMT ajudará na expansão adicional de tecnologias de

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

tradução de fala [...] disponíveis principalmente como sistemas monolíngues baseados em inglês, [...] transformando-os em sistemas multilíngues", o que implicará "muitas mudanças profundas e caras".

Mais recentemente, no fim de 2022, os modelos de linguagem em larga escala (Capítulo Modelos de linguagem), como o GPT-3 da OpenAI[33], têm desempenhado um papel importante no campo da tradução automática e prometem desempenhar um papel ainda maior no futuro. Esses modelos surgiram com o avanço das redes neurais e do aprendizado profundo. Desde sua introdução, os LLMs têm sido amplamente utilizados na TA, proporcionando melhorias significativas na qualidade e na fluidez das traduções geradas. Eles têm sido capazes de lidar com nuances linguísticas, contexto e ambiguidades, resultando em traduções mais precisas e naturais. Com o contínuo avanço da tecnologia, espera-se que os LLMs sejam capazes de melhorar a personalização das traduções, adaptando-se a estilos de escrita específicos e preferências individuais.

No entanto, embora os LLMs tenham apresentado avanços significativos na área da TA, ainda existem **desafios** a serem superados. A qualidade da tradução depende de vários fatores, como a disponibilidade de dados de treinamento de alta qualidade e a compreensão do contexto e nuances linguísticas. Além disso, os LLMs podem ser sensíveis a preconceitos (bias) presentes nos dados de treinamento, resultando em traduções imprecisas ou enviesadas.

Segundo a pesquisa da CSA, "a pós-edição como serviço diminuirá ao longo do tempo, sendo substituída pela tradução automática adaptativa em software de tradução mais dinâmico" (p. 23), e haverá uma demanda crescente por linguistas profissionais que possam interagir com a saída da tradução automática "Slator 2019 Language Industry Market Report" (2019, p. 22). Além disso, o relatório da Slator afirma que agora, com os altos níveis de qualidade e a ampla disponibilidade de ferramentas gratuitas de tradução automática, os clientes corporativos esperam mais do que uma tradução automática "apenas boa" e estão buscando "soluções personalizadas, adaptadas ao seu conteúdo, que possam ser adaptadas para seus fluxos de trabalho e preferências estilísticas específicas" "Slator 2019 Language Industry Market Report" (2019, p. 22).

Vale ressaltar que todos os relatórios afirmam que a maioria da indústria "ainda não espera que a qualidade da tradução automática atinja os níveis da tradução humana em um futuro próximo" TAUS (2020, p. 16), e, portanto, tanto a tradução humana quanto a interação humana com a tradução automática ainda são altamente demandadas.

Como podemos ver, os sistemas de TA estão atingindo níveis de qualidade significativamente altos e, por isso, estão sendo cada vez mais utilizados em diversas áreas de negócio. Com a TA se tornando ubíqua em nosso dia a dia, a necessidade de testar a qualidade desses sistemas se tornou essencial (Castilho et al., 2019a). Há muito espaço para a TA melhorar, e, portanto, uma boa prática na avaliação da TA é essencial para evitar afirmações exageradas e fornecer aos usuários um feedback honesto.

[33] https://openai.com/

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Capítulo 5

Sumarização Automática

Jackson Wilke da Cruz Souza
Paula Christina Figueira Cardoso
Crysttian Arantes Paixão
Pedro Henrique Paiola
Gabriel Lino Garcia
João Vitor Mariano Correia
João Renato Ribeiro Manesco
Danilo Samuel Jodas
Douglas Rodrigues
João Paulo Papa

Publicado em: 13/03/2024
Atualizado em: 16/04/2026

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

5.1 Introdução

Talvez resumir seja uma das tarefas mais comuns em nosso cotidiano. Não são raras as vezes que nos pegamos reportando algo a alguém: quando compartilhamos sobre o último filme que assistimos, ou o livro que lemos, ou quem sabe quando contamos para alguém sobre aquela história ou notícia que vimos ou vivenciamos.

Nessas tarefas, partimos do princípio que não conseguiremos reproduzir exatamente o que aconteceu, mas faremos o máximo de esforço para transmitir uma mensagem mais próxima do original, valorizando o **conteúdo** mesmo em detrimento da **forma**. Assim, por exemplo, podemos optar por reestruturar o fato acontecido ou vivenciado dentro da narrativa sintetizada: selecionamos o que julgamos ser o mais importante do evento; lançamos mão de inversões dos fatos, prevendo melhorar a compreensão do ouvinte/leitor da narrativa; ou ainda trocamos as palavras e construções frasais originais por outras semelhantes ou sinônimas.

Outra consideração importante nesse processo é levar em conta o tempo ou limite (textual) que temos à disposição para a construção da narrativa que foi resumida. Saber disso de antemão ditará o quão detalhista ou generalista precisaremos ser na síntese dos fatos, tendo em vista que quanto mais tempo e/ou espaço tivermos, mais próximo à realidade estará a narrativa sintetizada.

Ao chegar nesse ponto do nosso exemplo é necessário nos questionar sobre algo bastante relevante: o quanto das percepções pessoais e autorais ficam na narrativa quando comparada ao evento real? O fato de haver liberdade estilística na produção da síntese, não significa que a narrativa deva veicular conteúdo falso, já que ela precisa representar o evento real. Apesar disso, é possível, em alguns casos, acrescentar e evidenciar nossas avaliações e perspectivas. O esforço do autor é pensar em diferentes operações textuais

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

que incidem sobre o efeito de sentido de verossimilhança ao conteúdo, como a paráfrase ou ainda a cópia.

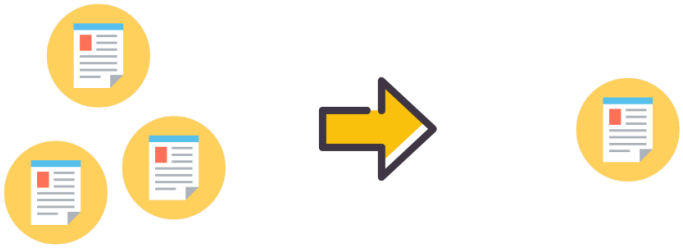
Do ponto de vista linguístico, os estudos sobre as operações associadas à transformação de texto-fonte em um texto-alvo sintetizado remetem ao conceito de **reescrita**. Para Fiad (2013), esse conceito é tido como o trabalho que é feito por um autor (com ou sem auxílio de um mediador) por meio de operações na linguagem que abrangem diferentes aspectos do texto conduzindo-o à sua modificação. Assim, podemos dizer que o processo de resumir algo está compreendido na reescrita.

Já do ponto de vista computacional, essa atividade que nos é tão comum e feita de maneira, por vezes, intuitiva, pode ser realizada automaticamente. Em processamento de linguagem natural (PLN), essa tarefa é conhecida como **sumarização automática** (doravante, SA). A SA pode ser definida como a produção automatizada de versões reduzidas de outros textos, resultando em sumários (ou "resumos"). Rino; Pardo (2003) elucidam algumas premissas para a SA, a saber:

1. Existência de um texto que deve ser condensado, tido como *texto-fonte*;
2. O texto-fonte dispõe de (i) uma ideia ou tópico principal, (ii) um conjunto de informações relacionadas entre si, (iii) um propósito comunicacional que organiza e relaciona as informações entre si e (iv) uma narrativa visando a informatividade de maneira coerente;
3. Identificação do conteúdo relevante no texto-fonte e construção de uma nova narrativa coerente no sumário, preservando a ideia inicial.

A partir dessas premissas, apresentamos um modelo de representação do processo de SA, na Figura 5.1.

Figura 5.1: Processo genérico de Sumarização Automática



Na Figura 5.1, a SA se inicia na consulta de um ou mais textos-fonte, que será(ão) submetido(s) a determinada análise para fazer a verificação do que é mais ou menos importante. Quando o processo de sumarização se baseia em apenas um texto-fonte, diz-se que a sumarização é **monodocumento**; quando baseado em dois ou mais textos-fonte, **multidocumento**. Como resultado, é elaborado o sumário, caracterizado por ser uma versão menor do(s) texto(s)-fonte. Mais adiante, detalharemos esses e outros aspectos e conceitos importantes à SA, como quantidade de textos-fonte a resumir e de línguas nesse processo, além de tendências e perspectivas metodológicas.

Com relação às tendências de **abordagens da SA**, ao longo do tempo, percebemos que as pesquisas em SA do começo dos anos 70 até o início da década de 90 tinham como motivação a recuperação de informação em textos. Por conta disso, os métodos utilizados para a sumarização nesse período se baseiam quase que exclusivamente em informações na

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

superfície textual, como a ocorrência de palavras-chave no título e no corpo do texto. No começo dos anos 90, percebeu-se que os métodos utilizados anteriormente não conseguiam dar conta de problemas linguísticos mais complexos, como resolução de correferência (Capítulo Resolução de correferência). Assim, métodos linguisticamente motivados começaram a ser implementados em sistemas de SA.

Já nas duas primeiras décadas dos anos 2000, notou-se uma tendência sócio-comportamental que também influenciou a SA: a democratização da Web. Por conta disso, nesse período, a sumarização se popularizou devido ao fato de haver um grande volume de informação produzida e circulante em relação ao pouco tempo que as pessoas tinham disponível para consumí-la. Mais recentemente, vimos a influência dos *Large Language Models* (LLMs) (Capítulo Modelos de linguagem) em PLN. Com relação à SA, a implementação desse tipo de abordagem apontou que é possível produzir sumários mais coerentes, coesos, mesmo em contextos que exigem manobras de reescrita dos textos. Em PLN há diferentes motivações que fomentam interseções com a SA. Diversas aplicações e subáreas, como compreensão de textos, recuperação da informação, indexação, mineração de textos e sistemas de pergunta-resposta, partem de texto(s)-fonte que necessita(m) ter seu(s) conteúdo(s) classificado(s) entre mais e menos importante e, a partir disso, propor uma nova versão do texto, ou em um tamanho menor ou um outro conteúdo. Além disso, os sumários também podem beneficiar pessoas que precisam ler biografias ou coleções de documentos em que seja necessário ter acesso à informação de maneira resumida.

Ao longo deste capítulo queremos introduzir conceitos importantes para a SA, bem como propor um material que você possa utilizar na construção de seus primeiros sumários. Além disso, nesta nova edição do capítulo você contará também com o acréscimo de novas reflexões sobre a utilização de LLMs na SA, sobretudo na abstrativa. Você irá perceber que em alguns momentos remontamos a história da área, pois entender as limitações que sistemas e métodos apresentaram ao longo da trajetória da SA nos permite entender quais os desafios atuais impostos ora pelas tecnologias desenvolvidas, ora pela própria língua em relação aos sistemas computacionais. Ainda, apresentamos exemplos de algoritmos introdutórios e simples para que você possa praticar a sumarização e, quem sabe, poder aplicá-la em outros cenários.

5.2 Por trás do processo

5.2.1 Sobre conceitos

Antes de discutir métodos e sistemas de geração e avaliação da SA, é necessário conhecer alguns conceitos importantes. Para tanto, destacamos as seguintes características da SA: (i) função, (ii) público-alvo, (iii) tipo, (iv) abordagem, (v) fonte, e (vi) idioma.

A versão sumarizada do texto-fonte é conhecida como sumário, que pode ser classificado como informativo, indicativo ou crítico, a depender da **função** comunicativa que exerce (Mani; Maybury, 1999). Os sumários informativos são aqueles que apresentam a informação original de maneira a preservá-la a ponto de dispensar a leitura do texto-fonte. Os sumários indicativos são orientações genéricas sobre o texto original, fazendo com que o leitor tenha um panorama sobre o conteúdo, sem ter acesso a seus detalhes, como o índice de um livro. Por fim, os sumários críticos são aqueles que permeiam entre a síntese e a avaliação do conteúdo, aproximando-se do gênero resenha.

Quando pensamos no **público-alvo**, os sumários podem ser caracterizados em genéricos ou específicos. Sumários genéricos são aqueles que extraem a informação do texto-fonte

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

levando em consideração apenas critérios técnicos, como importância do conteúdo. Já os sumários específicos, além de considerarem aspectos técnicos também devem levar em conta elementos relacionados ao leitor que afetam a construção do sumário, como o conhecimento prévio ou o interesse do público-alvo sobre o assunto do texto-fonte.

No que se refere ao **tipo**, os sumários podem ser classificados em extrativos e abstrativos (do inglês *abstracts*). Os sumários extrativos têm a característica de serem compostos por sentenças dos textos-fonte. Nesse sentido, o conteúdo não é submetido a nenhum tipo de edição linguística, mas apenas selecionado a partir de determinados critérios (como tamanho disponível do sumário e a importância do conteúdo, por exemplo). Já os sumários abstrativos são caracterizados por seu conteúdo (parcial ou integral) ter sido submetido a alguma operação linguística de reescrita, como a paráfrase.

Com relação à **abordagem**, em PLN há três grandes vertentes, organizadas em função da quantidade e profundidade de informação linguística utilizada no processo de sumarização. A abordagem profunda utiliza muito conhecimento linguístico na elaboração dos sumários. Nesse caso, podem ser levados em conta aspectos pertinentes à semântica, ao discurso e à pragmática, por exemplo, para tratar determinados fenômenos linguísticos. Já a abordagem superficial utiliza pouco conhecimento linguístico; quando o faz, utiliza majoritariamente informações dos níveis morfológico, morfossintático e sintático. Esse tipo de abordagem se baseia em conhecimento estatístico e empírico. Por fim, a abordagem híbrida seria a combinação das duas abordagens anteriores.

É importante destacar que há possibilidade de elaborar diferentes formatos de sumários a partir da abordagem escolhida. Os sumários produzidos sob a abordagem superficial tendem a ser extrativos, já que utilizam, na maioria das vezes, técnicas de contagem de *tokens* e/ou identificação de informações que estão na superfície textual. Ao passo que abordagens profundas e híbridas tendem a produzir sumários abstrativos, pois demonstram estratégias linguísticas de retextualização que podem resultar em reelaboração da forma com que o conteúdo está expresso.

Quanto à **fonte**, os sumários podem ser classificados em mono ou multidocumento. Os sumários monodocumento são resultantes apenas de um único texto que serviu de base para a sumarização. No caso dos sumários elaborados a partir de dois ou mais textos que dissertam sobre o mesmo assunto, tem-se a SA multidocumento. Este último tipo de sumarização, em especial, precisa lidar com fenômenos linguísticos que tornam o processo mais desafiador, como "redundância", "complementaridade" e "contradição" provenientes de fontes produzidas com estilos, vieses e perspectivas autorais diferentes durante a escrita dos textos[1].

Quanto ao **idioma**, os sumários podem ser produzidos a partir de uma única língua (monolíngue) ou com duas ou mais línguas (multilíngue).

Outro conceito importante na SA é a **taxa de compressão** dos sumários, isto é, quanto de informação será incluída no sumário. Essa quantidade é estabelecida em função da taxa de compressão, que é a razão entre o tamanho do sumário e o tamanho do texto-fonte (Mani, 2001). Um sumário com taxa de compressão de 70% apresenta tamanho equivalente a 30% do tamanho do texto-fonte, medido ou em função da quantidade de palavras ou de sentenças. Assim, se o texto-fonte possuir 1000 palavras e deseja-se construir um sumário com uma taxa de compressão de 0,7, o sumário terá 300 palavras. No caso da SA multidocumento, a taxa de compressão pode ser definida em relação ao maior texto ou à soma total de palavras ou de sentenças de todos os textos. Porém, em um cenário de

[1]Sugerimos a leitura de (Souza et al., 2011a), (Souza; Felippo, 2018), (Souza, 2019) e (Silva; Di Felippo, 2014) que se aprofundam nos referidos fenômenos linguísticos.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

sumarização abstrativa, é preciso ponderar que nem sempre os valores estabelecidos nas taxas de compressão farão segmentações de sentenças de maneira devida. Nesse sentido, com relação à informatividade, é preciso avaliar o conteúdo que já foi escolhido para compor o sumário, caso a taxa de compressão pretendida exceda a quantidade de palavras estimada para o sumário.

5.2.2 Sobre o processo de sumarização

A seguir, apresentamos uma possível análise de um sumário retirado do *corpus* CSTNews[2] (Cardoso et al., 2011). Os textos do Quadro 5.1 serviram de fonte de informação para a elaboração de sumários extrativos multidocumento do Quadro 5.2.

Quadro 5.1: Textos-fonte extraídos do *corpus* CSTNews

Texto-fonte 1	**Texto-fonte 2**
O médico pessoal do argentino Diego Maradona, Alfredo Cahe, revelou nesta segunda-feira que uma recaída da hepatite aguda de que sofre foi o motivo da nova internação do ex-craque. Maradona havia recebido alta no último dia 11, mas voltou a ser internado na sexta-feira e os boletins médicos não especificaram o que se passava com o ex-jogador –Cahe descartou pancreatite ou úlcera. "Maradona teve uma recaída na hepatite aguda. Agora está estável. Apesar de ter melhorado no domingo, deverá continuar internado", disse Cahe, em declarações ao jornal "La Nación". Maradona, 46, desenvolveu um hepatite tóxica por excesso de consumo de álcool, o que já o manteve internado durante 13 dias antes da primeira alta. Cahe disse ainda que Maradona não voltou a consumir bebidas alcoólicas e que as causas da recaída estão sendo investigadas.	BUENOS AIRES - Maradona voltou a ter problemas de saúde no fim de semana. Internado em um hospital em Buenos Aires, ele teve uma recaída e voltou a sentir dores devido a hepatite aguda que o atinge, segundo seu médico pessoal, Alfredo Cahe. "Agora está estável. Mesmo com esta melhora, ele continuará internado", disse o médico, que descartou a possibilidade do ex-jogador ter uma pancreatite (inflamação do pâncreas, órgão situado atrás do estômago e que influencia na digestão). Cahe reforçou que Maradona ainda tem problemas. "Os valores hepáticos dele na avaliação não estão equilibrados e ele não está bem. Mas não é nada grave", afirma, em entrevista ao diário La Nación. No domingo, Maradona assistiu ao empate por 1 a 1 no clássico Boca Juniors e River Plate pela televisão. Os torcedores do Boca, que compareceram em grande número ao Estádio La Bombonera, levaram muitas faixas e bandeiras com mensagens de apoio ao ídolo argentino. Sua filha, Dalma, foi ao estádio assistir ao jogo.

Os textos resumidos apresentam uma taxa de compressão de 70% do total de palavras do conjunto de textos-fonte. Assim, o humano poderia selecionar sentenças inteiras para compor seu sumário; caso a próxima sentença ultrapassasse muito o tamanho pré-estabelecido, permitia-se ter um sumário com menos palavras[3].

[2] O *corpus* CSTNews é um repositório de textos jornalísticos publicados no ano de 2007, originalmente escritos em português. Ele está organizado em 50 conjuntos de textos, em que cada conjunto possui de dois a três textos sobre um mesmo assunto. Cada conjunto de textos inclui anotações linguísticas (como morfossintática, semântica e discursiva, por exemplo), além de apresentar sumários humanos e automáticos. Esse repositório serviu de base para muitas pesquisas em SA em língua portuguesa, e está disponível em: http://nilc.icmc.usp.br/CSTNews/.

[3] Para saber mais detalhes sobre a criação de mais sumários humanos para o *corpus* CSTNews, consulte

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Quadro 5.2: Sumários extraídos do *corpus* CSTNews

Sumário humano multidocumento	**Sumário automático multidocumento**
BUENOS AIRES - Maradona voltou a ter problemas de saúde no fim de semana. <Texto-fonte 2> Internado em um hospital em Buenos Aires, ele teve uma recaída e voltou a sentir dores devido a hepatite aguda que o atinge, segundo seu médico pessoal, Alfredo Cahe. <Texto-fonte 2> Agora está estável. Apesar de ter melhorado no domingo, deverá continuar internado", disse Cahe, em declarações ao jornal "La Nación". <Texto-fonte 1> Maradona, 46, desenvolveu um hepatite tóxica por excesso de consumo de álcool, o que já o manteve internado durante 13 dias antes da primeira alta. <Texto-fonte 1> Cahe disse ainda que Maradona não voltou a consumir bebidas alcoólicas e que as causas da recaída estão sendo investigadas. <Texto-fonte 1>	Internado em um hospital em Buenos Aires, ele teve uma recaída e voltou a sentir dores devido a hepatite aguda que o atinge, segundo seu médico pessoal, Alfredo Cahe. <Texto-fonte 2> "Maradona teve uma recaída na hepatite aguda. Agora está estável. Apesar de ter melhorado no domingo, deverá continuar internado", disse Cahe, em declarações ao jornal "La Nación". <Texto-fonte 1> Cahe disse ainda que Maradona não voltou a consumir bebidas alcoólicas e que as causas da recaída estão sendo investigadas. <Texto-fonte 1>

Os textos-fonte do Quadro 5.1 noticiam o estado de saúde do ex-jogador de futebol Maradona e sobre seu processo de internação para tratar de um quadro de hepatite aguda. Apesar de serem textos jornalísticos, perceba que eles apresentam estilos de escrita diferentes, como o uso mais extensivo do discurso direto; ou ainda a construção narrativa dando ora mais detalhes, ora menos.

Os textos dos sumários extrativos também apresentam particularidades entre os sumarizadores: das 5 sentenças escolhidas pelo humano, 3 foram extraídas do Texto-fonte 1, admitindo que ele pode ter preferência por alguma fonte que julgue ser de sua confiança; ao passo que o sistema automático seleciona apenas 1 advinda do Texto-fonte 2, das 3 sentenças escolhidas. Diante disso, é possível inferir que o primeiro texto-fonte apresenta o conteúdo de maneira a ganhar a predileção de ambos os sumarizadores, seja pelo estilo, composição ou qualidade da informação.

Apesar de ambos os sumarizadores terem escolhido sentenças em comum, algo a se destacar é o fato de os textos sumarizados apresentarem estruturas narrativas discretamente diferentes. No sumário automático, a sentença inicial é apenas a segunda do texto sumarizado por um humano. Assim, é possível inferir que na sumarização manual elegeu-se uma sentença para abrir o texto que pudesse introduzir melhor o assunto, antes de chegar ao cerne da informação.

Também é possível identificar que os textos sumarizados, em relação aos textos-fonte, mantêm uma ordenação da informação: o início do texto apresenta informações mais importantes em relação ao seu desfecho. Nesse caso, os sumarizadores precisaram de alguma forma identificar as informações mais relevantes dos textos-fonte e, em seguida, refletirem como esse conteúdo se organizaria com relação ao tempo e à progressão temática no próprio sumário.

A identificação de conteúdo relevante não é uma tarefa trivial. Desde o advento da SA, um dos maiores desafios para as pesquisas foi desenvolver maneiras de identificar a

(Dias et al., 2014).

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

informação tida como a mais relevante no(s) texto(s)-fonte. Conforme foram superados alguns desafios, outros foram identificados e passaram a pautar muitos estudos, como a SA multilíngue, de atualização (update summarization) e abstrativa.

5.3 Avaliação na SA

Usualmente, a avaliação de desempenho dos sistemas de SA é medida por meio da análise de seus sumários, considerando os critérios de **informatividade** e **qualidade** dos textos produzidos. A informatividade geralmente é calculada de forma automática e consiste em verificar quanto da informação relevante dos textos-fonte é preservada no sumário automático. A avaliação da qualidade, por sua vez, é tradicionalmente realizada por humanos, pois o foco reside na análise de aspectos relacionados à gramaticalidade, coesão e coerência, foco e clareza referencial – elementos para os quais ainda não há métodos automáticos plenamente eficazes de avaliação. Ainda, ao final deste capítulo discutiremos brevemente os avanços que vêm sendo obtidos por meio dos LLMs.

Para medir a informatividade, a métrica de avaliação amplamente utilizada é a ROUGE (*Recall-Oriented Understudy for Gisting Evaluation*) (Lin, 2004). Essa medida compara automaticamente a quantidade de n-gramas (conjunto de palavras em sequência) em comum entre um sumário automático e um ou mais sumários de referência, ou seja, produzido por um humano. Porém, em cenários em que não houver sumários de referência, o texto-fonte pode ser adotado como critério para avaliação (Louis; Nenkova, 2013). A ROUGE possui quatro variantes principais, descritas a seguir:

a) ROUGE-N (*ROUGE-Ngram*): Avalia a sobreposição de n-gramas entre o sumário automático e o sumário de referência. A ROUGE-1, por exemplo, considera a sobreposição de unigramas (palavras individuais), enquanto o ROUGE-2 considera a sobreposição de bigramas (pares de palavras consecutivas), e assim por diante.

b) ROUGE-L (*ROUGE-Longest Common Subsequence*): Mede a similaridade baseada na sequência de palavras comuns mais longa entre o sumário automático e o texto de referência. A ROUGE-L não exige correspondência consecutiva, mas sim a mesma ordem das palavras tanto no sumário automático quanto no sumário de referência.

c) ROUGE-W (*ROUGE-Word)*: Calcula as correspondências consecutivas mais longas entre o sumário automático e o sumário de referência.

d) ROUGE-S (*ROUGE-Skip-bigram*): Essa métrica considera pares de palavras com um certo número de palavras intermediárias "puladas". Ela é útil para capturar dependências de longa distância entre palavras.

O resultado ROUGE é dado em termos de precisão (Equação 5.1), revocação (Equação 5.2) e medida-f (Equação 5.3), e possui correlação com a avaliação humana (Lin, 2004). A precisão (P) expressa a proporção de n-gramas coincidentes entre os sumários automático e de referência em relação ao número de n-gramas do sumário automático. A revocação (R) representa a proporção de n-gramas coincidentes entre os sumários automático e de referência em relação ao número de n-gramas do sumário de referência. Tais medidas são complementares e, por isso, costuma-se utilizar a medida-f (F) que representa a média harmônica entre precisão e revocação. Como precisão e revocação são inversamente relacionadas, uma tende a diminuir quando a outra sofre um aumento.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

$$\text{Precisão} = \frac{\text{Número de n-gramas em comum com o sumário automático}}{\text{Número de n-gramas do sumário automático}} \tag{5.1}$$

$$\text{Revocação} = \frac{\text{Número de n-gramas em comum com o sumário automático}}{\text{Número de n-gramas do sumário referência}} \tag{5.2}$$

$$\text{Medida f} = \frac{2 * P * C}{P * C} \tag{5.3}$$

Para exemplificar, considere o seguinte fragmento como texto-fonte "*o gato estava embaixo da cama*" e a sentença candidata "*o gato foi encontrado embaixo da cama*". Sem considerar os procedimentos de pré-processamento, as palavras em comum das duas sentenças são [o, gato, embaixo, da, cama]; então, os valores para revocação, precisão e medida-f considerando a ROUGE-1 são respectivamente: 0,83, 0,71 e 0,76.

Por ser rápida, barata e não sujeita à subjetividade, a ROUGE é uma das medidas mais populares para avaliar sumários (p.ex. (Cardoso, 2014), (Gambhir; Gupta, 2017), (Aliguliyev et al., 2019), (Parida; Motlicek, 2019), (Paiola et al., 2022)). A correlação dessa medida com o julgamento humano aumenta quando se utiliza vários sumários de referência. No entanto, a ROUGE aborda somente a capacidade de seleção de conteúdo dos sistemas, ignorando aspectos de qualidade linguística como coerência e gramaticalidade. Nesse caso, como o cálculo da pontuação ROUGE é realizado a partir da sobreposição de n-gramas, se um sumário apresentar o mesmo conteúdo do sumário de referência, contendo as mesmas ideias, mas utilizando expressões diferentes, a pontuação final não deve ser muito alta, apesar da qualidade do sumário. Essa, inclusive, é uma limitação da medida ROUGE na avaliação de sumários abstrativos (Gupta; Gupta, 2019).

Nesse sentido, em 2005, a DUC (*Document Understanding Conference)*[4] introduziu cinco propriedades linguísticas para avaliar a qualidade dos sumários automáticos e que não utiliza sumários de referência, a saber[5]:

(i) **gramaticalidade:** que diz respeito à ausência de erros de ortografia, pontuação e sintaxe;

(ii) **não redundância:** que se refere à ausência de informações repetidas;

(iii) **clareza referencial:** que diz respeito à clara identificação dos componentes da superfície textual que fazem remissão a outro(s) elemento(s) do sumário;

(iv) **foco:** se refere ao fato de que as informações de uma sentença devem se relacionar com as informações do restante do sumário;

(v) **estrutura e coerência:** que diz respeito à organização do sumário considerando sua textualidade.

Para avaliar de acordo com os critérios estabelecidos pela DUC, coleta-se a opinião de um grupo de juízes sobre um mesmo sumário e calcula-se a média para cada critério julgado. Cada anotador atribui uma nota que varia de 1 (muito ruim) a 5 (muito bom).

[4]Em 2008, a DUC passou a se chamar TAC (*Text Analysis Conference*). Disponível em: https://tac.nist.gov/.

[5]Original: https://www-nlpir.nist.gov/projects/duc/duc2006/quality-questions.txt.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Vale destacar que, embora esse tipo de avaliação não dependa do uso de sumário de referência, ela pode beneficiar sumários automáticos que sejam bastante diferentes daqueles de referência. Provavelmente, esses sumários automáticos teriam notas muito baixas pela ROUGE, mas, ainda assim, poderiam ser considerados informativos e coerentes. Por essa razão, alguns trabalhos de SA utilizaram tais medidas avaliativas, além da ROUGE (Ermakova et al., 2019; Pitler et al., 2010).

Conforme Luhn (1958), a produção de sumários é uma tarefa intelectual e que sofre influência da familiaridade com o assunto, atitude e disposição do produtor. O autor também sugere que a produção de sumários de referência pode depender dos interesses de quem o produz, dos interesses dos leitores e da importância subjetiva que ele atribui às informações textuais. Assim, algumas ponderações podem ser feitas a respeito disso: os humanos divergem na escolha de informações importantes; e existe a possibilidade de que a mesma pessoa, ao resumir novamente um texto-fonte, crie um sumário totalmente diferente do anterior (Hailu et al., 2020; Luhn, 1958).

Contudo, a necessidade de anotações ou avaliações manuais muitas vezes inviabiliza suas aplicações na avaliação de métodos de sumarização, em particular para bases de dados com uma quantidade massiva de textos. Por esse motivo, mesmo para trabalhos voltados à SA abstrativa é comum encontrar o uso de medidas mais simples, que podem ser calculadas de forma automática, como a ROUGE. Embora seja factual, há esforços recentes para desenvolver métricas mais sensíveis ao nível semântico, a partir de modelos de linguagem como avaliadores.

Além disso, no cenário atual em que a grande área de geração de texto em linguagem natural evolui rapidamente, críticas surgem a medidas de avaliação que dependem estritamente da sobreposição de unidades de n-gramas entre sumários de referência e candidatos, o que não é adequado para medir a qualidade de sumários abstrativos. Assim, têm surgido frentes de trabalhos que buscam criar protocolos de avaliação desde as métricas aplicadas até a qualidade dos conjuntos de dados utilizados.

Algumas métricas recentes buscam avaliar a similaridade entre textos com base em seus significados, aproximando a análise da semântica. Essas métricas utilizam representações vetoriais de palavras e sentenças (os chamados *embeddings*), o que permitem captar relações mais sutis entre textos com vocabulários diferentes, mas significados semelhantes. Um exemplo é o BERTScore (Zhang et al., 2020), que calcula a similaridade entre cada *token* do sumário automático e do sumário de referência utilizando *embeddings* extraídos de modelos como o BERT. Nesse sentido, essa métrica pode reconhecer paráfrases ou reformulações que mantêm o conteúdo original, superando algumas das limitações da ROUGE. Ainda que métricas como o BERTScore possam apresentar maior correlação com julgamentos humanos e serem mais robustas a variações linguísticas, seu uso ainda é restrito devido à necessidade maior de poder computacional para rodar modelos de linguagem para a extração de *embeddings*. Além disso, medidas de similaridade semântica ainda precisam de um sumário de referência, podendo penalizar um sumário candidato que apresente uma seleção de fatos diferente do apresentado no sumário de referência, mesmo que fosse considerado por um avaliador humano como um bom resumo do texto-fonte.

Mais recentemente, com a popularização dos LLMs, como o ChatGPT, têm ganhado espaço abordagens em que os próprios modelos são utilizados como avaliadores de sumários. Essa nova linha de pesquisa, conhecida como *LLM-as-a-Judge*, propõe que o modelo desempenhe o papel de juiz, avaliando diretamente a qualidade de um sumário com base em critérios específicos, como factualidade, coerência, fluência ou mesmo preferências humanas expressas em anotações anteriores. Um dos principais atrativos dessas abordagens

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

é a flexibilidade: com um único modelo, é possível avaliar diversos aspectos qualitativos do texto, com menor necessidade de métricas separadas ou protocolos elaborados (Gu et al., 2025).

Na SA, por exemplo, a utilização de LLMs já demonstrou que é possível avaliar sumários em múltiplas dimensões, como informatividade e coerência, por exemplo, atribuindo notas em escalas (por exemplo, de 1 a 5). As avaliações sobre esses métodos têm mostrado boa correlação com julgamentos humanos e podem ser utilizados mesmo quando há múltiplos sistemas sendo comparados (Gao et al., 2023). Além disso, a mesma abordagem pode ser estendida para domínios multimodais ou especializados, desde que os LLMs tenham sido ajustados para tais contextos.

Apesar do grande potencial, o uso de LLMs como avaliadores ainda enfrenta alguns desafios. Um deles é a necessidade de sumários de referência, especialmente em tarefas que requerem comparação com um texto ideal. Quando o sumário automático inclui informações implícitas ou inferenciais que não estão objetivamente no texto-fonte, o modelo pode ter dificuldade para julgar sua adequação sem informações externas ou conhecimento de mundo – o que pode levar a equívocos ou inconsistências. Outro ponto crítico diz respeito à possibilidade de alucinações: como os LLMs são treinados com grandes volumes de texto da Web, eles podem emitir julgamentos, correlações e/ou inferências com base em "conhecimentos" incorretos, incompletos ou não factuais. Por fim, ainda que tragam escalabilidade e redução de dependência de humanos, o uso de LLMs não é isento de custos: executar modelos de língua, sobretudo os comerciais, para avaliar milhares de sumários pode ser financeira e computacionalmente oneroso, principalmente em comparação com métricas tradicionais como a ROUGE.

Portanto, embora a utilização de LLMs como juízes na avaliação automática de sumários represente um avanço significativo, questões relacionadas à confiabilidade, à robustez e ao custo ainda precisam ser cuidadosamente consideradas. Como destacado no survey recente sobre LLM-as-a-Judge (Gu et al., 2025), há um esforço crescente da comunidade para padronizar metodologias, mitigar vieses e ampliar benchmarks específicos para avaliação automática. Ainda assim, o uso responsável e crítico dessas ferramentas é essencial para garantir que os sistemas de avaliação acompanhem de forma justa e precisa a evolução dos modelos de sumarização.

5.4 Aplicação em Sumarização Extrativa

De maneira geral, os sumários são produzidos a partir de três etapas[6], conforme a Figura 5.2.

Na primeira etapa, conhecida como **análise**, os textos-fonte passam por um processo de interpretação, resultando em uma representação formal deles. Na fase seguinte, tida como **transformação**, o sumário é elaborado a partir do ranqueamento de segmentos dos textos-fonte em função de algum critério que indique a relevância, sendo selecionados os segmentos com maiores pontuações até que se atinja a taxa de compressão desejada. Na etapa de **síntese**, gera-se, então, o sumário em língua natural a partir do conteúdo selecionado. Nessa etapa podem ser utilizados métodos de tratamento de correferência, fusão, linearização, justaposição e ordenação de sentenças. Essas três fases contribuem umas com as outras, de modo que alguns métodos que ocorrem na síntese também poderiam

[6]Para mais detalhes sobre as três etapas dessa arquitetura genérica, indicamos (Jones, 1993) e (Mani, 2001).

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Figura 5.2: Etapas do processo de sumarização



Fonte: Adaptado de (Sparck-Jones, 1998).

estar na fase de transformação, e vice-versa.

A partir dessa arquitetura genérica, apresentamos nesta seção dois algoritmos com estratégias distintas de seleção de conteúdo para gerar os sumários. É possível que para compreender algumas questões que levantaremos aqui sejam necessárias informações prévias. Pensando nisso, elaboramos o Apêndice A, em que discutimos aspectos relevantes ao pré-processamento dos textos para SA e que complementa o assunto abordado nesta seção. Ademais, para complementar a leitura, disponibilizamos um notebook[7] virtual com os algoritmos[8] apresentados neste capítulo. Sugerimos que a leitura desta seção seja acompanhada desse material suplementar.

5.4.1 Algoritmo baseado na frequência das palavras

Uma das unidades mínimas de análise linguística é a palavra, compreendida dentro do nível morfológico; quando combinadas entre si, é possível construir outros objetos de análise, como a sentença. Para a sumarização de textos escritos, as palavras e as sentenças podem ser o início do processo de identificação de conteúdo relevante. Os métodos clássicos de identificação de segmentos relevantes avaliam a importância de cada sentença em um texto-fonte com base em seu peso e, em seguida, selecionam aquelas com pesos mais elevados (acima de um limite mínimo) para formar o resumo (Baxendale, 1958; Edmundson, 1969; Luhn, 1958). As propostas mais recentes utilizam *word embeddings* (Hailu et al., 2020) e modelos de redes neurais profundas, como BERT (Liu; Lapata, 2019).

Nesta seção, apresentamos uma estratégia simples baseada na frequência de palavras para selecionar sentenças relevantes do texto-fonte para compor os sumários. Nessa perspectiva, quanto mais uma palavra se repete no texto, mais relevante ela é. Propusemos um algoritmo geral, organizado em sete passos, que seleciona as sentenças com maior número de palavras relevantes para compor um possível sumário. Vamos aplicar o algoritmo ao Texto-fonte 1 do Quadro 5.1 com o objetivo de determinar um *ranking* de sentenças para a composição do sumário.

Quadro 5.3: Algoritmo geral

[7] https://colab.research.google.com/drive/1qGfpxQJgnJy_UKKiHpKQSrACiFRwjyeP?usp=sharing Os resultados podem ser um pouco diferentes dos que são apresentados aqui, devido aos testes realizados durante a escrita do texto.

[8] *Grosso modo*, algoritmo pode ser definido como um conjunto de regras, compreendendo uma sequência finita de ações visando a resolução de alguma tarefa/problema.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Pré-processamento
Passo 1: Normalização do texto
Passo 2: Remoção de caracteres de pontuação
Passo 3: Remoção de *stopwords*
Geração do sumário
Passo 4: Cálculo das frequências das palavras que compõe o texto pré-processado
Passo 5: Determinação do peso de cada palavra
Passo 6: Cálculo do peso de cada sentença
Passo 7: Seleção das sentenças com maiores pesos até atingir a taxa de compressão

Ao executar o **Passo 1**, todo o texto será transformado para letras minúsculas para fins de padronização. Esse procedimento faz com que tenhamos palavras graficamente distintas a serem consideradas de maneira similar, como "hepatite" e "Hepatite".

Após a normalização, no **Passo 2**, devem ser removidos os caracteres de pontuação. Essa operação se justifica pelo fato de que uma palavra, acompanhada de uma pontuação, difere-se de outra com a mesma grafia sem pontuação (como "cahe" e "cahe,"). Os caracteres de pontuação podem divergir de um método para outro, a partir de diferentes linguagens de programação ou *software*. Aqui, consideramos a pontuação sendo o ponto final, a vírgula, a sequência de dois hífens (--) e as aspas (" ou ").

Já no **Passo 3**, removemos as *stopwords* utilizando a lista do Natural Language Toolkit – NLTK – (Bird, 2006). Por exemplo, o trecho "o médico pessoal do argentino diego" torna-se "médico pessoal argentino diego" após a remoção dos *tokens* "o" e "do".

É importante destacar que a remoção das *stopwords* se baseia na hipótese da distribuição (Capítulo Representação vetorial e semântica distribucional). Assim, na SA com abordagem lexical, acredita-se que após a remoção das *stopwords*, as palavras restantes representam melhor o conteúdo e aquelas que mais se repetem, são relevantes. Assim, no **Passo 4**, deve-se calcular a frequência de ocorrência dos *tokens*. Os cinco termos mais frequentes, em ordem alfabética do Texto-fonte 1 estão listados na Tabela 5.1.

Tabela 5.1: Cinco termos mais frequentes do Texto-fonte 1

Termo	Frequência	Peso
maradona	5	5/5 = 1,0
cahe	4	4/5 = 0,8
hepatite	3	3/5 = 0,6
internado	3	3/5 = 0,6
recaída	3	3/5 = 0,6

A partir da frequência das palavras, no **Passo 5** conseguimos determinar diferentes pesos para os *tokens* com base na frequência relativa. O cálculo é definido pela divisão da frequência de cada termo pela maior frequência calculada. No caso, o termo "maradona" é o que mais ocorre no texto, com frequência 5. Assim, todos os termos do texto devem ser divididos por 5, determinando o peso de cada um deles. Ressalta-se que essa não é a única forma de ponderar a relevância de uma palavra ou sentença em um texto. No exemplo, para fins de simplificação, optamos por considerar a fala do médico como uma única sentença por se tratar de um discurso direto compreendido entre aspas.

Uma vez calculados os pesos de cada uma das palavras, podemos estimar a pontuação das sentenças. Considerando o Texto-fonte 1, no **Passo 6,** definimos a pontuação por meio da soma dos pesos da frequência relativa de cada palavra restante em cada sentença.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Na Tabela 5.2, apresenta-se o peso final de cada sentença. Por exemplo, a Sentença 1 tem o peso 5,8, pois a soma dos pesos dos *tokens* (destacados no texto) resulta nesse valor.

Tabela 5.2: Sentenças do texto-fonte com seus respectivos pesos

	Sentença	**Marcações**	**Peso**
1	O médico pessoal do argentino Diego Maradona, Alfredo Cahe, revelou nesta segunda-feira que uma recaída da Hepatite aguda de que sofre foi o motivo da nova internação do ex-craque.	O (médico - 0,2) (pessoal -0,2) do (argentino - 0,2) (Diego - 0,2) (Maradona - 1,0), (Alfredo - 0,2) (Cahe - 0,8), (revelou - 0,2) (nesta - 0,2) (segunda-feira - 0,2) que uma (recaída - 0,6) da (Hepatite - 0,6) (aguda - 0,4) de que sofre fo (motivo - 0,2) da (nova - 0,2) (internação - 0,2) do (ex-craque - 0,2).	5,8
2	Maradona havia recebido alta no último dia 11, mas voltou a ser internado na sexta-feira e os boletins médicos não especificaram o que se passava com o ex-jogador – Cahe descartou pancreatite ou úlcera.	(Maradona - 1,0) (havia - 0,2) (recebido - 0,2) (alta - 0,4) no (último - 0,2) (dia - 0,2) (11 - 0,2), mas (voltou - 0,4) a ser (internado - 0,6) na (sexta-feira - 0,2) e os (boletins - 0,2) (médicos - 0,2) não (especificaram - 0,2) o que se (passava - 0,2) com o (ex-jogador - 0,2) (Cahe - 0,8) (descartou - 0,2) (pancreatite - 0,2) ou (úlcera - 0,2).	6,0
3	“Maradona teve uma recaída na hepatite aguda. Agora está estável. Apesar de ter melhorado no domingo, deverá continuar internado”, disse Cahe, em declarações ao jornal “La Nación”.	“(Maradona - 1,0) teve uma (recaída - 0,6) na (hepatite - 0,6) (aguda - 0,4). (Agora - 0,2) está (estável - 0,2). (Apesar - 0,2) de ter (melhorado - 0,2) no (domingo - 0,2), (deverá - 0,2) (continuar - 0,2) (internado - 0,6)”, (dis (Cahe - 0,8), em (declarações - 0,2) ao (jornal - 0,2) “(La - 0,2) (Nación - 0,2)”.	6,6
4	Maradona, 46, desenvolveu uma hepatite tóxica por excesso de consumo de álcool, o que já o manteve internado durante 13 dias antes da primeira alta.	(Maradona - 1,0), (46 - 0,2), (desenvolveu - 0,2) um (hepatite - 0,6) (tóxica - 0,2) por (excesso - 0,2) de (consumo - 0,2) de (álcool - 0,2), o que já o (manteve - 0,2) (internado - 0,6) (durante - 0,2 (13 - 0,2) (dias - 0,2) (antes - 0,2) da (primeira - 0,2) (alta - 0,4).	5,0
5	Cahe disse ainda que Maradona não voltou a consumir bebidas alcoólicas e que as causas da recaída estão sendo investigadas.	(Cahe - 0,8) (disse - 0,4) (ainda - 0,2) que (Maradona - 1,0) não (voltou - 0,4) a (consumir - 0,2) (bebidas - 0,2) (alcoólicas - 0,2) e que as (causas - 0,2) da (recaída - 0,6) estão (sendo - 0,2) (investigadas - 0,2).	4,6

No **Passo 7**, ocorre a seleção de conteúdo relevante. Para fins de exemplificação, a taxa de compressão utilizada foi de 60% em relação ao número de sentenças originais. Apesar disso, destacamos que geralmente a taxa de compressão é calculada em função do número de palavras do texto-fonte. Assim, o sumário poderá ter duas sentenças, conforme se mostra no Quadro 5.4.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Quadro 5.4: Sumário gerado com algoritmo de frequência de palavras

> Maradona havia recebido alta no último dia 11, mas voltou a ser internado na sexta-feira e os boletins médicos não especificaram o que se passava com o ex-jogador – Cahe descartou pancreatite ou úlcera.
>
> "Maradona teve uma recaída na hepatite aguda. Agora está estável. Apesar de ter melhorado no domingo, deverá continuar internado", disse Cahe, em declarações ao jornal "La Nación".

A partir desse algoritmo, é importante fazermos algumas considerações:

- **termos com maiores frequências:** a depender do estudo, talvez seja interessante remover os termos que ocorrem em maior frequência, pois podem não determinar exatamente a essência do texto. Suponha que estamos analisando textos com a palavra "covid". Se o *corpus* for criado a partir desse termo, é quase certo que todos os textos que o compõem devam possuí-lo. O pesquisador deve avaliar se é interessante tê-lo como fator determinante ou inseri-lo em sua *stoplist*;
- **frases curtas *versus* longas:** avaliando essa comparação, as frases com mais palavras tendem a ter um maior peso, porém podem não ser tão relevantes. Conforme demonstrado na Tabela 5.2, por exemplo, a Frase 1 tem 29 palavras e possui peso 5,8. Já a Frase 2 possui 33 palavras com peso 6 e, a Frase 3 apresenta 27 palavras e o maior peso 6,6. A forma como o algoritmo determina o peso de uma frase deve ser criteriosamente avaliada;
- ***stoplist***: a lista de *stopwords* pode interferir no resultado, uma vez que determina quais palavras devem ser removidas, influenciando nos pesos atribuídos pelos métodos aplicados. Portanto, os resultados dos métodos de SA podem variar conforme a *stopwlist* utilizada;
- **a ordenação do conteúdo selecionado**: a tarefa de ordenação de sentenças em sumarização multidocumento é difícil quando comparada com a sumarização monodocumento. O conteúdo é extraído de diferentes fontes e, portanto, nenhum documento pode fornecer a ordenação adequada (Barzilay et al., 2001). Como o sumário do Quadro 5.4 é monodocumento, o conteúdo selecionado foi organizado na ordem em que apareceu no texto-fonte.

5.4.2 Algoritmo baseado em Transformers

O modelo de *Bidirectional Encoder Representations from Transformers* (BERT) (Capítulo Modelos de linguagem) é considerado um dos melhores algoritmos para sumarização da atualidade, apresentando resultados expressivos (Rogers et al., 2021) em diferentes áreas. O algoritmo foi desenvolvido nos laboratórios da Google, a fim de implementar melhorias no algoritmo de busca, permitindo converter as intenções dos usuários (termo de busca) para a codificação adequada contribuindo para otimizar e melhorar a assertividade (LIVESO, 2020). Outra contribuição é a possibilidade do treinamento do algoritmo para atender a um determinado idioma. Algumas empresas direcionaram seus serviços para o treinamento dos idiomas e áreas específicas com demanda, como o português (NEURALMIND, 2020) e o meio empresarial (Julião, 2024). Em suma, o modelo BERT, por meio de técnicas avançadas de PLN, possibilitou uma melhor compreensão da informação textual

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

por parte dos modelos matemáticos, o que por sua vez permitiu melhorias e uma ampla gama de aplicações em diversas áreas.

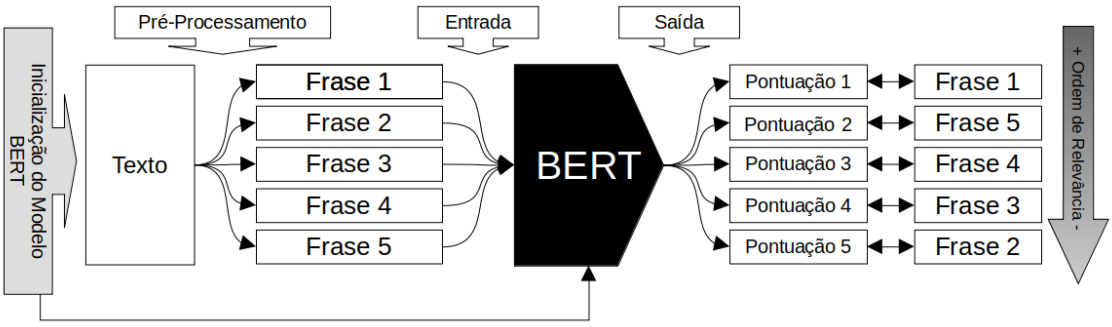
Diferente de seus antecessores, aqui o algoritmo se baseia em determinar as palavras-chave e variações para atribuir relevância aos trechos de um texto a fim de sumarizá-lo. Os antecessores, na maioria, analisavam as palavras individualmente, ou às vezes, considerando os termos ao entorno da palavra-chave requisitada. O BERT mudou essa estratégia, permitindo mapear o texto de forma geral. Dessa forma, o algoritmo possibilita até prever termos a partir de um contexto, ampliando as possibilidades de aplicações.

O processamento dos primeiros algoritmos era fundamentado em processar o texto, independente de o texto estar repetido ou não entre a coleção de documentos, para gerar uma lista das palavras-chave e atribuir pontuação às sentenças. Nesse sentido, até então, o modelo não possuía uma memória, demandando um novo processamento a cada texto a ser analisado. Essa questão foi contornada após a criação do BERT e apresentado por Devlin et al. (2019), sendo implementado por meio da tecnologia *Transformers* (Vaswani et al., 2017b).

O BERT pode ser aplicado tanto na sumarização abstrativa[9] quanto extrativa; porém, aqui, nos deteremos à segunda. O algoritmo é baseado em uma aplicação com treinamento bidirecional. Diferente dos métodos que focam na estrutura composicional das sentenças, que processam as palavras ora da esquerda para direita, ora da direita para esquerda (como os primeiros algoritmos de sumarização, com destaque para o Luhn (1958), Earl (1970) e Edmundson (1969)), no BERT o tratamento ocorre lendo os dados nos dois sentidos durante o treinamento. Desse modo, o algoritmo permite que a palavra seja analisada dentro de um "contexto", levando em consideração o que acontece ao seu redor.

Vaswani et al. (2017b) ressaltam a importância do treinamento bidirecional e demonstram que, uma vez treinado o modelo para o idioma desejado, reduz-se a necessidade de arquiteturas específicas, permitindo uma generalização na solução de desafios, como a escolha do conteúdo mais relevante nos sumários. Portanto, uma vez realizado um pré-treinamento, basta realizar um ajuste fino na rede para atender a casos específicos. No caso da tarefa de SA, o modelo pode ser utilizado para classificar e extrair as sentenças mais relevantes de um texto. Aqui, organizamos o processo em cinco passos, conforme a Figura 5.3.

Figura 5.3: Ilustração do processo simplificado de sumarização do BERT



De acordo com a Figura 5.3, no **Passo 1**: o BERT deve ser inicializado, carregando

[9]Quanto à sumarização abstrativa, sugerimos a leitura de (Barros, 2022), (Nascimento, 2023) e (Silva, 2022).

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

algumas configurações do modelo já treinado ("Inicialização do Modelo BERT"). No **Passo 2**, o texto a ser fornecido ao BERT passa por um pré-processamento, que compreende segmentação em sentenças e tokenização. O pré-processamento pode incluir outras etapas, como visto no algoritmo apresentado na seção anterior. Após os ajustes iniciais, o texto é submetido ao modelo na camada de codificação ("Entrada"). Ao entrar na rede, os *tokens* passam a compor um vetor numérico e constituem um espaço vetorial[10] para mapear o texto. Já no **Passo 3**, a partir dos vetores no espaço vetorial e do processamento do modelo BERT, em sua camada de decodificação ("Saída"), atribui-se um valor de importância de cada palavra no contexto do texto analisado. No **Passo 4**, com o valor de importância de cada palavra, torna-se possível atribuir um valor (ou "importância") à sentença que a possui ("Pontuação"). Uma vez definido esse valor para cada sentença, é possível classificá-las em ordem de relevância. Por fim, no **Passo 5**, a partir do nível de compressão informado no início do processamento pelo usuário, o BERT retorna as sentenças para compor o sumário de acordo com a taxa de compressão escolhida. Para exemplificar, são apresentados dois sumários gerados pelo BERT. O Sumário 2 segue a taxa de compressão do algoritmo anterior que é de 60% em relação ao número de sentenças originais e o Sumário 3 coleta uma sentença a mais, isto é, três sentenças.

Caso o sumário produzido não seja condizente com as expectativas, o modelo BERT pode sofrer ajustes finos para aprimorar os resultados. A tecnologia do BERT já fornece um modelo treinado a partir de dois *corpora* volumosos (a saber, *BookCorpus*[11] e Wikipedia). O treinamento do BERT consiste em fornecer ao algoritmo um *corpus* específico da área requisitada, o qual deve ser modificado para se adequar a essa etapa. As adequações previstas são *texto-previsão* - a criação de um conjunto de texto (frases); *previsão da próxima frase*; e *máscara* - modelagem de linguagem mascarada. A adequação texto-previsão consiste em ajustar o texto de entrada e a respectiva frase que deverá ser prevista pelo algoritmo; com isso, parte do treinamento é executado. A segunda parte do treinamento consiste na adaptação máscara em ajustar o texto em *tokens* (usando o WordPiece[12]), fazendo a permuta de alguns termos por máscaras (como "####") e os possíveis valores que podem substituí-las. Após o *corpus* ser ajustado, ele é submetido ao BERT para treinamento e geração do modelo adequado ao caso requisitado. De forma geral, esse ajuste fino ocorre em casos muito específicos, nos quais o texto pode ser de determinados domínios, como jurídico ou médico, onde podem ocorrer jargões e termos técnicos que não estão presentes na base geral de treinamento do modelo.

5.4.3 Avaliando os sumários

Até aqui, indicamos de maneira teórico-prática como funcionam os algoritmos de sumarização extrativa baseados em frequência de palavras e no modelo BERT. Agora, nosso objetivo é analisar como os sumários produzidos por esses algoritmos podem diferir entre si a partir de um texto-fonte com cinco sentenças.

[10] Para um melhor entendimento do espaço vetorial, sugere-se consultar o Capítulo Representação vetorial e semântica distribucional.

[11] https://paperswithcode.com/dataset/bookcorpus

[12] Wordpieces são partes de palavras, devendo não confundir-se com morfemas, já que não carregam nenhum significado. Wordpieces podem ser obtidas de maneira empírica e constituem um vocabulário induzido a partir de dados do texto. Para saber mais detalhes sobre esse conceito, confira o Capítulo Texto ou fala? e o Capítulo Sequência de caracteres e palavras. Neste capítulo utilizamos o pacote WordPiece proposto pela Google, e disponível em: https://huggingface.co/learn/nlp-course/chapter6/6?fw=pt.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

No Quadro 5.5 são apresentados os sumários extrativos gerados pelos métodos Frequência de Palavras e BERT. O texto-fonte é o mesmo do Quadro 5.1.

Quadro 5.5: Sumários gerados a partir dos métodos Luhn e BERT

Método	**Texto**	**Qtd. sentenças**	**Qtd. palavras**
Sumário 1 - Frequência de palavras	Maradona havia recebido alta no último dia 11, mas voltou a ser internado na sexta-feira e os boletins médicos não especificaram o que se passava com o ex-jogador – Cahe descartou pancreatite ou úlcera. "Maradona teve uma recaída na hepatite aguda. Agora está estável. Apesar de ter melhorado no domingo, deverá continuar internado", disse Cahe, em declarações ao jornal "La Nación".	2	60
Sumário 2 - BERT	O médico pessoal do argentino Diego Maradona, Alfredo Cahe, revelou nesta segunda-feira que uma recaída da hepatite aguda de que sofre foi o motivo da nova internação do ex-craque. Maradona, 46, desenvolveu uma hepatite tóxica por excesso de consumo de álcool, o que já o manteve internado durante 13 dias antes da primeira alta.	2	54
Sumário 3 - BERT - 3 sentenças	O médico pessoal do argentino Diego Maradona, Alfredo Cahe, revelou nesta segunda-feira que uma recaída da hepatite aguda de que sofre foi o motivo da nova internação do ex-craque. Maradona havia recebido alta no último dia 11, mas voltou a ser internado na sexta-feira e os boletins médicos não especificaram o que se passava com o ex-jogador – Cahe descartou pancreatite ou úlcera. Maradona teve uma recaída na hepatite aguda.	3	69

De maneira geral, observa-se que a taxa de compressão pode impactar diretamente a informatividade do sumário. Todos os três sumários gerados apresentam informações necessárias à compreensão do evento (no caso, o problema de saúde de Maradona e a recaída de seu quadro clínico). Apesar disso, destaca-se que, apesar de a informatividade ser um conceito, por vezes, subjetivo, ele está associado ao objetivo do sumário (se ele deve ser informativo, indicativo ou avaliativo, por exemplo).

Conforme já mencionado, é importante planejar a organização do conteúdo selecionado para que o texto final faça sentido e satisfaça a necessidade do usuário. Por exemplo, no Sumário 1, embora a segunda sentença tenha maior pontuação que a primeira, se ela iniciar o texto final, o leitor perderá o contexto de que Cahe é o médico de Maradona.

Num cenário em que comparamos os sumários com o texto-fonte, notamos que todos selecionaram sentenças do início e do meio do conteúdo original. Possivelmente essa escolha se baseou no fato de as sentenças que finalizam textos jornalísticos apresentarem

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

detalhamentos sobre o tópico principal, enquanto as primeiras sentenças apresentarem sumariamente o assunto.

Além de aspectos linguístico-textuais que devem ser observados nos sumários, como visto, há parâmetros estatísticos que podem ser utilizados durante a avaliação. Para tanto, os três sumários foram avaliados usando a medida ROUGE, mais especificamente ROUGE-1, ROUGE-2 e ROUGE-L. Os valores obtidos para cada sumário são apresentados na Tabela 5.3.

Tabela 5.3: Resultado da avaliação pelo método ROUGE

Algoritmos	Rouge 1			Rouge 2			Rouge L		
	R	P	F	R	P	F	R	P	F
Frequência de Palavras	0.447	1.00	0.618	0.443	1.00	0.614	0.447	1.00	0.618
Bert - 20%	0.394	1.00	0.566	0.390	1.00	0.561	0.394	1.00	0.566
Bert - 3 sentenças	**0.519**	1.00	**0.683**	**0.516**	1.00	**0.681**	**0.619**	1.00	**0.683**

Uma característica do texto jornalístico é que as informações localizadas no início do texto expressam o fato principal de uma notícia (Canavilhas, 2012), por isso, são selecionadas para compor o sumário. Tal consideração pode ser diferente para outros gêneros textuais. O resultado da métrica ROUGE, na comparação de sumários automáticos extrativos com seu texto-fonte (referência), justifica a precisão ser 1 para todos os casos. Se a comparação fosse em relação a um sumário humano com alguma reescrita, o resultado seria totalmente diferente.

Observa-se que os valores da medida-f não são muito diferentes entre si. Cada um dos sumários traz um contexto que pode ser importante para um tipo de leitor, mas não para outro: o Sumário 1 recupera a opinião do médico e o Sumário 2 apresenta o breve histórico da saúde de Maradona. Ao reduzir a taxa de compressão, o Sumário 3 foi mais abrangente que os demais, resultando em uma melhor avaliação em termos de medida-f.

5.5 Aplicação em Sumarização Abstrativa

A tarefa de sumarização abstrativa visa ir além da extração e concatenação de sentenças ou trechos do texto-fonte. Seu objetivo é gerar novas sentenças que expressem o conteúdo principal do texto, muitas vezes reescrevendo e reorganizando informações, prevendo maior fluidez e concisão. Embora mais flexível e com potencial para gerar sumários de qualidade superior, essa abordagem também apresenta maior complexidade e riscos, como a possibilidade de geração de informações incorretas (alucinações) ou desconectadas do conteúdo original.

Assim como na sumarização extrativa, o processo pode ser dividido em etapas. Tradicionalmente, considera-se o seguinte *pipeline*: extração de informações, seleção de conteúdo e geração de sentenças (Lin; Ng, 2019). A seguir, apresentamos brevemente dois grupos de abordagens para construção de sumarizadores abstrativos: as baseadas em estrutura e/ou semântica, e aquelas baseadas em modelos neurais, especialmente na arquitetura Transformer.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

5.5.1 Algoritmos baseados em estrutura sintática e/ou semântica

Os algoritmos desta categoria buscam representar as informações mais importantes do texto por meio de estruturas como árvores de dependência, grafos semânticos ou ontologias, para, então, gerar sentenças-resumo. Em geral, esses métodos combinam extração de estruturas linguísticas e alguma técnica de geração textual (Gupta; Gupta, 2019).

Como exemplo, destacam-se abordagens baseadas em árvores de dependência, que buscam representar a estrutura sintática das sentenças (veja Capítulo A ordem e a função das palavras em uma sentença). Normalmente estes métodos partem de uma sumarização extrativa, com a seleção das sentenças consideradas mais importantes, então sentenças similares são identificadas a partir de um analisador e posteriormente são construídas árvores de dependência para estas sentenças, buscando fundir as árvores de sentenças semelhantes. Por fim, ocorre um processo de linearização da árvore resultante, convertendo-a para texto.

Árvores de dependência, porém, buscam capturar estritamente a estrutura sintática das sentenças, não sendo possível captar, por si só, nuances de contexto. Por exemplo, considere o seguinte exemplo: "A menina postou uma foto com o celular" apresentado no Capítulo A ordem e a função das palavras em uma sentença. Esta sentença pode ser interpretada de, ao menos, duas maneiras: (i) a menina postou uma foto a partir do seu celular, ou (ii) ela postou uma foto em que ela aparece utilizando o celular. A sintaxe da sentença, por si só, além de poder ser ambígua, pode não expressar exatamente o que ocorreu, sendo necessário mais informações sobre o contexto.

Por outro ladoNeste sentido, abordagens baseadas em semântica consistem em obter uma representação semântica do texto e, a partir disso, gerar um sumário. Um exemplo de representação semântica são os grafos AMR (*Abstract Meaning Representation*) (Banarescu et al., 2013), que modelam o conteúdo do texto por meio de entidades e suas relações. Na Figura 5.4 é apresentado um exemplo de um grafo AMR construído para uma determinada sentença. Nela, percebe-se que o verbo "querer", por exemplo, possui dois argumentos, representados pelas arestas ARG0 (quem quer?) e ARG1 (o que quer?). Por não focar nos itens lexicais da sentença, no modelo AMR o verbo "desejar", por exemplo, não precisa necessariamente ser mantido no grafo, desde que o significado da sentença seja preservado. Dessa forma, uma mesma sentença pode gerar diferentes representações semânticas em AMRs, assim como uma mesma representação AMR pode ser transformada em sentenças distintas.

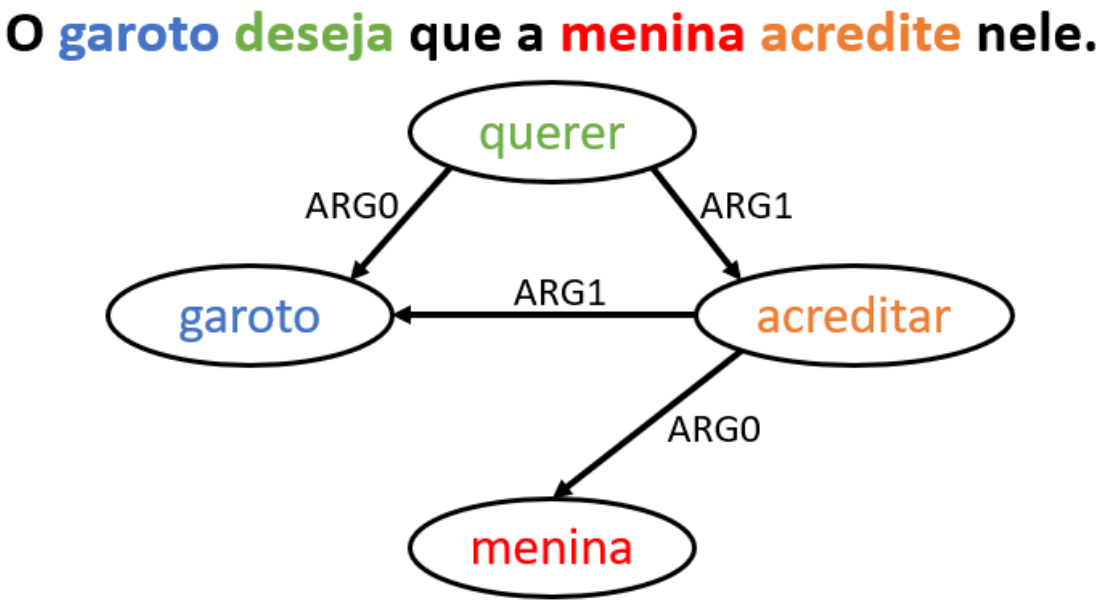
Em (Dohare et al., 2018), é apresentado um sistema de sumarização abstrativa baseado em grafos AMR. A abordagem dos autores se inspira no modo com que os humanos resumem textos. A partir da AMR do texto de entrada, são identificadas as entidades e eventos mais relevantes, seguidos pelas relações-chave entre essas entidades. Posteriormente, o sistema recupera informações de contexto relacionadas às relações selecionadas. O subgrafo resultante representa uma versão condensada da AMR original, e pode ser interpretado como uma sumarização semântica do texto, a partir da qual se gera o sumário final.

Cabe destacar que há um potencial interessante na combinação dessas abordagens com modelos de linguagem. Os Transformers, por exemplo, utilizam grafos semânticos como entrada para modelos gerativos (Hua et al., 2023; Yasunaga et al., 2022) ou como ferramenta para validação (Liu et al., 2024; Qiu et al., 2024; Roy et al., 2023) dos sumários produzidos. Essa integração, no entanto, ainda é uma área em desenvolvimento.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Figura 5.4: Um exemplo de sentença e sua AMR correspondente



Fonte: (Paiola, 2022)

5.5.2 Algoritmos baseados em Transformers

Com o avanço da arquitetura Transformer (Vaswani et al., 2017b), os métodos de sumarização abstrativa passaram a ser dominados por modelos baseados em aprendizado profundo e, mais recentemente, por modelos pré-treinados de larga escala. A aplicação desses modelos permitiu representar melhor a semântica dos textos, possibilitando sumários mais coesos e coerentes.

Um dos modelos nessa linha é o T5 (*Text-To-Text Transfer Transformer*) (Raffel et al., 2020), cujo diferencial está em tratar toda tarefa de PLN como um problema de transformação de texto para texto. Na tarefa de sumarização, o modelo recebe como entrada o texto-fonte precedido de um prefixo que indica a tarefa (por exemplo, "summarize:", como pode ser visto na Figura 5.5), e gera como saída o sumário correspondente. Adaptando esse modelo ao português brasileiro, foi proposto o PTT5 (Carmo et al., 2020b), uma versão do T5 pré-treinada em um grande *corpus* em português, o BrWac (Wagner Filho et al., 2018), sendo posteriormente ajustada para diferentes tarefas, incluindo a sumarização automática, como o PTT5-Summ (Paiola et al., 2022).

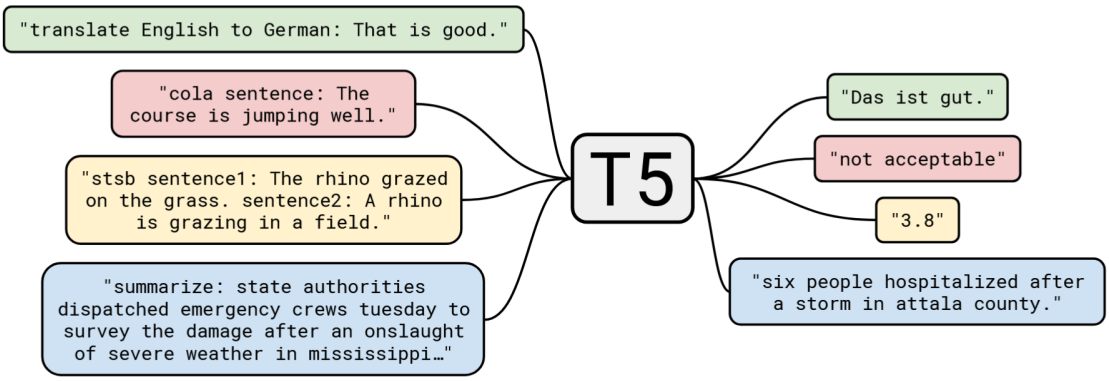
Modelos como o PTT5 permitem o uso de diferentes tamanhos de arquitetura (*small*, *base* e *large*), sendo possível realizar o *fine-tuning* específico para a tarefa de sumarização, com custos computacionais variados. O desempenho desses modelos está intimamente ligado ao conjunto de dados utilizado para o ajuste fino. Modelos ajustados em bases com sumários muito curtos, por exemplo, tendem a produzir resumos menores, não podendo ser aplicados para os casos em que necessitemos de sumários mais longos. Do mesmo modo, se o modelo foi treinado para sumarizar notícias, dificilmente ele conseguirá sumarizar adequadamente outros tipos de texto.

Além dos modelos especialistas treinados especificamente para a tarefa de sumarização, também vêm sendo explorados os LLMs, que são mais generalistas, capazes de realizar

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Figura 5.5: Um diagrama do *framework* de texto para texto proposto para o modelo T5



Fonte: (Raffel et al., 2020)

diversas funções, incluindo a própria sumarização. Tais modelos são geralmente treinados em múltiplas tarefas e idiomas, podendo ser aplicados diretamente à tarefa de sumarização mesmo sem ajuste fino, apenas com o uso de *prompts* específicos. Por exemplo, ao fornecer um comando como "Escreva um resumo em português com no máximo 150 palavras", é possível obter um sumário coerente e informativo, ainda que genérico, do ponto de vista funcional.

De modo geral, os LLMs podem conseguir resultados melhores, inclusive possibilitando a personalização de resultados mais facilmente, apenas passando instruções via *prompt*. Por outro lado, há uma série de questões que podem ser consideradas, como o custo de utilizar um LLM, seja de processamento, ao rodar em *hardware* local; ou de utilizar algum modelo fechado via API; até a questão de sigilo de dados, considerando o uso de modelos fechados. Além disso, considerando estudos como o de Sarmento; Oliveira (2024), podemos perceber que, apesar de modelos maiores, como o próprio GPT, realmente mostrarem um desempenho consideravelmente maior que modelos pequenos como PTT5, o mesmo pode não acontecer com LLMs na faixa de 8 a 9 bilhões de parâmetros, que chegam a apresentar resultados próximos ou até inferiores aos modelos especialistas de menor tamanho. Nesse trabalho, o LLaMa 3.1 (Grattafiori et al., 2024) de 8 bilhões de parâmetros obteve uma ROUGE-L de 33,8% na base de dados CSTNews, enquanto o PTT5 *Small*, com *fine-tuning* realizado pelos autores, alcançou 39,3%. O fato é que este último modelo contém 60 milhões de parâmetros, sendo mais de 100 vezes menor que o LLaMa.

Isso não quer dizer que não se deve considerar a aplicação de LLMs para sumarização. Pelo contrário, esse ponto reforça que eles não simplesmente substituem os modelos de linguagem menores ou até outras técnicas de sumarização, mas são importantes para avaliar o contexto de aplicação, com todas as suas particularidades.

Esses pontos evidenciam um dilema atual: modelos especialistas, como o PTT5, quando bem ajustados, tendem a oferecer um bom equilíbrio entre custo e desempenho, especialmente em cenários com restrições computacionais. Por outro lado, os LLMs oferecem maior flexibilidade e qualidade textual, mas com custos mais elevados e maior sensibilidade a variações no *prompt* podendo gerar grandes mudanças no resultado final (Zhang et al., 2025).

Diante desses novos modelos, novos desafios também surgem. Entre eles, destacam-se:

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

- a necessidade de conjuntos de dados de qualidade para o ajuste fino e avaliação (veja Capítulo Conjunto de dados, *dataset* e *corpus*);
- o risco de os modelos incorporarem vieses ou inserirem informações incorretas (veja Capítulo Questões éticas em IA e PLN);
- e o custo envolvido, que nem sempre é viável em contextos práticos.

Ainda assim, a aplicação de Transformers na SA abstrativa continua sendo um campo promissor e ativo de pesquisa, com potencial de aplicação real em sistemas voltados à leitura assistida, curadoria de notícias, educação e outras áreas que demandam a síntese de grandes volumes de informação textual.

5.5.3 Sumarização abstrativa na prática

Para ilustrar as possibilidades e desafios da sumarização abstrativa, preparamos exemplos de geração de sumários usando modelos de linguagem, assim como diferentes avaliações desses sumários. Os códigos que utilizamos para esses experimentos estão disponíveis[13].

5.5.3.1 Sobre o experimento

Os experimentos foram realizados com quatro modelos: **PTT5-Summ** (especialista em sumarização em português) e três LLMs generalistas (**Qwen3-8B**, **Sabiá-7B** e **Chatbode-7B**). Cada modelo recebeu textos jornalísticos como entrada, sendo solicitado a gerar um sumário abstrativo – ou seja, reescrevendo livremente com o objetivo de condensar a informação central.

Para a avaliação, utilizamos três perspectivas complementares:

- **Métricas automáticas**: ROUGE-1, ROUGE-2, ROUGE-L e BERTScore, que medem a sobreposição ou a similaridade semântica entre o sumário gerado e o sumário de referência.
- **Avaliação por LLM**: GPT-4o foi usado como "avaliador automático", atribuindo nota de 0 a 10 com justificativa, a partir de um *prompt* simples e do texto-fonte.

Vale ressaltar que, como já discutido neste capítulo, avaliar sumarização abstrativa é um desafio: sumários diferentes podem ser igualmente válidos, e métricas automáticas nem sempre captam nuances de qualidade textual, fidelidade factual ou concisão.

5.5.3.2 Gerando e avaliando sumários

A geração e avaliação dos sumários pode ser feita em poucas linhas, usando bibliotecas como `transformers`, `evaluate` e `bert_score`. Veja exemplos para dois modelos, o PTT5-Summ e o Qwen3-8B:

```
# Sumarização com PTT5-Summ
from transformers import T5Tokenizer, T5ForConditionalGeneration
import torch

# Carrega componentes do modelo pré-treinado para português
tokenizer = T5Tokenizer.from_pretrained('unicamp-dl/ptt5-base-portuguese-vocab')
```

[13] Jupyter notebook para Sumarização Abstrativa com Transformers disponível em: https://colab.research.google.com/drive/1cWhEWssFanREV94R2fc6N0IOlgkgeATY?usp=sharing.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

```
model = T5ForConditionalGeneration.from_pretrained('recogna-nlp/ptt5-base-summ')

# Prepara o texto para entrada no modelo
input_ids = tokenizer.encode(texto_fonte, return_tensors='pt')

# Gera o resumo com limite de tokens
summary_ids = model.generate(input_ids, max_length=256)

# Converte tokens de volta para texto e remove caracteres especiais
resumo = tokenizer.decode(summary_ids[0], skip_special_tokens=True)
```

Explicação passo a passo do que fizemos no código acima:

1° **Inicialização:** Carregamos o tokenizador e modelo T5 especializado em sumarização em português.

2° **Tokenização:** Convertemos o texto-fonte em representação numérica (`input_ids`).

3° **Geração:** O método `generate` cria o resumo usando arquitetura *encoder-decoder*.

4° **Decodificação:** Transformamos os *tokens* de saída em texto legível.

Agora, vamos ao código do LLM (`Qwen3-8B`):

```
# Sumarização com um LLM (Qwen3-8B)
from transformers import AutoTokenizer, AutoModelForCausalLM

# Carrega os componentes do LLM
tokenizer = AutoTokenizer.from_pretrained("Qwen/Qwen3-8B")
model = AutoModelForCausalLM.from_pretrained("Qwen/Qwen3-8B")

# Formata o prompt instrucional
prompt = f"Resuma o seguinte texto:\n{texto_fonte}"

# Processa entrada e gera saída
inputs = tokenizer(prompt, return_tensors="pt")
output = model.generate(**inputs, max_new_tokens=256)

# Decodifica a saída
resumo = tokenizer.decode(output[0], skip_special_tokens=True)
```

Vamos analisar passo a passo nosso código:

1° **Configuração:** Usamos o carregamento automático (`AutoTokenizer/AutoModel`) para carregar o *tokenizer* e o modelo de linguagem.

2° **Prompt:** Utilizamos instrução explícita para contextualizar a tarefa.

3° **Geração:** O modelo produz texto, *token* por *token*, em modo auto-regressivo.

Para comparar a qualidade dos resumos produzidos, utilizamos três tipos distintos de avaliação: (i) métricas automáticas baseadas em sobreposição, (ii) similaridade semântica e (iii) julgamento por um LLM. As avaliações também se tornam simples de realizar aproveitando as bibliotecas já existentes do Python.

Nossa primeira métrica será a ROUGE, ela tem como objetivo comparar a sobreposição de n-gramas entre o resumo gerado e um sumário de referência.

```
# Avaliação com ROUGE
import evaluate

# Carrega nosso método de avaliação
rouge = evaluate.load("rouge")
```

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

```
# Calcula sobreposição de unigramas, bigramas e subsequências mais longas entre resumo
# gerado e referência
results = rouge.compute(predictions=[resumo], references=[resumo_referencia])

# Visualizando o resultado com unigramas e bigramas
print("ROUGE-1:", results["rouge1"], "ROUGE-2:", results["rouge2"], "ROUGE-L:", results["rougeL"])
```

Agora, iremos aplicar a métrica BERTScore que vai além da sobreposição textual, calculando similaridade semântica entre os textos com base em representações vetoriais derivadas de modelos BERT.

```
# Avaliação com BERTScore
from bert_score import score as bert_score

# Compara vetores de embeddings do BERT, capturando equivalência semântica mesmo com
# palavras diferentes.
P, R, F1 = bert_score([resumo], [resumo_referencia], lang="pt", rescale_with_baseline=True)

# Foco no F1, que combina precisão e cobertura semântica.
print("BERTScore (F1):", F1.mean().item())
```

O BERTScore é uma métrica de avaliação de geração de texto que supera limitações das abordagens tradicionais baseadas em sobreposição lexical. Diferentemente da ROUGE (que conta n-gramas idênticos), ela opera no nível semântico usando representações vetoriais de linguagem. Assim, essa métrica é especialmente útil quando o sumário válido é muito diferente em termos de palavras, mas semelhante em conteúdo.

Por fim, empregamos o modelo GPT-4o como avaliador automático. Esse tipo de avaliação, conhecida como *LLM-as-a-judge*, busca aproximar o julgamento humano ao usar um LLM para avaliar outros modelos.

```
# Avaliação com LLM-as-judge
from openai import OpenAI
client = OpenAI()

# Um prompt simples e estruturado foi fornecido ao GPT-4o, solicitando uma nota de 0 a 10
# com justificativa, com base em critérios como coerência, fidelidade factual, completude
# e fluência.
messages = [
    {"role": "system", "content": "Você é um avaliador imparcial de resumos."},
    {"role": "user", "content": f"""Texto original:
    {prompt_text}

    Resumo gerado:
    {summary}

    Dê uma nota de 0 a 10 para este resumo com base em coerência, \
    fidelidade, completude e fluência. Justifique brevemente."""}
]


completion = client.chat.completions.create(
    model="gpt-4o",
    messages=messages
)

# A resposta inclui tanto a nota quanto uma explicação, o que permite compreender melhor
```

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

```
# os pontos fortes e fracos de cada sumário.
print(completion.choices[0].message.content)
```

A Tabela 5.4 mostra os principais resultados dos experimentos, unindo as três perspectivas de avaliação para dois textos jornalísticos.

Tabela 5.4: Resultado das avaliações dos sumários abstrativos

Modelo	Doc	ROUGE-1	ROUGE-2	BERTScore	Nota GPT-4o	Comentário GPT-4o (resumido)
internlm-chatbode-7b	1	0.133	0.019	0.138	8	Boa coerência, mas poderia ser mais completo.
ptt5-summ	1	0.222	0.093	0.271	3	Resumo superficial e incompleto.
qwen3-8b	1	0.160	0.016	0.136	9	Coerente e fiel, falta mais detalhes.
sabia-7b	1	0.228	0.116	0.054	5	Algumas informações corretas, mas erros e omissões.
internlm-chatbode-7b	2	0.442	0.174	0.468	6	Nome incorreto do presidente do BC; faltam detalhes.
ptt5-summ	2	0.231	0.064	0.309	5	Visão geral, mas faltam detalhes importantes.
qwen3-8b	2	0.413	0.210	0.487	9	Coerente e fiel, faltam detalhes de contexto.
sabia-7b	2	0.265	0.053	0.274	6	Boa introdução, mas incompleto e sem conclusão.

A seguir são apresentados o segundo documento, seu sumário de referência e os sumários candidatos gerados por cada modelo. Os resultados para o primeiro documento estão no Apêndice Apêndice B.

Quadro 5.6: Sumários gerados por cada modelo

Modelo	**Texto**
Documento2 Tendência dos juros no Brasil	"A tendência de queda da taxa de juros no Brasil é real, é visível", disse Meirelles, que participou na capital americana de uma série de reuniões e encontros com banqueiros e investidores que aconteceram paralelamente às reuniões do Fundo Monetário Internacional (FMI) e do Banco Mundial (Bird) no fim de semana. Para o presidente do BC, a atual política econômica do governo e a manutenção da taxa de inflação dentro da meta são fatores que garantem queda na taxa de juros a longo prazo. ... O presidente do Banco Central disse ainda que a crise política brasileira não foi um assunto de interesse prioritário dos investidores que encontrou no fim de semana.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Sumário de referência	Henrique Meirelles afirmou que a tendência de queda dos juros no Brasil é real e sustentada por políticas econômicas e controle da inflação. Ele destacou a importância de uma visão de longo prazo e disse que o trabalho do Banco Central garantirá uma redução gradual dos juros. Meirelles também celebrou o reconhecimento internacional do BC com o prêmio da revista The Banker, reforçando a credibilidade da instituição.
PTT5-Summ:	Meirelles diz que queda de juros no Brasil é real e é visível. Presidente do BC participou de reunião com banqueiros e investidores no fim de semana.
Qwen3-8B:	O presidente do Banco Central do Brasil, Henrique Meirelles, afirmou que a tendência de queda da taxa de juros no país é real e é resultado da política econômica e do controle da inflação. Ele destacou a importância do trabalho do BC, o reconhecimento internacional pela revista The Banker e ressaltou a necessidade de manter uma visão de longo prazo.
InternLM-Chatbode-7B:	O presidente do Banco Central do Brasil, Ilan Goldfajn, afirmou que a tendência de queda da taxa de juros no país é real e visível. Ele atribuiu essa tendência à política econômica do governo e à manutenção da taxa de inflação dentro da meta. Goldfajn também destacou a importância de considerar o médio e longo prazo, e que o trabalho do Banco Central para conter a inflação dentro da meta levará a uma queda gradual da taxa de juros. O Banco Central do Brasil recebeu o prêmio "Banco Central do ano" pela revista The Banker, que reconheceu o trabalho do instituição. Goldfajn também afirmou que a crise política brasileira não foi um assunto de interesse prioritário dos investidores que encontrou no fim de semana.
Sabiá-7B:	A tendência de queda da taxa de juros no Brasil é real, é visível, disse Meirelles, que participou na capital americana de uma série de reuniões e encontros com banqueiros e investidores que aconteceram paralelamente às reuniões do Fundo Monetário Internacional (FMI) e do Banco Mundial (Bird) no fim de semana. Para o presidente do BC, a atual política econômica do governo e a manutenção da taxa de inflação dentro da meta são fatores que garantem queda na taxa de juros a longo prazo. "Mas é importante que nós não olhemos para isso apenas no curto prazo. Temos que olhar no médio e longo prazos", disse Meirelles. Para ele, o trabalho que o Banco Central tem feito para conter a inflação dentro da meta vai gerar queda gradual da taxa de juros. BC do ano Neste domingo, Meirelles participou da cerimônia de entrega do (texto foi truncado por atingir o limite de *tokens* da resposta, ajustado em 256 *tokens*)

Observando os exemplos acima, fica claro que cada modelo tem seu "estilo" ao sumarizar. O PTT5-Summ, por exemplo, gera um texto sintético, buscando ser objetivo e breve, como foi treinado para fazer. Isso faz com que, ao ser avaliado por um LLM (neste caso, o GPT-4o), que recebe apenas o texto-fonte e espera uma cobertura completa das informações, sua nota seja mais baixa. Da mesma forma, métricas como ROUGE e BERTScore penalizam sumários muito curtos porque eles deixam de cobrir n-gramas e conteúdos do sumário de referência.

O Qwen3-8B e o InternLM-Chatbode-7B produziram textos mais próximos em extensão e detalhamento do sumário de referência, sendo melhor avaliados tanto por métricas automáticas quanto pelo GPT-4o – ainda que, no caso do InternLM, tenha ocorrido um erro factual ao trocar o nome do presidente do Banco Central. O nome citado corresponde ao presidente do Banco Central no período de 2016 à 2019, o que sugere uma possível

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

alucinação causada por viés adquirido nos dados de treinamento, provavelmente mais concentrados nesse período. O mesmo modelo também cometeu um erro de concordância no trecho "do instituição". O Sabiá-7B, por sua vez, apenas copia o texto, atingindode forma que acaba rapidamente o limite de *tokens* e terminando o texto abruptamente. Embora tenha recebido as piores avaliações nesse caso, é importante perceber que, apesar de realmente ele ter obtido as piores avaliações para este texto, aessa completa falta de sumarização não foi bem contemplada por nenhuma das medidas. Nesse contexto, a avaliação com LLMs pode ser uma alternativa mais adequada, ao incorporar aspectos como concisão como fator de avaliação, a partir de ajustes no *prompt*.

5.6 A Sumarização automática e a língua portuguesa

Ao analisar as publicações registradas em diferentes repositórios, percebe-se que a SA aplicada ao português brasileiro acompanhou as tendências metodológicas e de aplicações da literatura internacional.

Com relação a alguns **recursos** desenvolvidos, destacamos o TeMário (Pardo; Rino, 2003) por ser um *corpus* de textos jornalísticos acompanhados de seus respectivos sumários, e que inspirou metodologicamente outros *corpora* que foram elaborados para fins de sumarização, como o CSTNews e o RecognaSumm (Paiola et al., 2024). Sendo o CSTNews tido como padrão ouro por apresentar diferentes camadas de anotação linguística, além de sumários humanos que servem de referência para o processo automatizado, enquanto o RecognaSumm se destaca por possuir cerca de 135 mil amostras, abrindo a possibilidade de ajuste-fino de modelos de linguagem. Para apoiar pesquisas de SA de opiniões baseadas em aspectos, há o *corpus* OpiSums-PT[14] (López et al., 2015), que contém vários resumos extrativos e abstrativos de opiniões escritas em português brasileiro, nos quais cada resumo é derivado da análise de 10 opiniões sobre diferentes produtos.

Já com relação a algumas ferramentas, sublinhamos a importância da ferramenta CSTTool (Aleixo; Pardo, 2008), que utiliza o modelo discursivo *Cross-document Structure Theory* (Radev, 2000) [15] para solucionar desafios linguísticos da sumarização multidocumento. Outra ferramenta tida como precursora por muitos estudiosos da área é o Gistsumm (Balage Filho et al., 2007; Pardo, 2002), que gera sumários extrativos a partir da identificação do tópico principal dos textos, utilizando métodos superficiais (como palavras-chave).

Cabe destacar que nos últimos anos, com o avanço de modelos de língua e abordagens mais robustas em PLN, a SA passou a apresentar sumários potencialmente com mais qualidade linguística. Nesse sentido, foram encontrados trabalhos que utilizam LLMs (Barros, 2022; Paiola, 2022), apresentando respostas às questões ora não bem resolvidas por abordagens utilizadas no início da SA. Além disso, observaram-se estudos que aplicam a SA em outros domínios para além do jornalístico, como o jurídico (Feltrin et al., 2023) e códigos de programação (Pontes et al., 2022). Aplicações baseadas na dependência de língua e/ou domínio seriam bastante custosas do ponto de vista do PLN, o que pode, em partes, ser resolvido com os LLMs.

Muitos outros trabalhos foram e vêm sendo desenvolvidos em SA, já que se trata de uma área ainda em expansão. Nesse sentido, indicamos que conheçam os trabalhos capitaneados pelo Núcleo Interinstitucional de Linguística Computacional (NILC)[16], que deu suporte

[14] https://sites.icmc.usp.br/taspardo/sucinto/files/OpiSums-PT.zip
[15] Para saber mais detalhes sobre o modelo discursivo CST, confira o Capítulo Modelos discursivos.
[16] https://sites.google.com/view/nilc-usp/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

a muitas pesquisas[17] de diferentes níveis de formação (graduação e pós-graduação). Tais trabalhos utilizaram distintas abordagens que vigoravam como o estado da arte, como SA mono e multidocumento, mono e multilíngue, ora com abordagem mais linguística, ora com abordagem mais estatística. Outra indicação nossa é o Sistemas de Informação e Banco de Dados, da Universidade Federal de Campina Grande (SINBAD), que também abriga pesquisas recentes sobre SA. Por fim, outro grupo de destaque é o Grupo de pesquisa em sistemas inteligentes (GSI), da Universidade Tecnológica Federal do Paraná, que tem feito estudos em SA multidocumento e com modelos computacionais mais recentes, como o BERT.

5.7 Considerações finais

Neste capítulo, apresentamos um esboço sobre uma das áreas que tem evoluído no PLN, a SA. Como visto, o processo de automatização de sumários (extrativos ou abstrativos) não é uma tarefa simples. As dificuldades tendem a ser decorrentes de informações faltantes nos textos. Essa falta de dados exige ter-se algum conhecimento prévio do que é informado, de modo que se possa fazer associação com outros fatos relevantes, o que tende a gerar um melhor entendimento do sumário.

Ainda, apresentamos duas abordagens distintas para sumarização automática: a extrativa e a abstrativa. Inicialmente, detalhamos dois algoritmos voltados à sumarização extrativa. O primeiro, de abordagem mais simples, baseia-se na frequência de palavras para selecionar sentenças representativas do texto. O segundo utiliza o modelo BERT e aproveita o poder computacional de redes neurais profundas para realizar tarefas de forma contextualizada e precisa.

Em seguida, apresentamos a sumarização abstrativa, que avança em direção ao uso de modelos generativos capazes de produzir sentenças novas, com maior fluidez e coesão. Porém, essa flexibilidade aumenta os riscos de alucinação e de omissão de informações essenciais. Por isso, a avaliação de sumários abstrativos deve ir além de métricas tradicionais, como ROUGE, e considerar critérios como fidelidade semântica, completude e fluência.

Uma das tarefas que não pode ser negligenciada no processo de sumarização é a avaliação do produto gerado. Assim, pontuamos sobre os métodos de avaliação, com destaque para a ROUGE, além de uma abordagem mais linguística segundo a proposta da TAC/DUC. Cabe ressaltar que aqui também residem outros desafios à SA, já que aplicações de sumarização estão sendo implementadas em conteúdo gerado pelo usuário da Web (como avaliações de produtos e publicações em redes sociais). Nesse caso, concepções linguísticas sobre gênero e suporte textual deverão ser levadas em consideração para que os sumários gerados possam ser informativos, coerentes e coesos, para além de uma perspectiva normativista da língua.

Uma das motivações mais utilizadas para as pesquisas iniciais em SA era a falta de tempo que os usuários (especialmente na internet) tinham frente à quantidade de conteúdo que era constantemente disponibilizado. Atualmente, essa motivação continua viável, mas deve ser acolhida no interior da seguinte reflexão: material gerado por "usuários-autores" pode conter ainda mais enviesamento e incorrer em informações distorcidas ou falsas, podendo resultar em desinformação. Nesse caso, vislumbramos que a SA não poderá se afastar de pesquisas que abordem a língua de uma perspectiva funcional-descritiva, mas que também

[17] Indicamos acesso à página do projeto Sucinto, que abrigou diferentes pesquisas na área de SA ligadas ao NILC. Disponível em: https://sites.icmc.usp.br/taspardo/sucinto/index.html.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

dialoguem com soluções de identificação de conteúdo falso e desinformação.

Diante dessas ponderações, podemos concordar que não esgotamos o tópico de sumarização apenas neste capítulo, especialmente em um contexto em que há uma tendência nas pesquisas em PLN em utilizarem LLMs. Em decorrência da constante evolução da qualidade dos sumários, talvez torne-se necessário a criação de métodos mais apurados que permitam identificar diferenças de estilos de escrita e a identificação de conteúdo relevante. Esse processo de evolução das tecnologias permite imaginar os avanços que podem ocorrer por conta de trabalhos com redes neurais generativas. Fato é que todo esse processo demandará melhorias ou desenvolvimento de novos algoritmos, em especial dos voltados para a língua portuguesa.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Capítulo 6

Complexidade Textual e suas Tarefas Relacionadas

Sidney Evaldo Leal
Felipe Ribas Serras
Marcelo Finger
Sandra Maria Aluísio

Publicado em: 13/03/2024
Atualizado em: 16/04/2026

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

6.1 Introdução

Ratificando o redigido na subdivisão da obra acima epigrafada, os originadores da composição a seguir com afinco intentam discorrer sobre o significativo tópico da complexidade textual.

Humm... vamos recomeçar...

Neste capítulo, vamos tentar explicar o tema complexidade textual.

Como acabamos de ver, existem várias formas de dizer a mesma coisa, com graus de complexidade bem diferentes. O tema complexidade textual é largamente tratado em estudos do discurso, na educação, na psicolinguística, na linguística cognitiva, na fonoaudiologia, e no processamento de linguagens naturais (PLN). Aqui apresentaremos conceitos e soluções do ponto de vista do PLN.

Mas antes de tudo é importante dizer que a complexidade é sempre relativa, toda vez que falarmos e ouvirmos falar de complexidade, precisamos perguntar: complexo para quem?

Nesse ponto Dubay (2007) resume de forma certeira a definição: "Inteligibilidade é a facilidade de leitura de um texto criada pela escolha de conteúdo, estilo, estruturação e organização que atende ao conhecimento prévio, habilidade de leitura, interesse e motivação da audiência."

E já que acima aparece o termo **inteligibilidade**, vamos conversar um pouco sobre isso. Inteligibilidade textual vem da tradução do inglês *text readability* e às vezes também é traduzida como leiturabilidade; ambos os termos são bem representativos. Nós optamos por usar inteligibilidade pela relação com entender e dominar a língua (ser letrado) versus a habilidade de decodificar o sistema de escrita (ser alfabetizado). O mais importante é evitar o termo legibilidade, pois este está ligado com o que torna um texto fácil de ser lido (e não necessariamente entendido) como, por exemplo o tamanho e tipo da fonte, cor, estruturação em itens, etc. A inteligibilidade é inversamente correlacionada com a complexidade, isto é, quanto mais complexo um texto for, menos inteligível ele será, para o público alvo do texto.

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

6.2 Complexo para quem?

A complexidade só existe a partir do ponto de vista específico de quem está lendo, não é possível estudá-la sem o sujeito envolvido no processo da leitura. Mesmo em níveis de letramento próximos, pessoas diferentes podem achar o mesmo texto complexo e simples. Isso varia de acordo com o conhecimento de mundo adquirido e armazenado, a experiência, a habilidade de leitura e o grau de interesse no texto (Dubay, 2007). O Indicador de Alfabetismo Funcional (INAF)[1] é um ótimo retrato geral dos potenciais leitores adultos do Brasil (IPM, 2018). O levantamento é feito em média a cada dois anos, desde 2001 e classifica a população nos seguintes níveis de letramento:

- **Analfabeto**: Não consegue ler;
- **Rudimentar**: Localiza informações explícitas e literais;
- **Elementar**: Realiza pequenas inferências em textos de tamanho médio;
- **Intermediário**: Consegue interpretar textos e confrontar a moral da história com sua própria opinião ou senso comum;
- **Proficiente**: Interpreta e elabora textos de maior complexidade sem dificuldades.

É importante frisar que essa classificação em cinco níveis pode ser considerada relativamente arbitrária e agrupa internamente diversos níveis de letramento. O próprio INAF até 2011 utilizava apenas quatro níveis (Analfabeto, Rudimentar, Básico e Pleno). Mais um nível foi adicionado ao identificar que, após as ações do governo de combate ao analfabetismo, a maioria das pessoas subiu para o nível básico, porém os níveis superiores permaneceram estáveis. De 2001 a 2018, o nível analfabeto caiu de 12% para 8%, enquanto o nível proficiente se manteve em 12%.

Um ponto de atenção é que apesar de ter alfabetismo explícito no nome, o INAF avalia o nível de letramento (ou literacia) da população. Alfabetização está relacionada ao processo mecânico de reconhecer os grafemas, ligando-os aos fonemas, enquanto letramento é o uso social desse processo. Conforme Soares (1996): "Letramento é o resultado da ação de ensinar ou de aprender a ler e escrever: o estado ou a condição que adquire um grupo social ou um indivíduo como consequência de ter-se apropriado da escrita."

Outro conceito interessante ligado à habilidade de leitura versus motivação é o estado de fluxo (Csikszentmihalyi, 2008). Aplicando ao contexto da leitura, se o texto for demasiado simples e a habilidade do leitor for alta, a experiência vai se tornar enfadonha. Por outro lado, se a habilidade do leitor for pequena demais para o nível de complexidade ou desafio apresentado, o esforço exigido vai ser bastante desmotivador. O estado de fluxo seria o casamento do nível de dificuldade adequado para o nível de proficiência do leitor.

Por que não ensinar a ler em vez de simplificar um texto?

Esta é uma crítica recorrente e muito importante, logo é bom abordá-la aqui. Concordamos plenamente que é sempre melhor ensinar a ler do que simplificar um dado texto original para que ele seja acessível a uma pessoa com dificuldade de entendimento. Dito isso, são citadas a seguir duas grandes exceções para se utilizar a simplificação:

[1] https://alfabetismofuncional.org.br/

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

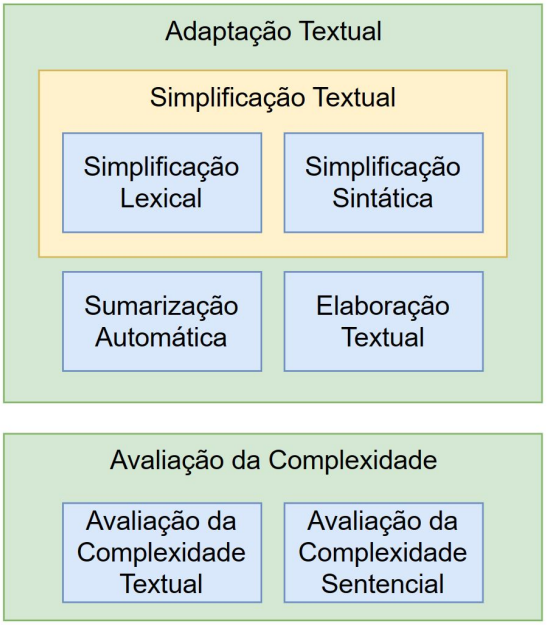
1. **Tempo versus Acesso.** Ensinar a ler exige tempo, enquanto simplificar pode permitir o acesso à informação no momento presente. Isso é uma verdade para a população adulta que possui dificuldades na leitura e, por diversos motivos, menos tempo para investir na própria educação. Além disso, para obter um resultado abrangente o suficiente, o investimento necessário na educação precisa partir do governo. Iniciativas isoladas conseguem bons resultados, mas quantitativamente o acesso à informação é maior simplificando os conteúdos publicados.

2. **Nível ideal de complexidade.** Conforme mencionado no tópico anterior, para um estudante em processo de aprendizagem, ser exposto a um texto demasiadamente difícil pode trazer mais prejuízos do que benefícios. A evolução do processo de ensino-aprendizagem pode ser muito mais eficiente se os textos fornecidos aos estudantes apresentarem um nível de desafio adequado.

6.3 Tarefas de PLN relacionadas à complexidade textual

As principais tarefas de PLN diretamente relacionadas com a complexidade textual estão representadas na Figura 6.1 e resumidas a seguir.

Figura 6.1: Principais tarefas de PLN relacionadas à complexidade textual



6.3.1 Adaptação textual

A Adaptação Textual é uma área de pesquisa de grande importância dentro da área de PLN, geralmente conectada com práticas educacionais, mas também com aplicações bem diversas como, por exemplo, auxiliar na recuperação de informações biomédicas (Jonnalagadda; Gonzalez, 2010). Ela permite alterar o conteúdo de um texto sem mudar seu significado, na maior parte das vezes. Possui duas grandes abordagens: Simplificação e

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Elaboração Textual (Aluísio; Gasperin, 2010; Burstein, 2009; Hartmann; Aluísio, 2020; Mayer, 1980).

6.3.2 Simplificação textual

A Simplificação Textual consiste no processo de reduzir a complexidade de um texto, enquanto se preserva o conteúdo informativo e significado, tornando o texto mais fácil de ser compreendido por leitores humanos ou ser processado por programas (Carroll et al., 1998; Max, 2006; Shardlow, 2014; Siddharthan, 2006).

Os primeiros avanços na área de simplificação textual automática surgiram com a ideia de dividir sentenças longas em sentenças menores para melhorar os resultados dos analisadores sintáticos (Chandrasekar et al., 1996). Desde os trabalhos iniciais em simplificação textual, a área prosperou, avançando as pesquisas em cenários de aplicação, línguas e métodos.

Tipos de simplificação textual

Arfé et al. (2018) definem o objetivo da simplificação textual como a adaptação da complexidade do texto (ou *readability*, em inglês) para as habilidades de um determinado grupo de leitores e, desta forma, *readability measures* (ou medidas de complexidade textual) foram desenvolvidas para alinhar/escolher textos para leitores, pois essas medidas podem predizer o quão difícil um texto será para seus leitores.

Desta relação entre *readability* e simplificação textual, surgem as abordagens profunda e superficial para *readability*, culminando nas abordagens cognitiva, temática (ou topical) e linguística para simplificação textual, explicadas abaixo.

Segundo as autoras acima, a abordagem superficial para complexidade textual se baseia no tamanho das palavras e sua frequência e no tamanho das orações para predizer a complexidade literal dos textos, ou seja, a compreensão do significado estrito de uma única proposição. Enquanto que a abordagem profunda, baseada em *features* como a presença e densidade de marcadores discursivos e correferência no texto, consegue predizer coerência e compreensão no nível inferencial, isto é, a integração entre segmentos de um texto e entre o texto e o conhecimento prévio do leitor.

Abaixo são definidos os três tipos de complexidade – a cognitiva, a linguística (envolvendo os níveis lexical e sintático) e a temática – que levam a três abordagens para simplificação textual, de mesmo nome.

A complexidade cognitiva está relacionada com a capacidade limitada de um leitor de identificar e compreender a estrutura global e local de um texto. A estrutura global é responsável por organizar as informações (ou tópicos) de um texto. As estruturas de textos informativos (jornalísticos, por exemplo) são mais variadas do que textos narrativos, podendo ser uma da seguinte lista: descrição, sequência, comparação e contraste, problema-solução e causa-efeito e, inclusive, aparecer de forma não-exclusiva, dificultando o seu reconhecimento. Essa dificuldade pode impedir um leitor de responder o que é dito no texto e de fazer um resumo dele, por exemplo.

A outra dificuldade se dá no processamento local de um texto, realizado pelo leitor, para conectar sentenças e identificar as suas relações (de contraste, exemplificação, causa, resultado, finalidade, dentre outras). As soluções para essas duas dificuldades são dois conjuntos de simplificações, chamadas de cognitiva no nível global e no nível local. No nível local são usadas para:

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

1. Aumentar a coesão via conectivos para explicitamente mostrar a relação entre sentenças;

2. Utilizar correferência para conectar as ideias.

No nível global, temos simplificações para:

1. Facilitar a retenção de novo conhecimento aprendido do texto via organização do conteúdo textual, ajudando o leitor a identificar a estrutura do discurso pelo uso de marcadores discursivos (linguísticos e tipográficos);

2. Organizar fatos e ideias presentes no texto pelo uso de subtítulos/seções que resumem o conteúdo dos parágrafos.

A complexidade temática/conceitual está associada à falta de conhecimento de mundo necessário para entender alguns temas. Existem projetos como o Newsela[2] (Xu et al., 2015) que realizam a simplificação conceitual, simplificando os conceitos expressos no texto. Por exemplo, o projeto Newsela inclui elaborações no texto para tornar certos conceitos mais explícitos ou redundâncias para enfatizar partes importantes do texto. Além disso, as operações reduzem e omitem informações que não são adequadas para determinado público-alvo.

Quanto às simplificações linguísticas, temos a lexical e a sintática. A complexidade lexical está relacionada ao desconhecimento do significado de palavras e expressões. A complexidade sintática está relacionada à capacidade ou não de processar alguns tipos de estrutura de sentenças. Na área de PLN, as simplificações linguísticas foram mais exploradas e muitos métodos criados, para várias línguas. Elas são detalhadas nas próximas seções.

6.3.3 Simplificação lexical

A Simplificação Lexical é uma forma de simplificação por meio da substituição de palavras raras ou complexas por hipônimos, hiperônimos ou sinônimos, equivalentes e mais simples, deixando a leitura com compreensão mais fácil para pessoas com baixo letramento (Aluísio; Gasperin, 2010), falantes não nativos de uma dada língua (Paetzold; Specia, 2016c), disléxicos e afásicos (Carroll et al., 1998), dentre outros (Boito, 2014). Um exemplo de sentença simplificada lexicalmente pode ser visto no Quadro 6.1.

A simplificação lexical geralmente é realizada com o apoio de dicionários compilados e recursos como WordNet (Fellbaum, 1998; Miller, 1995), de grandes *corpora* como a Simple Wikipedia (em uma abordagem de *ensembles*), e também de outras abordagens mais recentes baseadas em *word embeddings* (Paetzold; Specia, 2016a) e redes neurais para *ranking* (Paetzold; Specia, 2017).

Quadro 6.1: Exemplo de sentença simplificada lexicalmente

Original	Se acentuada e prolongada, a **hipertermia** pode causar a morte do animal.

[2] https://newsela.com/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

Simplificada	Se acentuada e prolongada, a **febre** pode causar a morte do animal

6.3.4 Simplificação sintática

A análise sintática é o estudo da disposição das palavras em uma oração e é dividida em funções sintáticas (sujeito e predicado) e constituintes (sintagmas nominais, verbais, preposicionais, adjetivas e adverbiais) (Candido Junior, 2013).

A Simplificação Sintática consiste em dividir orações longas (como exemplificado no Quadro 6.2) ou alterar a estrutura sintática das orações, eliminando fenômenos sintáticos considerados complexos para a inteligibilidade e compreensão de uma classe de leitores. Ainda segundo Candido Junior (2013), alguns exemplos comuns de fenômenos sintáticos são: reordenação de componentes de uma oração para facilitar a compreensão da informação principal veiculada, mudança de voz passiva para ativa, resolução anafórica de pronomes relativos, reordenação de orações e divisão de orações.

Quadro 6.2: Exemplo de sentença simplificada sintaticamente

Original	O uso de forragem conservada, cujas formas mais comuns são: ensilagem e fenação, é uma solução para alimentar o rebanho.
Simplificada	O uso de forragem conservada é uma solução para alimentar o rebanho. As formas mais comuns para conservar forragens são: ensilagem e fenação.

Embora haja um compromisso entre simplificação sintática e aumento do texto – é natural que quebrar uma oração longa em várias torne o texto mais longo, pois sujeitos devem ser adicionados, para vários públicos essa é uma adaptação necessária para permitir o entendimento do texto.

A principal ferramenta de simplificação para o português brasileiro foi desenvolvida durante o projeto PorSimples (Aluísio; Gasperin, 2010), e é chamada Simplifica (Candido-Junior et al., 2009; Scarton et al., 2010). Ela apoia autores na redação de textos mais simples, auxiliando tanto na simplificação lexical, que foi baseada em listas de palavras simples, quanto na sintática, realizada via regras baseadas no parser Palavras (Bick, 2000). Atualmente, não está disponível no site do NILC, mas sua interface pode ser vista no relatório do projeto[3].

Inicialmente, a tarefa era solucionada por meio de regras fixas programadas, porém a abordagem mais recente utiliza modelos de redes neurais recorrentes inspiradas na tarefa de Tradução Automática (*Machine Translation*), nos quais o texto original é "traduzido" em sua versão simplificada dentro da própria língua, utilizando o conhecimento adquirido com treinamento em grandes *corpora* (Scarton; Specia, 2018).

6.3.5 Sumarização automática

A Sumarização Automática pode ser definida como a diminuição da extensão dos textos mantendo os conteúdos principais. Ela tem um papel muito importante na simplificação de textos, principalmente para os níveis mais baixos de letramento, nos quais o tamanho

[3] https://fapesp.br/publicacoes/microsoft/microsoft_aluisio.pdf

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

do texto já é um fator desestimulante para a leitura. Diversos métodos de sumarização, na abordagem extrativa (na qual o sumário é composto de orações retiradas do texto original, sem alterações), foram avaliados no projeto PorSimples (Margarido et al., 2008) e foi escolhido o método Extração de Palavras-Chave por frequências de Radicais (EPC-R) para ser usado na ferramenta Facilita, desenvolvida no mesmo projeto (Watanabe et al., 2009a, 2009b).

Para mais informações sobre a tarefa de Sumarização, confira o Capítulo 5.

6.3.6 Elaboração textual

A Elaboração Textual visa melhorar a compreensão de um texto e ampliar ou explorar o vocabulário do leitor, adicionando informações como: sinônimos ou antônimos ao lado de palavras ou expressões complexas, definição de conceitos ou ainda tornar explícitas as conexões entre as ideias (Aluísio; Gasperin, 2010).

A elaboração lexical, em contraste com a simplificação lexical, não substitui as palavras e sim adiciona uma ou uma lista de palavras para explicar o significado de uma palavra complexa. Também pode inserir uma definição curta conforme exemplificado no Quadro 6.3. Trata-se de uma abordagem adequada em situações como o aprendizado de crianças e de uma segunda língua, pois explica e enriquece o vocabulário do estudante (Hartmann; Aluísio, 2020).

Quadro 6.3: Exemplo de sentença simplificada por elaboração textual

Original	A ensilagem é o processo de conservação do alimento por fermentação anaeróbia.
Simplificada	A ensilagem é o processo de conservação do alimento por fermentação anaeróbia **(sem a presença de ar)**.

6.4 Avaliação da Complexidade e suas Métricas

As formas de medir automaticamente a complexidade de textos ou sentenças representam por si só uma ampla área de pesquisa, como dito na introdução. A análise automatizada da complexidade de textos, também conhecida em inglês por ARA (*Automatic Readability Assessment*), tem um viés de aplicação prática, pois ajuda a indicar material de leitura adequado, por exemplo para uma dada série escolar, mas também pode contribuir para um melhor entendimento dos processos de leitura e compreensão em populações com processamento típico e atípico de linguagem.

Graesser et al. (2011) dividem as abordagens de predição e medição da complexidade (ou simplicidade) de textos em:

- **Tradicionais**: que usam uma única métrica ou a combinação linear de poucas métricas de dificuldade;
- **Modernas**: que analisam textos com múltiplas características em vários níveis linguísticos e cognitivos, e foram alavancadas por métodos de AM (aprendizado de máquina) nas últimas duas décadas.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Um exemplo da primeira abordagem é o Índice Flesch que será visto na Seção 6.4.1 e outro da segunda abordagem é o Coh-Metrix, apresentado na Seção 6.4.2.

Um dos grandes desafios para a aplicação dos métodos de AM em textos é a criação de *corpora* grandes e balanceados, anotados com as classes de interesse, por professores ou linguistas. O aprendizado do modelo usa a conversão dos textos em valores, geralmente numéricos, para serem usados nas fases de treinamento e avaliação dos métodos. Isso geralmente é obtido por meio da extração e seleção de métricas dos textos, em diversos níveis da língua, para em seguida utilizá-las como *features* nos métodos de aprendizado de máquina.

Há uma crítica frequente a essa abordagem de anotação da complexidade que usa preditores com base em *corpus* com julgamento de especialistas: o fato de a anotação não ser baseada no desempenho real da leitura de estudantes, por exemplo. Entretanto, se já há grande dificuldade em anotar um grande *corpus* com avaliação de professores, conseguir um *corpus* de alunos é mais ainda difícil e demorado (Vajjala; Meurers, 2016). O *corpus Touchstone Applied Science Associates* (TASA), na língua inglesa, é o único grande *corpus* disponível que atende essa crítica, no melhor do nosso conhecimento, por ser formado por 37.520 amostras de textos, com o tamanho de um parágrafo de tamanho médio de 288,6 palavras (desvio padrão de 25,4), cujas dificuldades foram avaliadas via tarefa de leitura de estudantes, sendo anotados também com a métrica DRP (*Degrees of Reading Power*)[4] (Graesser et al., 2011). Entretanto, a possibilidade de usar rastreamento ocular para capturar o processo de leitura de estudantes é muito bem-vinda e foi explorada na pesquisa de Leal (2021).

Para facilitar a apresentação, são mostradas nas próximas quatro seções as principais fontes de métricas para a tarefa de predição da complexidade (textual e sentencial): fórmulas clássicas, linguísticas, psicolinguísticas e de rastreamento ocular. Dentro de cada seção são descritas as principais métricas citadas na literatura.

6.4.1 Fórmulas clássicas

As primeiras fórmulas para avaliação de inteligibilidade textual surgiram na década de 1920 nos Estados Unidos, e por volta de 1980 já existiam mais de duzentas fórmulas diferentes (Dubay, 2007).

Mesmo com o advento das abordagens mais modernas para resolver a tarefa, essas fórmulas continuam a ter grande importância para as tarefas de PLN, e são usadas isoladamente ou em conjunto com outras *features*. As principais para o escopo deste capítulo são detalhadas a seguir. É importante ressaltar que essas fórmulas foram pensadas para a língua inglesa, portanto não devem ser utilizadas diretamente no português, mas já existem adaptações de algumas delas, como no caso do índice Flesch, por exemplo.

Índice Flesch

A fórmula *Flesch Reading-Ease Score* (FRES), para ser usada com textos em inglês:

$$\mathrm{F} = 206.835 - 1.015\left(\frac{\text{total palavras}}{\text{total sentenças}}\right) - 84.6\left(\frac{\text{total sílabas}}{\text{total palavras}}\right) \tag{6.1}$$

O valor da fórmula pode ser interpretado com a seguinte escala:

- 90-100: Muito simples,

[4] http://textcomplexity.questarai.com/getdrp/

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

- 80-89: Simples,
- 70-79: Relativamente simples,
- 60-69: Padrão,
- 50-59: Relativamente complexo,
- 30-49: Complexo,
- 0-29: Muito complexo.

É uma das mais antigas e utilizadas fórmulas de inteligibilidade e foi criada por Rudolph Flesch em 1948 (Dell'Orletta et al., 2011; Sjöholm, 2012). Foi adaptada para o português brasileiro em 1996 pelo NILC (Martins et al., 1996), adicionando 42 pontos a todos os escores da fórmula original em inglês:

$$\mathrm{F} = 248.835 - 1.015 \left(\frac{\text{total palavras}}{\text{total sentenças}} \right) - 84.6 \left(\frac{\text{total sílabas}}{\text{total palavras}} \right) \tag{6.2}$$

Flesch-Kincaid grade level

A fórmula *Flesch–Kincaid Grade Level* apresenta como resultado um número que corresponde a uma série no sistema educacional americano, facilitando a avaliação do nível de complexidade de livros e textos. Pode ser interpretada como o número de anos de educação necessários para a leitura de um dado texto:

$$\mathrm{FK} = 0.39 \left(\frac{\text{total palavras}}{\text{total sentenças}} \right) + 11.8 \left(\frac{\text{total sílabas}}{\text{total palavras}} \right) - 15.59 \tag{6.3}$$

Foi desenvolvida por J. Peter Kincaid em 1975 (Kincaid et al., 1975) a partir da anterior criada por Rudolph Flesch. É também uma função linear que utiliza a média de sílabas por palavras e média de palavras por sentença, estimando assim as complexidades lexical e sintática do texto (Dell'Orletta et al., 2011; Sjöholm, 2012).

Dale-Chall

Inspirada pela fórmula Flesch, a fórmula *Dale-Chall* acrescenta validação da dificuldade das palavras contra um dicionário com 3 mil palavras simples, sendo também considerada a média do tamanho das sentenças:

$$\mathrm{DC} = 0.1579 \left(\frac{\text{total palavras difíceis}}{\text{total palavras}} \times 100 \right) + 0.0496 \left(\frac{\text{total palavras}}{\text{total sentenças}} \right) \tag{6.4}$$

Foi criada em 1948 e atualizada posteriormente em 1995 por Edgard Dale e Jeanne Chall (Chall; Dale, 1995; Dell'Orletta et al., 2011).

Gunning Fog Index

Gunning Fog Index ou simplesmente FOG Index foi criada em 1952 por Robert Gunning:

$$\mathrm{GF} = 0.4 \left[\left(\frac{\text{total palavras}}{\text{total sentenças}} \right) + 100 \left(\frac{\text{total palavras complexas}}{\text{total palavras}} \right) \right] \tag{6.5}$$

Ao avaliar a dificuldade de inteligibilidade dos jornais por estudantes de graduação, ele escreveu que os textos estavam repletos de incertezas, névoa (*fog* em inglês) e complexidade

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

desnecessária (Dubay, 2014). Palavras complexas nesse contexto são as que possuem três ou mais sílabas.

Coleman–Liau

Baseada em caracteres em vez de sílabas por palavra, possibilita utilizações mais mecânicas em textos:

$$\mathrm{CLI} = 0.0588 \times \mathrm{L} - 0.296 \times \mathrm{S} - 15.8 \tag{6.6}$$

Na fórmula acima, L é a média da quantidade de letras por 100 palavras e S é a média do número de sentenças por 100 palavras (Coleman; Liau, 1975).

Brunét

O Índice de Brunét é uma variação da TTR (*Type Token Ratio*), mas insensível ao tamanho do texto:

$$\mathrm{W} = \mathrm{N}^{\mathrm{V}^{-0.165}} \tag{6.7}$$

Na fórmula acima N é o número de *tokens* e V é o total de palavras do vocabulário (ou *types*). Foi criado por Étienne Brunet em 1978 (Cunha, 2015; Thomas et al., 2005). Os valores típicos da métrica variam entre 10 e 20, sendo que uma fala mais rica produz valores menores.

Honoré

A Estatística de Honoré é outra variação da TTR, também insensível ao tamanho do texto:

$$\mathrm{R} = \frac{100 log \mathrm{N}}{1 - \frac{V_1}{V}} \tag{6.8}$$

Na fórmula acima N é o número de *tokens* e V_1 é o número de palavras do vocabulário que aparecem uma única vez e V é o número de itens lexicais (ou *types*). Foi criada por A. Honoré em 1979 (Cunha, 2015; Thomas et al., 2005), sendo que valores altos da fórmula indicam um vocabulário rico.

6.4.1.1 ALT - Análise de Legibilidade Textual

Com a devida ressalva já feita na introdução sobre o termo legibilidade, o ALT[5] (Moreno et al., 2022) é uma ferramenta recente bem interessante e publicamente disponível que traz a implementação das fórmulas vistas nesta seção com uma representação visual bem moderna. Na Figura 6.2 podemos ver o resultado da análise para um texto simples.

6.4.2 Métricas Linguísticas

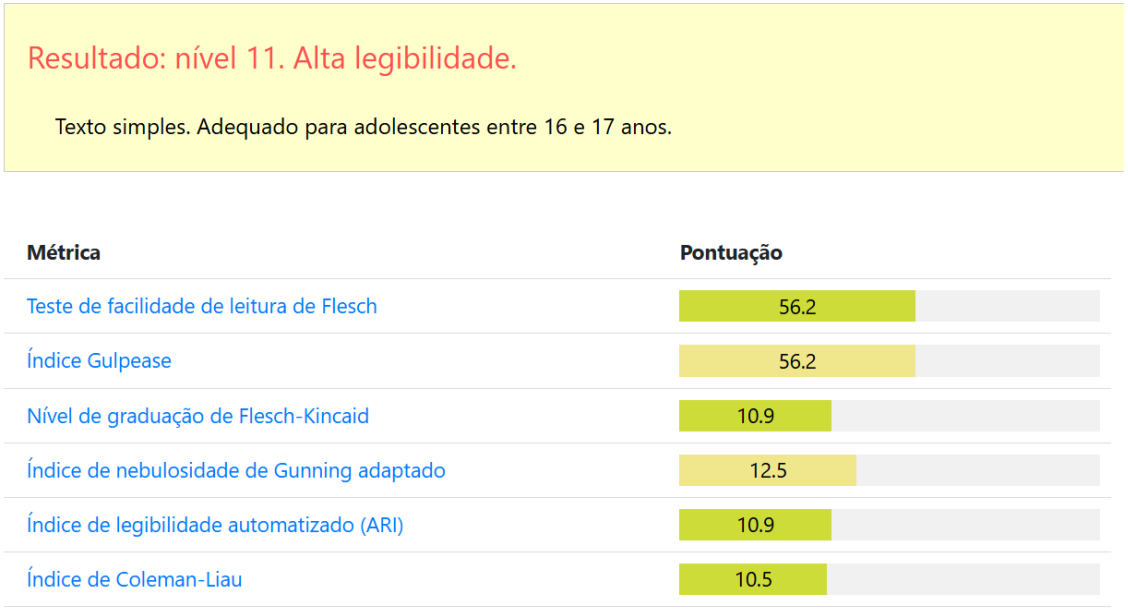
As métricas linguísticas extraem características nos níveis lexical, morfossintático, sintático, semântico e discursivo da língua. Existem ferramentas com esse fim específico, que facilitam bastante o processo. Trazemos nas seções seguintes, um resumo de seis ferramentas, tanto para o inglês como para o português.

[5] https://legibilidade.com/

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Figura 6.2: Tela de resultados da ferramenta ALT



Coh-Metrix

O Coh-Metrix[6] (McNamara et al., 2014) é uma ferramenta desenvolvida para a língua inglesa, que extrai de um texto métricas de coesão e coerência, permitindo avaliar a complexidade da sua leitura.

Os autores definem coesão como a relação entre as características do texto que guiam o leitor para a representação mental do significado, e coerência é a representação mental que o leitor cria durante a leitura (Graesser et al., 2004).

A versão 3.0 do Coh-Metrix implementa 106 métricas para a língua inglesa, agrupadas nas 11 categorias: *Descriptive, Text Easability Principal Component Scores, Referential Cohesion,*Latent Semantic Analysis* (LSA), Lexical Diversity, Connectives, Situation Model, Syntactic Complexity, Syntactic Pattern Density, Word Information e Readability*.

A tela da ferramenta pode ser vista na Figura 6.3), que traz no lado direito os valores de diversas métricas do pequeno texto informado no lado esquerdo.

Em contraste com as fórmulas clássicas que analisam o texto apenas no nível das palavras e sentenças e geram um único valor para quantificar a complexidade do texto, o Coh-Metrix utiliza uma análise multinível, alinhada com teorias de compreensão textual (Graesser et al., 2011):

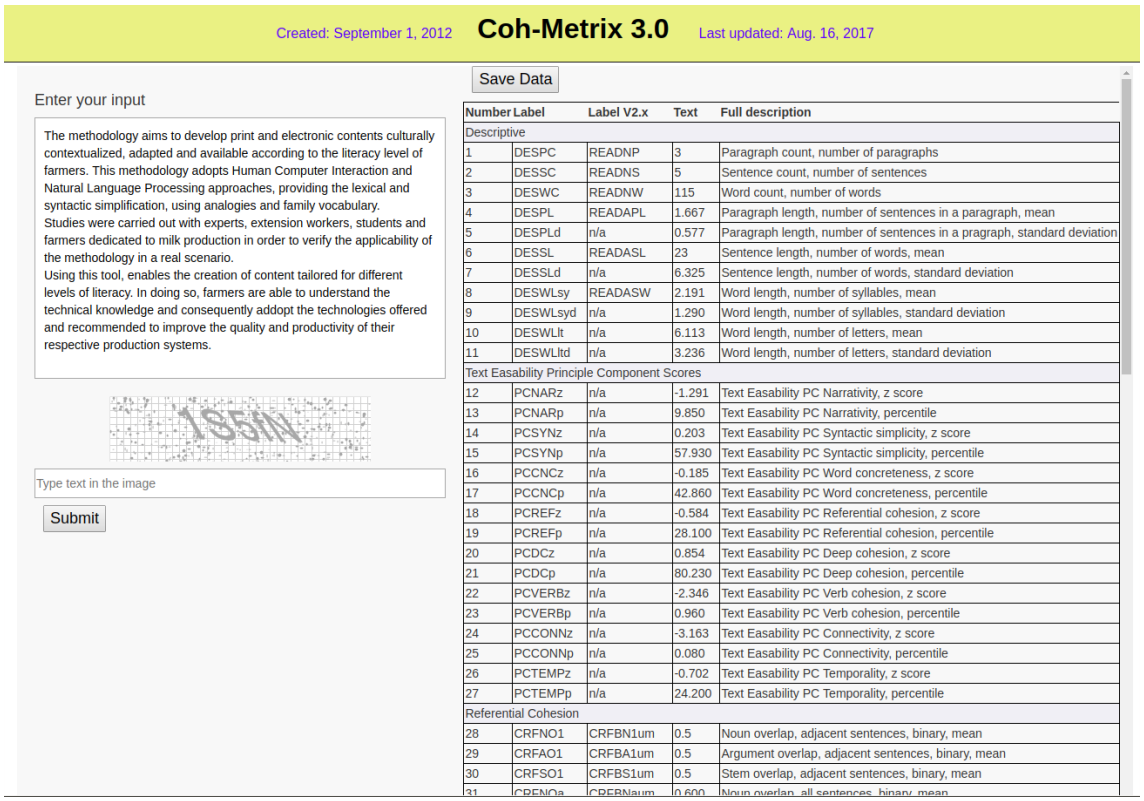
1. ***Words***: Como o conhecimento do vocabulário de uma língua tem um grande impacto sobre o tempo de leitura e compreensão, Coh-Metrix tem uma grande quantidade de métricas relacionadas a palavras, incluindo: análise de categorias gramaticais ou *Part of Speech* (PoS), frequência, medidas psicolinguísticas como concretude, familiaridade, idade de aquisição, imageabilidade, categorias semânticas obtidas da WordNet de Princeton[7];

[6] http://cohmetrix.com/
[7] https://wordnet.princeton.edu/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

Figura 6.3: Tela do Coh-Metrix com um texto de exemplo



Created: September 1, 2012 **Coh-Metrix 3.0** Last updated: Aug. 16, 2017

Enter your input

The methodology aims to develop print and electronic contents culturally contextualized, adapted and available according to the literacy level of farmers. This methodology adopts Human Computer Interaction and Natural Language Processing approaches, providing the lexical and syntactic simplification, using analogies and family vocabulary.
Studies were carried out with experts, extension workers, students and farmers dedicated to milk production in order to verify the applicability of the methodology in a real scenario.
Using this tool, enables the creation of content tailored for different levels of literacy. In doing so, farmers are able to understand the technical knowledge and consequently addopt the technologies offered and recommended to improve the quality and productivity of their respective production systems.

Type text in the image

Submit

Save Data

Number	Label	Label V2.x	Text	Full description
Descriptive				
1	DESPC	READNP	3	Paragraph count, number of paragraphs
2	DESSC	READNS	5	Sentence count, number of sentences
3	DESWC	READNW	115	Word count, number of words
4	DESPL	READAPL	1.667	Paragraph length, number of sentences in a paragraph, mean
5	DESPLd	n/a	0.577	Paragraph length, number of sentences in a pragraph, standard deviation
6	DESSL	READASL	23	Sentence length, number of words, mean
7	DESSLd	n/a	6.325	Sentence length, number of words, standard deviation
8	DESWLsy	READASW	2.191	Word length, number of syllables, mean
9	DESWLsyd	n/a	1.290	Word length, number of syllables, standard deviation
10	DESWLlt	n/a	6.113	Word length, number of letters, mean
11	DESWLltd	n/a	3.236	Word length, number of letters, standard deviation
Text Easability Principle Component Scores				
12	PCNARz	n/a	-1.291	Text Easability PC Narrativity, z score
13	PCNARp	n/a	9.850	Text Easability PC Narrativity, percentile
14	PCSYNz	n/a	0.203	Text Easability PC Syntactic simplicity, z score
15	PCSYNp	n/a	57.930	Text Easability PC Syntactic simplicity, percentile
16	PCCNCz	n/a	-0.185	Text Easability PC Word concreteness, z score
17	PCCNCp	n/a	42.860	Text Easability PC Word concreteness, percentile
18	PCREFz	n/a	-0.584	Text Easability PC Referential cohesion, z score
19	PCREFp	n/a	28.100	Text Easability PC Referential cohesion, percentile
20	PCDCz	n/a	0.854	Text Easability PC Deep cohesion, z score
21	PCDCp	n/a	80.230	Text Easability PC Deep cohesion, percentile
22	PCVERBz	n/a	-2.346	Text Easability PC Verb cohesion, z score
23	PCVERBp	n/a	0.960	Text Easability PC Verb cohesion, percentile
24	PCCONNz	n/a	-3.163	Text Easability PC Connectivity, z score
25	PCCONNp	n/a	0.080	Text Easability PC Connectivity, percentile
26	PCTEMPz	n/a	-0.702	Text Easability PC Temporality, z score
27	PCTEMPp	n/a	24.200	Text Easability PC Temporality, percentile
Referential Cohesion				
28	CRFNO1	CRFBN1um	0.5	Noun overlap, adjacent sentences, binary, mean
29	CRFAO1	CRFBA1um	0.5	Argument overlap, adjacent sentences, binary, mean
30	CRFSO1	CRFBS1um	0.5	Stem overlap, adjacent sentences, binary, mean
31	CRFNOa	CRFBNaum	0.600	Noun overlap, all sentences, binary, mean

2. ***Syntax***: Algumas sentenças do discurso oral são curtas, apresentam poucas orações relativas, poucas palavras nos sintagmas nominais e se apresentam na voz ativa, mas sentenças de textos escritos geralmente aparecem de forma oposta, demandando mais processamento da memória de trabalho. Coh-Metrix computa essas contagens e outras como similaridade de pares de sentenças adjacentes, que facilitam a leitura e compreensão;

3. ***Textbase***: A base textual está relacionada com o significado em vez da análise de palavras e da sintaxe. A correferência é um mecanismo importante para conectar as proposições, as orações e sentenças na base textual, assim Coh-Metrix traz várias métricas para o cômputo da correferência como a sobreposição de palavras de conteúdo, de substantivos e de radicais (*content word overlap, noun overlap, e stem overlap*, respectivamente). A diversidade lexical está relacionada com a coesão porque um número elevado de palavras diferentes em um texto significa que as palavras novas precisam ser integradas no contexto do discurso. Coh-Metrix também computa várias métricas relacionadas com o modelo estatístico para cálculo de similaridade chamado *Latent Semantic Analysis* (LSA), pois ele ajuda a medir o conhecimento implícito do leitor em adição às palavras explícitas usadas no texto;

4. ***Situation Model / Mental Model***: Textos narrativos incluem pessoas, objetos, ações, eventos, processos, planos e outros detalhes de uma estória, já em textos informativos o modelo mental é diferente, pois devem ajudar a entender como modelos da física, biologia e outras ciências funcionam. Assim, há métricas para avaliar se há quebras no entendimento desses modelos mentais que emergem de um texto;

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

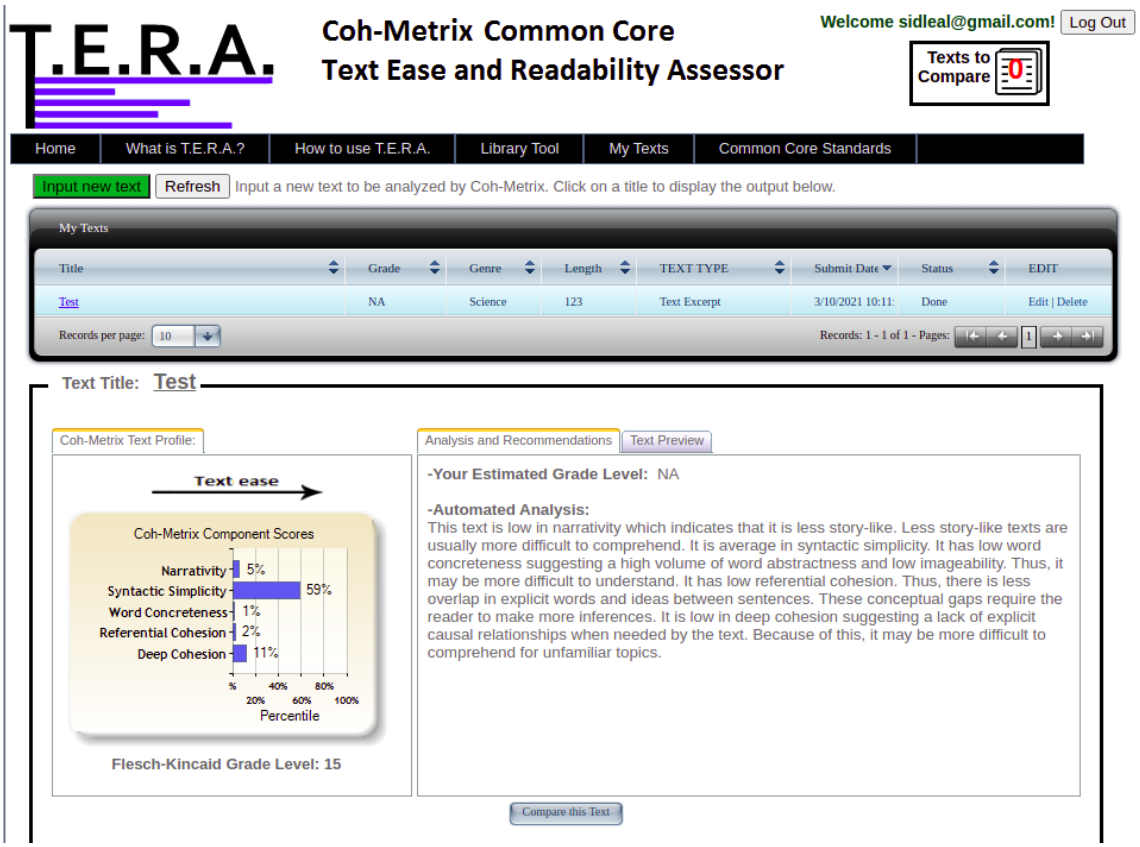
©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

5. ***Genre and Rhetorical Structure***: Exemplos de uma tipologia de gêneros são: narrativo, expositivo, persuasivo ou descritivo. Textos narrativos são mais fáceis de se ler, compreender e relembrar do que textos informativos. Coh-Metrix analisa se um texto pode ser classificado como narrativo ou informativo, via uma métrica chamada narratividade.

T.E.R.A

T.E.R.A.[8] (*Text Ease and Readability Assessor*) é uma ferramenta construída pelos mesmos autores do Coh-Metrix, também para língua inglesa. Utiliza o Coh-Metrix para avaliar amostras dos textos, reduzindo as métricas em cinco fatores, levantados via PCA (*Principal Component Analysis*) ((Graesser et al., 2011; McNamara et al., 2013)): *Narrativity, Syntactic Simplicity, Word Concreteness, Referential Cohesion (Textbase)* e *Deep Cohesion (Situation Model)*. Na Figura 6.4 é possível ver um exemplo da análise do texto com as cinco dimensões.

Figura 6.4: Tela de exemplo do Coh-Metrix-T.E.R.A.



Coh-Metrix-Port

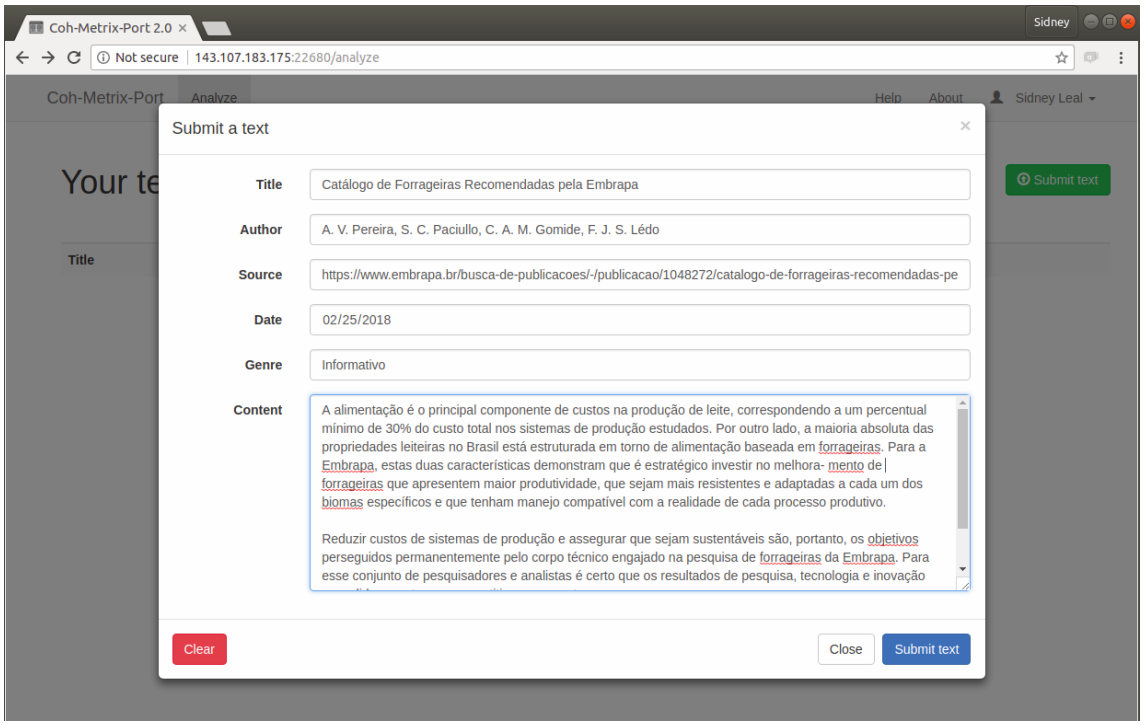
O Coh-Metrix-Port[9] (Scarton; Aluísio, 2010) é uma adaptação para o português brasileiro do Coh-Metrix, desenvolvida dentro do projeto PorSimples (Simplificação Textual do Por-

[8] http://www.commoncoretera.com/
[9] http://fw.nilc.icmc.usp.br:22680/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

Figura 6.5: Tela de avaliação do Coh-Metrix-Port 2



tuguês para Inclusão e Acessibilidade Digital), que teve como objetivo promover o acesso a textos da Web a pessoas com baixo letramento.

O Coh-Metrix-Port implementa 48 métricas específicas para o português brasileiro (Scarton et al., 2010), divididas nas categorias: contagens básicas, operadores lógicos, frequências, hiperônimos, tokens, constituintes, conectivos, ambiguidade, correferência e anáforas. A tela de cadastro dos textos, da versão 2.0[10], pode ser vista na Figura 6.5.

Coh-Metrix-Dementia

O Coh-Metrix-Dementia[11] (Cunha, 2015) é uma adaptação do Coh-Metrix-Port para análise automática de distúrbios de linguagem nas demências (como Doença de Alzheimer) ou no Comprometimento Cognitivo Leve (CCL).

Ele adiciona 25 novas métricas às 48 do Coh-Metrix-Port, nas categorias: disfluências, análise de semântica latente, diversidade lexical, complexidade sintática e densidade semântica.

Disponibiliza no total 73 métricas para o português brasileiro. Sua tela principal pode ser vista na Figura 6.6.

LIWC

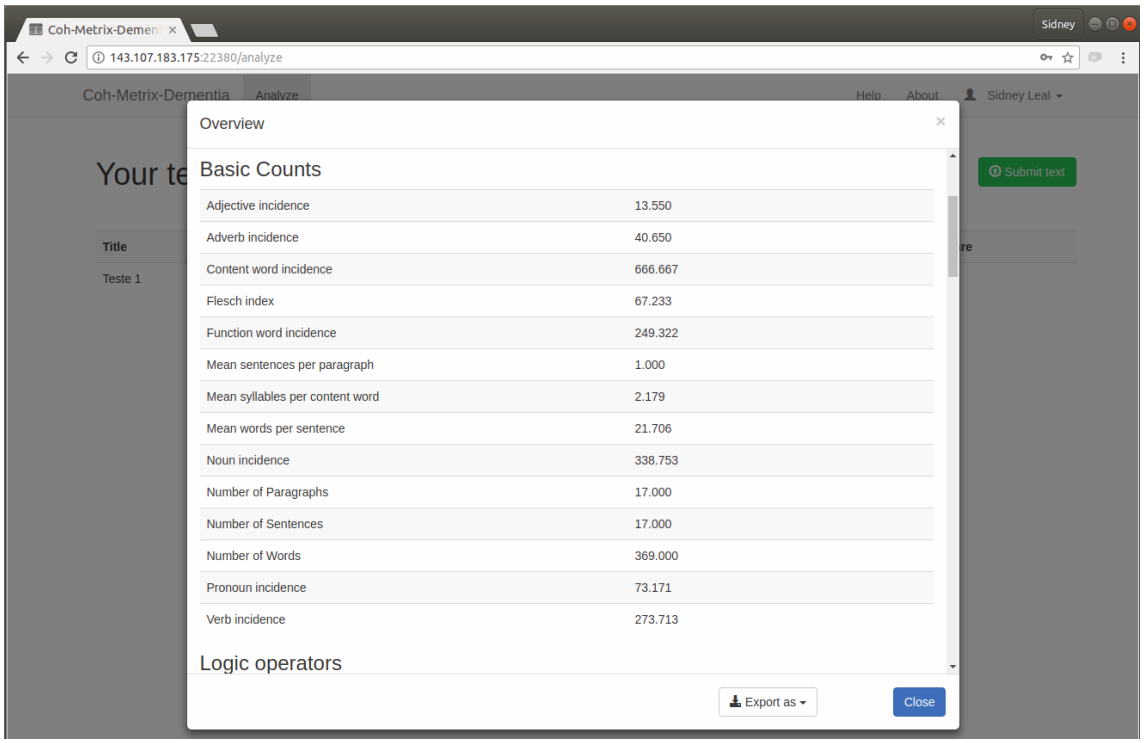
LIWC (*Linguistic Inquiry and Word Count*) é uma ferramenta baseada em dicionários para análise dos vários componentes emocionais, cognitivos e linguísticos em amostras de textos (Pennebaker et al., 2015), com categorias como: estatísticas comuns do texto,

[10] Refeita por Cunha (2015)
[11] http://fw.nilc.icmc.usp.br:22380/

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Figura 6.6: Exemplo da tela de saída do Coh-Metrix-Dementia



dimensão linguística, processos psicológicos, relatividade, assuntos pessoais e miscelânea, totalizando aproximadamente 100 métricas (Cunha, 2015).

A sua primeira versão foi criada em 1993, a segunda em 2001, a terceira em 2007 e a última em 2015. O dicionário da versão inglesa conta com 6400 palavras, radicais e emoticons (Pennebaker et al., 2015).

A tradução e adaptação do dicionário para o português brasileiro foi realizada em uma colaboração entre NILC, Checon Pesquisa e Unisinos no período de 2010 a 2012 e está disponível no site do projeto PortLex[12].

AIC

Também criada dentro do contexto do PorSimples (Maziero et al., 2008), a ferramenta AIC (Análise Automática de Inteligibilidade de *Corpus*) traz 39 métricas, com o principal diferencial de utilizar o analisador sintático PALAVRAS (Bick, 2000) para o cálculo delas. Elas estão organizadas em seis classes: estatísticas do texto, voz passiva, características das orações, densidade, personalização e marcadores discursivos (Cunha, 2015; Reis, 2017). A tela de saída pode ser vista na Figura 6.7.

6.4.3 Métricas Psicolinguísticas

As palavras possuem algumas propriedades subjetivas estudadas pela psicolinguística como: imageabilidade, concretude, familiaridade e idade de aquisição (Santos et al., 2017a), detalhadas abaixo:

[12] http://nilc.icmc.usp.br/portlex/index.php/pt/projetos/liwc

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

Figura 6.7: Exemplo da tela de saída da AIC, atualmente não disponível no site do NILC



Fonte: (Maziero et al., 2008)

- **Imageabilidade**: Envolve a facilidade e rapidez de evocar uma imagem mental da palavra.
- **Concretude**: É o grau com que uma palavra se refere a objetos, pessoas, lugares ou coisas que podem ser percebidas pelos sentidos, em contraste com os conceitos abstratos.
- **Familiaridade**: É o grau com que as pessoas conhecem e usam palavras no dia a dia.
- **Idade de Aquisição**: Estimativa da idade em que uma palavra foi aprendida, calculada via análise feita por adultos.

Essas propriedades têm um grande impacto na complexidade dos textos e sentenças, e trazem melhorias aos resultados de várias tarefas de PLN, como simplificação lexical e tarefas de classificação semântica quando utilizadas em conjunto com as demais métricas (Paetzold; Specia, 2016b).

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Santos et al. (2017a) anotaram automaticamente essas métricas em um banco[13] de 26.874 palavras do português brasileiro utilizando um método baseado em regressão e *Multi-View Learning*, com recursos fáceis de se obter em várias línguas.

6.4.4 NILC-Metrix

NILC-Metrix[14] é um conjunto de 200 métricas desenvolvido no Centro Interinstitucional de Linguística Computacional – NILC[15], do final de 2007 ao início de 2021 (Leal et al., 2023).

O principal objetivo dessas métricas é fornecer subsídios para avaliar a coesão, a coerência e a complexidade textual. O NILC-Metrix pode ajudar pesquisadores a investigar várias questões de pesquisa, por exemplo:

1. como as características do texto se correlacionam com a compreensão da leitura;
2. quais são as características mais desafiadoras de um determinado texto, ou seja, quais características tornam um texto ou corpus mais complexo;
3. quais textos possuem as características mais adequadas para desenvolver as habilidades dos alunos-alvo;
4. quais partes de um texto são desproporcionalmente complexas e devem ser simplificadas para atender a um determinado público.

Além de ser disponibilizado via interface web, o código fonte também foi disponibilizado sob licença AGPLv3[16], facilitando a incorporação do código das métricas em aplicações. As métricas são agrupadas em 14 categorias, seguindo sua similaridade e fundamentação teórica. São elas:

1. Medidas Descritivas (10 métricas)
2. Simplicidade Textual (8 métricas)
3. Coesão Referencial (9 métricas)
4. Coesão Semântica via LSA (11 métricas)
5. Medidas Psicolinguísticas (24 métricas)
6. Diversidade Lexical (15 métricas)
7. Conectivos (12 métricas)
8. Léxico Temporal (12 métricas)
9. Complexidade Sintática (27 métricas)
10. Densidade de Padrões Sintáticos (4 métricas)

[13] A base está disponível em: http://143.107.183.175:21380/portlex/index.php/en/component/content/article/2-uncategorised/23-psycholinguistic
[14] http://fw.nilc.icmc.usp.br:23380/nilcmetrix
[15] http://www.nilc.icmc.usp.br/
[16] https://github.com/nilc-nlp/nilcmetrix

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

11. Informações Morfossintáticas de Palavras (11 métricas)
12. Informações Semânticas de Palavras (11 métricas)
13. Frequência de Palavras (10 métricas)
14. Índices de Leiturabilidade (5 métricas)

6.4.5 Rastreamento ocular

As métricas do rastreamento ocular trazem uma abordagem recente e diferente das métricas mostradas anteriormente. Sua contribuição é bastante relevante, uma vez que permitem uma aproximação da percepção mais realista da complexidade pelos leitores. Elas foram fundamentais no trabalho de avaliação da complexidade no nível sentencial para o português brasileiro (Leal et al., 2020).

Os movimentos dos olhos podem ser interpretados como uma janela para o processamento do cérebro, refletindo os tempos cognitivos envolvidos em determinada tarefa. Por exemplo, durante a leitura os movimentos dos olhos são controlados por uma complexa interação entre os fatores de baixo nível (por exemplo, o quanto o olho consegue ver e interpretar a cada fixação) e de alto nível (por exemplo, o processamento sintático) (Barrett et al., 2015).

Rayner (1998) divide a pesquisa sobre os movimentos dos olhos (ou rastreamento ocular) em três grandes eras. A primeira era vai desde as primeiras observações sobre os movimentos dos olhos durante a leitura em 1879 até os anos 1920. Algumas importantes descobertas foram feitas nessa era como, por exemplo, o fato de que não percebemos nenhuma informação durante o reposicionamento do olhar, denominado sacada ou *saccade*, em inglês.

A segunda era coincide com o movimento behaviorista na psicologia experimental, com os trabalhos com focos mais práticos – estudos dos movimentos dos olhos em si ou em aspectos superficiais da tarefa investigada – e menos concentrados na utilização dos movimentos para inferir o processamento cognitivo.

Figura 6.8: Rastreamento ocular de plataforma, com óculos simples, e com óculos especial de realidade virtual



Fonte: (Fove, 2018; Imotions, 2017)

A terceira era começa em meados dos anos 1970, com melhorias nos sistemas de rastreamento que permitiram medidas mais acuradas e simples de obter (Figura 6.8). Nessa era, juntamente com os avanços das teorias de processamento da linguagem, os movimentos dos olhos começaram a ser utilizados para exame crítico dos processos cognitivos durante a leitura.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

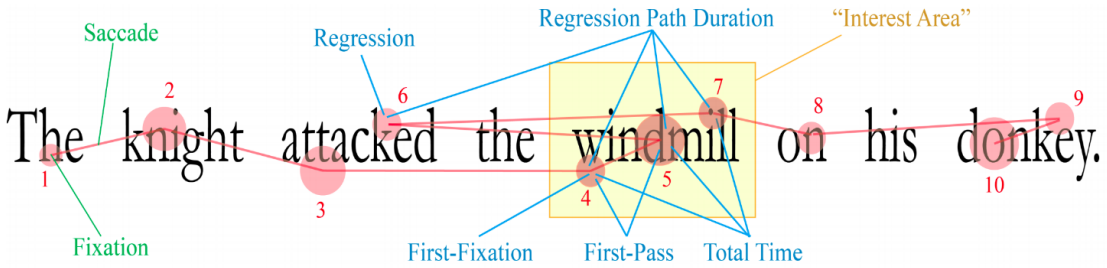
©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

No português brasileiro, o rastreamento ocular já é utilizado há algum tempo na área da psicolinguística. Por exemplo, Maia et al. (2007) utilizaram para investigar o papel do processamento morfológico na identificação de palavras, Leitão et al. (2012) utilizaram na investigação do processamento anafórico e Teixeira et al. (2014) para evidenciar o custo de resolução de pronomes nulos e plenos.

As características básicas dos movimentos dos olhos são:

- **Sacadas (*Saccades* em inglês)**: Os contínuos movimentos oculares, o reposicionamento do olhar (durante uma sacada nenhuma informação é percebida).
- **Fixações (*Fixations* em inglês)**: Os tempos de fixação em um ponto de atenção entre as sacadas.

Figura 6.9: Principais métricas de rastreamento ocular



Fonte: (Zelenina, 2015)

A partir dessas duas características é possível medir diversas outras informações relevantes para o processo de leitura e interpretação de textos. As principais métricas obtidas pelo movimento dos olhos são exemplificadas na 8, com a simulação do caminho feito pelo olhar em dez fixações numeradas sequencialmente, e detalhadas a seguir:

- ***First fixation duration***: Tempo da primeira fixação na palavra.
- ***First pass fixation duration***: Quando uma palavra é longa, pode ser necessário um segundo ponto de fixação dentro da própria palavra. Essa métrica é a soma dos tempos das fixações na primeira passada pela palavra.
- ***Total fixation duration***: Soma de todos os tempos de fixação na palavra.
- ***Average fixation duration***: Tempo médio de fixação, quando se tem mais de um ponto por palavra ou média por sentença.
- ***Regression***: Regressões no texto podem indicar necessidade de rever alguma informação para entendimento do ponto atual, por exemplo, para resolver uma correferência. É uma métrica muito importante para medir a complexidade textual e sentencial.
- ***Regression path duration***: Mede a extensão da regressão; quanto maior a regressão, maior o esforço despendido para a leitura, como resultado de um texto mais complexo.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

- ***Interest area***: Pontos de interesse, onde o leitor passou mais tempo fixando no texto. Calculado com a soma de todas as fixações.
- ***Skipping rate***: Algumas palavras são naturalmente saltadas durante a leitura, como artigos e preposições. Não saltar essas palavras pode indicar um leitor com menor proficiência na leitura.
- ***Number of fixations***: Quantidade de fixações na palavra; uma palavra simples só deve exigir uma única fixação.
- ***Second pass fixation duration***: Tempo de fixação na segunda vez que o leitor retorna à palavra.
- ***Spillover from previous word***: Nem sempre o processamento de uma palavra é completado antes que o olhar se mova para a próxima. Nesses casos ocorre o efeito de "transbordamento" do tempo para a palavra seguinte.

6.5 Recursos para Complexidade Textual e Sentencial no Português Brasileiro

Até onde foi possível verificar, existem poucos recursos disponíveis publicamente para as tarefas de complexidade no nível textual e sentencial para o português brasileiro. A seguir são apresentados resumos dos recursos e links para download, sempre que possível.

6.5.1 Recursos para Complexidade Textual

Nesta seção, consideramos os recursos com textos para a tarefa. O PorSimples traz textos simplificados em dois níveis, o RastrOS disponibiliza dados de rastreamento ocular em textos curtos e os *corpora* do projeto PorPopular são compilações representativas do português popular em jornais como o Diário Gaúcho[17] e Massa[18].

PorSimples

Corpus paralelo de textos originais e simplificados criado em 2009 no projeto PorSimples[19] (Simplificação Textual do Português para Inclusão e Acessibilidade Digital) do NILC ((Caseli et al., 2009)).

Um editor[20] foi desenvolvido para a tarefa de anotação, cuja tela principal pode ser vista na Figura 6.10. No lado esquerdo, fica o texto original e à direita, sua versão simplificada. Os textos jornalísticos foram simplificados em dois níveis por especialistas linguistas:

- **Natural**: textos para os quais o anotador escolheu livremente as operações de simplificação, inclusive podendo escolher não simplificar uma sentença.
- **Forte**: Anotadores seguiram o manual de simplificação também desenvolvido no projeto.

[17] https://diariogaucho.clicrbs.com.br/
[18] https://jornalmassa.com.br/
[19] http://www.nilc.icmc.usp.br/nilc/index.php/porsimples
[20] Após mais de uma década, o editor continua disponível e funcionando, pode ser utilizado em: http://fw.nilc.icmc.usp.br:23080/

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Figura 6.10: Exemplo da tela do editor de anotação do PorSimples



Fonte: (Caseli et al., 2009)

A primeira fase do *corpus* foi criada a partir de 104 textos do jornal Zero Hora. Na segunda fase, foram adicionados 50 textos do Caderno de Ciência do jornal Folha de São Paulo[21], resultando em 154 trios alinhados, com o total de 462 textos e mais de 185 mil tokens.

Na Tabela 6.1 podem ser vistos os números de sentenças de cada nível, e para os tokens confira a Tabela 6.2. Uma característica importante do *corpus* PorSimples foi a anotação dos fenômenos linguísticos nas sentenças; uma extração deles pode ser vista para cada nível na Tabela 6.3. Com essa informação é possível verificar que alguns fenômenos adicionam mais complexidade que outros, por exemplo, as orações apositivas foram as que mais diminuíram em número durante o processo de simplificação, já as subordinadas contraintuitivamente aumentaram nos níveis mais simples.

Tabela 6.1: PorSimples - Estatísticas de Sentenças

	Total	Mínimo/Texto	Máximo/Texto	Média/Texto
Original	2.985	5	46	19
Natural	4.080	5	62	26
Forte	4.974	7	72	32

[21] https://www1.folha.uol.com.br/fsp/

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Tabela 6.2: PorSimples - Estatísticas de Tokens

	Total	Mínimo/Sentença	Máximo/Sentença	Média/Sentença
Original	61.026	2	71	21
Natural	61.754	2	60	15
Forte	63.030	2	47	13

Tabela 6.3: PorSimples - Fenômenos Linguísticos

	Coordenadas	Subordinadas	Relativas	Passivas	Apositivas
Original	1.443	805	897	319	306
Natural	1.352	899	759	257	105
Forte	1.210	876	527	167	73

Um ponto muito importante é que esse *corpus* possui alinhamento no nível sentencial. Esse alinhamento foi utilizado para gerar o *corpus* PorSimplesSent, detalhado mais adiante.

Corpus de Complexidade Textual para Estágios Escolares do Sistema Educacional Brasileiro

Gazzola et al. (2019) compilaram um grande *corpus* com textos utilizados em diferentes etapas de ensino do Sistema Educacional Brasileiro. O *corpus* foi organizado nas quatro etapas utilizadas na Plataforma MEC de Recursos Educacionais Digitais (MEC-RED)[22] para classificação nos estágios escolares:

- Ensino Fundamental I (primeiro ao quinto ano);
- Ensino Fundamental II (sexto ao nono ano);
- Ensino Médio;
- Ensino Superior.

O *corpus* está publicamente disponível[23] e inclui livros-texto, notícias da seção Para Seu Filho Ler (PSFL) do jornal Zero Hora (que apresenta algumas notícias sobre os mesmos tópicos do jornal, mas escritas para crianças de 8 a 11 anos de idade), Exames do SAEB, Livros Digitais do Wikilivros em Português e Exames do Enem dos anos 2015, 2016 e 2017.

Em números, o *corpus* disponibiliza 2.067 documentos (min = 300 palavras, max = 596 palavras, média = 448).

CorPop e PorPopular

O CorPop (*corpus* de referência do português popular escrito do Brasil) foi criado em 2018 por Pasqualini (2018) durante seu doutorado. Ele traz uma compilação bem avaliada de textos selecionados com base no nível de letramento médio dos leitores do país, das seguintes fontes:

[22] https://plataformaintegrada.mec.gov.br/

[23] http://www.nilc.icmc.usp.br/nilc/index.php/tools-and-resources e https://github.com/gazzola/corpus_readability_nlp_portuguese

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

1. Textos do jornalismo popular do Projeto PorPopular (jornal Diário Gaúcho) consumido maciçamente pelas classes C e D;
2. Textos e autores mais lidos pelos respondentes das últimas edições da pesquisa "Retratos da Leitura no Brasil";
3. Coleção "É Só o Começo" (adaptação de clássicos da literatura brasileira para leitores com baixo letramento, realizada por linguistas);
4. Textos do jornal Boca de Rua[24], produzido por pessoas em situação de rua, com baixa escolaridade e baixo letramento;
5. Textos do Diário da Causa Operária[25], imprensa operária brasileira produzida também por pessoas dentro da faixa média de letramento do país.

O *corpus* possui 684 mil tokens. Está parcialmente disponível publicamente[26] (ferramentas e listas de palavras).

Além do CorPop, o projeto PorPopular[27] (Finatto, 2012) disponibiliza um *corpus* com amostras do jornal Diário Gaúcho e outro com amostras do jornal Massa.

MedSimples

A MedSimples[28] é uma ferramenta que auxilia na simplificação de textos sobre temas de saúde (Paraguassu et al., 2020). Inicialmente se baseou em terminologias e palavras potencialmente difíceis extraídas de um *corpus* sobre a doença de Parkinson, depois foram adicionados um *corpus* sobre oncologia e também a base do CorPop (Villar; Finatto, 2023).

Na ferramenta foi trabalhado o conceito de acessibilidade textual e terminológica, explorado no *e-book*[29] gratuito e publicamente disponível na internet (Finatto; Paraguassu, 2022).

RastrOS

RastrOS[30] é um *corpus* de 50 textos curtos com dados de rastreamento oculares de estudantes universitários durante a leitura silenciosa de parágrafos em português brasileiro. Ele foi criado para a evolução da tarefa de avaliação da complexidade sentencial (Leal et al., 2020, 2022; Vieira, 2020).

Além dos dados de rastreamento ocular de 37 participantes, ele também disponibiliza normas de previsibilidade semântica, obtidas por meio da aplicação de teste cloze com 393 participantes. O RastrOS pode ser obtido integralmente no repositório OSF (*Open Science Framework*)[31].

[24] https://jornalbocaderua.wordpress.com/
[25] https://causaoperaria.org.br/
[26] http://www.ufrgs.br/textecc/porlexbras/corpop/index.php
[27] https://www.ufrgs.br/textecc/porlexbras/porpopular/
[28] https://www.ufrgs.br/textecc/acessibilidade/page/cartilha/
[29] https://repositorio.ufu.br/bitstream/123456789/35193/1/eClasse_Acessibilidade_Textual.pdf
[30] http://www.nilc.icmc.usp.br/nilc/index.php/rastros
[31] https://osf.io/9jxg3/

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

6.5.2 Recursos para Complexidade Sentencial

Apresentamos a seguir os recursos disponíveis em português brasileiro para as tarefas que envolvem o trabalho com a complexidade no nível das sentenças.

PorSimplesSENT

O PorSimplesSENT é o primeiro *corpus* para o português brasileiro voltado para as tarefas de complexidade no nível sentencial. Foi compilado por Leal et al. (2018) a partir dos alinhamentos disponibilizados pelo *corpus* PorSimples.

O *corpus* está disponível publicamente[32] com diversos agrupamentos, de acordo com as decisões de alinhamento das sentenças que sofreram operação de divisão. Os grupos mais importantes são:

- PSS1 - Todas as divisões: repete a sentença original para cada sentença resultante da divisão;
- PSS2 - Maior divisão: utiliza somente a maior sentença resultante da divisão no alinhamento;
- PSS3 - Somente sentenças que não sofreram operação de divisão.

Como exemplo de aplicação, o agrupamento PSS2 foi utilizado por Leal et al. (2019) e Leal et al. (2020) na tarefa de avaliação da complexidade sentencial. As estatísticas do *corpus* estão listadas na Tabela 6.4.

Tabela 6.4: PorSimplesSENT - Estatísticas

Grupos	PSS1	PSS2	PSS3
Original-Natural	3.535	2.372	1.543
Natural-Forte	4.976	1.501	782
Original-Forte	2.105	1.095	275
TOTAL	10.616	4.968	2.600

6.6 Uso de Modelos de Língua para a Simplificação Textual

Nos últimos anos tivemos uma evolução interessante nas tarefas de PLN com o treinamento de grandes modelos de língua, nessa seção abordamos de forma bem preliminar a utilização desses modelos nas tarefas de avaliação e tratamento da complexidade.

Na primeira versão deste capítulo, fizemos um pequeno experimento utilizando um dos textos do *corpus* PorSimples. Solicitamos ao Bard da Google[33], ao Copilot da Microsoft[34] e ao ChatGPT 3.5 da OpenAi[35] para simplificar esse texto, e também para explicar as operações de simplificação que foram utilizadas. Um ano depois os modelos evoluíram bastante, repetimos o experimento com o Gemini 2.5 Pro[36] e com o ChatGPT 5 Pro[37].

[32] http://www.nilc.icmc.usp.br/nilc/index.php/tools-and-resources e https://github.com/sidleal/porsimplessent
[33] https://bard.google.com/
[34] https://www.microsoft.com/pt-br/microsoft-copilot
[35] https://openai.com/blog/chatgpt
[36] https://gemini.google.com/
[37] https://chatgpt.com/

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Quadro 6.4: Exemplos de simplificação utilizando modelos de língua

Original	O ano era 1978. As salas de cinema de todo o mundo exibiam uma produção do diretor Joe Dante em que um cardume de piranhas escapava de um laboratório militar e atacava participantes de um festival aquático. Quase 30 anos depois, banhistas assustados estão se afastando do principal balneário de Uruguaiana, na Fronteira Oeste. Mais de 20 pessoas foram mordidas por palometas (Serrasalmus spilopleura, espécie de piranha) que vivem nas águas da barragem Sanchuri, na margem da Br-472, a 40 quilômetros da cidade. - Os ataques se tornaram mais freqüentes. Por isso, aconselhamos mais cautela- diz o subprefeito Nei Pinto. As mordidas em pés e canelas de banhistas não são novidade. O que chamou a atenção das autoridades foi o aumento no número de ataques em relação aos outros anos.
Bard (Gemini 1.0)	No passado, um filme de piranhas aterrorizou as pessoas. Hoje, banhistas de Uruguaiana estão com medo de serem mordidos por palometas. Elas vivem em uma barragem a 40 quilômetros da cidade. Os ataques estão acontecendo com mais frequência. Por isso, devemos ter cuidado - diz o subprefeito Nei Pinto.
ChatGPT (GPT3.5)	Em 1978, um filme sobre piranhas assustou as pessoas em todo o mundo. Agora, em Uruguaiana, muitas pessoas estão sendo mordidas por palometas, que são como piranhas. Essas palometas vivem em um lugar chamado barragem Sanchuri, perto da estrada BR-472. Os ataques estão acontecendo com mais frequência do que antes, então é importante ter cuidado ao nadar lá.
Copilot (GPT4)	Em um filme, peixes que mordem pessoas fogem de um lugar onde os cientistas faziam testes. Muitos anos depois, pessoas que nadavam em um rio foram mordidas por peixes parecidos com os do filme. Os peixes se chamam palometas e vivem em um rio perto de uma cidade. - As pessoas precisam tomar cuidado porque os peixes estão mordendo mais - fala o chefe da cidade. Os peixes já mordiam as pessoas antes, mas agora estão mordendo mais ainda.
ChatGPT 5 Pro	Há muitos anos, um filme mostrou piranhas atacando pessoas. Agora, algo parecido está acontecendo de verdade. Num lugar de banho perto de Uruguaiana, no Rio Grande do Sul, mais de 20 pessoas foram mordidas por palometas. Palometas são peixes parecidos com piranhas. Esses peixes vivem na barragem Sanchuri e mordem, principalmente, os pés e as pernas de quem entra na água. A prefeitura pediu que todos tomem cuidado, porque os ataques ficaram mais frequentes.
Gemini 2.5 Pro	Em uma cidade chamada Uruguaiana, as pessoas que vão nadar em um rio estão com medo. É que um tipo de piranha, chamada palometa, está mordendo os banhistas. Mais de 20 pessoas já foram mordidas nos pés e nas pernas. Uma autoridade da cidade disse que os ataques estão acontecendo mais vezes e pediu para todo mundo tomar mais cuidado.

Os resultados completos estão disponibilizados no github llm-simplification-

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

experiment[38].

O *prompt* fornecido aos modelos foi: **"Simplifique o seguinte texto para que um aluno do quarto ano do ensino fundamental consiga compreender, após a simplificação, forneça passo a passo os detalhes das mudanças e motivos para fazer as adaptações no texto:"**, seguido do texto. O texto de teste precisou ser truncado no final do parágrafo anterior antes de completar 2000 caracteres, que é o limite do *prompt* nas interfaces testadas.

Além das simplificações e explicações dos modelos, deixamos também no github as duas simplificações feitas por humanos do PorSimples, para comparação. Reproduzimos no Quadro 6.4 os três primeiros parágrafos do texto original e a simplificação resultante para fins ilustrativos.

De forma geral todos os modelos testados tiveram resultados satisfatórios na tarefa de simplificação proposta, porém o Copilot optou por não situar o texto geograficamente e temporalmente (Em 1978/No passado), eliminando as entidades nomeadas da cidade onde aconteceu o incidente (Uruguaiana), por exemplo. Esta decisão deixou o resultado um pouco superficial.

O Bard e o ChatGPT 3.5 fizeram duas simplificações diferentes, mas ambas com boa qualidade. De forma interessante, optaram por manter entidades nomeadas diferentes, o Bard manteve o nome do subprefeito, enquanto o ChatGPT manteve o nome da barragem e da estrada.

Nas explicações das operações de simplificação utilizadas (que podem ser conferidas integralmente no link do github disponibilizado), o Bard explicou parágrafo por parágrafo, enquanto o ChatGPT focou nas operações utilizadas. Aqui cabe destacar que o Bard teve uma "alucinação" no momento da explicação e inventou uma operação que não aconteceu realmente.

O ChatGPT 3.5 explicou o termo palometas, usando uma exemplificação: "palometas, que são como piranhas", incluindo porém uma interpretação discutível, pois o texto original diz que palometas são uma espécie (no sentido biológico) de piranhas.

O GPT 5 Pro mostrou grandes avanços na simplificação, mas ainda mantém a interpretação que "palometas são peixes parecidos com piranhas". Outro ponto interessante dessa simplificação foi a inclusão de um glossário ao final (que pode ser visto no exemplo completo no github).

O Gemini 2.5 Pro manteve a tendência de resumir bastante, mas acertou que a palometa é um tipo de piranha. O resumo removeu a maioria das entidades nomeadas, mas justificou bem as mudanças.

Com este pequeno experimento, fomos surpreendidos positivamente pelos resultados dos grandes modelos de língua na tarefa de simplificação textual. Porém também ficou claro que ainda é necessária uma revisão humana desses resultados, pelo menos por enquanto.

6.7 A complexidade como parâmetro de descrição da linguagem

Até o presente momento, discutimos diversas aplicações de métricas de complexidade textual, seja como meios de classificar a adequação de um texto a um determinado público (como na fórmula *Flesch-Kincaid grade level*), seja como uma etapa no processo de adequá-lo ao público-alvo (como na ferramenta Simplifica), seja como forma de diagnóstico de

[38] https://github.com/sidleal/llm-simplification-experiment

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

distúrbios ou fatores que afetam a pessoa que produz a linguagem (como no Coh-Metrix-Dementia).

Um aspecto comum a todos esses casos é que o objeto central da nossa exploração é o falante – o indivíduo que utiliza a linguagem natural – e como uma métrica de complexidade pode ser um elemento na construção de tecnologias da linguagem que permitam auxiliar o indivíduo ou acessar informações sobre ele.

Entretanto, elementos externos à pessoa também influenciam a maneira como ela se expressa, como:

- a própria estrutura da comunicação humana: o fato de a pessoa querer ser entendida, de que ela tem que pensar e raciocinar enquanto está falando ou escrevendo e de que suas ideias podem mudar no processo;
- a língua utilizada: quais estruturas aquela língua oferece para construir a fala ou sua escrita, quais regras e restrições a língua impõe, e que outras línguas poderiam não impor;
- a situação em que a pessoa se encontra: as regras e convenções que definem como ela deve adaptar o uso da linguagem a contextos específicos.

Todos estes elementos representam níveis em que a linguagem humana é capaz de variar e sobre os quais agem diferentes forças ou restrições que influenciam a maneira como essa variação ocorre. Além disso, estas variações causam modificações palpáveis na complexidade do texto.

Por isso, podemos pensar na complexidade como um fator ou parâmetro da variação da linguagem que, se examinado, nos permitiria entender melhor o próprio fenômeno da variação linguística.

Ao observarmos, pela lente da complexidade, as dimensões em que a linguagem não varia, podemos deduzir coisas sobre a estrutura básica e comum da linguagem humana, subjacente às múltiplas formas em que essa habilidade se realiza, revelando leis e restrições que agem sobre o fenômeno da comunicação humana como um todo.

Assim, as mesmas métricas de quantificação da complexidade textual que nos permitem desenvolver tantas tecnologias importantes em benefício dos seres humanos, podem, também, nos permitir entender melhor o espaço sobre o qual essas tecnologias são desenvolvidas: a própria linguagem humana. Essa forma de pensar está intimamente ligada a duas áreas das ciências da linguagem: a **tipologia** e a **sociolinguística**.

A **tipologia** é a área do conhecimento que se debruça sobre a comparação estrutural entre as diferentes línguas e como a variação da competência da linguagem humana se dá no nível interlinguístico. A tipologia objetiva mapear e entender quais são os universais da linguagem, comuns a todas as línguas humanas, e modelar quais forças definem como e quais variações são possíveis sobre essa estrutura subjacente (Croft, 2002).

Ela ganhou sua forma moderna, principalmente, após o trabalho do linguista estadounidense Joseph H. Greenberg (Greenberg, 1966), no qual, ao comparar a estrutura gramatical de dezenas de línguas, ele mapeou um conjunto de universais linguísticos, comuns a todas – ou à maioria – das línguas da sua amostra. A tipologia está intimamente relacionada a uma visão funcionalista ou teleológica da linguagem, que a descreve a partir da sua função comunicativa. Mais recentemente, inclusive, muitos passaram a considerar a tipologia não como uma área da linguística, mas como uma abordagem à linguística, na sua versão fundida ao pensamento funcionalista, constituindo a chamada abordagem tipológico-funcional à linguagem (Croft, 2002).

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

A **sociolinguística**, por sua vez, é o ramo das ciências da linguagem interessado em como a linguagem varia e se modifica em diferentes contextos e situações sociais, indo além de como a gramática oficial da língua é definida, e buscando descrever como os diferentes grupos de uma sociedade realizam a língua em diferentes contextos, caracterizando as diversas forças que determinam e influenciam essa variação (Wolfram, 2006).

O **estudo dos registros**, em especial, é uma linha fortemente conectada à sociolinguística, que busca descrever como a linguagem varia entre registros – definidos aqui como variações da língua em função das características e parâmetros da situação em que ela é produzida (Biber, 2006).

Se a tipologia se debruça sobre a variação linguística no nível interlinguístico, a sociolinguística foca na variação no nível intralinguístico. Como veremos, entretanto, a complexidade parece ser um fenômeno transversal a esses níveis, embaraçando o limite entre essas duas áreas.

6.7.1 Questões sobre a complexidade como parâmetro

Um dos pioneiros do uso da complexidade linguística como parâmetro de variação, no sentido moderno, foi o linguista estado-unidense Charles F. Hockett, que, em (Hockett, 1958), apresentou duas conjecturas, nas quais a complexidade desempenha um papel central como parâmetro de variação: a hipótese da equicomplexidade e a hipótese do *trade-off* morfossintático:

- **Hipótese da Equicomplexidade**: todas as línguas humanas possuem a mesma capacidade comunicativa e a mesma complexidade geral, mesmo usando de mecanismos distintos para transmitir informações e construir estruturas complexas.
- **Hipótese do *trade-off* Morfossintático**: quanto mais complexa uma língua é morfologicamente – i.e. quanto mais informação ela transmite através de estruturas internas à palavra – menos complexa ela será sintaticamente – i.e. menos informação ela transmitirá através de estruturas externas às palavras.

Na visão de Hockett, a hipótese do *trade-off* morfossintático seria o principal mecanismo através do qual as línguas humanas manteriam a equicomplexidade.

Essas conjecturas exerceram um papel essencial no seu tempo, por virem como resposta a concepções prévias da linguagem bastante problemáticas, que caracterizavam as línguas indo-europeias como o pináculo de um dito processo evolutivo de complexidade, em detrimento das manifestações linguísticas de outros povos.

No mesmo compasso, várias outras conjecturas acerca da complexidade linguística, em especial da complexidade textual, foram surgindo após as ideias originais de Hockett, entre elas, podemos mencionar[39]:

- **Hipótese da conexão demográfica**: a complexidade média da linguagem produzida por uma determinada população refletiria a distribuição de falantes primários e secundários da língua naquela população.
- **Hipótese da simplificação mediante contato**: manifestações linguísticas que surgem do contato entre populações que falam línguas diferentes apresentariam menor complexidade geral, justamente como forma de facilitar a comunicação entre essas populações (McWhorter, 2001).

[39]Veja (Ehret et al., 2023) para uma discussão mais aprofundada sobre proposições em aberto.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

- **Hipótese dos fatores geográficos e filogenéticos**: línguas com relações de parentesco, ou línguas geograficamente próximas, com interação frequente, apresentariam semelhanças nos seus perfis de complexidade – i.e. na maneira como distribuem sua complexidade em diferentes níveis e estratégias.

Entretanto, a avaliação objetiva dessas conjecturas seria muito difícil sem métricas que permitissem de alguma forma quantificar a complexidade de uma língua ou do produto de uma língua, como uma fala, um texto ou uma sentença. Daí, vários pesquisadores se debruçaram sobre como definir e mensurar a complexidade linguística.

Trabalhos como os de Nichols (1998), McWhorter (2001) e Dahl (2004) foram pioneiros na definição de tais métricas, entretanto suas abordagens usualmente exigiam a anotação manual por especialistas, a partir de grandes amostras textuais ou extensas gramáticas de cada língua.

Contudo, fatores como (i) o desenvolvimento da linguística computacional e do PLN, (ii) o avanço do poder de processamento computacional e (iii) a democratização da internet e o consequente aumento do acesso a textos de mais línguas através de meios digitais, mudaram o cenário de pesquisa, levando à proposição de métricas de complexidade computacionais.

Essas novas métricas trouxeram a vantagem de poderem ser computadas de forma mais rápida e sem tanta dependência da anotação de especialistas, além de permitirem comparação em grande escala de diversas línguas. Com esses avanços, aumentaram as nossas possibilidades científicas do uso da complexidade como um parâmetro de estudo da linguagem, em especial para a análise e verificação das hipóteses como as que listamos nesta seção.

6.7.2 Métricas de Complexidade Baseadas em Compressibilidade

Atualmente, contamos com diversas métricas computacionais de complexidade de linguagem. Muitas das medidas apresentadas anteriormente neste capítulo servem também ao propósito do estudo da complexidade como parâmetro linguístico, da forma que o apresentamos nesta seção.

Entretanto, ainda estaria fora do nosso alcance apresentar aqui todas as métricas existentes e os tantos experimentos e resultados obtidos através delas. Para dar um vislumbre do tipo de resultados e conclusões que podemos obter nessas investigações, apresentaremos brevemente alguns resultados obtidos a partir das métricas baseadas em compressibilidade, uma família de métricas de complexidade textual que tem apresentado resultados bastante interessantes e prototípicos, apesar de serem apenas uma parcela muito pequena do conjunto de métricas que podemos considerar.

As métricas de complexidade baseadas em compressibilidade são um conjunto de métricas de complexidade textual, desenvolvidas em (Juola, 1998), (Juola, 2008) e (Ehret; Szmrecsanyi, 2016). Essas métricas consideram que a complexidade de um texto é a quantidade de informação nele contida. A complexidade em um nível linguístico – como morfológico ou sintático – é, então, a parte da informação total do texto sendo transmitida através das estratégias daquele nível específico.

Para poder acessar a quantidade de informação contida no texto e em cada um dos níveis de interesse, essas métricas utilizam algoritmos de compressão de dados simples, como o *gzip*. Esses algoritmos são capazes de perceber padrões de recorrência dentro de um conjunto de dados, reduzindo-o a uma quantidade de informação mínima, a partir da qual o arquivo original pode ser reconstruído. As métricas baseadas em compressibilidade

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

tomam, então, o tamanho de um texto comprimido como uma aproximação da quantidade de informação – e, portanto, da complexidade – nele contida.

Para aproximar a quantidade de informação em níveis específicos, realizam-se perturbações automáticas e aleatórias no conteúdo do texto, que visam afetar os padrões e estratégias através dos quais a informação é transmitida no nível de interesse, medindo o impacto gerado na quantidade de informação total do texto.

Quando aplicadas ao mesmo texto traduzido em diferentes línguas – visando verificar a hipótese do *trade-off* morfossintático – essas métricas nos apresentam resultados bastante interessantes. As Figuras 6.11 e 6.12 mostram uma renderização do *trade-off* em dois experimentos distintos. O primeiro, realizado em Juola (2008), mostra uma troca clara entre as complexidades sintática e morfológica, como previsto por Hockett, para algumas línguas europeias.

Já o gráfico da Figura 6.12, extraído de (Serras et al., 2024), mostra o mesmo *trade-off* para um conjunto mais amplo, composto majoritariamente de línguas nativas da América do Sul, com algumas línguas europeias de referência. Aqui, também é perceptível uma tendência inversa entre as complexidades sintática e morfológica, entretanto a dispersão é muito maior. Talvez isso possa ser explicado pela grande diversidade filogenética desse conjunto – de fato, é possível perceber que muitos dos dados parecem estar agrupados por semelhança filogenética – o que pode ser lido como um indício da hipótese dos fatores geográficos e filogenéticos, apresentada anteriormente.

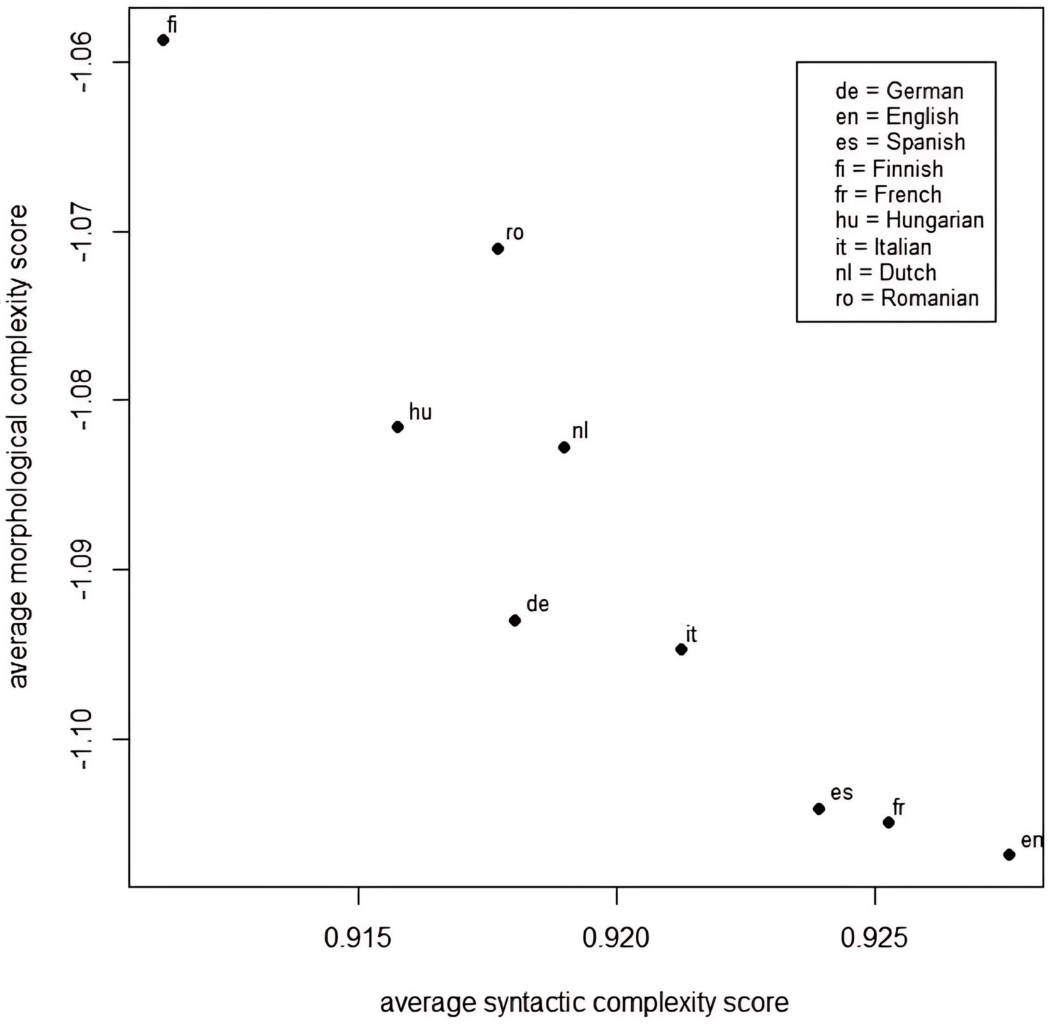
Um outro resultado interessante é que, quando as mesmas métricas são computadas sobre conjuntos de textos originários da mesma língua, mas de diferentes registros dentro dessa língua, i.e., sobre textos gerados em situações específicas distintas, observamos o mesmo padrão do *trade-off* morfossintático. Isso fica claro no gráfico da Figura 6.13, extraído de (Ehret, 2021). Nesse gráfico, vemos o *trade-off* morfossintático para vários registros ou gêneros distintos da língua inglesa. Também é possível ver uma separação clara entre registros mais e menos oralizados, no que tange ao seu perfil de complexidade.

Apesar de breve, essa comparação de resultados não só nos mostra a robustez dessas métricas, mas a capacidade que a quantificação da complexidade nos traz, ao permitir examinar hipóteses até então de difícil verificação e conectar fenômenos em escopos bastante distintos. Essas conexões talvez permitam que, nos próximos anos, construamos modelos cada vez mais profundos da linguagem e de como ela funciona em diferentes níveis. Talvez, esses novos modelos possam, inclusive, acarretar futuras melhorias nas tecnologias relacionadas à complexidade textual, apresentadas nas seções anteriores deste capítulo.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Figura 6.11: Gráfico apresentando o *trade-off* entre complexidade morfológica e sintática, usando métricas de complexidade de linguagem baseadas em compressibilidade, computadas sobre um conjunto de traduções paralelas de textos bíblicos em várias línguas europeias.

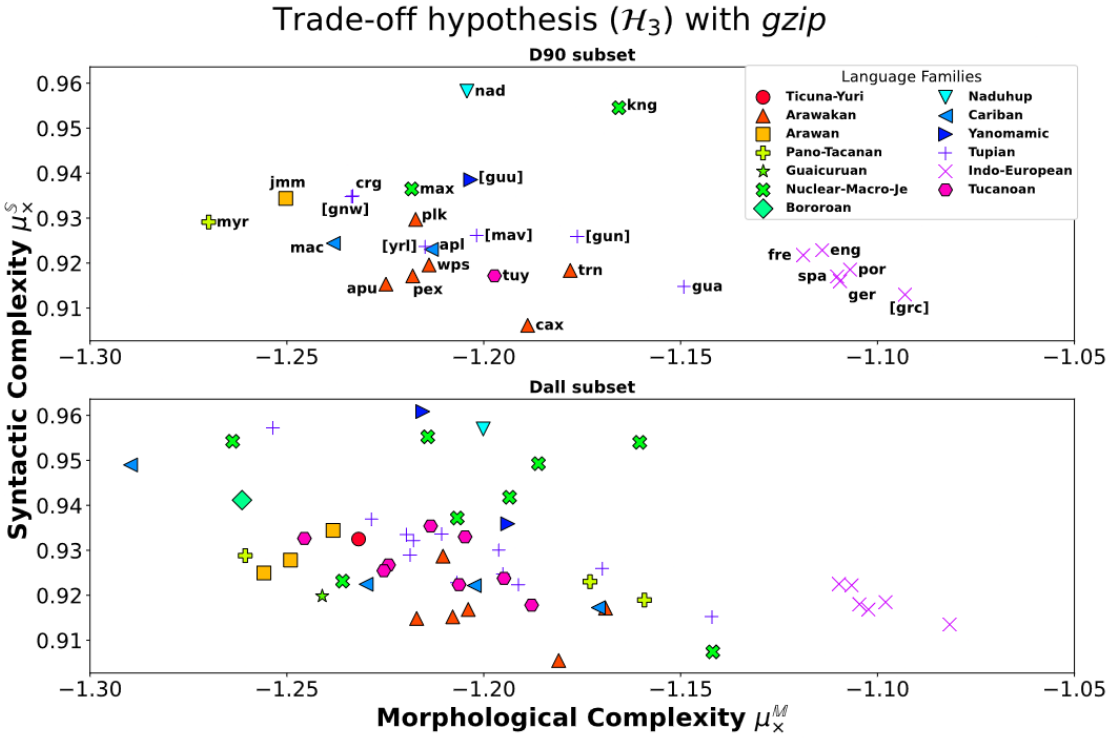


Fonte: (Juola, 2008)

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Figura 6.12: Gráfico apresentando o *trade-off* entre complexidade morfológica e sintática, usando métricas de complexidade de linguagem baseadas em compressibilidade, computadas sobre um conjunto de traduções paralelas de textos bíblicos em diversas línguas, principalmente nativas da América do sul. Os dois painéis representam técnicas distintas de seleção dos dados.

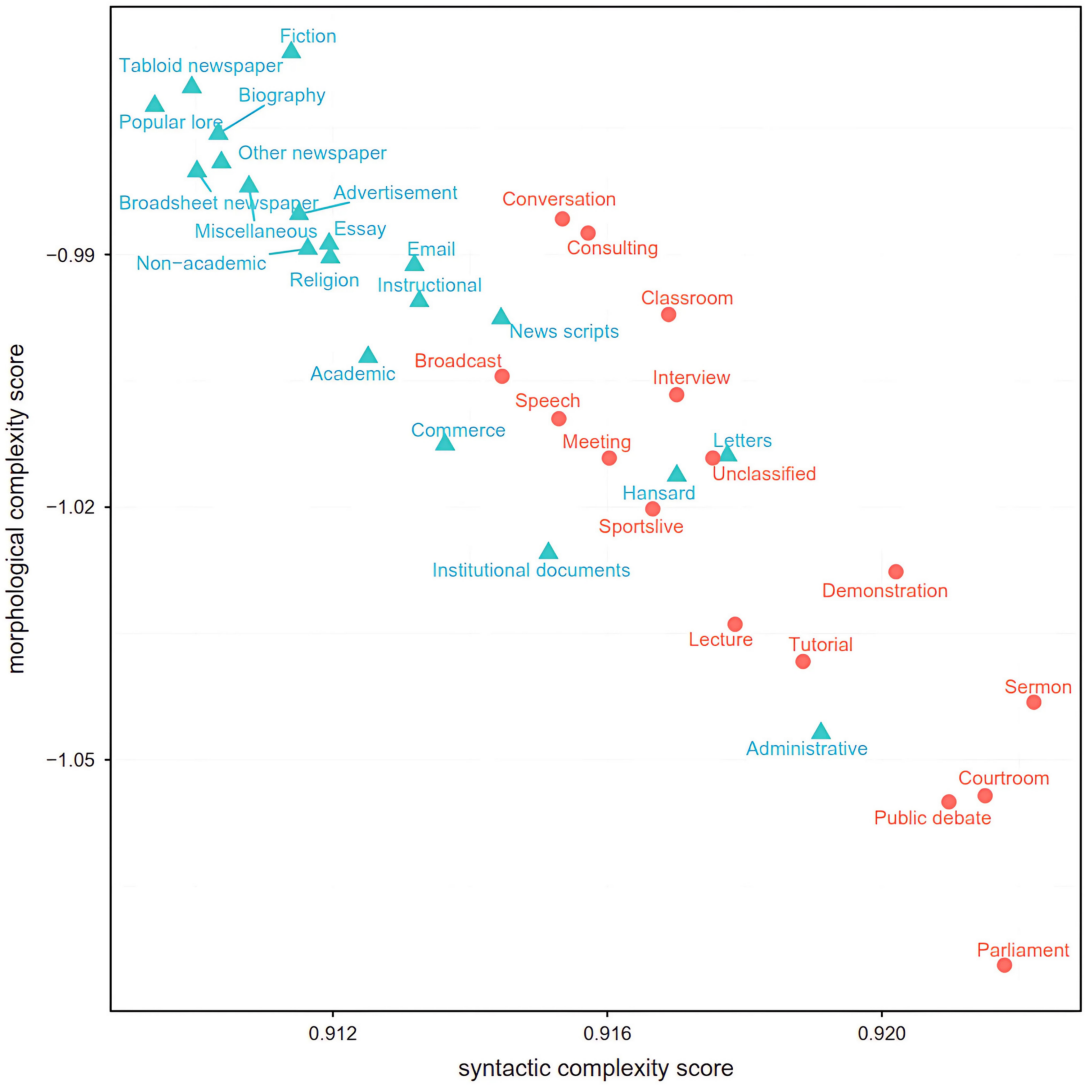


Fonte: (Serras et al., 2024)

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Figura 6.13: Gráfico apresentando o *trade-off* entre complexidade morfológica e sintática, usando métricas de complexidade de linguagem baseadas em compressibilidade, computadas sobre um conjunto de textos de vários registros ou gêneros distintos da língua inglesa. Os pontos triangulares azuis indicam registros/gêneros escritos e os pontos circulares alaranjados indicam registros/gêneros falados.



Fonte: (Ehret, 2021)

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

6.8 Considerações finais

Neste capítulo, apresentamos uma visão geral sobre o tema complexidade textual, começando pelas tarefas relacionadas, passando pelas diversas métricas disponíveis para avaliação automática, pelos recursos disponíveis atualmente para a avaliação das tarefas no português brasileiro e apresentamos também como grandes modelos de linguagem podem ser utilizados para a tarefa de simplificação textual. Por fim, apresentamos como a medição da complexidade textual pode ser usada como parâmetro na construção de modelos sobre a linguagem humana.

Como os grandes modelos de linguagem impactaram em várias tarefas do PLN nos anos recentes, apresentamos também como eles podem ser utilizados para a tarefa de simplificação textual.

Esperamos que esse capítulo sobre a complexidade textual e suas tarefas relacionadas desperte a curiosidade de mais alunos e pesquisadores e contribua de forma positiva para o desenvolvimento da grande área de PLN. Esperamos também que no futuro sejam criadas ferramentas incríveis para adequação dos textos para os leitores, mantendo sempre o nível de desafio adequado para o entendimento e motivação.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Capítulo 7

Correção Automática de Redação

Amanda Pontes Rassi
Priscilla de Abreu Lopes

Publicado em: 26/09/2023
Atualizado em: 16/04/2026

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

7.1 Introdução

A Correção Automática de Redação (CAR) é uma das diversas aplicações do Processamento de Linguagem Natural (PLN). De forma geral, ela pode ser definida como "o processo de avaliação e atribuição de nota em textos escritos em prosa, via programas computacionais" (Shermis; Burstein, 2013) [1].

Embora a correção manual de redações seja uma prática bastante antiga – os primeiros sistemas automáticos para essa finalidade surgiram na década de 60, para a língua inglesa – só mais recentemente foram adaptados para o português.

Em inglês, as áreas de *Automated Essay Scoring* (AES) e *Automated Essay Evaluation* (AEE) surgem como distintas, porém complementares e, às vezes, com alguma intersecção. A AES tem como principal objetivo automatizar a atribuição de nota, enquanto a AEE amplia esse foco, incluindo também a geração de feedbacks para o aluno, contribuindo de forma mais direta com o processo de aprendizagem da escrita.

A AES costuma ser traduzida para o português como **Avaliação Automática de Redação** (AAR) (Bittencourt Jr., 2020; Da Silva Jr., 2021; Lima et al., 2023), enquanto a AEE está associada ao termo **Correção Automática de Redação** (CAR), apesar de se tratar de um falso cognato. Neste capítulo, adotamos o termo CAR para nos referirmos a uma solução mais abrangente, que contempla tanto a avaliação quanto o feedback, isto é, que reúne as funções de AES e AEE. Para que seja considerada como solução completa de CAR, a aplicação deve contemplar pelo menos três etapas básicas:

i) a detecção de desvios no texto;
ii) a atribuição da nota, seja ela global ou por critério; e
iii) um feedback para o aluno.

Cada uma dessas etapas pode ser vista como uma aplicação independente no PLN. Por exemplo, existem várias ferramentas de auxílio à escrita, bem como corretores ortográficos e gramaticais, que executam exclusivamente a tarefa de identificação de desvios no texto; e isso constitui uma aplicação em si. Da mesma forma, o fornecimento de sugestões de melhoria se aproxima de aplicações de geração de linguagem natural (ou *Natural Language Generation*).

[1] Tradução nossa. Do original: "*the process of evaluating and scoring written prose via computer programs*".

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Ainda que essas funcionalidades possam ser oferecidas de forma independente, consideramos que uma solução pedagógica robusta de CAR deve integrar as três etapas mencionadas, as quais serão exploradas com mais profundidade ao longo deste capítulo.

Antes de abordar cada uma das etapas, porém, faremos uma breve explicação sobre o objeto de estudo da CAR, que é a redação escolar, definindo e exemplificando os principais gêneros e tipos textuais, os critérios avaliados e alguns modelos brasileiros de correção de redação.

7.1.1 O que é uma redação escolar?

A redação escolar é considerada um gênero textual, mas também pode ser distribuída em vários tipos e gêneros textuais. As redações, ou textos[2] de redação escolar, são geralmente utilizadas para avaliar as habilidades de escrita, interpretação, argumentação e criatividade dos alunos, bem como para desenvolver o pensamento crítico e a capacidade de expressão escrita. As redações podem abordar temas diversos, desde assuntos cotidianos até questões mais complexas e abstratas, e são uma forma importante de avaliar o progresso dos alunos ao longo do tempo.

Para fins didáticos e de correção de redação, é importante salientar a diferença entre tipo textual e gênero textual, já que as redações devem atender a um tipo específico e a algum gênero específico, a depender da proposta de redação. Por exemplo, a redação do Enem[3] é sempre do tipo argumentativo e do gênero dissertação-argumentativa.

Os **tipos textuais** (ou "modos textuais", para Marcuschi (2008, p. 154)) se referem à forma como o texto é organizado, ou seja, a sequência linguística e os aspectos lexicais, sintáticos, tempos verbais, relações lógicas que são mobilizados para constituir o texto. Existe um conjunto bastante limitado de tipos textuais, o qual abrange: narração, argumentação, descrição, exposição e injunção. Ressaltamos que um texto raramente apresenta apenas características de um mesmo tipo. Deste modo, classificamos um texto como sendo de um determinado tipo quando há predominância de elementos que o caracterizam.

Já os **gêneros textuais** são formas de comunicação que se desenvolvem em diferentes contextos sociais e culturais e se caracterizam pelo seu propósito ou objetivo comunicativo. Em outras palavras, cada gênero tem uma finalidade específica e é utilizado em determinadas situações comunicativas, dependendo de fatores sociais, culturais, dos falantes, da relação entre eles, do contexto, da finalidade da comunicação, dentre vários outros. Dependendo da situação comunicativa, cada gênero pode exigir um registro ou vocabulário específico, a norma culta ou coloquial, na modalidade escrita ou oral da língua.

Por esse motivo, os gêneros são mais fluidos, podendo surgir, modificar-se, mesclar com outros, desaparecer e reaparecer com outra roupagem em outro contexto ou época. São exemplos de gêneros textuais: bula de remédio, carta pessoal, diálogo informal, e-mail, edital de concurso, inquérito policial, piada, receita culinária, reportagem, resenha, sermão etc.

As redações podem apresentar diversos formatos e objetivos, dependendo do nível de

[2] Existe uma longa discussão conceitual e técnica sobre a definição do termo "texto" em Linguística Textual. Para o propósito deste capítulo, adotaremos como conceito de "texto" um conjunto de palavras e frases organizadas de forma coerente e coesa, com o objetivo de transmitir uma mensagem ou ideia. Em outras palavras, o texto é uma unidade de linguagem que tem um sentido completo e pode ser compreendido em um contexto específico.

[3] O Exame Nacional do Ensino Médio (Enem) é uma prova do Governo Federal que avalia o desempenho escolar dos alunos ao término do Ensino Médio. Essa prova avalia várias áreas do conhecimento e também a produção de uma redação.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

ensino e do tema proposto pelo professor ou pela instituição de ensino. Dentre os tipos e gêneros textuais mais comuns associados à redação escolar, convém mencionar:

- **Dissertação**, em que o autor apresenta e disserta sobre um determinado tema, apresentando informações e argumentos relacionados ao assunto;
- **Narração**, em que o autor conta uma história, relatando fatos e acontecimentos em alguma ordem que pode ser cronológica ou não;
- **Carta**, em que o autor se dirige a um destinatário específico, fazendo requisições, solicitações ou expressando suas opiniões e sentimentos;
- **Artigo de opinião**, em que o autor defende um ponto de vista sobre um tema específico, utilizando argumentos e evidências para sustentá-lo;
- **Resenha**, em que o autor faz uma análise crítica de um texto, obra ou produto.

Esses são apenas alguns dos tipos e gêneros de redação escolar mais comuns. Os demais incluem a descrição, a exposição, a crônica, o relatório, o conto, a fábula, entre outros.

7.1.2 O que é avaliado?

Vários aspectos do texto são avaliados em uma correção de redação, tais como o uso da norma padrão da língua portuguesa, a adequação ao tema e ao gênero, questões relacionadas à coesão, à coerência, à progressão textual etc. Cada modelo de correção organiza e nomeia seus critérios de avaliação de formas distintas, mas, basicamente, todos eles analisam:

Língua portuguesa Avalia a linguagem usada para expressar o conteúdo, verificando se há desvios ortográficos e/ou gramaticais, se a norma (culta ou coloquial) está de acordo com o tipo de texto exigido, se há problemas de estrutura sintática nas frases, orações e períodos, se o vocabulário foi usado adequadamente etc. Outros nomes para esse critério incluem “Escrita”, “Modalidade escrita”, “Norma culta”, “Norma padrão”, “Correção gramatical e adequação vocabular” ou “Expressão (modalidade)”.

Tema Esse aspecto avalia a adequação da redação em relação à temática proposta, verificando se a abordagem do tema foi completa, se tangenciou ou fugiu do tema proposto, se o abordou de forma superficial ou profunda etc. Também é chamado de “Abordagem temática”, “Desenvolvimento do tema”, “Proposta temática” ou “Progressão temática”.

Gênero Esse critério considera a adequação da redação em relação ao tipo textual e ao gênero textual exigidos na proposta. Também pode ser chamado de “Gênero textual”, “Adequação ao tipo textual”, “Organização do texto dissertativo-argumentativo” ou “Estrutura (gênero/tipo de texto)”.

Coerência Neste quesito, avalia-se a coerência entre as ideias, a ordem dos argumentos, a profundidade da argumentação, a clareza e autoria das ideias desenvolvidas, assim como verifica-se se há contradições no texto, se as informações são vagas e/ou muito generalistas, se falta informação, dentre outros. Também pode ser chamado de “Progressão textual”, “Defesa do ponto de vista”, “Coerência dos argumentos”, “Estrutura (coerência)”, “Indícios de autoria” e outros termos.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Coesão Avalia o uso correto ou incorreto, presença ou ausência, pertinência ou não de operadores coesivos, tais como conjunções, preposições, pronomes e expressões discursivas. O critério é também chamado de "Coesão e articulação", "Articulação das partes do texto", "Expressão (coesão)", "Conexão entre os parágrafos", "Uso de operadores argumentativos", "Recursos coesivos", dentre outros termos.

Além desses aspectos que são comuns a todos os modelos de correção, alguns professores, instituições de ensino e vestibulares também podem avaliar a "Leitura", ou seja, o uso e interpretação dos textos motivadores ou da coletânea que embasa a proposta de redação, e também a presença e adequação da "Proposta de intervenção", que é um critério exclusivo do Enem.

7.1.3 Alguns modelos brasileiros de correção

O principal modelo de correção de redação, no Brasil, é o Enem, responsável pela avaliação anual de milhões de alunos[4]. Mas também existem outros modelos de correção relacionados a vestibulares e universidades específicas, tais como Fuvest, Unesp, Unicamp, FGV e outros igualmente relevantes. Apesar de haver critérios gerais que são avaliados por todos eles, cada um tem autonomia para definir sua grade específica, os pesos de cada critério e sua própria forma de avaliação.

O Quadro 7.1 apresenta quatro modelos brasileiros de correção relacionados a vestibulares, indicando o gênero textual exigido, seus critérios de avaliação e faixas de nota possíveis.

Quadro 7.1: Modelos de correção de vestibulares.

Modelo	Tipo/Gênero	Critérios avaliativos	Nota mín. critério	Nota máx. critério	Nota global
Enem	Dissertação-argumentativa	Língua Portuguesa (Competência 1)	0	200	1000
		Abordagem temática e adequação ao tipo textual (Competência 2)	0	200	
		Progressão textual e defesa do ponto de vista (Competência 3)	0	200	
		Coesão e articulação (Competência 4)	0	200	
		Proposta de intervenção (Competência 5)	0	200	
Fuvest	Dissertação ou narrativa	Desenvolvimento do tema uso da coletânea e autoria	0	15	50
		Compreensão e atendiemnto da proposta quanto ao gênero e tipo de texto	0	10	

[4]No último Enem 2025, foram mais de 4,8 milhões de inscrições confirmadas: https://www.gov.br/mec/pt-br/assuntos/noticias/2025/julho/enem-2025-mais-de-4-8-milhoes-de-inscritos-confirmados

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

		Recursos linguísticos (coesão e coerência) e progressão textual	0	15	
		Convenções da escrita e adequação vocabular	0	10	
Unesp	Dissertação-argumentativa	Tema	–	–	28
		Estrutura (gênero/tipo de texto e coerência)	–	–	
		Expressão (modalidade e coesão)	–	–	
Unicamp	variados	Proposta temática	0	2	12
		Gênero	0	3	
		Leitura	0	3	
		Articulação escrita	1	4	

No modelo de correção do **Enem**[5], o aluno é avaliado quanto à produção de um texto do tipo dissertativo-argumentativo para um tema específico, que muda todo ano. A avaliação é dividida em 5 competências (critérios avaliativos), cada uma no intervalo de notas de 0 a 200. A soma direta das notas das competências leva à nota total, que fica no intervalo de 0 a 1000. Considerando os critérios básicos descritos na Seção 7.1.2, vale dizer que o Enem os divide da seguinte forma: (i) Língua Portuguesa, (ii) Tema e Gênero, (iii) Coerência e (iv) Coesão. Além desses, o modelo ainda avalia um quinto critério, que é a presença e adequação da "Proposta de Intervenção", que consiste na sugestão de ação ou medida interventiva para solucionar ou minimizar o problema associado ao tema proposto.

O vestibular da **Fuvest**[6] também exige um texto do gênero dissertativo-argumentativo para um tema específico que muda todo ano. A partir do vestibular de 2026, o candidato também tem mais uma opção de gênero (narrativo), além da dissertação, ambos sobre o mesmo tema. O modelo de avaliação até 2024 agrupava os critérios básicos em 3 (Tema/Gênero, Coerência/Coesão e Língua Portuguesa). A partir do vestibular 2026, essa grade está dividida em 4 critérios: (i) Desenvolvimento do tema, uso da coletânea e autoria, que vale de 0 a 15 pontos; (ii) Compreensão e atendimento da proposta quanto ao gênero e tipo de texto, que pode somar até 10 pontos; (iii) Recursos linguísticos (coesão e coerência) e progressão textual, também de 0 a 15 pontos; e (iv) Convenções da escrita e adequação vocabular, com pontuação máxima igual a 10. Com a soma de pontos dos 4 critérios, o aluno pode tirar de 0 a 50.

Já no modelo de correção da **Unesp**[7], os textos, que devem seguir o gênero dissertação-argumentativa, também são avaliados em três eixos, agrupados da seguinte forma: (i) Tema, (ii) Gênero/Coerência e (iii) Língua Portuguesa/Coesão. A pontuação individual ou peso por critério não é divulgado no material do candidato, mas é definido que a pontuação final fica entre 0 e 28 pontos.

A redação da **Unicamp**[8], a cada ano, varia a exigência dos tipos e gêneros textuais[9],

[5]Cartilha do participante do Enem 2022: https://download.inep.gov.br/download/enem/cartilha_do_participante_enem_2022.pdf.

[6]Manual do candidato do vestibular Fuvest 2023: https://www.fuvest.br/wp-content/uploads/fuvest2023_manual_candidato_retificado_29112022.pdf.

[7]Manual do candidato do vestibular Unesp 2023: https://documento.vunesp.com.br/documento/stream/MzQxOTk5NA%3d%3d.

[8]Manual do ingresso (https://www.comvest.unicamp.br/wp-content/uploads/2023/02/Manual_do_Ingresso_2023_Atualizado.pdf) e grade de redação (https://www.comvest.unicamp.br/vestibular-2023/grade-da-redacao/) do vestibular 2023 da Unicamp.

[9]Além da dissertação, outros gêneros textuais já exigidos pela Unicamp são: síntese e carta-convite (2015),

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

geralmente oferecendo duas alternativas das quais o candidato deve escolher uma para execução. A Unicamp agrupa os critérios básicos da seguinte forma: (i) Tema, (ii) Gênero e (iii) Língua Portuguesa/Coerência/Coesão. Além desses três eixos, também avalia a "Leitura", que corresponde à leitura e interpretação crítica dos textos fornecidos na proposta, sem contudo copiá-los ou parafraseá-los. Na avaliação, cada critério possui pesos diferentes: Tema varia entre 0 e 2 pontos, Gênero entre 0 e 3, Leitura entre 0 e 3 e Língua Portuguesa/Coerência/Coesão varia entre 1 e 4 pontos. A soma dos pontos de cada critério leva à nota final, cujo valor máximo é de 12 pontos.

Fora as diferenças já apontadas, convém ressaltar que todos os modelos penalizam o aluno (zerando a redação) no caso de falhas graves. No entanto, cada modelo define um conjunto específico de falhas graves, que podem ser: fuga ao tema, fuga ao gênero, assinatura na prova, desenho ou sinal gráfico, redação em língua estrangeira, caligrafia ilegível, recado para o corretor, parte desconectada, texto insuficiente, dentre outras situações[10].

Quanto ao título da redação, o Enem e a Unesp não exigem, mas também não proíbem; simplesmente desconsideram para a avaliação da redação. Já a Fuvest não menciona a exigência de título no Manual do candidato do vestibular Fuvest 2024[11], mas coloca como instrução no Caderno de prova[12]. Já a Unicamp pode ou não exigir título, a depender do gênero textual proposto.

Essa grande variedade de modelos de correção é considerada um dos grandes desafios para a CAR, não sendo recomendado treinar um modelo computacional que abarque todos os tipos de correção ou que misture redações dos vários tipos como dados de treinamento para algum modelo. A exigência por modelagens de nota específicas por modelo de correção não impede, no entanto, o reaproveitamento de parte das ferramentas de correção, como a detecção automática de desvios no texto, desde que modelos de correção distintos tenham diretrizes similares para esse tipo de tarefa.

Todas essas questões serão mais detalhadas ao longo deste capítulo, que está organizado da seguinte forma: a Seção 7.2 descreve como fazer a detecção de desvios em textos em português, demonstrando alguns tipos de desvios e os formalismos usados. A Seção 7.3 apresenta os principais trabalhos da literatura que realizam a atribuição da nota para redações em português. A Seção 7.4 demonstra as possibilidades de geração de um feedback para o aluno. Na Seção 7.5, discutimos as vantagens e desvantagens da correção manual e da correção automática, a fim de esclarecer ao leitor que ambas possuem potencialidades, mas também limitações. Por fim, nas Considerações finais (Seção 7.6), retomamos os pontos principais do capítulo.

7.2 Detecção de desvios no texto

Conforme apresentado na Introdução (Seção 7.1), consideramos que uma das etapas da Correção Automática de Redação (CAR) é a detecção ou identificação de desvios[13]. Essa

resenha e texto de divulgação científica (2016), carta argumentativa e texto de apresentação (2017), palestra e artigo de opinião (2018), abaixo-assinado e postagem em fórum (2019), roteiro de podcast e crônica (2020), discurso político e diário (2021), postagem para redes sociais e manifesto coletivo (2022), depoimento sigiloso e convocação (2023) e carta denúncia e discurso (2024).

[10] Para uma descrição completa e exemplos de todos os casos que zeram a redação em cada modelo de correção, sugere-se consultar os respectivos manuais do candidato ou cartilhas do participante.

[11] https://www.fuvest.br/wp-content/uploads/fuvest2024_manual_candidato.pdf

[12] https://acervo.fuvest.br/fuvest/2020/fuv2020_2fase_dia_1.pdf

[13] Adotaremos o conceito de "desvio" como sinônimo de "erro", mas evitaremos esse segundo termo para mitigar a carga negativa, os preconceitos e os julgamentos contidos na palavra "erro".

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

etapa nem sempre é realizada nos trabalhos de Avaliação Automática de Redação, ou, por vezes, os desvios são contabilizados para o cálculo da nota, porém não são apresentados ao aluno.

A detecção desses desvios costuma ser feita por meio de três abordagens distintas: baseada em regras (**abordagem simbólica**), baseada em modelos estatísticos (**abordagem estatística**) e a baseada em LLMs (**abordagem gerativa**). Os sistemas baseados em regras são mais adequados para identificar desvios gramaticais, o que é mais comum de ser cometido por falantes nativos da própria língua, enquanto os estatísticos capturam melhor os desvios de uso, que são erros mais comuns por não-nativos[14]. A abordagem gerativa tem se tornado uma opção muito usual desde a popularização dos LLMs (Subseção Modelos de Linguagem Neurais Modernos) e suas constantes melhorias.

Embora a abordagem simbólica (baseada em regras) seja considerada obsoleta para tarefas mais complexas, ainda é a mais utilizada atualmente para detectar desvios na área de CAR. Para outros tipos de tarefas, modelos estatísticos e neurais performam melhor e são mais escaláveis do que modelos simbólicos. A abordagem gerativa pode ser interessante como ferramenta de correção do texto, porém, ao apontar desvios, as respostas estão sujeitas a alucinações (situação exemplificada no Capítulo ChatGPT, MariTalk e outros agentes de conversação). Apesar do desempenho de modelos estatísticos e LLMs, a abordagem simbólica permite, com alta precisão, mostrar ao aluno o trecho exato que contém o desvio, explicar por que está errado e ainda fazer sugestões de correção.

Para o português, existem recursos disponíveis, tais como o CoGroo[15] e o LanguageTool[16], que são repositórios contendo regras gramaticais para a língua portuguesa. Esses recursos têm versões livres, gratuitas e de código-aberto, com extensão para navegadores web e também acopláveis a editores de texto.

Também há plataformas de correção de redação que desenvolveram seu próprio conjunto de recursos linguísticos e regras gramaticais, o que é uma boa opção quando há um padrão muito claro e estruturado que se possa expressar com regras simbólicas ou expressões regulares, que é o caso dos desvios mais comuns em redações.

Na Seção 7.2.1 caracterizamos alguns dos tipos de desvios mais comuns em redações. Posteriormente, na 7.2.2, apresentamos duas alternativas de formalismo para a definição de regras de detecção de desvios.

7.2.1 Tipos de desvios

Existem diversos tipos de desvios que podem ser marcados em uma redação, como os ortográficos, os gramaticais (ou sintáticos), os de uso de vocabulário ou registro, os desvios no uso de recursos coesivos, dentre outros. Para cada modelo de correção de redação, é possível criar uma taxonomia própria de tipos de desvios que se pretende identificar em um texto.

Ressaltamos que a criação de recursos para esse tipo de tarefa é um processo difícil, moroso e custoso, que depende de especialistas. Deste modo, é importante estabelecer um planejamento criterioso caso seja necessário criar recursos próprios.

Nesta seção exploramos alguns tipos de desvios, por serem os mais comuns, mas é importante esclarecer que os tipos de desvios não se limitam aos indicados neste capítulo.

[14] Para uma explicação detalhada dos vários sistemas que usam cada uma das abordagens simbólica e estatística para detecção de desvios, ver Leacock et al. (2010) e Gamon et al. (2013).
[15] https://cogroo.sourceforge.net/
[16] https://languagetool.org/pt-BR/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

Na Seção 7.2.1.1 descrevemos os desvios ortográficos, na Seção 7.2.1.2, os gramaticais e na Seção 7.2.1.3, outros tipos de desvios mais comuns em redações.

7.2.1.1 Desvios ortográficos

A grande maioria dos desvios ortográficos é facilmente detectável e tratável. O simples uso de um bom dicionário de língua portuguesa já indica quais palavras existem e quais não existem na língua. Portanto, identificar palavras com grafia desviante do léxico é uma tarefa relativamente simples.

O Unitex[17], por exemplo, dispõe de três dicionários muito completos para o português: o **Delas** (com cerca de 75.000 formas canônicas simples), o **Delaf** (com cerca de 9.000.000 entradas de palavras flexionadas) e o **Delacf** (com cerca de 4.000 entradas de formas compostas). Esse recurso pode ser usado como uma primeira etapa de identificação de desvios ortográficos, a fim de identificar palavras que existem no léxico do português e palavras desviantes.

Outros desvios ortográficos se dividem em:

- problemas de falta de sinal gráfico ou uso indevido de acentuação (ex: "prática" x "pratica");
- problemas de capitalização (uso de maiúscula onde deveria ser minúscula ou uso de minúscula onde deveria ser maiúscula);
- grafia incorreta de palavras homônimas ou parônimas (ex: "mas" x "mais" ou "há" x "a");
- problemas de segmentação (uso de hífen, palavras juntas que deveriam ser separadas ou palavras separadas que deveriam ser escritas juntas); e
- desvios com relação à nova ortografia.

Para todos esses casos, a abordagem baseada em regras precisa identificar corretamente o contexto em que a palavra-alvo está inserida. O que torna essa tarefa complexa e nem sempre bem sucedida é que a identificação do contexto linguístico em uma abordagem simbólica muitas vezes depende de um bom *parser* e um bom *tagger*. Conforme foi apresentado em capítulos anteriores (Capítulo Sequência de caracteres e palavras e Capítulo Ferramentas e recursos para o processamento sintático), essas ferramentas nem sempre têm uma ótima performance em português.

7.2.1.2 Desvios gramaticais

Os desvios gramaticais, também chamados de desvios sintáticos, correspondem aos problemas de estrutura sintática, ou seja, nas relações entre as palavras, que podem estar no escopo de uma sentença, um sintagma, um grupo ou uma *string*. Por exemplo, na sentença "As menina dançam", existe um desvio de concordância nominal entre "As" (plural) e "menina" (singular) e/ou um desvio de concordância verbal entre "menina" (singular) e "dançam" (plural).

Além dos desvios de concordância (que correspondem a cerca de 18,9%), também são comuns em redações escolares: os de vírgula e pontuação (44%), de formas verbais (6,8%), pronomes (5,8%), preposições (5,7%), crase (4,2%), segmentação (4,1%), regência (3,4%), conjunções (2,3%), determinantes (2%) e outros (2,3%)[18].

[17] http://www.nilc.icmc.usp.br/nilc/projects/unitex-pb/web/index.html

[18] Esses percentuais foram calculados a partir dos números absolutos da Tabela 8 de Ramisch (2020, p. 76), que anotou os desvios sintáticos em uma amostra de 1.045 redações.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Apesar de os desvios de pontuação e vírgula serem os mais frequentes, são também os mais difíceis de serem tratados, pois em geral a vírgula separa constituintes sintagmáticos, o que exigiria uma análise sintática por constituência, e esse tipo de *parser* é raro para o Português[19].

As regras gramaticais que exigem um contexto linguístico local, por exemplo, para avaliação da crase ou concordância nominal, geralmente funcionam melhor, ao passo que as regras que dependem de um contexto linguístico maior, com macrorrelações de dependência (Capítulo Sequência de caracteres e palavras), ou quando um *token* está muito distante do outro, tendem a performar mal.

7.2.1.3 Outros tipos de desvios

Além dos desvios ortográficos e gramaticais, há outros problemas textuais que podem ser marcados automaticamente em redações, como:

- Desvios de vocabulário e registro – quando o texto apresenta escolhas lexicais inadequadas ao gênero ou à situação comunicativa. Por exemplo, em uma dissertação argumentativa, o uso excessivo de marcas de oralidade ("tipo", "a gente") ou de pronomes em primeira pessoa ("eu acho", "na minha opinião") pode ser considerado um desvio de registro.
- Desvios de coesão – referentes à falta de recursos coesivos ou ao emprego incorreto de conectivos, pronomes ou expressões responsáveis por articular as ideias do texto. Um exemplo recorrente é o uso de "contudo" no sentido de conclusão, ou o uso do pronome "onde" em contextos não locativos ("A época onde…").
- Desvios de adequação ao gênero – quando o aluno não segue as convenções esperadas para o gênero textual proposto. Quando o aluno conta histórias ou traz longos trechos narrativos em dissertações argumentativas, por exemplo, pode configurar desvio.
- Desvios de progressão temática – quando há repetição excessiva de ideias, incoerências internas, rupturas bruscas na argumentação ou inserção de trechos desconectados do tema central, prejudicando a continuidade do texto.

Enfim, dependendo do modelo de correção adotado e da parametrização dos critérios de correção de cada banca, definem-se quais tipos de desvios devem ou não ser marcados no texto. A detecção desses desvios pode ser feita por meio de regras baseadas em listas lexicais, padrões sintáticos ou análise do contexto discursivo. Assim como nos casos ortográficos e gramaticais, as abordagens mais eficazes são aquelas que conseguem levar em conta o contexto em que a palavra ou expressão ocorre, reduzindo falsos positivos. Cabe destacar que, além da marcação de desvios, é igualmente relevante identificar e registrar aspectos positivos presentes na redação. A sinalização de usos adequados e escolhas bem-sucedidas contribui não apenas para reforçar práticas corretas, mas também para justificar a atribuição de notas mais altas nos critérios correspondentes.

7.2.2 Formalismos de regras

Há inúmeras maneiras de escrever regras de forma que o computador consiga lê-las e interpretá-las. Cada ferramenta pode criar seu próprio formalismo e mecanismo de infe-

[19] O *parser* PALAVRAS (Bick, 2000) dispõe de um módulo de análise sintática por constituintes.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

rência, mas também há alguns disponíveis gratuitamente e que podem ser usados para um projeto inicial.

O LanguageTool[20] implementa um mecanismo de inferência para regras formalizadas em XML (*Extensible Markup Language*). Para o português, o software disponibiliza cerca de 2.880 regras abrangendo várias categorias linguísticas, tais como: gramática geral, ortografia, pontuação, capitalização, tipografia, estilo, redundância, palavra composta, semântica, repetição, linguagem informal, uso de pronomes, dentre outras. Essas regras podem ser consultadas via repositório Language Tool Community[21].

Como exemplo, reproduzimos o formalismo de uma regra para identificar redundância quando se escreve "gelo gelado", na Figura 7.1[22].

Figura 7.1: Exemplo de formalismo de regra do LanguageTool

```
1  <rule id="GELO_GELADO" name="Gelo ">
2      <pattern>
3          <marker>
4              <token>gelo</token>
5              <token>gelado</token>
6          </marker>
7      </pattern>
8      <message>
9          Substitua «gelo gelado» por <suggestion>gelo</suggestion>.
10     </message>
11     <example correction="gelo">
12         Vendemos uma pedra de <marker>gelo gelado</marker>.
13     </example>
14 </rule>
```

Neste exemplo, consta o id e o nome da regra (linha 1), seguidos do padrão a ser buscado (linhas de 2 a 7), seguido da mensagem a ser mostrada (linhas 8 a 10), e de um exemplo de uso (linhas 11 a 13).

A complexidade das regras pode variar dependendo da complexidade do problema linguístico ou do padrão a ser buscado. No caso do código na Figura 7.1, o problema linguístico em questão – a redundância – é muito simples, e isso se reflete na simplicidade da regra, a qual procura basicamente dois *tokens*: o primeiro é "gelo", imediatamente seguido do segundo, que é "gelado". Por outro lado, problemas linguísticos mais complexos também exigem regras mais complexas que podem usar lemas, *tokens*, etiquetas morfológicas, morfossintáticas, expressões regulares, relações de dependência, entidades nomeadas, dentre outros.

Ressaltamos que a performance das regras do LanguageTool não é ótima, mas é um recurso útil para quem não quer começar essa tarefa do zero. Considerando que o software possui versão aberta[23], é possível corrigir e definir novas regras usando o mesmo formalismo e avaliá-las por meio da própria ferramenta.

Outra ferramenta que podemos indicar para esse tipo de tarefa é o módulo Python

[20] https://languagetool.org/pt-BR/

[21] https://community.languagetool.org/rule/list?lang=pt

[22] Fonte: regras para português no repositório languagetool (Github) (https://github.com/languagetool-org/languagetool/blob/50c9a5eb145f6289762fc64a2b8773629ca085e1/languagetool-language-modules/pt/src/main/resources/org/languagetool/rules/pt/pt-BR/style.xml#L142-L150).

[23] https://github.com/languagetool-org/languagetool

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

spaCy[24], que implementa três mecanismos para identificação de padrões em textos que podem ser bastante úteis na tarefa de detecção de desvios. Esses mecanismos fazem parte do sub-módulo chamado *Rule-based matching*, que permite a busca por um *token* em determinado contexto (chamado *Matcher*), por uma frase ou sintagma (chamado *Phrase matcher*), ou ainda por relações de dependências entre elementos da sentença (chamado *Dependency matcher*). Eles podem ser usados separadamente ou combinados entre si para garantir melhor acurácia na busca por padrões linguísticos.

Na Figura 7.2, apresentamos código Python que utiliza a classe `Matcher` do spaCy para a definição da regra que identifica a redundância "gelo gelado", que foi reescrita, e executa a busca por padrões em um texto.

Figura 7.2: Exemplo de formalismo da regra "gelo gelado" usando a biblioteca spaCy

```
1  import spacy
2  from spacy.matcher import Matcher
3
4  nlp = spacy.load('pt_core_news_sm')
5  matcher = Matcher(nlp.vocab)
6
7  rule_id = 'GELO_GELADO'
8
9  pattern = [
10     {'LOWER': 'gelo'},
11     {'LOWER': 'gelado'}
12 ]
13
14 message = 'Substitua «gelo gelado» por "gelo"'
15
16 matcher.add(
17     rule_id,
18     [pattern],
19     on_match=lambda matcher, doc, i, matches: print(message)
20 )
21
22 doc = nlp('Gelo gelado gela a garganta.')
23 matches = matcher(doc)
```

O exemplo inicia importando o módulo `spacy` e, especificamente, a classe `Matcher` (linhas 1 e 2). Em seguida um *pipeline* pré-treinado do spaCy para português é carregado (linha 4) e seu vocabulário é utilizado para inicializar uma instância da classe `Matcher` (linha 5). Em seguida, são definidos um identificador para a regra (linha 7), o padrão buscado de dois *tokens* (linhas 9 a 12), cada um representado por um dicionário, e a mensagem que deve ser impressa na tela caso o padrão seja identificado no texto (linha 14). Nas linhas de 16 a 20, a regra é adicionada à instância `matcher`, incluindo a definição de uma função para a impressão da mensagem na tela quando o padrão é encontrado (`on_match`). As linhas 22 a 23 definem um texto para teste de busca do padrão e a execução dessa busca.

O spaCy não conta com um repositório de regras pré-definidas para detecção de desvios. Contudo, por ser uma ferramenta de PLN, disponibiliza uma série de funcionalidades que podem contribuir para essa tarefa de maneira mais simples, i.e. sem a necessidade de

[24] https://spacy.io/

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

alterar a implementação dos mecanismos de busca já disponíveis.

A detecção de desvios no texto é uma etapa importante em CAR, especialmente por indicar e colaborar para a aprendizagem da escrita. Os desvios encontrados podem, inclusive, ser utilizados na etapa de atribuição de nota à redação. Na Seção 7.3 apresentamos as principais abordagens e tendências nessa área, além de citar os principais trabalhos dedicados a redações em português.

7.3 Atribuição de nota

A atribuição de nota a uma redação pode ser feita de forma global, ou seja, uma nota única para a redação inteira, ou por meio de notas individuais para cada critério de avaliação. O principal desafio desse processo está na grande diversidade de matrizes de correção existentes (ver Seção 7.1.3) – que variam quanto ao modelo adotado, à escala e às faixas de pontuação, bem como à quantidade e à distribuição dos critérios avaliativos.

Até recentemente, as abordagens faziam uso de *corpus* rotulado, i.e., conjuntos de redações que já foram avaliadas manualmente e possuem indicação de nota e/ou adequação da redação em relação ao critério avaliado. Desse modo, as técnicas utilizadas para atribuição de nota se enquadram na área de aprendizado supervisionado por classificação ou regressão. Hoje, com a popularização dos LLMs, muitas empresas optam por essa abordagem gerativa, por ser mais rápida e barata, apesar de menos controlável.

O *Project Essay Grade* (PEG) (Ajay et al., 1973) foi uma das primeiras ferramentas estáveis para a atribuição de notas em redações com boa performance dentro do contexto aplicado: redações universitárias curtas em inglês. No entanto, a falta de acesso a computadores foi, por algum tempo, impedimento para o desenvolvimento de outras soluções. Na metade da década de 90, dados os avanços tecnológicos de hardware e software, a área de AES viu um reaquecimento e, desde então, surgiram novos trabalhos consistentemente, inclusive apoiados por abordagens que tiveram ascensão a partir da década de 2010, como *deep learning* e Transformers.

Como mencionado na Seção 7.1, é importante conhecer o contexto e modelo de correção para realizar a atribuição de nota de forma efetiva. A despeito disso, diferentes estratégias podem ser reaproveitadas e combinadas para a avaliação de redações de modelos de correção distintos. Na Seção 7.3.1 apresentamos uma visão geral de técnicas e estratégias para a atribuição de notas em redações. Dada a relevância do Enem para o contexto de redações em português, a Seção 7.3.2 traz trabalhos especificamente voltados para a automatização da avaliação em redações desse modelo de correção.

7.3.1 Como atribuir nota a redações?

A abordagem clássica para atribuição de notas envolve a extração de atributos (*features*) a partir do texto, que são utilizados para descrever redações de um conjunto de treinamento, além da transformação e seleção desses atributos, em um processo nomeado engenharia de atributos. Os dados extraídos servem de entrada para um algoritmo de aprendizado para a geração de modelos capazes de atribuir nota a novas redações a partir dos valores de seus atributos.

A primeira versão do PEG (Ajay et al., 1973) utilizava atributos baseados em contagens de diferentes elementos do texto, categorizadas em: (i) **simples** (e.g. número de adjetivos na redação): redações com mais adjetivos são avaliadas com notas maiores por humanos

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

(relação linear); (ii) **enganosamente simples** (e.g. número de palavras na redação): redações muito curtas são penalizadas, porém, conforme o tamanho da redação aumenta, esse atributo perde importância para atribuição de nota (relação logarítmica); e (iii) **sofisticadas** (e.g. número de palavras que podem representar contextos maiores): o número de conectivos, por exemplo, pode indicar a complexidade de uma sentença.

Considerando a tarefa de atribuição de nota, é possível utilizar atributos que sejam independentes. Ferramentas como Coh-Metrix[25] e Linguistic Inquiry Word Count (LIWC)[26], são utilizadas em trabalhos como Ferreira et al. (2021) e Ferreira Mello et al. (2022) para a extração de informações linguísticas, como legibilidade e coesão.

Trabalhos que utilizam métricas independentes de conteúdo são capazes de representar critérios de avaliação como Coerência e Coesão. No entanto, critérios como Tema são melhor avaliados por atributos dependentes de conteúdo, como exemplo as matrizes de termos, descritas no Capítulo Conjunto de dados, *dataset* e *corpus*, e métricas calculadas a partir dessas matrizes, como a similaridade de cosseno utilizada entre tema e redação em Amorim; Veloso (2017). Em Louis; Higgins (2010) e Persing; Ng (2014), são propostos cálculos de atributos dependentes de conteúdo com base em recursos linguísticos pré-definidos e associados aos temas relacionados às redações utilizadas nos experimentos.

Ainda sobre a extração de atributos, vale mencionar o trabalho de Sousa et al. (2021) que, além de aspectos linguísticos, explora aspectos relacionados à construção da argumentação, por meio de mineração de argumentos. A combinação de diferentes estratégias para extração de atributos é bastante comum, conforme realizado por Amorim; Veloso (2017) que, além de aspectos linguísticos e associados ao tema, incluem métricas associadas ao correto uso da língua, calculadas com base em desvios identificados por ferramentas externas, como as mencionadas na Seção 7.2.

É importante ressaltar que a inclusão de atributos relacionados a critérios de avaliação específicos não é imprescindível para atribuição de nota global. No entanto, a partir do momento em que se propõe atribuir notas por critérios avaliativos, é interessante incluir atributos que representem cada critério, ou poderá haver discrepância significativa no resultado obtido entre critérios, como observado em alguns trabalhos (Amorim; Veloso, 2017; Fonseca et al., 2018).

Selecionado um conjunto de atributos e realizada a análise estatística dos dados, podemos seguir à etapa de treinamento de modelos. Não convém aqui sugerirmos esta ou aquela técnica ou algoritmo, uma vez que conjuntos de dados distintos podem apresentar resultados também diferentes para os mesmos algoritmos (Ferreira Mello et al., 2022; Ferreira et al., 2021; Fonseca et al., 2018; Marinho et al., 2022a). Ao treinar modelos para atribuição de nota, assim como modelos com outros objetivos, é fundamental definir mais de um algoritmo e configurações para, então, realizar uma comparação estatística entre os resultados obtidos.

Entre os trabalhos que utilizam a abordagem de extração de atributos, há modelos de classificação e regressão treinados com diversos algoritmos, como: regressão linear (Fonseca et al., 2018), *Suppport Vector Machines* (SVM) (Haendchen Filho et al., 2018, 2019), *Gradient Boosting* (Fonseca et al., 2018; Marinho et al., 2022a). A comparação entre os modelos se dá, principalmente, pela avaliação dos valores obtidos para métricas

[25]Coh-Metrix é uma ferramenta computacional que calcula métricas e índices para aspectos linguísticos e discursivos em um texto e que será melhor explorada neste capítulo na Seção 7.4.1. Disponível em: http://cohmetrix.memphis.edu/cohmetrixhome.

[26]LIWC é uma ferramenta computacional, que realiza análise de textos baseada em métricas. Disponível em: https://www.liwc.app/.

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

como precisão, revocação, medida-F, RMSE e Kappa de Cohen.

Embora seja possível obter resultados satisfatórios pela engenharia de atributos e treinamento de modelos por algoritmos clássicos de aprendizado de máquina, é notável o esforço humano necessário para o processo de extração e seleção de atributos, considerando que muitos dos conjuntos de atributos são compostos por algumas centenas de métricas. Com isso em vista, surgem trabalhos que utilizam outras técnicas para a representação de textos e algoritmos de redes neurais profundas para a tarefa de atribuição de nota.

Alikaniotis et al. (2016) propõem uma técnica de *word embeddings* treinada com base em notas de redações, que é utilizada com redes neurais LSTM. O trabalho relata melhores resultados obtidos em comparação com outras abordagens.

Em Fonseca et al. (2018), as *word embeddings* GloVe são combinadas com redes LSTM bidirecionais e os resultados são comparados, também, com uma abordagem que utiliza engenharia de atributos. Os autores relatam que, embora a técnica de redes neurais tenha gerado bons resultados, o modelo gerado a partir de atributos se mostrou superior em diferentes aspectos.

Mayfield; Black (2020) realizam *fine-tuning* de modelos pré-treinados (BERT e variações) para a atribuição automática de notas. Apesar de relatar resultados até 5% melhores do que modelos baseados em n-gramas, os autores discutem sobre o tempo de treinamento deste tipo de modelo, que é cerca de 100 vezes mais demorado do que outras abordagens, e sobre o impacto que isso pode ter em fluxos mais dinâmicos de trabalho.

Bittencourt Jr. (2020) define 14 técnicas baseadas em combinações de diferentes representações de palavras e arquiteturas de redes neurais profundas para a execução da tarefa de atribuição de nota a redações. Os experimentos são realizados com um conjunto composto por redações de 18 temas, sendo que cada técnica é utilizada para o treinamento de um modelo por tema (18 modelos por técnica). Também é proposta uma abordagem para treinamento de modelo multi-tema, ou seja, um modelo único para a atribuição de notas para redações de mais de um tema.

O trabalho de Marinho et al. (2022a) compara 3 tipos de abordagens: (i) engenharia de atributos com algoritmo de regressão, (ii) *doc embeddings* com algoritmo de regressão e (iii) *word embeddings* com LSTM. As abordagens (i) e (iii) apresentaram melhores resultados para critérios de avaliação distintos, sendo a abordagem (iii) eleita pelos autores como a melhor. Os resultados da abordagem (iii) ainda foram comparados com resultados de Amorim et al. (2018) e Fonseca et al. (2018), sendo relatado melhor desempenho desta abordagem na atribuição de nota por critério de avaliação.

Nos últimos anos, os LLMs também têm sido explorados para a atribuição automática de notas a redações. Esses modelos podem ser aplicados tanto em cenários de *fine-tuning* supervisionado (Yang et al., 2020b), a partir de bases de redações previamente anotadas, quanto em cenários de *few-shot* ou *zero-shot learning* (Shibata; Miyamura, 2025), em que o próprio modelo, orientado por instruções, realiza a atribuição de notas sem necessidade de treinamento extensivo (Pack et al., 2024). Estratégias híbridas também têm sido propostas, como a incorporação de *shallow features* ou de comparações par-a-par no processo de atribuição de nota, com ganhos de desempenho em relação a métodos tradicionais (Faseeh et al., 2024; Shibata; Miyamura, 2025).

A principal vantagem dessa abordagem é a redução do esforço manual na engenharia de atributos e a possibilidade de maior generalização a diferentes critérios avaliativos, gêneros textuais e modelos de correção. Além disso, *frameworks* inovadores como o Rank-Then-Score (Cai et al., 2025) e abordagens colaborativas entre humanos e LLMs (Xiao et al., 2025) têm demonstrado resultados promissores, combinando consistência estatística e

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

explicabilidade. Por outro lado, os LLMs ainda apresentam desafios importantes, como custo computacional elevado, necessidade de dados para ajuste fino, bem como riscos de variação e falta de interpretabilidade. Apesar dessas limitações, os LLMs podem se consolidar como alternativa barata para avaliação automática de redações em larga escala, inclusive em cenários multilíngues ou em línguas com menos recursos computacionais, como é o caso do português.

7.3.2 Atribuição de nota para redações do Enem

Especificamente para o português, a maior parte dos trabalhos relacionados à atribuição de notas (e CAR) utiliza *corpora* compostos por redações do modelo de correção do Enem como base de treinamento. Dada a importância e dimensão do exame no Brasil, há interesse particular em encontrar soluções para a atribuição de nota exclusivamente para esse modelo de correção.

Como descrito na Seção 7.1.3, o Enem exige a produção de um texto do gênero dissertativo-argumentativo sobre um tema específico que é avaliado em 5 critérios, também chamados de competências: (1) Língua portuguesa, (2) Abordagem temática e adequação ao tipo textual, (3) Progressão textual e defesa do ponto de vista, (4) Coesão e articulação e (5) Proposta de intervenção. Os trabalhos que treinam modelos de atribuição de nota para o Enem predizem uma nota global, porém alguns tentam também aperfeiçoar por competência.

Barbosa de Lima et al. (2023) fazem uma revisão sistemática da literatura envolvendo avaliação automática de redações do Enem. Entre as conclusões apresentadas, destacamos o foco dos trabalhos selecionados no uso de atributos extraídos do texto em vez do uso de modelos pré-treinados baseados em Deep Learning, o baixo detalhamento dos feedbacks gerados e a escassez de análise do impacto prático em aplicações no mundo real. Alguns dos trabalhos apontados por Barbosa de Lima et al. (2023) são detalhados a seguir, por apresentarem relevância teórica ou técnica.

Em Amorim; Veloso (2017), Fonseca et al. (2018), Marinho et al. (2022a) e Bittencourt Jr. (2020), o foco está na atribuição de notas para cada uma das competências, o que pode ser feito com base em modelos treinados para cada competência ou um modelo único que prediz as notas para cada uma delas. Já em Haendchen Filho et al. (2018), é explorada a atribuição de notas para a competência 2, especificamente.

Ao realizar estudo sobre a predição de notas para cada uma das competências do Enem, Haendchen Filho et al. (2019) notaram o significativo desbalanceamento do conjunto de redações e tornaram esse o foco de seu trabalho, a fim de analisar o impacto e tratamento de conjuntos de dados desbalanceados na tarefa de atribuição de nota.

Alguns trabalhos que utilizam redações em português não têm como foco direto a atribuição de nota, mas a proposta de técnicas mensuráveis relacionadas a critérios cuja avaliação pode ser mais complexa. Como exemplo, citam-se as contribuições de Ferreira et al. (2021), Sousa et al. (2021) e Ferreira Mello et al. (2022) para a avaliação das competências 3 e 4.

É notável que, no momento da escrita deste capítulo, não pudemos encontrar nenhum trabalho em que se dê atenção em particular para a melhoria de atribuição de nota na competência 5 do Enem.

Vale ressaltar que os conjuntos de dados utilizados pelos referidos trabalhos não são muito representativos, possuindo até alguns milhares de redações de uma baixa diversidade de temas. O maior conjunto relatado possui 56.644 redações, sem indicação de número

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

de temas (Fonseca et al., 2018). O conjunto de redações com maior número de temas relatado, que também é o segundo em número de redações, conta com 27.184 redações distribuídas entre 18 temas, sendo que o número de redações por tema varia entre 3.070 e 710 (Bittencourt Jr., 2020). Além disso, ambos os maiores conjuntos foram fornecidos por empresas privadas e, portanto, não são públicos.

O tamanho e distribuição do conjunto de dados são considerados obstáculos para o treinamento de modelos de atribuição de notas, especialmente quando utilizadas técnicas de *deep learning*. Mesmo com o uso de abordagens híbridas ou baseadas em redes neurais profundas, a comprovação e generalização de resultados é um desafio. No entanto, há uma iniciativa para a criação de um conjunto público de redações do modelo Enem para utilização em trabalhos de CAR: até agosto de 2022 era composto por 6.579 redações pré-processadas e divididas em 151 temas (Marinho et al., 2022b) [27].

Para o português brasileiro e para o contexto da avaliação do ENEM, não foram encontradas abordagens baseadas em LLMs voltadas para a correção ou avaliação de redações. Identificamos apenas o trabalho de Locatelli et al. (2025), que exploraram o comportamento de LLMs (GPT-3.5, GPT-4 e MariTalk) no contexto do ENEM, mas no sentido de comparar redações geradas pelos modelos com aquelas escritas por alunos. Portanto, não tiveram o intuito de avaliar automaticamente os textos.

Também não encontramos nenhum trabalho de PLN que tenha relatado a atribuição de notas em redações de outros modelos de correção, como Fuvest, Unicamp, FGV ou outros.

Enfim, acreditamos que ainda há espaço para trabalhos quanto à tarefa de atribuição de notas em redações em português. Contudo, para atingir a meta de soluções completas de correção de redação, apenas a nota é insuficiente do ponto de vista do processo de ensino e aprendizagem. Para suprir essa lacuna, a Seção 7.4 discute a terceira tarefa de CAR, referente ao provimento de feedback relacionado ao texto.

7.4 Feedback para o aluno

Conforme apresentado na Introdução (Seção 7.1), a última etapa da Correção Automática de Redação (CAR) é o fornecimento de um feedback para o aluno. Até pouco tempo atrás, a correção automática produzia basicamente uma nota como resultado da avaliação da redação. Mas isso já não era mais suficiente e foi surgindo a necessidade de explicar ou justificar essa nota. De acordo com Shermis; Burstein (2013), os primeiros trabalhos se limitavam a dar feedbacks sobre as características e propriedades linguísticas do texto. Pesquisas mais recentes vêm focando em aspectos mais complexos e profundos da língua, que vão além da superficialidade do texto[28].

Em uma **correção manual**, esse feedback é feito pelo próprio corretor da redação, na forma de comentário livre, em linguagem natural, sem seguir nenhum tipo de padronização, podendo tecer críticas, fazer sugestões, elencar pontos fortes e pontos a melhorar, abordar questões gerais ou específicas da redação, enfim, de formas bastante variadas.

Já em uma **correção automática**, as plataformas que dão algum tipo de feedback sobre a correção o fazem de forma sistematizada. Porém, são raras as empresas que fornecem esse tipo de devolutiva ao aluno. Lima et al. (2023) fizeram uma revisão sistemática da

[27] https://github.com/lplnufpi/essay-br

[28] Vale ressaltar que os feedbacks baseados em características e propriedades linguísticas do texto ainda são os mais praticados hoje pelas plataformas brasileiras ou que processam o português, então focaremos nessa abordagem ao longo desta seção.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

literatura sobre CAR e uma das lacunas que identificaram nos trabalhos para o português é o baixo detalhamento nos feedbacks retornados pelos modelos de avaliação.

Na prática, os corretores automáticos costumam apontar apenas estatísticas básicas do texto, tais como quantidade de conectivos (conjunções), variação lexical (taxa de *types* por *tokens*), quantidade de palavras de conteúdo (substantivos, adjetivos, verbos e alguns advérbios), tamanho médio das palavras, frases e parágrafos, dentre outros, o que geralmente não tem utilidade pedagógica para o aluno. A Seção 7.4.1 apresenta como essas informações são calculadas e exibidas.

Algumas plataformas de CAR também disponibilizam para o aluno sistemas ou *bots* baseados em assistentes de escrita ou ferramentas computacionais de auxílio à escrita. Na Seção 7.4.2 apresentamos como esses recursos e ferramentas são utilizadas em sistemas de CAR.

Mais recentemente, com o surgimento e popularização de LLMs, algumas empresas também já começaram a fornecer feedbacks gerados automaticamente por esses modelos gerativos. Também é possível gerar automaticamente as devolutivas a partir de elementos encontrados ou não encontrados no texto, instanciando palavras ou trechos do texto da redação. Mas isso só é possível se for usada uma abordagem simbólica. Nesse sentido, o feedback pode conter críticas referenciando os desvios apresentados na Seção 7.2 e/ou elogios aos pontos fortes, como será apresentado na Seção 7.4.3.

7.4.1 Estatísticas básicas do texto

Algumas plataformas e empresas privadas que oferecem serviço de CAR apresentam para o aluno contagens básicas do texto, tais como a quantidade de palavras, caracteres, sentenças, parágrafos e até a quantidade de palavras por classe gramatical (verbos, substantivos, adjetivos, preposições, conjunções etc.). Outras oferecem um pouco mais de informação baseada em estatísticas simples, como a proporção de palavras únicas (*types*) em relação à quantidade total de palavras no texto (*tokens*), alguma medida de similaridade entre as sentenças, desvio padrão dos parágrafos, dentre outras.

Um dos recursos disponíveis para recuperar essas informações é o NILC-Metrix (Leal et al., 2023), uma versão brasileira do Coh-Metrix. O NILC-Metrix[29] é a atualização mais recente do Coh-Metrix-Port (Scarton; Aluísio, 2010), contendo 200 métricas[30] distribuídas nas 14 categorias apresentadas no Quadro 7.2, as quais avaliam a coerência, a coesão, a inteligibilidade, a complexidade e outros aspectos:

Quadro 7.2: Categorias de métricas disponíveis no NILC-Metrix.

Categoria	Qtde.	Descrição das métricas
Medidas Descritivas	10	Quantificam sílabas por palavra, palavras por sentença, sentenças por parágrafo, assim como quantidades absolutas de palavras, sentenças e parágrafos.
Simplicidade Textual	8	Analisam proporção de palavras fáceis e difíceis em relação ao total de palavras, bem como sentenças longas e curtas.
Coesão Referencial	9	Avaliam quantidades de palavras de conteúdo, radicais de conteúdo, médias de referentes, de candidatos a referentes e proporção de pronomes anafóricos.

[29] http://fw.nilc.icmc.usp.br:23380/metrixdoc

[30] Definição, explicação e exemplos das métricas podem ser conferidos na Documentação do NILC-Metrix (http://fw.nilc.icmc.usp.br:23380/metrixdoc).

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Coesão Semântica	11	Calculam similaridade, entropia, desvio padrão e taxa de *giveness*, ou seja, quantidade de informação dada e nova baseada em LSA (Capítulo Representação vetorial e semântica distribucional).
Medidas Psicolinguísticas	24	Verificam 4 critérios: concretude, imageabilidade, familiaridade e idade de aquisição das palavras.
Diversidade Lexical	15	Calculam proporções e desvio padrão de *types* e *tokens*, tanto gerais quanto por PoS e por categorias de palavras lexicais e funcionais.
Conectivos	12	Verificam a proporção de vários tipos de conectivos (aditivos, causais, lógicos) e operadores lógicos (positivos e negativos) em relação ao total de palavras do texto.
Léxico Temporal	12	Consideram verbos flexionados nos diferentes tempos e modos verbais, bem como as formas regular e irregular de particípio.
Complexidade Sintática	27	São bastante diversas, considerando desde a quantidade de orações por sentença, coordenação e subordinação, até aspectos como voz ativa e passiva, aposto, distância na árvore de dependência, fórmulas de Frazier e Yngve, dentre vários outros.
Densidade de Padrões Sintáticos	4	Calculam proporção de verbos no gerúndio e também tamanhos de sintagmas nominais.
Informações Morfossintáticas de Palavras	42	Verificam quantidade média, proporção ou desvio padrão das palavras por PoS.
Informações Semânticas de Palavras	11	Incluem dados sobre substantivos abstratos, palavras polissêmicas, hiperônimos, nomes próprios e polaridade (positiva e negativa) das palavras.
Frequência de Palavras	10	Calculam frequências de vários tipos de palavra (de conteúdo, e.g.) com base na curva de Zipf e considerando diferentes *corpora*.
Índices de Leiturabilidade	5	Incluem índices e fórmulas já consolidados em PLN, tais como Fórmula Dale Chall, Índice de Brunet, Flesch, Gunning Fog e Estatística de Honoré.

Os cálculos dessas métricas geralmente resultam em um valor numérico, o qual não se faz útil para o aluno. Porém, há diferentes maneiras de devolver ao aluno um feedback textual com a interpretação de algumas dessas métricas. Por exemplo, se considerarmos os valores de 4 métricas de Simplicidade Textual, referentes a tamanho de sentença (a saber: *long_sentence_ratio*[31], *medium_long_sentence_ratio*[32], *medium_short_sentence_ratio*[33], *short_sentence_ratio*[34]), é possível criar um resultado interpretável para dizer ao aluno que ele constrói sentenças muito longas e isso pode prejudicar a compreensão das ideias do texto.

Tanto as estatísticas básicas quanto as métricas do NILC-Metrix podem ser utilizadas não apenas para devolver feedbacks aos alunos, mas também como atributos para calcular a nota da redação ou de alguns aspectos da redação, conforme apresentado na Seção 7.3.

7.4.2 Assistentes de escrita e ferramentas de auxílio à escrita

Para prover uma devolutiva ao aluno, também é possível recorrer a sistemas prontos de PLN, como os assistentes virtuais, assistentes de escrita ou ferramentas de auxílio à escrita. Essas soluções podem ser entendidas como aplicações finais, mas, na área de CAR, elas são usadas como recursos ou ferramentas intermediárias para subsidiar a solução completa de CAR.

Essas ferramentas são capazes de gerar, melhorar, reformular e personalizar qualquer tipo de conteúdo textual, incluindo redações. Algumas delas funcionam de forma síncrona *real-time*, fazendo correções e dando sugestões à medida que o texto está sendo escrito,

[31]Proporção de sentenças muito longas em relação a todas as sentenças do texto.
[32]Proporção de sentenças longas em relação a todas as sentenças do texto.
[33]Proporção de sentenças médias em relação a todas as sentenças do texto.
[34]Proporção de sentenças curtas em relação a todas as sentenças do texto.

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

enquanto outras funcionam a posteriori, ou seja, depois que o aluno submete sua redação à plataforma de correção, ele recebe uma devolutiva com críticas e/ou elogios.

Para a língua inglesa, há inúmeros assistentes de escrita e muitos deles conhecidos no Brasil porque as pessoas usam o inglês para escrever, por exemplo, artigos científicos. Um dos mais populares é o Grammarly[35], mas também há outros bastante usados, como Linguix[36], Ginger[37], Reverso[38], Writer[39], Hemingway App[40] e outros.

Para o português, também existem vários softwares comerciais, sendo a maioria paga. As ferramentas de auxílio à escrita, ao lado dos simplificadores textuais e dos sumarizadores automáticos, podem contribuir com a área de CAR, pois fornecem:

- **Correção ortográfica e gramatical:** Os sistemas podem usar regras ou modelos de linguagem treinados em um grande volume de textos em português para identificar erros ortográficos e gramaticais comuns.
- **Análise de contexto:** As ferramentas não apenas verificam palavras isoladas, mas também consideram o contexto da frase em que uma palavra está inserida. Isso ajuda a evitar falsos positivos e permite que o sistema forneça sugestões de correção mais precisas.
- **Sugestões de melhoria:** Quando uma palavra é identificada como incorreta ou quando uma construção gramatical suspeita é detectada, o assistente de escrita oferece sugestões para corrigir o problema. Essas sugestões podem incluir substituições de palavras, ajustes na estrutura da frase ou correções de pontuação.
- **Detecção de estilo:** Além de corrigir erros básicos, um assistente de escrita também pode oferecer sugestões para melhorar o estilo de escrita. Isso inclui alertas sobre uso excessivo de palavras, repetições, uso inadequado de voz passiva, entre outros aspectos.
- **Feedback de clareza:** As ferramentas também podem avaliar a clareza do texto, identificando frases longas e complexas que podem ser difíceis de entender, podendo sugerir dividir essas frases ou reformulá-las para tornar o conteúdo mais acessível.
- **Verificação de plágio:** Algumas soluções comerciais oferecem uma funcionalidade adicional para verificar a originalidade do texto, identificando trechos que possam ser semelhantes a outras fontes online. Isso é especialmente útil para evitar acidentalmente usar conteúdo plagiado.
- **Aprendizado contínuo:** Assim como outras ferramentas de PLN, os assistentes de escrita também continuam aprendendo e melhorando com o tempo. Eles são atualizados com novos dados e feedbacks dos usuários, o que ajuda a aprimorar seus modelos e a abordagem dos problemas linguísticos.
- **Extensões e integrações:** Muitos deles oferecem extensões para navegadores, complementos para processadores de texto e aplicativos móveis, o que permite aos usuários verificar seu conteúdo em tempo real enquanto escrevem em várias plataformas.
- **Personalização:** Em alguns desses sistemas, o usuário pode personalizar as configurações com base em suas preferências de estilo e escrita. Isso permite adaptar as sugestões e correções de acordo com o contexto e o público-alvo.

[35] https://www.grammarly.com/
[36] https://linguix.com/
[37] https://www.gingersoftware.com/
[38] https://www.reverso.net/tradu%C3%A7%C3%A3o-texto
[39] https://writer.com/grammar-checker/
[40] https://hemingwayapp.com/

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Conforme dito anteriormente, as melhores ferramentas de auxílio à escrita que existem hoje para o português são soluções comerciais de empresas privadas. Existem também alguns sistemas desenvolvidos a partir de pesquisas acadêmicas e científicas, mas nenhuma focada em redação. Por exemplo, o SciPo[41] (Feltrim et al., 2003), que é um sistema de auxílio à escrita de resumos acadêmicos em português, especialmente para teses e dissertações da área da Ciência da Computação. Outro exemplo é o WRITEME[42] (Leite et al., 2020), que é ferramenta de auxílio à escrita de READMEs que usa dados abertos dos repositórios mais populares do GitHub para gerar recomendações de seções, mas também não é focada e redação.

7.4.3 Identificação de pontos fortes e elogiáveis

Na Seção 7.2 falamos da detecção de pontos fracos e desvios no texto. Por outro lado, também é importante detectar pontos fortes e elogiáveis e demonstrá-los ao aluno para que ele continue usando a mesma estratégia nos próximos textos.

Esses pontos fortes podem ser identificados por meio de regras formais, mas também é possível usar diferentes estratégias para cada aspecto da avaliação.

Tendo identificado todos ou alguns aspectos (positivos ou negativos) do texto, é possível retornar essas informações ao aluno na forma de feedbacks construtivos para auxiliá-lo a se tornar um escritor mais habilidoso e confiante.

Avaliação de coesão Na Seção 7.2.1.3, falamos brevemente de como identificar usos corretos de recursos coesivos usando regras em contextos linguísticos menores, como dentro de uma sentença.

Também é possível criar regras formais que percorrem todo o texto procurando as ocorrências de conectivos, avaliar a sua distribuição ao longo do texto, calcular a variabilidade e diversificação deles e até procurar conectivos em pontos específicos da redação, como no início da conclusão, por exemplo.

Com o objetivo de fornecer um feedback baseado na avaliação da coesão do texto, uma solução simples é usar um *tagger* que identifique palavras etiquetadas como conjunções, preposições e advérbios, ou usando listas e léxicos específicos. A outra solução, que é um pouco mais rebuscada, é recorrer às métricas do NILC-Metrix que incidem sobre a coesão textual.

Identificação de repertórios Para avaliar a abordagem temática, referente à competência 2 do Enem, podemos elogiar a presença (ou criticar a ausência) de repertórios socioculturais, que são informações, fatos, citações, definições ou termos de alguma área do conhecimento, ou ainda experiências pessoais que, de alguma forma, contribuem como argumento para defender um ponto de vista.

Pelo Manual de leitura do Enem[43], os repertórios socioculturais podem ser legitimados (com citação da fonte) ou não legitimados (sem citação da fonte), ter uso produtivo (pertinente à discussão em mais de um momento do texto) ou não, pertencente ao tema ou não e ainda devem ser penalizados se forem exclusivamente baseados nos textos motivadores. Identificar automaticamente todos esses tipos e usos (corretos ou não) dos repertórios não é uma tarefa simples. Porém isso pode ser

[41] https://escritacientifica.sc.usp.br/scipo/
[42] https://repositorio.ufpe.br/handle/123456789/50043
[43] https://download.inep.gov.br/educacao_basica/enem/downloads/2020/Competencia_2.pdf

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

feito usando modelos de extração de entidades nomeadas (Capítulo 3), buscando, por exemplo, as citações de filósofos, sociólogos e outros estudiosos, ou buscando as menções a livros, filmes, séries, dentre outras entidades que funcionem como repertórios legitimados.

Avaliação da Progressão textual Para a avaliação da progressão textual, é possível treinar e usar modelos de tópico, a exemplo do *Hidden Topic Markov Models* (HTMM) (Gruber et al., 2007), que classificam as sentenças de um texto por tópicos ou assuntos, o que nos permite avaliar a progressão, a continuidade, a retomada e até a circularidade entre os assuntos, a partir da distribuição dos tópicos em um texto.

Blei; Moreno (2001) apresentam resultados dessa abordagem de segmentar um texto não estruturado em tópicos, testando em notícias do New York Times. Os autores propuseram uma combinação do tradicional modelo oculto de Markov (*Hidden Markov Model* – HMM) com o modelo de semântica latente de Hofmann (Hofmann, 1999), resultando em um novo método probabilístico que segmenta um texto em tópicos. Essa abordagem pode ser muito útil para avaliar o encadeamento das ideias de um texto, principalmente no caso de redação do Enem, pois o gênero dissertativo-argumentativo costuma seguir um padrão bem definido de: (i) apresentação do tema, introdução ao ponto de vista a ser defendido e breve menção aos argumentos a serem utilizados (no primeiro parágrafo); (ii) exposição do primeiro argumento (no segundo parágrafo); (iii) exposição do segundo argumento (no terceiro parágrafo); (iv) proposta de intervenção para solucionar o problema discorrido e retomada da tese na forma de conclusão (no quarto parágrafo).

Identificação de proposta de intervenção No Enem, para que uma redação receba nota máxima na Competência 5, o aluno precisa criar uma proposta de intervenção que contenha pelo menos 5 elementos: o agente (quem?), a ação (o quê?), o modo ou meio (como?), a finalidade (para quê?) e o detalhamento de algum dos elementos anteriores.

Para a correta identificação desses elementos, pode-se usar modelos de extração de informação (Capítulo 3) ou recorrer a extração de entidades nomeadas ou recursos linguísticos como listas e léxicos específicos.

O Inep disponibiliza a Cartilha do participante[44] com instruções sobre agentes que devem ser considerados nulos, ações interventivas que devem ser consideradas nulas, propostas de intervenção negativas ou condicionais, dentre outras orientações que podem se transformar em atributos para modelos.

Não encontramos nenhum trabalho para o português que reporte bons resultados quanto à identificação da proposta de intervenção e que valha ser replicado. É um dos campos de CAR que merece ser mais explorado.

Ao longo desta Seção 7.4, discutimos algumas formas possíveis de devolver um feedback ao aluno, que podem ser: indicando números, percentuais e estatísticas básicas do texto, ou acoplando um assistente de escrita ao corretor automático para fazer isso em tempo real, ou ainda instanciando elementos da redação (recursos coesivos, repertórios, sequência de tópicos, elementos da proposta de intervenção) em uma mensagem gerada automaticamente. Mas as possibilidades não se limitam a essas indicadas neste capítulo. Para outras formas de geração de feedbacks em redações escolares, ver Gamon et al. (2013).

[44] https://download.inep.gov.br/download/enem/cartilha_do_participante_enem_2022.pdf

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Tendo em vista todo o conteúdo apresentado na Seção 7.2, Seção 7.3 e Seção 7.4, pode surgir o questionamento sobre o papel (ou até extinção) das correções manuais, dados os avanços em CAR. Para tanto, na Seção 7.5, propomos uma discussão sobre prós e contras de cada uma das abordagens de correção: a manual e a automática, apresentando alguns casos de sucesso e defendendo uma correção híbrida que se beneficie do potencial de cada abordagem.

7.5 Correção manual vs(?) correção automática

A correção automática de redações (CAR) divide opiniões entre alunos, escritores, professores de redação, bancas de avaliação em série, especialistas em Linguística Computacional, cientistas de dados e desenvolvedores de sistemas. Ainda existe muito preconceito quando se trata de correção automática de redação, mas já é consensual aceitar as vantagens dos corretores ortográficos e gramaticais quando embutidos em outras soluções, como no pacote Office, no Gdrive, em redes sociais ou nos teclados dos smartphones.

A discussão principal gira em torno de seus prós e contras, se a correção automática deve substituir ou complementar a correção humana, sobre questões éticas relacionadas à correção automática, sobre a subversão dos valores pedagógicos e educacionais da avaliação manual para uma avaliação automática de textos; enfim, para uma discussão mais filosófica e profunda sobre todos esses aspectos, ver Elliot; Klobucar (2013) e Hakuta (2013).

Nesta seção, abordaremos apenas questões práticas relacionadas à correção manual e à correção automática de redações para, ao final, defender uma correção híbrida, que utilize as principais potencialidades de cada tipo, reconhecendo-se também suas limitações.

7.5.1 Avanços dos últimos anos

Até as décadas de 80 e 90, as avaliações de redação no Brasil eram holísticas, ou seja, o avaliador do texto atribuía uma nota global (de 0 a 100, por exemplo) para a redação, sem seguir rigorosamente nenhum critério previamente estabelecido. Por volta dos anos 2000, essas avaliações passaram a ser analíticas, tendo que explicitar todos os critérios e todos os conceitos que deveriam ser avaliados. Ao mesmo tempo, as avaliações passaram a ser em duplas às cegas, ou seja, cada redação deveria ser avaliada por dois corretores independentes, o que exigia maior sistematicidade e coerência entre eles.

Nessa transição de avaliação holística para analítica, as grades de correção de redações se tornaram mais padronizadas. E sabe-se que tarefas mais padronizadas são melhor executadas por máquinas do que por humanos.

Mesmo com a tentativa (por vezes, falha) de padronização das grades, ainda se percebe a falta de objetividade na definição de critérios por parte de alguns modelos de correção. Quando a grade de correção é muito aberta ou não apresenta os critérios bem definidos para cada faixa de nota, aumentam as chances de haver divergência entre duas avaliações cegas. Por outro lado, quando os corretores humanos passam por treinamentos rigorosos, tal como é feito no Enem, isso pode reduzir o número de inconsistências nas avaliações, mas ainda assim não elimina as divergências, já que pessoas diferentes podem ter interpretações diferentes sobre a mesma instrução. Prova disso são os índices de redações do Enem que vão para uma terceira correção[45], nos casos de discrepância de 80 ou mais pontos em uma

[45] Os índices de terceira correção variam a cada ano, pois dependem de vários aspectos, inclusive a mudança dos critérios do Inep para a terceira correção. A título de exemplificação, podemos citar o índice de 20,10% em 2012 disponível no Portal do MEC (http://portal.mec.gov.br/component/tags/tag/corre

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

competência ou de 100 ou mais pontos na nota final.

Posto isso, a correção automática no Brasil passou a ser considerada como uma alternativa à manual, já que esta última sempre foi passível de subjetividade e viés.

7.5.2 Vantagens da correção automática

Correções manuais estão sujeitas a subjetividade e viés, além do cansaço humano, a pressão por produtividade, a cobrança por eficiência, o desinteresse pela tarefa, dentre outros fatores que podem prejudicar a qualidade da avaliação ou comprometer sua validade.

Para além dessas questões de limitação humana, é necessário mencionar também o tempo e o custo da correção manual. De acordo com uma matéria veiculada no Portal G1[46] em 2016, os corretores humanos conseguem corrigir, em média, 74 redações por dia. Já Bittencourt Jr. (2020, p. 19) apresenta uma média de 12 minutos por correção, o que daria 40 redações por dia, considerando-se 8 horas de trabalho. E o custo de cada correção de redação do Enem para o Governo Federal era de R$15,88 em 2015. No mesmo ano foram corrigidas 6.54 milhões de redações, perfazendo um custo aproximado de R$104 milhões para o governo. Esse valor provavelmente está defasado, mas foi o último registro oficial encontrado.

Automatizar a correção de redações traz como vantagem a redução do custo de correção e elimina os fatores problemáticos relacionados ao trabalho humano.

Outro aspecto da correção que merece ser comparado é a confiança (ou *reliability*, em inglês). Os sistemas automáticos têm confiança de 100%, o que não pode ser afirmado para a correção manual. Isso significa que toda vez que a mesma redação passar pelo mesmo sistema de correção automática, receberá a mesma correção e a mesma nota. Isso parece óbvio, mas não é o que acontece na correção humana. Diferentes pessoas que corrigirem a mesma redação poderão naturalmente atribuir diferentes notas e/ou apontar diferentes aspectos a serem melhorados. O que também ocorre é que a mesma redação, quando corrigida pelo mesmo corretor humano em diferentes momentos, também pode receber avaliações muito diferentes, o que abre brecha para reclamações.

7.5.3 Vantagens da correção manual

Apesar de todos esses aspectos negativos em relação à correção manual, deve-se ressaltar o ponto forte desse tipo de correção, que é a possibilidade que o humano tem de observar todo e qualquer aspecto relacionado ao processo de construção de sentidos em um texto, o que a máquina não é capaz de fazer.

A produção textual é um processo sócio-cognitivo muito complexo que vai além da capacidade dos sistemas computacionais. A máquina não entende a redação, não interpreta o conteúdo veiculado pelo texto, mas apenas se comporta da forma como ela foi treinada para fazê-lo. Por mais que alguns modelos computacionais possam ser "interpretáveis", é impossível identificar e definir todos os fatores sociais, psicológicos, cognitivos, emocionais

cao). Também se pode inferir o índice de 43,52% em 2014, a partir DE "Ao todo, foram corrigidos 6.193.565 textos. [...] foram encaminhadas 2.695.949 redações para um terceiro corretor." disponível no Portal do MEC (http://portal.mec.gov.br/component/tags/tag/espelho-da-redacao). Ou uma estimativa de 29% em 2017 "O Inep estima que das 4,1 milhões de redações corrigidas, cerca de 1,2 milhão receberão a terceira correção." disponível no Portal do MEC (http://portal.mec.gov.br/component/tags/tag/correcao).

[46] https://g1.globo.com/educacao/enem/2016/noticia/corretores-de-redacao-do-enem-avaliam-em-media-74-redacoes-por-dia.ghtml

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

etc. que podem interferir tanto no processo de escrita por parte do aluno quanto no processo de correção por parte do avaliador.

Nesse sentido, considerando que a correção automática é limitada, é passível de erros e está mais voltada para a avaliação da forma do que do conteúdo, levanta-se o seguinte questionamento: O uso da correção automática não levaria o aluno a focar sua atenção apenas nos aspectos formais da escrita, excluindo os aspectos mais ricos da construção de sentidos no texto? Por trás desse questionamento, existe uma preocupação legítima de que o aluno não construa sua própria autonomia enquanto escritor, mas apenas seja "adestrado" a escrever de uma forma que o algoritmo lhe dê uma nota boa.

7.5.4 O exemplo da língua inglesa

Para a língua inglesa, algumas instituições educacionais (e.g. **ETS** – *Educational Testing Service*) utilizam modelos de AEE para auxiliar (e não substituir) a correção manual. Esses modelos computacionais são usados como uma segunda avaliação, complementar à avaliação humana. Por exemplo, a avaliação do **TOEFL** (*Test of English as a Foreign Language Internet-based Test*) é dupla, sendo uma feita por humanos e outra feita por sistemas automáticos. A nota final do aluno é dada pela média das duas avaliações. No caso de divergência entre as notas, a redação é enviada a terceiro corretor, semelhante ao que ocorre na avaliação do Enem. Mas, no caso do Brasil, tanto o primeiro quanto o segundo avaliador são humanos. Vale ressaltar que, segundo Bridgeman (2013, p. 227), "a experiência com o programa TOEFL iBT sugere que, quando há discrepância e a redação é enviada a um avaliador humano adicional, esse avaliador tende a concordar com a máquina mais do que com o outro humano"[47].

Um processo semelhante ocorre na avaliação do **GRE** (*Graduate Record Examination*), mas neste último caso a nota atribuída automaticamente é usada como se fosse uma validação para a avaliação humana. Em outras palavras, a correção automática é usada para monitorar a performance dos corretores humanos, a fim de identificar avaliadores desalinhados ou que precisam passar por novo treinamento.

A correção automática também pode auxiliar a correção manual no sentido de "nivelar" diferentes níveis de rigor. Sabe-se que diferentes avaliadores humanos podem ser sistematicamente mais rígidos ou mais permissivos em suas correções. Segundo Braun (1988, p. 1), "Quando o grau de leniência/severidade do avaliador pode ser atestado adequadamente, é possível calibrar estatisticamente os avaliadores e ajustar as pontuações corretamente [...] Essa calibração estatística parece ser uma abordagem econômica para aumentar a confiabilidade da nota quando comparada ao simples aumento do número de avaliadores por artigo."[48].

Nesse sentido, os modelos de AEE podem auxiliar a calibrar essas diferenças de rigidez, atribuindo um peso maior às correções dos avaliadores mais permissivos e um peso menor às correções dos avaliadores mais rigorosos.

[47]Tradução nossa. Do inglês: "*In fact, experience with the TOEFL iBT program suggests that when flagged discrepant scores are sent to an additional human rater, that rater tends to agree with the machine more often than she or he agrees with the other human score.*"

[48]Tradução nossa. Do inglês: "*When rater leniency/severity can be adequately documented, it is possible to statistically calibrate raters and adjust scores accordingly [...] This statistical calibration appears to be a cost-effective approach to enhancing scoring reliability when compared to simply increasing the number of readings per paper.*"

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

7.5.5 O que defendemos

Levantamos todos esses questionamentos ao longo da Seção 7.5 a fim de tornar explícitas as potencialidades da área de CAR, mas, ao mesmo tempo, esclarecer ao leitor sobre suas limitações, da mesma forma que a correção humana também possui vantagens e desvantagens.

Tendo considerado os vários aspectos das duas abordagens, defendemos neste capítulo uma correção híbrida, semelhante ao que é praticado para a língua inglesa (Seção 7.5.4), que possa se beneficiar dos pontos positivos da correção automática, mas mantendo a correção manual para garantir a responsabilização do humano sobre a avaliação.

7.6 Considerações finais

Neste capítulo exploramos uma das várias aplicações do Processamento de Linguagem Natural (PLN), a chamada Correção Automática de Redação (CAR), a qual abarca duas áreas de PLN em inglês, representadas pelas siglas AES (*Automated Essay Scoring*) e AEE (*Automated Essay Evaluation*).

Ademais, defendemos uma abordagem holística para a CAR, abrangendo, no mínimo, três fases essenciais: (i) a detecção de desvios no texto, (ii) a atribuição de nota, e (iii) a geração de feedback construtivo para o aluno. Apesar de termos dividido essas etapas para fins didáticos, é crucial reconhecer sua interdependência no decorrer do processo. Por exemplo:

1. os desvios gramaticais, ortográficos, de vocabulário etc. podem ser usados como atributos para o cálculo da nota;
2. as métricas e estatísticas básicas do texto, além de serem usadas para criar os feedbacks, também podem servir como atributos para o cálculo da nota;
3. a nota atribuída pelo modelo pode restringir, limitar ou ajudar a selecionar o feedback mais apropriado a ser exibido para o aluno;
4. A caracterização dos desvios (por tipos e quantidades) também pode ser usada para a geração do feedback.

Assim, ainda que tenhamos delimitado didaticamente essas três etapas, é importante ressaltar que, no contexto das tarefas de CAR, tais fases são intrinsecamente entrelaçadas e interdependentes, colaborando harmoniosamente para aprimorar a avaliação da redação.

Embora existam numerosos estudos nesses campos para o inglês e outras línguas, a documentação relevante para o português ainda é escassa e a maioria dos trabalhos acadêmicos confiáveis foi conduzida em pequenas amostras de dados. O progresso mais notável para textos em português provém de empresas e plataformas privadas que oferecem serviços de CAR. No entanto, os métodos e resultados dessas empresas nem sempre são divulgados, e, mesmo se o fossem, seria difícil compará-los devido à falta de uniformidade entre as soluções apresentadas.

Nesse sentido, a área de CAR ainda apresenta um vasto campo de trabalho a ser explorado por novos pesquisadores. Para o português, ainda faltam bons *datasets* de redações, que contenham, além dos textos, as notas por competência, anotação e apontamentos feitos por humanos; também faltam ferramentas robustas de detecção de desvios e de auxílio à escrita, bem como bons *parsers* e *taggers*; e faltam trabalhos que reportem bons resultados, com engenharia de atributos, comparação da performance dos algoritmos utilizados e uma análise aprofundada dos resultados.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Capítulo 8

Detecção Automática de Notícias Falsas

Renato Moraes Silva
Roney Lira de Sales Santos
Thiago Alexandre Salgueiro Pardo

Publicado em: 20/11/2024
Atualizado em: 16/04/2026

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

8.1 Introdução

Notícias falsas podem matar, podem destruir vidas e podem interferir nos rumos de povos e nações. Apesar das maravilhas que a tecnologia traz, a própria tecnologia (que inclui dispositivos e serviços, como smartphones, redes sociais e mensageiros instantâneos) tem permitido que notícias falsas se espalhem rapidamente, alcançando milhares de pessoas em pouco tempo, causando danos de todo tipo, afetando desde decisões políticas até a saúde física e mental da população. Grandes eventos e acontecimentos recentes, como eleições presidenciais e a pandemia de COVID-19, trazem muitas evidências do impacto dessas informações enganosas. Mais do que isso, a Inteligência Artificial, apesar dos avanços impressionantes, tem produzido sistemas que, como efeito colateral, podem criar desinformação. Assim, o problema das notícias falsas, embora antigo, ganhou uma nova dimensão com a tecnologia, tornando-se uma ameaça significativa para a sociedade.

Nesse contexto, a tarefa de detecção automática de notícias falsas tem grande apelo junto à sociedade (quer seja em nível individual, quer seja em nível governamental e de gestão), com potencial de trazer contribuições reais, com soluções aplicadas no dia a dia, podendo salvar vidas. Do ponto de vista científico, é também uma tarefa de grande interesse e com desafios instigantes, que incluem caracterizar o estilo de escrita de notícias falsas (e, complementarmente, das verdadeiras), explicar e mapear a distribuição dessas notícias e como elas afetam as várias camadas da sociedade, e como desenvolver sistemas computacionais que as identifiquem automaticamente.

Neste capítulo, apresenta-se o problema e relatam-se os esforços da área de Processamento de Linguagem Natural (PLN) para lidar com a questão, com foco especial na língua portuguesa do Brasil. Inicialmente, o problema é contextualizado e são apresentadas as bases teóricas para a compreensão do fenômeno das notícias falsas. Em seguida, são introduzidos os principais *corpora* existentes para o português, sobre os quais sistemas computacionais são desenvolvidos e avaliados. São então discutidos métodos e abordagens aplicados à identificação automática de notícias falsas. Finaliza-se com uma discussão sobre consequências sociais e considerações finais sobre este tópico tão relevante.

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

8.1.1 Conteúdo enganoso: definições

Conteúdo enganoso, desilusão, fraude e mentira são traduções aceitas para a expressão "*deception*", comum na língua inglesa para se referir a uma informação transmitida para criar uma falsa impressão ou conclusão. As notícias falsas, também popularmente chamadas pelo termo em inglês, *fake news*, são provavelmente o tipo mais citado de conteúdo enganoso. Elas são propositalmente divulgadas para enganar, sendo prejudiciais, incluindo os casos de quando o conteúdo é desconectado de sua origem e contexto (Lazer et al., 2018; Rubin, 2014).

Além das notícias falsas, Rubin et al. (2015) incluem os trotes/embustes (*hoaxes*, em inglês) e o conteúdo humorístico (que inclui ironia, sarcasmo e sátira, projetados para serem percebidos pelo leitor/ouvinte) como tipos de conteúdo enganoso. Santos (2022) e Chen et al. (2015) também citam os casos das revisões falsas de produtos (em sites de comércio eletrônico, por exemplo, para promover ou prejudicar produtos ou empresas) e os *clickbaits* (conteúdos que chamam a atenção, visando que o usuário clique em algum link).

Outra classificação interessante refere-se aos termos *misinformation*, *disinformation* e *malinformation* (Wardle; Derakhshan, 2017):

- *Misinformation*: as informações são falsas, mas não são criadas com a intenção de causar dano ao leitor. Esta definição abrange os erros não intencionais, tais como legendas de fotos erradas, datas, estatísticas, erros de tradução e até situações em que ironias e sátiras são levadas a sério.

- *Disinformation*: as informações são falsas e são criadas deliberadamente de forma a prejudicar uma pessoa, grupo social, organização ou país. Aqui são incluídos conteúdos de áudio ou visual (vídeos) fabricados. As notícias falsas fazem parte desse grupo, normalmente traduzido como "desinformação".

- *Malinformation*: informações baseadas na realidade, usadas para infligir danos a uma pessoa, organização ou país. Publicações de informações privadas (vazamentos), mudança deliberada de contexto, data ou tempo, assédio e discurso de ódio são exemplos desta definição.

Também se pode falar em meias verdades, quando fatos verdadeiros e falsos se misturam nas notícias em diferentes graus. Em geral, os fatos verdadeiros são utilizados de tal forma que os falsos sejam amparados, aumentando suas chances de serem percebidos como legítimos (Clem, 2017). Mais recentemente, entraram em cena as pós-verdades, quando os fatos objetivos em si são menos importantes do que o apelo às emoções que trazem e às crenças que evocam. Assim, os chamados "fatos alternativos" ganham relevância frente a "fatos reais" e sentimentos têm mais peso do que evidências (Saquete et al., 2019). O termo surgiu no âmbito político, mas logo foi popularizado em outros domínios.

Especialmente com os avanços da Inteligência Artificial generativa, outras modalidades de conteúdo enganoso ganharam força, envolvendo a produção de vídeos, imagens e áudios falsos ou manipulados digitalmente, contribuindo para a proliferação das chamadas *deepfakes*. Esse nome surgiu da junção dos termos *deep learning* e *fake*, sendo citado originalmente em 2017 por um usuário anônimo na rede social Reddit para se referir à substituição do rosto de uma pessoa por outra em um vídeo pornográfico (Rana et al.,

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

2022). Mais recentemente, tem-se falado em *deepfakes* textuais (Li et al., 2024) especificamente para se referir ao conteúdo escrito produzido por grandes modelos de língua (em inglês, *Large Language Models* ou LLMs).

Como se pode perceber, há todo um ecossistema em que as notícias falsas residem, sendo um ambiente em evolução, que muda conforme a sociedade, sua cultura e a tecnologia avançam. Conforme Saquete et al. (2019) discutem, esse ecossistema visa principalmente manipular a opinião pública, influenciar o comportamento das pessoas e, em última instância, trazer poder e ganhos financeiros aos envolvidos.

Esse capítulo foca especificamente nas notícias falsas em sua modalidade escrita, que têm sido o alvo de pesquisa da maior parte dos esforços atuais em PLN.

8.1.2 Notícias falsas no tempo e no espaço

Engana-se quem pensa que notícias falsas são um mal recente ou que se restringem a países com situações socioeconômicas mais frágeis.

A fabricação e a propagação de conteúdo enganoso são um fenômeno de longa data. Antigamente, as informações eram repassadas por meio de fofocas, no "boca a boca" e também nos tabloides (a chamada "imprensa marrom"), tanto para falar sobre outras pessoas de modo "inocente" como para, intencionalmente, prejudicar a reputação de outras pessoas ou empresas rivais. Até mesmo jornais e mídias televisivas podem, inadvertidamente, divulgar conteúdo enganoso, o que pode gerar severas consequências no mundo real. Por exemplo, em 1924, um documento forjado foi publicado por um jornal britânico renomado apenas quatro dias antes das eleições gerais, com o objetivo de desestabilizar o processo eleitoral[1]. Outro exemplo ocorreu depois de um acidente na Inglaterra, em 1989, quando muitas pessoas morreram devido à superlotação e falta de segurança em um estádio. Os jornais da época relataram que, entre as pessoas que morreram, algumas estavam alcoolizadas e brigaram com policiais que estavam tentando resgatar os feridos. Anos depois, provou-se que essas afirmações eram falsas[2].

Hoje as redes sociais e os aplicativos de mensagens instantâneas permitem que tais conteúdos enganosos alcancem uma audiência nunca imaginada antes da era da web, não discriminando países. Há casos famosos de eventos afetados por notícias falsas em todo o mundo, como a eleição americana de 2016, a saída do Reino Unido da União Europeia (que ficou conhecida pelo termo "*Brexit*") em 2020, a pandemia de COVID-19 entre 2020 e 2023, ou a recomendação de tratamento médico para a epidemia recorrente de dengue no Brasil. O Quadro 8.1 mostra um exemplo típico desse último caso, que transcreve uma fala divulgada por vídeo sobre tratamento para dengue, amplamente desmentida por diversas agências de checagem de fatos[3].

Quadro 8.1: Exemplo de notícia falsa sobre tratamento para dengue.

> Esprema 200 ml do limão taiti. Dá mais ou menos um copo de limão. Beba todo o copo de uma vez só. Basta apenas uma dose. A carga viral vai acabar, bem como todos os sintomas, nas próximas horas. Nos casos de dengue, em que as plaquetas começam a baixar, o que pode levar à dengue hemorrágica, é preciso tomar os 200 ml de limão para erradicar a carga viral

[1] https://www.theguardian.com/politics/1999/feb/04/uk.politicalnews6

[2] https://www.theguardian.com/football/2016/apr/26/hillsborough-disaster-deadly-mistakes-and-lies-that-lasted-decades

[3] Veja, por exemplo, https://g1.globo.com/fato-ou-fake/noticia/2024/02/14/e-fake-que-suco-de-limao-e-caldo-de-cana-curam-dengue.ghtml

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

primeiro e logo depois fazer o seguinte procedimento: tomar de hora em hora 50 ml de caldo de cana. Pode tomar esse volume de caldo de cana por umas seis horas consecutivas. Esses dois procedimentos vão eliminar a carga viral e aumentar muito rapidamente as plaquetas no sangue.

Atualmente, tais desafios têm ganhado proporções ainda maiores. Há contas forjadas e robôs nas redes sociais que replicam grandes volumes desse tipo de conteúdo (para fins diversos). O advento da Inteligência Artificial generativa, apesar de sua utilidade, também traz preocupações, pois há muita desinformação produzida (dadas as limitações ainda existentes nas técnicas empregadas e seu possível mal uso), trazendo ao cenário não apenas notícias falsas produzidas por humanos, mas também conteúdo enganoso gerado por máquinas (Su et al., 2024).

Soma-se a esse cenário a dificuldade humana de detectar conteúdo enganoso em geral. Pesquisas já mostraram que humanos não conseguem separar satisfatoriamente notícias verdadeiras de notícias falsas (Bond Jr.; DePaulo, 2006; George; Keane, 2006), alcançando entre 50% e 63% de sucesso, dependendo do que seja considerado enganoso (Rubin; Conroy, 2011). Algumas explicações para isso são:

- a falta de conhecimento prévio das pessoas sobre determinados assuntos, fato este relacionado a questões de formação e educação formal, muitas vezes;
- a percepção humana da veracidade da notícia pode ser afetada pelo fato de as informações veiculadas estarem alinhadas com suas crenças e opiniões pré-existentes, o que tem sido chamado de "viés da confirmação" (Thakar; Bhatt, 2024);
- algumas pessoas recebem informações de fontes socialmente próximas, o que as leva a considerar legítimas as informações compartilhadas, além do fato de os usuários raramente verificarem as informações que compartilham (Tandoc Jr. et al., 2018);
- a dificuldade, a falta de interesse ou a falta de disponibilidade do usuário para seguir dicas que são amplamente divulgadas por sites de notícias e redes sociais, como o Facebook, cujas orientações são mostradas no Quadro 8.2, conforme reproduzidas por Santos (2022).

Quadro 8.2: Diretrizes do Facebook para identificação de notícias falsas (Santos, 2022).

1. **Seja cético com as manchetes**. Notícias falsas frequentemente trazem manchetes apelativas em letras maiúsculas e com pontos de exclamação. Se alegações chocantes na manchete parecerem inacreditáveis, desconfie.
2. **Olhe atentamente para a URL**. Uma URL semelhante à de outro site ou um telefone podem ser um sinal de alerta para notícias falsas. Muitos sites de notícias falsas imitam veículos de imprensa autênticos fazendo pequenas mudanças na URL. Você pode ir até o site para verificar e comparar a URL com a de veículos de imprensa estabelecidos.
3. **Investigue a fonte.** Certifique-se de que a reportagem tenha sido escrita por uma fonte confiável e de boa reputação. Se a história for contada por uma organização não conhecida, verifique a seção "Sobre" do site para saber mais sobre ela.
4. **Fique atento a formatações incomuns.** Muitos sites de notícias falsas contêm erros ortográficos ou apresentam layouts estranhos. Redobre a atenção na leitura se perceber esses sinais.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

5. **Considere as fotos.** Notícias falsas frequentemente contêm imagens ou vídeos manipulados. Algumas vezes, a foto pode ser autêntica, mas ter sido retirada do contexto. Você pode procurar a foto ou imagem para verificar de onde ela veio.
6. **Confira as datas.** Notícias falsas podem conter datas que não fazem sentido ou até mesmo datas que tenham sido alteradas.
7. **Verifique as evidências.** Verifique as fontes do autor da reportagem para confirmar que são confiáveis. Falta de evidências sobre os fatos ou menção a especialistas desconhecidos pode ser uma indicação de notícias falsas.
8. **Busque outras reportagens.** Se nenhum outro veículo na imprensa tiver publicado uma reportagem sobre o mesmo assunto, isso pode ser um indicativo de que a história é falsa. Se a história for publicada por vários veículos confiáveis na imprensa, é mais provável que seja verdadeira.
9. **A história é uma farsa ou uma brincadeira?** Algumas vezes, as notícias falsas podem ser difíceis de distinguir de um conteúdo de humor ou sátira. Verifique se a fonte é conhecida por paródias e se os detalhes da história e o tom sugerem que pode ser apenas uma brincadeira.
10. **Algumas histórias são intencionalmente falsas.** Pense de forma crítica sobre as histórias lidas e compartilhe apenas as notícias que você sabe que são verossímeis.

8.1.3 Combate ao conteúdo enganoso

Como a propagação de conteúdo enganoso alcançou um ponto crítico, a partir das evidências de que as redes sociais têm uma parcela grande de responsabilidade pela proliferação desse tipo de conteúdo, somada à dificuldade humana de detectar tais conteúdos, surgiram iniciativas de combate a essa prática.

Agências de comunicação têm dado apoio a sites de checagem manual de fatos e a companhias de grande apelo digital. No Brasil, a checagem de fatos é em grande parte feita por jornalistas, em sites e agências como o Boatos.org, Fato ou Fake[4], E-Farsas[5], Aos Fatos[6] e Agência Lupa[7], entre outros. Eles costumam seguir uma metodologia de checagem de acordo com as diretrizes da Rede Internacional de Checagem de Fatos (em inglês, *International Fact-Checking Network* ou IFCN), que promove iniciativas de verificação de fatos[8].

Estudos sobre a produção e a disseminação de notícias falsas têm sido realizados buscando-se mapear o fenômeno e seu comportamento. Vosoughi et al. (2018) relatam que, em seus estudos, cerca de 1% das notícias falsas mais populares foram difundidas de 1.000 a 100.000 pessoas, enquanto as notícias verdadeiras raramente chegam a 1.000 pessoas, o que pode ser explicado pelo apelo emotivo e pelo grau de novidade das notícias falsas.

No âmbito das pesquisas de PLN, há esforços para estudar as características do conteúdo enganoso e como detectá-lo automaticamente. Considerando toda a história da área, as tentativas são relativamente recentes, tanto do ponto de vista teórico quanto do prático, sendo majoritariamente focadas no processamento do conteúdo escrito. Segundo Hauch et

[4] https://g1.globo.com/fato-ou-fake
[5] http://www.e-farsas.com
[6] https://aosfatos.org
[7] https://lupa.uol.com.br
[8] https://ifcncodeofprinciples.poynter.org

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

al. (2012), a automatização do processo de detecção de conteúdo enganoso é atraente por no mínimo dois motivos: (i) esses sistemas podem ser mais objetivos do que os humanos, que são propensos a erros e vieses; e (ii) julgamentos online de várias pistas a partir de vídeos ou áudios podem sobrecarregar o humano e levar a atrasos e erros.

Em PLN, o uso de aprendizado de máquina, tanto da linha mais clássica quanto da linha profunda, incluindo os LLMs, são atualmente abordagens bastante comuns para a identificação de notícias falsas. Essas abordagens visam aprender automaticamente as características textuais (simples ou complexas) que podem se correlacionar com a veracidade/falsidade da informação, usando-as para identificar a classe de notícias novas. Há também outras abordagens, por exemplo, aquelas que buscam identificar fatos e checá-los automaticamente, assim como usar técnicas baseadas em recuperação e extração de informação.

Normalmente, essas abordagens baseiam-se em dados encontrados em *corpora* que contêm notícias falsas e verdadeiras, assim como metadados, que servem de base para o desenvolvimento/treino e avaliação/teste dos métodos de detecção de notícias falsas (para mais detalhes sobre *corpora*, ver Capítulo Conjunto de dados, *dataset* e *corpus*). A seguir, os principais *corpora* existentes para a tarefa de detecção de notícias falsas para a língua portuguesa são brevemente apresentados.

8.2 *Corpora*

A análise de notícias falsas tem se beneficiado da disponibilidade de variados *corpora*. Para o inglês, existem vários *corpora* em diferentes domínios, como política (Gruppi et al., 2020, 2021; Vlachos; Riedel, 2014; Wang, 2017), medicina (Dai et al., 2020), opiniões sobre produtos ou hotéis (Fornaciari; Poesio, 2013; Ott et al., 2011; Salminen et al., 2022) e tópicos sociais, como aborto (Pérez-Rosas; Mihalcea, 2014). Apesar de existir uma quantidade significativamente menor de *corpora* em português, nos últimos anos esse problema tem diminuído com o surgimento de *corpora* específicos para o português do Brasil, com características distintas que influenciam diretamente a eficácia dos modelos de detecção de notícias falsas.

A Tabela 8.1 lista os principais *corpora* de notícias falsas para o português do Brasil de que se tem ciência (na data da escrita desse capítulo), organizados cronologicamente conforme seu surgimento.

Tabela 8.1: *Corpora* de notícias falsas em português do Brasil.

Corpus	Quantidade de notícias	Categorias envolvidas	Período do conteúdo	Tipo de material
Fake.br (Monteiro et al., 2018; Silva et al., 2020b)	7.200	falsa, verdadeira	2016 a 2018	Notícias
FACTCK.BR (Moreno; Bressan, 2019)	1.309	falsa, meia-verdade, verdadeira	2016 a 2019	Notícias
Bracis2019FakeNews – Twitter (Faustini; Covões, 2019)	8.981	falsa, verdadeira	—	Postagens do X (Twitter)

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Corpus	Quantidade de notícias	Categorias envolvidas	Período do conteúdo	Tipo de material
Bracis2019FakeNews – WhatsApp (Faustini; Covões, 2019)	177	falsa, verdadeira	—	Mensagens do WhatsApp
FakeTweet.Br (Cordeiro; Pinheiro, 2019)	299	falsa, verdadeira	—	Postagens do X (Twitter)
FakeWhastApp.BR (Cabral et al., 2021)	5.284	falsa, verdadeira	—	Mensagens do WhatsApp
FakeRecogna (Garcia et al., 2022)	11.902	falsa, verdadeira	2018 a 2021	Notícias
FakePedia (Charles et al., 2022)	12.398	falsa, verdadeira	2013 a 2021	Notícias
FakeTrueBR (Chavarro et al., 2023)	3.582	falsa, verdadeira	—	Notícias
Fake tweets Covid-19 (Geurgas; Tessler, 2024)	14.000	falsa, verdadeira	—	Postagens do X (Twitter)
FakeRecogna 2.0 (Garcia et al., 2024b)	52.800	falsa, verdadeira	2002 a 2023	Notícias

O primeiro *corpus* para o português de que se tem conhecimento é o Fake.br (Monteiro et al., 2018; Silva et al., 2020b), com notícias verdadeiras e falsas alinhadas, ou seja, cada notícia falsa tem uma notícia verdadeira relacionada. O Quadro 8.3 ilustra um caso de notícias alinhadas do *corpus*.

Quadro 8.3: Exemplo de notícia falsa e sua contraparte verdadeira no *corpus* Fake.br.

Notícia falsa	**Notícia verdadeira**
Polos magnéticos da Terra podem se inverter e causar colapso mundial: A Terra ficaria inabitável. Ao menos 100 mil pessoas morreriam por ano pela alta nos níveis de radiação espacial. Se o campo magnético continuar a diminuir e os polos magnéticos se inverterem, a Terra pode acabar como Marte – um local seco, árido e incapaz de preservar a vida.	Inversão dos polos magnéticos da Terra pode ocorrer mais rápido do que o previsto. Segundo afirmações, essas ocorrências são, a princípio, indistinguíveis das verdadeiras mudanças nos polos. Apesar dessas reversões não representarem qualquer ameaça à humanidade, os especialistas alertam que poderão gerar falhas nos satélites que orbitam a Terra.

Como se pode ver na Tabela 8.1, o maior *corpus* conhecido é o FakeRecogna 2.0 (Garcia et al., 2024b), com mais de 50.000 notícias.

Os *corpora* possuem características variadas. Por exemplo, alguns deles fornecem informações temporais, enquanto outros não. Informações temporais são importantes porque já se percebeu na área que os sistemas computacionais de detecção de notícias falsas podem ser fortemente ancorados nos tópicos de determinado período. Na coluna "Período do conteúdo" da Tabela 8.1, é apresentado o período abrangido pelos dados apenas quando essa informação está disponível. Além disso, a informação é baseada nos metadados encontrados no *corpus*, o que pode ser diferente da descrição do artigo principal do *corpus*.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Por exemplo, Garcia et al. (2024b) citam que as notícias do *corpus* FakeRecogna 2.0 foram coletadas entre 2019 e 2023, mas, na análise dos metadados, foram encontradas notícias publicadas em 2002.

Os dados também variam em termos de aplicação. Existem *corpora* que incluem notícias em seu sentido mais estrito (o jornalístico), como o FakeRecogna, FakeRecogna 2.0 e o Fake.br. Outros usam dados relativos a postagens em redes sociais, como o FakeTweet.Br. Além disso, há *corpora* voltados para dados de aplicativos de mensagens instantâneas, como o WhatsApp, incluindo o Bracis2019FakeNews e o FakeWhastApp.BR.

Outro ponto de variação entre os *corpora* são as classes de veracidade avaliadas. A maioria trabalha com uma classificação binária, distinguindo entre notícias falsas e verdadeiras. Porém, o *corpus* FACTCK.BR contém três classes: falsa, meia-verdade e verdadeira.

Por fim, algumas bases de dados, como a Fake.br, fornecem, além dos textos, metadados e atributos linguísticos. Esses recursos adicionais enriquecem as possibilidades de análise, fornecendo aos pesquisadores informações estruturadas que podem ser diretamente utilizadas nos modelos de classificação. Embora seja possível extrair alguns desses atributos diretamente dos textos, sua inclusão prévia no *corpus* facilita o processo de pesquisa e gera maior consistência nos resultados.

Na próxima seção, são discutidos alguns métodos de detecção automática de notícias falsas, focando-se nos trabalhos existentes para a língua portuguesa, muitos dos quais se baseiam nos *corpora* aqui apresentados.

8.3 Métodos de detecção automática

Na busca por identificar notícias falsas de forma automática, diversos estudos na literatura exploram uma ampla gama de abordagens. Muitos desses estudos podem ser categorizados de acordo com a taxonomia do comportamento enganoso proposta por Zhou (2005), reproduzida na Figura 8.1. Essa taxonomia ajuda a organizar o comportamento enganoso, auxiliando (i) na sistematização de sinais e atributos que podem indicar se uma informação é falsa e (ii) na criação de métodos computacionais.

A taxonomia se divide em Linguagem Verbal e Não Verbal. Os indicadores não verbais focam nas características exibidas enquanto uma pessoa produz conteúdo enganoso, enquanto que os verbais estão ligados ao conteúdo escrito e falado da língua.

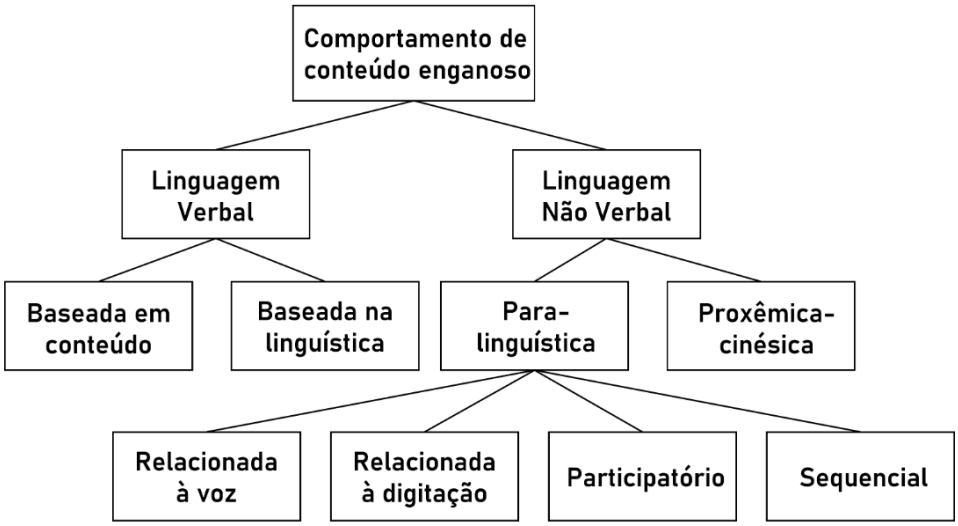
A paralinguística, um dos subgrupos da linguagem não verbal, foca em propriedades do discurso que não estão diretamente relacionadas ao conteúdo falado. Isso inclui propriedades relacionadas à voz, como o tom de voz e pausas na fala, relacionados ao teclado (velocidade de digitação, quantidade de vezes que apagou a mensagem, erros de digitação por proximidade de teclas), participatório (demora entre as respostas, mudança de assunto) e sequencial (de quem veio a iniciação de conversas). O segundo subgrupo da linguagem não verbal se refere ao comportamento proxêmico-cinésico do enganador. Ele envolve movimentos e posturas corporais, como gestos com as mãos e braços, expressões faciais e movimentos oculares, entre outros.

Apesar da linguagem não verbal ser importante, as pesquisas de PLN costumam atuar mais diretamente no grupo de Linguagem Verbal. Esse grupo é dividido em comportamento baseado em **conteúdo** e na **linguística**. Nas abordagens baseadas em linguística, são usadas pistas provenientes do texto para fazer a diferenciação entre conteúdo falso ou verdadeiro. Assume-se que os produtores de conteúdo enganoso podem inconscientemente deixar rastros linguísticos que sinalizam a falsidade. Por outro lado, a abordagem baseada

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Figura 8.1: Taxonomia do comportamento enganoso.



Fonte: Adaptada de (Zhou, 2005)

em conteúdo busca simular a metodologia de checagem de fatos, avaliando cada informação apresentada, muitas vezes remetendo a abordagens baseadas em grafos de conhecimento.

A seguir, descrevem-se alguns métodos de detecção de notícias falsas, principalmente os aplicados para o português e que se alinham ao grupo da Linguagem Verbal. Para fins didáticos, os métodos são organizados em abordagens computacionais mais claramente discerníveis (na época da escrita deste capítulo), sendo que se busca apresentar pelo menos um exemplo típico pertencente a cada abordagem.

8.3.1 Aprendizado de máquina com atributos linguísticos

De acordo com Conroy et al. (2015), a maioria dos enganadores usa a linguagem estrategicamente para evitar que sejam descobertos. Nesse caso, o objetivo da abordagem linguística é procurar por exemplos de vazamentos – ou os chamados "sinais de engano preditivo" – presentes no conteúdo de uma mensagem. Isso demanda uma análise mais profunda do texto na busca de padrões que possam indicar se uma informação é enganosa ou não.

Para identificar esses padrões, a análise pode utilizar atributos linguísticos que possam diferenciar o conteúdo verdadeiro do falso, extraídos de elementos como classes gramaticais, semântica, erros ortográficos, expressividade e diversidade de conteúdo para representar as amostras textuais, a frequência de uso de certas palavras e o padrão de capitalização do texto.

Monteiro et al. (2018) e Silva et al. (2020b) apresentaram a primeira abordagem linguística para a detecção de notícias falsas em português, seguindo o que talvez possa ser definida como uma das abordagens mais clássicas para o problema. Usando o *corpus* Fake.br, o processo de detecção de notícias falsas foi modelado como uma tarefa de aprendizado de máquina, em que as notícias são as instâncias, a classificação é binária (notícia falsa ou verdadeira) e os atributos das notícias utilizados para a classificação incluem informações sobre os *tokens* presentes nas notícias e atributos linguísticos derivados.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Para representar os *tokens* de cada notícia, foram utilizadas a representação de *bag of words* (com ponderação via *Term Frequency-Inverse Document Frequency* ou TF-IDF) e a média das representações vetoriais dos *tokens* (com 300 dimensões) produzidas pelos métodos Word2Vec (versão skip-gram) (Mikolov et al., 2013b) e FastText (Bojanowski et al., 2017) treinados para o português (Hartmann et al., 2017) (Capítulo Representação vetorial e semântica distribucional). Em relação aos atributos linguísticos, foram utilizados (i) atributos sugeridos por Zhou et al. (2004), que são pausalidade, emotividade, incerteza e não imediatismo, e (ii) atributos mais gerais como diversidade, tamanho médio das sentenças, tamanho médio das palavras e número de erros ortográficos na notícia. A pausalidade indica a ocorrência de pausas no texto, que é computada como o número de sinais de pontuação sobre o número de sentenças. A emotividade mede a expressividade da linguagem, calculada como a soma do número de adjetivos e advérbios dividida pela soma do número de substantivos e verbos. A incerteza é baseada nas ocorrências de verbos modais e no uso da voz passiva. O recurso de não imediatismo é baseado na frequência de uso dos pronomes de 1a e 2a pessoa. A diversidade é calculada como o número total de *tokens* de conteúdo diferentes sobre o número total de *tokens* de conteúdo (ou seja, é uma versão mais refinada da conhecida taxa *type-token*).

Os autores avaliaram diferentes abordagens de pré-processamento dos dados e técnicas de Aprendizado de Máquina, incluindo, por exemplo, regressão logística, árvores de decisão, floresta aleatória e máquinas de vetores de suporte, além de testarem métodos de *ensemble* e *stacking*. Um ponto a destacar é que, como as notícias verdadeiras do *corpus* utilizado costumam ser maiores do que as falsas, os autores truncaram os maiores textos para equalizar os tamanhos e evitar vieses na classificação automática. De forma geral, os autores concluem que os melhores resultados foram produzidos pela técnica de *ensemble*, aproximando-se de 97% de acurácia na classificação.

Há várias outras iniciativas para o português que seguiram uma abordagem similar, com alguma variação no conjunto de atributos linguísticos utilizados. Veja, por exemplo, os trabalhos de Faustini; Covões (2019) e Okano; Ruiz (2019). Há também trabalhos que buscaram avaliar técnicas independentes de língua, incluindo o português entre seus experimentos, como os trabalhos de Abonizio et al. (2020) e Faustini; Covões (2020).

8.3.2 Modelos de linguagem

Modelos de aprendizado profundo têm sido utilizados em várias pesquisas que tratam do problema das notícias falsas. Em particular, modelos de linguagem, como o BERT e T5 (Capítulo Modelos de linguagem), podem ser empregados tanto para gerar representações vetoriais dos textos, que servem como entrada para métodos de classificação, quanto para atuar diretamente como classificadores.

Nesse contexto, um estudo que ilustra bem essa abordagem é o de Vieira et al. (2025), que avaliou vários modelos de geração de representação vetorial, comparando modelos tradicionais de classificação que incluíram floresta aleatória, máquinas de vetores de suporte e rede neural perceptron multicamadas. Para a geração das representações vetoriais, eles aplicaram os seguintes modelos especializados para o português: BERTimbau, Teeny-TinyLlama e Tucano. Também aplicaram modelos voltados para o inglês, como BERT, RoBERTa e BART. Por fim, aplicaram modelos multilíngues: mBERT, XLM-RoBERTa e mBART. Todos os métodos com os diferentes modelos de representação foram testados na base de dados Fake.Br. De acordo com os autores, a melhor combinação em termos de desempenho foi o mBART com o método de máquinas de vetores de suporte, atingindo

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

uma acurácia de 97,43%. Entretanto, ao considerar o tempo de processamento, a combinação BERTimbau com o mesmo método apresentou uma acurácia de 97,22% e um tempo de processamento significativamente menor (aproximadamente 86% inferior), oferecendo assim um bom equilíbrio entre precisão e eficiência.

Na mesma linha, Garcia et al. (2024b) avaliaram o desempenho de modelos tradicionais e de aprendizado profundo sobre o *corpus* FakeRecogna 2.0. Entre os modelos tradicionais, testaram regressão logística, perceptron multicamadas, *naive* Bayes, *optimum-path forest* (Papa et al., 2009), floresta aleatória e máquinas de vetores de suporte. No que diz respeito aos modelos de aprendizado profundo, testaram LSTM, GRU, CNN e BERT. Para gerar a representação vetorial dos textos nos experimentos com métodos tradicionais, utilizaram a frequência dos termos, a representação TF-IDF e o FastText. Nos experimentos com modelos de aprendizado profundo, utilizaram o FastText para LSTM, GRU e CNN, o BERTimbau (Souza et al., 2020a) como a versão do BERT para o português, e o modelo PTT-5 (Portuguese, Tagalog, Turkish, Tamil e Telugu) (Carmo et al., 2020b) para o T5.

É interessante destacar que a abordagem proposta pelos autores se diferencia de outras metodologias ao implementar uma etapa de sumarização das notícias verdadeiras antes de treinar os modelos de classificação (em vez de truncar os maiores textos, como comentado anteriormente neste capítulo). Dessa forma, a diferença de tamanho entre notícias verdadeiras e falsas é resolvida. Os autores adotaram tanto a sumarização extrativa quanto a abstrativa. Garcia et al. (2024b) afirmam que a sumarização extrativa é vantajosa, pois tende a ser imune a inconsistências e "alucinações", já que o resumo final é composto pelas sentenças mais relevantes sem a geração de novas palavras ou expressões. Por outro lado, a sumarização abstrativa permite a produção de sentenças inéditas que podem diferir do texto original em termos de semântica e estrutura das sentenças, mas possui como vantagem ser mais semelhante à escrita humana.

Os melhores resultados obtidos pelos autores foram alcançados com o BERT, que apresentou acurácia aproximada de 98%. O modelo T5 apresentou resultados ligeiramente inferiores.

Além dos métodos de aprendizado profundo aplicados por Garcia et al. (2024b), outros estudos têm utilizado abordagens alternativas, como redes neurais de atenção hierárquica (Okano et al., 2020) e redes neurais de grafos (De Souza et al., 2024). No primeiro trabalho, a rede neural de atenção hierárquica alcançou uma acurácia de 90% (com textos truncados). No segundo trabalho, os autores propuseram um método que incorpora um mecanismo de atenção na abordagem de aprendizado não rotulado por propagação de rótulos utilizando uma rede neural baseada em grafo. O método foi avaliado em seis *corpora* de notícias falsas, representando diferentes cenários quanto à linguagem, fonte, tópicos e balanceamento. Nos testes realizados com o *corpus* Fake.br, o método alcançou acurácia de cerca de 78%.

Outros métodos já utilizados para a detecção de notícias falsas que empregaram textos em português do Brasil incluem variantes e aprimoramentos do BERT, como o modelo de linguagem DistilBERT. Um exemplo é o trabalho de Gôlo et al. (2023), que apresenta o MVAE-FakeNews (codificadores variacionais multimodais para detecção de notícias falsas). Este método multimodal representa os textos na detecção de notícias falsas por meio de aprendizado baseado em uma única classe, aprendendo uma nova representação a partir da combinação de modalidades promissoras para dados de notícias, combinando informações de linguagem, características linguísticas e tópicos, sendo que a modalidade de linguagem é derivada das informações do modelo DistilBERT. Neste trabalho, os autores utilizam o algoritmo de máquinas de vetores de suporte com uma classe para fazer a classificação. Os resultados indicam que o MVAE-FakeNews alcançou valores de acurácia melhores do que

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

diversos outros métodos, demonstrando que abordagens multimodais podem ser melhores do que abordagens unimodais que concatenam os dados de entrada. Nos experimentos com o *corpus* Fake.br, o melhor valor de acurácia obtido foi de 65% (utilizando 10% de notícias falsas rotuladas), com os métodos baseados em DistilBERT apresentando os melhores desempenhos.

O trabalho de Pires; Silva (2024) também avaliou variantes do modelo BERT, mais especificamente o BERTimbau e o mBERT, comparando seus resultados com o BERT. Eles testaram esses métodos nas seguintes bases de dados: Fake.Br, FakeTrueBR, FakeRecogna e Fakepedia. De acordo com os autores, os resultados obtidos indicam uma superioridade do BERTimbau sobre o BERT e sobre o mBERT na referida tarefa, com uma melhora média da medida F1 de 2,37% e 1,07%, respectivamente.

Outro trabalho que avaliou o mBERT é Fischer et al. (2022) comparando-o com outros modelos de classificação, como regressão logística, árvores de decisão, floresta aleatória, k-vizinhos mais próximos, SVM, naive Bayes, XGBoost, CNN, GRU e LSTM. O mBERT obteve uma medida F1 de 0,98 superando os outros métodos avaliados.

Em paralelo a essas abordagens, a aplicação de LLMs (Capítulo Modelos de linguagem) na detecção de notícias falsas tem ganhado atenção, especialmente em trabalhos em inglês, como os de Lucas et al. (2023) e Su et al. (2024), que discutem não apenas a detecção de notícias falsas com LLMs, mas também a geração desse conteúdo.

O uso de LLMs para detectar notícias falsas não é trivial, pois o treinamento desses modelos é custoso, resultando em atualizações pouco frequentes, o que muitas vezes os impede de fornecer informações atualizadas sobre determinados assuntos. No entanto, uma solução para isso é a técnica de *Retrieval-Augmented Generation* (RAG), que permite fornecer informações atualizadas aos LLMs. Essa técnica utiliza conceitos de recuperação de informação para buscar, em uma base de dados complementar, os conteúdos mais relevantes para uma pergunta (especificada na forma de um *prompt*) e, em seguida, passar essas informações ao LLM como contexto. Assim, o LLM gera respostas que, em vez de se basearem apenas em seu conhecimento prévio, são fundamentadas nos dados adicionais fornecidos.

Um exemplo relevante no contexto de notícias falsas é apresentado por Yue et al. (2024), que propõe o framework RARG (*Retrieval Augmented Response Generation*) para combater a desinformação online, especialmente no contexto da COVID-19. O RARG gera respostas baseadas em evidências científicas para refutar informações falsas, funcionando em duas etapas: (1) recuperação de evidências, onde documentos relevantes são coletados e reclassificados a partir de uma base com mais de um milhão de artigos acadêmicos; e (2) geração de respostas, em que um LLM é treinado para produzir respostas polidas e factuais com base nas evidências obtidas. Além disso, o modelo é otimizado com aprendizado por reforço com feedback humano (Capítulo Modelos de linguagem), garantindo o uso eficaz das evidências sem comprometer a qualidade das respostas.

Yue et al. (2024) fizeram experimentos com o RARG usando dados tanto de dentro quanto de fora do domínio da COVID-19 para a língua inglesa. Eles realizaram uma avaliação comparativa, de formas qualitativa e quantitativa, para analisar as respostas geradas por diferentes modelos. As métricas quantitativas utilizadas incluíram critérios de refutação, factualidade, polidez, e relevância da resposta em relação à alegação de entrada e à evidência. Foram avaliados modelos baseados em geração de texto, como BART, DialoGPT e GODEL, além de métodos de aprendizado por reforço, como PARTNER e MisinfoCorrect, e LLMs como Llama 2 e GPT-3.5. As análises revelaram que o modelo RARG obteve o melhor desempenho geral.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Para o português do Brasil, já existem algumas contribuições relevantes que indicam o potencial crescente dessa linha de pesquisa. Por exemplo, o estudo de Gôlo et al. (2024) explora modelos de código aberto, pouco aplicados ao contexto de notícias falsas. Os pesquisadores criaram um novo conjunto de dados de notícias políticas brasileiras e realizaram uma avaliação empírica de LLMs de código aberto e da OpenAI. O Gemma 2 do Google superou outros seis LLMs (Llama 3, Phi 3, Qwen 1, Qwen 2, Gpt 3.5 turbo e GPT 4), mostrando-se o mais promissor para essa tarefa, atingindo valores de F1 de 0,90. Além do desempenho de classificação, o estudo avaliou a capacidade dos LLMs de gerar explicações alinhadas ao raciocínio humano (com destaque para os modelos de melhor desempenho) e constatou que os modelos Gemma e Qwen produziram os melhores embeddings, separando com clareza notícias falsas e verdadeiras em visualizações via t-SNE.

Em outro estudo relacionado, Silva et al. (2025) avaliam o impacto de notícias falsas geradas por LLMs no desempenho de métodos de aprendizado de máquina na classificação. A pesquisa utilizou diversos modelos para classificar notícias falsas, incluindo algoritmos de aprendizado de máquina tradicionais (árvore de decisão, regressão logística, floresta aleatória, máquinas de vetores de suporte e gradiente descendente estocástico), um modelo BERTimbau ajustado, um LLM puro e dois LLMs combinados com RAG. Os experimentos foram realizados usando os corpora Fake.Br e FakeTrueBR pelo fato deles serem alinhados, isto é, cada notícia falsa possui sua contraparte verdadeira. Os pesquisadores avaliaram o desempenho desses modelos com notícias falsas sintéticas, criadas com o modelo de LLM Sabiá-3. Os resultados revelaram uma degradação significativa no desempenho dos modelos de aprendizado de máquina quando avaliados em condições de incompatibilidade, ou seja, quando treinados com conteúdo gerado por humanos e testados com notícias falsas geradas por LLMs, ou vice-versa. Por exemplo, a medida F1 do BERTimbau foi de 0,99 no experimento onde o conjunto de treino e teste eram compostos por notícias escritas por humanos, mas caiu para 0,69 quando o conjunto de treino era composto por notícias geradas por humanos e o conjunto de teste era formado por notícias geradas por LLMs. Observou-se ainda que classificadores baseados em LLM foram mais robustos sob condições de incompatibilidade, enquanto que o BERTimbau e os modelos tradicionais de aprendizado de máquina apresentaram melhor desempenho quando não havia essa diferença. Os autores afirmam que os resultados obtidos ressaltam a necessidade de abordagens mais resilientes e adaptativas diante da crescente sofisticação do conteúdo gerado por LLMs.

8.3.3 Checagem de conteúdo

Para abordagens de checagem automática de conteúdo, em que os fatos em si devem ser verificados (o que vai além de avaliar estilo de escrita e atributos linguísticos), Thorne; Vlachos (2018) sugerem que se extraiam triplas de informação dos documentos verdadeiros (por algum processo de extração de informação – ver Capítulo 3), normalmente formadas por <sujeito, verbo, objeto>, que constituem os fatos básicos mencionados. Nem sempre é trivial fazer isso, pois pode haver sentenças com número variável de argumentos (ausência de objeto ou mais de um objeto, por exemplo) ou com estruturas sintáticas mais complexas ou não usuais. Com base nessas triplas, pode-se construir um grafo de conhecimento, em que sujeitos e objetos dão origem a vértices do grafo e os verbos definem arestas entre eles. O grafo é, dessa forma, uma representação estruturada do conteúdo verídico, podendo ser utilizado para avaliação da veracidade de novas informações, o que pode ser feito de diferentes formas.

Santos; Pardo (2020) adaptam para o português a abordagem de Ciampaglia et al.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

(2015). Nessa abordagem, o grafo de conhecimento é constituído de triplas extraídas da Wikipédia, considerada como uma fonte de informações verídicas sobre o mundo. Apresentada uma nova informação a ser checada, as triplas dessa informação são extraídas e analisadas com o apoio do grafo de conhecimento: caminhos distintos entre sujeitos e objetos, com diferentes distâncias, fornecem suportes factuais diferentes para as afirmações. São consideradas verdadeiras as afirmações cujos sujeitos e objetos apresentam conexões diretas entre si ou caminhos de distância relativamente curta. A proximidade entre os conceitos envolvidos no grafo reflete, portanto, a veracidade da informação.

Os autores testam a abordagem com afirmações históricas e geográficas, em especial, capitais de estados e países e cônjuges de presidentes brasileiros. Em média, os resultados atingem valores de acurácia de 74%. Essa abordagem, apesar de exigir análise linguística mais refinada para produção do grafo de conhecimento e de produzir resultados inferiores em relação às outras abordagens, é interessante por permitir rastrear a veracidade de cada fato e explicar os resultados. Como amplamente discutido atualmente na Inteligência Artificial, a explicabilidade dos métodos é uma característica muito relevante (ver, por exemplo, o Capítulo Questões éticas em IA e PLN).

Há várias outras estratégias de uso de grafos que não necessariamente buscam os fatos explicitamente. Santos; Pardo (2021), por exemplo, testam outras formas de construção e percurso em grafos (tanto para notícias verdadeiras como falsas) e o uso de medidas de redes complexas para caracterização estrutural de notícias falsas e verdadeiras, que se mostram promissoras (e às vezes de mais simples aplicação), mas com resultados inferiores aos obtidos anteriormente.

8.3.4 Recuperação de informação

Santos; Pardo (2020) exploram uma estratégia baseada no comportamento humano de checagem de informação: usa-se um sistema de recuperação de informação (no caso, o Google – ver Capítulo 2) para buscar pelo título da notícia de interesse, analisando-se os resultados retornados na busca. Caso haja indício de falsidade da notícia correspondente, ela é classificada dessa forma.

Os autores analisam apenas os links e os *snippets*[9] retornados na 1a página de resultados. Se houver link de algum site de verificação de notícias (como os citados no início desse capítulo, por exemplo, Boatos.org, Aos Fatos e Agência Lupa) ou houver algum termo indicativo de falsidade nos *snippets*, a notícia é classificada como falsa. Os termos indicativos considerados pelos autores são “fake”, “falso”, “mentira”, “calúnia”, “inverídico”, “enganoso”, “farsa”, “ilusório”, “ilegítimo”, “boato” e “rumor”, além de suas variações possíveis de gênero e número.

Os autores avaliam essa estratégia com o *corpus* Fake.br, indicando resultados de classificação com acurácia de 68%. Apesar do resultado limitado, a simplicidade desse método torna-o interessante, pois faz uso de serviços de web já disponíveis. Além disso, essa estratégia evidencia a necessidade de modelagens mais sofisticadas para o devido tratamento do fenômeno das notícias falsas.

8.3.5 Análise de credibilidade

Uma outra linha de pesquisa relacionada à detecção de notícias falsas é a análise de credibilidade. Newell et al. (2017) afirmam que a credibilidade das atribuições feitas

[9] Os *snippets* são os trechos de texto retornados junto a cada link na página de resposta do buscador web.

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

na notícia é fundamental para garantir a confiança nos meios de comunicação. Nesse sentido, Burriss (1988) comenta que o jornalismo deve lidar com fatos verificáveis, sendo princípio básico a seguir, pois o público que recebe as notícias depende (i) da reputação da organização noticiosa, (ii) da reputação do repórter e (iii) da informação dentro da própria história, a fim de determinar a exatidão de uma notícia. Por isso, é necessário buscar padrões linguísticos no texto e metadados que possam medir a credibilidade de uma notícia.

Um exemplo de trabalho que segue essa linha é de Couto et al. (2024), que analisou padrões de rede de fontes de notícias brasileiras de alta e baixa credibilidade, focando em redes de desinformação. Esses padrões emergem de diversas fontes, como registros de domínios, certificados TLS, infraestrutura de hospedagem e geolocalização. Para conduzir essa análise, os autores criaram um conjunto de sites de baixa credibilidade, identificando aqueles que publicaram ao menos uma notícia cuja veracidade foi contestada por uma agência de checagem de fatos signatária da *International Fact-Checking Network* (IFCN). A partir dessa base, 31 atributos foram computados.

Segundo os autores, os resultados revelaram que apenas 12,2% dos sites de baixa credibilidade estão registrados e hospedados no Brasil, em contraste com 47% dos sites de alta credibilidade. Além disso, sites de alta credibilidade tendem a ter uma vida útil mais longa, domínios registrados por períodos maiores e certificados TLS (em inglês, *Transport Layer Security*) com validade estendida. Foi também identificado que sites de baixa credibilidade são mais propensos a utilizar TLDs (em inglês, *Top Level Domain*) alternativos. Os autores visualizaram a separação entre sites de alta e baixa credibilidade em duas dimensões utilizando a técnica t-SNE, e treinaram um modelo de floresta aleatória que previu a credibilidade dos sites com bom desempenho, atingindo uma acurácia de aproximadamente 83%.

Além de analisar as fontes de notícias, também é possível analisar a credibilidade de usuários, usando abordagens que buscam identificar sinais de qualidade do conteúdo (Souza Freire et al., 2021) e os disseminadores de conteúdo enganoso, sejam robôs ou usuários comuns (Chakraborty et al., 2016; Speicher et al., 2018; Varol et al., 2017; Yang et al., 2020a). Esse tipo de abordagem se baseia em informações sobre os usuários para prever seu comportamento e as chances de que as notícias que compartilham sejam falsas, incluindo análises demográficas, como idade e localização. Essa linha de investigação é bastante usada em postagens de redes sociais, onde pode ser analisado o número de postagens, o perfil do usuário, sua rede de contatos e os dados das postagens em si. Esse nível de meta-análise pode ser combinado com os demais tipos de métodos para a identificação de conteúdo enganoso.

8.3.6 Outras iniciativas relacionadas

Sendo um campo de pesquisa vasto, como pode ser percebido pelos diversos tipos de conteúdo enganoso e abordagens computacionais existentes para detecção de notícias falsas, há algumas outras iniciativas que são importantes de serem citadas.

Há trabalhos que focam suas análises em atributos de inteligibilidade das notícias (Capítulo 6), buscando identificar padrões de escrita que possam sinalizar conteúdo falso. Santos et al. (2020b), por exemplo, avaliam dezenas de medidas de inteligibilidade textual, como índices de legibilidade, coesão, diversidade lexical, complexidade sintática, organização discursiva e informações psicolinguísticas, entre outros. Em experimentos com o *corpus* Fake.br, os autores verificaram que essas medidas contribuem para a discriminação entre

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

notícias falsas e verdadeiras.

Alguns trabalhos analisam o impacto da análise de sentimentos (Capítulo 16) para a determinação da veracidade do conteúdo. Alonso et al. (2021) apresentam uma discussão ampla sobre o assunto. Em um caso aplicado ao português, Jeronimo et al. (2020) evidenciam que a classificação de subjetividade pode auxiliar na tarefa de detecção de notícias falsas, argumentando que notícias verídicas que visam compartilhar informações factuais e imparciais tendem a usar uma linguagem mais objetiva que depende menos de pressupostos ou expressão sentimental.

Em uma vertente que utiliza análise linguística mais profunda, há trabalhos que buscam considerar padrões de organização do conteúdo textual que possam sinalizar a veracidade desse conteúdo, fazendo uso, por exemplo, de modelos discursivos, como a famosa *Rhetorical Structure Theory* (RST) (Mann; Thompson, 1988) (Capítulo Modelos discursivos). Buscam-se padrões de relações discursivas utilizadas e de distribuição de elementos mais e menos importantes (os "núcleos" e "satélites", que são os termos utilizados pelo modelo RST) para caracterizar o conteúdo enganoso. Vargas et al. (2022c) apresentam uma revisão desses trabalhos.

Alguns trabalhos combinam diversas das linhas apresentadas neste capítulo. Por exemplo, o trabalho de Vargas et al. (2024) envolve diversas etapas para realizar raciocínio factual em nível sentencial, incluindo a identificação de conteúdo relevante para analisar, a busca de evidências que suportem as informações e a geração de explicação para os resultados obtidos, combinando métodos e recursos variados de PLN.

Por fim, para o leitor mais interessado, Vargas et al. (2025), Thakar; Bhatt (2024) e Farhangian et al. (2024) apresentam olhares complementares aos tópicos abordados neste capítulo, incluindo visão geral de métodos, comparações e discussões de desafios e tendências na tarefa de detecção automática de notícias falsas, podendo ser referências adicionais importantes.

8.4 Impactos para a sociedade

São muitos os impactos que a incidência de conteúdo enganoso traz para as sociedades em suas diversas camadas, em todos os domínios da vida. O Brasil não é exceção, infelizmente.

Algumas áreas, como a saúde, são extremamente sensíveis a isso. Um exemplo pode ser visto no estudo empírico conduzido por Galhardi et al. (2020) sobre a pandemia de COVID-19. Os autores analisaram os relatos de usuários enviados ao aplicativo brasileiro "Eu Fiscalizo" durante o período de março a abril de 2020. Foi observado que 65% dos relatos propagavam orientações caseiras e incorretas para evitar a disseminação do vírus. Eles também notaram que 20% continham métodos para curar a doença. Esse problema não foi exclusividade no Brasil. Por isso, o termo "infodemia" passou a ser usado em vários veículos de comunicação e até mesmo pelo Organização Mundial de Saúde, significando que o mundo estava enfrentando, além de uma pandemia do coronavírus, uma pandemia de desinformação (Salvi et al., 2021; Zarocostas, 2020).

A desinformação também pode causar prejuízos econômicos. Um caso foi a disseminação de uma informação falsa em 2020 sobre um investimento significativo em uma empresa brasileira. Essa notícia enganosa levou muitos investidores a aplicarem recursos na empresa, resultando em grandes prejuízos quando a fraude foi descoberta, o que provocou uma queda acentuada nos preços das suas ações[10].

[10] https://www1.folha.uol.com.br/mercado/2021/11/cvm-abre-processo-contra-ex-executivos-do-irb-por-

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

As notícias falsas podem gerar problemas ainda mais graves, chegando até a resultar em mortes. Existem vários casos que mostram esse perigo. Um caso impactante ocorreu em 2014, quando uma pessoa foi espancada até a morte após supostamente ser alvo de uma notícia falsa que a acusava de sequestrar crianças para rituais de magia negra[11]. Em um caso mais recente, em 2024, outra pessoa foi morta após a disseminação de uma notícia supostamente falsa que afirmava que ela matava cães[12].

Notícias falsas também podem aumentar a polarização ao interferir no processo de troca de opiniões e evitar a formação de consenso dentro de uma sociedade. Isso pode ser evidenciado pelo estudo promovido por Azzimonti; Fernandes (2023), que simularam uma rede social sintética, baseada no Twitter (renomeada para X), para analisar a disseminação de desinformação. Os autores observaram que robôs, programados para promover visões tendenciosas, podem influenciar uma pequena parte dos usuários ao disseminarem notícias falsas. Essas notícias são propagadas por esses usuários, contaminando outros que confiam em suas opiniões. Isso dificulta a difusão de informações corretas e amplia a divisão e polarização no debate público.

Conclusões similares às obtidas no trabalho acima foram apresentadas por Oliveira (2023) no contexto brasileiro. A autora analisou a disseminação de informações falsas a partir da compreensão de cultura como uma teia de significados, que se constrói continuamente na interação discursiva das pessoas. O estudo explora as dinâmicas discursivas e comunicacionais durante as eleições brasileiras de 2022, apontando a relação entre a circulação de desinformação e a polarização crescente no Brasil. Segundo a autora, informações falsas produzem "verdades socialmente compartilhadas", que nem sempre se confirmam na realidade factual, mas que moldam o comportamento das pessoas, gerando intolerância e dissonância cognitiva, fatores que contribuem para a polarização social. O estudo reforça que a disseminação de desinformação nas redes sociais facilita a criação de realidades paralelas, exacerbando os conflitos e fomentando divisões sociais. Oliveira (2023) argumenta que essas "verdades compartilhadas", baseadas em conteúdos falsos, interferem na capacidade de diálogo e na coesão social, contribuindo para o adoecimento psicológico e o afastamento entre grupos, refletindo diretamente na polarização da sociedade.

Os diversos impactos sociais discutidos nesta seção, que vão desde a saúde pública até a qualidade dos debates públicos, as repercussões econômicas e a violência, evidenciam a relevância do desenvolvimento de técnicas para a detecção automática de notícias falsas. O contínuo avanço da Inteligência Artificial, juntamente com o surgimento de métodos que facilitam a criação de conteúdos enganosos cada vez mais sofisticados, aumentam a necessidade de estudos que busquem aprimorar os mecanismos automáticos capazes de identificar e mitigar esses problemas. A implementação eficaz dessas técnicas é essencial para auxiliar na proteção da sociedade dos efeitos prejudiciais da desinformação.

Visando transferir conhecimento e tecnologia para a sociedade, alguns estudos científicos resultam na produção de softwares disponibilizados para o público em geral. A pesquisa relatada por Monteiro et al. (2018), por exemplo, resultou no sistema FakeCheck[13] (treinado com o *corpus* Fake.br), disponível online e podendo ser utilizado em qualquer navegador da web. Sua interface pode ser vista na Figura 8.2.

manipulacao-de-mercado.shtml

[11] https://g1.globo.com/sp/santos-regiao/noticia/2014/05/mulher-espancada-apos-boatos-em-rede-social-morre-em-guaruja-sp.html

[12] https://g1.globo.com/sp/mogi-das-cruzes-suzano/noticia/2024/02/24/Como-fake-news-sobre-assassinato-de-cachorros-causou-a-morte-de-um-jovem-e-a-prisao-de-7-pessoas-veja-o-que-se-sabe.ghtml

[13] https://nilc-fakenews.herokuapp.com/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

Figura 8.2: FakeCheck – um sistema para detecção automática de notícias falsas.



Fonte: (Monteiro et al., 2018)

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

8.5 Considerações finais

Este capítulo abordou os principais aspectos relacionados às notícias falsas, incluindo os diferentes tipos de desinformação, os métodos mais utilizados na detecção, algumas bases de dados relevantes na língua portuguesa e os impactos significativos das notícias falsas na sociedade. Embora essa seja uma área de estudo altamente relevante, há ainda muitos desafios e aspectos que não foram explorados.

O capítulo se concentrou principalmente no problema sob a perspectiva do PLN, mas é importante reconhecer que o problema pode ser multimodal, ou seja, pode envolver não apenas texto, mas também imagens, vídeos e áudio. Dessa forma, a detecção eficaz de alguns tipos de notícias requer a integração de áreas como visão computacional, reconhecimento de voz e análise de redes sociais. A colaboração multidisciplinar é essencial para enfrentar a complexidade crescente do problema.

Outro desafio refere-se à disponibilidade de dados em quantidade e com a qualidade necessárias, dada a dificuldade de coleta do tipo de dado desejado, o que pode comprometer a modelagem linguístico-computacional dos fenômenos de interesse e o treinamento eficaz de sistemas. Nesse momento, é relevante explorar em mais detalhes alguns pontos relacionados à construção dos *corpora*, como a existência de possíveis vieses de classificação (Vargas et al., 2025), o aspecto temporal e os fenômenos que se pretende abordar.

A cobertura temporal das notícias dos *corpora* é uma questão de grande importância, sendo que se tem observado na área que muitos sistemas acabam não conseguindo lidar bem com tópicos novos que não existiam ou que não estavam em voga quando os dados de treinamento foram coletados. Por exemplo, no *corpus* Fake.br, não há notícias sobre a COVID-19 (uma vez que as notícias desse *corpus* foram coletadas antes da pandemia relacionada), e, como consequência, os sistemas com ele treinados podem não ser capazes de identificar com boa acurácia as notícias falsas e verdadeiras sobre o assunto. O tempo também pode afetar os trabalhos da área de maneiras inesperadas. Usando o mesmo *corpus* como exemplo, há uma notícia sobre a prisão de um político que era falsa na época em que a notícia foi veiculada; entretanto, depois de algum tempo, o político citado de fato foi preso. Sistemas que não modelem apropriadamente a questão temporal podem classificar incorretamente notícias assim. Sistemas baseados em LLMs com RAG, por exemplo, podem facilmente incidir em algum erro assim se restrições temporais não forem especificadas.

A importância do fator temporal foi enfatizada em um estudo recente de Wanderley et al. (2025), que aplicou métodos de detecção de desvio de conceito no corpus FakeRecogna 2.0 e mostrou que o fenômeno de desvio de conceito está presente em notícias brasileiras, especialmente nas notícias falsas. Os autores afirmam que pesquisas que ignoram a natureza temporal das notícias falsas provavelmente apresentam métricas de desempenho infladas. Segundo eles, isso expõe uma falha metodológica crítica nas práticas de pesquisa atuais e sublinha o perigo de implementar modelos que não conseguem se adaptar à evolução das narrativas de desinformação.

Também é importante destacar a limitação de que a maioria dos *corpora* existentes abrangem notícias inteiramente falsas ou verdadeiras, mas não cobrem outros tipos de fenômenos, como meias verdades e pós-verdades. Por um lado, é mais difícil identificar e coletar notícias desses tipos; por outro, pode ser bastante desafiador desenvolver sistemas de detecção desses fenômenos, que (ainda) acabam sendo relativamente pouco privilegiados na área. Outros desafios são pontuados no estudo qualitativo dos *corpora* em português apresentado por Baracho et al. (2025), que afirmam que comparado aos *corpora* da língua

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

inglesa, há problemas em termos de volume desses corpora em português, de diversidade temática, de multimodalidade e, em alguns casos, de balanceamento de classes.

Em relação ao desempenho dos sistemas computacionais, um ponto que necessita de mais discussão é o impacto dos erros na detecção de notícias falsas. Muitas vezes, em pesquisas na área, busca-se uma alta acurácia, maximizando os valores de verdadeiros positivos e verdadeiros negativos (usuais em avaliações de aprendizado de máquina) e, consequentemente, melhorando a precisão e a cobertura dos sistemas (Capítulo Avaliação de tecnologias de linguagem). Os erros cometidos são, frequentemente, discutidos de forma agnóstica nas pesquisas. Entretanto, se um sistema de detecção de notícias falsas é adotado no dia a dia, os tipos de erros podem ter importâncias diferentes. O que seria menos danoso para a sociedade: classificar erroneamente uma notícia verdadeira como falsa ou uma notícia falsa como verdadeira? Levados ao extremo, o primeiro caso pode ser classificado como uma tentativa de desqualificação do produtor do conteúdo ou de censura do próprio conteúdo, enquanto o segundo caso pode legitimar mentiras e trazer consequências desastrosas (como as citadas na seção anterior). Com os avanços da Inteligência Artificial, esse tipo de discussão é cada vez mais importante na área (Capítulo Questões éticas em IA e PLN), não se devendo esquecer de considerar fatores humanos, como a resistência do público em confiar em sistemas automatizados ou a tendência a politizar o uso dessas ferramentas.

Por fim, é importante salientar que a tecnologia é apenas um dos pilares para o combate às notícias falsas. A educação e a legislação são outros pilares: na frente educacional, é necessário preparar as novas gerações para lidar com a tecnologia, com os meios digitais e com a Inteligência Artificial, para que conheçam suas potencialidades e os riscos inerentes; em termos de legislação, é preciso avançar nas leis que permitam a responsabilização dos envolvidos na produção e na disseminação de notícias falsas, sendo este um tópico muito desafiador por envolver questões que podem perpassar interesses variados (por exemplo, interesses comerciais e políticos) e por caminhar em uma separação às vezes sutil entre liberdade de expressão, censura e ações possivelmente criminosas.

8.6 Exercícios de fixação

1. O capítulo apresenta diversas definições de conteúdo enganoso. Explique o que são notícias falsas e qual sua relação com os termos *misinformation* e *disinformation*, fornecendo um exemplo para cada uma das últimas duas categorias.

2. Pesquisas mostram que os humanos têm dificuldade em separar notícias verdadeiras de falsas. Enumere e explique três razões principais que justificam essa dificuldade humana.

3. Na abordagem de aprendizado de máquina com atributos linguísticos para detecção de notícias falsas, os pesquisadores buscam por "sinais de engano preditivo" no conteúdo de uma mensagem. Explique o conceito de "emotividade" e "pausalidade" como atributos linguísticos usando um exemplo de notícia. Depois explique como um sistema baseado predominantemente em atributos linguísticos poderia erroneamente classificar notícias verdadeiras como falsas.

4. O capítulo apresenta métodos de detecção de notícias falsas que se concentram na análise do conteúdo da notícia (atributos linguísticos, checagem de fatos) e outros que analisam a credibilidade da fonte (reputação do site, padrões de rede, características

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

do usuário). Discuta a conexão entre essas duas abordagens. Em quais cenários a análise da fonte seria mais eficaz que a análise do conteúdo, e vice-versa? Como a combinação de ambas poderia resultar em um sistema de detecção mais robusto e confiável?

5. O capítulo descreve diferentes métodos de detecção automática de notícias falsas. Compare brevemente a abordagem de "checagem de conteúdo" com a abordagem de "recuperação de informação". Qual delas é mais facilmente explicável em seus resultados e por quê?

6. Um desafio importante na detecção automática é a questão temporal. Explique por que a temporalidade é um desafio para os sistemas de detecção e como a técnica de RAG, em conjunto com LLMs, pode oferecer uma solução para mitigar esse problema.

7. Com suas próprias palavras, crie uma notícia falsa sobre um assunto de seu interesse e teste alguns LLMs para checar se eles são capazes de identificar as notícias como falsas. Faça ajustes na escrita da notícia falsa e também no prompt utilizado, de forma a verificar a consistência dos modelos. Que lições você consegue tirar disso? Há diferenças entre os modelos? É possível enganar os LLMs, levando-os a validar notícias falsas? É possível ter plena confiança na avaliação dos LLMs?

8. Sistemas de detecção automática podem cometer dois tipos de erro: classificar uma notícia verdadeira como falsa (falso negativo) ou uma notícia falsa como verdadeira (falso positivo). Discuta sobre as implicações sociais e éticas de cada tipo de erro, considerando o impacto na liberdade de expressão, na confiança do público em fontes de informação e na segurança. Use exemplos de situações em que cada tipo de erro pode causar impacto negativo.

9. O capítulo foca no texto, mas ressalta que o problema da desinformação é multimodal, envolvendo imagens, vídeos e áudio. Se você fosse projetar uma plataforma de detecção automática de notícias falsas abrangente, que integrasse a análise de texto, imagem e áudio, quais seriam os principais desafios técnicos e de dados para essa integração, e quais disciplinas da Inteligência Artificial, além do PLN, seriam indispensáveis?

10. Algumas vertentes de conteúdo enganoso são muito desafiadoras para detecção automática. As pós-verdades são um desses casos. Explique o que são pós-verdades e como um método computacional poderia detectá-las.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Capítulo 9

PLN na Saúde

Adriana Pagano
Claudia Moro
Elisa Terumi Rubel Schneider
Lilian Mie Mukai Cintho
Yohan Gumiel

Publicado em: 26/09/2023
Atualizado em: 16/04/2026

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

9.1 Introdução

A área da saúde é uma das mais importantes em nossas vidas e, nos últimos anos, tem se beneficiado do uso da tecnologia para melhorar o diagnóstico, o tratamento e a gestão de pacientes. A aplicação de Processamento de Linguagem Natural (PLN) tem sido fundamental para avançar nessa área, pois permite a análise de grandes volumes de dados não estruturados gerados em ambientes clínicos (Turchioe et al., 2022).

O domínio da medicina abrange diversos tipos de texto, utilizados para distintas atividades produtoras de significado, que desenvolvemos em nosso convívio social. Chamamos essas atividades de socio-semióticas. Estudos da linguagem baseados em pesquisas antropológicas modelam essas atividades socio-semióticas em oito tipos (Matthiessen et al., 2008; Matthiessen, 2013).

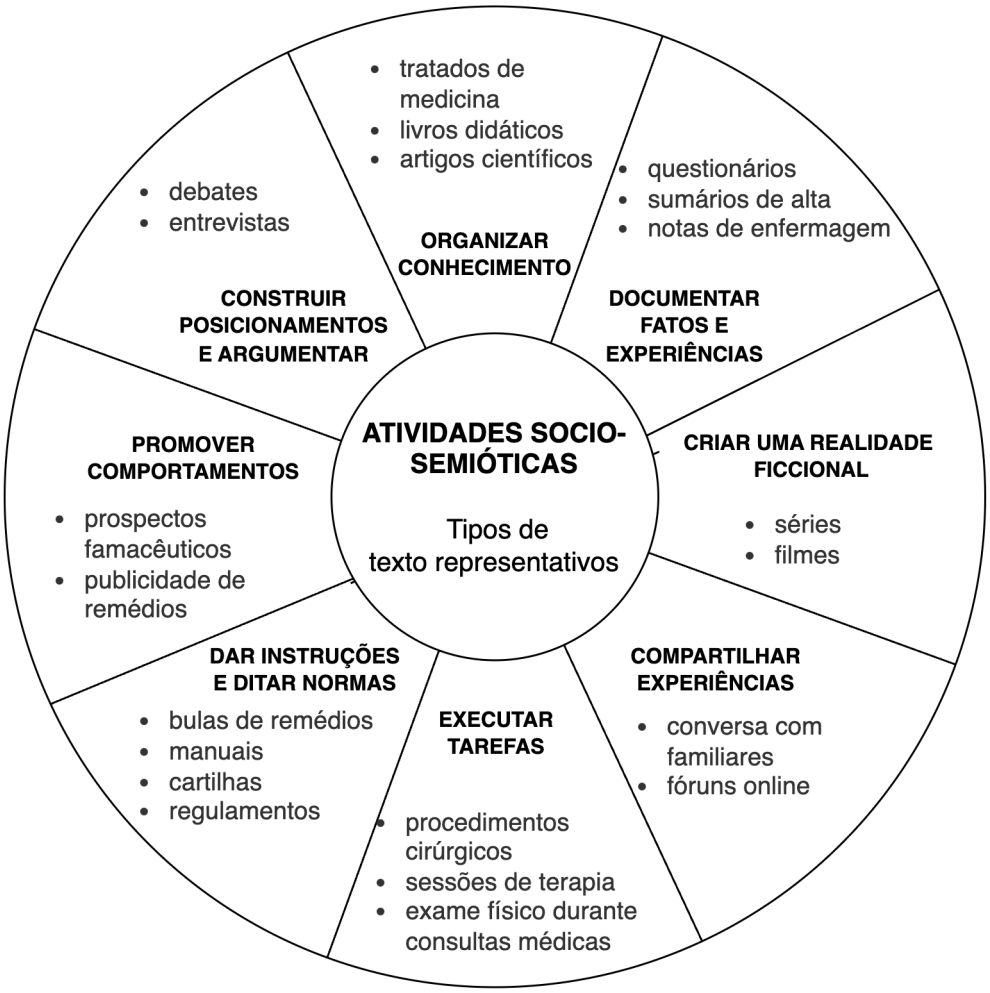
A Figura 9.1 mostra os oito tipos de atividades socio-semióticas e os tipos de texto mais representativos de cada um deles no domínio da medicina. Essas atividades são desenvolvidas por meio de textos escritos e falados, com funções específicas na nossa sociedade. Atividades nas quais a linguagem verbal tem um papel ancilar ou complementar são, por exemplo, a execução de procedimentos cirúrgicos, durante a qual ações podem ser verbalizadas ou não.

Mas, na grande parte das atividades humanas, a linguagem tem um papel constitutivo. Temos desde atividades que envolvem um uso especializado da linguagem para organizar a produção de conhecimento em tratados de medicina, livros didáticos e artigos acadêmicos, até atividades que envolvem um uso menos especializado, como o compartilhamento de experiências no âmbito privado, nas interações entre pacientes e familiares ou entre participantes de fóruns online sobre cuidados em saúde. Para a atividade de instruir e regular o comportamento, temos textos como bulas de medicamentos, cartilhas, normativas, manuais de instrução de equipamentos. Mesmo no domínio da medicina, há também textos pelos quais é construída uma realidade ficcional, como é o caso de séries e filmes que recriam interações em contextos médicos.

Uma atividade socio-semiótica muito relevante no domínio da medicina é documentar fatos e experiências, por meio de questionários aplicados ao paciente, registros de exames

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Figura 9.1: Tipos de texto no domínio da medicina



clínicos e relatos de profissionais da saúde, nos quais são documentadas percepções sobre a saúde do paciente. Esses textos são conhecidos em PLN como **narrativas clínicas** e abrangem notas de evolução de enfermagem, sumários de alta, boletins médicos, e notas em texto livre em campos próprios do prontuário eletrônico do paciente. Cada um desses tipos de texto pode oferecer informações valiosas a serem obtidas por meio do PLN mais adequado às características do texto. Artigos acadêmicos, por exemplo, podem ser usados para a extração de ontologias, que são estruturas semânticas que permitem uma representação formal de conceitos, suas propriedades e relações. Essas ontologias podem ser usadas para facilitar a compreensão de termos técnicos e complexos em diferentes áreas da saúde, permitindo que as informações sejam compartilhadas de forma mais clara e precisa (Jiang et al., 2020). Também podemos identificar padrões e relacionamentos entre os dados e a construção de modelos preditivos (Lee et al., 2019).

Narrativas clínicas, por outro lado, são textos não estruturados que oferecem informações valiosas sobre a história do paciente, incluindo seus sintomas, histórico médico, estilo de vida e outras informações relevantes. A mineração desses dados pode ser usada para identificar padrões e relacionamentos entre os dados, permitindo uma melhor compreensão da condição do paciente e a construção de modelos preditivos para prever possíveis

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

complicações ou doenças (Wu et al., 2018).

Assim como nas narrativas clínicas, áudios produzidos por pacientes podem dar pistas de condições de saúde que afetam voz e fala, especialmente as que afetam o sistema respiratório. No Capítulo 11, o leitor encontra modelos e técnicas de classificação de áudio para a voz e fala de pacientes com problemas respiratórios.

9.2 O texto livre em narrativas clínicas

Com o advento do Registro Eletrônico de Saúde (RES)[1], como é denominado no Brasil, ou em inglês, o *Electronic Health Record* (EHR), a quantidade de dados gerados relativos à atenção aos pacientes aumentou significativamente. Os prontuários eletrônicos podem conter dados estruturados, semiestruturados ou não estruturados, todos eles oferecendo uma grande quantidade de informações sobre o paciente. A mineração desses dados pode ajudar a identificar tendências e padrões em relação a diagnósticos, tratamentos e resultados, permitindo uma melhor gestão do cuidado do paciente e um melhor planejamento da assistência (Shickel et al., 2017).

Os dados clínicos presentes nas narrativas clínicas em texto livre (dados não estruturados) apresentam características únicas que dificultam sua análise e interpretação. Esses dados são frequentemente apresentados em linguagem médica especializada, repleta de termos técnicos, jargões e abreviaturas que podem variar entre os distintos profissionais de saúde. Esses textos também podem conter erros de digitação, ortografia ou gramática, tornando a interpretação ainda mais complexa (Dalianis, 2018). A Figura 9.2 apresenta um exemplo de narrativa clínica adaptada para fins de ilustração. Nela podemos observar que as informações podem ser estruturadas de acordo com categorias destacadas com cores e rotuladas na legenda da figura.

No escopo do que chamamos narrativas clínicas, há diferentes tipos de texto, os quais apresentam desafios específicos em termos do tipo de linguagem e também da relevância das informações registradas. Por exemplo, as notas de evolução de enfermagem podem ser mais descritivas e detalhadas do que outros tipos de texto, enquanto os sumários de alta podem fornecer informações importantes sobre a condição atual do paciente e seu histórico de tratamento. Já as notas de ambulatório podem ser mais informais e fragmentadas, o que dificulta sua análise por modelos treinados com outros tipos de texto em outros domínios. Isso demanda a anotação manual de narrativas clínicas de forma contarmos com modelos mais refinados.

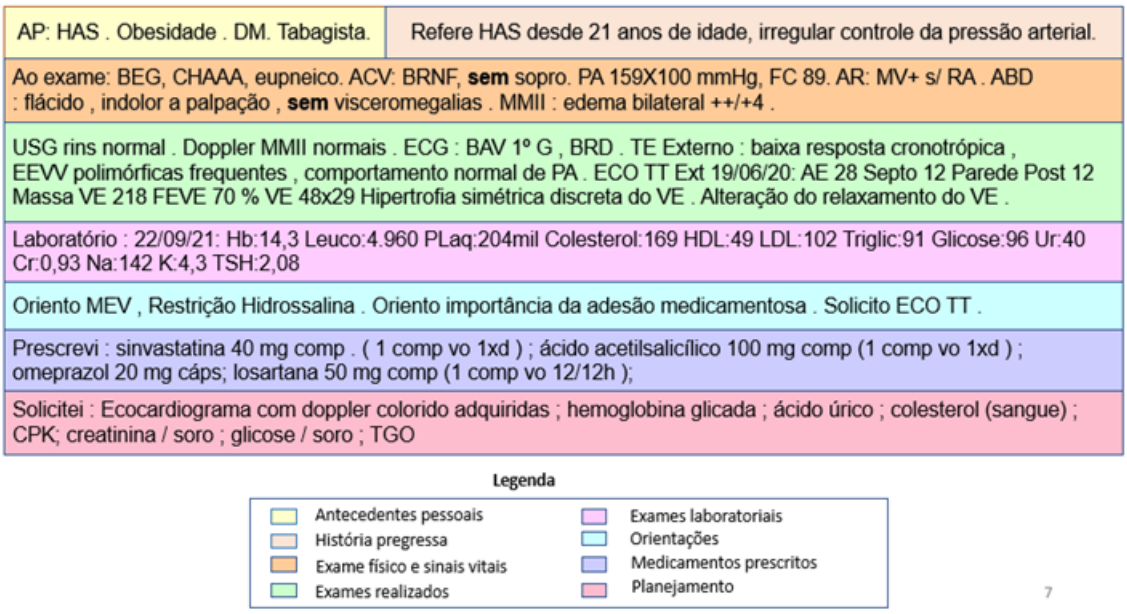
Como todo processo manual, a anotação de narrativas clínicas requer tempo e recursos, o que dificulta a construção de grandes *datasets* para treinamento de modelos de PLN. Como resultado, a aplicação de técnicas de aprendizado de máquina em dados clínicos sofre limitações pela disponibilidade de dados anotados manualmente (Koleck et al., 2019). Uma saída é utilizar modelos genéricos para pré-processamento, sendo a saída avaliada manualmente. Um exemplo deste tipo de trabalho é a anotação do *corpus* Depclin-Br, que vem sendo desenvolvida por uma equipe de cientistas da computação da PUCPR e de linguistas da Faculdade de Letras da UFMG. Trata-se de um conjunto de narrativas clínicas já anotadas em termos de entidades no domínio clínico e constituindo o *corpus* SemClinBr (Oliveira et al., 2022a). Uma parte desse *corpus* foi anotada morfossintaticamente com

[1] No Sistema Único de Saúde (SUS), as informações dos usuários são coletadas e armazenadas por meio do Prontuário Eletrônico do Cidadão (PEC). Nele, há campos pré-determinados que podem ser preenchidos com texto livre.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

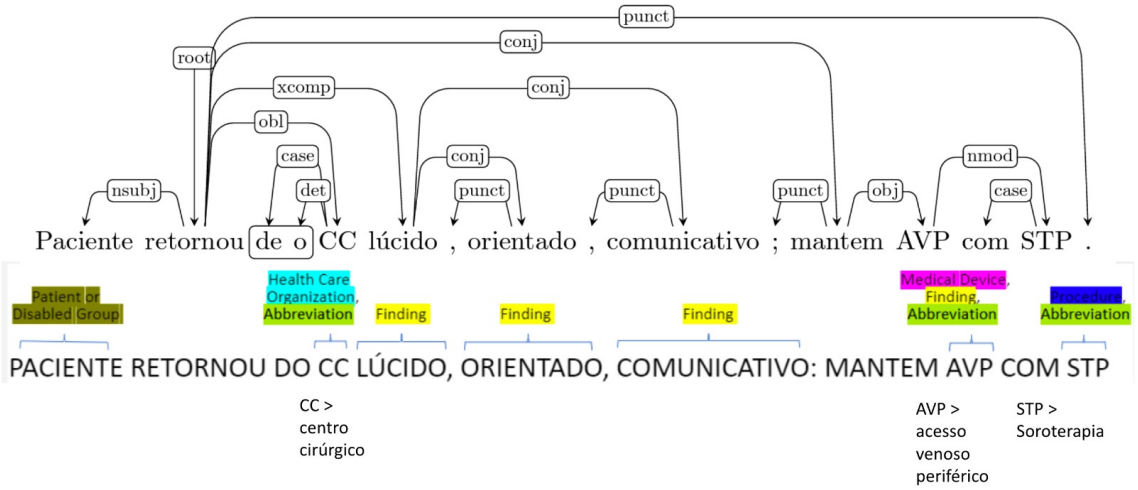
©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Figura 9.2: Exemplo de narrativa clínica elaborada para fins de ilustração. Na legenda, as categorias de informações que podem ser encontradas neste tipo de texto.



base num modelo genérico de português e a anotação revisada manualmente (Oliveira et al., 2022b). Essa primeira parte foi utilizada para refinamento do modelo genérico e anotação automática de um segunda parte do *corpus*. Uma vez concluída a anotação, dados do *corpus* DepClinBr, anotado com relações de dependência, podem ser minerados e utilizados para caracterizar as entidades nomeadas previamente anotadas no SemClinBr. A Figura 9.3 ilustra a correlação de anotações morfossintáticas e entidades.

Figura 9.3: Correlação de anotações morfossintáticas e entidades.



A construção de *corpora* de narrativas clínicas (dados não estruturados) está sujeita a restrições técnicas e regulatórias, que dizem respeito à privacidade de dados. Essa especificidade limita a capacidade de construção de grandes *datasets* para treinamento de modelos de PLN (Chen; Chen, 2022). Como foi apontado, para contornar essa limitação,

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

são utilizados modelos genéricos da língua, os quais precisam ser refinados com dados específicos do domínio em um processo de *fine-tuning*, para melhorar ainda mais sua precisão e relevância (Lee et al., 2019).

A seguir, veremos alguns exemplos de aplicações da PLN em dados clínicos.

9.3 Aplicações de PLN na Saúde

9.3.1 Predição

Uma das principais tarefas de PLN na área médica é a predição, que pode ser aplicada em diversas demandas do cuidado em saúde, como diagnóstico, tratamento, evolução, alta médica hospitalar, detecção de quedas, detecção de depressão e outras. Essas demandas envolvem a classificação de dados clínicos, como narrativas de pacientes, prontuários eletrônicos, relatórios médicos e outros dados de saúde, para ajudar os médicos e outros profissionais de saúde a tomar decisões mais precisas. A predição de diagnóstico, por exemplo, pode ajudar a identificar doenças em estágios iniciais, permitindo tratamentos mais eficazes e prevenindo complicações. A predição de tratamento pode ajudar a personalizar o tratamento para cada paciente, maximizando sua eficácia e minimizando efeitos colaterais. A detecção de quedas e depressão pode ajudar a prevenir acidentes e melhorar a qualidade de vida dos pacientes. Em resumo, a tarefa de predição é essencial para a aplicação bem-sucedida de PLN na área de saúde (Yan et al., 2022).

Alguns exemplos de trabalhos envolvendo predição e classificação em textos clínicos em português são (Gonçalves et al., 2023; Santos et al., 2021a; Silva et al., 2023a; Yang et al., 2022).

9.3.2 Desidentificação

Um aspecto crucial na aplicação de PLN na área médica é a desidentificação dos dados dos pacientes, associada a processos de anonimização ou pseudonimização. Esta envolve a remoção de informações que possam identificar o paciente, como nome, endereço, número de telefone e outras informações pessoais. A anonimização é necessária para garantir a privacidade dos pacientes e cumprir as regulamentações de proteção de dados, como a Lei Geral de Proteção de Dados (LGPD) no Brasil[2] e a General Data Protection Regulation (GDPR) na União Europeia[3].

A anonimização de dados clínicos é um processo desafiador, uma vez que esses dados contêm informações altamente sensíveis e complexas, como histórico médico, sintomas, exames, tratamentos e outros detalhes que podem identificar um paciente. Portanto, é necessário utilizar técnicas avançadas de PLN, como o uso de modelos de linguagem, para remover as informações sensíveis e garantir a privacidade dos pacientes (Jones et al., 2020).

Existem diversas técnicas que podem ser utilizadas na desidentificação dos dados clínicos, dependendo do tipo de informação que deve ser removida e do nível de anonimização desejado, por exemplo:

- Substituição de nomes próprios e outros identificadores pessoais por símbolos ou pseudônimos aleatórios;

[2] Lei Geral de Proteção de Dados Pessoais (LGPD), Lei nº 13.709/2018. Disponível em: https://www.gov.br › pt-br › acesso-a-informacao › lgpd

[3] *Data protection in the EU*. Disponível em: https://commission.europa.eu/law/law-topic/data-protection/data-protection-eu_en

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

- Remoção de informações geográficas específicas, como endereço e CEP;
- Substituição de datas de nascimento e outras informações temporais por intervalos ou idades aproximadas;
- Remoção de informações de contato, como números de telefone e endereços de e-mail;
- Remoção de informações de identificação de instituições, como o nome de hospitais e clínicas.

Além dessas técnicas, também é possível utilizar métodos mais avançados de PLN, como a detecção e remoção de termos médicos específicos ou o uso de técnicas de de-identificação baseadas em modelos de linguagem, que tentam preservar a integridade semântica dos dados, mesmo após a remoção ou substituição das informações pessoais.

A desidentificação dos pacientes permite que os dados clínicos sejam utilizados para fins de pesquisa e análise, sem comprometer a privacidade dos pacientes. Isso é fundamental no avanço da medicina, permitindo a análise de grandes volumes de dados na descoberta de padrões e tendências em doenças, tratamentos e outros aspectos da saúde (Liu et al., 2017). Em (Santos et al., 2021b) temos um exemplo de trabalho para o português nessa tarefa.

9.3.3 Extração de conceitos clínicos

A busca e extração de conceitos clínicos relevantes é uma tarefa essencial na aplicação de PLN na área médica. Essa tarefa envolve a identificação de entidades relevantes nos dados clínicos, como sintomas, diagnósticos, tratamentos, medicamentos e outros termos específicos da área da saúde. Essa identificação geralmente é feita por meio de técnicas de NER (do inglês, *Named Entity Recognition*) (Capítulo 3), que permitem a identificação e classificação automática de entidades em textos não estruturados. A Figura 9.4 ilustra um exemplo de entidades do tipo *Problema* reconhecidas em uma narrativa clínica elaborada para fins de ilustração.

Figura 9.4: Exemplo de entidades do tipo Problema (em azul) encontradas em narrativa clínica.

Extração de conceitos clínicos: problemas

Paciente DM tipo 2, HAS e obesidade, com queixa de descontrole glicêmico necessitando procurar o PS 3x em 2 semanas. Refere que não consegue perder peso mesmo realizando dieta alimentar e caminhadas e faz uso regular dos medicamentos prescritos. Relata que apresenta disúria e hematúria há 2 dias. Apresenta PA 120X80 mmHg, sem edemas de MMII. Faz uso de atenolol 25 mg 2X/dia, metformina 850 mg 2X/dia. Solicitado exames laboratoriais (hemoglobina glicada, colesterol total, LDL e triglicérides e exame de Urina I). Aumento metformina 850 mg para 3X/dia e prescrevo Ciprofloxacino 250 mg cada 12 horas, durante 7 dias. Oriento a manter os medicamentos em uso e à modificar o estilo de vida e retornar para acompanhamento ambulatorial.

Além da identificação de entidades, outras técnicas de PLN também podem ser utilizadas para a busca e extração de conceitos clínicos relevantes, como a detecção de negação e a resolução de ambiguidades. A detecção de negação, por exemplo, é útil para identificar quando um sintoma é negado pelo paciente ou um diagnóstico dado pelo médico nega alguma condição. A precisão na detecção de negação é fundamental para a interpretação dos dados clínicos (Nath et al., 2022).

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Outra técnica importante na busca e extração de conceitos clínicos é o mapeamento de terminologia, que consiste na associação dos termos clínicos encontrados nos textos com um conjunto de termos padronizados, como a Classificação Internacional de Doenças (CID) ou o *Systemized Nomenclature of Medicine* (SNOMED CT). Isso permite uma melhor organização e interpretação dos dados clínicos, facilitando a análise e a tomada de decisão médica (Fennelly et al., 2021).

A busca e extração de conceitos clínicos relevantes é fundamental para a análise de dados clínicos em larga escala, permitindo a identificação de padrões e tendências em doenças, tratamentos e outros aspectos da saúde. Além disso, essas técnicas de PLN também podem ser utilizadas para a construção de sistemas de suporte à decisão médica, que auxiliam os profissionais de saúde na escolha de tratamentos mais adequados para cada paciente (Demner-Fushman et al., 2009).

9.3.4 Relações temporais

Uma linha do tempo do paciente é uma representação gráfica que organiza as informações clínicas de um paciente de maneira cronológica. O interesse pela pesquisa em extração de relações temporais provém da característica longitudinal dos dados presentes nos RES. Esses registros contêm múltiplos textos clínicos referentes ao mesmo paciente, escritos em diferentes momentos (Gumiel et al., 2021).

A extração de relações temporais concentra-se na organização sequencial de menções em um texto, sendo essas menções eventos médicos ou expressões temporais.

No contexto clínico, eventos médicos são circunstâncias clínicas de relevância, cujo escopo é delimitado pelo contexto da aplicação. Por exemplo, para a extração de informações significativas para o diagnóstico, pode ser apropriado delimitar eventos como menções a tratamentos passados, sinais, sintomas, medicamentos em uso e exames realizados pelo paciente com os respectivos resultados. Já as expressões temporais envolvem menções de tempo, como a duração de um sintoma ou indicações de quando o paciente realizou determinada cirurgia. É notável que as expressões temporais só têm significado quando associadas a algum evento, enquanto os eventos podem fazer sentido quando relacionados entre si.

A fim de extrair essas menções do texto, são empregadas técnicas de Processamento de Linguagem Natural (PLN), como a Reconhecimento de Entidades Nomeadas. A tarefa de NER consiste em identificar e classificar automaticamente eventos e expressões temporais.

Com eventos e expressões temporais devidamente identificados, aplica-se a extração de relações temporais, uma técnica de PLN que se concentra na conexão de eventos entre si ou com expressões temporais. Desse modo, cada entidade acaba sendo relacionada a um período de tempo específico.

Ao considerar relações temporais no contexto clínico, diversas áreas de pesquisa emergem. Doenças crônicas, por exemplo, apresentam uma natureza longitudinal que torna a temporalidade extremamente relevante, pois existem fluxos de dados do paciente contínuos e extensos, nos quais podem ser extraídos padrões significativos (Sheikhalishahi et al., 2019). A progressão de uma doença e os eventos a ela associados são registrados cronologicamente, onde certos eventos são relevantes apenas em momentos específicos, como problemas médicos identificados durante um exame físico em uma consulta ambulatorial ou sintomas relatados (Sheikhalishahi et al., 2019). No caso de tratamento ineficaz de hipertensão com monoterapia, por exemplo, busca-se terapias com medicamentos combinados. Portanto, algumas informações sobre a progressão de doenças podem ser mais

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

facilmente discernidas por meio da extração de relações temporais (Gumiel et al., 2021).

A aplicação prática de uma linha do tempo na área da saúde pode ser utilizada para analisar a evolução do quadro clínico do paciente ao longo do tempo, identificar possíveis tendências e realizar previsões. Além disso, a linha do tempo do paciente pode ser integrada a sistemas de suporte à decisão médica, contribuindo para a seleção de tratamentos mais adequados para cada paciente.

9.3.5 Sumarização

A sumarização de evoluções clínicas é uma tarefa de PLN que tem como objetivo extrair as informações mais relevantes de um conjunto de dados clínicos, de forma a produzir uma versão resumida e legível dessas informações. A Figura 9.5 exibe um exemplo fictício de uma narrativa clínica sumarizada.

Figura 9.5: Exemplo fictício de uma narrativa clínica sumarizada, na qual as informações mais importantes foram mantidas.

Exemplo de Sumarização

Paciente hipertenso, dislipidêmico e sobrepeso, internou por 2 dias na clínica médica devido à crise hipertensiva e queixa de formigamento no MSD. Solicitado exames laboratoriais, ECG e cateterismo cardíaco. Sem apresentação de alterações enzimáticas e do ECG. Cateterismo sem lesões obstrutivas graves na CD, DP e VP, TCE sem lesões, DA e Dg1 sem lesões, CX, Mg1 e VPE sem lesões. Medicado e manteve a pressão arterial estável e sem queixas de dor. Mantido os medicamentos em uso: Atenolol 100 mg 1x/dia, Anlodipino 5mg 2xd e Atorvastatina 20mg a noite. Recebe alta hospitalar com a pressão arterial controlada e sem sintomas para dor. Oriento para a adesão medicamentosa, mudança de estilo de vida e acompanhamento ambulatorial.

Para realizar a sumarização de evoluções clínicas, são utilizadas técnicas de sumarização automática de texto (Capítulo 5), que podem ser baseadas em abordagens extrativas ou abstrativas[4]. Na abordagem extrativa, as frases mais importantes do texto original são selecionadas e combinadas para formar um resumo. Já na abordagem abstrativa, o resumo é gerado a partir da síntese das informações do texto original, gerando uma nova versão que não necessariamente contém as mesmas palavras e frases do texto original.

Para realizar a sumarização de evoluções clínicas, são utilizadas técnicas de processamento de linguagem natural, incluindo NER para identificar as entidades relevantes, PoS (*Part-of-Speech*) para identificar as partes do discurso e gramática do texto e também técnicas de análise sintática e semântica.

Essa tarefa de PLN é muito útil para os profissionais da área da saúde, pois permite que eles analisem brevemente as informações mais importantes dos pacientes, como histórico de doenças, exames realizados, tratamentos prescritos, entre outras informações clínicas (Gulden et al., 2019).

[4]Para projetos de sumarizadores em português, visite: https://sites.icmc.usp.br/taspardo/sucinto/

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

9.4 Recursos de PLN na área clínica

Como vimos, a aplicação de PLN na área clínica tem revolucionado a maneira como os dados médicos são processados e analisados. Com a crescente complexidade das informações clínicas e o volume crescente de dados gerados, a utilização de técnicas avançadas de PLN se torna essencial para extrair *insights* valiosos e melhorar a qualidade dos cuidados de saúde.

Nesta seção, vamos explorar alguns recursos de PLN em português utilizados no contexto clínico, com ênfase nos *corpora* e modelos.

9.4.1 *Corpora*

9.4.1.1 Semclinbr-corpus

SemClinBr (Oliveira et al., 2022a) é um *corpus* semanticamente anotado que contém um conjunto de dados clínicos em português brasileiro, com anotações detalhadas.

O *corpus* é composto por 1.000 notas clínicas provenientes de diversas especialidades médicas, tipos de documentos e instituições. Estas notas foram cuidadosamente anotadas para incluir informações valiosas sobre entidades e relações dentro dos textos clínicos.

SemClinBr inclui 65.117 entidades e 11.263 relações anotadas. As anotações cobrem uma variedade de aspectos clínicos, como condições médicas, medicamentos e procedimentos, além de incluir dicionários de negação e abreviações médicas.

A qualidade das anotações foi avaliada de forma intrínseca e extrínseca. A avaliação mostrou um alto nível de concordância entre anotadores, com índices variando de 0,71 a 0,92, e os resultados obtidos ao aplicar o corpus em tarefas de PLN demonstraram a utilidade e a confiabilidade das anotações.

SemClinBr fornece um recurso valioso para a pesquisa em PLN clínico, ajudando a desenvolver e avaliar novos algoritmos e técnicas.

Sendo um dos primeiros *corpus* clínico disponível em português brasileiro, o SemClinBr preenche uma lacuna significativa no campo e promove avanços na pesquisa biomédica e no desenvolvimento de tecnologias de PLN para o idioma.

O *corpus* está disponível para pesquisa acadêmica e pode ser solicitado neste endereço: https://github.com/HAILab-PUCPR/SemClinBr.

9.4.1.2 Detecção de Quedas

O *corpus* Detecção de Quedas (Santos et al., 2019a) é projetado para auxiliar na detecção de incidentes de quedas, que são uma das maiores categorias de relatórios de eventos adversos em hospitais.

A detecção eficiente desses eventos é crucial para melhorar a compreensão dos incidentes e, consequentemente, a qualidade do cuidado ao paciente.

O *corpus* contém 1.078 notas de progresso desidentificadas, que são registros clínicos detalhados sobre os pacientes durante sua estadia no hospital.

As notas foram anotadas para identificar incidentes de quedas, permitindo a avaliação de modelos de linguagem na tarefa de detectar tais eventos.

A avaliação do *corpus* foi realizada com vários modelos de linguagem, incluindo redes neurais recorrentes (RNNs) de última geração. Os experimentos mostraram que a abordagem de aprendizado profundo superou os trabalhos anteriores na detecção de eventos de queda.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

O *corpus* anotado está disponível para fins de replicação em https://github.com/nlp-pucrs/fall-detection, permitindo que outros pesquisadores reproduzam os experimentos e utilizem os dados para melhorar os sistemas de detecção de eventos adversos.

9.4.1.3 Casos Clínicos - Neurologia

O *corpus* de Casos Clínicos - Neurologia (Lopes et al., 2019) é formado por 281 textos de casos clínicos coletados dos números 1 e 2 do volume 17 da revista clínica Sinapse, publicada pela Sociedade Portuguesa de Neurologia.

Os textos foram pré-processados com ferramentas do NLPPort, um kit de ferramentas de PLN para o português, baseado no OpenNLP.

Cada texto foi tokenizado com TokPort, etiquetado com TagPort e lemas para cada par token-PoS foram obtidos com LemPort. Após o pré-processamento, foi realizada a anotação manual de entidades nomeadas clínicas.

O *corpus* está disponível para pesquisa acadêmica e pode ser solicitado neste endereço: https://github.com/fabioacl/PortugueseClinicalNER.

9.4.2 Modelos

9.4.2.1 BioBERTpt

Os modelos da família BioBERTpt (Schneider et al., 2020) são modelos adaptados do modelo BERT multilíngue para o português brasileiro, com foco em tarefas de PLN nas áreas clínica e biomédica.

Essas versões foram ajustadas utilizando um *corpus* de narrativas clínicas e artigos científicos biomédicos em português.

Os modelos foram avaliados em duas coleções anotadas de textos clínicos para reconhecer entidades nomeadas. Comparado com modelos BERT existentes, o BioBERTpt apresentou uma melhoria de 2,72% no F1-score, superando o modelo base em 11 de 13 entidades avaliadas.

O desenvolvimento de BioBERTpt é um passo importante na área, ao melhorar o desempenho de tarefas de PLN em idiomas com menos recursos e pesquisa, como o português brasileiro, que frequentemente é sub-representado em modelos de PLN avançados.

Os modelos estão disponíveis em: https://huggingface.co/pucpr.

9.4.2.2 BioBERTptRT

O BioBERTptRT (De Oliveira et al., 2024) é uma versão do BioBERTpt projetado para a classificação de sentenças de notas clínicas, categorizando sentenças de acordo com o padrão SOAP (Subjetivo, Objetivo, Avaliação e Plano), o que ajuda a estruturar e organizar as notas clínicas de maneira mais eficaz.

O modelo foi treinado e ajustado usando uma base de dados privada de 10.000 registros de saúde anonimizados contendo 234.673 notas clínicas, divididas em 1.183.345 sentenças únicas. Além disso, 100.021 sentenças foram rotuladas manualmente para uso no ajuste fino dos modelos.

Esse treinamento especializado ajudou o modelo a entender melhor o contexto e a terminologia médica em português.

O modelo BioBERTptRT demonstrou um desempenho superior em comparação com outros modelos BERT em português, alcançando resultados superiores em precisão, acurácia, revocação e F1-score na tarefa de classificação de sentenças.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

9.4.2.3 GPT2-Bio-Pt

GPT2-Bio-Pt (Schneider et al., 2021) é um modelo baseado em GPT-2 (*Generative Pre-trained Transformer 2*) especializado para o idioma português, desenvolvido para apoiar tarefas de PLN na área clínica e biomédica.

Um GPT-2 genérico em português foi ajustado para o domínio biomédico com a técnica de transferência de aprendizado, utilizando textos biomédicos escritos em português.

O modelo GPT-2 ajustado para o domínio biomédico foi testado em um conjunto de dados público, manualmente anotado para a tarefa de detecção de quedas de pacientes. O modelo especializado demonstrou um desempenho superior ao modelo GPT-2 genérico em português, com um aumento de 3,43 pontos em *F1-score*.

O modelo está disponível para pesquisa acadêmica e pode ser acessado neste endereço: https://huggingface.co/pucpr/gpt2-bio-pt.

9.4.2.4 BERT para Classificação ICD-10

(Coutinho; Martins, 2022) desenvolveram um modelo BERT especializado para a atribuição de códigos ICD-10 (Classificação Internacional de Doenças) para causas de morte, usando descrições em texto livre.

O modelo foi projetado para atribuir códigos ICD-10 a causas de morte com base em descrições de texto livre encontradas em certificados de óbito, relatórios de autópsia e boletins clínicos fornecidos pelo Ministério da Saúde de Portugal.

A abordagem inclui um procedimento de pré-treinamento que incorpora conhecimento específico do domínio clínico, permitindo a adaptação do BERT para a tarefa de codificação ICD-10.

O modelo também adotou uma estratégia de ajuste fino para lidar com o problema de desequilíbrio de classes, que é comum em tarefas de classificação quando algumas categorias são muito mais frequentes do que outras.

9.4.2.5 CardioBERTpt

CardioBERTpt (Schneider et al., 2023) é um modelo baseado em BERT adaptado especificamente para tarefas de PLN na área da cardiologia em português.

O modelo CardioBERTpt aproveita a transferência de aprendizado a partir de registros eletrônicos de saúde de um hospital terciário brasileiro especializado em doenças cardiológicas. Esta abordagem permite que o modelo capture nuances e terminologias específicas do domínio cardiológico.

Esse modelo demonstrou que a representatividade dos dados e um volume alto de dados de treinamento podem melhorar significativamente os resultados para tarefas clínicas.

O modelo está disponível para pesquisa acadêmica e pode ser acessado neste endereço: https://huggingface.co/pucpr-br/cardiobertpt.

9.5 Para onde estamos caminhando?

Embora a tecnologia de PLN na área clínica tenha avançado significativamente nos últimos anos, ainda existem vários desafios a serem superados. Alguns desses desafios incluem:

- Garantir a qualidade dos dados clínicos utilizados para treinar e testar os modelos de PLN, incluindo a devida anonimização e a padronização dos termos utilizados,

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

assegurando a ética e a privacidade dos dados clínicos;
- Desenvolver modelos de PLN capazes de lidar com textos clínicos mais complexos e heterogêneos, como notas de enfermagem, laudos médicos e textos escritos por pacientes;
- Integrar os modelos de PLN em sistemas de informação em saúde existentes, garantindo a interoperabilidade e a segurança dos dados;
- Garantir a aceitação e a adoção dos modelos de PLN pelos profissionais de saúde, demonstrando sua utilidade e eficácia na prática clínica.

É importante destacar que, embora o PLN possa ser útil na análise e interpretação de dados clínicos, ele não pode substituir a experiência e o conhecimento clínico de um médico ou de outros profissionais de saúde. A tecnologia pode ser uma ferramenta valiosa para auxiliar na tomada de decisões clínicas, mas não pode substituir o julgamento clínico humano. Ressalta-se que o desenvolvimento de tecnologias de PLN na área clínica seja visto como uma forma de complementar e melhorar o cuidado ao paciente, e não como uma substituição aos profissionais de saúde.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Capítulo 10

Detecção de Transtornos de Saúde Mental a partir de Texto

Ivandré Paraboni
Helena de Medeiros Caseli

Publicado em: 20/11/2024

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

10.1 Introdução

Transtornos de saúde mental como depressão e ansiedade são problemas bem conhecidos e uma crescente fonte de preocupação na sociedade moderna. Segundo dados do Global Burden of Disease Study de 2019[1], a depressão é o segundo fator que mais impacta na queda da expectativa de vida da população global. Aliado à gravidade do quadro, a Organização Mundial da Saúde (OMS) alerta que apenas um quarto dos indivíduos que sofrem algum transtorno de saúde mental recebem o atendimento adequado (WHO, 2021). No caso específico do Brasil, de acordo com Lopes et al. (2022), a prevalência da depressão no país cresceu 36,7% de 2013 a 2019, indo de 7,9% em 2013 para 10,8% em 2019. Durante a pandemia de Covid-19, outra pesquisa (Zhang et al., 2021) realizada em maio de 2020 com 483 adultos constatou que 70,3% tinha sintomas de depressão e 67,2% tinham sintomas de ansiedade.

Ao mesmo tempo, diversos estudos demonstram que indivíduos com transtornos de saúde mental são usuários regulares de redes sociais em proporção similar à população em geral, que é estimada em cerca de 70% entre indivíduos de meia idade e até 97% entre indivíduos mais jovens (Aschbrenner, 2020; Aschbrenner et al., 2018; Birnbaum et al., 2017; Brunette et al., 2019), e que indivíduos com estes transtornos frequentemente recorrem às redes sociais em busca do suporte de outros usuários com problemas semelhantes (Bucci et al., 2019).

Este cenário, e a observação de que transtornos de saúde mental frequentemente se refletem na linguagem empregada pelos indivíduos que sofrem dessas condições, levaram a um número expressivo de estudos no PLN e áreas relacionadas, tendo como foco a detecção de transtornos como depressão, ansiedade, bipolaridade, anorexia, ideação suicida e automutilação em redes sociais (Cohan et al., 2018; Lin et al., 2020a; Mann et al., 2020; Shen et al., 2017; Shen; Rudzicz, 2017; Trotzek et al., 2018; Yates et al., 2017; Yazdavar et al., 2017). Este capítulo visa apresentar um levantamento geral de estudos deste tipo, que procuram identificar casos de maior gravidade e eventualmente sinalizar a necessidade de um indivíduo buscar ajuda por meio da detecção computacional de transtornos de

[1] https://ghdx.healthdata.org/gbd-2019

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

depressão e ansiedade a partir de textos em português, em geral provenientes de redes sociais.

10.1.1 Definição do problema

A detecção computacional de indivíduos com transtornos de saúde mental a partir de textos provenientes de redes sociais e outras fontes é uma tarefa complexa e não totalmente resolvida na pesquisa em PLN, sendo tipicamente modelada como um problema de aprendizado de máquina (AM) supervisionado. Em abordagens deste tipo, um *corpus* de publicações (e.g., puramente textuais ou multimodais) rotulado com algum tipo de informação relativa à saúde mental (e.g., depressivo ou não depressivo) é utilizado para treinamento e teste de modelos de classificação do fenômeno de interesse.

Com base em um *corpus* textual rotulado com informações de saúde mental, há diversas formulações possíveis para o problema computacional a ser resolvido. Dentre estas, as mais comuns são a detecção de transtornos de saúde mental (e.g., decidir se um indivíduo é depressivo ou não, como em (Yates et al., 2017)), e a detecção de sintomas (e.g., decidir se um texto expressa um sintoma tradicional de depressão, como em (Yadav et al., 2020)). Outras formulações menos frequentes para o problema incluem o objetivo de determinar o grau de severidade de um determinado transtorno (Mann et al., 2020), ou diferenciar o discurso de indivíduos depressivos de outros meramente interessados em saúde mental (Santos et al., 2020c), ou identificar uma granularidade maior de sintomas e fatores protetivos e de risco para a depressão (Mendes; Caseli, 2024). Há, ainda, outra formulação popularizada pela série de desafios computacionais eRisk (Losada; Crestani, 2016) que é a tarefa de detecção precoce de risco de depressão e outros transtornos, em que o objetivo consiste em prever o estado de saúde mental de um indivíduo com base no menor volume possível de publicações em ordem cronológica, ou seja, o mais cedo possível.

Outro aspecto importante da formulação do problema é o objeto da classificação. Parte dos estudos da área é dedicada à classificação de estados de saúde mental de indivíduos (e.g., usuários de rede social). Este tipo de abordagem, aqui denominada de **nível autoral**, toma por base uma coleção de publicações de um indivíduo (e.g., uma sequência de postagens em rede social, ou *timeline*) que pode ou não conter informações relevantes para a tarefa. Abordagens orientadas ao indivíduo são empregadas quando o objetivo é estabelecer o estado de saúde mental (e.g., depressivo *versus* não depressivo) de uma pessoa. Por outro lado, abordagens aqui denominadas de **nível textual** modelam o problema de classificação de textos individuais (e.g., postagens de rede social) independentemente da sua autoria. Estas abordagens tendem a ser mais utilizadas quando o objetivo é filtrar ou selecionar textos de maior relevância para uma aplicação da área de saúde mental, como a detecção de sintomas de depressão em redes sociais. A Figura 10.1 ilustra estas possibilidades considerando o exemplo da detecção de depressão, e estas abordagens são discutidas em mais detalhes nas seções a seguir.

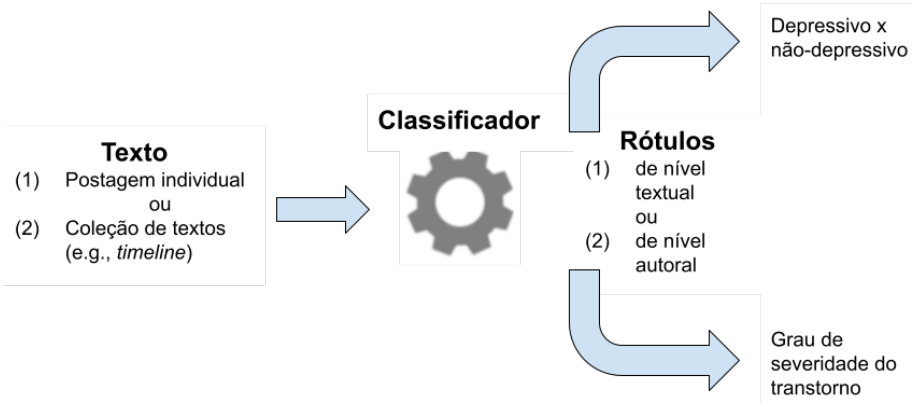
10.1.1.1 Classificação de nível autoral

A classificação de nível autoral recebe como entrada uma coleção de textos e outros dados produzidos por (ou relativos a) uma determinada pessoa (e.g., uma coleção de postagens em redes sociais, meta-dados etc.), e objetiva determinar o estado de saúde mental (e.g., depressivo ou não depressivo) desta pessoa. A base de dados de treino e teste nestes casos contempla indivíduos com um transtorno de saúde mental conhecido (ou pertencentes à classe positiva, na terminologia de AM), que podem ser selecionados por meio de validação

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Figura 10.1: Formulações do problema de classificação de depressão em nível textual e autoral.



externa (e.g., por meio de avaliação médica especializada ou uso de questionários, como em (Mann et al., 2020)), obtendo-se, por exemplo, um *corpus* de textos de indivíduo com e sem depressão. Este método de aquisição de dados, apesar de confiável, tende a ser de alto custo, o que na prática pode limitar o número de casos (ou instâncias) obtidos.

Como alternativa à validação externa, estudos da área de PLN têm explorado a ocorrência de **autorrelatos** em redes sociais, como em "Fui ao psicólogo hoje e ele me diagnosticou com esquizofrenia" para identificar rapidamente grandes volumes de indivíduos da classe positiva com custo mínimo. Autorrelatos deste tipo – que servem apenas para a seleção de usuários, mas que não são incluídos nos dados do *corpus* propriamente dito – são comuns em redes sociais, e embora sejam naturalmente mais sujeitos a imprecisões ou falsidade (o que de resto pode ser problemático também no caso da validação externa), é geralmente aceito que estas dificuldades tendem a ser compensadas pelo maior volume de dados obtido por este método (Coppersmith et al., 2015).

Dado entretanto que autorrelatos fornecem evidência apenas da possível **presença** de um transtorno, mas não da sua **ausência** – e que portanto não podem ser tomados por base para seleção de indivíduos que deveriam compor a classe negativa do aprendizado – faz-se necessário utilizar neste caso uma formulação distinta para o problema. Mais especificamente, ao invés de utilizar um conjunto de usuários (por exemplo) sem depressão, utiliza-se um grupo de usuários de controle selecionados aleatoriamente (e que portanto também inclui um certo número de indivíduos com depressão não detectáveis) de proporção várias vezes superior ao da classe positiva (normalmente proporcional à incidência do transtorno na população em geral). Assim, o problema computacional a ser resolvido não consiste em distinguir indivíduos com e sem depressão (e na verdade a informação sobre ausência de depressão é inexistente nos dados), e sim em **identificar indivíduos com probabilidade acima da média da população geral de virem a receber um futuro diagnóstico**. Esta probabilidade, que entre indivíduos da classe positiva será próxima de 100%, é necessariamente muito mais baixa em um grupo de controle aleatório, e pode assim ser capturada com métodos computacionais variados como o aprendizado profundo.

Em virtude destas características, o uso de autorrelatos para identificar indivíduos (provavelmente) diagnosticados com um transtorno de saúde mental configura um problema de AM fortemente desbalanceado. Mesmo com estas dificuldades, entretanto, a seleção de indivíduos com transtorno de saúde mental por meio de seus próprios autorrelatos é

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

possivelmente o método de coleta de dados mais popular na pesquisa na área de PLN (Cohan et al., 2018; Losada et al., 2019; Santos et al., 2024b; Yates et al., 2017).

10.1.1.2 Classificação em nível textual

A classificação de nível textual recebe como entrada um texto individual (e.g., uma postagem em rede social) e objetiva verificar se este texto expressa algum tipo de informação relacionada à saúde mental (e.g., se é de caráter depressivo ou não) sem levar em conta a identidade do seu autor. Modelos deste tipo são comumente empregados para detectar sintomas de depressão (Yadav et al., 2020), verificar a pertinência do texto para detecção de um transtorno (Santos; Paraboni, 2024), entre outras aplicações. A classificação em nível textual pode também ser tratada como um pré-requisito para a classificação de nível autoral, procurando filtrar textos de maior interesse para a tarefa em meio a uma grande massa de dados ruidosos (e.g., provenientes de redes sociais). Uma estratégia possível de combinação destes dois níveis de processamento é, por exemplo, o uso de aprendizado baseado em múltiplas instâncias (Mann et al., 2022).

Ao contrário da classificação em nível autoral, em que atribui-se um rótulo a uma coleção de textos (e.g., uma *timeline* de rede social) com base em um relato feito pelo próprio autor (ou usuário), a classificação em nível textual pressupõe a disponibilidade de textos rotulados individualmente com o tipo de informação a ser aprendida. Isso demanda um esforço adicional de anotação de *corpus*, ou algum método de rotulação automática implementado com uso de grandes modelos de linguagem (-cap-modelos-linguagem) e engenharia de *prompts* (Santos; Paraboni, 2023; Santos; Paraboni, 2024).

Um exemplo de projeto que trabalhou na classificação em nível textual foi o Amive[2] (Seção 10.5). Nesse projeto, postagens anônimas de páginas de segredos universitários foram coletadas por meio da busca por palavras-chave associadas à depressão e ideação suicida. Essas postagens foram, então, rotuladas com 21 sinais relacionados a depressão (sintomas, fatores de risco e protetivos) por estudantes e profissionais das áreas de psicologia, medicina e terapia ocupacional. Contudo, a anotação em uma granularidade maior, em dados de conteúdo sensível, por especialistas humanos é um trabalho bastante oneroso que demanda tempo e deve ser feito com cuidado para não afetar também a saúde mental dos anotadores.

10.2 Conjuntos de dados para detecção de depressão/ansiedade

A detecção de depressão e ansiedade a partir de textos faz o uso de *corpus* especialmente coletados e rotulados para este propósito. Recursos linguístico-computacionais deste tipo já foram desenvolvidos em quantidade expressiva para o idioma inglês, em alguns casos como subproduto de eventos do tipo *shared task* como o desafio CLPsych-2015 (Coppersmith et al., 2015) e a série eRisk (Losada et al., 2019). Iniciativas para o português incluem o estudo em (Mann et al., 2020) e o *corpus* SetembroBR (Santos et al., 2024b). Nesta seção apresentamos um levantamento de alguns recursos linguísticos-computacionais existentes na área. Para cada um dos estudos selecionados, a Tabela 10.1 apresenta o domínio, idioma e modalidade de dados, o número de instâncias de depressão/ansiedade, a relação entre usuários de controle (C) e diagnosticados (D) e o método de pareamento das classes. Detalhes adicionais são discutidos a seguir.

[2] https://www.amive.ufscar.br/

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Tabela 10.1: Exemplos de *corpora* para detecção de depressão/ansiedade

Corpus	Domínio	Idioma	Modalidade	Dep.	Ans.	C/D	Pareamento
CLPsych-2015	Twitter/X	En	texto	477	-	1	idade, gênero
RSDD	Reddit	En	texto	9210	-	11,6	comportamento
Shen et al.	Twitter/X	En	texto	1402	-	1	na
eRisk-2017	Reddit	En	texto	135	-	5,6	na
eRisk-2018	Reddit	En	texto	214	-	7	na
SMHD	Reddit	En	texto	14139	8783	9	comportamento
Mann et al.	Instagram	Pt	texto, imagens	82	-	1,7	na
SetembroBR	Twitter/X	Pt	texto, rede soc.	1283	1767	7	gênero, data, tamanho
Amive	Facebook	Pt	texto	780	-	-	na

Com base neste levantamento, observa-se que os *corpora* existentes tendem a ser baseados principalmente nos domínios Reddit e Twitter/X, e que a maioria dos recursos listados são do tipo textual e dedicados ao idioma inglês. Para o português temos o já citado estudo em (Mann et al., 2020), que se destaca por ser o único a considerar o domínio Instagram e baseado no idioma português, e também por incluir dados de imagens; o *corpus* SetembroBR (Santos et al., 2024b) a ser discutido em mais detalhes na Seção 10.4 e único do gênero a incluir informações relativas à rede social dos usuários representados na base de dados; e o *corpus* do projeto Amive, apresentado em detalhes na Seção 10.5. Além disso, observa-se que os *corpora* SetembroBR e SMHD são os únicos que contemplam explicitamente casos de transtorno de ansiedade[3]. É importante observar, entretanto, que diversos métodos de detecção de transtornos de saúde mental discutidos na próxima seção fazem uso de dados multimodais (e.g., imagens e características da rede social etc.) mas que este tipo de informação, via de regra, não é parte integrante de um *corpus* publicamente disponibilizado para pesquisa.

No que diz respeito ao volume de dados, observa-se que os *corpora* considerados apresentam grande variação de tamanho (aqui medida em quantidade de indivíduos diagnosticados), e que recursos baseados em dados da rede social Reddit tendem a ser maiores. Isso é explicável pela organização dos dados em grupos de discussão (e.g., sobre depressão) nessa plataforma e, consequentemente, pela maior facilidade de identificação de indivíduos diagnosticados. Dados no domínio Twitter/X, por outro lado, são consideravelmente menos estruturados, o que torna mais complexa a tarefa de seleção de indivíduos com transtorno de saúde mental. No caso do *corpus* do projeto Amive, a quantidade apresentada na tabela indica o total de *posts* do Facebook coletados para serem anotados pelos especialistas em saúde mental a partir de busca por palavras-chave relacionadas à depressão e ideação suicida. Neste sentido, esse *corpus* difere dos demais uma vez que as instâncias não se referem a indivíduos diagnosticados, mas sim a postagens que podem ter sinais relacionados à depressão. Pela mesma especificidade desse *corpus*, não faz sentido indicar a proporção C/D nesse caso.

Observa-se de modo geral uma grande variação também na proporção entre indivíduos de controle e diagnosticados, indicada na coluna C/D da tabela, e que é um aspecto crucial da modelagem do problema de classificação a ser resolvido. Dentre os estudos selecionados, o *corpus* CLPsych-2015 é o único, dentre os mais amplamente utilizados na área, que se limita a um balanceamento artificialmente ideal, criado com o objetivo de simplificar o uso

[3]Como descrito na Seção 10.5, ansiedade é um dos sintomas contemplados no projeto Amive (Preocupação, medo ou ansiedade). Contudo, no escopo do projeto Amive, ele foi considerado um sinal relacionado à depressão e não foi usado especificamente para classificação do transtorno de ansiedade.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

de métodos de AM no contexto da competição (ou *shared task*) de mesmo nome, enquanto que em todos os demais procurou-se considerar de alguma forma a necessidade de modelar um cenário mais realista. Dentre as estratégias adotadas, observa-se a modelagem de um grupo de controle de dimensão bastante superior ao do grupo de diagnosticados (Cohan et al., 2018; Losada et al., 2017, 2018; Santos et al., 2024b; Yates et al., 2017), o uso de múltiplas seleções de indivíduos aleatórios a partir de uma grande massa de dados (Shen et al., 2017), ou simplesmente o uso de inventários clínicos (Mann et al., 2020) como forma de selecionar indivíduos livres de depressão (a um custo possivelmente elevado, refletido no menor volume de indivíduos obtidos, porém dispensando o uso de um grupo de controle aleatorizado).

Outra questão fundamental na construção de um *corpus* deste tipo é o pareamento entre indivíduos diagnosticados e suas contrapartidas no grupo de controle, já que diferenças indesejáveis entre os dois conjuntos podem levar ao aprendizado de padrões espúrios. Por exemplo, caso indivíduos diagnosticados e de controle sejam extraídos de populações diferentes (e.g., com publicações feitas em períodos muito distintos), é possível que diferenças de vocabulário, decorrentes da evolução natural da língua, sejam tratadas como características relevantes para distinção entre as duas classes. Assim, como forma de assegurar que os dois grupos sejam minimamente comparáveis, o *corpus* CLPsych-2015 se destaca como o único a garantir pareamento de gênero e faixa etária entre indivíduos de ambas as classes, o que minimiza possíveis diferenças de vocabulário entre os dois grupos. O pareamento entre classes também é uma preocupação do *corpus* RSDD e sua extensão SMHD, em que indivíduos diagnosticados e de controle são selecionados com base no seu comportamento na rede social (e.g., grupos de discussão dos quais participam, frequência de postagens etc.). Este tipo de pareamento é um quesito importante no caso da plataforma Reddit, mas não possui contrapartida direta no domínio Twitter/X. No *corpus* SetembroBR, baseado na plataforma Twitter/X, foi realizado o pareamento por datas aproximadas de publicação, volume de postagens e gênero dos usuários. No caso do *corpus* do projeto Amive, uma vez que as postagens são anônimas, o pareamento com base em informações como gênero e idade não seria possível. Nesse caso, como o foco era o público alvo de estudantes universitários, buscou-se minimizar o impacto de algum viés variando-se as páginas de Segredos Universitários a partir das quais as postagens foram coletadas.

Com exceção do *corpus* SetembroBR a ser discutido em mais detalhes na Seção 10.4, os recursos aqui discutidos não levam em conta a noção de detecção **precoce** de transtornos, ou seja, anterior ao momento em que o indivíduo é oficialmente diagnosticado por um profissional da área de saúde. Em todos os casos discutidos, os dados constantes do *corpus* contemplam publicações feitas antes e depois do diagnóstico ou tratamento do indivíduo, e podem assim incluir um número potencialmente grande de indicadores a esse respeito. Por exemplo, uma vez que um indivíduo é diagnosticado ou começa a fazer tratamento para depressão, pode ser natural que ele passe a falar (ou falar mais) sobre este assunto, ou que relate as conversas que tem com o psicólogo, os efeitos da medicação etc. Todos estes indicadores, que em maior ou menor grau estão presentes em todos os *corpora* aqui discutidos, podem não só reduzir o grau de realismo do desafio computacional, como talvez reduzir a própria utilidade prática do método desenvolvido. Além disso, cabe observar que esta limitação está presente até mesmo nos *corpora* disponibilizados na série de desafios e-Risk (Losada et al., 2017, 2018, 2019), que mesmo possuindo o propósito explícito de detecção de primeiros sinais de transtornos de saúde mental como depressão, não fazem nenhuma distinção efetiva entre dados publicados antes ou depois do seu diagnóstico oficial, limitando-se apenas ao desafio computacional de classificar estes indivíduos com base no

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

menor volume de dados possível seguindo a ordem cronológica das publicações.

Finalmente, é importante destacar que *corpora* textuais rotulados com informações sobre saúde mental geralmente não são recursos computacionais totalmente acessíveis, como seriam *corpora* voltados a outros tipos de aplicação de PLN. Além da natureza sensível do problema e das considerações éticas associadas, algumas redes sociais como a plataforma Twitter/X proíbem a reprodução e divulgação de conteúdo gerado por seus usuários como forma de proteger a privacidade e o direito de arrependimento destes. Assim, *corpora* de tweets como SetembroBR, quando disponibilizados, são apenas coleções de identificadores numéricos a partir dos quais as postagens integrais podem ser recuperadas diretamente a partir da API da plataforma, caso ainda sejam disponibilizadas pelos seus autores. Este acesso, via de regra, acarreta algum esforço de programação e, no caso da plataforma Twitter/X, deixou de ser gratuito em fevereiro de 2023.

10.3 Modelos computacionais de detecção de depressão/ansiedade

A detecção de transtornos de saúde mental apresenta os desafios comuns à boa parte das aplicações de PLN. Em especial, observa-se que a distinção entre indivíduos diagnosticados e aqueles que pertencem ao grupo de controle com base em características linguísticas é naturalmente dificultada pelo fato de que ambos os grupos podem discutir qualquer tipo de assunto, incluindo questões relacionadas à saúde mental (deles próprios, de terceiros etc.), em uma variedade de contextos reais e imaginários, fazendo uso de humor, sarcasmo, metáforas, conjecturas e muitas outras formas de expressão. Além disso, no caso do aprendizado a partir de textos provenientes de redes sociais, observa-se que os dados de entrada são normalmente esparsos, isso é, com grande incidência de ruído na forma de postagens sem relação com o problema a ser tratado. É comum, por exemplo, deparar-se com *timelines* de usuários do Twitter/X que, apesar de sabidamente diagnosticados com depressão, não fazem qualquer referência a questões de saúde mental.

O que possui um texto para que seja classificado como relacionado a um determinado transtorno de saúde mental? De um lado, são conhecidos diversos indicadores linguísticos estabelecidos na literatura médica (Trifu et al., 2017). Indicadores deste tipo incluem o uso frequente de pronomes de primeira pessoa, tempo passado em verbos de ação, inversão da ordem de palavras em tópicos, ênfases, presença de sentenças curtas, impessoais, truncadas ou ríspidas, elipses, tautologias, repetições e ausência de comparação. O uso de termos absolutistas (e.g., todo, sempre etc.) também é mais fortemente associado a indivíduos com depressão (Johnstone, 2018), assim como o uso de expressões de conotação negativa em certas redes sociais (Rickard, 2018). Finalmente, alguns estudos exploram também o uso de informações não textuais como conexões da rede social, frequência e horário de publicações e outros (Choudhury et al., 2013; Shrestha; Spezzano, 2019).

Indicadores linguísticos de transtornos de saúde mental são normalmente modelados como sequências de *tokens* ou outras representações textuais tradicionais do PLN. Embora de valor indiscutível (por exemplo, para capturar o uso de referências pronominais etc.), a eficácia destes indicadores depende em grande parte da disponibilidade de um conjunto de treino suficientemente grande, onde todos os exemplos possíveis tenham ocorrido em número significativo para aprendizado automático. Assim, por exemplo, um modelo de AM reconhece a correlação entre sequências de *tokens* como "eu me sinto péssimo", ou "psicóloga disse que eu tenho que me cuidar melhor" e saúde mental. Por outro lado,

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

exemplos mais raros, mesmo que altamente significativos, tendem a ser ignorados. Este é o caso por exemplo do uso de linguagem figurativa como em "acho que vou pular da ponte" como uma expressão de ideação suicida.

Uma possível forma de compensar a insuficiência de exemplos (que é em certa medida inerente a qualquer abordagem supervisionada) seria o uso de grandes modelos de linguagem como base de conhecimento auxiliar. Neste sentido, estudos como (Santos; Paraboni, 2024) demonstram que modelos como GPT (Brown et al., 2020c) podem capturar com facilidade noções complexas como a do exemplo anterior, e que este conhecimento pode ser combinado com abordagens tradicionais (e.g., supervisionadas) para detecção de transtornos de saúde mental em dois níveis, ou seja, considerando tanto indicadores linguísticos de baixo nível como sequências de *tokens* e conceitos de alto nível como sintomas. Detalhes desta estratégia são discutidos na Seção 10.4.

O Quadro 10.1 categoriza uma série de exemplos de estudos da área de acordo com o tipo de problema considerado (d=depressão, a=ansiedade, s=sintomas de depressão, v=grau de severidade da depressão, *=outros), tarefa (det=detecção, e.d = detecção precoce, exp=explicação), domínio (Reddit, Twitter/X etc.), idioma (En=inglês, Ch=chinês, Pt=português), características de aprendizado (e=*embeddings*, t=tópicos, n=informações de rede, u=informações referentes ao usuário, l=atributos psicolinguísticos LIWC, p=*part-of-speech*, d=dicionário de domínio, s=sentimentos/emoções, i=imagens, b=*bag of words*, m=metadados, t=informação temporal, ou texto simples), e métodos de aprendizado (e.g., CNN=redes neurais convolucionais, LSTM=*long short-term neural networks*, MLP=*multilayer perceptron*, NN=outras arquiteturas neurais, LR = regressão logística, RF=Random Forest, DT=árvore de decisão, NB=Naive Bayes, XG=XGBoost etc.). Detalhes adicionais são discutidos a seguir.

Quadro 10.1: Exemplos de modelos computacionais de detecção de depressão e ansiedade

Estudo	Alvo	Tarefa	Domínio	Líng.	Atributos	Abordagem
(Yates et al., 2017)	d	det	Reddit	En	e	CNN
(Yazdavar et al., 2017)	s	det	Twitter/X	En	t	LDA
(Shen et al., 2017)	d	det	Twitter/X	En	n, u, i, t, d, l, e	LR
(Shen; Rudzicz, 2017)	a	det	Reddit	En	e, t, l	MLP
(Loveys et al., 2017)	a,*	det	Twitter/X	En	s	VADER
(Cohan et al., 2018)	d,a,*	det	Reddit	En	b, e	FastText
(Song et al., 2018)	d	exp	Reddit	En	d, s, t, m, p	MLP+RNN
(Trotzek et al., 2018)	d	e.d	Reddit	En	e, p, m, t, d	CNN
(Nascimento et al., 2018)	d	det	Blogs	Pt	s	DT
(Shen et al., 2018)	d	det	Weibo	Ch	s, p, t, m, i, u, n	NN
(Arora, 2019)	d,a	det	Twitter/X	En	d, s, t	ensemble
(Aragón et al., 2019)	d	e.d	Reddit	En	s	SVM
(Carneiro, 2019)	d	e.d	Reddit	En	t, m, n	RF
(Santos et al., 2020c)	d,a,*	det	Twitter/X	Pt	e, b	LR
(Burdisso et al., 2020)	d	e.d	Reddit	En	b	SS3
(Lin et al., 2020a)	d	det	Twitter/X	En	e, i	CNN
(Hitzler, 2020)	d	det	Twitter/X	En	b, s, t, i, n, l, u	RF
(Mann et al., 2020)	v	det	Instagram	Pt	i, e	NN
(Souza et al., 2021b)	a,d	det	Reddit	En	e	LSTM+CNN
(Ansari; Ji, 2022)	d	det	Reddit,Twitter/X	En	e,s	LR+LSTM
(Mann et al., 2022)	v	det	Instagram	Pt	e	BERT,LSTM
(Santos et al., 2024b)	d,a	det	Twitter/X	Pt	e, n	BERT,LSTM
(Santos et al., 2023)	d,a	det	Twitter/X	Pt	e	BERT,MoE
(Santos; Paraboni, 2023)	d	det	Twitter/X	Pt	texto	ChatGPT
(Oliveira; Paraboni, 2024)	d,a	det	Twitter/X	Pt	e, n	BERT,LSTM,LR
(Oliveira et al., 2024b)	d,a	det	Twitter/X	Pt	e, n	BERT,LSTM,LR
(Santos; Paraboni, 2024)	d	det	Twitter/X	Pt	texto	GPT
(Mendes; Caseli, 2024)	d	det	Facebook	Pt	e, l, p, d, b	SVM,LR,NB,RF,XG,BERT

No que diz respeito ao tipo de transtorno a ser detectado, observa-se que estudos focados na depressão (d) tendem a ser muito mais frequentes do que os focados no transtorno de ansiedade (a). Além disso, observa-se que o estudo de (Arora, 2019) trata da questão da

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

ansiedade depressiva, um transtorno misto que combina sintomas de depressão e ansiedade, e que o trabalho de (Souza et al., 2020b) investiga a comorbidade destes dois transtornos. Alguns estudos enfocam ainda a detecção de sintomas (Mendes; Caseli, 2024; Yazdavar et al., 2017) ou grau de severidade (Mann et al., 2020) destes transtornos.

Considerando-se a tarefa computacional a ser implementada, é possível observar que a maioria dos estudos identificados trata da detecção (det) de depressão/ansiedade ou sua detecção com base no menor volume possível de evidência (e.d), que é uma formulação específica do problema adotada na série de *shared tasks* eRisk (Losada et al., 2019), e/ou baseada nos *corpora* derivados destas competições. A preocupação com a necessidade de explicar os passos que levaram ao resultado da classificação é o foco do estudo de (Song et al., 2018).

Assim como no caso dos *corpora* discutidos na seção anterior, os estudos identificados são, com poucas exceções, baseados em dados nos domínios Twitter/X ou Reddit. Vários destes estudos desenvolvem seus próprios conjuntos de dados (especialmente quando considerando características não textuais da rede social ou outras não disponibilizadas em *corpora* de uso público), enquanto outros reutilizam um *corpus* de referência na área, como o RSDD (Yates et al., 2017), SMHD (Cohan et al., 2018) ou eRisk (Losada et al., 2017) no domínio Reddit, ou ainda o *corpus* de tweets discutido em (Shen et al., 2017).

Também conforme as características dos *corpora* discutidos na seção anterior, a maioria dos estudos identificados é dedicada ao idioma inglês (En). Há entretanto iniciativas isoladas para diversos outros idiomas, e que não são discutidos aqui porque fogem ao escopo deste livro. Estas iniciativas incluem os idiomas árabe (Almouzini et al., 2019), japonês (Tsugawa et al., 2015), coreano (Park et al., 2013), romeno (Briciu; Lupea, 2018), russo (Semenov et al., 2015), e tailandês (Katchapakirin et al., 2018). No caso da língua portuguesa (Pt), identificamos os estudos de (Mann et al., 2020, 2022), que tratam da questão da detecção de depressão em um *corpus* multimodal no domínio Instagram com base em questionários.

No que diz respeito aos modelos e métodos computacionais empregados, observa-se exemplos de estudos que exploram praticamente todo tipo de informação disponibilizada em redes sociais, incluindo dados textuais, temporais, estruturais da rede, e ainda características demográficas e comportamentais dos usuários, geralmente de forma combinada. A variedade de métodos de aprendizado também é significativa, com uma certa preferência por modelos neurais. Assim como em outras áreas do PLN, estudos mais recentemente passaram a utilizar modelos de língua pré-treinados do tipo BERT (Devlin et al., 2019), ELMo (Peters et al., 2017) e similares (Mann et al., 2022; Mendes; Caseli, 2024; Santos et al., 2024b), ou baseados em grandes modelos de linguagem e engenharia de *prompts* (Santos; Paraboni, 2023; Santos; Paraboni, 2024).

As próximas seções focam nos estudos especificamente desenvolvidos para o idioma português brasileiro, uma vez que este é o foco deste livro.

10.4 SetembroBR: detecção de depressão/ansiedade em português

Como forma de ilustrar o desenvolvimento de modelos de detecção de transtornos de saúde mental a partir de dados provenientes de redes sociais em Português, esta seção discute o caso do projeto SetembroBR, dedicado à detecção de depressão e transtorno de ansiedade no Twitter/X do Brasil.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

A Seção 10.4.1 descreve o *corpus* construído, e a Seção 10.4.2 apresenta algumas abordagens computacionais para a tarefa em questão.

10.4.1 *Corpus* SetembroBR

O *corpus* SetembroBR é uma base de dados mutimodal específica para o português, enfocando a detecção precoce de transtornos de depressão e ansiedade a partir de dados textuais e não-textuais no domínio Twitter/X[4]. A construção deste recurso partiu de um estudo-piloto em (Santos et al., 2020c), cuja definição da tarefa foi posteriormente reformulada e apresentada em sua forma final em (Santos et al., 2024b).

O *corpus* contempla *timelines* de usuários e meta-dados associados, como informações de suas relações na rede social e outros. Estes indivíduos foram selecionados com base em seus autorrelatos contendo indicação explícita do momento do diagnóstico (e.g., 'Na semana passada o meu psicólogo me informou que eu tenho depressão') feito por um profissional da área médica (e.g., psicólogos etc.).

Nesta abordagem, assim como em vários outros estudos da área (Cohan et al., 2018; Coppersmith et al., 2015; Yates et al., 2017), a seleção de usuários é feita de forma cuidadosa consultando-se a plataforma Twitter/X em busca de autorrelatos que correspondem a uma ampla gama de expressões regulares de interesse, e seguida de inspeção manual. Diferentemente de estudos prévios, entretanto, na presente abordagem a coleta dos dados (tweets) propriamente dita é feita também com inspeção manual, explorando-se a ordenação cronológica da plataforma Twitter/X de modo a restringi-los ao conjunto de dados anteriores ao evento relatado. Um exemplo de como esta porção de dados dita 'útil' (para fins de classificação) é delimitada é ilustrada pelo marcador [end] na Quadro 10.2 com base em um autorrelato indicado pelo marcador [msg].

Quadro 10.2: Timeline de um usuário com marcador [end] indicando o término da porção de dados a ser considerada na predição de depressão, em que todos os tweets abaixo do ponto [end] são descartados.

Data	Marcador	Texto
Mon March 25		Deixei meu celular em casa e agora a bateria está morta.
Tue March 26		Eu assisti esse filme duas vezes no ano passado.
Thu March 28	[end]	tão feliz que finalmente comprei meus óculos novos LOL
Mon April 1		Vou dormir agora. Amanhã é um grande dia.
Wed April 3		@usuário você nunca me contou isso.
Mon April 8		Pensando em ligar para ela de novo hoje à noite...
Fri May 3		Oiiii! como você está?
Sun May 5	[msg]	Mês passado o psiquiatra me diagnosticou com depressão :(

Neste exemplo, observa-se que no dia 5 de maio (na parte inferior da *timeline*), o usuário relata um diagnóstico recebido em uma data não especificada do mês anterior (abril). Assim, todos os tweets anteriores ao mês de abril, até o ponto indicado como [end], são coletados para suporte à tarefa de detecção precoce de depressão. O restante dos dados (a partir de 1o. de abril e incluindo o próprio autorrelato em [msg]) são descartados. Neste mesmo exemplo é possível ainda observar que, como em uma típica *timeline* de rede

[4]A escolha do domínio Twitter/X é motivada pela sua maior popularidade no Brasil.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

social, a grande maioria das postagens tende a apresentar pouca ou nenhuma relação com questões de saúde mental.

O *corpus* coletado segundo esta metodologia, detalhada em (Santos et al., 2024b), consiste em um conjunto de 46,8 milhões de tweets (totalizando cerca de 555 milhões de *tokens*) em Português, produzidos por 18.819 usuários únicos. Dentre estes usuários, 3.163 são considerados, em função de seus autorrelatos externos ao *corpus*, como tendo um futuro diagnóstico de depressão, ansiedade ou ambos. Os demais usuários formam o grupo de controle na proporção de 7 indivíduos aleatórios para cada indivíduo diagnosticado.

De acordo com as restrições da plataforma Twitter/X, o *corpus* é disponibilizado para reúso no formato de lista de identificadores numéricos de cada tweet divididos em conjuntos de treino e teste padronizados, a partir dos quais o pesquisador interessado pode recuperá-los utilizando-se a API da rede social. Estão disponíveis[5] informações para *download* e a rotulação do subconjunto Depressão com uso do modelo GPT 3.5., discutida na seção a seguir.

10.4.2 Modelos desenvolvidos no projeto SetembroBR

A seguir enumeramos alguns exemplos de uso do *corpus* SetembroBR aplicados à tarefa de detecção de depressão/ansiedade utilizando métodos computacionais variados. Para detalhes de cada abordagem e seus resultados, sugere-se consultar as publicações referenciadas em cada caso.

10.4.2.1 Abordagens supervisionadas tradicionais

A disponibilização do *corpus* SetembroBR incluiu a descrição de uma série de sistemas *baseline* construídos com uso de redes do tipo BiLSTM e do modelo de língua BERT (Devlin et al., 2019) e outros existentes na literatura da área para predição de depressão e transtorno de ansiedade a partir de dados textuais e não textuais. Dentre as arquiteturas consideradas, destaca-se o uso de LSTM que recebe como entrada uma representação em nível de usuário construída com uso de BERT. Nesta estratégia, a representação de indivíduo consiste de sequências consecutivas de 10 tweets iniciadas em uma posição aleatória da *timeline* do usuário a cada época. Os tweets de entrada são transformados em uma representação criada com uso do modelo BERTimbau (Souza et al., 2020a), e a última camada é usada como representação final da sequência. Esta arquitetura é seguida de uma camada BiLSTM com função de ativação ReLu que alimenta uma sequência de camadas totalmente conectadas com regularização *dropout*, e usando *softmax* como função de ativação na camada de saída. O desbalanceamento de classes é tratado com uso de uma função binária de entropia cruzada com pesos balanceados.

Esta arquitetura foi posteriormente aprimorada com uso de um modelo BERT treinado especificamente para o domínio do Twitter/X em português, denominado BERTabaporu (Costa et al., 2023). Este modelo foi treinado em uma massa de 238 mil tweets, totalizando cerca de 2,9 bilhões de *tokens*. Os resultados deste modelo foram comparados aos obtidos pelo tradicional BERTimbau (Souza et al., 2020a), treinado em cerca de 2,7 bilhões de *tokens*, para as tarefas de predição de transtornos de saúde mental e outras em . O modelo BERTabaporu propriamente dito está disponível para reúso no repositório Hugging Face[6].

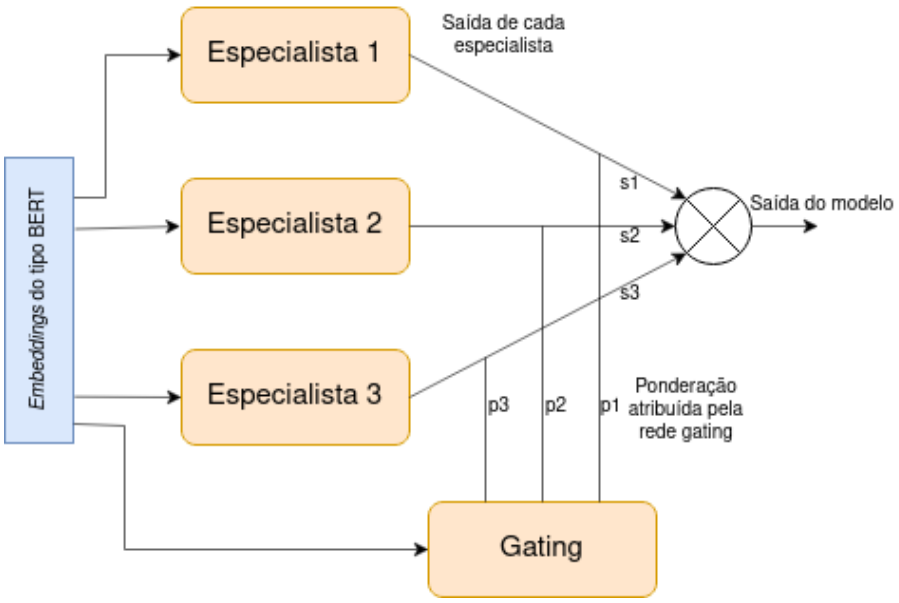
[5] https://drive.google.com/drive/folders/1MXFRs0u8iF1RNUWABTA0Oz8_Ix1skqZT?usp=sharing
[6] https://huggingface.co/pablocosta/bertabaporu-large-uncased

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

Finalmente, estes modelos foram combinados em uma arquitetura de comitê de máquina dinâmico de mistura de especialistas apresentada em (Santos et al., 2023), e que é baseada em (Jacobs et al., 1991). Esta arquitetura é constituída de três modelos especialistas com uma rede *gating* ponderadora na qual cada modelo especialista, assim como sua própria rede *gating*, consistem em uma rede do tipo LSTM que recebe como entrada uma representação em nível de usuário construída pelo modelo de representação distribuída BERT (Devlin et al., 2019). A Figura 10.2, extraída de (Santos et al., 2023), ilustra esta arquitetura.

Figura 10.2: Arquitetura de mistura de especialistas em (Santos et al., 2023)



Os experimentos apresentados em (Santos et al., 2023) confirmam a superioridade do modelo baseado em mistura de especialistas (MoE) usando BERT. Para a tarefa de detecção de depressão, MoE apresenta F1 médio de 0,65, enquanto a abordagem de engenharia de características apresenta F1 médio de 0,56. Para a tarefa de detecção de transtorno de ansiedade, são obtidos F1 médio de 0,60 e 0,53, respectivamente.

10.4.2.2 Abordagem *zero-shot* usando ChatGPT

Como alternativa às abordagens supervisionadas tradicionais, o estudo em (Santos; Paraboni, 2023) investigou a tarefa de predição de depressão com base nos dados do *corpus* SetembroBR em modalidade *zero-shot* com uso da ferramenta ChatGPT[7]. Para este fim, o modelo foi instruído a avaliar sequências aleatórias de 80 tweets anonimizados cada. A avaliação consistia em atribuir um escore em uma escala de 0 a 10 de acordo com a presença de indicadores linguísticos de depressão no texto, em que 0 significa nenhum indício e 10 significa indícios muito fortes. Com base nos textos assim rotulados, foi calculado então um *threshold* mínimo para diferenciar as classes positiva e negativa com base em dados de treinamento (de resto não empregados na presente abordagem *zero-shot*), obtendo-se assim o rótulo da *timeline* (ou seja, do autor dos textos) como um todo.

[7] https://chat.openai.com/chat

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

A abordagem *zero-shot* baseada em ChatGPT foi comparada com a arquitetura BERT tradicional discutida na seção anterior, obtendo resultado médio equivalente (F1 médio de 0,65 para a abordagem ChatGPT e 0,66 para BERT supervisionado), embora com grande diferença nas classes individuais. A abordagem baseada em ChatGPT apresentou ampla superioridade para a classe positiva (Diagnosticados) com F1 de 0,70 contra apenas 0,48 da abordagem BERT, enquanto que a abordagem supervisionada baseada em BERT foi superior para a classe Controle com F1 de 0,84 contra 0,60 da abordagem ChatGPT. O estudo em (Santos; Paraboni, 2023) argumenta que essa diferença é explicável pelo fato de que o modelo de linguagem subjacente ao ChatGPT não possuir uma definição clara do que seria a população média representada pela classe Controle do *corpus*, enquanto que a abordagem supervisionada possui um grande volume de exemplos da classe, e teria assim melhores condições de distingui-la da classe positiva.

Como resultado, observa-se que a abordagem *zero-shot* pode ser considerada superior tanto pela acurácia obtida para a classe de interesse (que é a positiva, não a aleatória), mas igualmente pelo fato de não fazer uso de *corpus* rotulado. Cabe destacar, entretanto, que os resultados em (Santos; Paraboni, 2023) são preliminares e não podem ser comparados diretamente com os estudos anteriores dado que a abordagem foi testada apenas em um subconjunto dos dados de teste do *corpus* em virtude do custo associado ao uso da ferramenta ChatGPT.

10.4.2.3 Engenharia de *prompts* com supervisão

Inspirado nos resultados iniciais obtidos com uso da ferramenta ChatGPT, o estudo em (Santos; Paraboni, 2024) explorou a questão da natureza do conhecimento útil para detectar depressão a partir de dados textuais, e a necessidade de filtrar o ruído proveniente da rede social, ou seja, as postagens ditas irrelevantes para a tarefa.

Conforme discutido na Seção 10.3, observa-se que parte da solução para o problema parece ser, por um lado, encontrada no uso de características de teor mais linguístico, como sequências de *tokens*, o uso de certas formas pronominais e afins. Por outro lado, abordagens puramente linguísticas são menos úteis na captura de indicadores de mais alto nível, ou mais próximos da aplicação, como no caso de sintomas de depressão. Casos deste tipo, que exigem um processamento mais profundo (ou um maior 'entendimento' do texto) são tratados de forma eficiente com uso de *prompts* submetidos a um grande modelo de linguagem. O estudo realizado tratou assim de combinar estas duas visões do problema, ou seja, a abordagem supervisionada e a engenharia de *prompts*.

O estudo em (Santos; Paraboni, 2024) utilizou o modelo de língua GPT 3.5. (Brown et al., 2020c) para avaliar uma amostra aleatória de tweets do *corpus* SetembroBR de acordo com a relevância para a previsão de depressão. Como resultado, estes dados foram categorizados como sendo de alta (1), média (2) ou baixa (3) relevância para a saúde mental. Esta rotulação automática foi então utilizada para treinar um classificador de postagens de alta, média e baixa relevância de acordo com o modelo de língua T5 (Raffel et al., 2020), e este classificador foi então aplicado à rotulação do conjunto completo de 19,42 milhões de postagens do sub*corpus* de depressão. Este conjunto de rótulos em nível de postagem, provisoriamente chamada de SetembroBRGPT foi parcialmente disponibilizada (no momento apenas para o subconjunto Depressão) para reúso juntamente com o *corpus* original (rotulado ao nível de usuário).

Como esperado, a maioria das publicações (65,2% na classe Diagnosticado e 71,3% na classe Controle) são consideradas de baixa relevância para a predição de saúde mental de

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

acordo com o modelo GPT. Por outro lado, publicações de alta relevância são relativamente raras (5,4% na classe Diagnosticados e 3,0% na classe Controle). No entanto, tão ou mais importante do que esta distribuição é a observação de que, ao consultar um LLM sobre a relevância de uma publicação para a saúde mental, estamos nos concentrando em grande parte na semântica, isto é, em sintomas e outros sinais clínicos bem conhecidos de depressão, mas isto não quer dizer que outros fatores possam ou devam ser ignorados. Em outras palavras, o *prompt* utilizado permite ao modelo identificar uma vasta gama de publicações que podem sugerir, por exemplo, distúrbios alimentares ou uso de linguagem negativa, ambos sabidamente relacionados à depressão (American Psychiatric Association, 2013), mas não considera explicitamente indicadores linguísticos de depressão mais refinados, como o uso de pronomes de primeira pessoa (Trifu et al., 2017), termos absolutos (Johnstone, 2018) e outros. Estes indicadores, que podem estar presentes em qualquer publicação de baixa ou alta relevância para a depressão, são também importantes preditores do estado de saúde mental, e sugerem que nenhuma publicação deveria em princípio ser descartada com base apenas nos resultados do LLM.

Como um meio de manter os dados de treino integralmente disponíveis para o classificador e, ao mesmo tempo, distinguir mensagens mais e menos relevantes (que mostram claramente diferentes distribuições no *corpus*, conforme discutido em (Santos; Paraboni, 2024)), foi proposta uma abordagem que usa os rótulos fornecidos pelo LLM para dividir os dados de treinamento em subconjuntos de alta, média e baixa relevância, a partir dos quais foram criados três vetores individuais de contagens TF-IDF reduzidos com seleção univariada. Além disso, como forma de capturar também a noção de ordem das postagens de diferentes graus de relevância, a esta representação foi anexado ainda um vetor de contagens de bigramas dos rótulos alto/médio/baixo computados. Esta adição foi motivada pela observação de que certas sequências de postagens (por exemplo, uma ocorrência de vários tweets consecutivos e altamente relevantes para a detecção de depressão) podem ser indicativas do estado de saúde mental do indivíduo.

A arquitetura proposta procura assim capturar (i) a noção de postagens mais ou menos importantes (e.g., postagens expressando sintomas de depressão etc.), (ii) os indicadores linguísticos de baixo nível que podem estar presentes em qualquer postagem, incluindo as que não tenham nenhuma relação com saúde mental, e ainda (iii) a noção de sequências de postagens consecutivas com maior ou menor importância para a tarefa. O conjunto final de 55.000 características empregadas para esse fim (i.e., 15.000 contagens de *tokens* mais 40.000 contagens de bigramas de rótulos) é ilustrado na Figura 10.3.

Figura 10.3: Modelo de 15.000 tokens e 40.000 bigramas de rótulos alto/médio/baixo.

Tokens (TF-IDF)			Bigramas de rótulos alto/médio/baixo
alto	médio	baixo	
6000	6000	3000	40000

Os resultados reportados em (Santos; Paraboni, 2024) apontam uma pequena vantagem da abordagem baseada em *prompts* (F1 médio de 0,66) sobre a arquitetura de mistura de especialistas em (Santos et al., 2023) (com F1 médio de 0,65). É possível, entretanto, que esta abordagem possa ser aperfeiçoada com uso de modelos do tipo BERT ao invés dos modelos de bigramas utilizados pelo componente supervisionado da arquitetura.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

10.5 Amive: detecção de sintomas de depressão em português

O projeto Amive[8] – um acrônimo para **Ami**go **v**irtual **e**specializado – foi desenvolvido na UFSCar[9] (Universidade Federal de São Carlos), de 2021 a 2023, por uma equipe multidisciplinar de pesquisadores e estudantes da computação e da área de saúde mental (psicologia, terapia ocupacional e psiquiatria). O projeto Amive visava combinar dados fisiológicos coletados por sensores presentes em um relógio inteligente (smartwatch) com as informações textuais das postagens de redes sociais para: (i) a identificação de um Possível Perfil Depressivo (PPD) e (ii) a oferta de intervenções personalizadas projetadas para o público-alvo de estudantes universitários (Cássia Alves et al., 2024).

Para tanto, um conjunto de sinais associados à depressão e seus fatores de risco e protetivos foram definidos pelos especialistas em saúde mental da equipe, conforme descrito na Seção 10.5.1. Focando no público-alvo de estudantes universitários, o *corpus* foi construído a partir de postagens em páginas do Facebook utilizadas por esse público, conforme descrito na Seção 10.5.2.

10.5.1 Sinais associados à depressão

Focando especificamente no uso das informações textuais para a etapa de identificação de PPD, a equipe de especialistas em saúde mental definiu um conjunto de 18 sintomas de depressão e 3 sinais adicionais (19, 20 e 21) descritos brevemente a seguir:

1. **Agitação ou inquietação** – Característica de não conseguir ficar parado fisicamente ou diante de uma situação. Percepção de que o corpo está agitado, não conseguir controlar o movimento dos membros (balançar a perna, andar de um lado para o outro, incapacidade de ficar sentado, agitar as mãos, puxar ou esfregar a pele, roupas ou outros objetos). Mente agitada e inquieta (muitos pensamentos, incômodo mental). Necessidade de fazer algo, de se mover.

2. **Alteração de peso ou apetite** – Relato de ganho ou perda de peso. Comer em excesso ou comer bem menos que o habitual. Muita ou pouca vontade de comer.

3. **Alteração de sono** – Caracterizado por não dormir de forma adequada. Dormir muito pouco. Dormir em excesso. Ter insônia. Acordar muitas vezes durante a noite. Sentir sono durante o dia. Acordar e não conseguir voltar a dormir. Cansaço e fadiga por não ter dormido.

4. **Alteração na eficiência ou funcionalidade** – Caracterizado pela diminuição na eficiência durante a realização de alguma tarefa. Perda de desempenho. Gastar mais tempo do que o habitual para realizar tarefas. Efetuar maior esforço para realizar tarefas. Falta de iniciativa. Interromper uma atividade. Desistir de realizar tarefas.

5. **Cansaço, desânimo, desencorajamento, fadiga, perda de energia ou lentificação** – Nesta categoria de sintomas juntamos diversos sintomas semelhantes, sendo que podem aparecer em conjunto na mesma sentença ou separadamente, devendo ser anotados de qualquer forma (em conjunto ou separados). Caracterizados por cansaço excessivo, falta de energia física ou mental, desgaste físico ou mental, fraqueza, falta de ânimo, falta de vontade, não ter motivação. Falta de coragem para resolver ou

[8] https://www.amive.ufscar.br/. Projeto apoiado pela FAPESP (# 20/05157-9).
[9] https://www.ufscar.br/

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

fazer algo. Perda de interesse para realizar alguma tarefa. Não ter estímulo. Perda do entusiasmo. Desejo de desistir diante de obstáculos percebidos como insuperáveis. Perda da agilidade. Demorar para realizar tarefas. Discurso, pensamento ou movimentos corporais lentificados. Maiores pausas antes de responder. Fala diminuída em termos de volume, inflexão, quantidade ou variedade de conteúdos. Não querer falar ou fazer. Preguiça. Falta de pressa, de vigor, de dinamismo.

6. **Desamparo, prejuízo social ou solidão** – Caracterizado pelo sentimento de não ter apoio e ajuda de outros. Sentimento de abandono. Necessidade de apoio. Prejuízo nas interações sociais. Isolamento. Não dar atenção a outras pessoas. Não querer ir para o trabalho, nem para escola para não ter que encontrar com pessoas, não fazer questão de participar de grupos, não querer encontrar amigos, perda do papel social. Desinteresse em estar com pessoas.

7. **Desesperança** – Caracterizado pela falta de esperança. Não enxergar perspectivas positivas seja no presente ou futuro. Não acreditar que algo pode ser bom, que o futuro pode ser melhor, que a própria pessoa pode melhorar. Sentimento de que não há saída ou que não resta mais nada.

8. **Desvalia ou baixa autoestima** – Caracterizado pelo sentimento de não ter valor. Inutilidade. Aversão a si mesmo. Sentimento de inferiorização, depreciação ou diminuição do valor pessoal. Sentimento de não ser merecedor de algo. Não reconhecer as próprias qualidades. Avaliações negativas e irrealistas do próprio valor. Sensação de fracasso. Imagem negativa sobre si mesmo. Não reconhecer a própria beleza (interior e exterior).

9. **Dificuldade para decidir** – Caracterizado pela dificuldade de tomar decisões. Incerteza. Dúvidas. Dificuldade para fazer escolhas.

10. **Deficit de atenção ou memória** – Caracterizado pela alteração da capacidade de concentração e memória. Não conseguir manter a atenção em algo. Distração excessiva. Não conseguir prestar atenção adequadamente. Perda de memória. Ter dificuldade para lembrar do que precisa fazer, do que fez. Esquecimento repentino e frequente. Capacidade diminuída para pensar.

11. **Irritação ou agressividade** – Caracterizado por sentir-se irritado, impaciente. Agressão verbal ou física. Provocação, hostilidade. Raiva persistente. Ataques de raiva.

12. **Perda ou diminuição do prazer ou da libido** – Caracterizado pela perda ou diminuição de prazer pela vida ou por atividades que antes eram agradáveis. Diminuição ou perda de satisfação pessoal ou com a vida. Diminuição ou perda do sentimento de realização. Não sentir alegria ou felicidade na realização de tarefas que antes geravam estes sentimentos. Caracterizado pela alteração no desejo sexual. Perda de vontade sexual. Diminuição do interesse sexual.

13. **Preocupação, medo ou ansiedade** – Neste marcador juntamos preocupação, medo e ansiedade, que são sintomas muito parecidos, podem aparecer em conjunto ou separadamente, mas a marcação se dará independente de estarem separados ou juntos. Podem ser caracterizados por: Preocupações exageradas e contínuas e/ou pensar excessivamente em algo e/ou pensamento fixo em algo e/ou medo em excesso

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

e/ou antecipação de algo que ainda não aconteceu e/ou angústia e/ou sentimento de que algo negativo vai acontecer e/ou avaliação negativa de um acontecimento que ainda não aconteceu e/ou catastrofização e/ou identificar perigo ou ameaça em algo que não apresenta perigo ou ameaça e/ou crises de pânico.

14. **Sentimento de culpa** – Sentimento de ser responsável por algum dano ou prejuízo causado a outra pessoa ou a si mesmo. Arrependimento por alguma atitude ou decisão tomada.

15. **Sentimento de vazio** – Caracterizado por algo que falta. Não conseguir perceber sentimento algum.

16. **Sintoma físico** – Qualquer sintoma físico apresentado. Dor de cabeça. Dor muscular. Batimentos cardíacos alterados para mais ou para menos. Respiração acelerada. Sudorese. Tremedeira. Tontura. Sensação de desmaio. Enjoo. Enxaqueca. Dormência. Falta de ar. Calor excessivo. Entre outros. Boca seca, flatulência, indigestão, diarreia, cólicas, palpitação, cefaleia, hiperventilação, suspiros. Alteração de pressão arterial.

17. **Suicídio ou auto-extermínio** – Caracterizado por ideias suicidas. Pensamentos recorrentes de morte. Planejamento da própria morte. Tentativa de suicídio. Desejo passivo de não querer acordar pela manhã. Acreditar que os outros estariam melhor se o indivíduo estivesse morto. As pessoas mais gravemente suicidas podem ter colocado seus negócios em ordem (p. ex., atualizar o testamento, pagar as dívidas), podem ter adquirido materiais necessários (p. ex., corda ou arma de fogo) e podem ter estabelecido um local e momento para consumarem o suicídio. Intenso desejo de pôr fim a um estado emocional extremamente doloroso.

18. **Tristeza ou humor depressivo** – Caracterizado pela tristeza, que pode durar mais do que o habitual. Sentir-se para baixo, desanimado, sem vontade. Negatividade, pessimismo. Melancolia. É importante que a gravidade e/ou a duração do sintoma de tristeza / humor depressivo sejam considerados na anotação.

19. **Fator de risco** – Fator relacionado ao ambiente ou a uma visão geral do mundo. Bullying. Sofrer muita pressão no trabalho. Racismo. Machismo. Pobreza extrema. Adoecimento ou comorbidade. Desemprego, problemas de moradia.

20. **Fator protetivo, cuidado em saúde e bem-estar** – Recursos que auxiliam no enfrentamento de problemas. Resiliência. Apoio de amigos. Apoio de familiares. Estar em acompanhamento ou realizar atividades que promovam o bem estar, que dão prazer ao usuário, sentimento de pertença ou sentido à vida.

21. **Morte ou suicídio de outro** – Aqui, com esta etiqueta, gostaríamos de deixar anotado quando uma pessoa está se referindo à morte de outras pessoas, estas postagens podem ser caracterizadas pelo relato de tentativas de suicídio ou o suicídio cometidos por outra pessoa, um familiar, conhecido ou pessoa pública, também devem ser anotados relatos de sentimento de luto e perda de pessoas queridas ou que a pessoa tenha apreço.

A motivação para uma investigação mais granular, e não apenas binária, de sinais de depressão em postagens textuais foi derivada do fato de que embora se caracterize por um

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

sofrimento psicológico clinicamente significativo que traz vários impactos nas funcionalidades e na qualidade de vida dos indivíduos, a depressão se manifesta de forma diferente nos indivíduos. Nesse sentido, a detecção de sintomas de depressão e fatores associados pode auxiliar os profissionais de saúde mental a oferecer a melhor forma de auxílio aos indivíduos identificados como PPD.

10.5.2 *Corpus* do projeto Amive e anotação manual

Um conjunto de postagens de páginas públicas do Facebook foi construído a partir da busca, em páginas de segredos universitários[10], por expressões e palavras-chave: "suicídio", "depressão", "me corto", "vontade de viver", "me matar" e "quero morrer". As postagens foram coletadas por meio da plataforma Crowdtangle, com data de postagem entre Janeiro de 2012 e Dezembro de 2021. 780 postagens foram selecionadas para anotação manual. Dado o conteúdo sensível, antes da anotação, um processo manual de anonimização adicional foi realizado para substituir referências a lugares, instituições, eventos, datas e cursos, por etiquetas genéricas como <CIDADE>.

As 780 postagens foram anotadas por 4 estudantes de psicologia, psiquiatria ou terapia ocupacional de gênero feminino e masculino. A tarefa de anotação consistia em selecionar trechos de texto característicos de um ou mais dos 21 sinais. Um rótulo especial também foi criado para indicar se a postagem como um todo era indicativa de um PPD: <PPD>. Assim, o *corpus* anotado resultante possui tanto a granularidade fina de instâncias para os 21 sinais de depressão quanto a anotação binária tradicionalmente usada na literatura.

A Tabela 10.2 traz a quantidade de instâncias anotadas para cada um dos 21 sinais de depressão. Vale destacar que um dos sinais mais frequentes foi o de "Suicídio ou autoextermínio" o que pode ter sido decorrente da escolha pelas expressões e palavras-chave usadas na coleta das postagens. Destaca-se também a grande quantidade de instâncias para os fatores de risco e protetivos, indicando a relevância deles no contexto de transtornos de saúde mental em universitários. Por fim, como esperado, o sintoma mais frequente é o mais característico do quadro depressivo: "Tristeza ou humor depressivo".

Tabela 10.2: Quantidade de instâncias (#) anotadas para os 21 sinais de depressão definidos no projeto Amive, em ordem crescente

Sinal	**#**
Agitação ou inquietação	5
Deficit de atenção ou memória	13
Alteração de peso ou apetite	15
Perda ou diminuição do prazer da libido	17
Sintoma físico	24
Dificuldade para decidir	26
Alteração de sono	27
Sentimento de vazio	34
Sentimento de culpa	48
Irritação ou agressividade	88
Cansaço, desânimo, desencorajamento, fadiga, perda de energia ou lentificação	106

[10]As páginas de segredos universitários publicam postagens anônimas de estudantes enviadas por meio de um formulário, também anônimo. Após o envio, as postagens ainda passam por uma moderação antes de serem publicadas.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Sinal	#
Desesperança	112
Alteração na eficiência ou funcionalidade	113
Preocupação, medo ou ansiedade	140
Desvalia ou baixa autoestima	168
Suicídio ou auto-extermínio	244
Desamparo, prejuízo social ou solidão	283
Tristeza ou humor depressivo	341
Morte ou suicídio de outro	47
Fatores protetivos	167
Fatores de risco	286

10.5.3 Identificação automática de sinais de depressão

Mendes; Caseli (2024) investigaram a identificação automática dos 21 sinais de depressão usando estratégias de aprendizado de máquina baseadas em atributos – SVM, regressão logística, naive Bayes, RandomForest e XGBoost – bem como o ajuste fino de modelos neurais – BERTimbau, mBERT, metalBERT e mentalRoBERTa. Entre os atributos investigados estão o uso de dicionários específicos (LIWC, AnewBR), categorias da PHQ-9, etiquetas de *part-of-speech* e traços morfológicos, métricas geradas pelo NILCMetrix e *embeddings* estáticas (Word2Vec, FastText, GloVe e LexVec, TF-IDF) e contextualizadas (BERTimbau).

De modo geral, as estratégias de melhor desempenho na identificação dos sinais de depressão foram as baseadas no ajuste fino dos modelos neurais, em especial o BERTimbau que foi o modelo de melhor desempenho para 8 dos 20 sinais[11]. Contudo, vale ressaltar que apenas 5 sinais obtiveram uma precisão média de 70% ou mais: Sintoma físico (70,0%), Suicídio ou auto-extermínio (71,2%), Tristeza ou humor depressivo (74,5%), Preocupação, medo ou ansiedade (79,4%) e Alteração de sono (90%).

Entre as principais conclusões de (Mendes; Caseli, 2024) tem-se o fato de que o desempenho dos modelos computacionais está mais associado à complexidade da identificação de alguns sinais do que à quantidade de instâncias disponíveis para treinamento. Por exemplo, com apenas 17 instâncias de um sinal aparentemente simples de ser expresso como "Alteração de sono" o modelo alcançou 90% de precisão enquanto que sinais que geralmente são expressos em uma linguagem mais complexa, mais abstrata e menos direta, como "Desesperança", mesmo com 5 vezes mais instâncias, foi identificado pelo melhor modelo com apenas 38% de precisão.

10.6 Considerações finais

A detecção de transtornos de saúde mental ou seus sintomas a partir de texto é uma tarefa computacional complexa e não totalmente resolvida do ponto de vista da pesquisa no PLN. Alguns **desafios**, exemplificados pelos projetos SetembroBR e Amive discutidos ao longo deste capítulo, são resumidos a seguir.

Em relação ao **quantidade e qualidade da anotação**, tem-se o alto esforço demandado pela anotação humana especializada, o que na prática pode impactar no número de

[11] O sinal de "Agitação ou inquietação" não foi avaliado nos experimentos reportados em (Mendes; Caseli, 2024) por conter apenas 5 instâncias.

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

instâncias obtidas, como no caso do projeto Amive. A sensibilidade dos dados também demanda atenção extra no processo de anotação manual, considerando **questões éticas** como a não publicização dos dados e o cuidado para não afetar a saúde mental dos próprios anotadores. Como alternativa, tem-se a anotação a partir de autorrelatos ou usando LLMs. Nesses casos, contudo, a confiabilidade das anotações pode não ser tão alta e, no caso dos LLMs, há que se considerar os **custos** computacional e financeiro envolvidos.

Há também que se considerar a presença de **vieses** como o inerente nos textos de indivíduos da classe positiva (e.g. depressivos ou ansiosos), que após serem diagnosticados podem passar a falar mais sobre o assunto. Por outro lado, os textos das redes sociais também permitem que uma grande variedade de assuntos sejam tratados nas postagens, o que acarreta em **ruídos** em relação ao problema de pesquisa em foco que para ser tratado deveria ter como base de dados apenas textos representativos de questões de saúde mental. Desse modo, é importante observar que os fenômenos de interesse, e que seriam indicativos de saúde mental, podem não estar totalmente representados na rede social. A frequência e teor das postagens de um indivíduo podem variar drasticamente a depender de múltiplos fatores externos, e que não são relacionados ao seu estado de saúde mental.

A rede social é, portanto, um **retrato naturalmente incompleto**, que talvez traga fragmentos de informação útil para a tarefa, ou talvez nem isso. Como consequência, não é razoável supor que um modelo computacional que receba como entrada apenas as postagens de um indivíduo possa obter altos níveis de acurácia na detecção de um transtorno de saúde mental ou de seus sintomas. Assim, modelos deste tipo são melhor entendidos como ferramentas de apoio ao usuário da rede social, aos seus pais ou responsáveis, ou mesmo a um profissional da área de saúde.

Considerando-se a natureza específica da atividade em redes sociais, observa-se também que modelos preditivos de saúde mental precisam lidar com **grandes volumes** de publicações, ou seja, *timelines* longas. Sequências longas deste tipo são uma limitação natural para métodos baseados em *fine-tuning* de modelos pré-treinados como BERT, e este problema é agravado pelo fato de que os dados normalmente incluem um grande número de publicações que podem ser pouco relevantes para a tarefa.

Por fim, vale mencionar um grande desafio para as pesquisas nessa temática que é a **não disponibilização pública** de dados de saúde mental. *Corpora* textuais rotulados com informações sobre saúde mental geralmente não são recursos computacionais totalmente acessíveis, como seriam *corpora* voltados a outros tipos de aplicação de PLN.

Esforços para contornar algumas destas dificuldades incluem, mais recentemente, o uso de métodos baseados em *prompt* de grandes modelos de linguagem. Entretanto, estudos iniciais deste tipo, apesar de promissores, ainda não obtiveram resultados significativamente superiores ao *fine-tuning* de modelos BERT.

Agradecimentos

Esse capítulo contou com apoio FAPESP # 2021/08213-0 e # 2020/05157-9 (Projeto Amive). O primeiro autor também agradece ao Centro de Inteligência Artificial (C4AI-USP) e ao apoio da Fundação de Apoio à Pesquisa do Estado de São Paulo (processo FAPESP # 2019/07665-4) e da IBM Corporation. A segunda autora agradece os especialistas do Comitê de Especialistas do Amive: Jair Borges Barbosa Neto, Larissa Campagna Martini, Augusto Rozendo Mendes, Tais Bleicher, Heloísa Cristina Figueiredo Frizzo, Ana Teresa Delfino, Karina Antonialli e Elaine Nabeth Louzada Torres.

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

Capítulo 11

Classificação de Áudio aplicada à Saúde

Detecção de Problemas Respiratórios

Marcelo Matheus Gauy
Larissa Cristina Berti
Flaviane Romani Fernandes Svartman
Beatriz Raposo de Medeiros
Celso Ricardo Fernandes de Carvalho
Marcelo Gomes de Queiroz
Marcelo Finger

Publicado em: 16/04/2026

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

11.1 Introdução

Classificação de áudio é um tipo de tarefa de aprendizado de máquina que consiste em atribuir uma classe a cada um dos áudios de um conjunto de dados. Um exemplo clássico de conjunto de dados para esta tarefa é o AudioSet (Gemmeke et al., 2017). O AudioSet consiste em aproximadamente 5000 horas de áudios do Youtube distribuídos em 527 classes (anotadas). Alguns exemplos de classe presentes no AudioSet são fala, respiração, tosse, ruídos de barcos e aviões, música clássica, entre outros. Modelos de aprendizado de máquina e, em particular, de aprendizado profundo, são utilizados para processar o sinal de áudio, seja ele em forma de onda ou espectrograma (Brigham; Morrow, 1967) e classificá-lo em uma das 527 classes. Trata-se, portanto, de uma tarefa de classificação, via aprendizado supervisionado, clássica.

O sucesso das técnicas de aprendizado de máquina (especialmente às de aprendizado profundo) tornou possível que pesquisadores buscassem usar essas técnicas na classificação de áudios de pessoas com problemas de saúde que possam afetar sua voz e fala (Fagherazzi et al., 2021). Dentre estes, destacam-se especialmente as condições de saúde que afetam o sistema respiratório de alguma maneira (Landry et al., 2025). Este capítulo tratará prioritariamente de modelos e técnicas de classificação de áudio para a voz e fala de pacientes com problemas respiratórios. Algumas especificidades distinguem este tipo de problema de classificação de uma tarefa clássica de classificação de áudio. Essas especificidades são dificuldades típicas no uso de aprendizado de máquina no contexto da saúde. A principal delas é a dificuldade, tanto por motivos práticos quanto por motivos éticos, de coletar dados de pacientes em hospitais. Modelos complexos de classificação necessitam, normalmente, de grande volume de dados para atingirem desempenho satisfatório. Em geral, é um grande desafio coletar áudios de pacientes no volume demandado por modelos do estado da arte de aprendizado profundo. Além do volume de dados coletados, a diversidade do conjunto de dados coletados também é um limitante importante em muitos casos. Esta diversidade vai desde possíveis comorbidades em pacientes, a outros fatores como

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

variações da doença ao longo do tempo. Além disso, o tratamento de áudios em situações reais pode levar a grandes variações nos tipos de ruído presentes, o que pode fazer com que modelos de excelente performance no conjunto de dados original falhem uma vez que ocorram alterações nestes ruídos.

O problema de lidar com o baixo volume de dados é tratado, em geral, através de técnicas clássicas de transferência de aprendizado. Modelos de classificação de áudio são pré-treinados em grandes volumes de áudio, como o AudioSet e, posteriormente, refinados para o conjunto de dados de menor volume em questão (Huang et al., 2022; Kong et al., 2020; Niizumi et al., 2022). A Seção 11.3.2.1 apresentará uma visão detalhada desta e outras técnicas para lidar com conjuntos pequenos de dados. O problema da baixa diversidade de dados presentes no conjunto de áudios de pacientes é mais complexo. No momento, não existem técnicas satisfatórias para tratar deste problema de forma geral. Isto torna a maior parte dos modelos de aprendizado profundo aplicados à saúde pouco úteis, uma vez que foram retirados de seu (restrito) contexto original (Cohen et al., 2021). Uma discussão sobre esse problema e como potencialmente lidar com ele será feito na Seção 11.4.

Descreveremos nas próximas subseções alguns problemas clássicos de saúde que afetam a voz e fala e podem ser detectados com modelos de classificação de áudio. Neste capítulo, focaremos principalmente na língua portuguesa, embora muitas das técnicas presentes aqui não sejam específicas para o português.

11.1.1 Detecção de COVID-19 através da voz e tosse

Durante a pandemia, diversas iniciativas de pesquisa (em nível internacional) foram criadas com o objetivo de detectar COVID-19 de forma automática, com aprendizado de máquina, através da voz, tosse e sons da respiração de pacientes (Brown et al., 2020a; Despotovic et al., 2021; Erdoğan; Narin, 2021; Han et al., 2021; Hassan et al., 2020; Laguarta et al., 2020; Mouawad et al., 2021; Orlandic et al., 2021; Pinkas et al., 2020). Esses trabalhos apontaram a possibilidade de detectar COVID-19 através de sinais de áudio, atingindo entre 70% e 90% de acurácia na detecção da doença. A maioria dos conjuntos de dados utilizados foram construídos por contribuição colaborativa (*'crowdsourcing'*). As técnicas de aprendizado de máquina variam desde modelos mais simples com *features* construídos manualmente até o uso de redes neurais recorrentes.

Os trabalhos citados acima têm foco principal na língua inglesa, uma vez que os áudios coletados são de falantes do inglês. Trabalhos similares para o português foram feitos no escopo do projeto SPIRA-BM (Finger et al., 2025), um projeto temático da FAPESP de número 23/00488-5 cujas informações podem ser encontradas em https://bv.fapesp.br/pt/auxilios/113158. O SPIRA-BM é uma sequência de outro projeto FAPESP (número 20/06443-5) chamado SPIRA (Aluísio et al., 2022). O SPIRA-BM, bem como o projeto SPIRA, seu predecessor, possuem uma natureza ligeiramente diferente, uma vez que o foco do SPIRA-BM está na insuficiência respiratória (sintoma típico de COVID-19), e não na detecção da COVID-19 em si. Inclusive, o projeto SPIRA-BM busca detectar não apenas a insuficiência respiratória proveniente de COVID-19, mas também a proveniente de outras fontes. A próxima subseção tratará dos resultados obtidos pelo projeto até meados de 2025 para o caso de insuficiência respiratória.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

11.1.2 Insuficiência Respiratória e seus efeitos na voz e fala

A insuficiência respiratória (IR) é um dos sintomas causados pela COVID-19 e a principal responsável pela hospitalização de pacientes de COVID-19. Por esse motivo, o projeto SPIRA reuniu médicos, fisioterapeutas, linguistas, fonoaudiólogos e cientistas da computação com o intuito de investigar a possibilidade de sua detecção automática através da voz e fala de pacientes. O projeto SPIRA-BM expandiu o escopo do projeto SPIRA original para além da IR proveniente de COVID-19, incluindo IR proveniente de outras fontes. No escopo desses projetos foram realizadas duas coletas de dados, uma na época da pandemia (*corpus* SPIRA que pode ser encontrado no Github) e outra um pouco depois das vacinas (*corpus* SPIRA 2.0 que não foi tornado público por questões de privacidade dos pacientes). Esses *corpora* consistem de sinais de áudio (voz e fala) de pacientes com IR e controles saudáveis e permitem a construção de modelos de aprendizado de máquina para a detecção da IR através do sinal de áudio. Mais detalhes sobre esses e outros *corpora* serão descritos na Seção 11.2. O projeto concluiu, através de análises dos áudios presentes, que os pacientes faziam pausas mais numerosas e mais longas que os controles, com uma distribuição, muitas vezes, em desacordo com o que seria esperado de um falante do português (Fernandes-Svartman et al., 2022). Em termos de parâmetros de voz, foi verificado que parâmetros relacionados à frequência fundamental são diferentes entre pacientes e controles (Berti et al., 2023).

Mais ainda, foram construídos diversos modelos de aprendizado profundo (Casanova et al., 2021b; Gauy et al., 2023; Gauy; Finger, 2021; Matheus Gauy et al., 2026) que buscam a detecção da IR nos *corpus* SPIRA e SPIRA 2.0. Em (Casanova et al., 2021b), os autores construíram uma rede convolucional, chamada de SpiraNet, que recebe o MFCC-grama[1]. De fato, ambos MFCC-grama e espectrograma foram testados como entrada da rede convolucional construída. O desempenho da rede ao receber MFCC-grama foi substancialmente superior. Com o MFCC-grama, a rede atingiu aproximadamente 87% de acurácia no *corpus* SPIRA. Posteriormente, a rede convolucional foi substituída por um Transformer (Vaswani et al., 2017b) que recebia MFCC-grama, chamado de MFCC-Transformer (Gauy; Finger, 2021). Esta alteração da rede levou a uma performance de aproximadamente 96% no mesmo *corpus* SPIRA. Posteriormente, o grupo SPIRA buscou usar técnicas de transferência de aprendizado que permitissem aumentar a complexidade dos modelos usados e, também, permitissem o uso do espectrograma direto, sem passar pelo MFCC. Assim, os pesquisadores usaram redes convolucionais e redes baseadas no Transformer que foram pré-treinadas no AudioSet. Esses novos modelos atingiram performance acima de 98.5% (Matheus Gauy et al., 2026), tanto no caso de IR proveniente de COVID-19 (*corpus* SPIRA), quanto no caso em que a IR provém de diversas causas potenciais (*corpus* SPIRA 2.0). Uma observação importante é que os modelos não apresentam o mesmo desempenho caso sejam retirados de seu contexto original de treinamento, com os modelos treinados no *corpus* SPIRA, não obtendo um bom desempenho no SPIRA 2.0 e vice versa (Gauy et al., 2023). Apesar destas dificuldades, as técnicas propostas pelo projeto serão de grande valor quando treinadas em *corpus* mais diversos. Como exemplo, trabalhos do projeto SPIRA foram vencedores no desafio de detecção de COVID-19 ComParE 2021 (Casanova et al., 2021a). Além disso, usando técnicas de interpretabilidade de modelos de aprendizado profundo, foi possível verificar que os modelos do projeto

[1]Em linhas gerais, o MFCC é o resultado do cálculo do espectro dos áudios, via FFT, duas vezes. Na primeira, obtemos o espectrograma. Uma segunda aplicação da FFT sobre o espectrograma, produz o MFCC-grama. Mais detalhes são apresentados na Seção 11.3.1.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

SPIRA dão atenção principalmente à voz e fala dos pacientes e não para os ruídos de fundo presentes nos áudios (Silva et al., 2024).

Considerando o sucesso obtido na detecção de IR pelo projeto, bem como o sucesso na detecção do COVID-19, foi criado o projeto temático SPIRA-BM que busca aprofundar a compreensão dos modelos de IR, potencialmente construindo modelos capazes de adicionalmente detectar a causa da IR, e também amplia o projeto para considerar outros problemas respiratórios, como asma e tabagismo. Na próxima subseção, discutiremos um pouco sobre esses problemas.

11.1.3 Outros problemas respiratórios

Inspirado no sucesso da detecção de IR, o projeto SPIRA-BM propôs estudar outros problemas respiratórios e seus efeitos na voz e na fala. Além da insuficiência respiratória, o projeto SPIRA-BM tem outras duas linhas gerais propostas. São elas: asma e tabagismo. Neste capítulo apresentamos esses problemas, o método de coleta de dados, os objetivos e seus impactos. Os resultados são parte do projeto temático SPIRA-BM (veja no link https://bv.fapesp.br/pt/auxilios/113158).

Consideremos o caso da asma: trata-se de uma doença crônica das vias aéreas, caracterizada por limitação ao fluxo aéreo expiratório que está associado a sintomas respiratórios como sibilância, dispneia, aperto no peito e tosse que variam ao longo do tempo quanto à sua frequência, ocorrência e intensidade. Esses sintomas são, frequentemente, desencadeados por fatores como exercício, exposição a alérgenos e/ou irritantes, alterações climáticas ou infecções respiratórias virais. O Brasil tem a oitava maior prevalência de asma no mundo, atingindo 12% da população brasileira. Na fase de criança e adolescência, a asma é mais prevalente em meninos enquanto, na fase adulta, tem prevalência maior entre mulheres. Em termos das tarefas de classificação de áudio que podemos considerar estão: i) a classificação entre asma não controlada e asma controlada; ii) a detecção de disfunção de prega vocal; iii) a previsão da probabilidade de exacerbação da asma dentro de um período de dois dias (estudo longitudinal). Os dois primeiros são problemas clássicos de classificação (binária) de áudio, com os dados de pacientes com asma controlada ou não controlada e com ou sem disfunção da prega vocal sendo coletados pelo projeto. A terceira tarefa é um estudo longitudinal, em que os pacientes com asma serão acompanhados ao longo do tempo e realizarão coletas de dados periódicas, com o intuito de permitir uma análise do risco de exacerbação da asma em um futuro próximo. Tal quadro de exacerbação, muitas vezes, leva à hospitalização. Estas tarefas, bem como outras análogas que possam ser consideradas, possuem o potencial, de auxiliar médicos e enfermeiros no tratamento preventivo da asma, se antecipando, por exemplo, a uma possível exacerbação dos sintomas.

No caso do tabagismo, busca-se: i) classificar o sujeito entre fumante ou não fumante; ii) estimar o nível de carga tabagística (quantidade de cigarros por dia vezes o número de anos como fumante); iii) determinar o nível de monóxido de carbono exalado pelos pacientes através do sinal de áudio. As tarefas ii) e iii) podem ser formuladas como problemas de regressão e, se necessário, quantizadas em problemas de classificação. Tarefas de regressão tendem a ser mais difíceis que tarefas de classificação, em particular, para modelos de análise de áudio. Um exemplo desta dificuldade pode ser encontrado em um trabalho do projeto SPIRA-BM que busca detectar o nível de saturação de oxigênio dos pacientes com insuficiência respiratória através do sinal de áudio (Matheus Gauy et al., 2026). Segundo Karelitz et al. (2017), a medida do monóxido de carbono no ar exalado

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

(COex) é uma avaliação simples e de baixo custo, além de ser um método não invasivo e objetivo que avalia a exposição recente a cigarro ou fumaça de cigarro. É bastante útil para detectar fumantes diários e mostra correlação com o consumo diário de cigarros. Dessa forma, estamos propondo o desenvolvimento de um método rápido e barato que auxilie os médicos e enfermeiros em programas de cessação de tabagismo.

11.1.4 Organização do capítulo

A Seção 11.2 apresentará uma descrição dos diversos *corpora* existentes atualmente, bem como suas limitações. Incluiremos os *corpora* construídos no escopo do projeto SPIRA e diversos outros *corpora*. Será feita uma divisão entre os *corpora* para COVID-19 e IR. Também serão brevemente apresentadas algumas características dos dados em processo de coleta de asma e tabagismo. A Seção 11.3 descreverá as técnicas de classificação de áudios para problemas respiratórios usados em geral. Trataremos das estratégias usuais de pré-processamento dos áudios e dos modelos de aprendizado profundo usados, descrevendo suas arquiteturas e técnicas para tratar de conjuntos pequenos de dados. A Seção 11.4 apresentará os potenciais impactos práticos do desenvolvimento de modelos de aprendizado de máquina para a detecção de problemas respiratórios e das limitações existentes nos modelos atuais e possíveis estratégias para o seu aprimoramento.

11.2 *Corpora* de áudios

Nesta seção, descreveremos os *corpora* existentes para detecção de problemas respiratórios. Ressaltamos que não temos como objetivo fazer uma descrição completa, escolhendo apenas alguns dos *corpora* principais. Para revisões mais detalhadas dos *corpora* existentes indicamos, por exemplo, (Shuja et al., 2021) (para a COVID-19) e (Kapetanidis et al., 2024; Xia et al., 2022) (para uma descrição um pouco mais geral). Ainda nesta seção, descreveremos os *corpora*: SPIRA (também chamado de primeira fase do projeto SPIRA), SPIRA 2.0 (também chamado de segunda fase do projeto SPIRA) e alguns *corpora* focados na COVID-19. Por último, comentaremos brevemente sobre *corpora* em construção no escopo do projeto SPIRA-BM para a insuficiência respiratória, asma e tabagismo.

11.2.1 *Corpus* de detecção de insuficiência respiratória

Os dois principais *corpora* já coletados no escopo do projeto SPIRA para a detecção de insuficiência respiratória são o SPIRA e o SPIRA 2.0.

11.2.1.1 *Corpus* SPIRA

Este *corpus* foi coletado pelo projeto SPIRA durante a pandemia da COVID-19 e consiste de dados de pacientes com insuficiência respiratória (causada pela COVID-19) e controles saudáveis. Os dados dos pacientes de insuficiência respiratória foram coletados em dois hospitais universitários da cidade de São Paulo. Os dados dos controles foram coletados através de um aplicativo na internet para a coleta dos áudios. A coleta consistiu em 4 áudios de cada paciente ou controle. O primeiro foi o consentimento do falante para o estudo. Usualmente, esse tipo de consentimento é feito por escrito, mas excepcionalmente na pandemia, foi permitido o consentimento por fala. O segundo áudio coletado foi uma frase elaborada por linguistas, para ser longa o bastante para conter pausas naturais, mas

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

simples o bastante para evitar que dificuldades de leitura levassem os pacientes a ter dificuldades de locução. A frase proposta foi: 'O amor ao próximo ajuda a enfrentar o coronavírus com a força que a gente precisa'. A função do segundo áudio é verificar se os pacientes com insuficiência respiratória necessitam fazer pausas adicionais, não previstas gramaticalmente em português brasileiro, para recuperar o ar, durante a fala. O terceiro áudio consiste em uma parlenda conhecida pelos falantes brasileiros como 'batatinha quando nasce...'. O objetivo desse áudio é averiguar propriedades da fala quando as pausas estão predeterminadas. O último áudio consiste em cantar 'Parabéns para você'.

Uma observação importante é que o projeto SPIRA coletou áudios de pacientes e controles de ambientes distintos. Em particular, os ruídos de fundo presentes nos áudios foram bastante diferentes e muito pronunciados no caso dos pacientes (pois foram gravados em hospitais). Para evitar que esses ruídos introduzissem vieses nos modelos de aprendizado de máquina, foram coletados alguns áudios adicionais no início de cada coleta nos hospitais. Esses áudios contêm apenas o ruído ambiente presente, sem a voz ou fala de nenhum paciente. O intuito dessa coleta adicional foi permitir que os ruídos ambientes fossem adicionados aos áudios de controle, dificultando a tarefa da rede em classificar os áudios baseada apenas no ruído presente. A Seção 11.3.1 tratará de maneiras de lidar com os diferentes tipos de ruído presentes nos áudios de pacientes e controles.

Ao todo, foram coletados mais de 6000 áudios de controle e 536 áudios de pacientes de insuficiência respiratória. Foi realizada uma triagem dos áudios de pacientes, para evitar, principalmente, que áudios contendo voz e fala dos coletores fossem mantidos no conjunto final. A triagem também foi feita, no caso dos estudos (Berti et al., 2023; Fernandes-Svartman et al., 2022) para excluir os dados em que fossem detectadas pausas ocasionadas por problemas de letramento. Como o número de controles total é muito maior que o de pacientes, também é feito um balanceamento do conjunto de dados, selecionando os controles (e pacientes) de forma a manter distribuições similares de homens e mulheres entre ambos[2]. Para a construção de modelos de aprendizado de máquina, os áudios remanescentes foram divididos em conjuntos de treino, validação e teste. A distribuição final dos dados está expressa na Tabela 11.1. O *corpus* original pode ser encontrado no Github. O conjunto de dados de controle completo (sem a triagem - mais de 6000 áudios e aproximadamente 18 horas de Português Brasileiro) pode ser encontrado no Zenodo.

Tabela 11.1: *Corpus* SPIRA

Dados		Treino	Validação	Teste
Controle	Masculino	59	8	22
	Feminino	84	8	26
	Duração média (s)	8.15	7.75	7.77
Pacientes	Masculino	83	8	28
	Feminino	66	8	32
	Duração média (s)	13.18	10.78	9.43
Áudios totais		292	32	108
Duração total (s)		3110	296	983

Fonte: (Casanova et al., 2021b)

[2]Segundo os autores, foi observado também um desbalanceamento das faixas etárias dos pacientes e controles, mas o tamanho do conjunto de dados tornou o balanceamento desse fator impraticável.

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

11.2.1.2 *Corpus* SPIRA 2.0

Este *corpus* possui uma estrutura similar ao anterior. A principal diferença é que ele foi coletado após o auge da pandemia e posterior às doses principais de vacinação. Assim, os dados de insuficiência respiratória não são dominados por pacientes de COVID-19, sendo provenientes de diversas causas. Outra diferença crucial é que tanto os dados dos controles quanto pacientes foram coletados em hospitais. Os dados foram coletados em 4 hospitais do estado de São Paulo. São eles: a Beneficência Portuguesa, o Hospital da Unimar, ambos em São Paulo, e a Santa Casa de Marília (*SC*) e CEES-Marília em Marília. Esta diferença torna o tratamento de ruído bem menos difícil que no *corpus* SPIRA. Apesar disso, os ruídos ambiente dos hospitais foram novamente coletados. Quanto aos outros áudios coletados de cada falante, novamente, o consentimento do estudo pelo falante foi em forma de áudio. A frase foi ligeiramente alterada para uma similar, que permitisse comparação posterior: 'O amor ao próximo ajuda a enfrentar *essa fase* com a força que a gente precisa'. A parlenda (com pausas predeterminadas) foi mantida. A canção 'Parabéns para você' foi substituída por uma vogal sustentada (vogal a, reproduzida por alguns segundos). A função da vogal sustentada é avaliar parâmetros de voz dos falantes, em oposição aos parâmetros de fala[3].

Uma desvantagem desse *corpus* sobre o anterior é seu tamanho. Por ter sido coletado após o auge da pandemia, a quantidade de pacientes nos hospitais que sofrem de insuficiência respiratória é bastante menor que durante a pandemia. Uma vantagem, porém, é que a insuficiência respiratória presente provém de diversas causas, não estando restrita apenas à COVID-19. Outra vantagem é que tanto pacientes e quanto controles foram coletados nos mesmos ambientes. Esse fato torna o tratamento do ruído bem mais simples.

Ao todo, foram coletados 116 áudios de controle[4] e 26 dados de pacientes com insuficiência respiratória de causas diversas. A Tabela 11.2 apresenta a distribuição dos dados de pacientes e controles por gênero usada em (Gauy et al., 2023). Novamente, existe um desbalanceamento. Como se trata de poucos dados, o balanceamento não é possível. Em (Gauy et al., 2023; Matheus Gauy et al., 2026), foi feito apenas o balanceamento dos conjuntos de validação e teste, mantendo o conjunto de treino com todos os dados remanescentes. Esse conjunto de dados não é público, como diversos outros conjuntos de dados de saúde. Observe que o *corpus* SPIRA 2.0 possui a limitação de ser excessivamente pequeno, não chegando a 20 minutos de duração total.

Tabela 11.2: *Corpus* SPIRA 2.0

Dados		
Controle	Masculino	36
	Feminino	80
	Duração média (s)	7.41
Pacientes	Masculino	14
	Feminino	12
	Duração média (s)	8.14
Áudios totais		144

[3]Aqui, temos uma distinção entre parâmetros de voz, que correspondem às características acústicas da elocução humana, e parâmetros de fala, que envolvem a produção de linguagem natural.

[4]Segundo os autores, foram removidos do conjunto de dados, aqueles pacientes que apresentaram sintomas de COVID longa, uma vez que eles poderiam enviesar os resultados (Robotti et al., 2021).

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Fonte: (Gauy et al., 2023)

11.2.1.3 Limitações atuais dos *corpus* de insuficiência respiratória

Além das limitações específicas já citadas em cada um dos *corpora*, temos limitações gerais, que estão presentes em diversos *corpora* de saúde. Estas limitações são relativas a baixa diversidade do *corpus* de áudios coletado, em parte pelo número total de áudios, em parte pela limitada distribuição espacial e temporal da coleta realizada. Os áudios coletados estão concentrados em São Paulo, e em um período bem específico de tempo, não contemplando, por exemplo, variações ao longo do tempo (ou de localização geográfica) na distribuição das doenças. Isto impede que modelos *end-to-end* de aprendizado de máquina, eficazes na prática, sejam construídos. Um exemplo desta dificuldade pode ser observada pelos dois *corpora* coletados. Modelos treinados no *corpus* SPIRA apresentam um desempenho pior que uma previsão aleatória, se testados no SPIRA 2.0 e vice-versa (Gauy et al., 2023).

11.2.2 *Corpus* de detecção de COVID-19

Durante a pandemia, diversos pesquisadores no mundo coletaram dados de voz, fala e tosse de pacientes para a detecção de COVID-19. Enquanto a maior parte da pesquisa no Brasil focou na tarefa de detecção de insuficiência respiratória (dados apresentados na Seção 11.2.1), pesquisadores no resto do mundo focaram principalmente na detecção de COVID-19 diretamente. Diversos *corpora* foram construídos por meio de coleta colaborativa (crowdsourcing) e foram apresentados em (Bhattacharya et al., 2023; Brown et al., 2020a; Despotovic et al., 2021; Orlandic et al., 2021). Embora existam problemas inerentes à coleta de dados de forma colaborativa (Garcia-Molina et al., 2016), essa é uma forma bastante eficaz de construir *corpus* de volume maior. Neste capítulo, apresentaremos dois *corpora* principais: o Coswara (Bhattacharya et al., 2023) (língua inglesa) e o ComParE (Schuller et al., 2021) (multilingual). Outros *corpora* foram construídos e podem ser encontrados nos artigos originais (Erdoğan; Narin, 2021; Han et al., 2021; Hassan et al., 2020; Laguarta et al., 2020; Mouawad et al., 2021; Pinkas et al., 2020; Watase et al., 2023).

11.2.2.1 *Corpus* Coswara

Este *corpus* foi construído durante a pandemia, entre abril de 2020 e fevereiro de 2022 (Bhattacharya et al., 2023). Ao todo, foram coletados áudios de 2635 indivíduos, dos quais 1819 são negativos para o Coronavírus, 674 são positivos para o Coronavírus e 142 contraíram e se recuperaram da COVID-19. Os áudios foram gravados através de um aplicativo online (https://coswara.iisc.ac.in/). Os participantes informaram o seu consentimento para a coleta de dados e responderam um questionário contendo informações demográficas (incluindo idade, gênero, localização, se são fumantes ou não e se a gravação foi feita com máscara), estado de saúde atual (incluindo sintomas como tosse, febre, diarréia, dificuldades de respiração, entre outros, e comorbidades como pneumonia e asma) e teste de COVID-19. Após responder o questionário, os participantes gravaram nove áudios de diferentes tipos. Dois deles tratam da respiração. Um é superficial, definido como alguns ciclos de respiração rápida sem forçar os pulmões. O outro é profundo, um tipo de respiração mais lenta que trabalha mais os pulmões. Os dois áudios para a tosse envolvem uma tosse superficial induzida, que não força o

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

pulmão, e uma tosse profunda, que envolve mais esforço pulmonar. Três tipos de vogais sustentadas foram gravados, com as vogais (como em *boot*), (como em *beet*) e (como em *bat*), sendo escolhidas. Para a fala, foi solicitado que os participantes contassem de 1 a 20, em velocidade normal (primeiro) e rápida (segundo). Assim, foram totalizados nove áudios de cada paciente. O tempo de coleta médio por participante foi 7 minutos. O total de horas gravadas foi de aproximadamente 65. Importante observar que uma fração significativa ($> 95\%$) dos participantes que reportaram teste positivo de COVID-19 foi gravada em hospitais e centros de saúde, tornando estas anotações de alta qualidade.

Os áudios coletados foram escutados por humanos para serem manualmente anotados entre qualidade excelente (nenhum ruído ambiente), moderados (pouco ruído ambiente) e ruim (ruído ambiente significativo). Aproximadamente 78% dos áudios foram considerados excelentes, com 12% considerados moderados e 10% considerados ruins. Os dados originais estão disponibilizados no Zenodo.

11.2.2.2 *Corpus* ComParE COVID-19

O *COMputational PARalinguistics challengE* de 2021 (COMPARE 2021) envolveu 4 tarefas, duas delas focadas na detecção de COVID-19. A primeira focou em dados de tosse e a segunda foi direcionada para a fala. A construção do *corpus* é similar ao Coswara e foi feita nos trabalhos (Brown et al., 2020a; Han et al., 2021). Novamente, foi usada a técnica de coleta colaborativa. Os dados foram coletados através de um aplicativo que podia ser acessado por diversas plataformas (móvel ou internet). Os participantes responderam um questionário com informações demográficas, médicas e o sintomas reportados. Para o desafio ComParE, foram incluídos apenas os áudios de tosse e fala dos participantes com a informação do teste de COVID-19. Para a primeira tarefa, focada em tosse, foram gravados 725 áudios de 343 participantes, totalizando uma hora e meia. Cada participante gravou entre uma e três tosses forçadas. Para a segunda tarefa, focada na fala, foram gravados 893 áudios de 366 participantes, totalizando pouco mais de três horas. Cada participante gravou de um a três áudios recitando a frase: *"I hope my data can help to manage the virus pandemic"* em um idioma (inglês, alemão, italiano, etc). O *corpus* ComParE não inclui áudios de português.

Para as tarefas de detecção de COVID no desafio ComParE 2021 foi proposta uma divisão entre treino, validação e teste. A Tabela 11.3 apresenta esta divisão. É importante observar que cada participante pode aparecer mais de uma vez nos conjuntos então a divisão deve ser respeitada.

Tabela 11.3: *Corpus* ComParE

Dados		Treino	Validação	Teste
Tosse	Covid positivo	71	48	39
	Covid negativo	215	183	169
	Total	286	231	208
Fala	Covid positivo	72	142	94
	Covid negativo	243	153	189
	Total	315	295	283

Fonte: (Schuller et al., 2021)

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

11.2.3 *Corpora* para detecção de outros problemas respiratórios

Trataremos brevemente de *corpora* para a detecção de problemas respiratórios como asma e tabagismo. Na literatura internacional, é comum usar o *corpus* Coswara para a detecção de asma (Looi et al., 2024). Além do Coswara, temos *corpora* privados construídos para detecção de asma (Balamurali et al., 2021; Xu et al., 2025) e infecções das vias aéreas (Balamurali et al., 2021).

No âmbito do Português brasileiro, o projeto SPIRA-BM estuda as condições de insuficiência respiratória, asma e tabagismo, como mencionado anteriormente. O *corpus* coletado para as três condições será similar ao SPIRA 2.0. Adicionalmente, são coletadas comorbidades de pacientes, bem como diversas informações demográficas, como idade, gênero. Também são coletados fatores como frequência respiratória, saturação do oxigênio no sangue (usualmente, via oxímetro). No caso de asma, os participantes respondem um questionário típico para avaliar a situação atual da doença. No caso do tabagismo, é coletado o nível de monóxido de carbono exalado. Para mais informações, veja (Finger et al., 2025).

11.3 Métodos de detecção de problemas respiratórios através de áudios

Nesta seção, apresentaremos as principais técnicas usadas na construção de modelos de aprendizado de máquina para a detecção de problemas respiratórios através do sinal de áudio (voz e fala) de pacientes. Na Seção 11.3.1, discorreremos sobre as etapas de pré-processamento necessárias para evitar a introdução de vieses no modelo (como tratamento de ruído, uniformização da duração dos áudios, divisão em treino, validação e teste) e para preparação de *features* que podem ser passados aos modelos (como o espectrograma e MFCC-grama). Na Seção 11.3.2, apresentaremos os principais modelos de classificação de áudio existentes, destacando as técnicas usadas para tratar com poucos dados (Seção 11.3.2.1) e as arquiteturas usualmente consideradas baseadas em redes convolucionais e redes do tipo Transformer (Seção 11.3.2.2).

11.3.1 Pré-processamento de áudios

Para o treinamento de modelos de aprendizado de máquina em geral, e, em particular, no caso de áudio, é essencial um pré-processamento adequado dos dados. Em termos das *features* inseridas nos modelos, a maioria dos modelos (em especial, os de aprendizado profundo) realiza um treinamento *end-to-end*, ou seja, o modelo recebe como entrada o áudio 'cru' e tem como saída a classe daquele áudio recebido. Mesmo assim, existem variações nos formatos de entrada dos áudios que são passados para os modelos. Um meio usado por alguns modelos é receber o áudio em forma de onda. Um exemplo de modelo com essa característica é o Wav2Vec e suas variantes (Baevski et al., 2020). Como alternativa à forma de onda, podemos passar ao modelo, como entrada, o espectrograma do áudio. O espectrograma é obtido através de uma transformada (rápida) de Fourier (Brigham; Morrow, 1967) que para cada janela de tempo, calcula um vetor de frequências relativas ao áudio. Ao realizar este cálculo para cada uma das janelas, obtemos uma representação do áudio no espaço de frequências, chamada de espectrograma. O espectrograma se assemelha a uma imagem (bidimensional) e os modelos de classificação de áudio que usam espectrograma como entrada são muito similares a modelos de processamento de

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

imagem (Baade et al., 2022; He et al., 2022; Huang et al., 2022; Kong et al., 2020). Adicionalmente, podemos fazer uma transformada de Fourier adicional sobre o espectrograma para obter o 'espectro' do espectrograma, chamado de MFCC-grama. O MFCC-grama é muito usado como entrada para os modelos de classificação de áudio (Gauy; Finger, 2021; Gourisaria et al., 2024; Vimal et al., 2021), apresentando muitas vezes desempenho superior ao espectrograma, em particular, nos casos em que o tamanho do *corpus* de áudios analisado seja pequeno.

Além do processamento do áudio para tornar a entrada compatível ao modelo de classificação, seja para espectrograma, MFCC-grama ou mesmo em forma de onda, devemos evitar que vieses sejam introduzidos nos modelos por características do processo de coleta de dados usado (entre outros motivos). Temos três fatores principais que introduzem viés nos áudios: ruído, duração dos áudios e distribuição demográfica. Eles serão descritos nas próximas subseções.

11.3.1.1 Formas de viés: ruído

O ruído é uma das principais fontes de viés. Em particular, como os *corpora* são usualmente pequenos e coletados em poucos locais, é bem possível que eles possuam ruído característico daqueles locais de coleta, o que provavelmente vai gerar vieses nos modelos de classificação. Um exemplo claro em que o ruído é determinante é o *corpus* SPIRA apresentado na Seção 11.2.1.1. No *corpus* SPIRA, os dados de controle e pacientes foram coletados em ambientes distintos e com ruídos claramente distintos. Assim, um modelo treinado diretamente naquele *corpus* para distinguir pacientes de controles, provavelmente, não faria mais do que classificar o tipo de ruído presente (Casanova et al., 2021b). Há duas técnicas gerais para tratar a presença de ruído. A primeira é eliminar o ruído presente nos áudios, por exemplo, através de um filtro que elimina frequências abaixo de um certo limiar. Embora esta técnica seja muito eficaz em diversos problemas, para a detecção de problemas respiratórios, ela pode resultar na remoção de parte do sinal relevante para a classificação dos áudios (como a respiração do paciente). Assim, uma técnica alternativa pode ser considerada: coletar o ruído presente nos ambientes e artificialmente acrescentá-lo nos áudios antes do modelo recebê-los. Esta foi a técnica usada no *corpus* SPIRA (Casanova et al., 2021b; Gauy; Finger, 2021). Observe que acrescentar ruído no áudios diminui a relação sinal-ruído. Assim, determinar a técnica mais adequada depende da realização de experimentos empíricos com ambas as técnicas.

11.3.1.2 Formas de viés: duração

O ruído não é a única forma de viés comumente presente em *corpus* de áudio para detecção de problemas respiratórios. Também é muito comum que vieses existam na própria duração dos áudios. Por exemplo, no caso de insuficiência respiratória, pacientes usualmente gravam áudios mais longos que controles, visto que fazem pausas (para respiração, provavelmente) mais longas e mais numerosas (Fernandes-Svartman et al., 2022). Assim, se os modelos de classificação receberem os áudios com a duração total, eles tenderão a focar excessivamente na duração dos áudios e não no conteúdo em si. Uma solução simples e eficaz para esse problema é quebrar os áudios em trechos de mesmo tamanho. Inclusive, podemos usar janelas de um tamanho fixo (por exemplo 4 segundos) com um salto menor (por exemplo, de 1 segundo), de forma que um áudio de 8 segundos se transforme em 5 áudios de 4 segundos. Este método é chamado de janelamento (Casanova et al., 2021b; Gauy et al., 2023; Gauy; Finger, 2021) e, além de prevenir o viés pela duração dos áudios,

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

é uma técnica simples e eficaz de aumento de dados. Em particular, ao lidar com conjuntos pequenos de dados, técnicas de aumentação são bastante importantes. Outra técnica de aumentação de dados para o caso de sinal de áudio é a inserção de ruído de diversos tipos nos áudios. Uma revisão da literatura sobre as muitas técnicas de aumentação de dados para problemas de classificação de áudio pode ser encontrada em (Ferreira-Paiva et al., 2022).

11.3.1.3 Formas de viés: distribuição demográfica

Além do ruído e da duração dos áudios, existe uma outra fonte bastante comum de vieses em modelos de classificação de áudio para detecção de problemas respiratórios. Essa fonte corresponde a diferentes características demográficas (por exemplo, sexo e idade) entre as diferentes classes. Por exemplo, podemos ter um *corpus* em que o número de pessoas do sexo feminino com a doença (asma, por exemplo) seja maior que o número de pessoas do sexo masculino. No conjunto de controle, porém, a relação se inverte. Nestes casos, os modelos de classificação teriam foco excessivo no sexo do paciente ao fazer a classificação. A solução para este viés é realizar uma divisão balanceada do *corpus* em conjuntos de treino, validação e teste, de forma que (pelo menos) a proporção de pacientes e controles nos diferentes grupos demográficos (especialmente gênero) seja similar em cada um dos conjuntos (de treino, validação e teste). A divisão em treino, validação e teste deve ser cuidadosa ao considerar técnicas como janelamento, por exemplo. Deve-se evitar, a todo custo, que áudios do mesmo participante caiam em conjuntos diferentes. Em geral, o conjunto de teste deve ser grande o bastante para permitir que os resultados medidos a partir dele sejam representativos, e o mais independente possível dos outros conjuntos. Sobre o conjunto de treino, existem menos restrições, e podemos usar diversas técnicas de aumentação de dados, por exemplo, para tentar melhorar os resultados na validação e, consequentemente, no teste.

11.3.1.4 Capacidade de generalização dos modelos

Por fim, gostaríamos de ponderar que a construção de modelos de aprendizado de máquina sobre conjuntos pequenos de dados é inerentemente problemática. É muito comum que modelos apresentem desempenho excelente no conjunto de teste mas que não generalizem na prática, quando testados em dados novos coletados por outros meios. Isto tende a ocorrer independentemente do cuidado ao pré-processar os dados e é uma característica de modelos de aprendizado de máquina, em especial, dos modelos de aprendizado profundo, por se tratarem de modelos caixa preta que encontram o caminho mais fácil para aprender a tarefa para a qual foram treinados (Cohen et al., 2021). Um exemplo desse tipo de problema, para o caso de insuficiência respiratória, é o trabalho (Gauy et al., 2023). Nesse trabalho, modelos treinados no *corpus* SPIRA foram testados no *corpus* SPIRA 2.0 e vice versa. Em nenhum dos casos, os modelos foram capazes de apresentar generalizações de forma eficaz. A Seção 11.4 irá expandir esta discussão, apresentando os impactos para a sociedade dos modelos construídos e como, potencialmente, recuperar a capacidade de generalização dos modelos.

11.3.2 Modelos de aprendizado profundo para classificação de áudios

Nesta seção, apresentaremos os principais modelos de aprendizado profundo para classificação de áudio, usados no contexto de detecção de problemas respiratórios. Na Se-

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

ção 11.3.2.1, discutiremos sobre os cuidados necessários ao serem considerados conjuntos pequenos de dados. Trataremos principalmente da técnica de realizar um pré-treinamento sobre o modelo em um conjunto de áudios relacionado e de grande volume, seguido de um rápido refinamento na tarefa em questão. Também discutiremos brevemente técnicas de aumentação de dados e regularização dos modelos. Na Seção 11.3.2.2, apresentaremos as arquiteturas comumente usadas pelos modelos, como as redes convolucionais e redes do tipo Transformer. Discutiremos as características das redes profundas de classificação de áudio, o tipo de entrada que recebem e a forma como são treinadas.

11.3.2.1 Tratamento de conjuntos pequenos de dados

Modelos do estado da arte de aprendizado profundo necessitam de grandes volumes de dados para seu treinamento. Quando tratamos de problemas de saúde, porém, atingir um grande volume de dados é extremamente difícil, visto que a coleta de dados é bastante custosa. Os *corpora* apresentados na Seção 11.2, em geral, não chegam a 100 horas de áudio, mesmo quando usam coleta colaborativa (*crowdsourcing*) e não chegam a 10 horas, quando coletados diretamente em hospitais. O contraste com conjuntos de dados gerais é nítido. O AudioSet, por exemplo, possui 5000 horas de áudios anotados em 527 classes. O *corpus* CommonVoice (Ardila et al., 2020), possui mais de 3000 horas, apenas de inglês, e mais de 30000 horas, incluindo mais de 100 idiomas.

A discrepância observada acima, sugere o uso de técnicas de transferência de aprendizado para permitir o uso de modelos de aprendizado profundo de complexidade maior. No contexto de processamento de linguagens naturais, o conhecido modelo BERT (Devlin et al., 2019) usou de uma técnica de pré-treinamento não supervisionado chamada de *masked language modelling*, que consistia em apagar uma proporção (15%) das palavras em uma frase para (pré-)treinar um modelo de forma não supervisionada para reconstruir a parte apagada. Posteriormente, o modelo foi refinado de maneira rápida e eficaz, mesmo em conjuntos com pequeno volume de dados, para outras tarefas de forma supervisionada. Esse método foi adaptado para o contexto de imagens e áudio alguns anos depois (Baade et al., 2022; He et al., 2022; Huang et al., 2022; Liu et al., 2020a, 2020b). O espectrograma dos áudios (ou as imagens) é visto como blocos de pixels (16×16 por exemplo) e aproximadamente 70% dos blocos são apagados. O modelo é treinado para reconstruir a parte apagada. Note que, como áudios e imagens possuem muito mais informação local, é necessário apagar uma fração muito maior dos mesmos de forma a induzir os modelos a compreenderem a estrutura global (e não apenas local) dos dados. Para permitir esta alteração, é necessário usar um modelo mais complexo (baseado na rede Transformer), tanto para a codificação (*Encoder*) quanto para a decodificação (*Decoder*) (Baade et al., 2022; He et al., 2022; Huang et al., 2022). Além do pré-treinamento não supervisionado, podemos também realizar um pré-treinamento supervisionado para classificação de áudio como feito em (Kong et al., 2020). Neste trabalho, as redes foram pré-treinadas de forma supervisionada no AudioSet e depois refinadas para outras tarefas, muitas vezes com poucos áudios. Qualquer que seja o método de pré-treinamento usado, a ideia é que a transferência de aprendizado permita que modelos complexos de aprendizado profundo sejam usados mesmo em conjuntos com pequeno volume de dados.

Outro cuidado que pode ser benéfico para o tratamento de *corpora* pequenos é a aumentação dos dados. Técnicas de aumentação dos dados ajudam a evitar que modelos se especializem no conjunto de treino sem serem eficazes para fora dele (Ferreira-Paiva et al., 2022). Isto é chamado de *overfitting* e costuma ocorrer com modelos mais complexos.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

No caso de áudio, a inserção de ruído é uma das principais técnicas de aumentação dos dados. Ela pode ser feita de diversas formas, como adicionar ruído gaussiano (branco), ou adicionar ruído de fundo comuns e não relacionados à tarefa em questão. Outra técnica é qualquer forma de segmentação dos áudios em áudios menores, como o janelamento, a remoção de partes silenciosas no começo e no final ou a translação do áudio por um fator aleatório. Também podemos alterar a velocidade de reprodução do áudio, a escala melódica dos áudios e ajustar o volume. Diversos tipos de filtros podem ser aplicados, para remover frequências mais baixas ou altas por exemplo. Alternativamente, podemos mascarar algumas frequências ou trechos do áudio aleatoriamente. Uma revisão das técnicas de aumentação de dados pode ser encontrada em (Ferreira-Paiva et al., 2022).

Além das técnicas acima, existem dois outros métodos gerais para tratar de conjuntos com pequeno volume de dados. O primeiro consiste em aplicar algum tipo de regularização ao modelo. Uma técnica de regularização comum é o acréscimo na função de perda de um termo que controle o tamanho dos pesos da rede, segundo alguma norma como a euclideana. Outra técnica de regularização é o uso do *dropout* (Srivastava et al., 2014). Esta consiste em remover uma proporção aleatória dos neurônios artificiais de uma camada da rede, forçando a mesma a construir um modelo mais robusto. O segundo método, em muitos casos imprescindível, é usar redes menos complexas. Redes de maior complexidade, com maior quantidade de parâmetros, possuem uma tendência mais forte ao *overfitting*. Uma solução é diminuir a complexidade do modelo construído, reduzindo o número de parâmetros, por exemplo, e evitando assim que a rede memorize o conjunto de treino.

Existe um problema comumente presente em conjuntos com pequeno volume de dados que não possui solução até o momento. Trata-se da tendência de que a distribuição dos dados do *corpus* seja diferente da que será encontrada na prática no momento de aplicação do modelo (Cohen et al., 2021). Este problema é bastante comum para problemas de saúde, visto que a população e as doenças passam por transformações ao longo do tempo. Além disso, a distribuição das doenças é diferente em localizações diferentes e conjuntos com pequeno volume de dados normalmente foram coletados em poucos locais. Na Seção 11.4, discutiremos as implicações dessa dificuldade prática e como poderíamos tentar contorná-la.

11.3.2.2 Arquiteturas

Em termos de arquiteturas comumente usadas em modelos de aprendizado profundo para detecção de problemas respiratórios, as redes convolucionais e as redes baseadas no Transformer são as mais comuns. As redes recorrentes tiveram diminuída a sua relevância com o advento do Transformer, assim como ocorreu em outros domínios. Em geral, os áudios são passados para as redes em forma de onda, ou mais comumente, como espectrograma. Em alguns casos, é usado o MFCC-grama. Tanto espectrograma quanto MFCC-grama são entradas parecidas com imagens, de forma que os métodos para análise de áudio usados se assemelham aos métodos usados para imagens.

11.3.2.3 Redes convolucionais: *Pretrained Audio Neural Networks*

Dentre as arquiteturas que usam redes convolucionais, as *Pretrained Audio Neural Networks* ou PANNs, propostas em (Kong et al., 2020), se destacam. Mais precisamente, as PANNs são redes pré-treinadas de forma supervisionada em grandes volumes de áudios (usualmente no AudioSet) e não necessariamente tem uma arquitetura definida. No caso

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

de modelos de detecção de problemas respiratórios as PANNs baseadas em redes convolucionais CNN6, CNN10 e CNN14 foram bastante utilizadas. Além de serem pré-treinadas, elas possuem a vantagem de serem 3 modelos similares de redes convolucionais com grau crescente de complexidade. Isto permite evidenciar o efeito do baixo volume de dados presente nos *corpora* considerados. A CNN6 é uma rede convolucional com 6 camadas, enquanto a CNN10 possui 10 camadas e a CNN14 possui 14 camadas. Todas as três recebem o espectrograma dos áudios como entrada. As camadas convolucionais da CNN6 usam kernels 5×5, enquanto que a CNN10 e CNN14 usam kernels 3×3. Cada uma dessas camadas é seguida de uma camada de *Batch normalization* (Ioffe; Szegedy, 2015) e a não linearidade escolhida é a *ReLu* (Nair; Hinton, 2010). Quatro dessas camadas estão presentes na CNN6 e entre duas delas uma camada de *average pooling* (Kong et al., 2019) está presente. Na CNN10 e CNN14, dois blocos convolucionais são aplicados em sequência antes do *average pooling* ser realizado. A CNN10 possui 4 pares desses blocos, enquanto a CNN14 tem 6 pares desses blocos. No último bloco convolucional, é usada uma camada de *Global pooling* (soma do *average* e *max pooling*) no lugar do *average pooling*. Uma penúltima camada totalmente conectada, intermediária entre os blocos convolucionais e a saída sigmoide (também totalmente conectada) com as 527 classes, está presente. Para prevenir o *overfitting*, a técnica de *dropout* é aplicada após cada camada. A CNN6 tem 4.8 milhões de parâmetros, a CNN10 tem 5.2 milhões e a CNN14 tem 80.7 milhões.

11.3.2.4 Redes convolucionais: SpiraNet

A SpiraNet, proposta em (Casanova et al., 2021b), também usa camadas convolucionais em sequência e recebe espectrograma ou MFCC-grama (mais eficaz para a detecção de insuficiência respiratória). Como a SpiraNet foi treinada do zero e não recebeu nenhum tipo de pré-treinamento, seu desempenho costuma ser inferior ao das PANNs. A tendência futura da área é o uso de modelos pré-treinados, de forma a facilitar o aprendizado da rede em *corpus* de poucas horas (ou mesmo minutos) de áudio.

11.3.2.5 Redes Transformer: *Masked Autoencoder*

Diversas arquiteturas recentes eficazes, baseadas no Transformer, foram construídas. Discutiremos principalmente um tipo de arquitetura chamada de *Masked Autoencoder* (Baade et al., 2022; He et al., 2022; Huang et al., 2022; Niizumi et al., 2022). Esta arquitetura consiste em duas redes do tipo Transformer, uma para codificação e outra para decodificação. A construção da rede tem por objetivo permitir o pré-treinamento não supervisionado sobre grandes volumes de áudios (ou imagens). O codificador recebe o espectrograma dos áudios como blocos (usualmente 16×16). O codificador é alguma variação do *Vision Transformer* (Dosovitskiy et al., 2021). O decodificador é um Transformer clássico que vai reconstruir o espectrograma original, a partir da saída do codificador. O pré-treinamento consiste em apagar uma proporção alta (70% por exemplo) do espectrograma do áudio original e transmitir apenas a parte não apagada para o codificador. Como apenas uma fração menor dos dados é transmitida ao codificador, ele usualmente possui mais parâmetros que o decodificador. O decodificador deve reconstruir o espectrograma original a partir da saída. Esta arquitetura permite que o modelo aprenda a estrutura global dos áudios de forma auto-supervisionada. Este modelo pré-treinado é um dos mais eficazes classificadores de áudio existentes atualmente. Naturalmente, para as tarefas de detecção de problemas respiratórios, esta arquitetura também costuma se sobressair.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

11.3.2.6 Redes Transformer: MFCC-Transformer

O MFCC-Transformer, usado em (Gauy; Finger, 2021), também usa um Transformer como base, tratando os tokens do Transformer como blocos de 1 frame pelo número de faixas de frequência. Ele recebe o MFCC-grama o que o torna mais eficiente do que se recebesse o espectrograma. Embora MFCC-Transfomer possua versões pré-treinadas no português brasileiro (Gauy; Finger, 2023), a técnica de pré-treinamento utilizada é menos eficaz (construída antes do surgimento do *Masked Autoencoder*) e o volume de áudios do português usados no pré-treinamento também é menor.

Um caminho potencial futuro é realizar um pré-treinamento adicional sobre um modelo pré-treinado. Esse pré-treinamento adicional poderia usar dados do português brasileiro, por exemplo, para preparar o modelo para melhor compreender o idioma (caso estejamos tratando de detectar problemas respiratórios de falantes do português). A técnica de pré-treinamento adicional foi proposta em (Niizumi et al., 2023, 2025), no modelo chamado *Masked Modeling Duo* ou M2D. Resumidamente, esse modelo usa duas redes em paralelo, uma que codifica os dados apagados e outra que codifica os dados não apagados e prevê o resultado codificado pelo outro modelo. A rede é treinada para maximizar o acordo entre os dois métodos. Esta técnica permite que o modelo seja pré-treinado de forma não supervisionada em um grande volume de dados e, posteriormente, que seja realizado um pré-treinamento adicional em uma tarefa mais próxima da tarefa alvo desejada. Em geral, esse pré-treinamento adicional leva a uma melhora de performance na tarefa final. No entanto, algumas dificuldades existem ao selecionar os parâmetros para o pré-treinamento adicional.

11.3.2.7 Outras arquiteturas

Além das redes apresentadas acima, existem diversos modelos de processamento de áudio cujo enfoque é o processamento de fala. Dentre essas, se destacam os modelos do tipo Wav2Vec (Baevski et al., 2020), que recebem o áudio em forma de onda e usam tanto camadas convolucionais (para extrair a informação do áudio), como camadas no estilo do Transformer. O Wav2Vec usa um quantizador para os áudios que permite extrair componentes discretas que compõe o mesmo (se assemelham a fones, ou seja, a sons da fala). Isto torna o modelo um transcritor de fala bastante potente. Outros modelos que têm foco no processamento de fala são o HuBERT (Hsu et al., 2021) e o XLS-R (Babu et al., 2021). Um tipo de modelo recente que foca em processamento multimodal e pode se tornar relevante para as tarefas de detecção de problemas respiratórios é o OmniVec (Srivastava; Sharma, 2024).

11.3.3 Avaliação dos modelos de detecção de problemas respiratórios

Em termos de avaliação dos modelos, algumas métricas, como a acurácia, são comumente usadas. Estas são métricas que costumam ser relevantes para qualquer problema de classificação. As seguintes métricas são comuns (VP corresponde a verdadeiro positivo, VN corresponde a verdadeiro negativo, FP corresponde a falso positivo e FN a falso negativo):

$$\text{Acurácia} = \frac{VP + VN}{VP + FP + VN + FN}$$

$$Recall = \frac{VP}{VP + FN}$$

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

$$\text{Precisão} = \frac{VP}{VP + FP}$$

$$\text{Especificidade} = \frac{VN}{VN + FP}$$

$$\text{G-médio} = \sqrt{\text{Especificidade} \times \textit{Recall}}$$

$$\text{F-measure} = 2 * \frac{\text{Precisão} \times \textit{Recall}}{\text{Precisão} + \textit{Recall}}$$

$$\text{MCC} = \frac{VP \times VN - FP \times FN}{\sqrt{(VP + FP)(VP + FN)(VN + FP)(VN + FN)}}$$

Em geral, a acurácia é a métrica mais comum. Porém, muitos *corpora* possuem um desbalanceamento apresentando mais controles que pacientes. Isto torna medidas alternativas, que consideram a precisão, o *recall* e a especificidade, como o G-médio ou a *F-measure*, muito mais acuradas para a avaliação dos modelos. No caso de problemas de saúde, poderiam ser consideradas ainda variações dessas medidas que põem pesos maiores para falsos negativos do que falsos positivos por exemplo. Curvas ROC e a área abaixo dela também são medidas comumente usadas para problemas de classificação (Hanley; McNeil, 1982), e não é diferente no caso particular de problemas respiratórios.

A Tabela 11.4 apresenta um resumo geral dos resultados obtidos para os modelos de detecção de insuficiência respiratória no escopo do projeto SPIRA. A Tabela 11.4 foca nos dados em português brasileiro, não apresentando modelos treinados em dados de outros idiomas. Em termos de métricas, ela apresenta apenas a acurácia, a título de comparação entre os modelos. Outras métricas podem ser encontradas em (Matheus Gauy et al., 2026).

Tabela 11.4: Resumo de trabalhos para a detecção de insuficiência respiratória

Referência	*Corpus*	*Features*	Modelo	Acurácia	Fraqueza
(Casanova et al., 2021b)	SPIRA	MFCC	SpiraNet	87%	Varia significativamente dependendo do ruído
(Gauy; Finger, 2021)	SPIRA	MFCC	MFCC-Transformer	96.5%	Modelo não pré-treinado
(Gauy et al., 2023)	SPIRA	MFCC	pretrained MFCC-Transformers	97.4%	Técnica de auto-supervisionamento é ineficaz
(Gauy et al., 2023)	SPIRA 2.0	MFCC	pretrained MFCC-Transformers	95%	*Corpus* muito pequeno para uma análise completa
(Matheus Gauy et al., 2026)	SPIRA	Audio Spectrogram	Audio-MAE	99.9%	Dados de pré-treinamento apenas em inglês
(Matheus Gauy et al., 2026)	SPIRA 2.0	Audio Spectrogram	CNN6	98.6%	Dados de pré-treinamento apenas em inglês. *Corpus* pequeno como (Gauy et al., 2023)

Fonte: Retirada de (Matheus Gauy et al., 2026).

11.4 Considerações finais e impactos para a sociedade

Os resultados apresentados indicam a viabilidade da detecção de problemas respiratórios através da voz e fala de pacientes. Os modelos analisados, embora apresentem um grau de complexidade elevado para o tamanho do *corpus* de treinamento, se mostraram bastante eficazes nas tarefas. O uso de diversas técnicas para prevenir o *overfitting* é crucial para atingir esse objetivo, especialmente o uso de modelos pré-treinados em grandes volumes de

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

áudios. A possibilidade de identificar problemas respiratórios de forma automática, com uma simples gravação de um paciente em um celular, é de grande valia para a sociedade e uma grande ajuda para médicos, enfermeiros e outros agentes de saúde.

Naturalmente, cuidados são necessários ao usar modelos caixa preta para detecção de problemas de saúde. Uma dificuldade natural é que esses modelos, por serem caixa preta, não permitem uma interpretação fácil de seus resultados, e não fornecem justificativas para a tomada de decisão. Embora existam técnicas como o Grad-CAM (Selvaraju et al., 2017) para facilitar a interpretação desses modelos caixa preta, e muitos autores as usaram para mostrar que os modelos de saúde em muitos casos identificam fatores relevantes nos dados (Moujahid et al., 2022; Panwar et al., 2020; Silva et al., 2024; Sobahi et al., 2022), é um problema em aberto, em geral, produzir modelos de aprendizado profundo que apresentem resultados interpretáveis. Parte desse problema pode ser evitado, deixando claro que esses modelos devem ser usados para uma análise preliminar, como a triagem de pacientes e a opinião final sempre deve ser de um médico especialista na área. Mesmo assim, seria bastante valioso criar modelos mistos que apresentem, como saída, além da previsão caixa preta do modelo, diversos parâmetros de voz e fala característicos e os resultados da comparação desses parâmetros de fala e da voz de pacientes com os mesmos parâmetros da fala e da voz de pessoas saudáveis. As análises dos linguistas e fonoaudiólogos sobre os áudios de pacientes e controles para determinar fatores de voz e fala que distinguem um grupo do outro são cruciais para uma melhor compreensão do problema (Berti et al., 2023, 2025; Fernandes-Svartman et al., 2022).

Outro cuidado, ainda mais relevante no uso desses modelos, consiste no fato de que é muito difícil gerar modelos que generalizem para fora do contexto original de seu conjunto de treinamento. Em particular, quando temos *corpora* pequenos como os de saúde, esses normalmente apresentam características específicas que permitem que o modelo use fatores específicos dos dados coletados e não generalize para fora daquele conjunto. Um exemplo claro desse efeito para problemas respiratórios pode ser encontrado em (Gauy et al., 2023). Os autores demonstraram que modelos de aprendizado profundo treinados no *corpus* SPIRA (coletado durante a pandemia), apresentam desempenho péssimo se testados no *corpus* SPIRA 2.0 (coletado após a vacinação e envolvendo insuficiência respiratória proveniente de causas além da COVID-19). O mesmo ocorre se o processo inverso for feito (treino no SPIRA 2.0 e teste no SPIRA). Ressalta-se que os modelos considerados em (Gauy et al., 2023) possuem todos excelente desempenho nos conjuntos de teste extraídos do *corpus* no qual foram treinados. Uma vez retirados de seu contexto original, porém, esses modelos passam de acurácias acima de 95% para acurácias menores que a de um chute aleatório. Esta dificuldade prática de implementar modelos de aprendizado profundo para problemas de saúde é documentada de forma mais ampla em (Cohen et al., 2021).

Para o caso de detecção de insuficiência respiratória e, potencialmente, para outros problemas respiratórios, é conhecido que diversos fatores de voz e fala são bastante diferentes entre pacientes e controles. Por exemplo, a distribuição das pausas nos enunciados de pacientes é bastante diferente da distribuição das pausas nos enunciados dos controles. A duração das pausas na fala dos pacientes também é bem maior que a dos controles. Considerando que diversos desses fatores que podem ser extraídos dos áudios devem ser semelhantes não apenas na fala dos controles mas na fala de todos os falantes do português brasileiro (a distribuição das pausas, por exemplo, é prescrita pela gramática do português, na medida em que há lugares na sentença em que sua ocorrência resultaria em agramaticalidade), é um caminho promissor construir modelos que não tratem dos áudios

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

diretamente mas desses fatores específicos de voz e fala que sabemos serem relevantes e mais estáveis, independentemente da forma de coleta. Naturalmente, esse método estaria abandonando a estratégia *end-to-end* tão comum na construção de modelos de aprendizado profundo. As dificuldades de construir *corpus* de saúde de grande volume e de criar modelos de aprendizado profundo que generalizem para fora de seu contexto podem tornar esse método alternativo, ao tentar controlar o tipo de entrada que é fornecida aos modelos, uma estratégia necessária para criar modelos mais robustos na prática.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Capítulo 12

PLN no Direito

Perspectivas e Desafios com Textos Jurídicos e Legais

Maria José Bocorny Finatto
Aline Macohin

Publicado em: 26/09/2023
Atualizado em: 16/04/2026

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

12.1 Introdução

Neste capítulo, tratamos de diferentes aspectos associados ao trabalho computacional com textos produzidos na esfera do Direito. As tarefas de PLN envolvidas, em geral, são a análise textual e a representação de conteúdos por meio de diferentes técnicas, mas há várias abordagens e estudos, voltados para diferentes finalidades. O nosso objetivo é apresentar apenas algumas perspectivas e desafios no âmbito de trabalhos que exploram materiais produzidos em português, considerando somente o cenário do Direito Brasileiro. Afinal, o Direito, de país para país, tem especificidades linguísticas e culturais que repercutem muito sobre seus textos, discursos e tipo de vocabulário.

Enquanto este capítulo introduz o gênero jurídico e discute alguns desafios para o PLN, o Capítulo 31 trata da tarefa de Reconhecimento de Entidades Nomeadas nos textos legais. Por sua vez, o Capítulo 32 propõe métodos de geração automática de ementas (resumos de processos judiciais).

Iniciamos este capítulo apresentando alguns aspectos sócio-históricos do Direito Brasileiro, que acabam influenciando suas práticas de escrita e os seus conteúdos textuais. Em seguida, situamos exemplos de reconhecimento e exploração do vocabulário jurídico, dos seus modos de dizer e, especialmente, das suas terminologias. Vamos partir de dois diferentes cenários textuais: as leis e sentenças judiciais. A primeira parte de exemplos tem a ver com um trabalho que denominamos reconhecimento terminológico (**RT**). O **RT** envolve a identificação dos usos e dos significados de um vocabulário especializado ao longo de acervos textuais, os quais formam um corpus de estudo. Esse trabalho, atualmente, é baseado em fontes escritas disponíveis em formato digital e se beneficia muito das técnicas da Linguística de *Corpus* (Sardinha, 2000) e do PLN. Depois dessa parte, mais dedicada ao vocabulário e terminologias, segue um exemplo de estudo em PLN, na área conhecida como **Análise de Sentimentos** (Capítulo 16).

O território de materiais para estudo e de enfoques, em Direito, é extremamente amplo, isso se ficarmos restritos aos trabalhos que lidam com os textos jurídicos da atualidade. Ainda assim, vale mencionar que uma série de estudos históricos sobre a linguagem jurídica brasileira, tratando de seus conceitos e até preconceitos, têm sido muito úteis para uma crítica social e política sobre o Direito. Para esses estudos históricos, os processos sobre crimes no período colonial e do império, reunidos em *corpora* que se exploram com apoio

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

computacional, têm mostrado a importância de se fazer uma linha de tempo de ações e de entendimentos até os dias de hoje. No Brasil, temos já, por exemplo, diferentes pesquisas filológicas e linguísticas dedicadas a estudos de processos criminais dos séculos 17, 18 e 19.

Entretanto, para se trabalhar com textos antigos em português, há todo o processo de *normalizar* e padronizar a apresentação escrita das "palavras antigas", para então podermos fazer o seu processamento. A normalização de textos é, assim, um desafio multidisciplinar de uma nova área de estudos denominada *Humanidades Digitais* e que inclui o PLN no tratamento de acervos antigos.

Conforme nos coloca o artigo de Cameron et al. (2023), os desafios são muitos. Afinal, geralmente trabalha-se com textos em forma de arquivo provenientes de manuscritos que foram "decifrados" e transcritos. Isso significa enfrentar muitas questões associadas à variabilidade da escrita. Afinal, uma mesma palavra podia apresentar-se de vários modos, em um mesmo documento, escrito por uma mesma pessoa, como nos casos de ÁGUA/AGUA/AGOA ou UMA/HUA/HUMA. Há exemplos interessantes desses tipo de estudo histórico, no âmbito do Direito Penal e da Medicina Legal, com processos judiciais que envolveram crimes contra mulheres no Brasil do século 19. O artigo de Teixeira et al. (2022) mostra um exemplo de estudo bem interessante nesse tema do histórico dos tratamentos jurídicos para as práticas de violência contra as mulheres.

Mas, voltando à atualidade dos textos e da linguagem do Direito, veremos, mais adiante, como exemplos, alguns textos jurídicos brasileiros, buscando ilustrar suas peculiaridades. Vamos destacar: **a)** o texto do Estatuto da Criança e do Adolescente (ECA), conforme apresentado na Lei 8.069-90, promulgada em 13 de julho de 1990 e atualizada em 2021; **b)** o texto da nossa Constituição do Brasil, de 1988 (CF88); e, **c)** um conjunto de Sentenças Judiciais dos chamados "tribunais de pequenas causas", os Juizados Especiais Cíveis. Conforme pretendemos deixar claro, esses três tipos de fontes, em suas características linguísticas e textuais, podem estar associados a diferentes tarefas de PLN, desde a descrição do português até a pontos específicos de *Recuperação da Informação*, área conhecida como *Information Retrieval* (Capítulo 2).

Por isso, um outro exemplo que trazemos neste capítulo é o da análise de conteúdos em sentenças judiciais via *Análise de Sentimentos*. Este tipo de técnica pode ser muito útil para identificar, por exemplo, padrões de sentenças judiciais favoráveis ou desfavoráveis a um determinado assunto. A utilidade dessa técnica para os profissionais do Direito é grande, pois um profissional geralmente faz buscas para entender como um determinado tribunal já vem decidindo sobre um assunto específico. Nesse trabalho de pesquisa, são buscadas retornadas inúmeras sentenças e documentos. Vale destacar que existem tribunais no Brasil inteiro, que lidam diversos assuntos (trabalhistas, penais, civis, eleitorais, entre outros). Nesses órgãos são protocolados milhares de novos processos diariamente e neles existe uma base de milhões de processos já julgados, muitos já em formato digital. A criação de um método que possa filtrar, por exemplo, as causas que foram consideradas favoráveis, em um dado tema, tende a reduzir o trabalho de leitura individual de cada sentença, ajudando o profissional a buscar e encontrar a informação que precisa.

12.2 O Direito – uma moldura para a significação

Conforme já mostraram os estudos (Motta, 2021, 2022), o Direito se manifesta através da língua, pois são as palavras que emprega e os enunciados que produz que conferem e confirmam a sua existência peculiar (Maciel, 2001) como uma prática social e área de conhecimento. Assim, temos uma relação intensa entre o Direito e a língua em uso

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

pelas pessoas que nele atuam. Isto é, pelo emprego de certas palavras[1], com um sentido particular e pela forma como suas proposições e teses são enunciadas vemos todo um cenário de valores. Isso é tão importante que temos, há bastante tempo, uma área de estudos específica conhecida como *jurilinguística* (veja mais em Cornu (1990)).

Estudiosos dessa área da linguística (Montoro, 1998, p. 1998) explicam que a linguagem jurídica, sempre com destaque para escrita, compreende diversas "espécies" de práticas que se subdividem, conforme uma dada finalidade e foco. Vejamos um detalhamento dessas espécies ou modos de se apresentar conforme seus propósitos (Petri, 2017, p. 47):

1. linguagem legislativa – a linguagem dos códigos - como o Código Penal ou Civil, das normas; sua finalidade: criar o Direito;
2. linguagem judiciária, forense ou processual - é a linguagem dos processos e sentenças; sua finalidade: aplicar o Direito;
3. linguagem convencional ou contratual – é a linguagem dos contratos, por meio dos quais se criam direitos e obrigações entre as partes;
4. linguagem doutrinária – é a linguagem dos mestres, dos doutrinadores, cuja finalidade é explicar os institutos jurídicos, é ensinar o Direito;
5. linguagem cartorária ou notarial – a linguagem jurídica que tem por finalidade registrar os atos de Direito.

Assim, embora se possa pensar em uma linguagem jurídica em geral, quando lidamos com sentenças produzidas em processos judiciais, temos linguagem judiciária, forense ou processual. Quando trabalhamos com os textos de leis, decretos e portarias, temos a linguagem legislativa. Cada tipo de suporte e/ou instrumento jurídico tende a adotar usos diferenciados e um vocabulário diferenciado. E esses elementos podem ser importantes quando se trabalha com o processamento em larga ou pequena escala desses textos. Como há uma especificidade de discursos envolvida, considerar os seus elementos linguísticos e modos de dizer próprios poderá nos ajudar a desempenhar tarefas de um modo mais produtivo. Afinal, o "Direito é, por excelência, entre as que mais o sejam, a ciência da palavra. Mais precisamente: do uso dinâmico da palavra" (Xavier, 2002, p. 1).

Conforme já mencionamos no início deste capítulo, em cada país, a linguagem jurídica tende a realizar um uso particular da língua nacional. Por isso, a linguagem do Direito de um país se diferencia da de um outro – como acontece com a linguagem jurídica dos diferentes países de Língua Portuguesa. Embora haja elos comuns, o Direito brasileiro é bastante distinto do de países como Angola ou Portugal. E, mesmo os textos jurídicos, em suas diversas formas (no Brasil conhecidos como petições, recursos, decisões judiciais etc.) podem adotar nomes e modelos de apresentação diversos. Conforme a cultura jurídica e as tradições de cada país, os produtores dos textos jurídicos serão também "autorias" diferentes, conforme o que é estabelecido no ordenamento legal de cada país.

O Direito no Brasil é regido pelo sistema da *civil law*, isso significa que uma a lei escrita tem preponderância sobre a jurisprudência – que são as decisões dos juízes – lembrando que os juízes são encarregados de verificar e direcionar a aplicação das leis. No Brasil, quem produz as leis são os membros do poder legislativo, eleitos, democraticamente, pelo povo. Os textos das leis são discutidos e votados, e então aprovados para entrarem em vigor. Os membros do poder executivo, também eleitos pelo povo, devem executar as leis

[1]Veja mais sobre a delimitação de "palavras" e de unidades de processamento no Capítulo Sequência de caracteres e palavras. Neste capítulo sobre Direito e PLN, vamos acentuar a noção de palavra como uma unidade da língua escrita, situada entre dois espaços em branco, ou entre espaço em branco e sinal de pontuação.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

aprovadas. Vejamos um resumo sobre como se organiza o Direito brasileiro, atualmente, em suas hierarquias:

- **Constituição Federal**: é a lei maior do país e define os direitos e deveres dos cidadãos, além de estabelecer a organização e o funcionamento dos poderes Executivo, Legislativo e Judiciário.
- **Legislação infraconstitucional**: são as leis, decretos e normas que regulamentam assuntos específicos, como por exemplo, o Código Civil, Código Penal, Código de Defesa do Consumidor, entre outros.
- **Poder Executivo**: é composto pelo Presidente da República, Vice-Presidente e ministros. É responsável pela administração do país e pela implementação das leis.
- **Poder Legislativo**: é formado pelo Congresso Nacional (Senado Federal e Câmara dos Deputados). É responsável por criar, modificar e aprovar as leis.
- **Poder Judiciário**: é composto pelos tribunais e juízes. É responsável por aplicar a lei em casos concretos, solucionar conflitos e garantir os direitos fundamentais dos cidadãos.

Apesar de o sistema jurídico brasileiro ser o *civil law*, há grande influência da jurisprudência nas decisões judiciais, principalmente quando agrupadas pelos tribunais e transformadas em *súmulas*. A súmula é um tipo de documento que consiste em um verbete que registra a interpretação pacífica ou majoritária adotada por um Tribunal a respeito de um tema específico. Portanto, quando textos legislativos e documentos processuais tornam-se objetos do PLN, com vistas a obter conhecimento para os profissionais do Direito, será preciso compreender esses elementos e valores diferenciados. Sem isso, há o risco de "misturar alhos com bugalhos".

12.3 Entre as terminologias e as palavras no Direito do Brasil

Grosso modo, um RT equivale à identificação e à sistematização de denominações associadas a conceitos conforme utilizadas em um dado campo ou área do conhecimento. Geralmente, o RT envolve a produção de uma "lista" de nomes (termos) vinculados aos seus significados (conceitos). Além disso, junto de cada item dessa "lista", tem-se um conjunto de informações que ajudam a contextualizar e a entender o seu uso ao longo de um conjunto de documentos escritos.

Desse modo, vamos pensar nesse processo ao longo de um conjunto de documentos jurídicos - em um dado tipo - tendo em mente a situação particular do uso de suas palavras. A Terminologia descritiva, para nós um ramo da Linguística Aplicada brasileira, e os terminólogos dedicam-se a estudar – descrever e compreender - os diferentes fenômenos linguísticos da comunicação técnico-científica, o que se estende ao Direito, em seus variados cenários.

O que diferencia uma terminologia ou um "termo técnico" de uma palavra "comum" é, em primeiro plano, o seu ambiente comunicativo. E, repetindo a ideia de uma das maiores autoridades da nossa área da Terminologia descritiva internacional (Cabré, 2005), podemos dizer: uma palavra não é um termo técnico-científico, ela está nessa condição em determinados contextos, que conferem a ela um significado "especial". Esse significado ou modo de compreensão especial, chamaremos, grosso modo, de conceito.

Vejamos um exemplo, com a palavra/item CRIANÇA, muito corriqueira no nosso dia a dia. Como seu significado básico, geralmente, entendemos algo como "pessoa não adulta".

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Mas, quando empregada e "significada" em um dado ambiente comunicativo de especialidade, como é o caso do Direito brasileiro, essa palavra "comum" assume contornos semânticos diferenciados.

No contexto do nosso Estatuto da Criança e do Adolescente, documento brasileiro conhecido como ECA, que corresponde à Lei 8.069-90, atualizada em 2021, que podemos enquadrar no domínio do Direito Civil do Brasil, temos o seguinte:

> "Art. 2º Considera-se **criança**, para os efeitos desta Lei, a pessoa até doze anos de idade incompletos, e adolescente aquela entre doze e dezoito anos de idade" (BRASIL, 1990, grifo nosso).

Como se percebe, há um significado "especializado", jurídico, uma delimitação em termos de anos de idade, que se soma ao nosso entendimento mais comum de **criança**. E você deve estar se perguntando: no que isso importará ou como poderia repercutir em um trabalho computacional sobre o tema das crianças em leis e documentos em português? A resposta é: isso importa muito! Se você comparar esse conceito com os que estabelece a OMS, Organização Mundial da Saúde, a faixa etária de uma pessoa considerada como **criança** é outra, pois compreende pessoas até 19 anos de idade. Isto é, os traços/valores de uso da palavra, que adquire estatuto terminológico, são variáveis. Além disso, temos uma conceituação jurídica específica e bem particular associada a um dado termo que, à primeira vista, não pareceria ser um termo, e sim uma palavra comum.

No caso do segmento de lei acima, o ECA, podemos considerar que há uma definição específica para CRIANÇA, que se opõe à de ADOLESCENTE. Além disso, essa definição é circunscrita, isto é, ela vale apenas em um dado contexto ou "*frame* de significação". Assim, teríamos um problema, para aquelas pessoas que se interessassem pelo Direito das Crianças, seja em sistemas jurídicos específicos, como o do Brasil, ou que busquem um mapeamento sobre esse tema no âmbito do Direito Internacional. Afinal, há variação no modo como se delimita o conceito em foco e isso pode ser um problema também para o tratamento automático dessas informações, não é mesmo?

Vamos supor uma aplicação de PLN que pudesse nos ajudar a dar conta de uma busca de informações sistematizadas sobre esse tema, o Direito das Crianças, mas restrita ao cenário brasileiro. Como vimos, em Direito, temos uma definição que tende a ser circunscrita, isto é, ela vale apenas em um dado contexto, correspondendo a um valor que estabelece frente a todo um CONJUNTO DE OUTROS TERMOS E CONCEITOS com ela relacionados. Isso é o que chamamos de sistema conceitual, que tem a ver como uma rede de conceitos e de terminologias que se entrelaçam. Como vimos, o ECA está subordinado à Constituição do Brasil, e ainda podemos ter, por exemplo, leis estaduais ou municipais – ou mesmo códigos e portarias – que "valem como leis locais" sobre, por exemplo, o tratamento de crianças em estabelecimentos de Saúde em diferentes estados do Brasil. Além das normas, também pode haver interpretações jurídicas unânimes ou diversas sobre assuntos relacionados à criança e disponibilizadas em sentenças judiciais.

12.4 Um caso concreto: em pequeníssima escala

Para realizar um ensaio de um reconhecimento terminológico (RT), podemos explorar um conjunto de textos que servem de referência ou espelhamento em uma dada área de conhecimento (veja um passo a passo detalhado de um RT com a *Constituição do Brasil* em Finatto et al. (2022)). Lidando com textos jurídicos, como vimos, será importante

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

levar em conta suas naturezas e tipologias. Assim, vamos supor um RT associado, por hipótese, ao tema "Direitos das Crianças no Brasil". Esse RT poderia envolver identificar, em diferentes documentos relevantes previamente selecionados, os seguintes elementos:

a) TERMOS e seus respectivos CONCEITOS

b) TERMOS e seus respectivos FORMATOS LINGUÍSTICOS

c) TERMOS, CONCEITOS e respectivos TERMOS E CONCEITOS RELACIONADOS.

Nos itens a) e b), acima, entra em jogo uma questão muito importante: a variação terminológica. Essa variação que um TERMO pode ter tem a ver com as diferentes formas das denominações e das conceituações, dentro de uma dada especialidade ou subárea. Você poderá perguntar sobre o escopo do RT: vamos explorar esse tema nos âmbito do Direito Civil até o Direito Criminal? Ou vamos ficar apenas em um dado recorte legislativo? Essa decisão é muito importante e deve ser bem ponderada antes do trabalho começar.

Para administrar a variabilidade de termos e conceitos, sem a ideia de condená-la, pois o enfoque linguístico e conceitual em um reconhecimento terminológico (RT) é sempre descritivo, teremos, para nos socorrer, os vocabulários controlados e/ou padronizados. Esses vocabulários mostram padrões de denominações que geralmente são colocados pela autoridade de órgãos profissionais associados a uma dada especialidade. Nesses vocabulários, encontramos as "terminologias padronizadas" e também as "normas técnicas" de uma área. Assim, uma forma de denominar um respectivo conceito/significado pode ser formalmente estabelecida em um dado contexto, de modo a se garantir precisão e boa correlação com outros termos e conceitos relacionados. Essa padronização será importante especialmente em situações de trocas de conhecimento e de trocas em geral, como no caso da comunicação jurídica internacional. Ainda assim, todos os terminólogos sabem que as variabilidades dos usos da língua, apesar dos esforços normativos, sempre se apresentam e que elas devem ser descritas e administradas.

Guardadas as devidas diferenças, é semelhante o caso, por exemplo, do conceito de CRIANÇA frente ao conceito de ADOLESCENTE no nosso Estatuto da Criança e do Adolescente, o ECA. Ainda que pareça um "detalhe", as crianças não poderão ser confundidas, em um cenário legal e jurídico, com adolescentes, muito menos com pessoas adultas, salvo condições especiais definidas naquele texto, que funciona como uma moldura de significação para suas terminologias e conceituações.

O mesmo vemos nos casos dos nomes "oficiais" para algumas doenças, que inclusive correspondem a um código numérico, conhecido como CID ou Classificação Internacional de Doenças. A ideia, nesse contexto de padronização das terminologias da área da Saúde, é evitar confusões de nomes e tentar garantir que todos possam ter um mesmo entendimento – ou conceito uniforme – de um dado TERMO + CONCEITO/DESCRIÇÃO DE SEU SIGNIFICADO. Abaixo, alguns exemplos dessa padronização da CID para o termo SARAMPO e seus tipos – uma doença, no Brasil, geralmente associada a **crianças**.

- CID 10 – B05 – Sarampo
- CID 10 – B05.0 – Sarampo complicado por encefalite
- CID 10 – B05.1 – Sarampo complicado por meningite

Dada a relevância e necessidade de tratar esse assunto, alguns tribunais como o Supremo Tribunal Federal e Superior Tribunal de Justiça criaram um site denominado "Tesauro" como forma de ferramenta para controle terminológico que tem por objetivo a padronização da informação. Nesta ferramenta, o tesauro, são apresentados os termos, conceitos,

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

termos relacionados, mas também categorias, termos genéricos e termos específicos. A partir deste mapeamento do tesauro, é possível orientar que os servidores públicos redijam os documentos judiciais com uma terminologia uniforme, para auxiliar na pesquisa e recuperação da informação posteriormente. Para saber mais sobre o tema dos *tesauros e sua interface com as terminologias*, recomendamos consultar o trabalho de Vargas; Van der Lann (2011).

Em função dos diversos contextos a que os tribunais se vinculam, há uma variedade maior de termos relacionados em comparação à legislação e cada tribunal pode apresentar informações diversas nos *tesauros* para o mesmo termo. *Tesauros* são listas de assuntos, palavras-chave e de terminologias de uma dada área de conhecimento. Essas listagens,que também se estudam há bastante tempo na Terminologia descritiva do Brasil (Van der Laan, 2002), dão suporte à indexação e catalogação de documentos em bibliotecas e em diferentes acervos físicos, como também nas bases de dados. Geralmente, quem produz esses tesauros são os bibliotecários, documentalistas e cientistas da informação que lidam com a catalogação de informações técnicas e científicas. Veja este exemplo de como funciona um tesauro, quando se busca pelo item CRIANÇA no *tesauro* do Supremo Tribunal Federal. Nessa busca, temos o seguinte resultado:

- **Termo Genérico:** Menor
- **Termos relacionados:** Adulto, Castigo Físico, Direito da Criança e do Adolescente, Educação, Educação Infantil, Estatuto da Criança e do Adolescente (ECA), Guarda de Menor, Investigação de Paternidade, Poder Familiar.
- **Categoria:** DCT Direito Constitucional, ECA Estatuto da Criança e do Adolescente.

O que se informa aqui não é um conceito para CRIANÇA, mas se aponta que ele tem um correspondente ou equivalente genérico nesse âmbito. Isto é, CRIANÇA = MENOR (genérico). Em seguida, nos termos relacionados, vemos assuntos em que se inclui esse item.

Feitas essas explicações sobre peculiaridades das terminologias jurídicas, em suas diferentes circunstâncias e variabilidades de uso e de significações, um reconhecimento terminológico (RT) pode ser visto como um tipo de trabalho de mediação de comunicação, realizado por profissionais de uma área, terminólogos, linguistas, informatas, entre outros. Salienta-se, assim, a ideia de uma mediação terminológica (Conceição; Zanola, 2020).

Além disso, um RT pode ser um trabalho multidocumento, por exemplo, com diferentes tipos de leis e normas, e multitemático, quando se tenta cobrir diferentes temas ou assuntos que se entrelaçam. Pode, ainda, apontar ligações entre documentos de diferentes naturezas, extrapolando-se o reconhecimento de um dado tópico para diferentes fronteiras. Um exemplo seriam os materiais sobre temas e políticas de Saúde Pública voltadas para crianças e os documentos jurídicos que estabelecem seus direitos. Outro exemplo de trabalho seria verificar como um determinado tribunal interpreta e aplica a legislação sobre crianças, diante de alguns problemas específicos, em âmbitos regionais ou locais.

12.5 Outros casos/exemplos: Direito Ambiental

Um reconhecimento terminológico (RT) legislativo também poderia servir de apoio para um recurso didático voltado para o cidadão comum, sem formação em Direito, ou mesmo para diferentes estudantes universitários interessados na legislação ambiental do Brasil.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Nesse caso, vamos imaginar um conjunto composto, por exemplo, por 800 leis, as quais versam sobre diferentes aspectos ambientais. Vamos supor que estamos trabalhando em um RT para uso de jornalistas novatos que lidam com temas ambientais. Como explorar essas 800 leis para chegar, por exemplo, a um conjunto de seus termos e conceitos conforme sejam mais comumente empregados nessas leis? Como apresentar a informação de forma a melhor atender o nosso suposto usuário jornalista, que, sem ter formação em Direito ou Biologia, precisaria ler e entender a legislação? Bastaria perguntar ao ChatGPT?

Naturalmente, hoje, dada a larga prática de digitalização desse tipo de documento e a garantia de seu acesso a todo o cidadão, parece ser fácil encontrar e percorrer uma base de dados com leis ambientais. O Senado Federal do Brasil, por exemplo, oferece todo um banco de leis, decretos e outros documentos afins para acesso público. Basta a pessoa acessar um site determinado e salvar os documentos no seu computador. Feito isso, "bastaria" a pessoa – o jornalista novato que imaginamos – ler, calmamente e com cuidado, todas as 800 leis do nosso caso imaginário e ir fazendo um registro, em um arquivo de texto, de suas terminologias e conceituações à medida que avance com a leitura.

Outra opção seria o "nosso" jornalista consultar um dicionário especializado sobre esse tema. A constituição de um dicionário especializado também serve como um exemplo de um ponto final possível, entre outros, quando se produz um reconhecimento terminológico (RT). Dicionários terminológicos são um ponto final ou um produto possível, entre outros, quando se elabora um RT Assim, ao empreender um RT específico, podemos nos beneficiar de RTs anteriores ao nosso.

Como a legislação, em alguns aspectos, pode ser principiológica ou apenas fornecer diretrizes, ou ainda possuir conflitos de termos entre normas diversas, pode ser necessário associar um outro reconhecimento terminológico (RT) para identificar o entendimento prevalente. Uma legislação principiológica é um tipo de norma jurídica que se baseia mais em princípios do que em regras detalhadas. Em vez de definir condutas específicas de forma minuciosa, ela apresenta valores, diretrizes gerais e objetivos fundamentais que devem orientar a interpretação e aplicação do Direito. Neste caso, podemos associar a um RT Judicial, alguns *tesauros* de tribunais, como o mencionado anteriormente.

Naturalmente, além dessas fontes padronizadas, há dicionários especializados em terminologias e glossários descritivos sobre o tema do Direito Ambiental do Brasil. Um exemplo é o *Dicionário de Direito Ambiental* do Grupo Termisul da UFRGS, publicado em segunda edição em 2008. A produção desse dicionário demandou construir, desde 1994, toda uma base legislativa sobre temas do meio ambiente associada à obra, a Base Legis[2]. Essa base, que começa com o texto do Código de Águas do Brasil, de 1930[3].

Mas, voltando aos tesauros, segundo o *tesauro* do Supremo Tribunal Federal, temos as seguintes informações:

- **Termo Genérico:** Direito (Ciência Jurídica)
- **Termo Relacionado:** Lei da Biossegurança

[2]A Base Legis Termisul-UFRGS é composta de textos da Legislação Ambiental do Brasil, Alemanha, Argentina, Estados Unidos, França, Paraguai e Uruguai. Também inclui códigos brasileiros, constituições dos países anteriormente mencionados e dos demais países que empregam a língua portuguesa (Angola, Cabo Verde, Guiné Bissau, Moçambique, Portugal, São Tomé e Príncipe e Timor Leste), Atos Internacionais relativos ao meio ambiente (Agenda 21, Convenção de Estocolmo, Declaração do Rio e Protocolo de Kyoto). Todos os textos possuem uma descrição de conteúdos e pontos principais, podendo ser baixados em formato TXT.

[3]Você pode acessar todos os textos legislativos ambientais reunidos pelo grupo Termisul- UFRGS em http://www.ufrgs.br/termisul, na aba "Recursos" , "Bases textuais" e "Base legis".

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

- **Categoria:** DAM Direito Ambiental

Como termo, há poucas informações sobre Direito Ambiental, mas ao acessar a categoria "DAM Direito Ambiental", temos cerca de 200 termos relacionados. Vale destacar que os termos apresentados fazem parte do contexto ao qual o tribunal pertence e os processos judiciais que julga. Um RT legislativo pode ter mais termos que os apresentados no tesauro do Supremo Tribunal Federal, mas em contrapartida, o *tesauro* do tribunal, pode ter um nível de detalhamento maior.

Já para o Superior Tribunal de Justiça do Brasil, ao se consultar o termo/assunto **Direito Ambiental**, temos os seguintes resultados:

- **Termos relacionados:** Bioética, Meio Ambiente, Órgão público ambiental, Princípio da Precaução, Princípio da Prevenção, Princípio do In Dubio Pro Natura, Teoria do Risco Integral.

A vantagem do tesauro elaborado por alguns tribunais é que esse instrumento reduz esforços na aplicação do PLN ao Direito, uma vez que já associa os termos mais frequentes nas decisões judiciais e associa esses termos às decisões judiciais existentes. Dado que muitos dos tesauros não associam conceitos aos termos, ao trabalhar com RT de outras fontes, como o legislativo, podem ser obtidas estas informações.

12.6 Aplicação: Análise de Sentimentos em Direito: desafios e exemplos

A análise de sentimento em textos jurídicos envolve a aplicação de técnicas de PLN para determinar o tom emocional, qualificativo ou opinativo presente nos documentos legais.

Dentre diferentes possibilidades de aplicação da técnica de análise de sentimento em textos jurídicos, podemos trabalhar com sentenças judiciais. Assim, analisa-se o contexto das sentenças judiciais e identifica-se se o juiz foi favorável ou desfavorável ao pedido de cada parte. Para esse tipo de trabalho, alguns passos devem ser seguidos como: 1) Coleta de Dados; 2) Pré-Processamento do Texto; 3) Rotulação de Dados; 4) Escolha da técnica de análise de sentimento; 5) Execução da técnica e 6) Avaliação e Validação.

Na coleta de dados, é necessário escolher um repositório que possua os textos das decisões judiciais, sejam elas de forma resumida ou na íntegra. Dentre as opções públicas e gratuitas, estão os diários de justiça dos tribunais, uso de APIs (*Application Programming Interface*) públicas como o DataJud do Conselho Nacional de Justiça e decisões disponibilizadas nos sistemas de busca dos tribunais. Para automatização desta coleta, é preciso o conhecimento de técnicas de *web crawling* e *web scraping* (Macohin; Carneiro, 2020). Essas técnicas consistem na automatização do download das páginas e arquivos que possuem decisões judiciais e posterior filtragem da informação que se deseja usar), respectivamente.

Um exemplo de informação que pode ser obtida de um tribunal pode ser verificado abaixo. Neste exemplo foram suprimidas algumas informações que pudessem identificar os envolvidos. Veja que lidamos com um tipo de texto que contém *uma parte* denominada *ementa* e outra parte que é o *acórdão*, com o resultado do processo.

A partir do download desta página, o objetivo é extrair a informação do *acórdão*, último parágrafo da imagem, onde consta se foi dado ou negado provimento ao pedido do autor. Neste caso, foi negado provimento ao autor do recurso, como se verifica através do uso das palavras "negou provimento".

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Figura 12.1: Exemplo de resumo de acórdão judicial de um tribunal superior brasileiro.



PROCESSO

REsp / SC
RECURSO ESPECIAL
2019/ -0

RELATOR	ÓRGÃO JULGADOR	DATA DO JULGAMENTO	DATA DA PUBLICAÇÃO/FONTE
Ministro	T2 - SEGUNDA TURMA	/11/2022	DJe /12/2022

RELATOR PARA ACÓRDÃO

Ministro

EMENTA

ADMINISTRATIVO. AÇÃO CIVIL PÚBLICA AJUIZADA PELA UNIÃO COM VISTA A OBTER O RESSARCIMENTO POR DANO PATRIMONIAL DECORRENTE DE EXPLORAÇÃO DE MINÉRIO (BASALTO) SEM AUTORIZAÇÃO. PRETENSÃO SUJEITA À INCIDÊNCIA DA PRESCRIÇÃO QUINQUENAL. RECURSO ESPECIAL A QUE SE NEGA PROVIMENTO.
1. No caso dos autos, a União ajuizou ação civil pública com o objetivo de obter ressarcimento pela lavra ilegal de basalto.
2. O Tribunal Regional Federal manteve a sentença de improcedência do pedido, pois, "Em se tratando de ação civil pública movida pelo Poder Público em face de particular (não abrangido pelo conceito de agente público), objetivando a reparação de dano decorrente da extração ilegal de recursos minerais, deve ser aplicado o prazo prescricional quinquenal delineado na Lei da Ação Popular (Lei n. 4.717/1965)". Inconformada, a União recorrente defendendo o afastamento da prescrição.
3. O entendimento adotado pelo Tribunal Regional Federal da 4ª Região não merece reparos, pois a imprescritibilidade da pretensão de ressarcimento ao erário se aplica somente em casos excepcionais, como é o do ato doloso de improbidade administrativa; e a incidência da prescrição, como regra, consagra o princípio da segurança jurídica (e até mesmo o da ampla defesa), não sendo cabível o sacrifício de **direito** fundamental do particular como medida de compensação da ineficiência da máquina pública.
4. Recurso especial a que se nega provimento.

ACÓRDÃO

Vistos, relatados e discutidos esses autos em que são partes as acima indicadas, acordam os Ministros da SEGUNDA TURMA do Superior Tribunal de Justiça, na conformidade dos votos e das notas taquigráficas, o seguinte resultado de julgamento:
"Prosseguindo-se no julgamento, após o voto-vista do Sr. Ministro Francisco Falcão, acompanhando o Sr. Ministro-Relator, dando provimento ao recurso especial, a Turma, por maioria, **negou provimento** ao recurso especial, nos termos do voto do Sr. Ministro Og Fernandes, vencidos os Srs. Ministros Herman Benjamin e Francisco Falcão. Lavrará o acórdão o Sr. Ministro Mauro Campbell Marques."
Os Srs. Ministros Mauro Campbell Marques e Assusete Magalhães votaram com o Sr. Ministro Og Fernandes.

Em posse dos trechos das decisões judiciais que se deseja analisar, de forma automatizada, se o desfecho foi favorável ou desfavorável, pode-se iniciar a fase de pré-processamento do texto. A fase de pré-processamento pode contemplar diversas subtécnicas. Mas, para fins de exemplificação, vamos citar apenas a tokenização, a remoção de pontuações, conversão de todas as letras para minúsculas e remoção de *stop words*.

A tokenização consiste em dividir o texto em palavras ou unidades menores, chamadas de *tokens* (veja mais detalhes no Capítulo Sequência de caracteres e palavras), que são conjuntos de caracteres separados por um espaço em branco. Um critério para separação dos *tokens* pode ser o espaço entre as palavras. Já a remoção de pontuações visa eliminar pontuação e caracteres especiais que não são relevantes para a análise de sentimento.

Em seguida, a conversão para minúsculas consiste em transformar todas as palavras em minúsculas para garantir consistência nas comparações. Por fim, a remoção de *stop words* consiste na remoção de palavras que são comuns - as palavras gramaticais ou instrumentais - e não contribuem significativamente para uma análise de sentimento, como “a”, “o”, “em”, “por” etc. Vale destacar que a lista de *stop words* deve ser na mesma língua do texto analisado. Veja o Quadro 12.1 abaixo.

Quadro 12.1: Exemplo de pré-processamento de um trecho de decisão judicial.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

> **Texto Original:** "Vistos, relatados e discutidos esses autos em que são partes as acima indicadas, acordam os Ministros da SEGUNDA TURMA do Superior Tribunal de Justiça, na conformidade dos votos e das notas taquigráficas, o seguinte resultado de julgamento: 'Prosseguindo-se no julgamento, após o voto-vista do Sr. Ministro Francisco Falcão, acompanhando o Sr. Ministro-Relator, dando provimento ao recurso especial, a Turma, por maioria, negou provimento ao recurso especial, nos termos do voto do Sr. Ministro Og Fernandes, vencidos os Srs. Ministros Herman Benjamin e Francisco Falcão. Lavrará o acórdão o Sr. Ministro Mauro Campbell Marques.' Os Srs. Ministros Mauro Campbell Marques e Assusete Magalhães votaram com o Sr. Ministro Og Fernandes."
>
> **Após Pré-processamento:** "vistos relatados discutidos autos partes acima indicadas acordam ministros segunda turma superior tribunal justiça conformidade votos notas taquigráficas seguinte resultado julgamento prosseguindo julgamento após voto-vista sr ministro francisco falcão acompanhando sr ministro-relator dando provimento recurso especial turma maioria negou provimento recurso especial termos voto sr ministro og fernandes vencidos srs ministros herman benjamin francisco falcão lavrará acórdão sr ministro mauro campbell marques srs ministros mauro campbell marques assusete magalhães votaram sr ministro og fernandes"

Ainda na fase de pré-processamento, é possível aperfeiçoar a tarefa e incluir novas *stop words*, com o objetivo de limpar mais ainda o texto e facilitar futuramente a identificação das palavras positivas ou negativas na sentença. No Quadro 12.1, verifica-se que as palavras "srs", "sr", "vistos", não influenciam na interpretação da decisão judicial e podem ser removidas.

A próxima fase, rotulamento de dados, consiste em classificar manualmente algumas decisões como positivas ou negativas, para fins de validação futura se a classificação automatizada está desempenhando um bom resultado. A partir do Quadro 12.1, facilmente esta decisão seria classificada como "NEGATIVA". Outras opções de rótulo seriam "POSITIVA" e "NEUTRA". Os casos de neutro poderiam ser utilizados, por exemplo, quando o juiz decidiu parcialmente pelo provimento (aceitação do pedido).

Já a fase da escolha da técnica de análise de sentimento, consiste em selecionar qual abordagem será utilizada, se baseada em regras ou baseada em aprendizado de máquina. Na abordagem baseada em regras, é criado um conjunto de regras e heurísticas que determinam o sentimento ou polaridade com base em palavras-chave, padrões gramaticais e outras características linguísticas. Por exemplo, certas palavras negativas podem indicar um sentimento negativo. Já, na abordagem baseada em aprendizado de máquina, é treinado um modelo de aprendizado de máquina usando-se os dados linguísticos rotulados. Algoritmos como *Naïve Bayes, Support Vector Machines* (SVM) ou redes neurais podem ser usados para construir um modelo.

Para dar continuidade ao exemplo mencionado, utilizaremos a abordagem baseada em regras. Nesse caso, utilizamos um dicionário prévio com palavras positivas e negativas. Quando são usados dicionários, deve ser considerada a língua do texto. Como exemplo de dicionário de palavras positivas, negativas e neutras em português, temos o SentiLex-PT[4] (Carvalho; Silva, 2017).

Na fase de execução da técnica e a partir do texto pré-processado anteriormente, será preciso verificar se cada palavra do texto consta no dicionário como palavra positiva, nega-

[4] O SentiLex-PT está disponível em: https://b2share.eudat.eu/records/93ab120efdaa4662baec6adee8e7585f.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

tiva ou neutra. Segundo o SentiLex-PT, foi encontrado o seguinte resultado apresentado no Quadro 12.2.

Quadro 12.2: Palavras positivas e negativas segundo o SentiLex-PT.

Dicionário:
Positivas: [“acordam”, “conformidade”, “aprovar”]
Negativas: [“negou”, “vencidos”]
Sentimento:
Positivas: 3 (acordam, conformidade, aprovar)
Negativas: 2 (negou, vencidos)
Sentimento Geral: 1 (POSITIVO) (Positivas - Negativas = 3 - 2)

O sentimento geral é calculado realizando-se uma subtração do número de palavras positivas com o número de palavras negativas. Um resultado com valor positivo indica um sentimento positivo, já um resultado com valor negativo indica um sentimento negativo e um valor próximo de zero indica um sentimento neutro. Este cálculo pode ser aperfeiçoado ao dividir o número encontrado pelo total de palavras (*tokens*) existentes no texto ((palavras positivas - palavras negativas) / total de palavras). Ou seja, se há 10 palavras no texto, 3 são positivas e 0 negativas, indica uma maior “probabilidade” que o texto realmente seja positivo. Por outro lado, se há 50 palavras no texto, apenas 1 negativa e nenhuma positiva, há probabilidade de ser um falso negativo. Essa divisão pode indicar que novos aperfeiçoamentos no dicionário podem ser necessários.

Neste exemplo acima, verifica-se que o resultado não reflete a resposta correta (NEGATIVO) e ajustes devem ser feitos. O dicionário SentiLex-PT pode ser adaptado para a linguagem jurídica, uma vez que “acordam” “conformidade” e “aprovar”, não indicam necessariamente que o juiz está dando provimento (julgando como positivo) a decisão judicial, logo, devem ser desassociados do sentimento “POSITIVO”. Outro ajuste que pode ser feito também é não associar a palavra “vencidos” ao sentimento “NEGATIVO”, uma vez que é comum, quando há divergência entre o grupo de juízes votantes, aparecer a palavra “vencidos”. Outro ajuste que também pode ser feito no dicionário é associar palavras frequentemente encontradas juntas e que reflitam a intenção da decisão judicial, por exemplo “negou provimento”, “negado provimento”, “não provido o recurso”, ” não dado provimento”, entre outros.

Feitos estes ajustes, teremos o seguinte resultado apresentado no Quadro 12.3.

Quadro 12.3: Resultado após os ajustes.

Dicionário:
Positivas: [“ ”]
Negativas: [“negou provimento”]
Sentimento:
Positivas: 0 ()
Negativas: 1 (negou provimento)
Sentimento Geral: -1 (NEGATIVO) (Positivas - Negativas = 0 - 1)

Lembrando de que essa abordagem é uma simplificação e pode não capturar todas as nuances de sentimento e/ou as polaridades em textos jurídicos complexos. Principalmente quando há uma variedade de pedidos sendo julgados com decisões diferentes para cada pedido. O contexto legal específico também pode influenciar a interpretação das palavras. Portanto, ajustes e validações são sempre necessários.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Por fim, com relação à fase de avaliação e validação, se foi utilizada a abordagem baseada em regras, como a acima exemplificada, é mais simples validar com a amostra rotulada previamente e comparar a taxa de acertos e erros. Já no caso da abordagem baseada em aprendizado de máquina, é possível utilizar parâmetros estatísticos para demonstrar a precisão e desempenho do modelo.

Como mencionado anteriormente, é necessário fazer ajustes em cada fase da execução da análise de sentimentos devido às peculiaridades discursivas dos textos jurídicos que nem sempre constam nos dicionários de referência existentes. A partir das informações obtidas na última fase, de avaliação e validação, pode-se sugerir que novos aperfeiçoamentos sejam feitos nas fases anteriores, para que o algoritmo tenha um desempenho similar e até superior ao de uma atividade humana.

Dentre os desafios para aplicar este tipo de técnica em decisões judiciais, está principalmente na coleta dos dados. Os tribunais, no geral, não possuem repositórios com estas informações prontas, estruturadas, segmentadas e públicas. Isso, por si só, já dificulta iniciar qualquer trabalho de Processamento de Linguagem Natural. Apesar do problema poder ser contornado com a criação de algoritmos de *web crawling* e *web scraping*, alguns tribunais fazem uso de *captchas* que impedem o acesso automatizado e massivo às informações. Embora haja iniciativas do Conselho Nacional de Justiça, como o DataJud, para centralizar e fornecer informações estruturadas por meio de uma API, ainda não há a íntegra das decisões disponibilizadas. Entretanto, como o DataJud continua em constante evolução, é possível que, futuramente, esse recurso seja disponibilizado. A limpeza e seleção das informações contidas em uma página HTML ou em um arquivo PDF também é bastante custosa e trabalhosa. Por isso, somente a partir destes esforços se torna possível dar seguimento à aplicação, em melhores condições, da técnica de análise de sentimentos em Direito.

12.7 Considerações finais

Este capítulo tentou situar o "mundo textual" oferecido e formatado pelo Direito do Brasil. Com isso, o objetivo foi trazer alguns exemplos de trabalhos e estudos com a sua linguagem e as suas práticas de escrita. Comentamos um pouco dos estudos históricos com a linguagem empregada em processos brasileiros antigos sobre crimes contra mulheres, mencionamos que a tarefa de reconhecimento terminológico (RT) pode ser uma demanda importante da sistematização de termos e de conceitos dos Direitos da Criança e do Direito Ambiental. Por fim, trouxemos um breve exemplo de estudo de sentimentos ou de polaridades, em PLN, que serve para ajudar a detectar os tipos de decisões que os juízes brasileiros tomam em determinados tipos de processos federais.

Buscamos salientar que é importante considerar as características e elementos de diferentes tipos de documentos, sejam leis, códigos, processos ou sentenças, produzidos por diferentes instâncias jurídicas. E, visto esse panorama, caso você possa se interessar especificamente por sentenças de tribunais administrativamente menores, como os Juizados Especiais Cíveis (JECs) - conhecidos como "tribunais de pequenas causas", vale conhecer o trabalho de doutorado de Motta (2022). A autora estudou elementos da complexidade das sentenças dos JECs, quanto ao vocabulário, terminologias e sintaxe, frente aos preceitos da legislação que estabelece que tais sentenças devem ser escritas em linguagem simples, que possibilitem fácil compreensão sobre o que se decide em uma causa. Afinal, o cidadão comum recorre aos JECs geralmente sem advogados, em meio a causas de valor limitado. Além de ampla análise, Motta (2022) oferece, no seu trabalho, acesso a todo

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

um *corpus* de sentenças por ela reunido e analisado com a ferramenta NILC-METRIX [5] (Leal et al., 2023). Também os *corpora* que ela usou como contraponto para ponderar a complexidade/facilidade de linguagem dessas sentenças estão disponíveis nos anexos da sua tese.

O Direito é um mundo feito de palavras e modos de dizer, o que oferece um terreno fértil para os nossos trabalhos de análise linguístico-textual, em geral, e, em especial, para diferentes tarefas do PLN. Os resultados desses trabalhos tendem a auxiliar tanto os profissionais quanto o cidadão e a sociedade, que são os principais focos e beneficiários das ações do Direito.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

[5] http://fw.nilc.icmc.usp.br:23380/nilcmetrix

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Capítulo 13

Reconhecimento de Entidades Nomeadas no Domínio Legal

Um Panorama para a Língua Portuguesa

Ellen Souza
Hidelberg O. Albuquerque
Nádia F. F. Silva
Matheus Cerqueira
Rafael O. Nunes
André C. P. L. F. de Carvalho
Adriano L. I. Oliveira

Publicado em: 20/11/2024
Atualizado em: 16/04/2026

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

13.1 Introdução

Neste capítulo, são apresentadas as principais iniciativas na área de reconhecimento de entidades nomeadas no domínio legal, com foco na língua portuguesa. São descritos os principais *corpora*, modelos de linguagem neurais e de larga escala, além de exemplo de aplicação de entidades nomeadas na recuperação de documentos legislativos.

O Capítulo 3 apresenta conceitos relacionados à Extração de Informação (EI), que é, normalmente, dividida em diversas tarefas de interesse, com foco no tipo de informação a ser extraída do texto, entre elas, o Reconhecimento de Entidades Nomeadas (REN), a Extração de Relações (ER) e a Extração de Eventos (EE). Com relação ao REN, são apresentadas informações históricas, conceituação formal, abordagens para rotulação, métricas de avaliação, principais entidades e *corpora*, e estado da arte dos modelos de REN para a língua portuguesa.

O Capítulo 12, por sua vez, apresenta diferentes aspectos associados ao Processamento de Linguagem Natural (PLN) na esfera jurídica. As tarefas de PLN envolvidas, em geral, são a análise textual e a representação de conteúdos por meio de diferentes técnicas, mas há várias abordagens e estudos voltados para diferentes finalidades. São descritos desafios e perspectivas no âmbito de trabalhos que exploram materiais produzidos em português, considerando somente o cenário do Direito brasileiro. Também, é exemplificada uma aplicação de análise de sentimentos em Direito.

13.2 Contextualização

Nos últimos anos, houve um grande crescimento no uso de técnicas de PLN para a área jurídica, que produz uma grande quantidade de dados no formato de texto (Zhong et

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

al., 2020). Em várias atividades do meio jurídico, é necessário extrair informações desses textos. No entanto, a crescente quantidade de documentos legais, associada ao tamanho desses documentos, torna muitas vezes impraticável sua análise manual. A Inteligência Artificial (IA), em particular PLN, provê técnicas que podem automatizar a extração de conhecimento de textos. Por conta disso, PLN é cada vez mais utilizada para lidar com a grande quantidade de documentos produzidos pelas organizações legislativas e judiciais. Em muitos países, há um considerável acúmulo de casos jurídicos a serem processados e o número de documentos gerados é enorme (Kapoor et al., 2022). A Câmara dos Deputados brasileira, desde sua fundação, já processou mais de 144 mil projetos de lei e tem processado aproximadamente 30 mil projetos de lei a cada ano (Brandt, 2020). Para cada projeto, diversos documentos são produzidos e agregados, em diferentes etapas, até sua discussão e votação.

O levantamento anual do Conselho Nacional de Justiça (CNJ) revelou um aumento significativo no número de projetos de IA no Poder Judiciário Brasileiro em 2023[1], como parte do Programa Justiça 4.0. Em comparação com 2022, houve um crescimento de 17% no número de tribunais com projetos de IA. Das 140 soluções tecnológicas mapeadas, 63 já estão em uso ou aptas a serem utilizadas e 46 estão em fase final de desenvolvimento. Já em 2023, 63% dos tribunais participaram no desenvolvimento de soluções baseadas em IA, em um crescimento de 27% em relação à 2022[2]. Os principais motivadores para o uso da IA pelos tribunais incluem o aumento da produtividade, a busca por inovação, a melhoria na qualidade dos serviços e a redução dos custos. Em 2018, em parceira com a Universidade de Brasília (UnB), o Supremo Tribunal Federal (STF) desenvolveu o Victor (Hartmann Peixoto, 2020), uma IA que separa e classifica as peças processuais mais usadas nas atividades do STF e identifica os temas de repercussão geral de maior incidência. Além do projeto Victor, o STF desenvolveu a VitorIA que identifica, no acervo do Tribunal, os processos que tratam do mesmo assunto e os agrupa automaticamente, e a RAFA 2030 (Redes Artificiais Focadas na Agenda 2030), uma IA, lançada em 2022, para apoiar a classificação de processos de acordo com os Objetivos de Desenvolvimento Sustentável da Agenda 2030 da Organização das Nações Unidas[3]. Em 2026, o CNJ apresentou o Promptus[4], um repositório nacional para padronização de instruções (*prompts*) de IA no ecossistema da Plataforma Digital do Poder Judiciário, visando aumentar, entre outras, a segurança jurídica na Justiça brasileira.

Em 2019, a Câmara dos Deputados lançou o Ulysses, um conjunto institucional de iniciativas de IA com o propósito de aumentar a transparência, melhorar a relação da Câmara com os cidadãos e apoiar a atividade legislativa com análises complexas (Almeida, 2021). Inicialmente, o Ulysses analisava, classificava e distribuía os pedidos dos parlamentares entre as 22 áreas de conhecimento da Consultoria Legislativa da Câmara. Posteriormente, em parceria com pesquisadores do Instituto de Ciências Matemáticas e de Computação da Universidade de São Paulo (ICMC-USP), novos algoritmos[5] foram desenvolvidos, permitindo a busca por documentos similares (Albuquerque et al., 2024b; Souza et al., 2021a; Vitório et al., 2022, 2024), análise do posicionamento de cidadãos sobre um projeto de lei

[1] https://www.cnj.jus.br/programa-justica-4-0-divulga-resultados-de-pesquisa-sobre-ia-no-judiciario-brasileiro/

[2] <www.cnj.jus.br/wp-content/uploads/2025/09/relatorio-entregas-do-programa-justica-4-0-versao-interativa.pdf>

[3] https://portal.stf.jus.br

[4] https://www.cnj.jus.br/tecnologia-da-informacao-e-comunicacao/justica-4-0/conheca-o-conecta/promptus/

[5] https://github.com/ulysses-camara

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

em tramitação (Maia et al., 2022; Silva et al., 2021), reconhecimento de entidades relevantes para o domínio legislativo (Albuquerque et al., 2022, 2023a; Costa et al., 2022), além da construção de modelos de linguagem específicos para o domínio legislativo (Garcia et al., 2024a) e um grande *corpus*, o Ulysses Tesemõ (Siqueira et al., 2024), composto por mais de 3,5 milhões de arquivos, totalizando 30,7 GiB de texto bruto, coletados de 159 fontes que abrangem dados judiciais, legislativos, acadêmicos, notícias e outros documentos relacionados.

O domínio legal abrange uma grande variedade de textos jurídicos, incluindo legislação, jurisprudência e trabalhos acadêmicos (Maxwell; Schafer, 2008). A natureza dos documentos legais também acrescenta um novo desafio às aplicações de PLN que pode não estar presente em outros domínios, pois esses documentos são, tipicamente, muito longos, desestruturados, contém ruídos, e são escritos usando jargão e uma linguagem específica do domínio (Kapoor et al., 2022). Há iniciativas em várias tarefas de PLN e IA, como: Recuperação de Informação Legal (Chalkidis et al., 2021; Souza et al., 2021a), Sumarização de Texto (Kornilova; Eidelman, 2019), Previsão de Julgamentos (Chalkidis et al., 2019), Segmentação Semântica (Malik et al., 2021) e Reconhecimento de Entidades Nomeadas (Albuquerque et al., 2022; Alles, 2018; Araujo et al., 2018; Castro, 2019; Costa et al., 2022).

Legal Named Entity Recognition ou reconhecimento de entidades nomeadas no domínio legal (REN-Legal), tem como objetivo detectar e rotular todas as instâncias de entidades nomeadas específicas, juridicamente relevantes, dentro de textos legais (Bonifacio et al., 2020; Cabrera-Diego; Gheewala, 2023; Cardellino et al., 2017). Considerando que os relatórios jurídicos contêm um grande número de termos e terminologias complexas, identificar entidades é um aspecto crucial na organização de informações e conhecimento dentro do domínio. É uma tarefa com grande importância, pois consiste no primeiro passo da análise semântica do texto, com potencial aplicação em diversas tarefas. Contudo, estudos sobre REN-Legal para a língua portuguesa evidenciam que os modelos utilizados para este idioma enfrentam desafios não encontrados para outros idiomas, o que pode ser explicado pelo baixo volume de *corpora*, ferramentas e modelos pré-treinados desenvolvidos para a língua portuguesa (Albuquerque et al., 2023b; Castro, 2019), necessitando de um esforço maior, principalmente no desenvolvimento de recursos e abordagens, como ocorre com a língua inglesa (Pirovani, 2019).

13.3 Iniciativas de REN-Legal

A revisão sistemática apresentada por Firdaus Solihin; Makarim (2021) mapeou 4.798 pesquisas sobre Extração de Informação no domínio legal, entre janeiro de 1990 e dezembro de 2020, extraindo dados de 107 estudos selecionados. Os autores relatam que 53% das pesquisas utilizaram *corpora* compostos de decisões judiciais, 23% de leis ou regulações, 21% utilizaram outros documentos legais, e os demais 4% utilizaram *corpora* abertos pré-existentes. Os estudos mostraram que as pesquisas utilizando REN-Legal tiveram maior destaque (acima de 60 estudos), ressaltando sua importância para a EI.

Para se ter uma ideia do avanço do estado da arte, para a edição deste capítulo foi executada uma pesquisa amostral das publicações sobre REN-Legal utilizando o *Google Scholar*[6]. Foram utilizadas duas *strings* de busca sobre REN, uma sem informação de

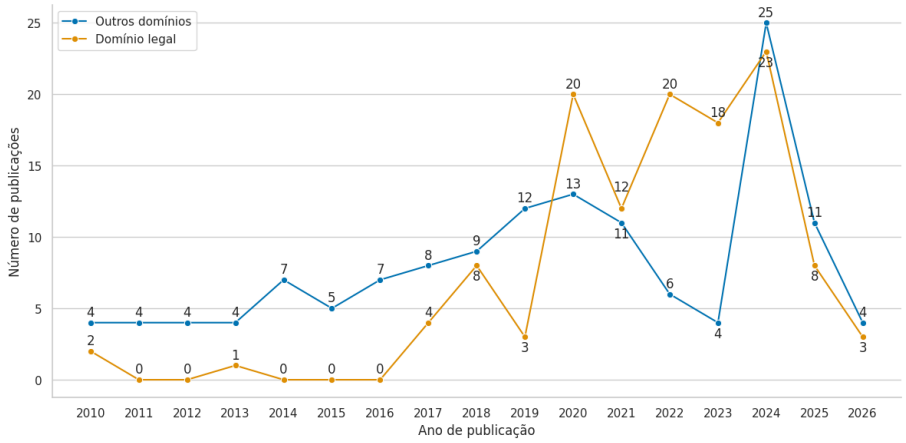
[6]A escolha pelo motor de busca foi baseada na sua praticidade, amplitude (e.g., artigos de eventos, revistas e trabalhos acadêmicos), e facilidade de replicação.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

domínio, e a outra com o domínio legal[7]. Para nossos propósitos, foram selecionados arbitrariamente os 100 primeiros artigos de cada resultado, filtrados pelo título e pelo resumo, no período de 2010 a 2026[8]. A Figura 13.1 apresenta o resultado amostral do avanço das pesquisas sobre REN-Legal, principalmente a partir de 2017, ressaltando a importância dada a este domínio especialmente entre 2020 e 2024, período em que o número de publicações em REN-Legal se manteve próxima ou superior às demais pesquisas na maior parte dos anos.

Figura 13.1: Amostra da evolução do estado da arte de REN-Legal



Além da natureza global e complexidade documental próprias do domínio, alguns fatores como diferenças culturais, idioma, linguagem, padronização de documentos, entre outros, podem ser determinantes para o sucesso das aplicações desenvolvidas. Também é preciso considerar seus subdomínios jurídicos existentes (e.g., processual, penal, trabalhista, previdenciário etc.), e suas esferas de atuação (e.g., tipos de tribunais, jurisdição etc.) (Akbik et al., 2016; Cabrera-Diego; Gheewala, 2023; Firdaus Solihin; Makarim, 2021). Neste sentido, alguns trabalhos se destacam por sua relevância para as tarefas de REN-Legal.

Dale; Mazur (2007) exploram o uso de aprendizado de máquina supervisionado para desambiguação de entidades em documentos legais, utilizando 13.460 documentos em inglês da Bolsa de Valores Australiana, aplicados na estimação de frequência de conjunções de entidades candidatas. Foram extraídos 16 tipos de entidades utilizando anotação automática, com desambiguações resolvidas manualmente. Dos nove algoritmos de aprendizagem avaliados, o classificador IBk[9] (baseado no algoritmo *K-nearest neighbors* (KNN) (Aha et al., 1991)) obteve o melhor resultado, com 84% de medida F1.

Dozier et al. (2010) utilizaram REN para identificação de informações legais em 43.936 documentos de processos legais norte-americanos. Os documentos foram anotados manualmente utilizando cinco entidades legais, classificadas através de abordagens híbridas baseadas em pesquisa (utilizando listas de entidades de interesse), regras contextuais (com regras dedutivas) e modelos estatísticos (com indicadores ponderados). O classificador *Support Vector Machine* (SVM) (Cortes; Vapnik, 1995) obteve o melhor resultado com 95% de medida F1.

[7]Dado que a maioria dos trabalhos sobre REN são escritos em inglês, foram utilizados os termos neste idioma: "*named entity recognition*" e ("*named entity recognition*"+*legal*), respectivamente.

[8]Pesquisas executadas em 24/07/2024, e atualizadas em 28/03/2026 sem o uso de filtros da ferramenta. Para 2026, os resultados representam apenas as publicações recuperadas até março deste ano.

[9]https://weka.sourceforge.io/doc.stable-3-8/weka/classifiers/lazy/IBk.html

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

Chalkidis et al. (2017) exploram a automação da extração de elementos contratuais, construindo um *corpus* de ∼3.500 contratos em inglês, manualmente anotados com 11 entidades, das quais três são do domínio legal. Utilizando regras manuais e classificadores lineares como Regressão Logística (Tolles; Meurer, 2016) e SVM, os melhores resultados foram alcançados através de um método híbrido de aprendizado de máquina e regras pós-processamento, com medida F1 acima de 84%. Posteriormente, o *corpus* desenvolvido foi utilizado para gerar vetores de palavras utilizando ∼750.000 contratos (Chalkidis; Androutsopoulos, 2017), com melhorias utilizando os modelos *Long Short-term Memory* (LSTM) (Hochreiter; Schmidhuber, 1997), *Bidirectional LSTM* (BiLSTM) (Schuster; Paliwal, 1997) e *Conditional Random Field* (CRF) (Lafferty et al., 2001). Os resultados obtidos alcançaram medida F1 de 87% na classificação por *tokens*, e 88% na identificação de elementos nos contratos.

Chalkidis et al. (2020) exploram a adaptação do modelo *Bidirectional Encoder Representations from Transformers* (BERT) (Devlin et al., 2019) para o domínio legal, através de estratégias de pré-treinamento e ajuste fino do modelo em um *corpus* de 354.621 textos legais em inglês, composto por leis, decisões de tribunais e contratos de diversas fontes. Para as tarefas de REN, foram utilizados dados de cerca de 2.000 contratos, chamado de Contracts-NER. As entidades utilizadas foram as mesmas dos trabalhos anteriores dos autores. O estudo conclui que o modelo LEGAL-BERT, especialmente uma versão menor (LEGAL-BERT-small), foi eficaz para tarefas específicas do domínio legal, apresentando desempenho comparável a modelos maiores.

Pilán et al. (2022) apresentam o *corpus* TAB (*Text Anonymization Benchmark*), desenvolvido para a anonimização de dados pessoais utilizando REN. Composto por 1.268 casos judiciais em inglês do Tribunal Europeu dos Direitos Humanos, o *corpus* foi anotado manualmente utilizando oito entidades universais. No processo de anonimização, entidades marcadas como sensíveis foram categorizadas de acordo com informações como crença filosófica, opinião política, entre outras. Juntamente com o *corpus*, a pesquisa introduz três métricas específicas de avaliação de anonimização: identificadores diretos, identificadores quase-indiretos, e identificadores diretos e quase-indiretos. Um modelo BERT ajustado ao *corpus* alcançou o resultado de 99,9% de revocação para identificadores diretos.

Aejas et al. (2023) apresentam um método para automação de contratos escritos em linguagem natural no contexto de cadeia de suprimentos, utilizando *blockchain*, REN e ER. Foram utilizados 200 contratos públicos em inglês, anotados manualmente com nove entidades do domínio. Os experimentos foram executados com modelos de *deep learning* na avaliação do *corpus* e na criação de um meta-modelo comum para as diferentes estruturas de REN e ER, utilizadas na representação de regras de negócio padronizadas. A acurácia encontrada nestes modelos foi de 99,4% e 99,6%, utilizando BERT e BiLSTM, respectivamente.

Pesquisas relevantes sobre REN-Legal também foram executadas em outras línguas, além da língua inglesa. Na língua alemã, Leitner et al. (2019) apresentam uma metodologia detalhada para REN em documentos legais, com 19 classes de entidades semânticas específicas, generalizadas em sete tipos de classes, que foram avaliados utilizando os modelos CRF e BiLSTM. Mais recentemente, Darji et al. (2023) adapta o modelo GermanBERT para o domínio legal, utilizando as mesmas entidades (Leitner et al., 2019), com resultados melhores que os modelos originais em 14 tipos de entidades específicas. No grego, Angelidis et al. (2018) combinam os modelos BiLSTM, CRF e Regressão Logística para REN em um *corpus* governamental, utilizando seis tipos de entidades. A pesquisa desenvolveu um vocabulário de referências para integrar dados legais com outras fontes na internet,

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

criando assim uma rede de informações legais acessíveis. Na língua chinesa, Chen et al. (2020) desenvolveram um sistema de extração de entidades e suas relações semânticas em documentos criminais relacionados a drogas, utilizando os modelos RoBERTa (Liu et al., 2019c), BiLSTM e a técnica Seq2Seq (Sutskever et al., 2014). A pesquisa de Çetindağ et al. (2022) apresenta o primeiro estudo sobre REN-Legal para o idioma turco. Foi utilizado um *corpus* de casos de tribunais com oito tipos de entidades (04 universais e 04 legais), os modelos BiLSTM e CRF, além de seis diferentes combinações de vetores de palavras. Na língua espanhola, Samy (2021) apresenta uma metodologia para anotação de entidades em textos legais escritos em espanhol peninsular, utilizando uma abordagem híbrida com expressões regulares, listas de fontes externas e ferramentas para REN. Cinco tipos de entidades foram utilizadas.

O REN também foi explorado em outros idiomas, como árabe (Shamma et al., 2020), finlandês (Oksanen et al., 2019a, 2022), francês (Andrew; Tannier, 2018; Barriere; Fouret, 2019; Sleimi et al., 2021), holandês (Heusden et al., 2023; Schraagen et al., 2017), italiano (Agnoloni et al., 2022; Lorè et al., 2023), romeno (Pais et al., 2021, 2024), além dos idiomas russo (Averina et al., 2022), urdu (Iftikhar et al., 2019), vietnamita (Bach et al., 2019), japonês (Onaga et al., 2023), entre outros, incluindo abordagens multilíngues (Kulkarni et al., 2023; Moreno Schneider et al., 2022; Quaresma; Gonçalves, 2010).

13.4 Iniciativas de REN-Legal em língua portuguesa

Albuquerque et al. (2023b) realizaram um mapeamento das técnicas, métodos e recursos de REN para a língua portuguesa. Foram realizadas buscas manual e automática no período de janeiro de 2010 a junho de 2022, recuperando 447 estudos primários, com 45 estudos incluídos na revisão. Os autores observaram um crescente interesse no desenvolvimento de pesquisas em REN a partir de 2019. Vinte e um estudos primários focaram na análise comparativa entre técnicas e ferramentas, tendo o modelo BiLSTM como o mais utilizado, seguido por métodos mais tradicionais que utilizam apenas CRF. Trabalhos mais recentes utilizaram o modelo BERT. Além disso, 24 estudos descreviam *corpora* de REN para vários domínios, entre eles o legal, com 8 estudos primários mapeados. A Figura 13.2 apresenta a evolução de trabalhos mapeados[10], agrupados de acordo com a variante da língua portuguesa, sendo PT-EU para textos escritos exclusivamente com a variante Europeia e PT-BR com a variante Brasileira.

13.4.1 *Corpora* para REN-Legal

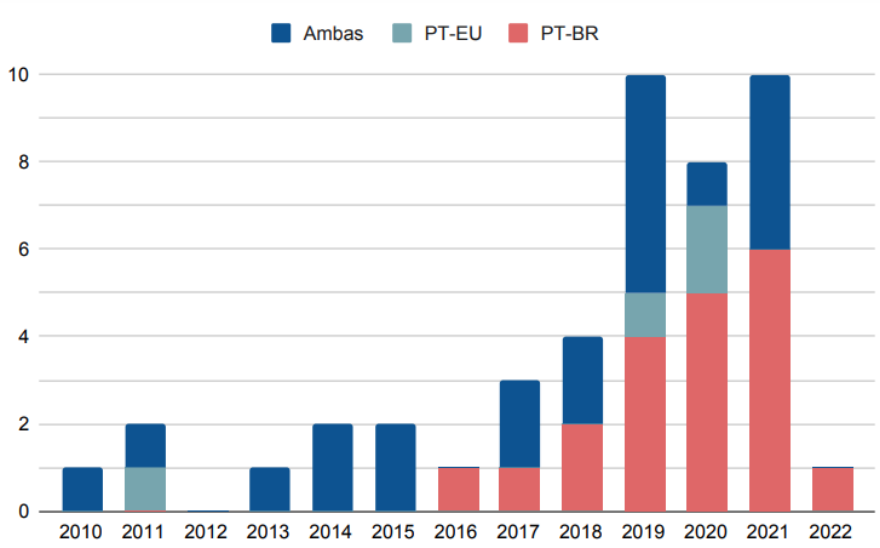
Esforços para a criação e compartilhamento de recursos, como *corpora/datasets* e modelos, vêm sendo feitos pela comunidade de PLN e EI (Firdaus Solihin; Makarim, 2021; Oksanen et al., 2019b). Estes recursos tornam-se cruciais para impulsionar novas pesquisas e o aprimoramento de pesquisas em curso (Bonifacio et al., 2020; Bordino et al., 2021). O Quadro 13.1 apresenta informações sobre alguns *corpora* para REN-Legal em língua portuguesa encontrados na literatura, destacadas a seguir.

[10]Segundo a pesquisa, o baixo número de pesquisas no ano de 2022 é justificado pelo período de execução do mapeamento, finalizado em junho de 2022.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Figura 13.2: Distribuição temporal dos artigos mapeados de REN para a língua portuguesa, entre janeiro/2010 e junho/2022 (Albuquerque et al., 2023b)



Quadro 13.1: *Corpora* REN-Legais em língua portuguesa com acesso público (PB) ou privado (PV)

Corpora/ (Autores)	Entidades	#Docs/Anotação	Arquitetura	Acesso
LeNER-Br / Araujo et al. (2018)	Legislação, Local, Jurisprudência, Organização, Pessoa, Tempo	70 / manual	BiLSTM, CRF, Glove	PB[11]
DOU-corpus / Alles (2018)	Cargo, Data, Evento, Lei, Lugar, Número, Organização, Pessoa, Processo, Valor-monetário	470 / híbrida	OpenNLP	PV
Castro (2019)	Fundamento, Função, Local, Organização, Pessoa, Tribunal, Valor_Acordo, Valor_Causa, Valor_Condenação, Valor_Custas, Vara	1.305 / híbrida	Elmo, GloVe	PV
DrugSeizures-Br / Bonifacio et al. (2020) Bonifacio (2020)	Droga, Local, Pessoa, Organização, Outro, Tempo	6.218 / automática	BERT, ELMo	PV
KauaneJunior / Fernandes et al. (2020b)	DecMorDm, IncMorDm, InitIntArr, InitMon, ValLegFee, ValMorDm	3.022 / manual	BiLSTM, BIGRU, CRF	PV
Fernandes et al. (2020a)	Local, Organização, Pessoa	63.380 / automática	CRF, L-BFGS	PB[12]
Rodríguez; Bezerra (2020)	Pessoa	10.000 / manual	NLTK	PB[13]
Aposentadoria / Neto; Faleiros (2021)	Ato, Nome_Ato, Classe, Cod_Matricula_Ato, Empresa_Ato, Fund_Legal, Cargo, Quadro, Padrão, Processo	5.515 sentenças / híbrido	CNN, LSTM, CRF, GloVe	PB[14]
Batista et al. (2021)	CPF_CNPJ, Estado_civil, Nacionalidade, Nome, OAB, RG	7.966 / híbrida	BiLSTM, CRF, Glove, Flair	PV
Mota et al. (2021)	Jurisprudência, Legislação	74 / manual	BiLSTM, CRF, Flair, Spacy	PV
Correia et al. (2022)	Citação acadêmica, Precedente, Pessoa, Referência legislativa	594 / manual	BiLSTM, CRF	PV
Lima; Borges (2022)	Aditamento Contratual, Aviso de Licitação, Aviso de Revogação/Anulação de Licitação, Aviso de Suspensão de Licitação, Extrato de Contrato, Extrato de Convênio	3.850 / manual	CRF, CNN, LSTM, Word2Vec	PB[15]

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

UlyssesNER-Br / Albuquerque et al. (2022), Costa et al. (2022)	Data, Evento, Fundamento, Local, Organização, Pessoa, Produto de Lei	950 + 1000 comentários / manual	BERT, BiLSTM, CRF, Glove	PB[16] (PL e C-corpus)
CDJUR-BR / Brito et al. (2023)	Endereço, Norma, Pena, Pessoa, Prova, Sentença	1.216 / manual	BERT, BiLSTM, CRF, Spacy	PB[17]
PersoSEG / Guimarães et al. (2024)	**Obs.: iniciam com *"Ato"*** Abono_Permanencia, Cessao, Exoneracao_Comissionado, Exoneracao_Efetivo, Nomeacao_Comissionado, Nomeacao_Efetivo, Retificacao_Comissionado, Retificacao_Efetivo, Reversao, Substituicao, Tornado_Sem_Efeito_Apo, Tornado_Sem_Efeito_Exo_Nom	127 / manual	CNN, CRF, LSTM	PB[18]
UlyssesLegalNER-Br / Albuquerque et al. (2026)	Data, Evento, Fundamento, Jurisprudência, Local, Organização, Pessoa, Produto de Lei, Valor Monetário	560 / híbrida	BERT, BiLSTM, CRF, Glove	PB[19]

O *corpus* LeNER-Br (Araujo et al., 2018) é composto por 70 documentos anotados manualmente para a tarefa de REN, sendo 66 provenientes de tribunais judiciais brasileiros, tais como: STF, Superior Tribunal de Justiça, Tribunal de Justiça de Minas Gerais e Tribunal de Contas da União. Outros quatro documentos são legislativos, como a Lei Maria da Penha. Além das entidades Pessoa, Local, Tempo e Organização, o *corpus* contém etiquetas específicas para as entidades Legislação, específica para leis, e Jurisprudência, para decisões relacionadas a casos legais. Um modelo BiLSTM-CRF (Lample et al., 2016) com arquitetura Glove (Pennington et al., 2014) foi retreinado com o *corpus*, alcançando as medidas F1 de 97,04%, 88,82% e 92,53% para entidade Legislação, para Jurisprudência e para todas as entidades, respectivamente.

O DOU-Corpus (Alles, 2018) contém entidades existentes nos documentos do Diário Oficial da União (DOU), um boletim oficial do Governo Federal brasileiro. Para o domínio, foram definidas novas entidades como Cargo, Lei, Número, Organização, Processo e Valor-monetário. A entidade Lei é definida pelos seguintes exemplos: Lei n 5548/97 e LEI 1865/89. A entidade Número descreve todo numerário explorado nos textos do DOU. A entidade Processo contém uma expressão que possui a palavra "processo" com um sequencial de números de um certa formatação. O *corpus* é composto por 60 mil trechos extraídos de diários dos meses de agosto de 2015 à outubro de 2017, anotados utilizando um processo de anotação híbrida (manual e automática). As diferentes seções que compõem o diário foram avaliadas individualmente, alcançando medida F1 de 44,50% com a ferramenta OpenNLP[20].

O *corpus* da Justiça Trabalhista (Castro, 2019) é composto por 1.305 documentos obtidos do Processo judicial eletrônico (PJ-e), distribuídos entre atas de audiência, sentenças e acórdãos, e selecionados a partir de processos distribuídos em todas as 24 regiões da justiça do trabalho brasileira, entre os anos de 2008 e 2018. Utilizando uma abordagem de

[19] https://github.com/peluz/lener-br
[19] https://journals.sagepub.com/doi/10.1177/0010414019858964#supplementary-materials
[19] https://github.com/marciamrodriguez/AtosAdministrativos_RodriguezMarcia2018
[19] https://avio11.github.io/resources/aposentadoria/aposentadoria.html
[19] https://github.com/alice7lima/PLN/tree/main/Projeto
[19] https://github.com/ulysses-camara/ulysses-ner-br
[19] https://github.com/mauriciobritojr/CDJUR-BR
[19] https://github.com/UnB-KnEDLe/persoseg-corpus
[19] https://github.com/ulysses-camara/ulysses-ner-br
[20] https://opennlp.apache.org/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

anotação híbrida, foram anotadas as entidades Função, Fundamento, Local, Organização, Pessoa, Tribunal, Vara e Valores (Acordo, Causa, Condenação e Custas). A categoria Função corresponde ao papel das pessoas mencionadas nos documentos, tais como advogado, juiz, reclamado, entre outras funções. Fundamento refere-se a qualquer dispositivo jurídico que possa ser referenciado no documento. Tribunal e Vara visam a identificação de um órgão específico. Foi treinado um modelo de REN-Legal para a Justiça do Trabalho do Brasil, alcançando a medida F1 de 93,81%, utilizando as arquiteturas ELMo (Peters et al., 2018) e GloVe.

A pesquisa desenvolvida por Fernandes et al. (2020a) analisa o impacto de discursos legislativos na captação de votos em sistemas eleitorais portugueses por localidade. Foi utilizado um *corpus* composto por 60.000 discursos parlamentares escritos em português europeu, no período de 1999 a 2015. REN-Legal foi utilizada na análise dos discursos, identificando as entidades Local, Organização e Pessoa, que serviram para medir a frequência e o contexto em que os parlamentares mencionavam seus distritos eleitorais ou outros locais relevantes. Foi utilizado o modelo CRF para identificação automática de entidades, ajustado com o algoritmo *Limited-memory Broyden-Fletcher-Goldfarb-Shanno* (L-BFGS) (Malouf, 2002), a partir de dados sequenciais anotados manualmente. Além disso, para resolver ambiguidades de entidades geográficas, foi utilizada a técnica de Resolução de Topônimos com um dicionário de localidades e regras heurísticas. Para fundamentar a hipótese da pesquisa, foi utilizada uma avaliação qualitativa baseada em entrevistas. Os resultados indicaram que o localismo nos discursos é mais presente conforme a necessidade eleitoral, independente da magnitude do distrito eleitoral, e que, quanto maior o número de candidatos por distrito, maior a quantidade de localismo presente nos discursos.

DrugSeizures-Br (Bonifacio et al., 2020) é um *corpus* contendo 6.218 petições relacionadas a procedimentos de apreensão de drogas. Estas petições foram produzidas pelo Ministério Público do Estado de Mato Grosso do Sul (MPMS) entre os anos de 2015 e 2019. O *corpus* inclui uma tabela de metadados relacionados às petições que compreendem entidades de interesse do MPMS. As entidades anotadas estão divididas em 25 tipos diferentes envolvendo, por exemplo, nome e quantidade da droga apreendida, dados da pessoa denunciada (nome, documentos etc.), local da apreensão, entre outros. Os tipos estão agrupados em 6 categorias: Local, Pessoa, Tempo, Organização, Droga e Outro. Para anotação, foi aplicado um algoritmo simples de busca de *string*: a partir de uma entidade atrelada a uma denúncia, o algoritmo busca a *string* da entidade no texto da denúncia. O *corpus* foi avaliado utilizando a medida F1, com os modelos BERT Multilingual e BERT Acórdãos, alcançando 73,83% e 72,78%, respectivamente, utilizando todas as categorias. Considerando um cenário seletivo, no qual a categoria Outros foi removida do *corpus*, os modelos alcançaram medida F1 de 80,94% 81,04%.

O *corpus* KauaneJunior (Fernandes et al., 2020b) é composto por 3.022 decisões judiciais sobre a matéria "negativação indevida", do Tribunal de Justiça do Estado do Rio de Janeiro no ano de 2016. Os documentos foram anotados manualmente, utilizando diretrizes específicas de anotação baseadas em exemplos. Foram utilizados seis tipos de entidades legais: três relacionadas à definição de valor, aumento ou diminuição do dano moral (ValMorDm, IncMorDm e DecMorDm, respectivamente), uma para valor de honorários advocatícios (ValLegFee), além de duas entidades relacionadas a prazos iniciais, tanto para correção monetária (InitIntArr), como para juros de mora (InitMon). Foram utilizados cinco modelos de aprendizagem na avaliação da tarefa: *Bidirectional Gated Recurrent Units network* (BiGRU) (Graves et al., 2013), BiLSTM, CRF, e duas combinações destes modelos. Tanto na análise por entidades quanto no resultado geral, o modelo

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

BiLSTM-CRF obteve o melhor desempenho, alcançado a medida F1 de 94,79%.

O *corpus* Aposentadoria (Neto; Faleiros, 2021) contém entidades nomeadas de 10 classes associadas a atos de aposentadoria de servidores públicos do Diário Oficial do Distrito Federal (DODF), validadas pelo Tribunal de Contas do Distrito Federal e Territórios. O *corpus* faz parte de um grande conjunto de dados criado por colaboradores do projeto *KnEDLe*[21] da Universidade de Brasília. As entidades mapeadas foram Ato, Nome_Ato, Classe, Cod_Matricula_Ato, Empresa_Ato, Fund_Legal, Cargo, Quadro, Padrão, Processo e estão anotadas no formato CoNLL-2003 (Sang; De Meulder, 2003). A divisão entre treinamento, validação e teste foi feita em ordem cronológica dos DODFs, para não dividir entidades do mesmo documento do DODF em diferentes conjuntos. O conjunto de treinamento foi formado por 3.860 sentenças, o de validação por 828 sentenças e o de teste por 827 sentenças.

O *corpus* UlyssesNER-Br (Albuquerque et al., 2022) é composto por 150 projetos de lei (PL-corpus) e 800 consultas legislativas (ST-corpus) da Câmara dos Deputados brasileira. Anotados manualmente em três fases e por três grupos de anotadores, o *corpus* possui sete categorias semânticas, algumas contendo tipos específicos. A avaliação da concordância entre anotadores alcançou coeficiente Kappa de 91% na média. Além de cinco categorias baseadas no HAREM (Santos; Cardoso, 2007b), Pessoa, Localização, Organização, Evento e Data, o *corpus* possui duas categorias específicas para o domínio legislativo: Fundamento e Produto de Lei. Fundamento refere-se a entidades relacionadas a leis (FUNDlei e FUNDApelido), resoluções, decretos, bem como a entidades específicas do domínio, como projetos de lei (FUNDprojetodelei), que são propostas de lei em discussão pelo parlamento, e consultas legislativas (FUNDsolicitacaotrabalho), realizadas pelos parlamentares. Produto de Lei refere-se a sistemas (PRODUTOsistema), programas (PRODUTOprograma) e outros produtos (PRODUTOoutros) criados a partir da legislação. O *corpus* foi avaliado com os modelos HMM e CRF, obtendo melhor resultado com o CRF, alcançando a medida F1 de 81,04%.

Posteriormente, o *corpus* UlyssesNER-Br foi expandindo com texto informal gerado pelo usuário (Costa et al., 2022). Para a expansão, foram utilizados, aproximadamente, 1000 comentários públicos de cidadãos (C-corpus), relacionados a projetos de lei em tramitação, coletados de enquetes do portal da Câmara dos Deputados. Foram anotadas manualmente as mesmas entidades da primeira versão (Albuquerque et al., 2022), com exceção da FUNDsolicitacaotrabalho, a qual só está presente nos documentos de consultas legislativas. O coeficiente de anotação Kappa alcançou a média de ∼80%. O modelo BERT com ajuste fino, treinado com a junção do C-corpus e PL-corpus, melhorou o resultado individual do treinamento para C-corpus, alcançando medida F1 de 73,90%.

Correia et al. (2022) disponibilizaram um *corpus* anotado manualmente para REN em português no contexto do Judiciário brasileiro. O *corpus* contém 594 decisões publicadas pelo STF e anotadas por 76 estudantes de Direito. O processo de anotação foi composto por três fases, duas de treinamento e uma fase final. Os autores utilizaram um processo de anotação detalhada de decisões jurídicas, na qual foram anotados dois níveis de entidades jurídicas aninhadas. No primeiro nível, foram anotadas quatro entidades mais abrangentes (categorias): Precedente, Citação acadêmica, Referência legislativa e Pessoa. No segundo nível, foram anotadas entidades específicas de cada uma das quatro categorias. No total, foram identificadas 24 entidades aninhadas (tipos). Dois grupos de estudantes, ao longo de quase um ano, realizaram a tarefa de anotação sob rigorosa supervisão. Ambos os grupos

[21] https://unb-knedle.github.io/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

realizaram anotação na mesma coleção de 1.363 trechos, com um coeficiente de concordância Kappa médio de 70%. Os modelos CRF e BiLSTM-CRF construídos obtiveram a mesma medida F1 de 90%.

O artigo de Brito et al. (2023) apresenta uma metologia de anotação manual de documentos jurídicos que serviu como base para construção do CDJUR-BR, um *corpus* de REN-Legal com 1.216 documentos de processos judiciais brasileiros, contendo 21 entidades específicas das categorias Pessoa, Prova, Pena, Endereço, Sentença e Norma. Um comitê com professores da área do Direito e da Computação, juntamente com os especialistas do domínio, definiu os principais parâmetros do CDJUR-BR, bem como garantiu a adequada aplicação da metodologia. Foram selecionados documentos das classes-CNJ mais representativas dos processos de primeiro grau, encerrados em 2019 no Tribunal de Justiça do Estado do Ceará (85% dos processos de primeiro grau são das classes selecionadas). O processo de anotação foi realizado em duas fases, sendo a primeira de treinamento. Foram definidos três grupos de anotadores formados por especialista. A avaliação da concordância alcançou coeficiente Kappa de 0,69 para 73% dos documentos, e os demais documentos passaram por revisões por especialistas e etapas de refinamento. Com a CDJUR-BR, os resultados apontaram superioridade do modelo BERT com medida F1 média de 58%. Testes comparativos entre CDJUR-BR e LeNER-BR indiciaram que o modelo REN treinado com a CDJUR-BR é superior em reconhecer entidades de outros *corpora*.

O *corpus* PersoSEG (Guimarães et al., 2024) é composto por 127 documentos legais em português, no período de 2001 a 2015, obtidos do DODF. A pesquisa utiliza técnicas de REN aplicadas às tarefas de segmentação e rotulação em documentos legais. Foram utilizadas 12 tipos de entidades manualmente anotadas, representando atos da administração pública, como por exemplo, abono de permanência, (Ato_Abono_Permanencia), nomeação ou exoneração de cargos comissionados (Ato_Nomeacao_Comissionado e Ato_Exoneracao_Comissionado) ou efetivos (Ato_Nomeacao_Efetivo e Ato_Exoneracao_Efetivo), entre outras. O *corpus* foi avaliado utilizando combinações dos modelos CRF, CNN e LSTM, com melhor medida F1 média de 75,65% para REN com o modelo CRF.

A pesquisa desenvolvida por Albuquerque et al. (2026) estende o *corpus* UlyssesNER-Br originalmente voltado para o domínio legislativo, para um *corpus* jurídico mais abrangente, unificado e de dados abertos, cobrindo três subdomínios: projetos de lei, jurisprudência e leis. O *corpus* UlyssesLegalNER-Br é composto por 560 documentos públicos, anotados por meio de uma abordagem híbrida, que combina sugestões automáticas com curadoria humana especializada. Com o objetivo de preservar a compatibilidade com a versão anterior, os autores utilizaram as entidades do UlyssesNER-Br, e acrescentaram duas novas categorias, Jurisprudência e Valor Monetário, totalizando 9 categorias semânticas e 23 tipos específicos. A pesquisa revisou sistematicamente as anotações anteriores a partir da versão original do PL-corpus (público), atualizando o esquema de anotação, corrigindo inconsistências e incorporando as novas entidades. O *corpus* ampliado passou então a ser formado por três *subcorpora*: o PL-corpus_v3, voltado para o domínio legislativo, ampliado para 300 documentos e com cobertura temática mais ampla, incluindo textos do Senado Federal; o Juris-corpus, com 160 decisões de diferentes cortes superiores e órgãos do sistema jurídico brasileiro; e o Leis-corpus, o primeiro *corpus* de REN composto exclusivamente por leis brasileiras, num total de 100 documentos, contendo, entre outros, a Constituição Federal de 1988, três constituições estaduais, além de outros códigos. Os *corpora* foram avaliados utilizando um protocolo experimental sistemático, tanto individualmente como de forma unificada, com modelos de aprendizado de máquina consolidados

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

no estado da arte (CRF, BiLSTM, Glove, BERTimbau). O estudo incluiu ainda uma análise complementar de eficiência computacional destes modelos. O BERTimbau apresentou os melhores resultados em todos os *corpora*, alcançando a maior medida F1 de 96,18% no *corpus* unificado. A análise dos resultados também permitiu observar como diferentes arquiteturas se comportam em *corpora* jurídicos distintos, reforçando o papel do *corpus* desenvolvido como um recurso multigênero que reduz a fragmentação de *corpora* jurídicos em português, oferecendo uma base consistente para estudos futuros em diferentes granularidades e tipos documentais. Todos os recursos do projeto (dados e modelos) foram publicados abertamente.

13.4.2 Modelos de linguagem neurais para REN-Legal

O Capítulo Modelos de linguagem descreve os conceitos relacionados aos modelos de linguagem, detalhando o funcionamento dos modelos probabilísticos, neurais e de grande escala (*Large Language Models* ou LLMs). Nesta subseção, são apresentados estudos e modelos neurais específicos para o REN-Legal em língua portuguesa.

A pesquisa de Bonifacio et al. (2020) avalia o uso de modelos de linguagem neural para as tarefas de REN-Legal, utilizando os modelos pré-treinados para a língua portuguesa ELMo e BERT, ajustados para o domínio legal. Os ajustes foram feitos utilizando o *corpus* Acordãos-TCU, composto de 298 mil documentos do Tribunal de Contas da União (TCU)[22]. Os modelos foram avaliados em diferentes cenários, utilizando os *corpora* HAREM, LeNER-Br e DrugSeizures-Br. Os resultados demonstraram que os modelos treinados no domínio foram melhores. O modelo BERT obteve o melhor desempenho, alcançando a medida F1 de 89,39%, superando o LeNER-Br em cerca de quatro pontos na média.

O trabalho de Zanuz; Rigo (2022) apresenta os primeiros modelos BERT ajustados para REN-Legal em língua portuguesa. Dois modelos pré-treinados, `bert-base-portuguese-cased` e `bert-large-portuguese-cased`, foram utilizados no treinamento. O ajuste fino foi realizado com 10 épocas sobre os dados de treinamento e o melhor modelo, considerando a medida F1, foi selecionado para avaliar os resultados no conjunto de teste. Os modelos propostos foram comparados com resultados anteriormente obtidos para o *corpus*: LSTM+CRF, BERT-Multi, BERT-Acordaos e ELMo-Acordaos. Para o *corpus* LeNER-Br, os modelos BERT pré-treinados para língua portuguesa (Souza et al., 2020a) ajustados para o domínio alcançaram medida F1 de 91,14% no BERTimbau-Large e 90,62% no BERTimbau-Base.

O estudo de Albuquerque et al. (2023a) avaliou duas arquiteturas de *deep learning*, BiLSTM-CRF e BERT+Fine-Tuning, aplicadas a tarefa de REN-Legal em português brasileiro, utilizando os *corpora* UlyssesNER-Br e LeNER-Br. Para o UlyssesNER-Br, foi utilizado o conjunto de documentos que contém apenas projetos de lei (PL-corpus) visto que estes estão disponíveis publicamente. O modelo BERT+Fine-Tuning alcançou resultados superiores, com medida F1 de 88,27% para o LeNER-Br e 83,17% para o UlyssesNER-Br/PL-corpus, com um aumento de cerca de 7% neste último em relação à primeira versão.

O trabalho de Nunes et al. (2024b) explora como o REN pode ser aprimorado usando técnicas semi-supervisionadas. A abordagem consiste em uma estratégia de autoaprendizagem para ajustar um modelo BERT projetado para REN, usando documentos legislativos escritos em português brasileiro como estudo de caso. O UlyssesNER-Br foi utilizado para treinamento e avaliação. O BERTimbau com autoaprendizagem obteve a medida F1 de 86,70 ± 2,28.

[22] https://www.kaggle.com/datasets/ferraz/acordaos-tcu

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

A pesquisa de Garcia et al. (2024a) investiga a aplicação de métodos de PLN na adaptação de modelos para o domínio legal em língua portuguesa. O modelo RoBERTaLexPT foi desenvolvido aplicando técnicas de deduplicação de dados (identificação e eliminação de cópias duplicadas de dados) ao modelo RoBERTa, utilizando os *corpora* LegalPT (composto de seis *corpora* legais), o *corpus* CrawlPT (composto por três *corpora* do domínio geral) e o Tesemô (Siqueira et al., 2024). Os experimentos incluíram pré-treinamento e *fine-tuning* dos modelos em tarefas de REN-Legal e classificação com *benchmarks*, comparando os resultados dos modelos genéricos com específicos. Os resultados mostraram que o RoBERTaLexPT treinado com a combinação de *corpora* legal e geral, superou outros modelos legais, incluindo o RoBERTaTimbau, alcançando medida F1 média de 90,73% para LeNER-Br e 88,56% UlyssesNER-Br (categoria).

A Tabela 13.1 apresenta os resultados obtidos para a tarefa de REN-legal em língua portuguesa utilizando modelos de linguagem neural. É possível verificar que, a partir de 2020, com o trabalho de Bonifacio et al. (2020), os trabalhos assumiram o estado da arte utilizando o modelo BERT.

Tabela 13.1: Modelos de linguagem neural estado da arte para REN-Legal em português

Corpus	Modelo	Medida F1
Justiça do Trabalho	ELMO+GLOVE (Castro, 2019)	93,81%
DrugSeizures-Br	BERT (Bonifacio et al., 2020)	89,39%
LeNER-Br	BERTimbau-Large+Fine-Tuning (Zanuz; Rigo, 2022)	91,14%
CDJUR-BR	BERT (Brito et al., 2023)	58,00%
UlyssesNER-Br (PL-Corpus)	RoBERTaLexPTbase (Garcia et al., 2024a)	88,56% (categoria)

As avaliações recentes de REN-Legal em português mostram que ganhos de desempenho precisam ser interpretados com cautela, sobretudo quando há risco de contaminação entre treino, validação e teste ou inconsistências de anotação nos *corpora*. Essa preocupação acompanha discussões mais amplas em PLN sobre sensibilidade a particionamentos e estabilidade de métricas (Gorman; Bedrick, 2019; Sainz et al., 2023), e foi evidenciada por Nunes (2025), ao analisar os *corpora* LeNER-Br e UlyssesNER-Br/PL-corpus em um *benchmark* com 22 modelos de diferentes arquiteturas, a partir de um processo de decontaminação que envolveu remoção de sentenças duplicadas entre partições e revisão de anotações parciais e ambíguas. No LeNER-Br, a limpeza reduziu em 11,2% o total de entidades anotadas, enquanto no UlyssesNER-Br/PL-corpus também foram removidas instâncias tanto no nível de categorias quanto no de tipos. Os efeitos foram expressivos: no LeNER-Br, 18 dos 22 modelos apresentaram diferença estatisticamente significativa entre as versões contaminada e decontaminada, com perdas de micro-F1 de até 10,14 pontos; no PL-corpus, as oscilações chegaram a variar de $-17{,}71$ a $+16{,}23$ pontos em categorias e de $-10{,}00$ a $+14{,}18$ em tipos. Em outras palavras, a contaminação não apenas pode inflar resultados absolutos, mas também modificar a leitura comparativa entre os próprios modelos.

Na versão decontaminada, os melhores resultados de micro-F1 foram obtidos por RoBERTaLexPT-base no LeNER-Br, por Legal-portuguese-R-large no UlyssesNER-Br/PL-corpus em categorias e por Albertina-100m no nível de tipos, o que sugere que o tamanho do modelo, por si só, não explica o desempenho observado. Ao mesmo tempo, a análise de erros mostrou que a REN-Legal continua marcada por dificuldades

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

típicas do domínio jurídico, como ambiguidades entre classes semanticamente próximas (e.g., LOCAL e ORGANIZAÇÃO, no LeNER-Br), problemas no reconhecimento de entidades compostas (principalmente em referências legais), e forte sensibilidade à qualidade das anotações (e.g., PRODUTODELEI, FUNDAMENTO e ORGANIZAÇÃO, no UlyssesNER-Br/PL-corpus). Esse quadro reforça duas implicações metodológicas centrais: a curadoria do *corpus* deve integrar o protocolo experimental desde o início, com deduplicação, revisão de anotações e tratamento de ambiguidades; e a comparação entre modelos não deve se apoiar apenas em médias de medida F1. Nesse sentido, a análise estatística com teste de Wilcoxon pareado e teste de Friedman com pós-teste de Nemenyi indica que pequenas diferenças numéricas nem sempre correspondem a vantagens consistentes, o que torna mais adequada uma avaliação que combine desempenho, robustez estatística e controle rigoroso dos dados.

13.5 Modelos de linguagem de grande escala (LLM) para REN

Conforme descrito na Seção Tendências, os LLMs se diferenciam dos demais modelos de linguagem neural pré-treinados devido a (i) quantidade enorme de parâmetros, o (ii) enquadramento na categoria de métodos de IA Gerativa, e (iii) habilidades emergentes, as quais não costumam ser observadas em modelos menores. Neste sentido, destacamos aqui algumas aplicações de LLMs em tarefas de PLN, independente de língua. Por fim, na Seção 13.5.1, detalhamos as iniciativas de REN-Legal em língua portuguesa.

O trabalho de Wang et al. (2021) foi um dos primeiros a investigar como o modelo de linguagem GPT-3 (Brown et al., 2020c) pode ser utilizado como uma ferramenta de rotulagem de dados de baixo custo para treinar outros modelos de PLN. Os autores conduziram um estudo empírico em nove tarefas de PLN, como sumarização, análise de sentimento e classificação de tópico. A pesquisa demonstrou que (i) o uso do GPT-3 para rotulagem de dados pode reduzir os custos de rotulagem em até 96% em comparação com a rotulagem humana, (ii) dados rotulados pelo GPT-3 podem superar o desempenho do próprio GPT-3 em configurações de *few-shot learning*, (iii) a combinação de rótulos gerados pelo GPT-3 com rótulos humanos pode melhorar ainda mais o desempenho dos modelos e, por fim, (iv) a estratégia de rotulagem ativa proposta no artigo, na qual os rótulos de baixa confiança do GPT-3 são revisados por humanos, melhora a qualidade dos dados anotados.

Especificamente para REN, Wang et al. (2023) apresentaram o GPT-NER para adaptar modelos de linguagem de grande escala à tarefa de REN, transformando-a em uma tarefa de geração de texto com *tokens* especiais para entidades. Foi introduzida uma estratégia de auto-verificação para mitigar alucinações do modelo. Foram realizados experimentos em cinco conjuntos de dados de REN para língua inglesa em domínios variados. O GPT-NER alcançou desempenho comparável aos *baselines* supervisionados utilizando *few-shot learning*, superando modelos supervisionados com dados de treinamento escassos. De maneira similar, o trabalho de Wei et al. (2024) apresenta o ChatIE, um framework de extração de informações com *few-shot learning* através de conversas com o ChatGPT[23]. O ChatIE foi avaliado em três tarefas de EI, entre elas a tarefa de REN, a qual foi transformada em um problema de pergunta e resposta com duas etapas. Na primeira etapa, são identificados os tipos de elementos em uma sentença. Na segunda etapa, é realizada a EI para cada tipo de elemento identificado. Para a tarefa de REN, foram usados dois conjuntos de dados relacionados a notícias, um em inglês e outro em chinês.

[23] https://chatgpt.com/

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

O ChatIE alcançou excelente desempenho, superando alguns modelos em vários conjuntos de dados.

No domínio legal, o trabalho de Hussain; Thomas (2024) investiga a aplicação de modelos de LLM na extração de entidades específicas do domínio jurídico, em documentos de casos legais indianos. O estudo avalia a eficácia de várias arquiteturas de LLMs de última geração na identificação de entidades jurídicas. O conjunto de dados utilizado para avaliação foi o InLegalNER, um *corpus* contendo 14 entidades, sendo 11 entidades específicas do domínio legal, tais como Petição, Vítima, Número do Caso, Juiz, Advogado, entre outras. Foram avaliados quatro modelos de larga escala: LLaMA 3 (AI@Meta, 2024), Gemma (Team et al., 2024), Mistral (Jiang et al., 2023) e Phi-3 (Abdin et al., 2024), utilizando a técnica de *few-shot learning*, na qual o *prompt* foi elaborado instruindo o LLM a gerar respostas em formato JSON, incluindo o texto extraído e os rótulos de entidade correspondentes. Os modelos foram avaliados utilizando precisão, revocação e medida F1 como métricas. O LLaMA 3 alcançou a medida F1 de 59,17%, Gemma de 63,53% , Mistral de 63,76%. Phi-3 de 54,40%. Os principais resultados indicam que os modelos Mistral e Gemma se destacaram em termos de equilíbrio entre precisão e revocação, essenciais para a identificação precisa de entidades.

13.5.1 LLM para REN-Legal em língua portuguesa

Visando encontrar alternativas para os métodos tradicionais de anotação manual, reduzindo os custos e esforços de anotação, o trabalho de Oliveira et al. (2024c) investiga o uso de LLM baseados em *prompt* e supervisão fraca para a tarefa de REN-Legal em português. O estudo utilizou o DODFCorpus I, disponibilizado pelo projeto KnEDLe da Universidade de Brasília. Foi utilizado o *corpus* "Atos de Contratos e Licitações", que contém documentos públicos diários que relatam as ações do governo do Distrito Federal do Brasil, particionado em 783 atos de treinamento, 379 atos de validação e 380 atos de teste. Foram comparadas três abordagens: (i) baseada em *prompt*, com *few-shot learning*, apresentando três exemplos de dados rotulados ao GPT-3, (ii) supervisão fraca, com sistema baseado em regras e heurísticas para gerar conjuntos de dados rotulados rapidamente e (iii) anotação humana. As 783 instâncias de treinamento foram rotuladas pelo GPT-3, realizando a predição de 1.565.108 *tokens*, com um custo final de 31,30 dólares utilizando o modelo Davinci. Quatro modelos de linguagem neural pré-treinados, com ajuste fino, foram usados para a construção do modelo de REN: BERTimbau, Lener-BR, RoBERTa e DistilBERT-PT[24]. Para as onze entidades do *corpus*, considerando a média dos quatro modelos e a métrica de medida F1, os resultados foram: 52,8% (GPT-3), 67,9% (supervisão fraca), 71,3% (anotação humana) e 64,6% (GPT-3 + supervisão fraca). Considerando a métrica de pontuação de preservação, que preserva o desempenho apresentado pelos modelos treinados por humanos, os resultados foram: 74,0% (GPT-3), 95,6%(supervisão fraca), 90,7% (GPT-3 + supervisão fraca) e 83,9% (GPT-3 + 30% de anotação humana). O BERTimbau superou o LenerBR em alguns casos, mesmo este último tendo sido adaptado para documentos legislativos brasileiros.

O trabalho de Nunes et al. (2024a) explora a aplicação do paradigma de Aprendizagem em Contexto (em inglês, *In-Context Learning* ou ICL), para melhorar o desempenho dos LLMs para o REN-Legal em língua portuguesa. O Sabiá, um LLM em português (Pires et al., 2023), foi utilizado para extrair entidades nomeadas no domínio. Dois *corpora* de REN para o domínio legal (LeNER-Br e UlyssesNER-Br) foram utilizados como entrada para a

[24] https://huggingface.co/adalbertojunior/distilbert-portuguese-cased

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

tarefa de *prompt*. Para a seleção das amostras, foram utilizadas três heurísticas: (i) top K exemplos similares, (ii) top K exemplos similares por tipo de entidade, e (iii) K amostras aleatórias. O modelo generativo foi ensinado a catalogar as entidades e o resultado foi comparado com a anotação original. No *corpus* LeNER-Br, a heurística K similares por categorias apresentou medida F1 de 51% com oito exemplos. No resultado por tipos, três classes de entidades superaram 50% de medida F1: Pessoa com 68%, Organização com 53%, e Jurisprudência com 53%. No *corpus* UlyssesNER-Br, a mesma heurística atingiu medida F1 de 48% com oito exemplos, necessitando de onze exemplos para alcançar 51%. No resultado por tipo, três classes também obtiveram mais de 50% de medida F1: Data com 81%, Local com 69%, e Pessoa com 60%. De acordo com os autores, as classificações incorretas são resultados de uma formatação similar de termos que se referem a conceitos diferentes (Jurisprudência e Legislação, por exemplo), ambiguidade semântica (o "jornal diário catarinense" pode ser um Local ou uma Organização) e estruturas e jargões específicos do domínio (Plenário, junto com referências a jurisprudência ou legislação, separadas por um traço, confunde o modelo). A amostragem aleatória apresentou o pior desempenho.

O trabalho de Coelho et al. (2024) aborda a extração de informações no domínio jurídico brasileiro, especificamente extraindo características estruturadas de pareceres jurídicos relacionados a reclamações de consumidores. O estudo compara dois métodos: técnicas tradicionais de aprendizado supervisionado e o uso de LLM, especificamente o ChatGPT, nas tarefas de classificação de texto e de REN. O *corpus* utilizado nos experimentos contém 959 pareceres jurídicos manualmente anotados, redigidos por juízes de primeira instância no Tribunal de Justiça do Rio de Janeiro. As entidades anotadas fora: Dano Moral, Dano Material, Honorários Advocatícios e Restituição. Utilizando a versão GPT-3.5-turbo, a técnica de *prompt* envolveu três principais itens: detalhes das informações a serem extraídas, formato de saída e exemplos opcionais (*few-shot*). Os resultados obtidos foram em termos de acurácia e raiz quadrática média dos erros (em inglês, *Root Mean Squared Errors* ou RMSEs): BERT (*Fine-tuning*) 96,2% e 825,8, BERT-CRF (*Fine-tuning*) 98,2% e 639,8, BERT-LSTM (*Feature-based*) 95,4% e 738,6, BERT-LSTM-CRF (*Feature-based*) 98,9% e 277,2, ChatGPT Entity Extractor 98,4% e 1230,9. Os resultados dos experimentos mostraram que ambas as abordagens (aprendizado supervisionado tradicional e ChatGPT) alcançaram desempenhos semelhantes em termos de métricas de avaliação tradicionais. No entanto, o uso do ChatGPT reduziu substancialmente a complexidade e o tempo necessário para o processo de EI.

A Tabela 13.2 apresenta os resultados obtidos para a tarefa de REN-legal em língua portuguesa utilizando, exclusivamente, LLMs.

Tabela 13.2: LLMs estado da arte para REN-Legal em português.

Corpus	Modelo	Métrica
DODFCorpus I	GPT-3 Davinci + *few-shot learning* com 3 exemplos (Oliveira et al., 2024c)	52,8% (F1)
LeNER-Br	Sabiá + *in-context learning* com 8 exemplos (Nunes et al., 2024a)	51,0% (F1)
UlyssesNER-Br	Sabiá + *in-context learning* com 11 exemplos (Nunes et al., 2024a)	51,0% (F1)
Reclamações de Consumidores	GPT-3.5-turbo + *few-shot learning* (ChatGPT Entity Extractor) (Coelho et al., 2024)	98,4% (Acurácia)

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

13.6 Aplicação de REN em sistema de recuperação de informação legislativa

O Capítulo 2 destaca a importância da área da Recuperação de Informação (RI) para o PLN. Pesquisas na área de *Legal Information Retrieval* ou recuperação de informação no domínio legal (RI-Legal) focam principalmente em aplicações que utilizam a linguagem judiciária, forense ou processual (Sansone; Sperlí, 2022). O Capítulo 12 destaca que a esfera de aplicação do subdomínio legislativo é a criação de textos legais, como leis e estatutos, entre outros. Sistemas RI-Legal têm sido criados ou adaptados para diversas tarefas legislativas (Heusden et al., 2023; Smywiński-Pohl et al., 2021).

Como mencionado anteriormente, o projeto Ulysses (Seção 13.2) é um conjunto de iniciativas de IA aplicados no processo legislativo brasileiro, criado pela Câmara dos Deputados, o qual recebeu recentemente novos recursos em várias frentes de trabalho[25]. Entre estes recursos, foi utilizado o *pipeline* para um sistema de RI-Legal proposto por Souza et al. (2021a), visando automatizar tarefas de busca de documentos relevantes. REN-Legal foi inserido neste contexto.

Figura 13.3: *Pipeline* do UlyssesNERQ, adaptado de Albuquerque et al. (2024b)



O UlyssesNERQ (Albuquerque et al., 2024b) atualiza o *pipeline* do sistema de RI-Legal original, utilizando NER-Legal e a técnica de expansão de consulta. Através desta técnica, o sistema de RI modifica a *string* de consulta original, inserindo informações que possam maximizar seus resultados (Seção 2.5). A Figura 13.3 detalha o processo utilizado. O sistema recebe como entrada a *query* com solicitação de consulta legislativa e o *corpus* UlyssesNER-Br. Após a aplicação de tarefas de pré-processamento de texto na consulta, é feita a identificação de duas entidades legislativas, FUNDprojetodelei e FUNDsolicitacaotrabalho. Em caso de identificação destas entidades, a consulta é expandida com os conteúdos dos documentos correlacionados: (i) para a entidade FUNDprojetodelei, a consulta é expandida com a ementa dos projetos de lei encontrados; (ii) para FUNDsolicitacaodetrabalho, a expansão é feita com todo o conteúdo da solicitação legislativa encontrada. Em caso de não ter sido encontrada nenhuma entidade, a consulta não é modificada. O Quadro 13.2 apresenta exemplos fictícios de consultas legislativas, com e sem expansão utilizando entidades. Ao final, são retornados os *top K* documentos mais relevantes encontrados.

[25]Recursos disponíveis publicamente em https://github.com/ulysses-camara

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

Quadro 13.2: Exemplos de expansão de consultas legislativas utilizando REN, adaptado de (Albuquerque et al., 2024b). Expansões de consulta em negrito.

Consulta legislativa (*query*)	Entidades encontradas	Consulta expandida
Solicito revisão dos projetos de lei PLP 438/2014 e PLP 137/2015.	PLP 438/2014 PLP 137/2015	Solicito revisão dos projetos de lei PLP 438/2014 e PLP 137/2015. **Dispõe sobre o procedimento para a criação, a incorporação, a fusão e o desmembramento de Municípios, nos termos do § 4º do art. 18 da Constituição Federal, altera a Lei nº 5.172, de 25 de outubro de 1966, e dá outras providências. Dispõe sobre o procedimento para a criação, a incorporação, a fusão e o desmembramento de Municípios, nos termos do § 4º do art. 18 da Constituição Federal, e dá outras providências.**
Solicito complementação de parecer em função da apensação do PL 5165/2019 ao mesmo.	PL 5165/2019	Solicito complementação de parecer em função da apensação do PL 5165/2019 ao mesmo **Altera o art. 3º da Lei n. 13.855, de 8 de julho de 2019, para modificar a sua cláusula de vigência**
Parlamentar solicita aprovação	–	Parlamentar solicita aprovação

Para validação do UlyssesNERQ, os experimentos utilizaram 32 configurações de pré-processamento de texto nas *strings* de busca com as técnicas de *stemming*, contagem de frequência, modelos de linguagem *n-gram*, e diversas combinações. A expansão de consulta foi então avaliada aplicando três modelos de REN-Legal ajustados para o domínio: CRF (Albuquerque et al., 2022), BERT (Albuquerque et al., 2023a), e uma adaptação do modelo Bertikal (Polo et al., 2021) ajustado para as entidades legislativas. Além destes modelos, foram utilizadas duas técnicas de expansão de consulta, a primeira utilizando sinônimos, e a segunda utilizando termos relacionados ou representativos ao conteúdo, através do algoritmo RM3 (Nogueira et al., 2019). Os resultados foram verificados isoladamente e através de análises estatísticas. A métrica utilizada foi a medida de Revocação para 20 documentos (Recall@20). O melhor resultado foi alcançado pelo modelo BERT, alcançado 74,58% de revocação na análise individual, e média de 65,19% ± 0,0840, ultrapassando os resultados do *pipeline* original. Este modelo foi novamente combinado com as técnicas utilizadas, alcançando o mesmo resultado individual anterior, mas com média um pouco melhor utilizando a combinação de RM3 + BERT, com 65,35% ± 0,0810 de revocação. Comparando com o resultados do pipeline original, houve uma melhoria de cerca de 1,94% para os melhores resultados, e de 8,58% no resultado geral.

13.7 Considerações finais

Neste capítulo, foram detalhadas as principais iniciativas na área de reconhecimento de entidades nomeadas no domínio legal, com foco na língua portuguesa. Destacou-se o crescimento do uso de técnicas de Processamento de Linguagem Natural nesse campo, especialmente no Brasil. Foram discutidos os desafios inerentes à extração de informações de documentos jurídicos, considerando a grande quantidade e a complexidade desses textos. Além disso, foram mencionadas aplicações relevantes de inteligência artificial nas áreas jurídica e legislativa, como os projetos Victor e Ulysses, respectivamente. Apresentou-se um panorama abrangente das iniciativas em diversas línguas relacionadas ao REN no domínio legal, com suas abordagens e métodos, como aprendizado supervisionado e modelos neurais. Foram descritos *corpora* importantes para o REN-Legal em língua portuguesa e avaliação de modelos treinados especificamente para esse domínio, destacando seu desempenho em *corpora* brasileiros. Por fim, explorou-se a aplicação de REN em sistemas de recuperação de informação legislativa, como o UlyssesNERQ, que aprimora as consultas

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

por meio da identificação de entidades nomeadas, aumentando a eficiência na busca de documentos relevantes.

No campo de pesquisa sobre REN, diversas áreas têm se destacado com foco na melhoria contínua dos modelos. Investigações com LLMs têm se mostrado promissoras, permitindo avanços no processamento de grandes volumes de dados e na adaptação a tarefas específicas. Contudo, o desbalanceamento entre classes de entidades, no qual certas categorias apresentam uma representação desproporcionalmente menor nos dados, continua sendo um desafio. Esse fenômeno, aliado à existência de entidades pouco representativas, podem impactar negativamente o desempenho dos modelos. Estratégias como *Data augmentation*, que visa aumentar a quantidade de dados de treinamento de forma sintética, e *Active learning*, que melhora a eficiência do aprendizado ao selecionar amostras mais informativas, têm sido exploradas para mitigar esses problemas. Por fim, a dependência de domínio, dada a especificidade dos textos legais, impõe a necessidade de adaptar os modelos a contextos muito particulares, o que exige uma personalização contínua dos modelos para cada subdomínio do Direito.

Agradecimentos

Gostaríamos de expressar nossa gratidão à Câmara dos Deputados Brasileira, ao Instituto de Ciências Matemáticas e de Computação da Universidade de São Paulo (ICMC/USP) e ao Instituto Nacional de Inteligência Artificial (IAIA) por possibilitarem o acesso a informações cruciais para o desenvolvimento da pesquisa, assim como a todos os colaboradores do projeto Ulysses. Além disso, agradecemos também à colaboração científica dos autores dos trabalhos citados neste capítulo, que gentilmente nos responderam com dados e recursos de suas pesquisas.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Capítulo 14

Geração Automática de Ementas Jurídicas

Kenzo Sakiyama
Roseval Malaquias Junior
Raphael Montanari
Rodrigo Nogueira
Roseli A. F. Romero

Publicado em: 20/11/2024

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

14.1 Introdução

14.1.1 Contextualização

Segundo dados do Conselho Nacional de Justiça (CNJ) (Justiça - CNJ, 24DC), ao final de 2022, havia 81,4 milhões de processos em trânsito no judiciário brasileiro, um aumento de 7,5% em relação ao ano anterior. O relatório também destaca que, nos casos mais extremos, o tempo de análise pode levar vários anos. Além do esforço humano necessário para a análise e elaboração dos documentos, é preciso alocar recursos financeiros para sua manutenção ao longo de todo o processo. Neste cenário, o aprimoramento de sistemas computacionais jurídicos, com a automação do fluxo de documentos, torna-se essencial para agilizar a tramitação no judiciário brasileiro.

Quadro 14.1: Exemplo de ementa. A verbetação está destacada em negrito.

DIREITO CONSTITUCIONAL, TRIBUTÁRIO, PREVIDENCIÁRIO E PROCESSUAL CIVIL. CONTRIBUIÇÃO SOCIAL. ART. 2º DA LEI Nº 7.856, DE 25 DE OUTUBRO DE 1989. MAJORAÇÃO DE 8% PARA 10%, DA ALÍQUOTA DA CONTRIBUIÇÃO SOCIAL. SUCUMBÊNCIA. HONORÁRIOS ADVOCATÍCIOS. CUSTAS PROCESSUAIS.
1. Firmou-se em Plenário do Supremo Tribunal Federal o entendimento no sentido de que, em se tratando de "lei de conversão da Medida Provisória nº 86, de 25 de setembro de 1989, da data da edição desta é que flui o prazo de noventa dias previsto no art. 195, § 6º, da CF, o qual, no caso, teve por termo final o dia 24 de dezembro do mesmo ano, possibilitando o cálculo do tributo, pela nova alíquota, sobre o lucro da recorrente, apurado no balanço do próprio exercício de 1989" (RE 197.790-6-MG, Rel. Min. ILMAR GALVÃO).
2. Adotados os fundamentos desse precedente, o RE, no caso, é conhecido e provido.
3. As autoras ficaram vencedoras, quanto a outros pedidos, nas instâncias ordinárias, sem que o R.E. abordasse esses pontos. Sendo maior a sucumbência da ré, pagará honorários advocatícios às autoras.
4. Custas em proporção.

Este Capítulo trata da geração automática de **ementas**, documentos amplamente utilizados em tribunais brasileiros, cuja função é fornecer um resumo de processos judiciais. As

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

ementas oferecem uma representação sucinta e simplificada das decisões, permitindo que profissionais do Direito compreendam rapidamente o conteúdo sem a necessidade de ler o acórdão completo. Conforme demonstra o exemplo no Quadro 14.1, estes documentos usualmente contêm dois elementos principais: **verbetações** e **parágrafos enumerados**.

As **verbetações** são cabeçalhos constituídos por frases curtas, em caixa-alta, que descrevem os assuntos mais relevantes da decisão. Além de resumir o conteúdo, elas auxiliam na tarefa de busca e recuperação de jurisprudências (precedentes) (Guimarães; Santos, 2016). Por outro lado, os **parágrafos enumerados** expõem as teses da decisão, sendo que o último apresenta geralmente a conclusão do tribunal. Essa estrutura contribui para a agilidade no acesso à informação jurídica, facilitando a consulta à jurisprudência.

Vale destacar que, devido à sua estrutura e estilo, as verbetações assemelham-se a palavras-chave e a textos sumarizados. No entanto, diferentemente das palavras-chave, elas podem consistir em frases curtas. Além disto, enquanto resumos são redigidos de maneira fluida, as verbetações seguem um estilo linguístico único e rígido em comparação aos resumos. Um desafio adicional na redação das verbetações é que muitos dos termos utilizados não aparecem no restante da ementa, tornando sua produção uma tarefa que vai além da simples extração de informação.

Considerando as características do sistema judiciário brasileiro, os componentes das ementas e sua grande disponibilidade no Brasil, o objetivo deste Capítulo é apresentar métodos de PLN que propiciam a automatização da escrita de verbetações empregando os parágrafos enumerados como entradas para um modelo supervisionado generativo. Em específico, como Transformers (Brown et al., 2020c; Carmo et al., 2020b; Scao et al., 2022; Zhang et al., 2022) se mostraram efetivos em tarefas de geração de texto, neste capítulo os mesmos são avaliados na tarefa de geração de verbetações. O Quadro 14.2 ilustra exemplos de verbetações geradas por Transformers, nos quais fica evidente a capacidade desse modelo generativo, obtendo resultados semelhantes à verbetações escritas por especialistas.

Quadro 14.2: Exemplos de verbetações geradas por Transformers.

Original: PROCESSUAL CIVIL. EMBARGOS DE DECLARAÇÃO NO RECURSO ESPECIAL. ALEGAÇÃO DE OMISSÃO. RESSALVAS DE FUNDAMENTAÇÃO CONTIDAS NO VOTO-VISTA. OMISSÃO NÃO CARACTERIZADA. **Gerada:** PROCESSUAL CIVIL. EMBARGOS DE DECLARAÇÃO. OMISSÃO. VÍCIO INEXISTENTE. REJEIÇÃO.
Original: AGRAVO REGIMENTAL NO HABEAS CORPUS. TRÁFICO DE DROGAS. PRISÃO PREVENTIVA. PERICULUM LIBERTATIS. CRISE MUNDIAL DE COVID-19. NÃO DEMONSTRADA A EXISTÊNCIA DE RISCO À INTEGRIDADE FÍSICA DO ENCARCERADO. AGRAVO REGIMENTAL NÃO PROVIDO. **Gerada:** AGRAVO REGIMENTAL NO HABEAS CORPUS. TRÁFICO DE DROGAS. PRISÃO PREVENTIVA. PERICULUM LIBERTATIS. AGRAVO REGIMENTAL NÃO PROVIDO.

Este Capítulo apresenta mais uma aplicação de PLN na área do Direito, assim como as apresentadas no Capítulo 12. São utilizados modelos de linguagem (descritos no Capítulo Modelos de linguagem) em uma tarefa de geração de texto, similar à tarefa de sumarização (Capítulo 5). A metodologia proposta descreve como a coleta de dados foi realizada, apresenta o modelo transformer utilizado e também a técnica de validação e métrica adotadas. Além disto, emprega-se conceitos de Recuperação de Informação (Capítulo 2) para avaliação final das verbetações geradas.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

14.1.2 Organização do Capítulo

Este Capítulo está organizado da seguinte forma. A Seção 14.2.1 introduz conceitos preliminares relacionados à aplicação. Em seguida, a Seção 14.3 descreve a metodologia utilizada para a geração de verbetações. A Seção 14.4 detalha a metodologia adotada. Na sequência, os resultados obtidos são apresentados e discutidos na Seção 14.5. Por fim, conclusão é apresentada na Seção 14.6.

14.2 Conceitos Preliminares

14.2.1 Transferência de Aprendizado e Ajuste Fino

Uma propriedade importante para a popularidade de Transformers em PLN é a **Transferência de Aprendizado** (TA). Como o próprio nome sugere, a técnica visa transferir conhecimento de um domínio para outro. Em outras palavras, busca-se generalizar o conhecimento adquirido em um domínio A para um domínio B. É importante notar que essa transferência é mais vantajosa se os domínios estiverem de alguma forma relacionados (Zhuang et al., 2020).

Esta propriedade é útil em casos em que há uma grande diferença na quantidade de dados disponíveis no domínio e na tarefa alvo. Por exemplo, pensando no contexto de PLN, há uma abundância de textos disponíveis publicamente na web, porém, apenas uma fração é relacionada ao Direito (uma fração ainda menor, se considerarmos textos em português). Neste caso, a TA permite que um modelo seja treinado em um domínio com dados abundantes e, depois, treinado novamente no conjunto de dados destinado à tarefa alvo. Este segundo treinamento, por sua vez, é chamado de **Ajuste Fino**.

Um exemplo de TA no contexto de PLN utilizando Transformers é encontrado no modelo BERTimbau (Souza et al., 2020a) (Transformer codificador). Inicialmente, ele foi treinado no *corpus* genérico brWaC (Wagner Filho et al., 2018). Ambos, modelo e *corpus*, foram apresentados no Capítulo Modelos de linguagem. O treinamento inicial foi feito em um grande volume de texto, visando aprender estruturas linguísticas comuns da língua portuguesa. Este treinamento é baseado em modelagem de linguagem (também discutido no Capítulo Modelos de linguagem). Similar ao modelo BERT, o modelo BERTimbau, gera representações para os *tokens* que podem ser utilizadas em diferentes tarefas de PLN, como classificação, reconhecimento de entidades nomeadas, agrupamentos etc.

Vale notar que o treinamento inicial não necessariamente corresponde à tarefa final na qual o modelo de aprendizado de máquina será empregado. Além disto, o pré-treinamento possibilita um bom desempenho na tarefa final, mesmo que esta não possua uma quantidade abundante de dados para o ajuste fino. Além disso, após essa etapa inicial de treinamento, os modelos pré-treinados podem ser disponibilizados publicamente [1][2]. Este fato é interessante, pois possibilita que outros pesquisadores tenham à sua disposição modelos pré-treinados, prontos para serem utilizados em suas tarefas alvo de interesse.

14.2.2 *Bilingual Evaluation Under-study* (BLEU)

Para avaliar quantitativamente sistemas de geração de texto, muitas métricas foram propostas ao longo dos anos, como por exemplo a medida BLEU.

[1] https://pytorch.org/hub/
[2] https://huggingface.co/models

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

A métrica BLEU (Papineni et al., 2002) foi proposta originalmente para avaliar sistemas de tradução. De maneira simplificada, seu funcionamento envolve analisar a sobreposição entre as palavras (ou *tokens*) no texto gerado e aquelas nos textos de referência humanos. Na prática, as palavras são representadas por n-gramas (usualmente de uma a quatro palavras), após a tokenização dos textos. Os valores da métrica estão em uma faixa de 0 a 1, na qual uma pontuação mais próxima de 1 indica uma melhor qualidade no texto gerado.

Outra característica interessante é que a métrica considera todo o *corpus* de avaliação para seu cálculo. Assim, as estatísticas são obtidas a partir de todos os textos gerados e referências, sem realizar qualquer tipo de agregação, como uma média, para determinar a pontuação final.

Formalmente, o BLEU está definido da seguinte forma:

$$BLEU = Penalidade \times \Big(\prod_{i=1}^{4} Precisão_i\Big)^{1/4} \tag{14.1}$$

$$Penalidade = \min\Big(1, \exp\big(1 - \frac{\text{tamanho da referência}}{\text{tamanho do texto gerado}}\big)\Big) \tag{14.2}$$

$$Precisão_i = \frac{\sum_{\text{tg}\in\text{Gerados}} \sum_{i\in\text{tg}} \min(c^i_{gerado}, c^i_{ref)}}{C^i_t = \sum_{\text{tg'}\in\text{Gerados}} \sum_{i'\in\text{tg'}} c^{i'}_{gerado}} \tag{14.3}$$

na qual i define o i-ésimo n-grama, tg corresponde a um texto gerado, c^i_{gerado} corresponde ao número de ocorrências simultâneas do i-grama na referência e no texto gerado, c^i_{ref} corresponde ao número de ocorrências da i-grama no texto de referência, $c^{i'}_{tg}$ denota a quantidade de n-gramas no texto gerado e C^i_t informa o número total de i-gramas dos textos gerados.

Na fórmula apresentada, o componente de Penalidade (Equação 14.2) penaliza a produção de textos curtos, já que seu valor diminui à medida que os textos gerados são menores. Já que o número de termos no texto gerado influencia o denominador da Equação 14.3, quando esse valor é pequeno, facilita-se maximizar a precisão. Logo, a penalidade atua para incentivar a geração de textos mais longos. Note que os tamanhos são calculados considerando todos os exemplos, tanto os gerados quanto os de referência.

Em sequência, a Equação 14.3 ilustra um cálculo de precisão no qual verifica-se qual porcentagem dos n-gramas ocorre tanto no texto gerado quanto no texto de referência.

A Equação 14.1 define o BLEU, combinando componentes mencionados anteriormente. Como se pode observar, a métrica BLEU é a média harmônica das precisões dos n-gramas (tipicamente, entre 1 e 4), com uma penalização pelo tamanho do texto gerado.

Em suma, o BLEU destaca-se por sua simplicidade e rapidez de cálculo, além de estar correlacionado positivamente com as avaliações humanas (Papineni et al., 2002). No entanto, é importante reforçar que sua principal limitação está na falta de consideração dos aspectos semânticos dos textos comparados, já que sua metodologia se baseia no casamento de n-gramas. Além disso, a métrica é sensível ao método de tokenização utilizado.

14.3 Geração de verbetações

Neste Capítulo, serão apresentados os componentes da metodologia proposta para o desenvolvimento desta aplicação. Iniciando pelos dados utilizados, a Seção 14.3.1 discute a coleta dos documentos. Em seguida, a metodologia empregada para a geração automática

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

de verbetações e sua avaliação é discutida na Seção 14.3.2. Mais detalhes a respeito da metodologia podem ser encontrados em (Sakiyama, 2023).

14.3.1 Aquisição de dados

Em 2022, como parte de um movimento em direção à transparência administrativa, o Supremo Tribunal de Justiça (STJ) lançou a plataforma Dados Abertos[3]. Este domínio público destina-se ao compartilhamento de decisões judiciais de diversos tribunais do Brasil, sendo estas julgadas por ministros do STJ. Além de promover a transparência, a plataforma foi criada visando reduzir o uso de *web scrappers* no portal oficial do STJ e estimular a pesquisa e o desenvolvimento de ferramentas baseadas em inteligência artificial que dependem da disponibilidade de dados.

As ementas coletadas estão classificadas em seções. Estas representam as seções do STJ e se diferenciam pelos temas jurídicos julgados. Por exemplo: impostos (primeira seção), comércio (segunda seção) e crimes no geral (terceira seção) (STJ, 24DC). Desta forma, os documentos obtidos abrangem uma ampla variedade de temas do contexto jurídico brasileiro. Para o desenvolvimento desta pesquisa foram coletados um total de 726.384 documentos da plataforma (a coleta ocorreu em agosto de 2022) e foram analisadas as ementas dos processos obtidos.

14.3.2 Geração textual

Nesta Seção, é apresentada a metodologia utilizada para geração das verbetações e sua avaliação.

14.3.2.1 Preparação de exemplos para geração de texto

Para a preparação de entradas para o modelo generativo escolhido, verbetações e parágrafos enumerados foram separados, por meio da identificação de frases iniciadas por letras maiúsculas. Assim, as verbetações originais extraídas passam a compor o conjunto de referência (conjunto ouro ou *gold standard*) a ser usado para avaliar a qualidade das verbetações geradas pelas estratégias estudadas.

Feito o processamento adicional, o *corpus* final foi dividido em três conjuntos: treino (70%), validação (10%) e teste (20%). Tal divisão foi feita de forma estratificada, para preservar as proporções das origens (seções de julgamento) dos documentos. Este procedimento foi feito visando uma avaliação utilizando a técnica de validação cruzada (mais detalhes no Capítulo Avaliação de tecnologias de linguagem).

A fim de ilustrar aos leitores os tamanhos dos parágrafos enumerados e das verbetações, na Tabela 14.1 é sumarizada a análise estatística descritiva da variação dos tamanhos dos *tokens* presentes no conjunto de treinamento. Dentre as medidas foram calculadas a média, o desvio padrão e os 3 quartis: Q1, Q2 e Q3. O conjunto em questão possui cerca de 1,6 milhão de *tokens*. Os parágrafos enumerados têm, em média, 203,26 *tokens*, enquanto as verbetações têm, em média, 55,84 *tokens*. Os valores do desvio padrão revelam que os tamanhos dos parágrafos enumerados e das verbetações variam consideravelmente.

[3] https://dadosabertos.web.stj.jus.br/

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Tabela 14.1: Variação dos tamanhos dos parágrafos enumerados e das verbetações com base em medidas da estatística descritiva para *tokens* do conjunto de treino (78.375 exemplos).

	Média	**Desvio Padrão**	**Q1**	**Q2**	**Q3**
Parágrafos enumerados	203,262	183,332	92	155	253
Verbetação	55,842	32,136	34	49	69

14.3.2.2 Transformers para geração de verbetações

Como mencionado anteriormente, as verbetações podem conter termos que não estão presentes no corpo da ementa (parágrafos enumerados). De fato, ao analisar os documentos do conjunto de validação, observou-se que apenas cerca de 10% dos termos das verbetações estão contidos no restante do texto. Sendo assim, optou-se por abordar a geração de verbetações como sendo uma geração de sequência-a-sequência (ou texto-para-texto). Utilizou-se o conteúdo dos parágrafos enumerados como entrada para os modelos Transformers, enquanto que as verbetações originais foram consideradas como as saídas desejadas.

Estudos anteriores (Carmo et al., 2020b; Rosa et al., 2021a; Souza et al., 2020a) apontam que modelos pré-treinados para a linguagem da tarefa tendem a superar os modelos multilíngues nas mesmas tarefas. Em experimentos detalhados em (Sakiyama, 2023), o modelo PTT5 (Carmo et al., 2020b)[4], ajustado para português brasileiro, destacou-se como o mais eficaz, e seus resultados serão reportados aqui. Este modelo foi pré-treinado no *corpus* brWaC e foi utilizada sua versão base (com 220M de parâmetros). Tendo treinado o modelo PTT5, a geração de verbetações foi feita usando uma simples decodificação gulosa (que consiste sempre em escolher o *token* mais provável na geração).

14.3.2.3 Avaliação dos textos gerados

A avaliação dos textos gerados pelo modelo de língua PTT5 foi feita empregando-se a metodologia descrita no Capítulo Avaliação de tecnologias de linguagem. Utilizou-se a técnica de amostragem *holdout* para validação e a métrica BLEU (*Bilingual Evaluation Under-study*) para quantificar a qualidade dos textos gerados.

A fim de aprofundar a avaliação, foram realizadas outras análises comparativas entre o texto gerado e as entradas originais. Por exemplo, os tamanhos das entradas geradas pelo melhor modelo foram comparados com os das entradas originais, bem como a quantidade de palavras (*tokens*) copiadas e geradas. Palavras 'novas' (ou geradas) são aquelas que não aparecem nos parágrafos enumerados, utilizados na geração de verbetações.

14.4 Avaliação utilizando RI

Para a avaliação final das verbetações geradas, elas foram concatenadas aos seus documentos originais para simular um caso de uso real em uma tarefa de Recuperação de Informação (RI). Foram empregadas tanto a decodificação gulosa quanto a decodificação por amostragem na geração das verbetações nesta avaliação. Mais detalhes a respeito da avaliação experimental podem ser encontrados em (Sakiyama, 2023).

[4] https://huggingface.co/unicamp-dl/ptt5-base-portuguese-vocab

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

14.4.1 Formulação da tarefa

Os documentos apresentados na Seção 14.3.1, juntamente com o texto da ementa, contêm metadados úteis, entre eles, o tema da decisão. Tais temas, ou temas de recursos repetitivos, consistem em identificadores únicos que correspondem a questões jurídicas comuns. Cada documento pode estar associado a mais de um tema, e há mais de mil temas distintos listados no STJ[5]. Exemplos desses temas estão ilustrados no Quadro 14.3.

Quadro 14.3: Exemplos de temas de recursos repetitivos, listados pelo STJ.

Tema 192: Incidência da pensão alimentícia sobre o décimo terceiro salário e o terço constitucional de férias .
Tema 312: Prazo para restituição ao consorciado desistente das parcelas pagas ao grupo de consórcio .
Tema 1166: Natureza material do crime de apropriação indébita previdenciária e sua consumação com a constituição definitiva do crédito tributário na via administrativa.

Com base nos temas, foi utilizada a definição binária de relevância para formular a seguinte tarefa de RI: dado um documento de consulta Q, os documentos relevantes R devem pertencer ao mesmo tema de Q. Esta formulação simula o caso de uso em que um advogado busca processos semelhantes a um caso base, por exemplo. A formulação apresentada é semelhante à utilizada por Ostendorff et al. (2021), na qual foram criados pares de relevância (documento de consulta e documento relevante) usando decisões da Suprema Corte dos Estados Unidos. As decisões foram categorizadas por livros de casos ou categorias pré-definidas.

Em resumo, na avaliação proposta, ementas são usadas para buscar outras ementas. Para imitar o cenário real nos experimentos, foram empregadas ementas completas (incluindo verbetações e parágrafos enumerados). Foram utilizadas ementas completas em todos os experimentos, exceto o primeiro, que será discutido na próxima Seção.

Entre as ementas coletadas, apenas 801 possuíam anotações de tema, com 99,3% apresentando valores ausentes. Estas ementas foram removidas do conjunto de treinamento dos geradores de texto e utilizadas para compor pares de consulta e documentos relevantes nos experimentos com RI. Foram filtrados temas que apareceram pelo menos duas vezes, resultando em 482 consultas.

A Figura 14.1 mostra uma visualização do número de documentos relevantes (com o mesmo tema) por documento de consulta. Observa-se que a maioria das consultas possui entre 1 e 3 documentos relevantes. No caso mais extremo, uma consulta pode ter até 10 documentos relevantes.

Para criar o *corpus* final de busca para a tarefa mencionada, foram combinados os 801 documentos com tema ao conjunto de teste discutido na Seção 14.3.2.1, totalizando 23.194 documentos. Esta combinação foi feita visando tornar a tarefa de RI mais desafiadora, pois a mesma pode incluir falsos negativos ao *corpus* de busca, que acabam prejudicando as métricas de RI. Falsos negativos são documentos que têm o mesmo tema da consulta, mas são considerados irrelevantes por não possuírem anotação de tema (dado ausente).

14.4.2 Experimentos realizados

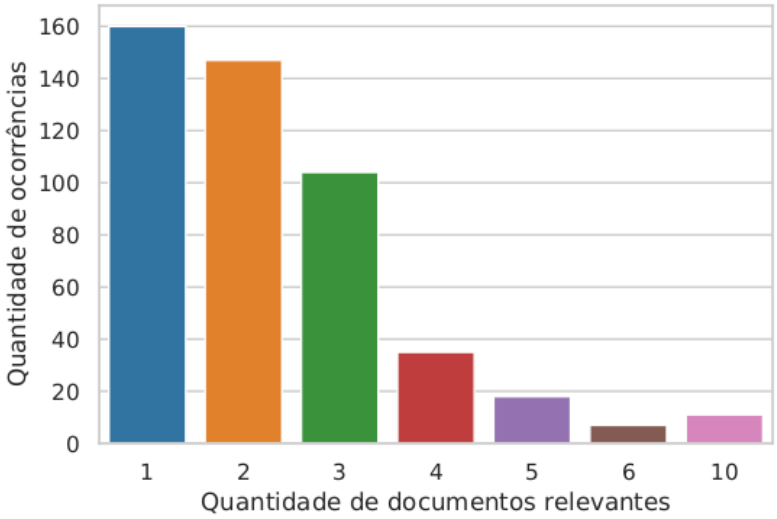
A avaliação baseada em RI foi conduzida por meio de dois experimentos distintos. Todos os experimentos utilizaram métodos de decodificação simples (decodificação gulosa) e mé-

[5] https://scon.stj.jus.br/SCON/recrep/

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Figura 14.1: Visualização do número de documentos relevantes (de mesmo tema) por documento de consulta.



todos não determinísticos (decodificação com amostragem). Seus detalhes serão descritos a seguir.

1. **Comparação de documentos com e sem verbetações**: a influência das verbetações na tarefa de RI foi avaliada, analisando-se métricas de RI para buscas, usando documentos com e sem suas verbetações originais, tanto para documentos de consulta quanto para documentos do *corpus* de busca. Este experimento teve como objetivo validar o papel das verbetações no contexto de uma tarefa de RI.

2. **Avaliação das verbetações geradas**: neste experimento, foi investigado o efeito de usar verbetações geradas no lugar das originais. O texto gerado foi concatenado ao início de ambos os documentos de consulta e dos documentos do *corpus* de busca. O objetivo deste experimento é verificar se as verbetações artificiais melhoram as métricas de RI em comparação com a ausência de verbetações. Para este experimento, foram usadas verbetações geradas com decodificação gulosa.

14.4.3 Métodos de RI avaliados

Dois métodos tradicionais de RI para a tarefa descrita anteriormente: TF-IDF e BM25 (Robertson; Walker, 1999) foram considerados para a avaliação. Tais métodos são amplamente utilizados em sistemas de busca populares (como Lucene[6]), e são frequentemente adotados como *base* em comparações de desempenho de algoritmos (Pradeep et al., 2020; Rosa et al., 2021b). Além disto, conforme investigado por estudos anteriores (Lima et al., 2021; Mandal et al., 2021), métodos de representação esparsa tendem a apresentar melhor desempenho em tarefas similares no domínio jurídico. Tendo apresentado os modelos investigados, a seguir serão apresentados mais detalhes sobre suas execuções.

[6] https://lucene.apache.org/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

14.4.4 Ordenação dos documentos

Ao utilizar o método TF-IDF para realizar as buscas, calculou-se a similaridade de cosseno entre consultas e documentos e os documentos do *corpus* foram ordenados a partir de suas semelhanças em ordem decrescente. Já o método BM25 utiliza o Princípio de Classificação por Probabilidade (*Probability Ranking Principle*) para estimar a relevância de um documento para uma consulta (Crestani et al., 1998). Logo, os documentos foram ordenados com base em suas relevâncias estimadas.

14.4.5 Métricas avaliadas

Para avaliar de forma quantitativa o desempenho da busca e, em adição, a qualidade das verbetações geradas, foram computadas métricas tradicionais de RI. As métricas utilizadas foram: *Mean Reciprocal Rank* (MRR), *Normalized Discounted Cumulative Gain* (NDCG), Revocação, Precisão e *Mean Average Precision* (MAP).

Utilizou-se os 10 primeiros documentos ordenados para o cálculo das métricas mencionadas. Este limite foi escolhido para simular um cenário restritivo, em que o usuário investiga somente até 10 documentos. Conforme (Russell-Rose et al., 2018), profissionais do Direito geralmente aceitam analisar até 50 documentos em suas consultas. Portanto, este estudo avalia um cenário ainda mais desafiador do que o mencionado pelos autores.

14.5 Resultados e discussões

Esta Seção apresenta e discute os resultados obtidos na geração de verbetações e sua avaliação usando RI.

14.5.1 Comparação entre verbetações originais e geradas

Como apontado na Seção 14.3.2.2, as verbetações foram geradas usando o modelo PTT5, o qual obteve uma pontuação BLEU de 38,607 no conjunto de teste preparado. Nesta Seção, serão mostrados exemplos (Quadros 14.4, 14.5 e 14.6) dos textos gerados e os mesmos também serão comparados com as verbetações originais.

Ao avaliá-las qualitativamente (Quadro 14.4), observa-se que, com pontuações BLEU próximas a 40%, as verbetações geradas não apresentam erros ortográficos ou léxicos. Elas simulam o estilo de escrita das verbetações originais e são bastante similares a elas. Além disto, fica evidente que as verbetações geradas não são cópias perfeitas das originais, uma vez que nem todos os termos das verbetações originais estão nas geradas. Assim, os resultados da geração textual com Transformers foram muito positivos, se consideramos o tamanho do conjunto de treinamento utilizado neste caso (menos de 100 mil exemplos).

Quadro 14.4: Exemplos de verbetações geradas, destacando-se termos em comum (idênticos) entre as verbetações.

Original: AGRAVO REGIMENTAL NO HABEAS CORPUS . TRÁFICO DE DROGAS. PRISÃO PREVENTIVA. PERICULUM LIBERTATIS. CRISE MUNDIAL DE COVID-19. NÃO DEMONSTRADA A EXISTÊNCIA DE RISCO À INTEGRIDADE FÍSICA DO ENCARCERADO. AGRAVO REGIMENTAL NÃO PROVIDO.
Gerada: AGRAVO REGIMENTAL NO HABEAS CORPUS. TRÁFICO DE DROGAS. PRISÃO PREVENTIVA. PERICULUM LIBERTATIS. AGRAVO REGIMENTAL NÃO PROVIDO.
Original: ADMINISTRATIVO. RECURSO ESPECIAL REPETITIVO. SERVIDOR PÚBLICO.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

ARTIGO 46, CAPUT, DA LEI N. 8.112/1990. TESE DEFINIDA NO TEMA 531/STJ. AUSÊNCIA DE ALCANCE NOS CASOS DE PAGAMENTO INDEVIDO DECORRENTE DE ERRO DE CÁLCULO OU OPERACIONAL DA ADMINISTRAÇÃO PÚBLICA. POSSIBILIDADE DE DEVOLUÇÃO. SALVO INEQUÍVOCA PRESENÇA DA BOA-FÉ OBJETIVA.
Gerada: ADMINISTRATIVO. RECURSO ESPECIAL REPETITIVO. TEMA 531/STJ. ERRO DE CÁLCULO OU OPERACIONAL. PAGAMENTO INDEVIDO. BOA-FÉ OBJETIVA. DEVOLUÇÃO. DESCABIMENTO.

Quadro 14.5: Mais exemplos de verbetações geradas. Termos em comum foram destacados em verde. Parafrases ou sinônimos foram destacados em amarelo.

Original: CONFLITO NEGATIVO DE COMPETÊNCIA. JUSTIÇA COMUM ESTADUAL E JUSTIÇA DO TRABALHO. AÇÃO DE OBRIGAÇÃO DE FAZER C/C COBRANÇA. CONTRIBUIÇÃO SINDICAL DE SERVIDORES PÚBLICOS MUNICIPAIS. ART. 114, III, DA CF/1988 COM REDAÇÃO DADA PELA EC 45/2004. COMPETÊNCIA DA JUSTIÇA DO TRABALHO. PRECEDENTES.

Gerada: CONFLITO NEGATIVO DE COMPETÊNCIA CONTRIBUIÇÃO SINDICAL. AGENTE PÚBLICO ESTATUTÁRIO. ART. 114, III, DA CONSTITUIÇÃO FEDERAL DE 1988. EMENDA CONSTITUCIONAL 45/2004. COMPETÊNCIA DA JUSTIÇA DO TRABALHO.

Quadro 14.6: Exemplo de verbetação em que são introduzidos termos não presentes na verbetação original.

Original: TRIBUTÁRIO E PROCESSUAL CIVIL. RECURSO ESPECIAL. INTERPOSIÇÃO DE RECURSO, TEMPESTIVAMENTE, POR MEIO DE FAC-SIMILE. AUSÊNCIA DE APRESENTAÇÃO DA PETIÇÃO ORIGINAL, NO PRAZO PREVISTO NO ART. 2º DA LEI 9.800/99. RECURSO ESPECIAL INTERPOSTO VIA E-MAIL. INADMISSIBILIDADE. NÃO EQUIPARAÇÃO AO FAC-SIMILE. PRECEDENTES DO STJ. RECURSO ESPECIAL NÃO CONHECIDO.

Gerada: : TRIBUTÁRIO E PROCESSUAL CIVIL. RECURSO ESPECIAL, POR FAC-SIMILE. INTERPOSIÇÃO, NA ORIGEM, DE AGRAVO DE INSTRUMENTO, POR INTERPOSIÇÃO DE RECURSO POR MEIO DE FAC-SIMILE, ORIGINAIS APRESENTADOS DENTRO DO PRAZO LEGAL. IMPOSSIBILIDADE. ART. 2o DA LEI 9.800/99. RECURSO INTEMPESTIVO. PEÇA INCOMPLETA. NÃO APRESENTAÇÃO. PRESCRIÇÃO. PRAZO PARA INTERPOSIÇÃO DE RECURSO. PRECEDENTES DO STJ. ART. 543-C DO CPC/73. TERMO INICIAL. SÚMULA 83/STJ. AGRAVO INTERNO IMPROVIDO.

Dando prosseguimento às comparações, ao analisarmos o Quadro 14.5, é possível observar uma característica interessante. Houve casos em que o PTT5 gerou paráfrases ou expandiu siglas presentes na verbetação original. Tal fato é facilmente perceptível no Quadro 14.5, observando os termos "SERVIDOR PÚBLICO" e "AGENTE PÚBLICO". Por outro lado, siglas como "CF/1998" foram expandidas para "CONSTITUIÇÃO FEDERAL DE 1998", assim como "EC" foi expandida para "EMENDA CONSTITUCIONAL". Os comportamentos observados são justificados pelo funcionamento de modelos de linguagem baseados em Transformers, uma vez que, durante a geração de texto, os mesmos tendem a gerar *tokens* que aparecem em contextos semelhantes.

Como última discussão de exemplos, o Quadro 14.6 ilustra potenciais alucinações do modelo avaliado. Como o modelo PTT5 está na família de Transformers generativos, o mesmo não está isento de alucinações, como é o caso dos mais conhecidos (GPT's, por

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

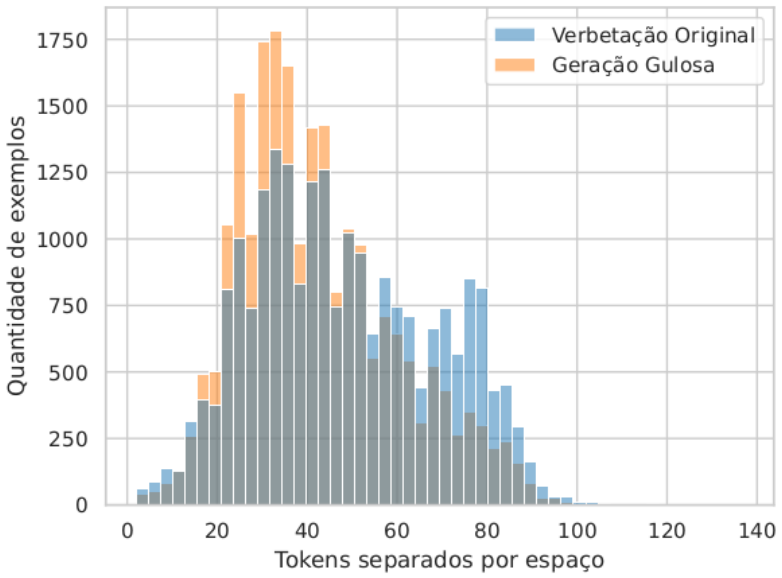
©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

exemplo). Dentro do contexto desta aplicação e focando no Quadro 14.6, é possível notar a presença de termos e citações de artigos que não constam no texto original.

Embora esses termos possam surgir em contextos parecidos, não se deve ignorar a possibilidade de que tenham sido gerados por alucinações do Transformer. Sem o conhecimento específico do domínio, não é possível assegurar que tais termos "novos" estejam realmente associados ao texto de entrada. Pode-se apenas supor que eles ocorrem em contextos semelhantes. Assim, existe sim o risco de adicionar textos factualmente incorretos às verbetações geradas, devido ao processo generativo.

A investigação da métrica BLEU permite estimar a quantidade de termos em comum entre as verbetações de referências e as geradas, ao comparar a interseção de n-gramas em ambas. Como a métrica é computada considerando o *corpus* de teste inteiro, ela apresenta uma boa estimativa do casamento de n-gramas entre todas as verbetações originais e geradas. Para aprofundar a comparação entre as verbetações, na Figura 14.2 é apresentada uma comparação entre o número de *tokens* das verbetações originais e das verbetações geradas com decodificação gulosa.

Figura 14.2: Histogramas de comparação entre a quantidade de *tokens* das verbetações originais e das verbetações geradas.



Embora as distribuições apresentadas pelos dois histogramas sejam semelhantes, fica evidente que as verbetações geradas tendem a ter uma concentração maior de exemplos abaixo de 60 *tokens*. Além disto, ao analisar todas a verbetações mostradas anteriormente (Quadros 14.2, 14.4, 14.5 e 14.6) fica evidente que as mesmas tendem a ter menos palavras do que as referências. Assim, podemos concluir que as verbetações geradas tendem a ser mais curtas que as originais.

14.5.2 Experimentos com RI

A Seção anterior apresentou comparações qualitativas entre as verbetações geradas pelo método PTT5 e as originais. Resta verificar como as verbetações originais e geradas se comportam na avaliação, usando um caso real de RI descrito na Seção 14.4. Vale observar

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

que as métricas mostradas em todas as tabelas desta Seção consistem em médias obtidas para cada métrica considerando as 482 consultas.

14.5.2.1 Comparação de documentos com e sem verbetações

A Tabela 14.2 mostra as métricas obtidas ao realizar o primeiro experimento com documentos com e sem as verbetações originais. Observa-se que o uso de verbetações melhora todas as métricas avaliadas, com maiores ganhos para o método BM25. Tais resultados eram esperados, pois a adição de termos discriminatórios nos documentos beneficia os métodos esparsos. Assim, os resultados indicam que o uso de palavras-chave tem um impacto positivo na tarefa de busca. O maior ganho foi observado na métrica R@10 (4,2 pontos percentuais) para TF-IDF e na MAP@10 (6,1 pontos percentuais) para BM25.

Tabela 14.2: Métricas de RI obtidas para cada modelo avaliado, avaliando a influência das verbetações originais. '*' indicam diferenças estatisticamente significativas, de acordo com um teste-T pareado (p-valor $< 0,05$).

	TF-IDF				
	MRR@10	**NDCG@10**	**R@10**	**P@10**	**MAP@10**
Sem verbetações	0,806	0,745	0,790	0,191	0,691
Verbetações originais	0,825*	0,780*	0,832*	0,201*	0,729*
		BM25			
Sem verbetações	0,819	0,805	0,878	0,212	0,754
Verbetações originais	0,879*	0,858*	0,916*	0,222*	0,815*

14.5.2.2 Avaliação das verbetações geradas

Iniciando os experimentos com as verbetações geradas, a Tabela 14.3 apresenta a avaliação das verbetações geradas com decodificação gulosa. Observa-se que, ao adicionar as verbetações artificiais aos documentos, houve melhorias em quase todas as métricas para ambos os modelos avaliados. Para o método TF-IDF, a diferença foi estatisticamente significativa em todas as métricas avaliadas.

Tabela 14.3: Métricas obtidas para cada modelo de recuperação, ao utilizar documentos sem verbetação e utilizando as verbetações geradas com decodificação gulosa. '*' indicam diferenças estatisticamente significativas, de acordo com um teste-T pareado (p-valor $< 0,05$).

	TF-IDF				
	MRR@10	**NDCG@10**	**R@10**	**P@10**	**MAP@10**
Sem verbetações	0,806	0,745	0,790	0,191	0,691
Geradas	0,822*	0,770*	0,815*	0,196*	0,718*
		BM25			
Sem verbetações	0,819	0,805	0,878	0,212	0,754
Geradas	0,854*	0,819*	0,877	0,211	0,768

Quanto ao BM25, houve ganhos em quase todas as métricas. No entanto, a diferença foi estatisticamente significativa apenas para duas métricas (MRR@10 e NDCG@10). Embora

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

tenham ocorrido melhorias para o TF-IDF, os resultados sugerem que o modelo BM25 foi mais sensível a falsos positivos e falsos negativos, introduzidos no *corpus* de busca por verbetações ruidosas geradas pelo PTT5. Assim, houve quase nenhuma alteração nas métricas baseadas em revocação e precisão.

As métricas analisadas mostram que ambos os modelos tiveram desempenho inferior em relação ao experimento anterior utilizando as verbetações originais (Tabela 14.2). Este resultado já era esperado, pois os resultados com verbetações originais atuam como um limite superior para este experimento. Mesmo com este cenário, o uso de verbetações artificiais melhorou ambos os métodos de RI em comparação à ausência de verbetação, evidenciando o potencial da geração automática de verbetações.

14.6 Conclusão

Nesta aplicação, utilizou-se modelos Transformers para a automação da escrita de verbetações presentes em ementas de processos judiciais brasileiros. O Transformer usado, PTT5, alcançou uma pontuação BLEU superior a 37% na geração textual. Em outras palavras, ao comparar as verbetações geradas com as originais, foi possível comprovar que as verbetações geradas foram capazes de reproduzir com fidelidade as características principais das verbetações escritas por especialistas.

Além disso, notou-se que as verbetações geradas capturaram com sucesso o estilo linguístico das verbetações originais. As frases curtas e objetivas, presentes nas verbetações, também são características das geradas. O modelo também foi capaz de reconhecer e expandir siglas comuns (como "CF" e "EC").

Como a geração não é perfeita, também foi identificada a ocorrência de alucinações do modelo generativo usado e seus potenciais malefícios. Verificou-se também que as verbetações geradas tenderam a conter menos termos que as originais. Mesmo com estas ressalvas, os resultados qualitativos e quantitativos são evidências de que a geração automática de textos jurídicos é de fato possível.

Note que, em virtude das ressalvas (principalmente das alucinações), a automação completa da escrita de verbetações não é recomendada. Contudo, a metodologia apresentada aqui pode ser útil para auxiliar na escrita de ementas. Como exemplo, os Transformers podem ser incorporados a sistemas existentes, gerando automaticamente sugestões de verbetações (ou *templates*), que podem, então, serem corrigidas por especialistas, conforme a necessidade.

Considerando que o setor judiciário brasileiro convive com um crescente volume de processos para análise e que todo processo demanda tempo e recursos financeiros, qualquer esforço rumo à automatização é sem dúvida muito bem-vindo. Sendo assim, acredita-se que os resultados aqui apresentados possam estimular outros trabalhos no domínio jurídico, mostrando o potencial de modelos generativos em domínios com grandes demandas sobre recursos textuais.

Agradecimentos

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Código de Financiamento 001. Os autores agradecem também ao CEMEAI, pelo acesso ao *cluster* Euler, para a realização de experimentos.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Capítulo 15

PLN e Humanidades Digitais

Renata Vieira
Helena Freire Cameron
Fernanda Olival
Fatima Farrica
Maria José Bocorny Finatto
Ana Paula Banza
Ana Sofia Ribeiro
Cassia Trojahn

Publicado em: 13/03/2024
Atualizado em: 20/11/2024

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

15.1 Introdução

A área de humanidades digitais (HD) tem ganhado força e adeptos nas últimas décadas, em paralelo com o desenvolvimento de ferramentas digitais que ampliam as possibilidades de armazenamento, acesso e processamento de dados. Essas capacidades estendem os horizontes de atuação de pesquisadores, permitindo a captura, organização e análise de um volume muito grande de dados. Com esta interação, que pode ser classificada de tecnológica, as humanidades ganham visibilidade, atravessam fronteiras disciplinares e enfrentam desafios sem precedentes.

Este capítulo apresenta as bases comuns para projetos na área de HD, relacionados ao Processamento de Linguagem Natural (PLN), em particular sobre fontes textuais com grafia pré-contemporâneas, o que imprime uma maior complexidade no processamento. Além disso, nosso enfoque é a relação entre as HD e o PLN no contexto da língua portuguesa, entendendo que o alcance da área de HD vai muito além da língua, e do escopo das questões e projetos aqui mencionados.

Na área de HD, relativamente aos trabalhos baseados em fontes textuais, encontramos uma grande variação, tanto nos períodos históricos das fontes, no seu suporte (manuscritos em papel, impressos, fotografados etc.), como no seu estágio de digitalização, que pode variar entre imagens digitais, textos em PDF e textos digitalizados em outros formatos. Todas essas variações adicionam esforços extras de processamento. Desta forma, apresentaremos um panorama geral e discutiremos os requisitos de preparação e organização das fontes, objetos de análise que depois de transcritas e digitalizadas, podem ser submetidas a processamentos mais avançados. O objetivo é mostrar não apenas como o PLN é útil e relevante nesse domínio, mas também como a área de HD é rica em despertar novas questões para o desenvolvimento do PLN.

Este capítulo está organizado da seguinte maneira. Na Seção 15.2 apresentamos os tópicos e os desafios relacionados à preparação das fontes, desde a digitalização até a

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

anotação. Na Seção 15.3, discutimos os processos de transformação de fontes em dados, nos quais várias tarefas conhecidas de PLN (discutidas em outros capítulos deste livro) podem ser empregadas. Esses processos de transformação de textos em dados organizados possibilitam ao pesquisador realizar análises diferenciadas sobre textos, por exemplo, ao agrupar tipos específicos de informação, ou rapidamente quantificar fenômenos observáveis. Na Seção 15.4, apresentamos alguns projetos relacionados às HD, em especial aqueles desenvolvidos pelas autoras. São projetos que apresentam diferentes fontes e objetivos de pesquisa, mas nos quais o PLN se faz presente em vários níveis, desde a preparação e a anotação até a extração, organização e partilha da informação. Por fim, apresentamos nossas considerações finais na Seção 15.5.

15.2 Preparação das fontes

Trabalhos em HD que se baseiam em textos podem se beneficiar das atuais técnicas de Inteligência Artificial (IA) e PLN para um melhor acesso às fontes. Os textos em HD podem ter relevância por suas características históricas e literárias, mas também podem estar relacionados a estudos sociais ou de outra natureza. Na verdade, é difícil distinguir os limites das HD, mas se pode dizer que a área está ligada ao emprego e à compreensão de novas maneiras de se desenvolver pesquisa, com base em recursos digitais e computacionais. Em relação ao processamento da língua, essa influência tecnológica se observa em uma variedade de processos referentes ao tratamento das fontes e sua informação, tais como:

- **Digitalização de documentos**: preparação no nível mais básico, para o tratamento computacional, e assim possibilitar a sua leitura por meio de programação.
- **Adição de metadados**: inserção de níveis de informação extra-textuais, que podem ser importantes para a estruturação e o armazenamento, e também para o acréscimo de interpretações sobre o conteúdo de uma coleção de interesse. Os metadados podem trazer informações de contexto, como autores, datas, locais de produção, origem, ou ainda informações como volume, páginas, etc. Ou podem ser de análise linguística: morfológica, sintática ou semântica (por exemplo, identificando entidades ou eventos).
- **Normalização**: a normalização para uma língua padrão e contemporânea pode ser necessária em situações como fontes históricas, textos de redes sociais que usam muitas abreviações, símbolos associados à emoções, ou comentários como *hashtags*, etc.
- **Anotação de *corpora***: muitas vezes o estudo de uma fonte ou *corpus* necessita a adição de informações extras sobre o registro, que são representadas como metadados (mencionado acima). Os processos podem ser automáticos ou manuais. No entanto, geralmente o desenvolvimento de um processo automático requer um processo de anotação manual anterior, quer para o seu desenvolvimento (treino de algoritmos de aprendizado), quer para a avaliação da correção da anotação realizada por máquina.

Os métodos usuais de tratamento textual, desenvolvidos em pesquisas de PLN e IA, podem requerer adaptação a diferentes necessidades de investigação, estilo textual, objetivos da pesquisa e também ao uso pretendido e seus usuários. Idealmente, são necessárias novas interfaces para que os métodos desenvolvidos sejam usados de forma facilitada e

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

fora do contexto das estruturas de programação. Apresentaremos, a seguir, esses passos iniciais para projetos de HD que lidam com fontes textuais e, posteriormente, discutiremos sobre as técnicas de PLN.

15.2.1 Digitalização

Muitos projetos na área de HD necessitam lidar com a digitalização de manuscritos originais. A área de paleografia, em particular, lida com a leitura e transcrição de versões manuscritas para um suporte atual. Mesmo registros em suportes muito antigos, como a escrita cuneiforme, podem ser transcodificados para o digital (Liu et al., 2016), e a partir daí, podem passar por processos computacionais.

A paleografia, deste modo, também se desenvolveu de habilidade tradicional do intelecto, e manual, para uma prática digital (Ackel, 2021). Algumas ferramentas de transcrição, como por exemplo o Transkribus[1] ajudam na digitalização de manuscritos (Kahle et al., 2017), através de um sistema de algoritmos HTR (*Handwritten Text Recognition*).

Alguns documentos impressos podem necessitar a aplicação de tecnologia OCR (*Optical Character Recognition*). A qualidade da saída de sistemas OCR varia muito dependendo do formato e da qualidade da entrada. É um problema básico, mas não totalmente resolvido (Strien et al., 2020). Por exemplo, nas primeiras levas de digitalização de acervos arquivísticos e bibliotecários, muitas digitalizações foram feitas em formatos de imagem, sem a preocupação de possibilitar a pesquisa dos seus conteúdos de texto. Por meio de diferentes ferramentas de transcrição automática com OCR, como o eScriptorum[2], podemos fazer essa transcrição de imagem para texto de forma automática.

Muitas vezes, as saídas de OCR precisam passar por processamento extra de correção, por vezes manual, para tornar a fonte adequada para as próximas etapas. Esse processamento se faz necessário para que outras ferramentas, como as de anotação de texto, tenham maior eficácia na aprendizagem dos *tokens* anotados. É preciso preparar os textos, por exemplo, retirando números de páginas, textos de cabeçalhos ou rodapés, e evitar a translineação/hifenização de palavras. Se o texto a ser anotado for composto por um conjunto de textos, pode ser necessário que estejam devidamente separados. Outro problema que pode emergir são os textos organizados em mais de uma coluna nas páginas, pois será preciso ordená-los numa segmentação textual contínua. Essas etapas devem ser consideradas em projetos de HD envolvendo textos e PLN e é, portanto, necessário antever-se o tempo e recursos necessários para o seu desenvolvimento.

15.2.2 Metadados

Uma vez digitalizada a fonte ou *corpus* de estudo, é necessário pensar na organização dos seus metadados. O material digital deve ser bem descrito, ou seja, deve conter a informação sobre a qual acervo pertence, identificar unicamente cada arquivo e, quando pertinente, associar autoria, data e outros elementos pertinentes. Os metadados também podem descrever a estrutura do documento, quando necessário, identificar volumes, capítulos, páginas, por exemplo. É importante, ainda, separar os metadados dos dados originais. Para materiais transcritos, por exemplo, devem ser identificados os cabeçalhos, notas de rodapé, numeração de páginas ou comentários adicionados aos originais. Esta organização e distinção dos elementos extra-textuais são essenciais para garantir o bom

[1] https://www.transkribus.org
[2] https://www.resilience-ri.eu/blog/resilience-tool-escriptorium/

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

processamento posteriormente. Os metadados são essenciais para conectar uma fonte a outras e as conectar aos dados abertos ligados (Nair; Jeeven, 2004).

Há vários padrões de metadados bastante difundidos. Por exemplo, a Text Encoding Initiative (TEI) é um consórcio que desenvolve e mantém coletivamente um padrão para a representação de textos em formato digital. O seu principal resultado é um conjunto de diretrizes que especifica métodos de codificação para textos legíveis por máquina, principalmente em ciências humanas, ciências sociais e linguística[3]. Outro exemplo é o grupo de trabalho sobre "Publicação Sustentável de Metadados", ativo na rede europeia da *Digital Research Infrastructure for the Arts and Humanities* (DARIAH) desde 2016. Esse grupo tem como objetivo apresentar informações sobre a reprodução e o compartilhamento de metadados entre institutos e pesquisadores no campo das ciências humanas e sociais.

Vários trabalhos trataram da representação de metadados por meio de vocabulários semânticos, para descrever quer recursos digitais, como vídeos, imagens, páginas web, quer recursos físicos, como livros ou obras de arte. Esse vocabulários podem ser usados para fornecer interoperabilidade para o uso de metadados na nuvem de dados ligados. Exemplos são o Dublin Core[4], VoID (Vocabulary of Interlinked Datasets[5]), schema.org, DCAT (Data Catalog Vocabulary[6]) e PROV-O (Provenance Ontology[7]). Em particular, o DCAT é um vocabulário projetado para facilitar a interoperabilidade entre catálogos de dados publicados na Web. Permite a descrição de conjuntos de dados e serviços, usando um modelo e um vocabulário padrão que facilitam o consumo e a agregação de metadados de vários catálogos. Os metadados agregados do DCAT podem servir como um arquivo de manifesto[8] como parte do processo de preservação digital.

15.2.3 Normalização

Em HD, as fontes textuais, quer manuscritas, quer impressas, sendo de uma época anterior ao estágio atual da língua, apresentam variações ortográficas ou morfossintáticas, não só em relação ao padrão atual como também dentro da mesma época.

A variação gráfica no século XVIII, por exemplo, é muito expressiva, quer em textos manuscritos, quer em impressos, especialmente ao nível da representação das vogais e de ditongos nasais, por exemplo "am-ão", em "fizerão", ou "oens-ões", em "embarcaçoens", etc. Também as consoantes duplas como "ll-l", em "delles", "tt-t", em "Cratto", entre outras, têm uma elevada frequência nesses textos. Note-se, também, a existência do fenômeno das chamadas grafias cultas ou pseudo-etimológicas, que já não integram o cânone atual da língua portuguesa. Estas podem trazer uma maior complexidade ao processamento lexical, como "th-t", em "athé", "ch-c" em "Christo" e "y-i", por exemplo na palavra "Rey". A palavra "coro", que frequentemente é registada como "choro" em textos do século XVIII, é um exemplo da complexidade da variação gráfica deste período.

Do ponto de vista do leitor, a abordagem destes documentos requer um forte conhecimento do contexto para a sua desambiguação e exige, frequentemente, conhecimentos linguísticos e históricos para que possa desvendar o conteúdo de forma adequada. A variação gráfica tem uma elevada incidência nos documentos desta época. Grande parte

[3] https://tei-c.org
[4] https://www.iso.org/standard/71339.html
[5] http://vocab.deri.ie/
[6] https://www.w3.org/TR/vocab-dcat-3/
[7] https://www.w3.org/TR/prov-o/
[8] *Manifest file*, arquivo que contém metadados para um grupo de arquivos que são parte de um conjunto ou de uma unidade coerente.

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

das palavras tem pelo menos uma variante gráfica e algumas palavras têm múltiplas num mesmo texto. A título ilustrativo, veja-se a palavra "circunstância", com nove variantes gráficas recolhidas num mesmo *corpus* textual (Cameron et al., 2023).

Do ponto de vista do PLN, a variação, quer pela sua grande incidência, quer pela sua imprevisibilidade, vem trazer uma perturbação acrescida. Variantes de uma mesma palavra são tratadas como se fossem unidades lexicais diferentes, diminuindo a eficácia do processamento. Para que se possa processar o texto de forma mais célere e eficaz, frequentemente têm de ser desenvolvidas tarefas intermediárias de normalização, quase sempre de forma manual. Unificam-se as variantes gráficas em torno de um lema, normalmente a forma existente na ortografia padrão atual. Essa tarefa, morosa e minuciosa, muito ganharia com processos automáticos ou semi-automáticos de normalização sistemática de textos desses estágios da língua portuguesa. Essa tarefa não só os tornaria aptos para futuros processamentos, mas também permitiria que leitores não historiadores ou não linguistas pudessem ter acesso mais facilmente a textos valiosos culturalmente e de valor patrimonial, com interesse para várias áreas do saber. Compreender essas diferenças e conseguir traduzir ou associar escritos antigos aos padrões atuais é um passo importante para outros níveis de processamento (Cameron et al., 2020).

Outra forma de mitigar as variantes de escrita é criar modelos de linguagem que incluam *corpora* de outros períodos de tempo, com suas variantes de ocorrência natural, numa fase adicional de treinamento (ajuste por *fine tuning*), para adaptar o modelo às variantes (Arevalo; Fonteyn, 2021). Em *corpora* históricos transcritos, a presença de vários critérios de transcrição pode dificultar a operação. Além disso, há a dificuldade em reunir um volume substancial de textos de uma mesma época que permita aprendizagens eficientes e com elevada acuidade.

15.2.4 Anotação de textos

A anotação de textos é um processo que por vezes pode ser complexo, necessitando de equipes interdisciplinares, de modo a prever todas as vertentes. A anotação requer a identificação de categorias de interesse para o estudo realizado, isso implica em impor alguma objetividade sobre a língua, sendo que a interpretação é fundamentalmente um processo subjetivo.

A anotação textual ocorre em diferentes níveis, sintático, semântico lexical, e discursivo. Como para qualquer outra área, também nas HD, é útil identificar elementos de interesse para embasar diferentes estudos. Geralmente a anotação sintática permite um tipo de análise diferenciada da anotação semântica. Estudos linguísticos de filologia podem se beneficiar das anotações de nível morfológico (lexical estrutural) e sintático (sentencial), como poderemos conferir em exemplos de projetos apresentados nas Seções 15.4.1 e 15.4.4.

A anotação de entidades é frequentemente adotada, uma vez que se trata de uma tarefa com desenvolvimento bastante avançado, com utilidade para estudos de literatura, história, geografia, entre muitos outros campos do conhecimento. Conforme a complexidade do fenômeno e a respectiva disponibilidade de recursos, a anotação pode ser manual ou automática.

Identificar personagens, instituições ou os locais mencionados em documentos históricos ou obras literárias possibilita uma série de análises. Quase sempre permite uma resposta mais consistente a perguntas básicas, mas essenciais em qualquer estudo rigoroso, como são as associadas a questões como "quem?","o quê?", "onde?". Isso poderá ser verificado nas Seções 15.4.2, 15.4.3 e 15.4.5. No processo de anotação, é fundamental respeitar o

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

contexto nas decisões que o anotador deve tomar, pois há muitos casos que se prestam a interpretação ambígua.

15.3 Transformação de textos em dados

Naturalmente, com a evolução dos estudos em HD, são produzidos dados diferenciados, mais elaborados, mais numerosos, e úteis aos investigadores em humanidades. Amplia-se, assim, a capacidade de análise, pela possibilidade de trabalhar com um volume maior de informação e pela capacidade de organizar essa informação de forma mais ágil e rápida em estruturas bem definidas. O PLN atua como área responsável por possibilitar essa transformação de textos em dados. São comuns nesses processos a presença dos seguintes elementos:

- extração de informação para a criação de conjunto de dados;
- uso de técnicas de representação e organização de conhecimento, para organizar os dados;
- criação de bases de dados interligadas, de forma a estimular o re-uso e a colaboração.

Contudo, tais processos, quando aplicados em pesquisas na área das humanidades, requerem não apenas ferramentas atuais de PLN, mas também uma interação mais próxima com os pesquisadores das respectivas áreas. Isso é necessário para o desenvolvimento das adaptações de ferramentas a diferentes objetivos e para a construção de interfaces adequadas ao seu uso. É, de fato, uma elaboração interdisciplinar. É importante aliar os interesses dos pesquisadores em humanidades às possibilidades mais atuais e eficientes de obtenção de informação por meio da aplicação de tecnologias da linguagem, especialidade dos pesquisadores em PLN. Discutiremos a seguir os elementos, mencionados acima, concernentes à transformação de textos em dados.

15.3.1 Extração de informação

Entre as subáreas de PLN, a extração de informação é o processo mais fortemente relacionado ao objetivo de transformação de dados não estruturados (textuais) em dados estruturados (tabelas ou banco de dados). O capítulo 17 deste livro apresenta em detalhes essa área do PLN. Enquadram-se nela as tarefas de reconhecimento de entidades nomeadas e de identificação e classificação de eventos, aplicadas, por exemplo, nos projetos descritos nas Seções 15.4.2, 15.4.3 e 15.4.5 para o estudo de fontes históricas.

O reconhecimento de entidades nomeadas possibilita a identificação de personagens importantes de um dado período; possibilita também mapear áreas de atuação, fazer relações com questões de cartografia e sistemas de informação geográfica. As instituições de diferentes perfis também podem ser identificadas. Se usarmos técnicas mais avançadas, os relacionamentos entre essas entidades também poderão trazer informações relevantes para os pesquisadores. A identificação de eventos e sua respectiva classificação podem ajudar a abordar zonas e contextos significativos, como, por exemplo, no projeto *Monsoon* (Seção 15.4.3) que realiza uma análise dos eventos inerentes a conflitos na coleção "Livro das Monções".

Em textos das áreas das humanidades que estejam já digitalizados e normalizados, podem ser aplicadas ferramentas de PLN já desenvolvidas para diferentes tarefas, como

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

por exemplo, reconhecimento de entidades nomeadas, extração de eventos, resolução de correferências, e sistemas de respostas a perguntas.

Uma outra maneira de fazer uso de ferramentas prontas, em textos não normalizados, com características diferenciadas do padrão contemporâneo, seria adaptar os modelos de linguagem já treinados. Essas adaptações podem ser feitas para datas históricas ou para a diversidade de mídias. Uma adaptação pode melhorar a performance dos sistemas, uma vez que os grandes modelos são treinados em versões contemporâneas da língua. Não conhecemos alguma adaptação de modelos para o português histórico, mas em breve deverão estar disponíveis, pois esta é uma área em rápida evolução.

15.3.2 Representação de conhecimento e ontologias

Ontologias são artefatos conceituais interpretados com diferentes nuances em filosofia, lógica e sistemas de informação. A ideia central é representar conceitualizações de forma estruturada. Em HD, a nuance mais comumente relacionada é aquela empregada em sistemas de informação, ontologias como estruturas conceituais que podem ajudar na organização dos dados e na comunicação entre sistemas. O nível de detalhe de uma ontologia pode variar entre estruturas taxonômicas, definição de instâncias, relações e axiomas lógicos. Elas auxiliam a esclarecer (e combinar) diversas conceitualizações que compõem diferentes disciplinas.

Ontologias têm sido usadas em diversos domínios do conhecimento em diferentes tarefas:

- organizar e anotar grandes quantidades de dados (anotação semântica);
- integrar dados de diversas fontes (integração semântica);
- representar conhecimento de domínios complexos;
- dar ancoragem para o raciocínio automático:
- suportar buscas em grandes quantidades de dados (busca semântica).

Em HD, além das tarefas citadas acima, ontologias têm sido utilizadas para representação de metadados, pois, geralmente, é necessário estender ontologias de domínio existentes ou desenvolver novas para atender às especificidades de cada *corpus* ou fonte. Essas descrições formais permitem a integração de dados de múltiplas fontes, de maneira independente de software e de esquema. Um passo essencial para melhorar a qualidade dos dados é usar vocabulários padronizados e ontologias para representação de dados e metadados (Guizzardi, 2020).

Uma ontologia bastante referenciada em HD, em especial na área do patrimônio cultural, é o CIDOC-CRM. É entendida como uma ferramenta teórica e prática para a integração de informação. O objetivo é ajudar os pesquisadores, os administradores e o público a explorar questões complexas, relacionadas ao passado, disponíveis em conjuntos de dados diversos e dispersos. O CIDOC-CRM faz isto fornecendo definições e uma estrutura formal para descrever os conceitos e relações implícitos e explícitos, utilizados em documentações do patrimônio cultural.

15.3.3 Dados ligados e dados *FAIR*

Os esforços de partilha de dados produzidos nas pesquisas ainda têm um longo caminho a percorrer em termos de padronização. É muito importante tornar os dados compatíveis com os princípios *FAIR* (*Findability*/Encontrabilidade, *Accessibility*/Acessibilidade,

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Interoperability/Interoperabilidade e *Reuse*/Reutilização) (Wilkinson et al., 2016). Esses princípios correspondem a um conjunto de 15 recomendações que visam facilitar a reutilização de dados por humanos e máquinas. Eles são independentes de domínio e podem ser implementados principalmente através de:

- (F) atribuição de identificadores exclusivos e persistentes a conjuntos de dados, descrevendo-os com metadados ricos que permitem sua indexação e descoberta;
- (A) uso de protocolos abertos e de padrões para acesso a conjuntos de dados;
- (I) uso de linguagens formais e de vocabulários padronizados *FAIR* para representar (meta)dados;
- (R) uso de metadados ricos sobre licença de uso, proveniência e qualidade de dados.

Portanto, o primeiro passo para o cumprimento dos princípios *FAIR* é atribuir metadados aos conjuntos de dados e definir esquemas precisos de metadados. De fato, 12 dos 15 princípios *FAIR* (Wilkinson et al., 2016) se referem a metadados. Para dar um passo adiante no aprimoramento *FAIR* dos dados, os esquemas de metadados devem se basear em modelos semânticos (ou seja, ontologias) para uma representação de metadados mais rica (Guizzardi, 2020). Graças à sua capacidade de tornar os tipos de dados explícitos, em um formato que pode ser processado por máquinas, as ontologias são essenciais para tornar os dados *FAIR*, mesmo para dados já publicados na Web (Jacobsen et al., 2020).

A ideia dos dados ligados é que os esforços possam ser reaproveitados e as bases enriquecidas com dados já identificados e organizados por outras iniciativas. A *open linked data cloud* (LOD)[9] fornece uma visão geral dos conjuntos de dados vinculados que estão disponíveis na Web. Representa um grafo de conhecimento, composto por diversos padrões abertos, como URI, URL, HTTP, HTML, RDF, a linguagem de consulta SPARQL, entre outros. Esse padrões são mantidos por curadores de dados amadores e profissionais na indústria e na academia[10].

Por exemplo, a *Europeana* é uma biblioteca virtual desenvolvida por países da União Europeia (Borin; Donato, 2023; Coneglian; Santarem Segundo, 2017). O protótipo agrega milhões de itens digitais, todos eles em domínio público, e está ligado ao LOD Cloud[11]. Em (Koho et al., 2021) é proposta uma infraestrutura semântica compartilhada, baseada na ideia de representar conceitos de guerra como uma sequência espaço-temporal de eventos, nos quais participam soldados, unidades militares e outros atores. O esquema de metadados usado é uma extensão do CIDOC-CRM, complementado por várias ontologias de domínio histórico militar, e é ligado ao LOD Cloud.

Em termos de avaliação, várias estruturas foram propostas para avaliar o grau de *FAIRness* de um determinado objeto digital (Sun et al., 2022). Em vários deles, a avaliação é realizada respondendo a um conjunto de perguntas, também chamadas de métricas ou indicadores, ou preenchendo uma lista de verificação[12], como o *FAIRshake* (Clarke et al., 2019) ou o *FAIR Data Maturity Model* (FAIR Data Maturity Model Working Group RDA, 2020).

Outros autores (Devaraju et al., 2020; Wilkinson et al., 2019) propuseram abordagens automatizadas. Recentemente, além de avaliar o grau de *FAIRness* dos dados, as propostas

[9] https://lod-cloud.net
[10] https://www.w3.org/wiki/LinkedData
[11] https://www.europeana.eu/en
[12] https://fairassist.org/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

abordaram a avaliação de vocabulários e ontologias (Cox, 2021; Garijo; Poveda-Villalón, 2020).

Nesta seção discutimos alguns elementos do caminho de construção de bases de pesquisa a partir de textos. Muitos projetos na área de HD são desenvolvidos com o objetivo de mapear informações textuais volumosas, dispersas, ou ainda não totalmente exploradas em novas descobertas. Uma vez mapeadas e organizadas, elas podem estar disponíveis para outros investigadores. As áreas de extração de informação, construção de ontologias e dados ligados caminham em paralelo, embora esforços interdisciplinares ainda sejam necessários para sua integração. Na próxima seção serão apresentados exemplos de projetos que caminham nesse sentido.

15.4 Exemplos de projetos de HD envolvendo o processamento da língua portuguesa

Apresentamos aqui exemplos de trabalhos envolvendo HD e PLN com foco em língua portuguesa. Esta não é uma apresentação completa; há diversos outros projetos e trabalhos importantes e reconhecidos nesse domínio. Em especial, damos atenção aqui aos projetos relacionados com as autoras deste capítulo, e que colaboram no contexto do Laboratório Chronos[13], o Laboratório de Humanidades Digitais do Centro Interdisciplinar de História Culturas e Sociedades, CIDEHUS, da Universidade de Évora, Portugal. São, em geral, projetos atuais em desenvolvimento, com exceção do primeiro, que tem uma relevância histórica no desenvolvimento dessa área no Brasil.

15.4.1 O *corpus* Tycho Brahe

O *corpus* histórico português Tycho Brahe (Sousa, 2014) é um *corpus* eletrônico de textos escritos em português por autores nascidos entre 1380 e 1978. O objetivo principal do projeto foi possibilitar, de forma ampla, a recuperação de informações filológicas e linguísticas dos textos. Esse pode ser considerado um dos trabalhos pioneiros na área de *corpus* histórico, com adição de anotação linguística e que tenha envolvido recursos computacionais na sua criação, sendo um importante contributo para essa área (Britto et al., 2002; Finger, 2000).

No site do projeto[14], 95 textos se encontram disponíveis para pesquisa. Os textos contam com anotação linguística em dois níveis: anotação morfológica *(part of speech)* (60 textos, total de 2.204.889 palavras); e anotação sintática (31 textos, total de 1.311.834 palavras). Estão disponíveis ainda versões sem anotação. A coleção engloba o português brasileiro e o europeu, contendo, por exemplo, a Gazeta de Lisboa e o Jornal da Bahia.

15.4.2 As Memórias Paroquiais

As Memórias Paroquiais constituem uma das principais fontes sobre o Portugal metropolitano dos meados do século XVIII. São, aliás, das mais citadas pelos historiadores. A coleção é composta pelas respostas dos párocos a um inquérito lançado em 1758, três anos depois do grande sismo de 1755. Incluía 60 perguntas sobre o território da freguesia, pelo que reúne uma plêiade diversificada de informações, desde elementos administrativos e

[13] https://sites.google.com/view/hdlabcidehus
[14] https://www.tycho.iel.unicamp.br/corpus/

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

demográficos aos recursos existentes: piscícolas, minerais, aquíferos com propriedades específicas, produções agro-pecuárias, etc. Interessam tanto ao historiador como ao botânico ou ao sismógrafo, para referir apenas alguns dos campos abrangidos pela observação dos párocos informantes.

Estas respostas, encadernadas no século XIX, fez com que os volumes passassem a ser conhecidos como "Dicionário Geográfico de Portugal". Aliás, a iniciativa de 1758 também tinha em vista dar continuidade a um projeto de dicionário com esse perfil, do Padre oratoriano Luís Cardoso (1697-1769), e que o terremoto de 1755 interrompera. A atual designação só se impôs no limiar do século XX (Olival et al., 2022).

Os originais manuscritos, num total de 43 volumes e um de índices, encontram-se na Torre do Tombo, desde o século XIX. Entre 1993 e 2003, os textos foram microfilmados e em 2005 esses rolos de 35 mm[15], em preto e branco, foram digitalizados e disponibilizados online.

Por volta de 2007-2008, o CIDEHUS iniciou a tarefa de transcrever as Memórias das localidades a Sul do Tejo. O trabalho foi concluído em 2023 em um projeto colaborativo. Por vezes, até aproveitou transcrições já publicadas, sempre que foi possível entrar em contato com o Autor(a) da edição. Parte desse trabalho encontra-se disponível online, em acesso aberto, e com licenças *Creative Commons*, que facilitam a reutilização.

Nos últimos dois anos começaram a ser aplicadas técnicas de PLN ao conjunto transcrito. Principiou-se por anotar categorias simples e básicas: pessoas, locais e organizações (Vieira et al., 2021); já em 2023, usaram-se categorias de anotação mais complexas e com várias subdivisões. Teve-se em vista responder de modo mais adequado às necessidades do historiador e de outros cientistas ou pessoas interessadas. Também foi efetuado um ensaio de reconhecimento de entidades nomeadas, usando sistemas previamente desenvolvidos e que foram calibrados para esse efeito (Santos et al., 2024a). Baseavam-se em técnicas de aprendizagem de máquina e modelos de linguagem.

Está ainda em curso a normalização da grafia do texto, quer em versão com léxico explicativo para o grande público, quer em versões para processamento, trabalhando-se na possibilidade de normalizar de forma automatizada (Cameron et al., 2023).

Nos próximos anos, será dada continuidade à tarefa de transformar estes textos em dados confiáveis e de ligá-los a outros repositórios de conhecimento.

15.4.3 O projeto MONSOON

Extração de eventos é uma tarefa bastante conhecida de PLN, com recursos desenvolvidos para o português (Sacramento; Souza, 2021). O desafio, neste caso, é adaptar essas ferramentas à língua portuguesa do século XVII, esforço que está sendo explorado no projeto *Monsoon*: o Estado da Índia Habsburgo em perspectiva digital (1580-1640).

Este projeto pretende estudar as dinâmicas internas desse espaço macro que constituía o Estado da Índia português, combinando o estudo da presença formal portuguesa com a presença de comunidades portuguesas que se instalaram na Ásia fora do território administrado pela Coroa portuguesa, buscando relacionar estas duas esferas com o Outro – seja o europeu ou o asiático ou africano. Por outro lado, o projeto tem o objetivo de criar uma metodologia de análise semi-automática que possa ser replicada em coleções documentais semelhantes de outras potências europeias (Ribeiro, 2022).

Este projeto tem como base documental a correspondência trocada entre o monarca português e as suas instituições com o vice-rei da Índia. Este conjunto ficou conhecido

[15]Agradece-se à Dra. Anabela Ribeiro, do Arquivo Nacional da Torre do Tombo, estas informações.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

como Livros das Monções, dado que era na altura da monção no Oceano Índico que os navios portugueses podiam partir de Goa para Portugal ou podiam chegar ao Índico. Só no regime climático de monções existem ventos no Índico que permitem que navios sem motor aí circulem. Este conjunto documental, dividido entre o Arquivo Nacional Torre do Tombo, em Lisboa, Portugal, e o Arquivo Histórico do Estado de Goa, em Pangim, Índia, possui informações sobre os mais variados aspectos da presença portuguesa na Ásia e abarca uma geografia que vai desde Moçambique e toda a costa oriental africana até ao Japão.

Este *corpus* documental é composto por um grande número de volumes, sendo que nenhum deles possui um índice sobre os assuntos tratados. Por meio da extração automática de eventos, os historiadores podem facilmente ter acesso a passagens relevantes a partir de uma palavra relacionada com uma determinada ação. É possível identificar e classificar semanticamente eventos, o que nos permite ter uma percepção clara dos temas abordados nesta correspondência, de forma eficiente e rápida. A partir da aplicação de extração de eventos numa amostra destas cartas, compreendemos que, em determinada cronologia, havia no Oceano Índico uma conjuntura generalizada de conflito, quer devido à concorrência de outras potências ultramarinas europeias, como a Holanda e a Inglaterra, quer entre os portugueses e outras entidades políticas asiáticas (Albuquerque et al., 2024a).

Através do sistema de extração de eventos, temos acesso a estatísticas de classificação que nos permitem uma percepção mais evidente dos assuntos representados em cada carta e assim analisar conjunturas históricas. Isso, no que diz respeito à metodologia da análise histórica, é algo relativamente novo.

Por outro lado, este projeto utiliza também a anotação de entidades nomeadas de quatro tipos: pessoas, locais, grupos de pessoas e instituições. Nos próximos passos deste projeto pretendemos conjugar as entidades extraídas por meio de anotação de texto e ligá-las aos eventos extraídos no texto, podendo assim articular agentes, instituições e locais às ações em que estão envolvidos.

15.4.4 A História do Futuro

Nos projetos em curso sobre a "História do Futuro" (Banza, 2022), de Padre António Vieira, após alguns testes com a ferramenta de transcrição de manuscritos Transkribus (Kahle et al., 2017), concluiu-se que, tratando-se de um texto relativamente curto e dadas as características do manuscrito, de leitura muito difícil, em parte devido às más condições do suporte, o treino necessário para uma utilização rentável da ferramenta não se justificaria neste caso, ao contrário de projetos envolvendo um grande volume de textos, uma vez que seria sempre necessária uma correção manual. Assim, o texto foi lido e transcrito manualmente para suporte digital, aplicando-se apenas posteriormente determinadas ferramentas de análise, de acordo com os objetivos pretendidos.

A primeira transcrição, feita a partir da leitura paleográfica do manuscrito, sendo maximamente conservadora, mantém um número muito significativo de traços do português seiscentista, e do português do Padre António Vieira em particular, que dificultam muito ou impossibilitam mesmo a aplicação da maioria das ferramentas de PLN. Assim, optou-se por, a partir da primeira, realizar uma segunda transcrição, igualmente em formato , mas normalizadora. Dadas as características do texto, a normalização foi, mais uma vez, feita manualmente, depois de definidos critérios rigorosos que, sem desvirtuarem as principais características do texto e da sua época, o tornassem apto a processamento computacional e a aplicação de ferramentas automáticas.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Sobre a versão normalizadora, tem sido possível aplicar com êxito algumas ferramentas e testar outras. Inicialmente, constituiu-se um *corpus* não lematizado, que foi anotado lexicalmente (permitindo, por exemplo a elaboração de um léxico da obra) e morfologicamente (o que permitirá rastrear automaticamente alguns aspetos inovadores na obra de Vieira neste domínio, como é o caso do uso dos pronomes clíticos), com recurso à ferramenta LX-Tagger[16], disponível no repositório PORTULAN-CLARIN[17].

Nesses domínios, a anotação automática, verificada depois manualmente, revelou-se particularmente eficaz. Elaboraram-se, ainda, com auxílio da ferramenta AntConc[18], listas lexicais, alfabéticas, por frequência descendente, ou ordenadas pelo final de palavra, permitindo ver a disponibilidade do sistema sufixal. O *corpus* contém ainda um valioso intertexto latino, que foi processado manualmente de modo a constituir um sub-*corpus* autônomo.

Ainda no domínio da análise linguística, prevê-se a utilização do esquema de Dependências Universais, por exemplo, para a análise morfossintática, cujos testes têm revelado um desempenho bastante razoável, apesar de, como nos demais casos, não se poder excluir a revisão manual. Encontram-se também ainda em fase de avaliação ferramentas que permitam identificar similaridades semânticas (Lopez-Gazpio et al., 2017) entre o texto consensualmente aceito como "História do Futuro" e outros, fisicamente distintos, mas conceitualmente idênticos. Neste aspecto, as ferramentas presentemente disponíveis têm-se revelado pouco adequadas, quer ao tipo de texto, quer ao tipo de similaridades que se procura detectar. O objetivo será, neste caso, adaptar os modelos atuais à"História do Futuro", conseguindo, essencialmente, mais eficiência, em termos de precisão e cobertura. Ainda assim, algumas ferramentas já disponíveis, como por exemplo, para o reconhecimento de entidades nomeadas ou para a análise de sentimentos, úteis para a análise histórico-cultural e literária da obra, poderão necessitar de uma menor adaptação.

15.4.5 Manuais médicos do século XVIII

Com a digitalização de acervos bibliográficos de bibliotecas físicas, impulsionada pelas HD (Finatto, 2023), tornou-se mais fácil acessar obras médicas antigas impressas em português. Todavia, tais obras precisam ter seus conteúdos sistematizados para que possam ser compreendidas e situadas em relação ao conhecimento médico e científico atual.

Mesmo com a digitalização e com as transcrições dos textos, conforme sua apresentação original, a interpretação e sistematização dessa informação requer consulta a toda uma série de materiais de apoio: compêndios de História da Ciência e da Medicina, edições filológicas, bases de dados com textos do português de diferentes épocas e vários tipos de dicionários. Assim, no âmbito do projeto "Humanidades digitais e conhecimento médico em língua portuguesa no século XVIII", são analisados e disponibilizados[19] textos médicos desse período, reunidos em um *corpus* de arquivos de transcrições (Lazzari; Finatto, 2023). Esse *corpus* é de acesso público e oferece manuais escritos por médicos, enfermeiros e cirurgiões. Nele, têm destaque as obras do médico Curvo Semedo (1635-1719) e o primeiro manual de Enfermagem publicado em português, de 1741.

O objetivo desse projeto é descrever e sistematizar terminologias, conceitos e modos de dizer relacionados a doenças e seus tratamentos, criando-se bases para um futuro hiper-

[16] http://lxcenter.di.fc.ul.pt/tools/en/conteudo/LXTagger.html
[17] https://portulanclarin.net
[18] https://www.laurenceanthony.net/software/antconc/
[19] https://sites.google.com/view/projeto38597

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

dicionário (Wives, 1997) de epidemiologia histórica luso-brasileira. Nesse hiperdicionário, a ideia é associar nomes de doenças, partes do corpo, perfis de pessoas/doentes, dados geográficos e datas, com nomes de remédios, de processos e de tratamentos que chegam aos seus correspondentes terapêuticos atuais. Esses dados podem ser correspondidos à identificação das Entidades Nomeadas ao longo dos textos (Zilio et al., 2022).

Naturalmente, há todo um trabalho de geração e tratamento desse *corpus* histórico, que é feito também com auxílio de recursos e técnicas de PLN. Isso porque, no período dessas obras, verificava-se toda a sorte de variação de escrita e de tipografia, de modo que em uma mesma página, de um mesmo autor, encontra-se, por exemplo a palavra "água" escrita como "agua" ou "agoa", além de haver itens de vocabulário hoje desconhecido, como o termo anatômico "madre" por "útero". Por isso, o projeto já tenta aproximar, com apoio computacional, a escrita antiga da sua forma atualizada, investindo-se na normalização automática dos textos (Zilio et al., 2024).

Embora sejam enfrentados problemas com a qualidade das informações automaticamente extraídas em dados brutos, pois lida-se com textos apenas na ortografia antiga, plenos de variações nas formas das palavras conforme escritas na época, ferramentas genéricas de PLN já se mostram capazes de identificar informações relevantes (Quaresma; Finatto, 2020). São ferramentas com bom potencial para ajudar especialistas que lidam com esses textos na criação de bases de conhecimento histórico. Enfim, o processamento desse *corpus* ajuda a identificar e refinar a informação nele contida. E isso pode ser muito relevante para várias aplicações, produzindo-se verdadeiros "mapas de navegação" da sua organização e conteúdos.

A ligação de ontologias é, por fim, outro dos objetivos desse projeto de pesquisa. Assim, as antigas terminologias de Saúde poderão ser encontradas e mapeadas em seus equivalentes em ontologias da atualidade (Schriml et al., 2012). Este mapeamento será mais uma forma de enriquecer os recursos derivados do *corpus* e estabelecer quais as áreas da Medicina, da Anatomia e que tipo de doenças e medicamentos eram conhecidos daquela época. Atualmente, as terminologias e expressões relacionadas, conforme usadas nos manuais de Curvo Semedo, já estão contrastadas com o vocabulário de Enfermagem empregado em 1741 (Finatto et al., 2023). Esse contraste, entre outros, pode ajudar a identificar diferentes perspectivas, gêneros textuais, conceitos e saberes dos vários práticos e profissionais de Saúde em ação e formação no século XVIII.

15.5 Considerações finais

Há um grande potencial de aplicações de PLN na área de humanidades digitais, muitas ainda a serem exploradas. Como visto neste capítulo, o PLN se faz presente no tratamento de problemas em diferentes fases, desde as etapas iniciais relacionadas à transcrição de manuscritos, ao tratamento de textos como imagem com técnicas como o OCR, a adição de metadados, e a normalização textual. Para as fases de processamento, a partir de um determinado texto preparado, as tecnologias de linguagem podem prestar auxílio a tarefas diversas como tradução, recuperação e extração de informações, criação de bases de conhecimento, e sua associação a ontologias ou outros dados. O ideal é que exista uma atenção especial aos padrões de disponibilização de dados, como os propostos pelos princípios *FAIR* de Dados Abertos.

Nessa área de intersecção entre humanidades e tecnologias é essencial um trabalho interdisciplinar. São várias questões relevantes a serem combinadas: os objetivos de um pesquisador da área de humanidades, como linguística e literatura, ciências sociais, his-

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

tória, e, na perspectiva do processamento de língua, os objetivos de encontrar, aplicar ou mesmo desenvolver soluções apropriadas para cada tipo de problema, maximizando a correção e a performance das soluções encontradas de acordo com a disponibilidade de recursos.

É crucial que uma perspectiva de IA, centrada no ser humano, seja levada em consideração para fornecer interfaces de usuário adequadas para preparar e acessar as fontes e os dados extraídos. Os projetos, muitas vezes considerando coleções distintas, ou mesmo possuindo objetivos diferentes, podem se beneficiar com a troca de experiências e com o uso das mesmas ferramentas para lidar com os textos e seus conhecimentos codificados.

No contexto do evento principal da língua portuguesa, o PROPOR, tem sido organizado o *Workshop* de Humanidades Digitais e Processamento de Língua Natural, que em 2024 estará na sua terceira edição. Nos anais desse *workshop* disponíveis *on-line*[20], podem ser encontrados mais detalhes sobre os projetos mencionados neste capítulo, e muitos outros.

Agradecimentos

Os trabalhos aqui mencionados contam com o suporte da Fundação de Ciência e Tecnologia (FCT) - projetos CEECIND/01997/2017, 2022.07730.PTDC, UIDB/ 00057/2020 (https://doi.org/10.54499/UIDB/00057/2020), e e do Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) - processos 308926/ 2019-6 (PQ) e 401770/2022-2 (PDE).

[20] https://sites.google.com/view/dhandnlp-propor

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Capítulo 16

PLN em Redes Sociais

Brenda S. Santana
Larissa A. de Freitas

Publicado em: 26/09/2023
Atualizado em: 13/03/2024

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

16.1 Introdução

O Processamento de Linguagem Natural (PLN) desempenha um papel cada vez mais significativo no cenário das redes sociais. O volume de dados advindos de redes sociais a todo instante é imenso. Dentre os dados gerados, podemos citar os dados textuais, os quais variam desde conversas informais até discussões complexas. Nas redes sociais, as pessoas expressam ideias e opiniões de maneiras diversas. Isso inclui o uso de gírias, abreviações, emojis e outros elementos da linguagem cotidiana. Tratar esse tipo de dado não é uma tarefa trivial e é desafiador para os sistemas de PLN.

Os sistemas de PLN são ferramentas indispensáveis para compreender, analisar e extrair informações. O estudo dos estilos de linguagem utilizados nas redes sociais ajuda a melhorar a compreensão de textos informais. As redes sociais são fontes valiosas de informação e seus conteúdos podem ser utilizados como *corpora* para treinar e testar algoritmos de PLN, permitindo que pesquisadores e desenvolvedores trabalhem com exemplos reais e relevantes. Visto que o Brasil é um dos países com maior presença nas redes sociais, e o português é o idioma predominante nessas interações, tem-se aqui uma área muito fértil para o desenvolvimento de estudos de aplicações de abordagens e PLN.

Neste capítulo buscamos abordar algumas das principais áreas de aplicação de PLN em redes sociais, discutindo os desafios encontrados. Ainda, buscamos apresentar alguns dos recursos disponíveis para suporte no desenvolvimento de estudos voltados para as tarefas apresentadas, focando em dados em língua portuguesa. Para tanto, este capítulo se organiza da seguinte maneira: na Seção 16.2, apresentamos a definição de redes sociais e descrevemos sobre os conteúdos nela postados; na Seção 16.3, apresentamos as principais áreas de aplicação de PLN que utilizam essas redes sociais. E, na Seção 16.4, apresentamos as considerações finais.

16.2 Redes Sociais

Uma rede social é definida como um conjunto de dois elementos: atores e suas conexões (Wasserman; Faust, 1994). Nos últimos anos, as redes sociais (como: Facebook, Reddit, Youtube, Twitter/X, Whatsapp e Instagram) têm revolucionado a forma como indivíduos, grupos e comunidades interagem. Nelas, são compartilhados textos, fotos, vídeos e outros

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

tipos de conteúdo. Assim, as redes sociais estabelecem um ambiente rico e dinâmico que oferece inúmeras oportunidades para o estudo e o aprimoramento de abordagens em PLN. Segundo Recuero (2009), o estudo das redes sociais na Internet objetiva analisar como as estruturas sociais surgem, de que tipo elas são e como são compostas.

De acordo com Farzindar; Inkpen (2018), usar PLN em textos provindos de mídias tradicionais (como jornal, rádio e televisão) tem sido um tópico de pesquisa popular nos últimos 25 anos. Hoje, usar PLN em textos provindos de redes sociais é uma área de pesquisa que requer adaptações dos métodos tradicionais, já que os textos provindos de redes sociais têm várias peculiaridades, principalmente devido a sua natureza. Ainda, eles podem estar escritos em diferentes idiomas e pertencerem a diferentes fontes.

As redes sociais se popularizaram no Brasil em 2004, com a criação do Orkut[1]. Desde lá, novas redes surgiram e com elas a percepção da necessidade e viabilidade de aplicação de abordagens em PLN para o estudo de conteúdos e comportamentos gerados nesse meio. Dentre as áreas de aplicação dessas abordagens, destacam-se a detecção de discurso de ódio e linguagem ofensiva, a detecção de ironia/sarcasmo/humor, a detecção de notícias falsas, a análise de sentimento, entre outras (Ferreira et al., 2017).

Na literatura, existe uma predominância do Twitter/X como fonte de dados, isso se deve, provavelmente, ao fato de ele oferecer uma API[2] que, de forma muito simples, consegue acessar mensagens publicadas e os dados associados a seus usuários (por exemplo, o número de seguidores deste). No caso do Facebook, é necessário criar um aplicativo e obter a autorização dos usuários para que seus dados possam ser acessados/capturados (Coello; Junqueira, 2019), o que pode ser visto como um limitante na extração de informações desta rede.

No ano de 2023, algumas mudanças ocorreram nas APIs do Twitter/X e do Reddit. No Twitter/X, os pesquisadores terão que se adaptar às restrições da versão gratuita ou assinar alguns dos planos pagos para manter suas atividades. Já, no Reddit, o uso da API[3] passou a ser cobrado. Portanto, é de se esperar que mudanças aconteçam nas pesquisas que utilizam *corpora* advindos dessas redes sociais.

Como mencionado anteriormente, o conteúdo postado nas redes sociais pode variar muito de acordo com a plataforma, o público-alvo e a intenção por trás da postagem. Abaixo descrevemos brevemente as redes sociais mais utilizadas em trabalhos de PLN sobre redes sociais em língua portuguesa.

16.2.1 Facebook

O Facebook[4] é atualmente a maior rede social do mundo, com 2.9 bilhões de usuários ativos em 2023. Ela permite que os usuários criem perfis pessoais, adicionem amigos, compartilhem textos, fotos, vídeos e atualizações de status. Os usuários podem interagir com as postagens de outros usuários através de curtidas, comentários e compartilhamentos. Além disso, essa rede social também permite a criação de páginas para empresas, tornando-se uma ferramenta importante para marketing e divulgação.

[1] http://orkut.com/
[2] https://developer.twitter.com/en/docs/twitter-api
[3] https://www.reddit.com/dev/api/
[4] https://www.facebook.com/

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

16.2.2 Reddit

O Reddit[5] é uma plataforma online de compartilhamento de conteúdo e discussões, organizada em comunidades chamadas *subreddits*. Os usuários podem enviar postagens, comentar, votar em conteúdos e interagir uns com os outros. Essa plataforma abrange uma ampla variedade de tópicos e interesses, permitindo que os usuários encontrem comunidades específicas que correspondam aos seus interesses. É um espaço onde os usuários podem trocar informações, debater, compartilhar histórias, memes e muito mais.

16.2.3 Youtube

O Youtube[6] é uma plataforma de compartilhamento de vídeos que permite que os usuários compartilhem e assistam vídeos de uma variedade gêneros, incluindo filmes, programas de TV, vídeos musicais, documentários, entre outros. Também é uma ferramenta de marketing importante para muitas empresas e indivíduos, os quais usam a plataforma para compartilhar conteúdo promocional e aumentar a conscientização sobre seus produtos ou serviços.

16.2.4 Twitter/X

O Twitter/X[7] é um serviço de microblogging que pode ser utilizado para transmitir pequenas atualizações de status (Russell, 2011). Nele podem ser analisados os vínculos entre amigos e seguidores, grafos sociais e descobertas de mais informações sobre os usuários, inspecionando as entidades presentes em seus tweets. Os tweets são mensagens curtas (contendo até 280 caracteres, incluindo texto, imagens, GIFs, vídeos e links para outros sites) e públicas postadas no Twitter/X. Eles têm um alcance imediato e podem se tornar viral rapidamente, dependendo do conteúdo e da quantidade de interação que recebem de outros usuários. Isso faz do Twitter/X uma plataforma poderosa para disseminar informações, ideias e tendências em tempo real.

16.2.5 Whatsapp

O Whatsapp[8] permite que os usuários troquem mensagens privadas. Apesar de ser usado principalmente para conversas individuais, o WhatsApp possui recursos de grupos de conversação, onde podem participar até 256 usuários, e encaminhamento de mensagens (Cabral et al., 2021). Concebido como um aplicativo de mensagens instantâneas, o WhatsApp evoluiu para uma plataforma multifacetada, permitindo não apenas conversas privadas, mas também a formação de grupos e comunidades, compartilhamento de mídia, chamadas de voz e vídeo e até mesmo recursos empresariais.

16.2.6 Instagram

Instagram[9] é uma rede social para compartilhamento de fotos e vídeos. Nela também é possível acompanhar (seguir) outras contas, curtir, comentar e compartilhar publicações.

[5] https://www.reddit.com/
[6] https://www.youtube.com/
[7] https://twitter.com/. O Twitter mudou de nome em 2023 e agora se chama X. Por esse motivo, ao longo do texto, iremos usar o termo Twitter/X quando mencionarmos essa rede social.
[8] https://www.whatsapp.com/
[9] https://www.instagram.com/

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Todas as publicações realizadas no aplicativo são mostradas por meio do *feed* e o usuário pode visualizar as postagens das contas que ele segue. Ainda, esta rede social oferece diversas outras funcionalidades, como: *boomerang*, *live* e *stories*.

16.3 Áreas de Aplicação

Abaixo são descritas quatro áreas de aplicações que surgiram com a finalidade de compreender, analisar e extrair informações de textos que são publicados diariamente nas redes sociais, são elas: detecção de discurso de ódio e linguagem ofensiva, análise de sentimento, detecção de notícias falsas e detecção de ironia/sarcasmo/humor.

16.3.1 Detecção de Discurso de Ódio e Linguagem Ofensiva

A partir de definições encontradas na literatura, termos diferentes, porém semelhantes, podem ser enquadrados como discursos simbolicamente prejudiciais (por exemplo, discurso perigoso, discurso tóxico, discurso de ódio, discurso intolerante e outros). Certos discursos possuem o potencial de causar danos significativos, inclusive críticos, e podem ser considerados tóxicos (Tirrell, 2018). Discursos tóxicos podem assumir diversas formas, podendo ser um discurso persistente ou momentâneo, afetar indivíduos ou a sociedade como um todo, causando danos temporários ou permanentes. O impacto de toxinas discursivas é de natureza social, afetando comunidades e prejudicando indivíduos pertencentes aos grupos-alvos. Essas toxinas podem incluir palavras ofensivas, insultos, discriminação, discurso de ódio, difamação, ameaças ou qualquer forma de linguagem que busque macular, menosprezar ou ferir a dignidade e a integridade de indivíduos pertencentes ao grupo-alvo. De acordo com Kumar et al. (2023), comentários tóxicos são a principal forma de ódio e assédio online.

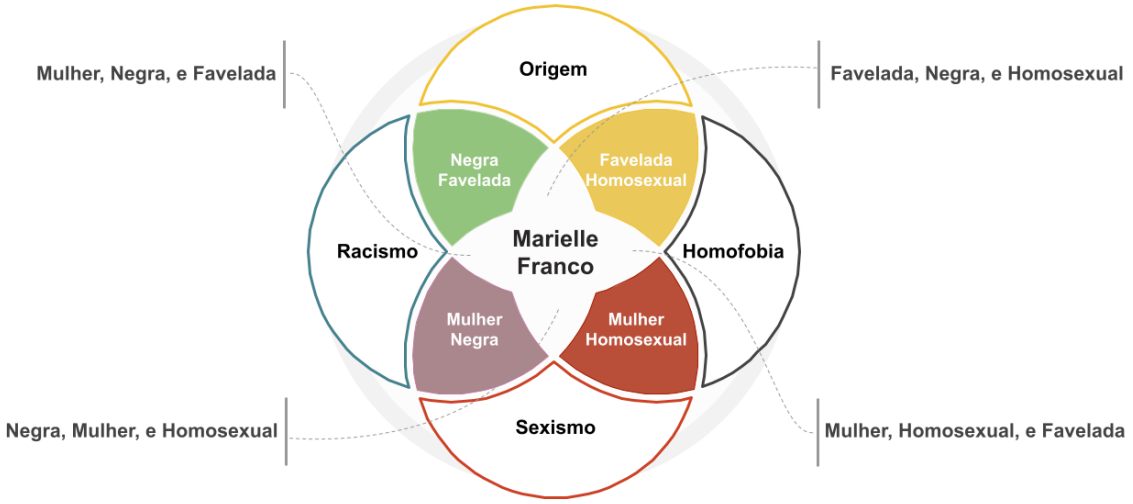
Os grupos-alvos de discursos tóxicos podem variar dependendo do contexto e da natureza do discurso. Grupos frequentemente alvos de discursos tóxicos incluem minorias étnicas e raciais, comunidade LGBTQIA, mulheres, religiões minoritárias, portadores de deficiência, refugiados e imigrantes, e grupos políticos ou ideológicos. Contudo, qualquer grupo ou indivíduo pode ser alvo de discursos tóxicos, e a disseminação desse tipo de linguagem é prejudicial para a sociedade como um todo. Indivíduos pertencentes a diferentes grupos discriminados podem ainda ser alvos de discursos interseccionais que os atacam por múltiplas frentes. A Figura 16.1, adaptada de Santana (2023), ilustra um caso em que ataques interseccionais dirigidos a uma entidade foram disseminados na Internet. A imagem busca ilustrar o que aconteceu após o assassinato em 2018 da socióloga e política brasileira Marielle Franco, quando uma rede de ódio e desinformação gerou diversos comentários online atacando sua imagem por diversas características que ela possuía, e até outros traços que foram indevidamente atribuídos a ela. Teixeira; Zamora (2019) destacam que Marielle - mulher negra, assumidamente bissexual, favelada, defensora política dos direitos humanos - foi, sem dúvida, atravessada por todo tipo de opressão desencadeada pelo sistema machista, racista e classista. Os ataques registrados neste caso foram motivados pelo ódio.

A análise e a detecção de diferentes tipos de discursos tóxicos são um tópico de crescente interesse tanto para a área de PLN quanto demais áreas de interesse social como um todo. De acordo com Guimarães et al. (2020), quando focamos em comentários tóxicos, especialmente em notícias, o Facebook é a rede social que mais se destaca. O Reddit também introduz uma inclinação em relação à linguagem tóxica e ofensiva (Mohan et al.,

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Figura 16.1: Exemplo de interseccionalidade entre diferentes formas de discurso de ódio observado no caso de Marielle Franco.



Fonte: Adaptada de (Santana, 2023)

2017). Por esse motivo, o conteúdo do Reddit tem sido usado para estudar microagressões (Breitfeller et al., 2019; Mollas et al., 2022) e depressão (Pirina; Çöltekin, 2018). De acordo com estudos realizados por Kumar et al. (2023), perfis que postam comentários tóxicos representam 3,1% de todas as contas que postam comentários no Reddit. Entretanto, ainda de acordo com os autores, apesar de seu percentual relativamente pequeno, tais contas desempenham um papel ativo e de alto impacto na plataforma.

Apesar dos diversos avanços pelos quais a área de PLN vem passando, a detecção de discursos tóxicos ainda é um desafio latente. O desenvolvimento de algoritmos de PLN e Aprendizado de Máquina (AM) para detectar esses tipos de conteúdo depende da disponibilidade de *corpora* anotados para treinamento. Conforme identificado por Trajano et al. (2023) quase todos os sistemas de detecção de toxicidade usam modelos de aprendizado supervisionado que requerem uma grande quantidade de dados rotulados[10]. Entre estes *corpora*, podemos ressaltar recursos para a língua a portuguesa como o ToLD-Br[11] desenvolvido por Leite et al. (2020).

O ToLD-Br (Leite et al., 2020) é um conjunto de dados capturado do Twitter/X entre julho e agosto de 2019 com a ferramenta GATE Cloud's Twitter Collector[12]. Elaborado para estudos sobre classificação automática de comentários tóxicos, este conjunto de dados tem como objetivo equilibrar o viés de anotação. Para tanto, 42 anotadores foram selecionados, com base em suas informações demográficas. Este *corpus* apresenta um conjunto de 21000 tweets em português manualmente anotados por três diferentes anotadores em sete categorias: LGBTQ+fobia, obsceno, insulto, racismo, misoginia e/ou xenofobia. Exemplo 16.1 apresenta uma instância de tweet obsceno e Exemplo 16.2 uma instância de tweet insulto contidos no *corpus* ToLD-Br.

[10]Em https://hatespeechdata.com alguns *corpora* anotados sobre discurso de ódio, abuso online e linguagem ofensiva são catalogados e podem ser utilizados como base para o desenvolvimento de estudos que buscam detectá-los.
[11]https://github.com/JAugusto97/ToLD-Br
[12]https://cloud.gate.ac.uk

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

Exemplo 16.1:

> "Aonde tem um monte que fala mal, mas ninguem vai embora do morro." acha que alguem mora aqui por que quer, c*****o!? Que idéia. [...]

Exemplo 16.2:

> [...] VAI SE F***R IRMÃO VC NÃO É FELIZ PQ NAO QUER

O estudo de discursos tóxicos é de suma importância por várias razões, abordando questões sociais, éticas e técnicas. O volume de informações gerado a partir das redes sociais e plataformas online aumenta a exposição a discursos tóxicos, o que pode causar impactos negativos na saúde mental e emocional dos usuários. Compreender e identificar esses discursos é fundamental para criar um ambiente digital mais saudável e seguro para todos. Discursos tóxicos frequentemente incluem discursos de ódio e manifestação de linguagem imprópria, ou seja, atos que podem promover a violência, intolerância e discriminação contra grupos específicos. A análise desses discursos permite identificar padrões prejudiciais e trabalhar para mitigar seus efeitos negativos.

Adicionalmente, muitos discursos tóxicos envolvem a disseminação intencional de informações incorretas, desinformação e notícias falsas (Seção 16.3.3). Ao estudar esses discursos, podemos desenvolver técnicas para detecção precoce de conteúdo enganoso, ajudando a manter a qualidade da informação nas redes. Além disso, técnicas de PLN auxiliam na detecção de linguagem irônica, sarcástica e outros formatos frequentemente usados nas redes sociais para mascarar discursos de ódio (Seção 16.3.4). Essas abordagens avançadas permitem que plataformas de redes sociais aprimorem suas ferramentas de moderação, identificando automaticamente discursos de ódio e adotando medidas para removê-los ou sinalizá-los.

Muitas iniciativas têm sido empreendidas com o intuito de possibilitar a detecção automatizada de discursos de ódio nas diferentes plataformas. Conforme mencionado por Fortuna; Nunes (2018), esse crescente interesse não se restringe apenas à ampla cobertura midiática, mas também à crescente relevância política do tema. No entanto, os autores também destacam desafios latentes, como a falta de técnicas automáticas adequadas e a escassez de dados confiáveis sobre o discurso de ódio, que continuam motivando pesquisas nessa área. Analisando estatísticas brasileiras, Dadico (2020) explana que os dados indicam que o ódio sobrevitimiza pessoas de grupos identificados por critérios de raça, cor, etnia, sexo, orientação sexual, identidade de gênero, origem nacional e regional, sem-teto ou deficiência, entre outros atributos que os expõem a maior vulnerabilidade social. Apesar da normalização do ódio, esse discurso é parte de uma narrativa socio-histórica que traz em si os modos de pensar de uma cultura. É pela língua que nos mostramos como somos, e enquanto ela pode ser um instrumento de empoderamento, também pode gerar exclusão, opressão. O avanço de estudos de aplicação de abordagens de PLN para a detecção de tais conteúdos é essencial. Entretanto, avanços nesta área de estudos dependem fundamentalmente de conjuntos de dados anotados, ferramentas de análise de texto e modelos específicos disponibilizados para tal.

Para o português, Fortuna et al. (2019) criou um conjunto de dados para a classificação do discurso de ódio, o HLPHSD[13]. As instâncias deste conjunto foram coletadas através do uso da API do Twitter/X. Para isso, foram usadas palavras-chave e hashtags como #dyke

[13] https://b2share.eudat.eu/records/9005efe2d6be4293b63c3cffd4cf193e

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

ou #womensPlaceIsInTheKitchen coletadas entre janeiro e março de 2017 (majoritariamente). Este conjunto de dados contém conteúdo de 1156 usuários diferentes e abrange diferentes tipos de discriminação, com base sobre religião, gênero, orientação sexual, etnia, e migração. Nele foram feitas duas anotações: binária ("é discurso de ódio" ou "não é discurso de ódio") e hierárquica ("racismo", "sexismo", ou "homofobia"). Exemplo 16.3 apresenta uma instância de tweet que é discurso de ódio e Exemplo 16.4 uma instância de tweet que não é discurso de ódio contidos no *corpus* HLPHSD.

Exemplo 16.3:

> Os negros deveriam voltar para suas terras!!

Exemplo 16.4:

> Carne e feijão preto são deliciosos!

Na anotação binária, cada tweet foi anotado por três diferentes anotadores. Por fim, uma votação majoritária para determinar classificação final foi realizada nos 3059 tweets. Os autores realizaram experimentos utilizando uma LSTM combinada com *embeddings* pré-treinados para realizar uma classificação base a partir deste conjunto de dados e assim demonstrar seu potencial de uso. O resultado obtido foi a medida-F de 78%.

Outro conjunto de dados disponível na literatura que foi elaborado para estudos sobre a classificação do discurso de ódio é o HateBR[14], elaborado por Vargas et al. (2022b). O HateBR é composto por 7000 textos sobre o domínio político coletados através da API do Instagram. Neste conjunto de dados, constam postagens de seis contas pré-definidas (gênero - 4 mulheres e 2 homens, posição política - 3 liberais e 3 conservadores). Sua anotação foi feita de três maneiras: binária ("é ofensivo" ou "não é ofensivo"), granularidade ("levemente ofensivo", "moderadamente ofensivo" e "altamente ofensivo") e grupos de discursos de ódio ("partidarismo", "sexismo", "intolerância religiosa", "apologia pela ditadura", "gordofobia", "homofobia", "racismo", "anti-semitismo" e "xenofobia"). Exemplo 16.5 apresenta uma instância de texto que é discurso de ódio e Exemplo 16.6 uma instância de texto que não é discurso de ódio contidos no *corpus* HateBR.

Exemplo 16.5:

> Vagabunda. Comunista. Mentirosa. O povo chileno nao merece uma desgraça desta

Exemplo 16.6:

> Pois é, deveria devolver o dinheiro aos cofres públicos do Brasil. Canalha.

Tal como diversas outras tarefas de aplicação de abordagens de PLN, apesar dos esforços recentes, a detecção de discurso de ódio em português fica muito atrás do inglês (Jahan; Oussalah, 2023). A detecção de discursos de ódio em língua portuguesa é, sem dúvida, uma área promissora de pesquisa no campo do PLN. Redes sociais são um terreno fértil para a disseminação de discursos de ódio em qualquer idioma. Dada a popularidade da língua portuguesa nas redes sociais, tem-se aqui uma área muito fértil para o desenvolvimento de estudos de aplicações de abordagens e PLN. A detecção de discursos de ódio

[14] https://github.com/franciellevargas/HateBR

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

em português envolve desafios únicos, como a diversidade linguística, o uso de gírias e expressões regionais, além das particularidades culturais. Isso torna a pesquisa nessa área empolgante e relevante, não apenas do ponto de vista técnico, mas também do ponto de vista social e ético.

Embora muito do que é visto em discursos tóxicos seja também discurso de ódio, cabe ressaltar que outras formas de toxicidade também são manifestas através de discursos. Há também o que chamamos de linguagem ofensiva. Diferentemente de discursos de ódio, os quais são voltados para indivíduos ou grupos específicos de pessoas com base em características identitárias, a linguagem ofensiva tem a intenção de magoar, insultar ou provocar os sentimentos das pessoas, sem necessariamente ter um objetivo discriminatório. É importante notar que a linha entre discurso de ódio e linguagem ofensiva nem sempre é clara, e o contexto em que o conteúdo é apresentado pode influenciar a percepção do quão prejudicial ele é. Isto é, todo discurso de ódio é uma linguagem ofensiva, mas nem toda linguagem que é ofensiva é também um discurso de ódio. Ambos podem ser prejudiciais e problemáticos em diferentes aspectos, e muitas vezes é necessário avaliar cuidadosamente o conteúdo para entender suas implicações e tomar medidas apropriadas para mitigar seus efeitos negativos. Tal qual os demais discursos considerados tóxicos, é importante também o desenvolvimento de meios de detecção de linguagem ofensiva.

Conjuntos de dados voltados para a detecção deste tipo de linguagem podem ser usados em um contexto que não é necessariamente de ódio. Para o português brasileiro, Trajano et al. (2023) construíram um conjunto de dados voltados a detecção de linguagem ofensiva, nomeado OLID-Br[15]. Exemplo 16.7 apresenta uma instância de texto que é insulto e sexismo e Exemplo 16.8 uma instância de texto que é insulto e ideologia contidos no *corpus* OLID-Br.

Exemplo 16.7:

> Pior do que adolescentezinhas de merda...são pessoas que levam filmes tão a sério! O livro/filme é dela, ela faz o que quiser! E por mais ruim que seja, ta rendendo milhões (:

Exemplo 16.8:

> “Heterofobismo?” Pelo que eu saiba héteros nunca foram perseguidos, mortos e torturados por centenas de anos pela sua orientação sexual. Em alguns países do Oriente, gays são presos e , as vezes, mortos só por causa de orientação sexual. Isso NUNCA aconteceu e nem vai acontecer com héteros. A implantação nas escolas da ideia de que existem casais diferentes é ótima! O conhecimento é a arma contra a ignorância e tenho certeza que vai ajudar muitas crianças confusas por aí! “Heterofobismo?” Estude mais um pouco antes de abrir a boca e passar vergonha, por favor.

Inspirado em outros *corpora* similares (do inglês, ***O**ffensive **L**anguage **I**dentification **D**atasets* ou OLID), construídos para outros idiomas, o OLID-Br reúne dados de diferentes fontes: Twitter/X, YouTube, e ainda de outros conjuntos de dados em português anotados com um esquema de anotação distinto do proposto. Os conjuntos de dados utilizados foram o OffComBR de Pelle; Moreira (2017), NCCVG[16] de Nascimento et al.

[15] https://github.com/DougTrajano/olid-br
[16] https://github.com/LaCAfe/Dataset-Hatespeech

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

(2019), HLPHSD de Fortuna et al. (2019), e ToLD-Br de Leite et al. (2020). O conjunto de dados OLID-BR contém anotações para cinco tarefas, são elas: (1) classificação de comentário tóxico: classificação binária utilizada para identificar se um comentário é ou não tóxico; (2) detecção do tipo de toxicidade: classificação multi-rótulo que identifica os rótulos de toxicidade presentes em um comentário tóxico; (3) classificação de alvo de toxicidade: classificação binária que prevê se um comentário tóxico é direcionado ou não; (4) identificação do tipo de alvo de toxicidade: classificação multiclasse que identifica o tipo de alvo de um comentário direcionado; e (5) categorização de *spans*: tarefa voltada a detecção de *spans* (parte de um texto) em um comentário tóxico. O conjunto de dados contém 6.354 (extensível para 13.538) comentários rotulados usando um esquema de anotação de três camadas com granulação fina compatível com conjuntos de dados em outros idiomas, o que permite o treinamento de modelos multilíngues.

16.3.2 Análise de Sentimento

Com a proliferação das redes sociais e das plataformas de avaliação online (tais como: TripAdvisor[17], Booking[18] e Airbnb[19]), assim como em diversos sites de e-commerce, uma infinidade de textos opinativos são publicados diariamente. Estes textos têm grande potencial para apoiar os processos de tomada de decisão (Zhang et al., 2023). A Análise de Sentimento (AS) estuda as opiniões, sentimentos, avaliações, apreciações, atitudes e emoções em relação a entidades como produtos, serviços, organizações, indivíduos, problemas, eventos, tópicos e seus diferentes aspectos expressos em textos (Liu, 2012). Nesta área desenvolvem-se aplicações em diversos campos do conhecimento como: política, finanças e marketing[20].

Existem muitos nomes e tarefas ligeiramente diferentes, por exemplo, análise de sentimento no nível de aspecto, reconhecimento/classificação de emoções etc. A AS visa encontrar soluções computacionais para extrair e analisar as opiniões das pessoas sobre uma entidade e seus diferentes aspectos (Pereira, 2021). Como as opiniões podem ser categorizadas com polaridades (por exemplo, positivo e negativo), a AS pode ser considerada uma tarefa de classificação de texto (Zhang et al., 2023). E como se trata de uma tarefa de classificação de texto, de acordo com Tan et al. (2023), três tipos de abordagens podem ser utilizadas, são elas: AM, Aprendizado Profundo (AP) e Aprendizado Conjunto (AC), usualmente referido como *ensemble learning*. Abordagens baseadas em AM, como classificador Ingênuo de Bayes (em inglês, *Naïve Bayes* ou NB) de Zhang (2004) e Máquina de Vetor de Suporte (em inglês, *Support Vector Machine* ou SVM) de Cortes; Vapnik (1995), usam modelos matemáticos para prever sentimentos. Já, as abordagens baseadas em AP, como Redes de Memória Longa de Curto Prazo (em inglês, *Long Short-Term Memory* ou LSTM) de Hochreiter; Schmidhuber (1997), utilizam Redes Neurais Artificiais para prever sentimentos. O AC combina vários classificadores para obter um melhor desempenho de AS.

No trabalho de Pereira (2021) é apresentada uma pesquisa de AS em língua portuguesa. Nele são apresentados os principais tipos de abordagens de AS, as quais podem ser baseadas em AM (classificação também proposta por Tan et al. (2023)), em léxico de sentimento, em conceitos, e híbrida. Abordagens baseadas em AM utilizam algoritmos de AM tradicionais.

[17] https://www.tripadvisor.com.br/
[18] https://www.booking.com/
[19] https://www.airbnb.com.br/
[20] Veja exemplo na área do Direito no Capítulo 12.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Já, abordagens baseadas em léxico de sentimento obtêm o grau de polaridade de opinião ou emoção de um léxico de sentimento. As abordagens baseadas em conceito usam redes de conceito (por exemplo: ontologias) para realizar a análise semântica do texto. Por fim, as abordagens híbridas, combinam as abordagens mencionadas anteriormente.

Em geral, a AS tem sido investigada principalmente em três níveis de granularidade: documento, sentença ou aspecto (Liu, 2012). No nível de documento, um sentimento é atribuído ao documento como um todo (Exemplo 16.9 com polaridade positiva). No nível de sentença, um sentimento é atribuído a cada sentença do documento (Exemplo 16.10 com polaridade positiva). No nível de aspecto, um sentimento é atribuído a cada aspecto de determinada entidade. É uma análise mais refinada, onde os aspectos podem ser atributos ou componentes de uma entidade (Exemplo 16.11 com polaridade positiva para o aspecto **café da manhã**).

Exemplo 16.9:

> O café da manhã é **incrível**. O hotel é um **ótimo** lugar para relaxar e curtir cada momento.

Exemplo 16.10:

> O café da manhã é **incrível**.

Exemplo 16.11:

> O **café da manhã** é **incrível**.

Abordagens de AS são altamente dependentes do uso de ferramentas de PLN, pois precisam interpretar textos em linguagem natural. Logo, desenvolver soluções específicas para a língua portuguesa está diretamente condicionado ao desenvolvimento de recursos linguísticos para a língua. Segundo Lo et al. (2017), o português é uma das línguas com poucos recursos linguísticos disponíveis, apesar de estar entre as línguas mais utilizadas na Web.

Dentre os recursos para o português brasileiro, podemos citar os léxicos de sentimentos: OpLexicon[21] (Souza et al., 2011b), OpenWordNet-PT[22] (De Paiva et al., 2012), SentiLex[23] (Silva et al., 2012), Reli-Lex[24] (Freitas, 2013), Brazilian Portuguese LIWC Dictionary[25] (Balage Filho et al., 2013), Word NetAffect-BR[26] (Pasqualotti, 2015), Personalitatem Lexicon (Machado et al., 2015), AffectPT-BR[27] (Carvalho et al., 2018) e LexReli (Machado et al., 2018).

- O OpLexicon (Souza et al., 2011b) possui 30.322 palavras (23.433 adjetivos e 6.889 verbos) e foi construído com base em um *corpus* do português brasileiro (composto por 346 resenhas de filmes e 970 textos jornalísticos), no *thesaurus* denominado TEP[28] (do português, Thesaurus Eletrônico Básico para o Português do Brasil) de Dias-da-Silva; Morales (2003) e no léxico de sentimento de Hu; Liu (2004) traduzido para o português.

[21] https://www.inf.pucrs.br/linatural/wordpress/recursos-e-ferramentas/oplexicon/
[22] https://github.com/own-pt/openWordnet-PT
[23] https://b2find.eudat.eu/dataset/b6bd16c2-a8ab-598f-be41-1e7aeecd60d3
[24] https://www.linguateca.pt/Repositorio/ReLi/
[25] http://143.107.183.175:21380/portlex/index.php/pt/projetos/liwc
[26] https://www.inf.pucrs.br/linatural/wordpress/recursos-e-ferramentas/wordnetaffectbr/
[27] https://github.com/LaCAfe/AffectPT-br
[28] http://www.nilc.icmc.usp.br/tep2/

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

- A base de dados da OpenWordNet-PT (De Paiva et al., 2012) é o resultado da tradução da base de dados da WordNet de Princeton[29], portanto, contém uma base de dados com grande abrangência, possui 62034 sentidos de pares de palavras e 45421 palavras únicas.
- A versão 2 do SentiLex (Silva et al., 2012) é composta por 82347 formas flexionadas, organizadas em adjetivos (16863), substantivos (1280), verbos (29504) e expressões idiomáticas (34700).
- O ReLi-Lex (Freitas, 2013) é derivado do *corpus* ReLi de Freitas et al. (2012), que é composto por resenhas de livros publicadas na internet e possui 1600 resenhas de treze livros (sete autores), este léxico contém 609 entradas.
- O Brazilian Portuguese LIWC Dictionary (Balage Filho et al., 2013) é um léxico disponível para a língua portuguesa, construído a partir do LIWC de Pennebaker et al. (2001), ou seja, foi resultado de tradução automática, utilizando diversos dicionários bilíngues português-inglês e possui 127.149 instâncias.
- O WordNetAffect-BR (Pasqualotti, 2015) é um vocabulário de emoções que possui 289 palavras (adjetivos e substantivos).
- O Personalitatem Lexicon (Machado et al., 2015) contém lexemas de conotação afetiva baseada nos traços de personalidade e foi construído com base no *Linguistic Inquiry e Word Count* (LIWC) 2.015.
- O AffectPT-BR (Carvalho et al., 2018) tem um total de 1.139 palavras atribuídas na categoria "afeto", 479 em "posemo" e 661 em "negemo".
- O LexReli (Machado et al., 2018) é uma combinação de três léxicos, OpLexicon (Souza et al., 2011b), SentiLex (Silva et al., 2012) e Brazilian Portuguese LIWC Dictionary (Balage Filho et al., 2013), especializado em identificar a polaridade de aspectos em textos opinativos sobre livros e contém 1.543 entradas.

Além dos léxicos há também os *corpora* anotados para a tarefa de AS: ReLi (Freitas et al., 2012), comentários sobre hotéis publicados no TripAdvisor (Freitas, 2015), comentários sobre produtos publicados no Buscapé[30] (Avanço; Nunes, 2014), comentários sobre restaurantes (Farias et al., 2016), TweetSentBR[31] (Brum; Nunes, 2018), UTLCorpus[32] (Sousa et al., 2019) e tweets sobre a pandemia de COVID-19 (Vargas et al., 2020). Exemplo 16.12 apresenta uma instância de texto que é positivo e Exemplo 16.13 uma instância de texto que é negativo contidos no *corpus* ReLi. Exemplo 16.14 apresenta uma instância de texto que é positivo, Exemplo 16.15 uma instância de texto que é negativo e Exemplo 16.16 uma instância de texto que é neutro contidos no *corpus* TweetSentBR. Exemplo 16.17 é um comentário sobre restaurante onde o aspecto **sobremesa** é positivo e o **atendimento** é negativo.

Exemplo 16.12:

> impossível abandonar o livro pela metade

Exemplo 16.13:

> a tradução deveria ser CRAPúsluco

[29] https://wordnet.princeton.edu/
[30] https://github.com/avanco/LBC
[31] https://metatext.io/datasets/tweetsentbr
[32] https://github.com/RogerFig/UTLCorpus

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

Exemplo 16.14:

> Essa mulher que faz voz de siri e do google tradutor é mó linda #TheNoite

Exemplo 16.15:

> Nunca fiquei tão bravo numa eliminação quanto hoje, Mas fazer o que, né? #HellsKitchenBR

Exemplo 16.16:

> Vamos acorda esse prédio!!!!!! #AltasHoras @luansantana

Exemplo 16.17:

> As **sobremesas** deste restaurante são **deliciosas**, mas o **atendimento** é muito **ruim**.

No ano de 2022, foi proposto um desafio sobre AS no nível de aspecto (em inglês, *Aspect-based Sentiment Analysis* ou ABSA) para língua portuguesa no IberLEF[33] denominado ABSAPT[34]. A proposta do ABSAPT foi inspirada em competições propostas em outros idiomas, como SemEval (Pontiki et al., 2014, 2015, 2016) para o inglês e EVALITA (Mattei et al., 2020) para o italiano. Além disso, tinha como público-alvo acadêmicos, pesquisadores e profissionais de empresas privadas. Na competição participaram cinco equipes de diferentes universidades e institutos do Brasil. O *corpus* disponibilizado na competição foi desenvolvido por Freitas (2015) e Corrêa (2021). Os participantes usaram diferentes tipos de abordagens para resolver a tarefa de ABSA, a qual foi dividida em duas, identificação de aspectos e extração de polaridade destes aspectos. O time da UFSCAR (Assi et al., 2022) propôs uma solução baseada em regras e léxico de sentimento, os times do NILC (Machado; Pardo, 2022) e da UFPR (Heinrich; Marchi, 2022) propuseram soluções baseadas em AM, utilizando algoritmos de AM tradicionais como *Conditional Random Field* (CRF) e os times Deep Learning Brasil (Gomes et al., 2022), PiLN (Ricarte Neto et al., 2022) e UFPR (Heinrich; Marchi, 2022) propuseram soluções baseada em AP, utilizando Transformers (Silva et al., 2022a). Enfim, estratégias como estas, especialmente para línguas com poucos recursos como o português, são extremamente importantes.

16.3.3 Detecção de Notícias Falsas

Uma notícia falsa é uma mensagem transmitida conscientemente por um remetente para promover uma falsa crença ou conclusão por parte do destinatário (Fuller et al., 2006). Segundo Oliveira et al. (2020), a classificação de notícias falsas pode ser vista como uma execução de uma classificação binária entre falso ou verdadeiro. A principal diferença entre a definição dos problemas de classificação de notícias falsas é em função dos diferentes esquemas de anotação ou contextos de aplicação em diferentes conjuntos de dados. Em geral, os dados são coletados de declarações anotadas em sites de verificação de fatos, com

[33] https://sites.google.com/view/iberlef2022/
[34] https://sites.google.com/inf.ufpel.edu.br/absapt2022/

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

os rótulos "verdadeiro" ou "falso". No Brasil, algumas agências de checagem são: Agência Lupa[35], Aos Fatos[36], Fato ou Fake[37] e Comprova[38].

A detecção de notícias falsas, também conhecidas como *fake news*, em redes sociais é uma área de pesquisa crítica e desafiadora. A aplicação eficaz de técnicas de PLN nesse contexto é crucial para preservar a integridade da informação online e combater a desinformação que pode trazer consequências sérias tanto na política, quanto na economia, e ainda na sociedade como um todo. O PLN desempenha um papel fundamental no desenvolvimento de abordagens eficazes para lidar com esse problema (veja Capítulo 8). Apesar do problema de disseminação de notícias falsas estar presente em todas as redes sociais, algumas tendem a ter o compartilhamento deste tipo de conteúdo mais dissipado. De acordo com Cabral et al. (2021), o Whatsapp facilita a disseminação rápida de desinformação. No Brasil, cerca de 35% das notícias falsas são compartilhadas através do WhatsApp (Newman et al., 2020), e 40,7% destes mensagens são compartilhadas após serem desmentidas (Resende et al., 2019).

Na literatura, encontramos alguns *corpora*, descritos na língua portuguesa, anotados para a tarefa de detecção de notícias falsas em língua portuguesa, são eles: COVID-19[39], FakeTweetBr[40] de Cordeiro; Pinheiro (2019), Fake.br-Corpus[41] de (Monteiro et al., 2018) e FakeWhatsApp[42] de Cunha (2021). O COVID-19 contém notícias sobre a cura da COVID-19 postadas no Twitter/X. O FakeTweetBr é um *corpus* de notícias falsas também advindo do Twitter/X. O Fake.br-Corpus contém notícias classificadas em seis grandes categorias (política, TV e celebridades, sociedade e notícias diárias, ciência e tecnologia, economia, religião) extraídas do G1[43], Folha de São Paulo[44] e Estadão[45]. O FakeWhatsApp possui mensagens anônimas do WhatsApp de grupos públicos do português brasileiro para detecção automática de desinformação textual e de usuários maliciosos. Exemplo 16.18 apresenta uma mensagem falsa e Exemplo 16.19 apresenta uma mensagem verdadeira contidas no *corpus* FakeWhatsApp.

Exemplo 16.18:

> Olhem reportagem da TV italiana RAÍ Essa matéria é de 2015 e apresenta a preocupação da comunidade científica a respeito das pesquisas feitas pelo Instituto de virulogia de Wuhan com o coronavirus de morcegos, combinando - o geneticamente com o SARS para fazê-lo mais contagioso, a fim de estudar seus efeitos.

Exemplo 16.19:

> Em novembro de 2015 a televisão italiana apresentou uma matéria que denunciava que cientistas chineses haviam desenvolvido um super virus capaz de desencadear pneumonia aguda em seres humanos.

[35] https://piaui.folha.uol.com.br/lupa/
[36] https://www.aosfatos.org/
[37] https://g1.globo.com/fato-ou-fake/
[38] https://projetocomprova.com.br/
[39] https://wp.ufpel.edu.br/midiars/datasets/
[40] https://github.com/prc992/FakeTweet.Br
[41] https://github.com/roneysco/Fake.br-Corpus
[42] https://github.com/cabrau/FakeWhatsApp.Br
[43] https://g1.globo.com/
[44] http://m.folha.uol.com.br/
[45] https://www.estadao.com.br/

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Um dos trabalhos recentes que utiliza Aprendizado Profundo (AP) na detecção de notícias falsas é o trabalho de Narde (2021). Nele foram utilizados diferentes modelos (ELECTRA de Clark et al. (2020), RoBERTa de Liu et al. (2021), XLM-R de Conneau et al. (2020), Multilingual BERT de Devlin et al. (2019) e BERTimbau de Souza et al. (2020a)) para detectar notícias falsas em redes sociais. O modelo BERTimbau (Souza et al., 2020a) com 6 épocas foi o que obteve acurácia e medida-F superior a todos outros os modelos utilizados nos experimentos, medida-F de 95%.

16.3.4 Detecção de Ironia/Sarcasmo/Humor

Apesar dos avanços na área de Análise de Sentimentos (Seção 16.3.2), ela ainda se depara com vários desafios. Entre eles, destaca-se o entendimento de figuras de linguagem (ironia/sarcasmo/humor). As figuras de linguagem são difundidas em quase qualquer gênero de texto e são especialmente comuns nos textos da Web e das redes sociais, em plataformas como o Twitter/X (Ghosh et al., 2015).

Existe uma linha tênue entre os conceitos de ironia, sarcasmo e humor. Reyes et al. (2012) define a ironia como uma "ligeira fronteira no significado do sarcasmo e da sátira". Gibbs; Colston (2001) afirmam que o sarcasmo, combinado com jocosidade, hipérbole, perguntas retóricas e eufemismo, são tipos de ironia. Tradicionalmente, a ironia é conhecida como o oposto do significado literal (Grice, 1975).

Dentre os trabalhos aplicados à tarefa de detecção de ironia para a língua portuguesa, podemos citar: Carvalho et al. (2009), Freitas et al. (2014) e Silva (2018). Nos trabalhos de Carvalho et al. (2009) e Freitas et al. (2014) são propostas pistas linguísticas para detectar ironia. Em Carvalho et al. (2009), os autores mostraram que é possível identificar opiniões irônicas em comentários, com precisão relativamente alta (de 45% a 85%), usando padrões linguísticos relativamente simples, tais como: emoticons, expressões onomatopeicas para risos, sinais de pontuação, aspas e interjeições positivas ("viva", "parabéns", "força" etc.). Ainda, em Freitas et al. (2014) padrões linguísticos foram aplicados no *corpus* sobre o assunto "fim do mundo" extraído do Twitter/X (Exemplo 16.20 de tweet irônico com emoticons e Exemplo 16.21 de tweet irônico com sinais de pontuação e expressões onomatopeicas para risos presentes no *corpus* sobre o assunto "fim do mundo").

Exemplo 16.20:

> O fim domundo não chegou ainda porque esta engarrafado naBR116. #GauchanoFimdoMundo :)

Exemplo 16.21:

> Todo mundo tirando sarro do fim do mundo...oquei!!!...mas quero ver amanhã, qualquer temporalzinhovai todo mundo se mijar de medo...hahaha

Esses padrões estão relacionados ao português brasileiro, mas a metodologia pode ser facilmente transferida para análises em outras línguas. Isso foi feito no trabalho de Freitas et al. (2020), no qual um subconjunto de padrões foi testado em *corpora* de diferentes idiomas (inglês, italiano e espanhol). Em Silva (2018), o autor descreve sobre o processo de geração de um *corpus* de ironia para a língua portuguesa, bem como, a criação de um modelo pré-treinado de uma Rede Neural Convolucional (em inglês, *Convolutional Neural*

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Network ou CNN) para detectar ironia. A CNN foi capaz de adaptar-se e detectar automaticamente as figuras de linguagem em questão. Tal abordagem mostrou-se satisfatória para detecção de ironia, obtendo medida-F de 89,78%.

No ano de 2021, foi proposto um desafio sobre detecção de ironia para língua portuguesa no IberLEF[46] denominado IDPT[47]. A proposta do IDPT foi inspirada em competições proposta em outros idiomas, como SemEval (Hee et al., 2018) para o inglês, IronITA (Cignarella et al., 2018) para o italiano, IroSvA (Bueno et al., 2019) para o espanhol e IDAT (Ghanem et al., 2019) para o árabe. Participaram da tarefa seis equipes de universidades e de empresas de quatro diferentes países: Brasil, China, Portugal e Espanha. Os *corpora* disponibilizados na competição contêm textos (tweets e notícias) sobre diferentes temas. Exemplo 16.22 de notícia irônica, Exemplo 16.23 de notícia não irônica, Exemplo 16.24 de tweet irônico e Exemplo 16.25 de tweet não irônico contidos no *corpus* do desafio IDPT.

Exemplo 16.22:

> A criatura chegou ao Rio de Janeiro e deve conceder entrevista na tarde de hoje.

Exemplo 16.23:

> A superlotação de hospitais da região metropolitana de Belo Horizonte preocupa a capital, já que a cidade é referência hospitalar da macrorregião Central e, caso não haja vagas na cidade de origem, os pacientes são transferidos para BH. No momento, cerca de 30% dos leitos de terapia intensiva e enfermaria da capital estão ocupados com pacientes de outras cidades.

Exemplo 16.24:

> Agora deu vontade de me alistar… #sqn #exercitosuperfaturaatecerveja https://t.co/Pp6vx36jzN

Exemplo 16.25:

> Agora a festa dos europeus comemorando o Mundial é contagiante hein?

O conjunto de dados de treinamento foi desenvolvido por Freitas et al. (2014), Silva (2018) e Schubert; Freitas (2020). Os participantes usaram abordagens tradicionais de AM (como: SVM, NB e outros) e/ou AP (como: Transformers). Os times que atingiram os melhores resultados foram o BERT4EVER (Jiang et al., 2021) e PiLN (Anchiêta et al., 2021). BERT4EVER (Jiang et al., 2021) utilizou Transformers e obteve uma acurácia balanceada de 92% para conjunto de dados de notícias. Para o conjunto de dados composto por tweets, a equipe PiLN (Anchiêta et al., 2021) utilizou *superficial features* e SVM e obteve uma acurácia balanceada de 52%.

[46] https://sites.google.com/view/iberlef2021/
[47] https://sites.google.com/inf.ufpel.edu.br/idpt2021/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

16.4 Considerações finais

Falar sobre aplicações de PLN em redes sociais é de grande importância por diversas razões. As redes sociais desempenham um papel fundamental na comunicação e interação social na sociedade moderna. Compreender como o PLN é aplicado nessas plataformas é essencial para entender as dinâmicas sociais e o impacto da tecnologia na vida das pessoas. Este capítulo forneceu uma visão geral sobre aplicações de abordagens de PLN em conteúdos de redes sociais. Demos ênfase ao desenvolvimento de pesquisas desenvolvidas com foco na língua portuguesa, dado o foco deste livro e de esta ser ainda uma língua com recursos escassos para algumas tarefas. Nessa primeira versão, deixamos de cobrir tópicos relevantes e atuais como reconhecimento/classificação de emoções, rastreio de transtorno mental, e detecção de postura. Reconhecendo a importância destes tópicos, pretendemos contemplá-los na versão seguinte deste livro.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Capítulo 17

PLN e Responsabilidade Social

Aplicações da FrameNet-Br

Maucha Andrade Gamonal
Lívia Vicente Dutra
Mariane de Carvalho Pinto
Tiago Timponi Torrent

Publicado em: 16/04/2026

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

17.1 Introdução

Este capítulo apresenta um panorama da aplicação da Semântica de *Frames* e da infraestrutura computacional da FrameNet Brasil em tarefas de Processamento de Linguagem Natural (PLN) voltadas à construção de sentido em aplicações práticas, como no enfrentamento à violência de gênero e na acessibilidade cultural. No Capítulo Semântica com técnicas simbólicas, as autoras discutem métodos simbólicos voltados ao processamento e à compreensão de textos em linguagem natural, apresentando três modelos como exemplos de bases de conhecimento semântico: a FrameNet (Ruppenhofer et al., 2016), a WordNet (Fellbaum, 1998) e a ConceptNet (Speer et al., 2017). A FrameNet e a WordNet, bem como suas versões em português, são classificadas como recursos léxico-semânticos, enquanto a ConceptNet é caracterizada como uma base de conhecimento de senso comum. Naquele capítulo, as principais referências teóricas e os conceitos fundamentais dessas abordagens são detalhados e contextualizados, oferecendo um suporte conceitual para as aplicações que apresentamos aqui. Assim, ao longo da Seção 17.2, retomaremos e aprofundaremos aspectos do modelo FrameNet Brasil introduzidos anteriormente, além de inserir novos conceitos necessários para sustentar as aplicações descritas.

Partindo da teoria desenvolvida por Fillmore (1982) e da FrameNet como sua aplicação computacional, discutimos como os *frames* e suas relações semânticas estruturam a compreensão linguística e servem de base para a anotação e o processamento de dados em português brasileiro. A tarefa de PLN que orienta este capítulo é a **anotação semântica baseada em *frames***, aplicada a diferentes domínios e modalidades comunicativas. Trata-se de uma tarefa central para o desenvolvimento de modelos computacionais capazes de operar sobre dados reais em contextos sensíveis como a saúde pública e a acessibilidade audiovisual.

O capítulo busca mostrar como demandas sociais do dia a dia podem ser objetos de aplicações computacionais ao envolverem desde a identificação de padrões de violência de gênero em registros de saúde até a avaliação de traduções intersemióticas em produtos audiovisuais. A primeira parte do texto apresenta os fundamentos teóricos da Semântica de *Frames*, detalhando o funcionamento da FrameNet Brasil e seus refinamentos semânti-

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

cos, como as Relações Qualia Ternárias[1] que se dão por meio da ferramenta de anotação Webtool. Em seguida, são discutidos dois estudos de caso que ilustram o potencial da anotação semântica para enfrentar problemas reais.

O primeiro estudo está relacionado ao projeto **DataToStopGBV**, voltado à identificação de casos de violência baseada em gênero a partir de registros do sistema público de saúde. Por meio da modelagem de domínios específicos (saúde e violência), da modelagem de *frames* e da anotação manual e automática dos dados, o projeto mostra como a estrutura semântica da FrameNet pode ser empregada para apoiar políticas públicas de prevenção à violência.

O segundo estudo apresenta o *dataset* ***Audition***, um recurso multimodal anotado semanticamente com o objetivo de apoiar a geração e a avaliação de traduções audiovisuais acessíveis, como a audiodescrição. A partir da articulação entre modos comunicativos, como áudio original, vídeo e audiodescrição, o *Audition* permite análises de similaridade semântica e contribui para o avanço de tecnologias assistivas baseadas em teoria linguística.

Apesar de compartilharem a mesma base teórica e metodológica – a Semântica de *Frames* e a FrameNet Brasil –, os dois estudos de caso exploram abordagens de anotação semântica distintas: enquanto o projeto **DataToStopGBV** faz uso de **anotações semânticas automáticas**, com apoio de modelos supervisionados, o *dataset* ***Audition*** foi construído com **anotação manual** especializada sobre os dados multimodais. Essa diferença reflete tanto os objetivos como as particularidades de cada tarefa.

Ao integrar teoria linguística, recursos computacionais e compromisso social, o capítulo demonstra como a Semântica de *Frames* pode orientar o desenvolvimento de ferramentas e metodologias eficazes para o processamento semântico em português brasileiro, com aplicações relevantes na saúde pública e na acessibilidade cultural.

17.2 Semântica de *Frames*

Como apresentado no Capítulo E o significado?, a Semântica ocupa-se do estudo do significado expresso por palavras, expressões e enunciados. Dentro desse campo, a Semântica de *Frames,* também discutida na Subseção FrameNet, é uma teoria desenvolvida por Charles J. Fillmore (Fillmore, 1982), que propõe uma maneira particular de se entender o significado lexical. Segundo essa abordagem, compreender o sentido de um item lexical envolve mais do que conhecer sua definição no dicionário: é necessário também ativar um conjunto de conhecimentos socialmente compartilhados sobre o mundo e sobre as situações em que ele costuma ser usado. Esse conjunto é chamado, na teoria, de *frame*.

Na definição de Fillmore (1982, p. 111), um *frame* é "qualquer sistema de conceitos relacionados de tal maneira que, para entender qualquer um deles, é preciso entender toda a estrutura na qual ele se insere"[2]. Em outras palavras, para o autor, o significado de uma palavra está associado a uma rede de conceitos interligados, que funciona como uma espécie de pano de fundo conceitual para a construção de sentido. Essa rede de conceitos, os *frames*, está ancorada na forma como experienciamos e categorizamos o mundo ao nosso redor e possibilita a atribuição de significado no momento da comunicação.

[1] Conjunto de relações semânticas inspiradas na Teoria do Léxico Gerativo (Pustejovsky, 1995), utilizadas na FN-Br para modelar aspectos internos dos conceitos, como finalidade, composição e origem. Este tópico será abordado na Seção 17.3.2 deste capítulo.

[2] "[...] By the term 'frame' I have in mind any system of concepts related in such a way that to understand any one of them you have to understand the whole structure in which it fits." [Tradução nossa]

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Ao ouvir uma palavra como "medicar", por exemplo, não pensamos apenas na ação de administrar um remédio mas também em tudo que geralmente a acompanha: há um profissional de saúde, um medicamento, uma condição a ser tratada. O interessante é que, mesmo quando esses elementos não aparecem explicitamente na frase, eles ainda são mobilizados para a sua compreensão. Se alguém diz "O plantonista medicou a criança", por exemplo, já se pressupõe que a criança estava com algum sintoma ou doença, que foi administrada uma substância com fins terapêuticos e que esse ato se deu num contexto de cuidado em saúde.

Isso acontece porque vivemos em uma sociedade em que as práticas de atenção à saúde são parte do cotidiano, ou seja, esse tipo de situação bem como os participantes e as circunstâncias que a compõem integram nosso repertório sociocultural. Assim sendo, quando o verbo "medicar" é usado, ele ativa uma rede de conhecimentos compartilhados, isto é, o *frame* que permite tal entendimento. Na Semântica de *Frames*, diz-se que o verbo "medicar" é uma Unidade Lexical que **evoca** um *frame* do cenário da saúde.

Ao destacar os conhecimentos compartilhados que sustentam o significado das palavras, a Semântica de *Frames* amplia a visão tradicional sobre a compreensão linguística. Em vez de analisar apenas sentenças isoladas, essa abordagem valoriza o papel da experiência humana e do contexto cultural para explicar como usamos e entendemos a linguagem no cotidiano. Por isso, Fillmore (1985) caracteriza essa teoria como uma **semântica da compreensão**, em contraste com a semântica das condições de verdade, que se concentra em avaliar se uma sentença é verdadeira ou falsa com base na realidade. Para ele, a linguagem não é simplesmente um espelho do mundo, mas uma ferramenta por meio da qual construímos e negociamos sentidos, ancorados na cognição e nas práticas culturais. Por isso, de modo geral, a teoria propõe uma pergunta central: como as pessoas organizam sua experiência por meio da linguagem? Para respondê-la, os estudos em Semântica de *Frames* analisam como as Unidades Lexicais são utilizadas em contextos autênticos e quais *frames* são acionados nesses usos.

Desde 1997, essa abordagem tem fundamentado a construção da FrameNet[3] (Fillmore et al., 2003), projeto de lexicografia computacional que descreve o léxico com base nos *frames* evocados pelas ULs em ocorrências reais extraídas de *corpora*. Fundada por Charles J. Fillmore em 1997, no International Computer Science Institute (ICSI), localizado em Berkeley, Califórnia, a iniciativa se apoia na Semântica de *Frames* (Fillmore, 1982; Fillmore, 1985) e tem como objetivo investigar como os sentidos são construídos nas línguas naturais. O projeto tem como foco o desenvolvimento de um banco de dados lexical capaz de descrever a língua inglesa sob uma perspectiva semântica e sintática, utilizando exemplos reais extraídos de *corpora*. Como desdobramento desse trabalho, Fillmore também contribuiu para a proposição do ***Constructicon***[4], recurso voltado à descrição de construções gramaticais que evocam *frames*, e que busca representar a integração entre forma e sentido em estruturas mais complexas e idiomáticas.

Com o tempo, os recursos criados pela FrameNet começaram a ser adotados em diferentes áreas, incluindo o avanço de aplicações em linguística computacional, especialmente em tarefas de PLN, e o projeto foi expandido para outras línguas. Um dos principais exemplos dessa expansão é a FrameNet Brasil (FN-Br), um laboratório de linguística computacional

[3]A ferramenta de consulta da FrameNet original, bem como informações sobre os dados anotados, estão disponíveis em: https://framenet.icsi.berkeley.edu/.

[4]Seu desenvolvimento foi adotado por diferentes framenets, como as do japonês, sueco e português brasileiro, embora não tenha avançado da mesma maneira no projeto original em inglês. A versão do *Construction* da FrameNet Brasil pode ser consultada em: https://webtool.frame.net.br/report/cxn.

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

vinculado à Universidade Federal de Juiz de Fora (UFJF), em Minas Gerais. O projeto brasileiro desenvolve anotações semânticas para o português brasileiro e também propõe extensões à base de *frames* e à estrutura de dados originalmente concebidas pela Berkeley FrameNet.

17.3 FrameNet Brasil

Para compreender a metodologia de anotação utilizada pela FN-Br, é fundamental conhecer as nomenclaturas essenciais ao projeto. Algumas foram apresentadas na Subseção FrameNet e outras serão detalhadas ao longo desta seção. A primeira delas é o ponto de partida de toda anotação, a **Unidade Lexical** (UL), que corresponde à associação de um lema ao *frame* evocado por ele. Na FN-Br, o lema "medicar.v" está associado ao *frame* `Intervenção_em_saúde`[5], apresentado na Figura 17.1.

Figura 17.1: O *frame* `Intervenção_em_saúde`



Fonte: Webtool 4.1

Na Figura 17.1, é possível visualizar a estrutura básica de um *frame* na interface gráfica

[5]De acordo com a metodologia das FrameNets, os nomes dos *frames* são grafados em fonte `Courier`.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

da FrameNet Brasil, a **Webtool**[6] (Torrent et al., 2024), plataforma responsável pelo gerenciamento do banco de dados da FN-Br. Na ferramenta, os anotadores têm acesso aos *frames* e às ULs do banco de dados e podem interagir com esses elementos diretamente na interface, o que facilita a análise e a anotação.

No topo da página de cada entrada, é exibido o nome do *frame*, seguido por uma breve definição que o caracteriza. No caso do *frame* `Intervenção_em_saúde`, por exemplo, trata-se de um evento em que um profissional da área da saúde realiza uma intervenção médica, como administrar um medicamento, com o objetivo de tratar ou aliviar uma condição ou doença. Logo abaixo da definição, são apresentados os participantes, os instrumentos ou as circunstâncias que compõem a cena representada pelo *frame*, denominados de **Elementos de Frame** (EFs). Na FrameNet, os EFs são divididos em dois grupos: os **nucleares**, que são essenciais para a composição do *frame*, e os **não nucleares**, que desempenham funções circunstanciais.

No *frame* `Intervenção_em_saúde`, são apresentados como EFs nucleares o Profissional_em_saúde[7], que realiza os procedimentos, o Paciente, que sofre a condição ou doença, a Condição_em_saúde, que motiva a intervenção, a própria Intervenção e o Resultado obtido. Os outros elementos que aparecem na Figura 17.1, como Efeitos_colaterais e Tempo, são considerados não nucleares.

Para realizar a anotação de texto corrido[8] – modelo de anotação mais seguido atualmente na FN-Br –, os anotadores do projeto recebem lotes de sentenças para análise. Conforme a metodologia descrita por Ruppenhofer et al. (2016), que fundamenta o modelo adotado pela FrameNet, a tarefa central nesse tipo de anotação é identificar, em cada sentença, as ULs e os *frames* que elas evocam. Na anotação, ao selecionar uma UL na interface, são exibidos os *frames* associados a ela na base de dados. O anotador, então, lê as descrições desses *frames* para escolher o mais adequado ao contexto da sentença e, a partir dessa escolha, o sistema gera uma camada para marcar os Elementos de *Frame* (EFs), que devem ser destacados manualmente no texto.

Durante o processo de anotação, pode acontecer de EFs nucleares não serem instanciados explicitamente como material linguístico nos enunciados em análise. Quando isso acontece, dizemos que houve uma **instanciação nula**. Na FrameNet, esse fenômeno é dividido em três tipos: **Instanciação Nula Definida** (DNI, do inglês *Definite Null Instantiation*), **Instanciação Nula Indefinida** (INI, do inglês *Indefinite Null Instantiation*) e **Instanciação Nula Construcional** (CNI, do inglês *Constructional Null Instantiation*).

A primeira delas, a DNI, diz respeito à omissão de um EF que pode ser recuperado com base no contexto. Na sentença do Exemplo 17.1[9], por exemplo, o verbo "observar" está anotado no *frame* `Percepção_ativa`, que tem dois EFs nucleares: o Perceptor_agentivo, que é aquele que realiza a ação de observar, e o Fenômeno, que é aquilo que é percebido.

Exemplo 17.1:

> Um cachorro **observa**[PERCEPÇÃO_ATIVA] de longe.

Exemplo 17.2:

[6] A ferramenta de anotação semântica WebTool, bem como os dados anotados, estão disponíveis em: https://webtool.frame.net.br/. O repositório oficial da FrameNet Brasil no Github pode ser acessado em: https://github.com/FrameNetBrasil.

[7] Seguindo a metodologia das FrameNets, nomes de EFs são apresentados em Versalete.

[8] Ao longo do texto, as Figuras 17.9 e 17.14 discutem exemplos de anotações de texto corrido.

[9] As sentenças que aparecem ao longo desta seção para ilustrar os conceitos discutidos advêm de curtas-metragens do Audition e foram anotadas por meio da Webtool.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

As pessoas descem.

A sentença do Exemplo 17.1, no entanto, expressa apenas o PERCEPTOR_AGENTIVO ("um cachorro") enquanto o FENÔMENO está ausente. Como aquilo que é observado ("as pessoas") pode ser facilmente recuperado pelo contexto discursivo – mais especificamente pela sentença do Exemplo 17.2, que ocorre logo antes no *corpus* –, essa omissão é classificada como DNI.

Já a INI ocorre quando um EF nuclear não está explicitamente expresso na sentença nem pode ser determinado com base no contexto imediato. Isso faz com que o elemento ausente seja interpretado de maneira genérica ou indefinida. Na sentença do Exemplo 17.3, por exemplo, o *frame* `Pessoas_por_idade` é evocado pela UL "crianças.n"[10], e um de seus EFs nucleares é a IDADE, que, nesse caso, não está lexicalmente realizado nem pode ser inferido com precisão a partir do contexto. No entanto, como o *frame* é formado por ULs que já pressupõem uma faixa etária aproximada, a ausência do EF IDADE pode ser considerada INI e DNI. Isso demonstra que, em certos casos, há sobreposição entre os dois tipos de instanciação nula, exigindo decisões interpretativas do anotador.

Exemplo 17.3:

As **crianças**$^{\text{PESSOAS_POR_IDADE}}$ se reúnem ao redor do globo.

A Instanciação Nula Construcional (CNI), por sua vez, ocorre quando a própria estrutura da língua licencia a omissão de um EF nuclear. No português brasileiro, isso acontece, por exemplo, nas sentenças com sujeito desinencial ou nas construções imperativas. É o caso da sentença do Exemplo 17.4, em que, por se tratar de uma oração com sujeito desinencial, a expressão do agente não é obrigatória no português. Nesse caso, o verbo "brincar" evoca o *frame* `Atividade`, cujo único EF nuclear é o AGENTE, ou seja, aquele que realiza a ação. No exemplo, no entanto, o AGENTE da ação está lexicalmente omitido pela estrutura. Por isso, essa omissão é classificada como uma ocorrência de CNI.

Exemplo 17.4:

Brincam$^{\text{ATIVIDADE}}$ de lançar as bolinhas dentro da lata.

Existem casos, ainda, em que a própria UL que evoca o *frame* já inclui um dos EFs em seu radical, caracterizando a ocorrência de Incorporação (INC). Na sentença do Exemplo 17.5, por exemplo, o lema "elefante" está associado ao *frame* `Animais`, cujo EF nuclear é o próprio animal referido. Nesse caso, não há um termo separado na sentença que realize esse EF, pois ele já está presente na própria UL "elefante.n", configurando assim uma INC (do inglês, *incorporation*).

Exemplo 17.5:

O **elefante**$^{\text{ANIMAIS}}$ azul.

Essa questão da ausência linguística ganha novos contornos quando se consideram dados multimodais, como imagens e vídeos. Em contextos audiovisuais, elementos que parecem ausentes no texto verbal muitas vezes estão presentes na imagem, o que desafia os limites

[10] Na nomenclatura das FrameNets, antes do ponto é indicado o lema e, após o ponto, a classe gramatical: .n (substantivo, *noun*), .v (verbo, *verb*), .a (adjetivo, *adjective*), entre outros.

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

da anotação textual tradicional. Estudos realizados em nosso laboratório, como o de Belcavello (2023), demonstraram que objetos visuais também são capazes de evocar *frames* ou complementar o que foi evocado pela narração. Diante disso, tornou-se necessário desenvolver um tipo específico de anotação que integrasse essas diferentes modalidades semióticas. Essa anotação multimodal será abordada na última seção deste capítulo, em que se destaca sua importância para ampliar a representação semântica na FrameNet Brasil e, inclusive, repensar o escopo das instanciações nulas.

Considerando a estrutura interna dos *frames* e de seus elementos apresentada até aqui, a FrameNet Brasil categoriza os *frames* em cinco grandes grupos, definidos a partir da natureza do conteúdo semântico que representam (Torrent et al., 2022). No Lutma[11], recurso de criação de *frames* da FrameNet Brasil para não especialistas, eles são definidos da seguinte forma:

(i) Entidade: um conceito, uma coisa, um ser vivo.

(ii) Evento: envolve pelo menos um participante e ocorre em um local e hora.

(iii) Estado: uma situação estável.

(iv) Relação: uma relação entre duas ou mais entidades, eventos ou estados.

(v) Atributo: um atributo com um valor é aplicado a uma entidade.

Na prática, esses grupos podem ser ilustrados com exemplos concretos presentes na própria base da FN-Br. O grupo **Entidade**, por exemplo, inclui o *frame* `Pessoas`, que é evocado por ULs gerais para indivíduos (como "povo.n" e "gente.n"). O grupo **Evento** pode ser representado por `Fazer_exame_em_Saúde`, que diz respeito ao ato de realizar um exame médico e é evocado por ULs como "aferir.v" e "coletar.v". Já o grupo **Estado** é exemplificado por `Condições_em_Saúde`, formado por ULs como "diabético.a" e "cardiopata.a". O grupo **Relação**, por sua vez, inclui *frames* como `Parentesco`, que reúne ULs que expressam relações familiares (como "irmão.n" e "mãe.n"), enquanto o grupo **Atributo** pode ser exemplificado com o *frame* `Cor`, que associa características como "azul.a" e "acinzentado.a" a uma entidade.

Além de contribuir para a sistematização da base de dados da FrameNet Brasil, essa divisão também facilita a compreensão do conteúdo e do uso de cada *frame*, antecipando o tipo de estrutura conceitual que será encontrada. Porém, para entender como os *frames* se conectam e formam uma rede, é preciso olhar para as relações entre os *frames*, que serão discutidas a seguir.

17.3.1 As relações *frame-a-frame*

Na FrameNet, cada *frame* faz parte de uma grande rede de significados. Isso quer dizer que, ao modelar um novo *frame*, é possível conectá-lo a outros já existentes, formando um sistema interligado. Essas conexões são chamadas de **relações *frame*-a-*frame*** e são fundamentais para mostrar como eles se organizam e se relacionam na rede.

As principais categorias dessas relações são: Herança, Subframe, Uso, Precedência, Causativo_de, Incoativo_de e Perspectiva. Cada uma delas indica um tipo específico de ligação entre os *frames* – como quando um *frame* mais geral se conecta a um mais específico, por exemplo, ou quando dois *frames* representam diferentes pontos de vista de uma

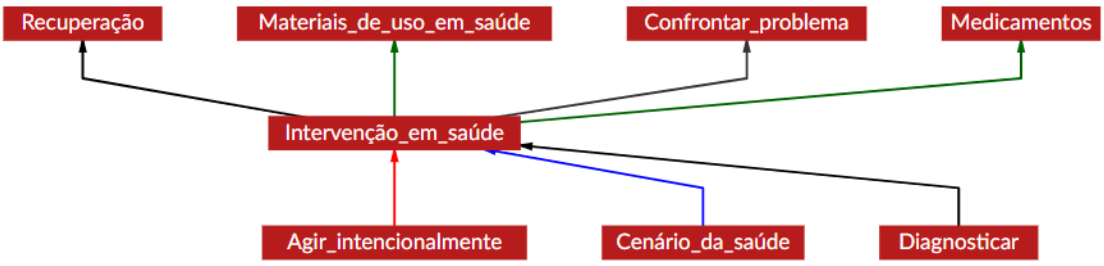
[11] Disponível em: https://lutma.frame.net.br/. Acesso em: 28 jul. 2025.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

mesma situação. Na FrameNet Brasil, essas relações podem ser visualizadas por meio do **Grapher**, uma ferramenta gráfica disponível na WebTool. No Grapher, cada tipo de relação aparece com uma cor diferente, facilitando a identificação e o estudo dessas conexões. A Figura 17.2, por exemplo, ajuda a visualizar como o *frame* `Intervenção_em_saúde` se conecta a outros na base.

Figura 17.2: Relações do *frame* `Intervenção_em_saúde`



Fonte: Webtool 4.1

A relação de **Herança** (*Inheritance*), representada pela seta vermelha, indica quando um *frame* filho herda as características de um *frame* mãe, elaborando um ou mais de seus EFs. Um exemplo disso é `Intervenção_em_saúde`, que herda as características de `Agir_intencionalmente` (veja Figura 17.3).

Figura 17.3: O *frame* `Agir_intencionalmente`



Fonte: Webtool 4.1

Acerca disso, em `Agir_intencionalmente`, o EF Pessoa é definido de forma genérica, como qualquer indivíduo que realiza uma ação. Já no *frame* `Intervenção_em_saúde`, esse papel é especificado: o agente da ação passa a ser o Profissional_em_Saúde, caracterizado como alguém que atua com a intenção de melhorar a condição clínica de um paciente. A mesma situação se dá com outros elementos herdados: o que antes era simplesmente a Ação, por exemplo, passa a ser identificada como a Intervenção realizada. Essa

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

especificação ajuda a representar de forma mais detalhada o tipo de situação descrita pelo novo *frame*.

A relação de **Subframe** (*Subframe*), indicada pela seta azul, conecta um *frame* que codifica um evento complexo a outros que representam subeventos delimitáveis (ou seja, partes) desse evento. É o caso de `Cenário_da_saúde`, que representa situações em que um paciente precisa de cuidados médicos. Na FrameNet Brasil, esse cenário está relacionado, por meio da relação de Subframe, a *frames* como `Diagnosticar`, `Intervenção_em_saúde` e `Recuperação`, que descrevem, respectivamente, as etapas de diagnóstico, realização da intervenção e recuperação do paciente, momentos específicos desse cenário.

Os *frames* `Diagnosticar`, `Intervenção_em_saúde` e `Recuperação` se articulam também por meio da relação de **Precedência** (*Precedes*), indicada pelas setas pretas na Figura 17.2, que organiza os eventos em uma sequência temporal. No caso, `Diagnosticar` ocorre antes da intervenção, que é seguida pela `Recuperação`, formando assim uma ordem cronológica que reflete o processo dos cuidados médicos ao longo do tempo.

Por sua vez, a relação de **Uso** (*Using*), indicada pelas setas verdes, ocorre quando a compreensão de um *frame* depende de outro que funciona como seu contexto de fundo. É o caso de `Medicamentos` e `Materiais_de_uso_em_saúde`, por exemplo, que pressupõem a cena da `Intervenção_em_saúde` para serem plenamente compreendidos, visto que tanto os medicamentos quanto os materiais são utilizados no contexto de uma ação médica. Assim, `Intervenção_em_saúde` fornece o cenário necessário para interpretar o uso desses elementos, dando sentido às práticas que envolvem seu manuseio ou aplicação.

As relações Causativo_de (*Causative_of*), Incoativo_de (*Inchoative_of*) e Perspectiva (*Perspective_on*), por outro lado, não podem ser demonstradas a partir do *frame* `Intervenção_em_saúde`. Por isso, para exemplificá-las, é necessário recorrer a outros *frames*.

A relação **Causativo_de**, representada por uma seta amarela no Grapher, indica um vínculo de causalidade entre dois *frames*. Por exemplo, na FN-Br, o *frame* `Causar_condição_em_saúde`, que indica a ação de causar (intencionalmente ou não) uma condição em saúde, é causativo do *frame* `Condições_em_saúde`, que reúne ULs relacionadas a doenças ou condições que um paciente apresenta.

Já a relação **Incoativo_de**, representada por uma seta marrom, marca uma mudança de estado. O *frame* `Condições_em_saúde`, por exemplo, tem como incoativo o *frame* `Tornar-se_paciente_em_saúde`, que descreve a transição de um indivíduo, a partir de um diagnóstico, para a condição de estar doente.

A relação de **Perspectiva**, por fim, acontece quando diferentes *frames* representam a mesma situação sob pontos de vista distintos. Ela pode ser exemplificada a partir do *frame* `Emoções`, que descreve um experienciador em determinado estado emocional, geralmente provocado por algum estímulo. A partir dele, surgem duas perspectivas distintas: `Foco_no_estímulo`, que enfatiza o estímulo responsável por causar a emoção, e `Emoção_com_foco_no_experienciador`, que destaca o ponto de vista de quem sente a emoção em relação ao que a provocou.

As relações *frame*-a-*frame* ajudam a entender como os *frames* se organizam e se ligam dentro da rede semântica da FrameNet Brasil, mostrando diferentes formas de associação entre eles. No entanto, nem todas as conexões importantes acontecem nesse nível. Em muitos casos, é preciso observar ligações mais específicas.

Nesse sentido, para ampliar o refinamento semântico da base, a FrameNet Brasil incluiu novos tipos de relações. Um exemplo é a **relação EF-*frame*** (FE-to-*frame*), que modela o fato de que um EF, em determinado *frame*, pode estar conceitualmente associado a

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

outro *frame* na rede. Em outras palavras, essa relação expressa que certos EFs, ao serem instanciados por ULs, tendem a evocar outros *frames*. Assim, se `Intervenção_em_saúde` inclui o EF Paciente, como vimos, a relação EF-*frame* pode indicar que esse Paciente frequentemente remete ao *frame* `Pessoa`, já que, em geral, os pacientes são pessoas.

Nesse mesmo esforço de ampliar a capacidade descritiva da rede, também foi proposta a **relação metonímica**. Gamonal (2017) propõe que a metonímia seja representada na FN-Br por meio de relações entre EFs de um mesmo *frame*, permitindo capturar casos em que um EF substitui outro com base em inferências pragmáticas apoiadas em conhecimento compartilhado. Por exemplo, em uma sentença como "O Hospital Central atendeu dezenas de pacientes ontem", a expressão "Hospital Central" é usada metonimicamente para se referir à equipe de profissionais da saúde que prestam o atendimento naquele local.

Para representar essa relação na base, primeiramente, mapeia-se que o EF Profissional_de_saúde – no *frame* `Atendimento_em_saúde` – pode ser instanciado em `Profissionais_em_saúde`; em seguida, estabelece-se uma relação metonímica entre os EFs Profissional e Lugar_de_trabalho nesse *frame*, indicando que o local onde o profissional atua pode ser usado para representá-lo. Temos primeiramente uma relação entre *frames* (EF - *Frame*) seguida de uma relação metonímica considerada *intraframe* por se dar entre EFs de um mesmo *frame*. Na literatura da Linguística Cognitiva, considera-se que haja um mapeamento "intradomínio".

Além dessas, outras relações foram desenvolvidas com foco nas conexões entre ULs de *frames* distintos, a fim de capturar nuances semânticas importantes, principalmente em domínios mais especializados. Nesse cenário, ganham destaque as relações qualia. Elas foram incorporadas pela FN-Br com base na proposta de Pustejovsky (1995) e servem para capturar sentidos mais detalhados, especialmente quando se trata de áreas mais técnicas ou específicas. Como esse tipo de relação desempenha um papel central na proposta de anotação discutida neste trabalho, ele será abordado de forma mais detalhada na próxima seção.

17.3.2 As relações qualia ternárias

Além das relações *frame*-a-*frame*, que foram inicialmente desenvolvidas pela Berkeley FrameNet, a FrameNet Brasil adicionou um novo tipo de relação semântica à sua base: as relações qualia. Elas foram implementadas com base nos estudos de Pustejovsky (1995), cuja ideia central é a de que o significado de uma palavra pode depender de diferentes formas pelas quais as pessoas entendem e relacionam os objetos ao seu redor. Para Pustejovsky (1995), essas formas são organizadas em quatro tipos de relações, definidas como apresentado no Quadro 17.1.

Quadro 17.1: Relações qualia ternárias (Torrent et al., 2022, p. 8)[12]

Formal – descreve a categoria básica do item e fornece a informação que distingue uma entidade dentro de um conjunto maior dentro de seu domínio semântico. Constitutivo – expressa uma variedade de relações sobre a constituição interna de uma entidade.

[12]Formal – describes the basic category for the item and provides the information that distinguishes an entity within a larger set inside its semantic domain.
Constitutive – expresses a variety of relations concerning the internal constitution of an entity.
Telic – concerns the typical function or purpose of an entity, i.e., what the entity is for.
Agentive – concerns the origin of an entity, its creator, or its coming into being. [Tradução nossa]

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

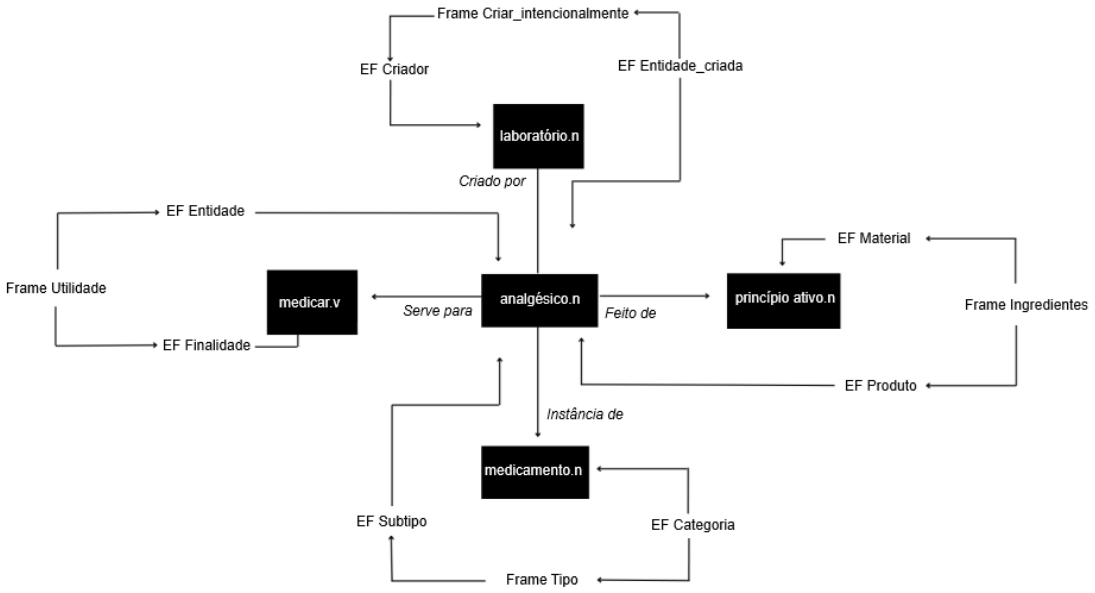
Télico – diz respeito à função ou propósito típico de uma entidade, ou seja, para que a entidade serve.
Agentivo – refere-se à origem de uma entidade, seu criador ou seu processo de surgimento.

Essas quatro dimensões – formal, constitutiva, télica e agentiva – ajudam a explicar como os sentidos das ULs se constroem e se conectam no uso da linguagem. Na FN-Br, foi adotada uma forma própria de representar essas relações semânticas em seu banco de dados. Em vez de simplesmente indicar que duas Unidades Lexicais estejam conectadas por uma relação qualia, o projeto propôs uma modelagem mais detalhada, utilizando *frames* para mediar essas conexões. Essa abordagem, chamada de **qualia ternária**, foi desenvolvida por Costa (2020) e permite representar essas relações de forma mais precisa e estruturada.

Nesse modelo, cada relação qualia é representada dentro de um *frame* específico, que funciona como um mediador para a conexão entre duas ULs. Segundo Belcavello et al. (2020), as duas ULs envolvidas são associadas a diferentes EFs num mesmo *frame*. Assim, o *frame* estabelece uma ponte semântica entre os termos, deixando claro que tipo de relação existe entre eles e quais papéis cada um desempenha. Outro aspecto importante dessa modelagem é a direção da relação: ela é sempre unidirecional, ou seja, vai de uma Unidade Lexical para outra, e não o contrário.

Como apresentado por Dutra (2024), a aplicação das relações qualia pode ser ilustrada com o exemplo da UL "analgésico.n". No banco de dados da FN-Br, essa unidade está conectada a outras cinco, por meio de diferentes tipos de relações qualia, e cada uma dessas conexões é mediada por um *frame* específico, como mostra a Figura 17.4.

Figura 17.4: As relações qualia a partir da UL *analgésico.n*



Fonte: Adaptado de (Dutra, 2024)

Por exemplo, a **relação agentiva** liga *analgésico.n* à UL "laboratório.n" usando o *frame* `Criar_intencionalmente`. Nesse caso, o analgésico é entendido como aquilo que

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

é criado (EF Entidade_Criada), enquanto o laboratório aparece como o agente da ação (EF Criador). A **relação constitutiva**, por outro lado, conecta "analgésico.n" e "princípio ativo.n" por meio do *frame** `Ingredientes`. *Nele, o"analgésico.n" ocupa o papel de Produto, e "princípio ativo.n", o de Material, indicando do que o analgésico é feito. Já a* ***relação formal*** *é modelada no* frame* `Tipo`, em que "analgésico.n" aparece como um Subtipo e "medicamento.n" como uma Categoria, representando a ideia de que analgésico é um tipo de medicamento. Por fim, a **relação télica** envolve o *frame* `Utilidade`. Nesse caso, "analgésico.n" é associado ao EF Entidade (ou seja, um item com uma finalidade), enquanto a UL "medicar.v" representa sua Finalidade.

A organização das relações na FN-Br, tanto no nível dos *frames* quanto das Unidades Lexicais, constitui a base do trabalho de anotação desenvolvido no projeto. Para realizar essas anotações de forma padronizada e organizada, a equipe utiliza a WebTool. Na próxima seção, apresentamos um exemplo prático dessa aplicação por meio de um projeto que utiliza a FrameNet Brasil para apoiar pesquisas sobre o enfrentamento da violência contra a mulher.

Exercício prático: Explorando as relações qualia ternárias

Agora é o momento de aplicar os conceitos discutidos! Acesse a lista de *frames* disponíveis na WebTool e escolha um que lhe pareça interessante. Em seguida, selecione uma das Unidades Lexicais no *frame* (listadas logo após os EFs) e, com base nela, reflita sobre os aspectos que compõem seu significado. Para isso, **associe a UL escolhida às relações qualia que forem possíveis**. Para cada dimensão, pense em uma UL que esteja semanticamente relacionada à original e, sempre que possível, pesquise essa nova UL na WebTool para verificar a que *frame* ela pertence – ou seja, que *frame* poderia ser responsável por **mediar** essa ligação entre as duas ULs.

Exemplo: A UL "pizza.n" pertence ao *frame* `Criação_culinária` e se relaciona com "farinha.n", que pertence ao *frame* `Ingredientes`. Essas duas ULs estão conectadas por meio desse *frame*, que faz a mediação de uma relação qualia do tipo **constitutiva**, já que *farinha* é parte da composição da *pizza*.

Para ajudar no processo, as perguntas do Quadro 17.2 foram organizadas segundo cada uma das dimensões qualia e podem servir como guia para sua análise.

Quadro 17.2: Perguntas para cada uma das dimensões qualia

Relação qualia	Pergunta-guia
Construtiva	Do que isso é feito? Quais são seus componentes?
Formal	O que o define como sendo daquele tipo?
Télica	Para que serve? Qual é sua finalidade ou função?
Agentiva	Quem o produziu ou como foi criado?

17.4 Projeto DataToStopGBV: Semântica contra a violência de gênero

O **projeto DataToStopGBV** é uma iniciativa que nasceu da colaboração entre Vital Strategies Brasil e a FrameNet Brasil em 2023. A Vital Strategies Brasil é a filial brasileira

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

da organização global de saúde Vital Strategies, que trabalha em conjunto com o governo e a sociedade buscando implementar medidas que minimizem desafios ligados a questões de saúde pública. A parceria entre a **Vital Strategies Brasil** e a **FrameNet Brasil** se deu com uma intenção clara: desenvolver ferramentas para alerta precoce e intervenção em casos de violência de gênero, que pudessem oferecer maior apoio às equipes de saúde, autoridades locais e criadores de políticas públicas.

Este projeto foi também o foco principal da pesquisa de mestrado "Evaluating The Contribution of Framenet to Gender-Based Violence Identification" (Dutra, 2024). Esta seção apresenta um panorama das discussões iniciadas no âmbito dessa pesquisa.

17.4.1 Contexto e motivação

A **Violência baseada em gênero (VBG)** afeta diversos grupos sociais, no entanto a incidência de casos da violência direcionada às mulheres faz com que este seja o maior foco de pesquisas e normalmente delimita as mulheres como grupo de foco. Seguindo, então, a mesma linha que a Organização Mundial da Saúde (OMS) (WHO, 2024) e o Instituto de Igualdade de Gênero Europeu (EIGE) (EIGE, 2024), o projeto também optou por focar neste grupo em um primeiro momento. Este não é o único ponto de alinhamento entre esta pesquisa e órgãos internacionais que visam aos direitos humanos. A compreensão de que a VBG é uma questão de saúde pública também é defendida pela OMS e corroborada por diversos outros estudos que julgam que a busca por soluções para a sua prevenção é uma prioridade (Garcia-Moreno; Watts, 2011; Öhman et al., 2020; Sweet, 2014).

No Brasil, profissionais da saúde desempenham papel estratégico na identificação e registro de casos de VBG, sendo obrigados por lei a notificar episódios suspeitos ou confirmados ao Sistema de Informação de Agravos de Notificação (SINAN). No entanto, mesmo com a obrigatoriedade legal, o número de subnotificações ainda é alto. Alguns estudos buscam entender as causas que levam a este cenário e evidenciam que o problema parte da falta de preparo para lidar com tais casos, desde o atendimento até a notificação. Entre as principais razões para a subnotificação se destacam o medo de um possível confronto ou represália e também entraves estruturais, tais como a ausência de protocolos, a falta de mecanismos legais de proteção para os profissionais responsáveis pela notificação, dificuldades na identificação de sinais de violência e o conflito com o sigilo profissional (Garbin et al., 2015; Kind et al., 2013).

Considerando este cenário, Vital Strategies Brasil e FrameNet Brasil propuseram o projeto DataToStopGBV, que visa ao desenvolvimento de recursos que possam levar ao alerta precoce de possíveis casos de violência contra mulheres que não foram expressamente reportados aos órgãos competentes. A proposta parte da **análise semântica de dados inseridos nos campos abertos de prontuários eletrônicos do SUS**, com o objetivo de **identificar padrões semânticos**, a partir da **metodologia da FrameNet**, para compreender como tais padrões de violência aparecem nos registros de saúde. Essa proposta vai na contramão da maioria das pesquisas neste campo, que focam em dados parametrizados (Malta et al., 2023; Mascarenhas et al., 2020; Pinto et al., 2021).

No Brasil, a inserção e o registro de informações sobre a população em contextos de saúde pública são realizados por meio de diferentes sistemas. Entre os principais sistemas voltados à coleta de dados sobre atendimentos em saúde e casos de violência estão o **e-SUS AB**, o **Sistema de Informações sobre Mortalidade** (SIM) e o **SINAN**.

O e-SUS AB é responsável por registrar os atendimentos realizados na atenção primária à saúde, a partir de informações do **prontuário eletrônico do paciente**. Já o SIM

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

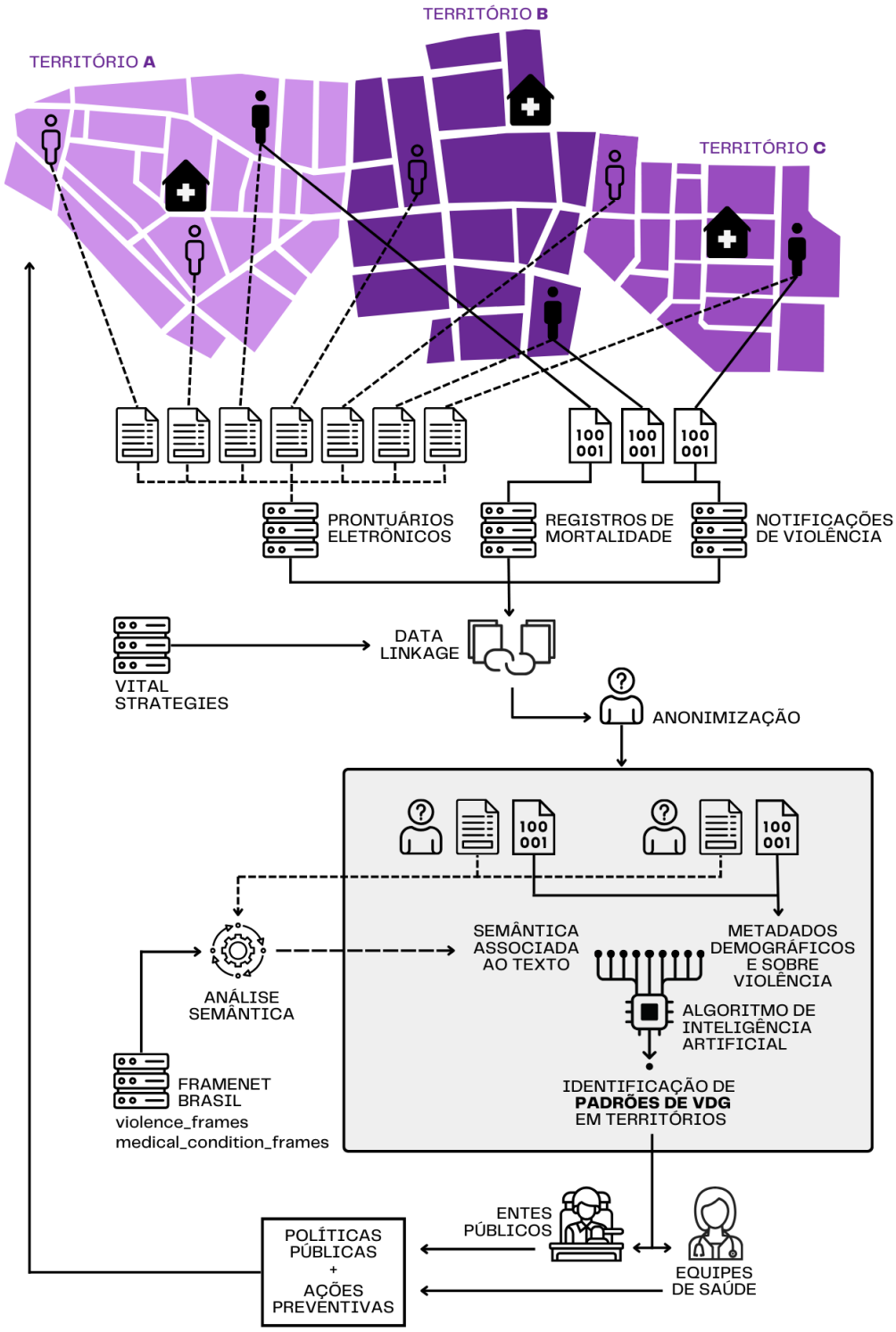
reúne **dados sobre óbitos** tendo como principal fonte a Declaração de Óbito. Grande parte das informações inseridas nesses sistemas segue um formato padronizado: os profissionais devem selecionar códigos a partir de listas pré-estabelecidas, como a **Classificação Internacional de Doenças** (CID-10), o que acaba por limitar o nível de detalhamento dos casos registrados. No entanto, tanto o e-SUS quanto o SINAN oferecem **campos de texto aberto**, que permitem uma descrição mais contextualizada das situações atendidas. Apesar de trabalhar com um sistema diferente, o Capítulo 9, por exemplo, também aborda o trabalho com campos de textos abertos que permitem uma descrição narrativa dos atendimentos aos pacientes. Esses são os campos relevantes para o desenvolvimento deste projeto.

Para atingir esse objetivo, o projeto firmou uma parceria com a **prefeitura de Recife**, que garantiu acesso aos registros do e-SUS, SINAN e SIM para estudo. Com isso, foi necessário o desenvolvimento de uma metodologia capaz de conectar tais sistemas, garantir a segurança dos dados usados e aplicar a metodologia da FrameNet para identificar padrões que pudessem ser usados para criar políticas públicas (veja Figura 17.5).

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Figura 17.5: *Pipeline* do projeto DataToStopGBV



Fonte: (Dutra, 2024, p. 5.)

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Desta forma, uma vez que os dados foram disponibilizados, a primeira etapa do processo consistiu no *linkage* dos bancos de dados da saúde pública de Recife – e-SUS, SINAN e SIM – o que permitiu relacionar os registros de uma mesma pessoa em diferentes unidades, além de mapear possíveis padrões de uso do sistema por vítimas de violência. Tais registros, no entanto, contêm dados extremamente sensíveis que são, por lei, protegidos. Portanto, existe uma preocupação ética relacionada ao acesso e armazenamento de tais tipos de dados, temas estes que também são discutidos nos Capítulos 9 e Questões éticas em IA e PLN. Dessa forma, para que o acesso aos dados pudesse ser autorizado, todos os conjuntos utilizados foram previamente anonimizados.

A próxima etapa da metodologia descreve o foco da contribuição da FrameNet Brasil para o projeto que envolve o desenvolvimento do modelo semântico baseado na Semântica de *Frames* e FrameNet. Esta etapa inclui a **modelagem de domínios específicos da Saúde e da Violência** e a **anotação semântica manual e automática** dos dados que permitiu a análise dos registros em campos abertos e identificação de possíveis padrões semânticos ligados a casos de violência baseada em gênero.

A seguir, iremos destacar o processo linguístico-computacional aplicado neste projeto, o recorte de interesse para este capítulo, que descreve a modelagem dos domínios específicos da Saúde e da Violência, o processo de anotação aplicado, assim como os experimentos que avaliaram a contribuição da metodologia para a identificação de casos e de padrões de violência.

17.4.2 A contribuição da FrameNet para identificação da Violência Baseada em Gênero

A construção de uma rede semântica especializada se caracteriza pela **modelagem de um domínio específico**. No contexto da FrameNet, tal processo constitui a criação de uma representação cognitiva estruturada, que organiza conceitos centrais de um determinado tópico. O domínio, então, corresponde a um grupo de *frames* que cobrem eventos, participantes, objetos, entre outros, que são organizados e relacionados entre si via relações *frame*-a-*frame*. O processo de modelagem de um domínio envolve diferentes etapas que passam pelo estudo de um dado tópico, as ULs e *frames* relacionados a ele, as relações entre *frames* e entre ULs – as relações qualia – e, por fim, a anotação semântica que valida a modelagem.

17.4.2.1 Modelagem dos domínios

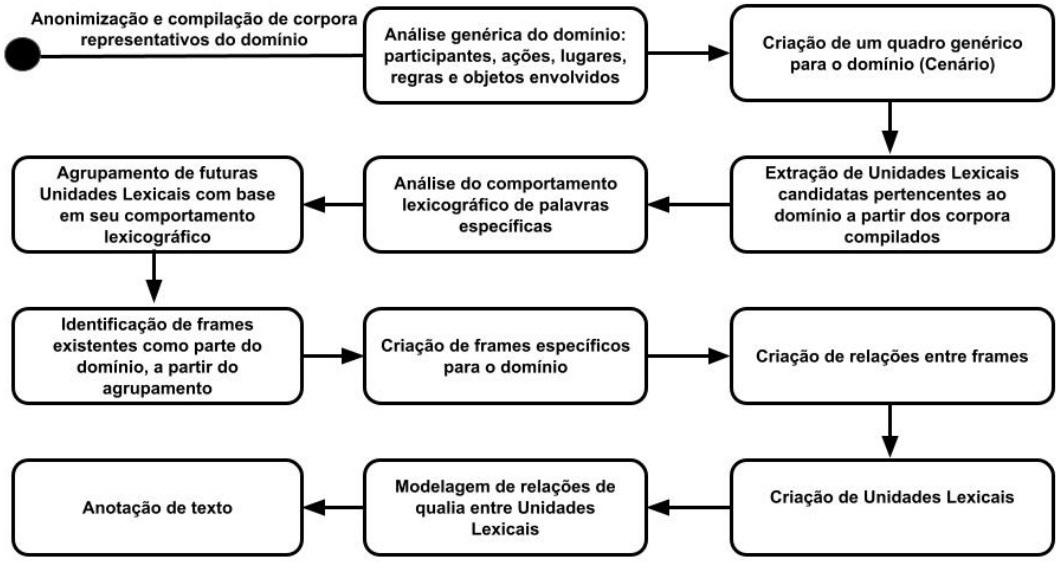
Este processo foi definido inicialmente por Costa (2020), mas adaptado e ampliado para o escopo deste projeto como é descrito por Dutra (2024). A Figura 17.6 ilustra o passo a passo de 12 etapas da modelagem de domínio específico seguida neste projeto para ambos os domínios.

Como descrito na imagem, o primeiro passo na modelagem dos domínios foi a anonimização e compilação dos *corpora* representativos do domínio: os campos abertos dos registros disponibilizados por Recife do e-SUS e SINAN. Conforme já discutido, tais registros possuem informações extremamente sensíveis que poderiam levar à identificação dos pacientes, o que por si já levanta preocupações. No entanto, quando consideramos que tais pacientes podem ser potenciais vítimas de violência e que sua identificação poderia prejudicar sua segurança, esse fator aumenta o cuidado com que os dados precisam ser trabalhados. Portanto, para garantir a proteção das informações, todos os membros envolvidos assinaram acordos de confidencialidade. Além disso, os dados passaram por

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Figura 17.6: Passo a passo da metodologia para modelagem computacional de um domínio específico



Fonte: (Dutra, 2024, p. 25).

um processo de anonimização em três etapas, e seu acesso foi restrito até mesmo entre os participantes, com apenas amostras anonimizadas sendo repassadas para os pesquisadores da FrameNet Brasil. Uma vez que as amostras anonimizadas que compuseram os *corpora* de análise foram disponibilizadas, um estudo dos domínios foi feito para que um panorama de conceitos relacionados a eles pudesse ser desenhado. Isso implicou a compreensão de quais são os possíveis participantes, ações, lugares, objetos que poderiam estar presentes no domínio. A partir deste estudo, foi modelado o primeiro *frame* de um domínio: o de cenário. No caso dos domínios da Saúde e Violência, *frames* existentes foram ajustados para ocupar essa posição. O *frame* `Cenário_de_interações_médicas` foi reestruturado para ser mais amplo e se tornou `Cenário_da_saúde`. Enquanto o *frame* `Violência` foi ajustado para `Cenário_da_Violência`.

Os próximos passos estão relacionados com lemas que são candidatos a ULs. Primeiramente, a ferramenta Keywords do *software* Sketch Engine (Kilgarriff; Al., 2014) foi usada para extrair termos mais relevantes ao domínio. Isso foi feito a partir de uma comparação dos *corpora* específicos com um *corpus* genérico, o Portuguese Web 2020. De acordo com a frequência de ocorrência de um termo, ele era considerado tendo maior ou menor probabilidade de estar relacionado ao domínio. Dessa forma, palavras como "dor de cabeça.n" ou "abuso.n" se destacavam como tendo maior probabilidade de pertencer aos domínios em comparação a palavras como "dia.n" ou "almoço.n", por exemplo. A lista inicial de tais termos foi composta pelas 1.000 palavras e 1.000 multiwords[13] (Capítulo Expressões multipalavras) com maior chance de pertencer ao domínio.

Uma vez que esta etapa foi concluída, os termos da lista gerada foram analisados quanto

[13]Expressão composta por duas ou mais palavras que juntas possuem sentido próprio, como por exemplo, "pé de moleque". Expressão que denomina um doce e não o pé de um menino.

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

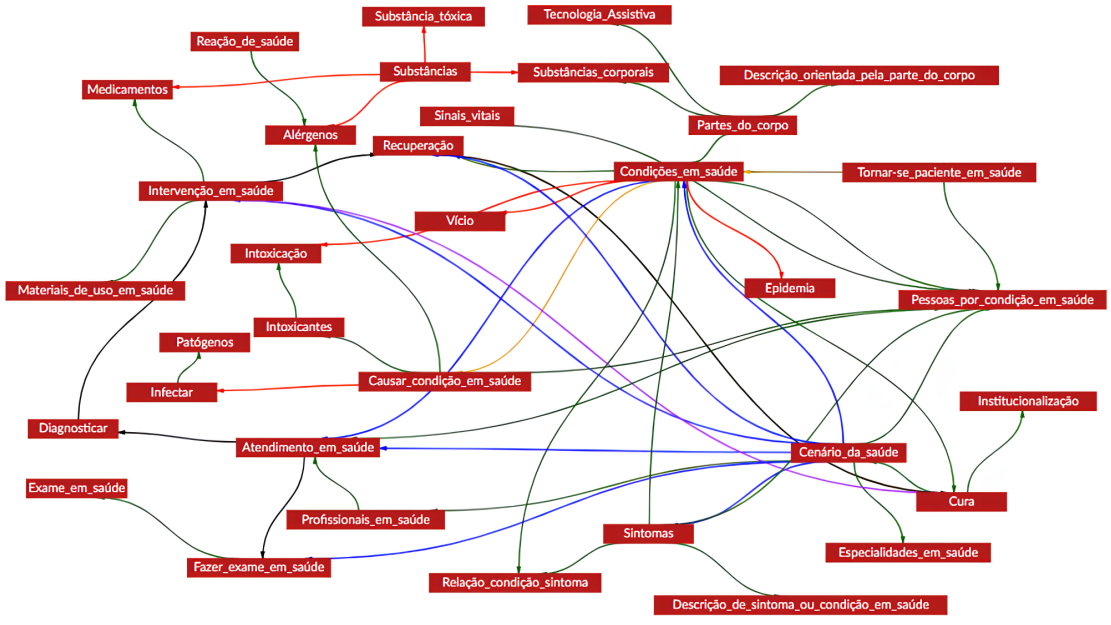
https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

ao seu comportamento lexicográfico e o agrupamento das futuras ULs foi feito com o objetivo de identificar ULs pertencentes a *frames* existentes ou novos *frames*. A modo de ilustração, houve uma grande ocorrência de ULs relacionadas ao *frame* `Parentesco` – "pai.n", "mãe.n", "filho.n" – no entanto, tal *frame* não faz parte diretamente de nenhum dos domínios em foco. Já ULs que nomeiam doenças – "gripe.n", "enxaqueca.n", "infecção.n" puderam ser relacionadas diretamente ao `Condições_em_saúde`.

Em um segundo contraponto, havia também um grande número de ULs que poderiam estar relacionadas a este *frame*, por se referirem a condições de saúde – tais como "dor de cabeça.n", "febre.n" e "coriza.n". No entanto, sua análise apontou a necessidade da modelagem de um novo *frame*, o de `Sintomas`. Esta etapa do processo permitiu uma melhor compreensão de quais *frames* existentes deveriam ser incluídos nos domínios e quais precisariam ser modelados para formar o domínio.

Com isso, os próximos passos foram a modelagem dos *frames* sugeridos e das relações entre *frames* que precisavam ser estabelecidas dentro da rede semântica da FrameNet Brasil. O domínio da Saúde possui um total de 35 *frames*, desses, 17 foram modelados a partir das demandas deste projeto. Dos que já estavam no banco, seis precisaram ser modificados. Já o domínio da Violência possui 48 *frames*, sendo um modificado e 5 novos. Com os *frames* de cada domínio definidos, foi possível conectá-los via relações *frame*-a-*frame*. As Figuras 17.7 e 17.8 ilustram as relações, sendo que cada cor, como explicado anteriormente, representa um tipo de relação – vermelho: herança, azul: subframe, verde: uso, amarelo: causativo de, marrom: incoativo de, preto: precedência, rosa: perspectiva de e roxo: ver também. As descrições da implicatura de cada uma das relações pode ser revista na Seção 17.3.1.

Figura 17.7: Representação do Domínio da Saúde



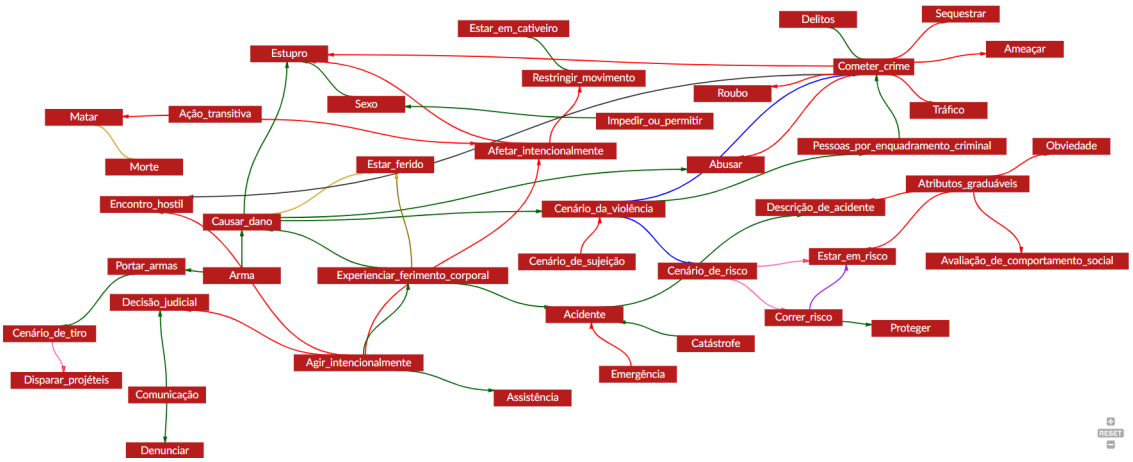
Fonte: Webtool 4.1

Os próximos passos estão ligados às ULs. Após a definição das relações entre os *frames*, os lemas agrupados como possíveis ULs foram associados aos *frames* na base de dados e às relações qualia foram definidas. Para que isso pudesse ser feito, no entanto, os lemas e

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Figura 17.8: Representação do Domínio da Violência



Fonte: Webtool 4.1

wordforms – possíveis variações de cada lema tais como flexão e conjugação - precisavam existir na Webtool para que as estruturas fossem reconhecidas na próxima fase: a da anotação.

Para este projeto, no entanto, além das *wordforms* tradicionais foram acrescentadas grafias alternativas e abreviações. Isso foi feito devido ao caráter dos *corpora* que possuem jargões específicos do âmbito da saúde e foram "construídos" durante consultas em unidades de saúde brasileiras que geralmente estão sob alta demanda, o que pode levar a erros ortográficos e de digitação. Ignorar tais ocorrências levaria à perda significativa de dados que não seriam reconhecidos.

Com todos os lemas e *wordforms* adicionados, as ULs foram criadas e as relações qualia definidas. Como introduzido na Seção 17.3.2, as relações qualia relacionam ULs o que permite a expansão de suas informações semânticas. Existem quatro tipos de relações qualia: agentiva, formal, constitutiva e télica que na FrameNet Brasil são intermediadas por *frames* como ilustrado no Quadro 17.1. Até este momento, foram criadas 4.698 relações entre ULs pertencentes aos domínios em foco.

DESAFIO: *Frame* de cenário para domínios específicos

Com base na descrição acerca do processo de modelagem de um domínio, propomos o seguinte desafio: **a proposta de um domínio específico e a modelagem de um frame de cenário para o domínio proposto.** No momento, a base de dados da FrameNet Brasil possui por volta de 1.200 frames e cinco domínios específicos: turismo, esporte, saúde, violência e educação.

Para a modelagem do frame de cenário, o domínio precisa ser estudado para que seja possível definir o frame e os seus elementos de frame – nucleares e não-nucleares. Na Webtool 4.1 é possível acessar outros frames de cenário, para melhor entendimento se necessário. A modelagem do frame deve ser feita pela plataforma Lutma. A plataforma é intuitiva e nela também estão disponíveis vídeos tutoriais.

Dúvidas podem ser enviadas para o email dos autores deste capítulo.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

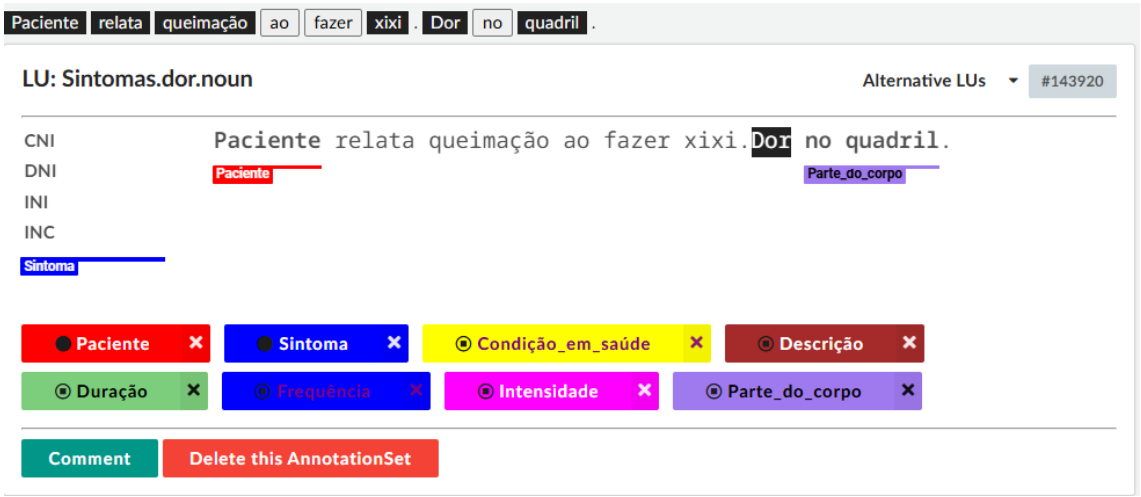
17.4.2.2 Anotação Computacional

A **anotação computacional** neste projeto foi dividida em duas etapas. A primeira delas foi a anotação manual feita na plataforma Webtool. Dada a natureza dos dados, uma versão da plataforma foi criada com acesso restrito aos integrantes do projeto. O primeiro objetivo da anotação manual foi validar a modelagem do domínio. A anotação computacional é o último passo do processo e permite uma análise geral da efetividade do domínio, tais como a suficiência dos *frames* – em cobertura e em estrutura.

A metodologia de anotação utilizada foi baseada na definida por Ruppenhofer et al. (2016) descrita na Seção 17.3. O objetivo do projeto e as características dos *corpora* levaram a pequenos ajustes, tais como (1) o foco apenas na camada semântica; (2) a ausência da anotação de Instanciações Nulas (NI); (3) liberdades quanto à localidade sintática; e (4) a anotação por campo e não por sentença.

Na Figura 17.9, observa-se a anotação da UL "*dor.n*" no registro "*Paciente relata queimação ao fazer xixi. Dor no quadril.*"[14] que exemplifica as alterações nas diretrizes de anotação. Neste exemplo, nota-se que a anotação contém apenas informações semânticas. Outro ajuste feito diz respeito à presença de dois períodos no campo de anotação, quando normalmente"*Paciente relata queimação ao fazer xixi*" e "*Dor no quadril*" estariam em campos de anotação distintos. Observa-se também que a marcação do elemento de *frame* `PACIENTE` que poderia ter sido anotado como uma Instanciação Nula Definida - pelo contexto do prontuário - ou uma Instanciação Nula Indefinida - por não se saber quem é este paciente, foi marcada na sentença no período anterior. No entanto, como foi optado por não anotar Instanciações Nulas, a anotação do elemento de frame foi feita fora da localidade sintática - neste caso, em outro período.

Figura 17.9: Exemplo de anotação



Fonte: Adaptada de (Dutra, 2024)

Ao todo foram anotadas 2.352 sentenças, sendo que foram criados 6.313 *Annotation Sets* no domínio da Saúde e 8.309 no domínio da Violência. Este processo foi realizado por sete anotadores com diferentes níveis de experiência. Reuniões semanais também foram

[14] Devido ao conteúdo sensível dos dados utilizados nesta pesquisa, sentenças reais não puderam ser utilizadas. Desta forma, o exemplo usado foi criado para ilustrar o tipo de ocorrência presente no *corpus*.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

realizadas para discutir o processo de anotação e esclarecer possíveis dúvidas.

A segunda etapa de anotação consistiu na anotação automática dos dados. A partir da amostra obtida pela anotação manual, uma versão do **rotulador automático LOME** (Xia et al., 2021) foi treinada do zero – com cerca de 200 *frames* a mais do que os usados no artigo citado – e usada para a anotação em larga escala dos *corpora*. O LOME é um modelo desenvolvido para extração de informações multilíngues com um *pipeline* que inclui um ***parser* semântico da FrameNet**. O modelo foi usado para anotar os registros do e-SUS AB, SINAN e CID. Ele é capaz de identificar as ULs alvos, rotular o *frame* e seus elementos. A avaliação da anotação utilizou as mesmas métricas propostas por Xia et al. (2021): *exact match* (em) e *span match* (sm), todas considerando *precision*, *recall* e *F1*. A métrica **em** exige correspondência exata de todo o trecho da anotação. Já a **sm** avalia apenas a detecção correta dos limites do trecho, sem considerar rótulos. Os resultados foram similares aos alcançados por Xia et al. (2021), o que aponta uma boa performance do modelo, uma vez que foram usadas informações de domínios mais específicos, *frames* novos e mais de 5.000 ULs que haviam sido anotadas pela primeira vez.

17.4.3 Avaliação da identificação da Violência Baseada em Gênero

Para avaliar os benefícios do uso da Framenet para a identificação de casos de VBG, três experimentos foram feitos. Os experimentos contrastavam o uso da metodologia da FrameNet classificando casos como sendo ou não casos de VBG em uma configuração que se apoiava nos dados dos campos de texto abertos e fechados.

Os registros que foram utilizados para a classificação se enquadram em quatro grupos: (1) violência – casos com código CID para violência ou notificação SINAN; (2) não violência – sem evidências de violência, muitas vezes associados a doenças; (3) provável violência – casos com notificação de violência sem código CID para agressão 30 dias antes da mesma; e (4) desconhecidos – não se relaciona a nenhuma das categorias. Foram considerados 801 casos, sendo 167 deles classificados como violência.

A partir dos experimentos, foram feitas as avaliações quantitativas e qualitativas. A avaliação quantitativa focou na comparação das performances dos experimentos. Já a avaliação qualitativa observou os atributos considerados mais relevantes para o modelo e foi complementada com a análise dos *frames* e unidades lexicais com maior ocorrência na anotação para ambos os domínios.

A avaliação do processo de modelagem e anotação manual dos domínios também foi avaliada no âmbito de Dutra (2024). No entanto, para este recorte, o foco será nos resultados diretamente ligados à avaliação da contribuição da FrameNet para a identificação de casos de VBG. Os experimentos e os processos avaliativos aplicados serão discutidos a seguir.

17.4.3.1 Os experimentos

Os três experimentos usaram o modelo **Support Vector Machine** (SVM) (Cortes; Vapnik, 1995), um modelo frequentemente usado em tarefas de classificação. A escolha foi motivada devido a sua capacidade de lidar com dados com muitas variáveis e flexível para separar padrões tanto em casos lineares quanto não lineares, ou seja, em que não há uma divisão simples entre as classes. Uma das vantagens do SVM é a possibilidade de ranquear a importância dos atributos na escolha da classificação, o que foi feito com o objetivo de apoiar a análise qualitativa e reforçar a relevância da informação semântica da FrameNet.

O primeiro e o segundo experimentos utilizaram o mesmo pipeline de seis etapas: (1) a anotação com LOME; (2) a identificação das LUs, (3) a criação dos atributos; (4) a

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

filtragem dos vetores com base em sua ocorrência; (5) a geração da representação TF-IDF[15] dos vetores; e (6) a aplicação da análise do componente principal (PCA)[16] às representações TF-IDF. A diferença entre os dois experimentos está nos dados que foram usados. O primeiro experimento considerou apenas os campos abertos de texto anotados, enquanto o segundo, além dos campos abertos anotados, considerou também um grupo de parâmetros do campo fechado anotado para *frames*: orientação sexual e gênero, necessidade de prótese, raça e idade.

Por fim, o terceiro experimento consistiu apenas na criação de atributos a partir das informações dos campos parametrizados, ou seja, os dados considerados não passaram pelo processo de anotação. Consideramos vinte dos campos disponíveis, dentre eles características do paciente, bloco e número CID, localidade e variáveis temporais acerca dos atendimentos.

17.4.3.2 Avaliação Quantitativa

A avaliação quantitativa dos modelos foi feita por meio de validação cruzada. Esta é uma técnica de avaliação que permite o treinamento e teste do modelo com diferentes combinações dos conjuntos de dados utilizados (Kohavi, 1995). As métricas consideradas foram *F1 score*, *recall* e *precision*. O resultado final é uma média calculada a partir das diferentes combinações feitas. Para este projeto, o conjunto de dados foi dividido em cinco conjuntos, portanto foram feitas cinco interações e a média foi calculada a partir do resultado delas. O resultado pode ser visto na Tabela 17.2, que apresenta o desempenho final da validação cruzada, seguido da variação, que indica a dispersão dos resultados entre as diferentes rodadas do processo.

Tabela 17.2: Resultados das performances dos experimentos

Experimento	***F1***	***Recall***	***Precision***
1	0.772 (0.113)	0.838 (0.071)	0.756 (0.190)
2	0.771 (0.114)	0.832 (0.071)	0.759 (0.189)
3	0.461 (0.089)	0.701 (0.173)	0.345 (0.057)

Fonte: (Dutra, 2024)

Os resultados obtidos apontam que, quando a anotação semântica seguindo as diretrizes da FrameNet Brasil foi considerada, a performance do modelo obteve um desempenho melhor na identificação de casos de VBG. O primeiro experimento apresentou o melhor desempenho, com alta *recall* e equilíbrio entre *precision* e *F1*, o que é ideal no contexto da pesquisa, em que falsos positivos são preferíveis a falsos negativos. O segundo modelo, que inclui dados parametrizados, teve desempenho semelhante, que confirma a vantagem do uso da metodologia. Já o terceiro experimento teve resultados inferiores, com alta taxa de falsos positivos e grande variação nos testes. Esses achados reforçam a hipótese de que informações semânticas extraídas de texto livre acrescida de informações do modelo semântico da FrameNet são mais eficazes na identificação de casos de VBG.

[15]Técnica que calcula a importância de uma palavra em um texto, levando em conta sua frequência no documento e quão rara ela é em todo o conjunto de textos.

[16]Método que reduz o número de variáveis nos dados, transformando-as em componentes principais e mantendo a maior parte da informação original.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

17.4.3.3 Avaliação Qualitativa

Para a avaliação qualitativa, foram analisados o primeiro e o terceiro experimento. O foco foi nos atributos que mais influenciaram a classificação dos casos como violência ou não violência. No entanto, é importante ressaltar que neste momento não foi possível determinar para qual classe cada atributo foi mais relevante, apenas que ela teve peso na decisão de classificação do modelo. De modo complementar à análise semântica do primeiro experimento, uma lista dos *frames* e ULs mais evocados dos domínios da Saúde e da Violência de acordo com a anotação automática do LOME também foi analisada em busca de mais padrões. Essa lista considerou apenas a anotação de casos confirmados de VBG.

A análise dos atributos mais relevantes no terceiro experimento revelou uma concentração em dados de localização das unidades de saúde, indicando um viés, já que esse fator não tem correlação direta com a ocorrência de violência; pelo contrário, alimenta estereótipos. O atributo *raça* também teve destaque, o que pode refletir um viés decorrente da inconsistência no preenchimento do campo, muitas vezes deixado em branco. Esse resultado sugere que o modelo tende a considerar mais atributos pessoais, e não o contexto do atendimento.

Por outro lado, alguns códigos CID relacionados ao motivo da busca por atendimento também foram identificados, com destaque para códigos que podem estar associados a casos de violência sexual ou situações que exigem acompanhamento, como o pré-natal. Apesar disso, tais variáveis foram menos frequentes. Isso reforça novamente que o uso dos campos parametrizados não traz informações que podem ser cruciais sobre o atendimento, e também tendem a reforçar estereótipos ao focar no indivíduo e não na causa para a busca pelo serviço de saúde. Esses dados podem ser vistos na Figura 17.10.

Já a análise do primeiro experimento, apesar de preliminar, apresentou melhor resultado quanto à identificação de possíveis padrões de violência. Os atributos de maior relevância estão listados na Figura 17.11. Dentre eles, destacam-se *frames* genéricos como `Relações_pessoais` e `Medo`, que, embora fora do domínio de violência, podem indicar relações com o agressor ou o cenário emocional envolvido. Entre os *frames* do domínio da violência, apenas `Experienciar_ferimento_corporal` foi identificado, o que pode estar relacionado à baixa frequência desses *frames* nos registros do e-SUS. Ainda assim, sua presença sugere que sinais de violência podem surgir indiretamente por meio de dados de saúde. No domínio da saúde, destacaram-se `Condições_em_saúde`, `Medicamentos` (com múltiplos elementos de *frame* relevantes) e `Intervenção_em_saúde`, com ULs como *prescrever.v* e *administrar.v*. Tais resultados reiteram que os registros médicos podem fornecer pistas importantes para a identificação precoce da violência baseada em gênero.

A análise da lista extraída da anotação do LOME dos casos de violência trouxe mais possíveis padrões que merecem atenção. No domínio da saúde, dois grupos de ULs de `Condições_em_saúde` se destacaram. O primeiro relacionado à gravidez. Este fato, no momento, leva a duas hipóteses: é possível que este resultado se dê devido a um enviesamento do modelo, uma vez que o grupo de estudo é composto por mulheres. Uma alternativa, no entanto, é de que a gravidez aumenta a vulnerabilidade da mulher, fazendo com que ela seja mais propensa a permanecer em um relacionamento abusivo. O segundo grupo que chama atenção é o de ULs relacionadas a transtornos mentais. Novamente, existe a possibilidade de enviesamento do modelo, já que mulheres podem buscar ajuda profissional com maior frequência nestes casos do que homens. No entanto, tal estado pode também ser uma consequência da violência experienciada.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Figura 17.10: Resultado dos atributos mais relevantes no terceiro experimento de usando apenas dados parametrizados



Fonte: (Dutra, 2024)

No domínio da Violência, destacam-se também dois grupos de ULs. O primeiro relacionado à automutilação e ao suicídio. Tais dados foram encontrados com alta frequência e demandam um estudo isolado, mas já exibem forte indício de serem uma consequência da violência. O segundo grupo está relacionado a abuso sexual, resultado esse que foi reforçado nos dados do domínio da Saúde devido à alta incidência de exames e doenças ligadas a doenças sexualmente transmissíveis. Esses resultados, tanto da avaliação semântica do modelo, quanto dos resultados obtidos a partir da anotação LOME, reforçam mais uma vez o valor que a análise de tais dados via FrameNet pode trazer para a compreensão de padrões de violência nos registros de saúde e identificação de casos de VBG.

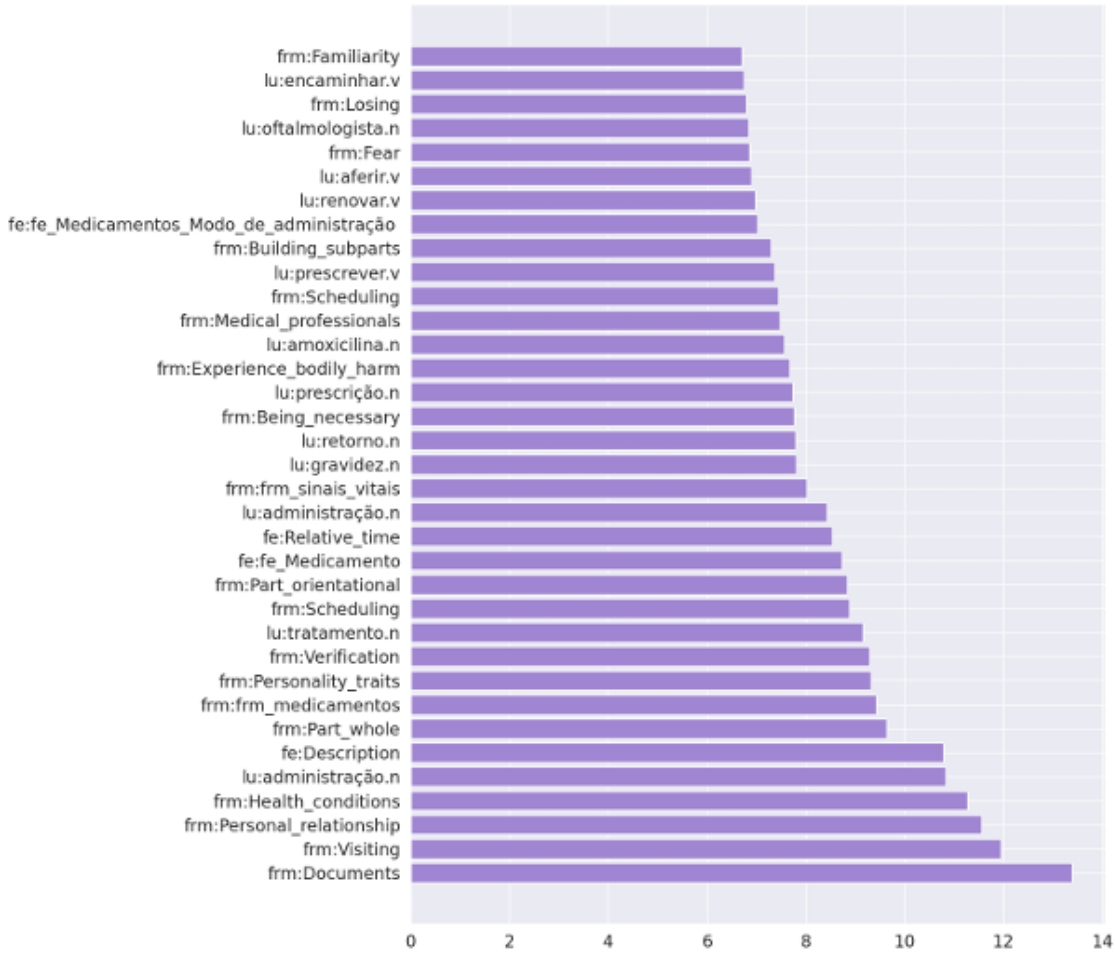
17.4.4 Impacto Social e técnico

A partir da descrição da metodologia aplicada a este projeto e dos resultados discutidos, são visíveis os benefícios do uso da metodologia da FrameNet para a interpretação semântica de registros nos sistemas de saúde e a identificação de padrões de violência. Considerando

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Figura 17.11: Resultado dos atributos mais relevantes no primeiro experimento usando a anotação semântica do campo aberto



Fonte: (Dutra, 2024)

que a maior parte das pesquisas que buscam a prevenção de casos de Violência Baseada em Gênero focam em dados estatísticos, o projeto DataToStopGBV é o pioneiro em avançar este campo de pesquisa para a análise dos campos abertos com resultados que comprovam a eficácia do uso da FrameNet no estudo de casos notificados e, também, na identificação de casos subnotificados.

No entanto, como pode ser observado nos resultados, para além de padrões relacionados à violência física e psicológica, foram identificados também padrões relacionados à saúde da mulher vítima de violência. Tais dados levam para dois possíveis desdobramentos futuros: a busca da mulher por atendimentos relacionados a sua saúde que mascaram a violência sofrida e as condições de saúde que se sobressaem em mulheres que são vítimas de violência. O estudo e a análise destes dados em conjunto com profissionais da saúde podem levar a desdobramentos com impactos significativos para a prevenção da VBG.

Para além deste âmbito, dada a eficácia da metodologia aplicada, existe também a possibilidade da ampliação deste projeto para outras esferas que extrapolem a VBG. Esse

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

movimento pode partir da expansão do grupo foco, fazendo com que não apenas os casos de VBG contra mulheres sejam analisados, e pode alcançar outros tipos de violência tais como a violência infantil.

Na perspectiva técnica, este projeto também mostra a capacidade de adaptação da metodologia da FrameNet, não apenas para a criação de domínios semânticos específicos – trabalhos já desenvolvidos anteriormente (Costa, 2020; Gamonal, 2013) –, mas também para objetivos específicos. Observou-se, neste projeto, a adaptação da base de dados em duas esferas: a criação de uma versão espelhada com acesso restrito e a adesão de lemas e *wordforms* que incluíam abreviações e grafias alternativas. Para além, houve a adaptação das diretrizes de anotação manual, que foram ajustadas e ainda assim alcançaram performance significativa quando reproduzidas automaticamente.

A metodologia deste projeto é adaptável não apenas com relação aos métodos da FrameNet, mas ao modo de expansão territorial. Portanto, após possíveis ajustes tanto no âmbito do léxico regional ou de uma possível diferença no sistema público de saúde usado, o trabalho desenvolvido com os dados da cidade de Recife pode ser ampliado para outras cidades e/ ou estados brasileiros.

Dessa forma, o trabalho conjunto entre essa iniciativa, profissionais da saúde e órgãos governamentais para a elaboração de políticas públicas que possam ser implementadas é imperativo. Com a expansão desta pesquisa e possível aplicação em maior escala, esperamos o fortalecimento das políticas públicas para intervenção preventiva de casos de Violência Baseada em Gênero e o decrescimento de casos de subnotificação.

17.5 Multimodalidade, acessibilidade cultural e sentido compartilhado no Audition: *dataset* multimodal de curtas-metragens acessíveis

A **comunicação humana** em contextos reais de uso é fundamentalmente **multimodal**. A noção de multimodalidade, tal como entendida neste capítulo, refere-se à participação simultânea e coordenada de diferentes **modos comunicativos** (verbal, visual, sonoro, corporal entre outros) na construção de sentido. Assumimos uma concepção fenomenológica do termo, como discutido por Torrent; Coneglian (2025), por estarmos interessados nos meios (multi)semióticos pelos quais os indivíduos constroem significado na experiência situada da linguagem.

Essa concepção ultrapassa a ideia de uma simples junção de semioses, pois envolve uma estrutura compartilhada de significados, na qual os elementos provenientes dos diferentes modos se articulam de forma integrada. Em outras palavras, a multimodalidade, tal como concebida em nossas pesquisas, diz respeito à articulação entre modos em torno de um sentido coerente, ativado e inferido no uso situado da linguagem.

É a partir dessa perspectiva que passamos a relacionar a Semântica de *Frames* e a FrameNet Brasil a um contexto multimodal. Partimos do pressuposto de que, assim como o texto verbal evoca *frames*, imagens, sons e gestos também o fazem. Essa premissa abriu caminho para a criação da ReINVenTA (*Research and Innovation Network for Vision and Text Analysis*), uma rede de pesquisa dedicada ao processamento semântico computacional de objetos multimodais.

A atuação da ReINVenTA concentra-se em quatro eixos principais: (i) a ampliação da cobertura da FrameNet para o português brasileiro, incorporando dados multimodais; (ii) a constituição de um *gold standard dataset* de objetos multimodais anotados semanticamente

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

e validados em perspectiva psicolinguística; (iii) o desenvolvimento de algoritmos e modelos para rotulação automática e descoberta de conhecimento em objetos multimodais; e (iv) a proposição de melhores práticas para tradução audiovisual acessível, com ênfase na audiodescrição fílmica.

Esses objetivos se materializam em três *datasets* distintos que compõem, até o momento, o conjunto de dados da rede.

O primeiro é o Frame² (Belcavello et al., 2024), composto por episódios do programa *Pedro pelo Mundo*. Esse *dataset* foi anotado para *frames*, elementos de *frames* e categorias de objetos reconhecíveis por algoritmos de visão computacional para as modalidades de vídeo, áudio original e legendas.

O segundo é o Framed Multi30k (Viridiano et al., 2024), que expande o *dataset* Multi30k com descrições em português e anotações semânticas visuais. Neste *dataset*, foram produzidas cinco descrições em português e uma descrição traduzida do inglês para cada uma das cerca de 30 mil imagens do *dataset* Flickr 30k. Além disso, o *dataset* inclui anotações automáticas de *frames* para todas as legendas em inglês e português, bem como anotação manual de *frames*, elementos de *frame* e *bounding boxes* aplicadas às imagens do *dataset* Flickr 30k Entities.

O terceiro é o Audition, *dataset* de curtas-metragens brasileiros acessíveis, que será apresentado a seguir.

17.5.1 O *dataset* Audition

O *Audition* é um *dataset* multimodal em desenvolvimento, composto por curtas-metragens brasileiros anotados semanticamente com base na teoria da Semântica de *Frames* e no modelo da FrameNet Brasil. Sua criação foi motivada pela necessidade de recursos que integrem os modos sonoro (verbal e não verbal) e visual em português brasileiro com ênfase na promoção do acesso a conteúdos audiovisuais, especialmente para pessoas com deficiência visual.

Embora os modelos atuais de PLN venham avançando significativamente no tratamento de textos escritos, ainda há pouca atenção às necessidades de usuários que dependem de traduções intersemióticas – como a audiodescrição – para ter acesso a obras audiovisuais, por exemplo.

O Capítulo Conjunto de dados, *dataset* e *corpus* oferece um excelente panorama sobre a preparação e a avaliação de dados linguísticos para PLN, discutindo, entre outros pontos, a distinção entre *corpus* e *dataset*. Quem já leu o capítulo deve ter alguma opinião formada sobre o conjunto de dados multimodais que configura o *Audition*. A partir dessa discussão, propomos uma reflexão: o *Audition*, enquanto conjunto de dados multimodais, deve ser classificado como *dataset* ou também pode ser considerado um *corpus*?

De acordo com o Capítulo Conjunto de dados, *dataset* e *corpus*, um *dataset* é um conjunto de dados linguísticos anotados por pessoas ou feito por máquinas e revisto por pessoas. Já um *corpus* linguístico pode ter sido anotado automaticamente. A depender dos objetivos da pesquisa, receber um material já categorizado previamente pode facilitar a tarefa de estudo desejada. Essa questão diz respeito à utilidade do material em discussão.

Outro critério importante é a origem dos dados. Para configurar-se como *corpus*, o conjunto de textos deve ser composto por material linguístico naturalmente produzido, ou seja, não elaborado especificamente para fins experimentais ou científicos. Em contrapartida, *datasets* podem ser criados com a finalidade explícita de alimentar modelos, validar hipóteses ou compor bancos de dados anotados, como é o caso do *dataset* Multi30k

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

(Viridiano et al., 2024).

Em se tratando de dados multimodais, como os que compõem o *Audition*, as distinções conceituais entre *corpus* e *dataset* devem ser compreendidas de maneira ampliada. Isso porque os dados anotados não se restringem à linguagem verbal, mas envolvem também o modo visual e sonoro, organizado em alinhamento temporal com base nas unidades significativas do vídeo.

Nesse cenário, a noção de dado linguístico é estendida para abranger modos semióticos diversos que participam conjuntamente da construção de sentido. O alinhamento semântico entre esses modos – textual, visual e sonoro – permite explorar relações intermodais e avaliar traduções intersemióticas na construção de sentido da obra fílmica.

Tendo em vista esse enquadramento conceitual, o *Audition* pode ser caracterizado como um *dataset* multimodal padrão-ouro, cuja cobertura inclui anotação manual de transcrições derivadas de textos verbais (orais e escritos) e de conteúdo visual dinâmico (vídeo). Além do áudio original, ele contempla diversos modos de tradução intersemiótica, tais como audiodescrição, legendas, *closed captions* e texto sobreposto à imagem (*on-screen text*), todos anotados semanticamente por meio de *frames*.

O Quadro 17.3 apresenta a relação entre as modalidades contempladas, os tipos de dados e as anotações semânticas realizadas em cada caso.

Quadro 17.3: Cobertura Multimodal do *Dataset* Audition

Modalidade	Tipo de dado	Anotação semântica
Áudio original (OA)	Texto verbal (Fala, narração)	*Frames*, LUs, FEs
Conteúdo de Vídeo (VC)	Imagens em movimento (vídeo)	*Frames*, FEs, *Visual Entity* (via *bounding boxes*)
Audiodescrição (AD)	Texto verbal (roteiro escrito)	*Frames*, LUs, FEs
Closed Captions (CC)	Texto verbal escrito + som não verbal	*Frames*, LUs, FEs
Legendas	Texto verbal (tradução escrita das falas)	*Frames*, LUs, FEs
On-screen Text	Texto verbal (de origem gráfica sobreposto à imagem)	*Frames*, LUs, FEs

Todos esses modos, que atuam como recursos importantíssimos de acessibilidade cultural, estão alinhados semântica e temporalmente, permitindo a observação da sincronicidade e da evocação de *frames* na construção de sentido. Esse alinhamento, detalhado na Seção 17.3, oferece subsídios para experimentos de similaridade semântica entre modos, contribuindo para avaliação da qualidade de traduções intersemióticas. Assim, o *Audition* é tanto um recurso para o desenvolvimento de sistemas de tradução audiovisual acessível, como também uma ferramenta para o estudo da construção e compartilhamento de sentido entre diferentes modalidades comunicativas.

Para garantir um bom *dataset* linguístico, o Capítulo Conjunto de dados, *dataset* e *corpus* destaca cinco características centrais. A seguir, cada uma dessas dimensões será discutida a partir das especificidades do *Audition*.

(i) consistência.

"Por consistência, entende-se que fenômenos semelhantes devem ser anotados,

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

> *isto é, analisados da mesma maneira. A consistência permite que algoritmos de aprendizado de máquina generalizem a partir dos dados e que as avaliações sejam confiáveis".* (Seção Características de um bom *dataset* linguístico)

A anotação semântica no *Audition* foi realizada com base na metodologia de anotação *full text* da FrameNet e nas *guidelines* desenvolvidas pela FrameNet Brasil para anotação semântica multimodal. Essas diretrizes asseguram a rotulação de *frames*, unidades lexicais (LUs), entidades visuais (*visual entities*) e elementos de *frame* (FEs) a todas as modalidades comunicativas do *dataset*.

A equipe responsável pela anotação é composta por estudantes de graduação com bolsas de Iniciação Científica, previamente treinados em anotação semântica baseada em *frames*. Todo o trabalho é supervisionado por anotadores experientes vinculados à FrameNet Brasil.

Adotamos uma abordagem de anotação como definida pelo manifesto de dados perspectivista[17] e por Basile et al. (2021). Desse modo, reconhecemos que a seleção de frames pode variar legitimamente entre anotadores em função de fatores como conhecimento de mundo, contexto situacional e a própria perspectiva interpretativa assumida.

Para diminuir inconsistências e assegurar qualidade da anotação, o processo inclui: **tutoriais iniciais de formação**, **encontros semanais** com espaço para discussão coletiva de decisões, **resolução de dúvidas** e **refinamento colaborativo das anotações**. Esse modelo de curadoria ativa e engajamento interpretativo é o que proporciona a consistência no *Audition*.

(ii) variedade:

> *"Um bom dataset deve ser variado (ou balanceado) com relação aos fenômenos para os quais foi construído"* (Seção Características de um bom *dataset* linguístico)

O *Audition* contempla filmes curtos de diferentes gêneros – ficção, animação, documentário e performance – com diversidade estilística e discursiva. Há ainda **variação** entre os modos comunicativos incluídos, o que amplia o escopo de análise de fenômenos intersemióticos. A diversidade de contextos e gêneros assegura que o *dataset* seja útil para tarefas no PLN e na tradução audiovisual acessível. O quantitativo por gênero cinematográfico e modos comunicativos está exibido na Figura 17.12.

(iii) representatividade:

> *"Esperamos que os textos que compõem o corpus/dataset sejam representativos do tipo de texto que será alvo da aplicação, isto é, ao qual o modelo baseado nesse dataset será aplicado"* (Seção Características de um bom *dataset* linguístico)

Como conjunto de dados voltado ao estudo e treinamento de tarefas computacionais em tradução audiovisual acessível em português brasileiro, o *Audition* foi construído a partir de obras cinematográficas acessíveis. Embora haja vários recursos disponíveis, a opção por audiodescrição foi a prioridade e está presente em todos os curtas que compõem o *dataset*.

(iv) documentação:

> *"Um bom dataset é um dataset bem documentado, que informa a origem do conteúdo textual, as classes de anotação e as diretrizes usadas para anotar, as*

[17] https://pdai.info/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

> *características e/ou formação dos anotadores".* (Seção Características de um bom *dataset* linguístico)

A documentação do *Audition* inclui quatro etapas principais: (a) seleção e liberação das obras cinematográficas com audiodescrição; (b) formações teóricas e práticas voltadas à anotação estrutural e à anotação semântica multimodal; (c) realização da anotação estrutural, que incluiu a transcrição e o alinhamento temporal entre os modos transcritos e o conteúdo visual dinâmico; e (d) anotação semântica multimodal, envolvendo a rotulação de texto e de vídeo em termos de *frames*, elementos de *frames*, unidades lexicais e entidades visuais.

Complementarmente, a documentação inclui metadados sobre os anotadores, incluindo formação acadêmica e grau de experiência com a anotação. Esses registros são importantes para proporcionar a transparência do processo, facilitar a replicação de experimentos, além de possibilitar comparações entre modalidades comunicativas e avaliação crítica do material produzido.

(v) tamanho:

> *"o tamanho é um aspecto que precisa ser levado em conta. E por tamanho não me refiro apenas ao tamanho do corpus em tokens ou palavras, mas também à quantidade de fenômenos rotulados".* (Seção Características de um bom *dataset* linguístico)

Como bem pondera (Capítulo Conjunto de dados, *dataset* e *corpus*), o critério do tamanho de um *dataset* diz respeito tanto à quantidade total de dados quanto ao número efetivo de itens anotados. Nesse sentido, ele deve ser suficiente para permitir treinamento e avaliação confiáveis, conforme o tipo de anotação realizada. Discussões sobre o tamanho não são conclusivas, como destacam (Berber Sardinha, 2004; Biber, 1993; Bowker; Pearson, 2002) e Capítulo Conjunto de dados, *dataset* e *corpus*. Isso porque o tamanho precisa ser representativo diante do propósito assumido. Desse modo, por mais que a quantidade seja algo relevante, especialmente no contexto dos *Large Language Models* (LLMs), esse tópico não é fechado. Ele não deve ser tomado isoladamente como critério de qualidade ou suficiência, pois precisa ser pensado em articulação com a finalidade da tarefa para a qual está sendo construído.

O *Audition* é um material que está em fase de conclusão. Contém cerca de 240 minutos de material audiovisual, segmentado, até o momento, em mais de 10.093 unidades visuais anotadas semanticamente via *bounding boxes*. Reúne aproximadamente 5 mil sentenças advindas de diferentes modos de acessibilidade e atuantes em diferentes gêneros cinematográficos. Essas sentenças resultaram, até o momento, em mais de 13 mil conjuntos de anotações (*annotations sets)* e cerca de 118 mil elementos de *frames* anotados. A Figura 17.12 apresenta esses números organizados por gêneros cinematográficos[18].

Checklist do Audition: construindo e avaliando um *dataset* linguístico multimodal

Baseado nos critérios da Capítulo Conjunto de dados, *dataset* e *corpus**

Na Figura 17.13, propomos um *checklist* para acompanhamento da criação do *Audition*. Pedimos que você, leitora ou leitor, preencha os campos com base nas informações já

[18] Por decisão metodológica, optamos por classificar a animação como um gênero cinematográfico distinto da ficção, com base nas reflexões de (Gordeeff, 2023). Além disso, o Audition inclui um exemplar de performance.

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

Figura 17.12: Panorama quantitativo parcial do *dataset* Audition por gênero cinematográfico

		Ficção	Anima.	Docum.	Perfo.	Total *Parcial*
Áudio original (OA)	Sentenças	664	337	537	1	1.539
	Annotations sets	1.387	686	2.148	-	4.221
	EFs	12.451	6.112	21.269	-	39.832
Audio descrição (AD)	Sentenças	1.012	251	491	126	1.880
	Annotations sets	3.748	557	1.642	413	6.360
	EFs	31.445	4.848	15.060	3.585	54.938
Legenda	Sentenças	344	286	521	-	1.151
	Annotations sets	678	16	1.502	-	2.196
	EFs	5.869	145	13.361	-	19.375
***Closed Caption* (CC)**	*Sentenças*	59	-	73	-	132
	Annotations sets	73	-	138	-	211
	EFs	685	-	1.111	-	1.796
On-screen text	*Sentenças*	98	39	27	-	164
	Annotations sets	115	13	45	-	173
	EFs	1.150	65	405	-	1.620
Conteúdo de vídeo (VC)	*Semantic boxes*	3.415	1.576	5.067	35	10.093

apresentadas nesta seção. Após a leitura das seções seguintes, retorne a este quadro para revisar a complementar seus apontamentos. Este exercício pode ajudá-la(o) a refletir sobre os passos metodológicos necessários para a criação de um *dataset* linguístico (e multimodal) de qualidade e a identificar, de forma crítica, como eles se materializam em projetos reais, como este.

17.5.2 Anotação semântica multimodal

A anotação semântica do *Audition* fundamenta-se nos princípios da Semântica de *Frames* (Fillmore, 1982), segundo os quais o significado de uma palavra ou expressão está ancorado em um "quadro" ou "moldura" conceitual, do inglês *frame*. Na linguagem verbal, o *frame* é ativado por unidades lexicais, as ULs, que podem ser verbos, substantivos, adjetivos etc. Cada *frame* envolve um conjunto de papéis semânticos denominados elementos de *frame* (EFs), que representam participantes, objetos, ações e relações relevantes para a cena conceitual evocada.

No contexto multimodal do *Audition*, parte-se da premissa de que não apenas a linguagem verbal, mas imagens, gestos e sons não verbais também evocam *frames*. Assim, a tarefa dos anotadores consiste em identificar e registrar essas evocações em cada modo comunicativo, possibilitando a comparação e o alinhamento semântico entre eles. A anotação

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Figura 17.13: *Checklist* do Audition

Critério	Você identificou no Audition?	Comentários e observações
Problema ou tarefa	( ) Sim ( ) Não ( ) Parcialmente	
Etiquetas de anotação (tagset)	( ) Sim ( ) Não ( ) Parcialmente	
Diretrizes de anotação (guidelines)	( ) Sim ( ) Não ( ) Parcialmente	
Compilação ou criação de corpus adequado à tarefa	( ) Sim ( ) Não ( ) Parcialmente	
Maneira de codificar a anotação	( ) Sim ( ) Não ( ) Parcialmente	
Pessoas responsáveis pela anotação	( ) Sim ( ) Não ( ) Parcialmente	
Ferramenta de anotação (se for o caso)	( ) Sim ( ) Não ( ) Parcialmente	
Estratégias para otimizar o processo de anotação (se for o caso)	( ) Sim ( ) Não ( ) Parcialmente	
Maneiras de avaliar o resultado da anotação	( ) Sim ( ) Não ( ) Parcialmente	

deste *dataset* é realizada integralmente por humanos e segue uma **abordagem orientada pelo texto** (*text-oriented*), na qual o processo de anotação parte da análise dos áudios transcritos para guiar a rotulação semântica do conteúdo visual.

No caso específico do *Audition*, adotamos como diretriz metodológica que o conteúdo de vídeo fosse anotado para situações, quais sejam **eventos, estados, processos** e **relações**. Esse recorte está alinhado com o objetivo do *dataset* de oferecer uma base para a análise da construção de sentido em sequências audiovisuais, priorizando a ação, a mudança de estado e os processos perceptivos. Em contraste, em (Belcavello et al., 2024), o protocolo de anotação previu também a marcação de entidades, por meio de *frames* centrados em participantes ou objetos nas cenas de vídeo. Essa diferença metodológica reflete a flexibilidade do modelo da FrameNet Brasil frente a diferentes objetivos analíticos e tarefas de anotação.

A ferramenta utilizada para anotação é a **Webtool 4.1**, adaptada especificamente para a multimodalidade. Essa plataforma permite trabalhar com os modos em sincronia e oferece funcionalidades para:

- transcrever o material audiovisual e alinhá-lo ao tempo de exibição do conteúdo de vídeo;
- visualizar simultaneamente os segmentos verbais e visuais;
- criar *bounding boxes* associadas a entidades visuais evocadoras de *frames*;

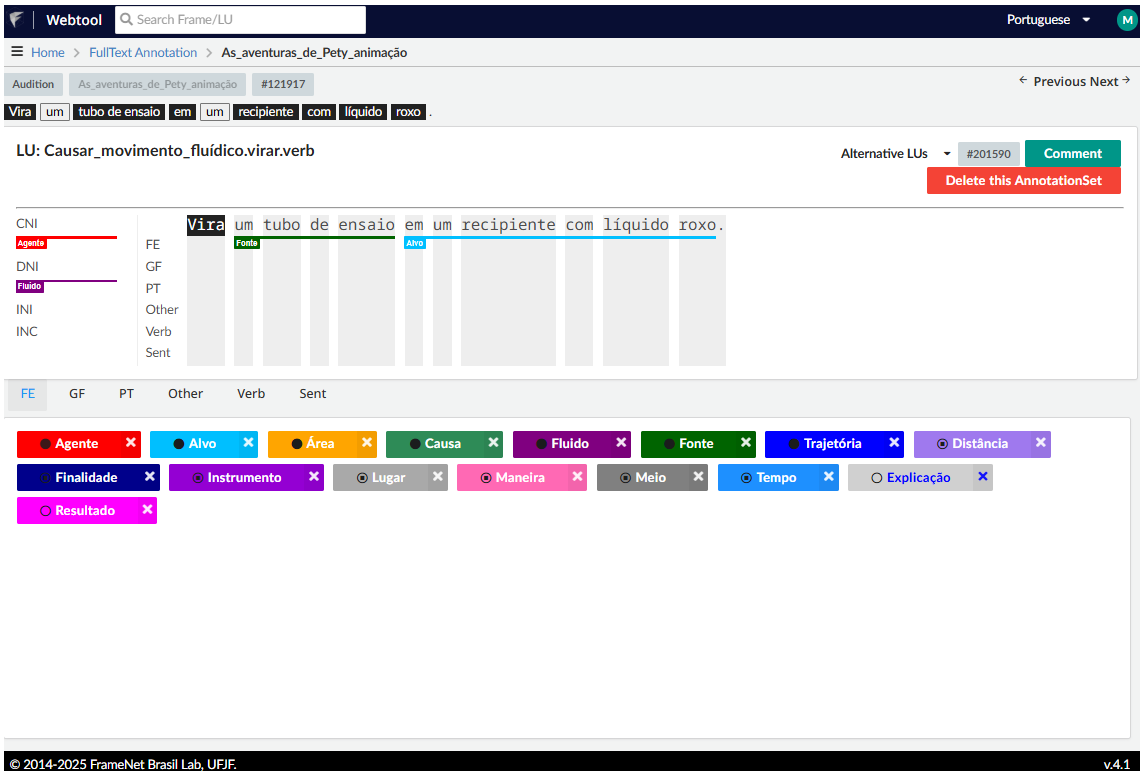
https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

- selecionar unidades lexicais e entidades visuais;
- associar essas unidades e entidades a *frames* e a seus respectivos elementos, com base na rede semântico-lexical da FrameNet Brasil.

A seguir, serão apresentados exemplos de anotação de texto e de vídeo, com o objetivo de ilustrar os dois processos adotados no *Audition*. O exemplo da Figura 17.14 é um caso de anotação linguística. No caso das modalidades verbais (como Áudio Original e Audiodescrição), a cada UL selecionada, o anotador associa o *frame* correspondente e especifica os elementos de *frame* (EFs) que estiverem lexicalmente instanciados ou inferíveis a partir do contexto verbal ou visual.

Figura 17.14: Anotação linguística no *dataset* Audition



Na sentença "Vira um tubo de ensaio em um recipiente com líquido roxo", uma das ULs anotadas é "virar.v", que evoca `Causar_movimento_fluídico`. Esse *frame* prevê EFs como:

- AGENTE: quem faz com que o fluido se mova;
- ALVO: local onde o fluido termina;
- ÁREA: configuração na qual o movimento do fluido ocorre;
- CAUSA: evento ou força que faz com que o fluido se mova;
- FLUIDO: entidade que muda de localização e se movimenta de maneira fluida;
- FONTE: local que o fluido ocupa inicialmente.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

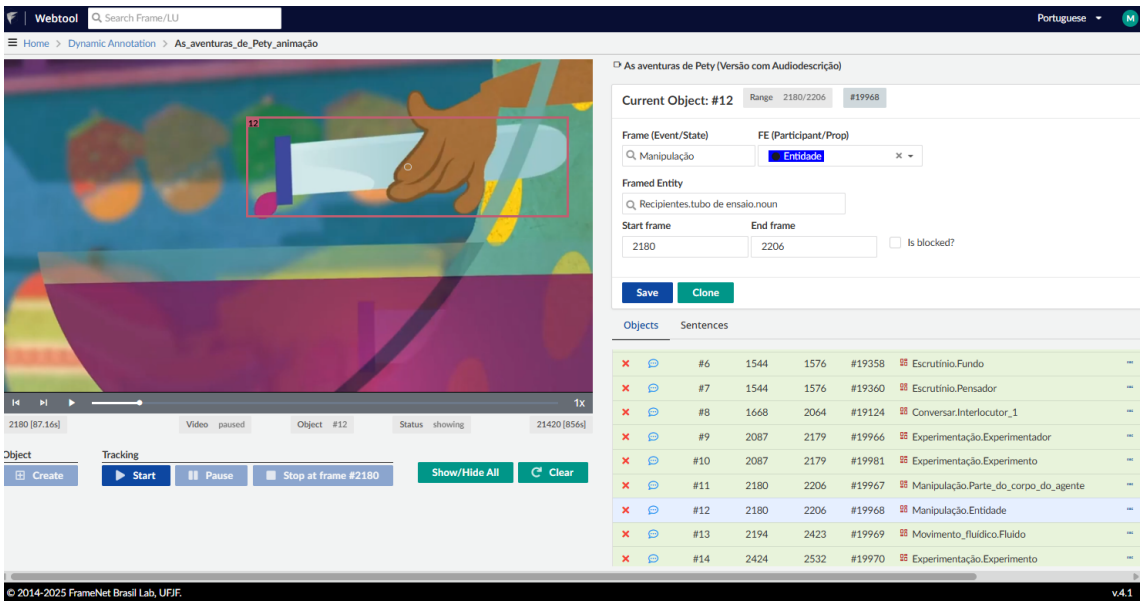
©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

- TRAJETÓRIA: A trajetória pela qual o fluido se move.

Nesta sentença, o EF AGENTE é anotado como CNI, pois sua ausência é justificada pela desinência verbal. E o EF FLUIDO foi anotado como DNI, dado que pode ser inferido pelo conteúdo do vídeo. A camada ilustrada na Figura 17.14 corresponde à anotação de texto, de onde se destacam as ULs e seus respectivos *frames* e EFs. A metodologia da FN-Br prevê anotação conjunta de camadas semântica e sintática, ambas organizadas a partir da UL (unidade predicadora). No âmbito da ReINVenTA, no entanto, a anotação ocorre apenas na camada semântica, também chamada de anotação em primeira camada.

No caso do conteúdo de vídeo, a anotação é realizada por meio da marcação de *bounding boxes*, que delimitam visualmente as entidades e situações relevantes no conteúdo audiovisual. Para auxiliar essa tarefa, foi incorporado o YOLOv3 (You Only Look Once), algoritmo de detecção de objetos em tempo real amplamente utilizado em tarefas de visão computacional, como detecção de pessoas, objetos, animais etc. No entanto, no *Audition*, o aproveitamento direto das detecções automáticas foi limitado, de modo que o procedimento predominante consistiu na delimitação manual das entidades visuais, associadas a EFs a partir do *frame* escolhido.

Figura 17.15: Anotação de conteúdo visual dinâmico no *dataset* Audition



A Figura 17.15, referente ao recorte do vídeo correspondente à sentença da AD apresentada na Figura anterior, ilustra um exemplo de anotação de conteúdo de vídeo. Nela, o anotador criou uma *bounding box* para delimitar o artefato em cena. Esse artefato foi anotado como EF ENTIDADE no *frame* `Manipulação`, que descreve eventos em que um agente manipula uma entidade com um propósito específico. Além da anotação do *frame* semântico que designa o evento ou estado, o objeto visual também é rotulado quanto ao seu próprio *frame*. Assim, a entidade delimitada é um tubo de ensaio, que está associado ao *frame* `Recipiente` na base da FN.Br. Esse duplo nível de anotação, que especifica tanto o evento quanto a entidade, promove refinamento semântico para o *dataset*. Essa infraestrutura de anotação possibilita o registro das evocações semânticas nos diversos modos presentes no *dataset*, oferecendo base para análises intermodais, comparação entre

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

recursos de acessibilidade e avaliação de preservação de sentido entre os modos, como será explorado na próxima seção.

17.5.3 Validação, Aplicações em PLN e Impactos na Tradução Audiovisual Acessível

A **primeira versão do *dataset* Audition**[19] foi empregada em um experimento de avaliação de similaridade semântica entre modos comunicativos, descrito em (Gamonal et al., 2025). O objetivo foi mensurar, por meio de uma **métrica baseada em *frames***, o grau de convergência semântica entre os diferentes modos anotados no *corpus* – especialmente o áudio original, a audiodescrição e o conteúdo de vídeo. Trata-se de um exemplo de **validação extrínseca**, tal como definida no Capítulo Conjunto de dados, *dataset* e *corpus*, pois avalia a capacidade do *dataset* de operar uma tarefa concreta de PLN: a comparação semântica multimodal.

A métrica empregada é uma **medida de similaridade semântica baseada em *frames***, inicialmente explorada pela ReINVenTA em (Viridiano et al., 2024). Seu objetivo é quantificar o alinhamento conceitual entre diferentes modos comunicativos, operando em três etapas principais:

(i) construção de uma **matriz de associação** que representa relações semânticas entre *frames* da FrameNet (como Herança, Subframe, Uso e Perspectiva), segundo sua proximidade conceitual;

(ii) aplicação de um algoritmo de ativação propagada (***spread activation***) que dissemina a ativação a partir dos *frames* anotados, com decaimento conforme a distância entre os nós no grafo semântico;

(iii) geração de **vetores de ativação** para cada instância anotada, nos quais cada dimensão representa um *frame* e seu valor de ativação acumulado.

A similaridade entre duas instâncias é, então, calculada por meio da similaridade do cosseno entre esses vetores, resultando em um *score* normalizado entre 0 (nenhuma similaridade) e 1 (similaridade total). A métrica permite, assim, avaliar não apenas a sobreposição direta de *frames*, mas também a proximidade conceitual entre eles, oferecendo uma medida interpretável da **aderência semântica** entre os modos.

Os resultados do experimento, sintetizados na Tabela 17.4, compararam os resultados entre áudio original (OA), audiodescrição (AD) e conteúdo de vídeo (CV) a partir de: (i) unidades lexicais (ULs); (ii) *frames* e elementos de *frame* (EFs) e (iii) todos os componentes anotados em conjunto (UL + *frame* + EFs), conforme apresentado na Tabela 17.4.

Tabela 17.4: Similaridade de cosseno entre representações semânticas baseadas em frame de OA, AD e VC de conteúdo visual dinâmico no *dataset* Audition

	pairs	avg	var	stdev
AD LU	323	0.42	0.05	0.22
AD Frame and FE	323	0.38	0.05	0.22

[19] O acesso aos dados da primeira versão do *dataset* estão disponíveis em: https://huggingface.co/datasets/FrameNetBrasil/Audition.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

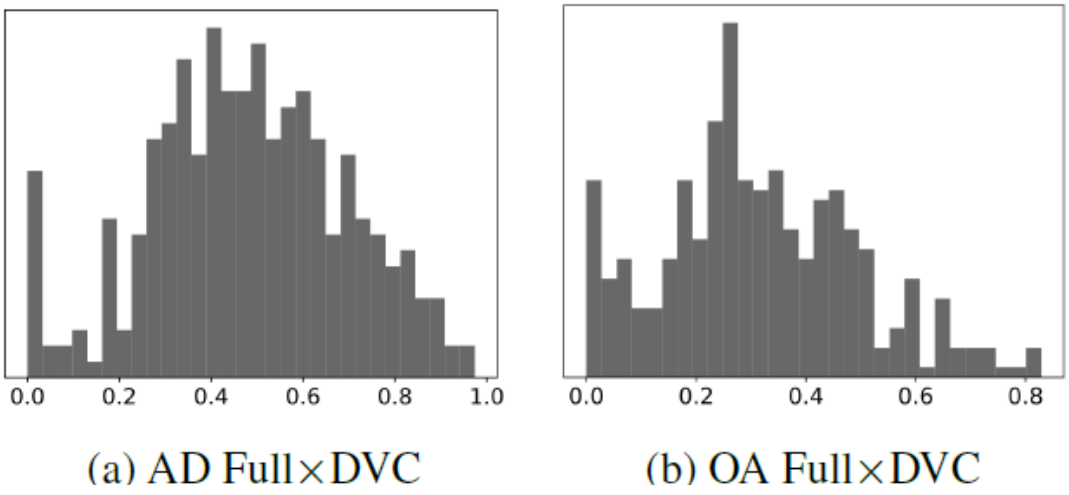
	pairs	avg	var	stdev
AD Full	323	0.49	0.05	0.21
OA LU	357	0.23	0.03	0.17
OA Frame and FE	357	0.29	0.04	0.19
OA Full	357	0.32	0.03	0.18

Fonte: (Gamonal et al., 2025) (no prelo)

As médias de similaridade por cosseno foram superiores entre as representações de AD e VC (máximo de 0.49), em relação às de OA e VC (máximo de 0.32). Esses resultados validam a hipótese de que a audiodescrição desempenha, de fato, a função de mediação semântica entre o conteúdo visual e o público-alvo, e demonstram que o *dataset* pode ser utilizado para mensurar o nível de similaridade semântico com rigor metodológico.

Essa tendência é reforçada pela distribuição dos valores de similaridade apresentados na Figura 17.16.

Figura 17.16: Distribuição de valores de similaridade entre AD full, AO Full e CVD



Fonte: (Gamonal et al., 2025) (no prelo)

A AD mantém maior convergência semântica com o vídeo do que o áudio original, como esperado. As comparações que incluem todos os componentes anotados (*full*) produziram as maiores médias de similaridade, o que reforça a validade da anotação semântica e do próprio *design* do *dataset* para finalidades analíticas.

Esses resultados indicam o potencial do *Audition* para servir a aplicações de PLN no campo da acessibilidade, como:

- **Geração automática de audiodescrição**, com base na associação entre o conteúdo visual segmentado por meio das *bounding boxes* e suas respectivas evocações semânticas;
- **Detecção de similaridade semântica entre modos**, útil para tarefas de reformulação, sumarização ou avaliação de traduções intersemióticas;

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

- **Classificação e extração de informações** em ambientes multimodais, como identificação de entidades visuais e suas funções em narrativas audiovisuais;
- **Avaliação de coerência narrativa** entre trilha sonora, falas, imagens e recursos de acessibilidade, especialmente em contextos educativos e culturais.

No campo da Tradução Audiovisual Acessível, o *Audition* representa um avanço metodológico por permitir a avaliação sistemática da construção de sentido entre modos de tradução intersemiótica. O estudo semântico por meio de *frames* entre modos – como AD, legenda, *closed caption* e *on-screen text* – permite avaliar a eficácia desses recursos e propor melhorias na sua elaboração e padronização.

Além disso, por ser um *dataset* anotado manualmente e estruturado a partir de uma teoria semântica atrelada aos estudos sociocognitivos da linguagem, o *Audition* caminha para se tornar um recurso padrão-ouro para o desenvolvimento de tecnologias voltadas à acessibilidade cultural. O alinhamento temporal e semântico entre os modos proporciona base para aplicações em *machine learning* supervisionado, incluindo o treinamento de modelos multimodais com foco em acessibilidade.

17.6 Considerações finais

Este capítulo buscou apresentar como a aplicação computacional da Semântica de *Frames*, a FrameNet Brasil, pode contribuir com aplicações transformadoras no campo do Processamento de Linguagem Natural no que tange a causas sociais. Para isso, discutiu-se o papel da anotação computacional em duas frentes – manual e automática – como estratégia central para a construção de sistemas computacionais mais interpretáveis e sensíveis aos contextos de uso, partindo de dois projetos: *DataToStopGBV* e *Audition*.

No que concerne ao projeto DataToStopVBG, os resultados evidenciam os benefícios do uso da metodologia da FrameNet para interpretar registros abertos nos sistemas de saúde, auxiliando na identificação de possíveis casos de VBG e possíveis padrões de violência. Ao avançar além das abordagens estatísticas tradicionais, o projeto revelou possíveis correlações entre violência e questões de saúde das vítimas, abrindo caminhos para investigações futuras e intervenções em parceria com profissionais da saúde. Dessa forma, os resultados revelam como a aplicação da FrameNet pode ser uma aliada no enfrentamento da violência.

Neste momento, duas vertentes do projeto estão em andamento. Primeiramente, o estudo mais profundo de possíveis padrões semânticos ligados à violência e a condições de saúde dentro desse contexto. Espera-se que esses dados sejam validados no âmbito da análise da FrameNet, em um trabalho conjunto com profissionais da saúde. Paralelamente a esse estudo, está sendo desenvolvido um painel com dados sobre a subnotificação de violências e padrões semânticos relacionados à violência de gênero, com o objetivo de gerar mapeamentos descritivos sobre tais casos no município de Recife. Por fim, o próximo passo envisionado para esta iniciativa consiste no estudo dos dados com grupos de foco mais específicos, tais como violência autoprovocada e infantil.

No âmbito do projeto *Audition*, os resultados não foram diferentes, apontando para as vantagens da aplicação da metodologia da FrameNet para tarefas de PLN com foco em acessibilidade. Os resultados da análise apontam que a audiodescrição tem maior convergência semântica com o conteúdo visual do que o áudio original, o que valida os benefícios da anotação multimodal empregada, além de indicar a possibilidade da aplicação

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

para a geração automática de audiodescrição. O *Audition*, então, se destaca como *dataset* promissor para tecnologias de tradução audiovisual acessível e referência no campo da acessibilidade cultural.

A próxima etapa central do projeto é a finalização do *dataset*. Uma versão inicial foi compilada como prova de conceito, permitindo validar o recurso por meio da medida de similaridade semântica. Paralelamente, estamos desenvolvendo um guia semântico para a audiodescrição fílmica, com o objetivo de apresentar as contribuições da tecnologia linguística para a experiência cinematográfica acessível. Essa iniciativa destaca a importância do uso de *frames* semânticos na estruturação de roteiros de audiodescrição, promovendo aderência ao conteúdo visual e o cuidado com a experiência fílmica do usuário. Espera-se, assim, que o *Audition* seja amplamente utilizado em tarefas voltadas à promoção da inclusão cultural por meio da tradução audiovisual acessível em português brasileiro.

Além da descrição dos projetos em desenvolvimento na FrameNet Brasil, este capítulo procurou evidenciar o valor da aplicação da Semântica de *Frames* como fundamento para o desenvolvimento de recursos linguísticos e metodologias voltadas a demandas sociais urgentes. Os estudos de caso apresentados ilustram caminhos distintos de aplicação da anotação semântica que demonstram a amplitude do alcance que a aplicação da FrameNet como recurso de PLN possui. Espera-se, então, que esta discussão contribua para a expansão do escopo de atuação em PLN no português brasileiro, especialmente em iniciativas que tragam benefícios concretos para a população.

Agradecimentos

O projeto DataToStopGBV foi apoiado pelo Programa de Aceleração da Fundação Patrick J. McGovern, pela Fundação José Luiz Setúbal e pelo Instituto Galo da Manhã. Também expressamos nossa gratidão aos nossos parceiros da Secretaria Municipal de Saúde do Recife — Luciana Caroline, Marcella Abath, Natalia Barros e Yana Lopes — pela valiosa colaboração e apoio contínuo ao longo deste projeto. O *dataset* Audition integra a ReINVenTA (Rede de Pesquisa e Inovação para Análise Visual e Textual de Objetos Multimodais), financiada pela FAPEMIG (RED-00106-21) e pelo CNPq (420945/2022-9), e contou também com apoio da CAPES (88887.936139/2024-00) e do CNPq (151361/2023-1). M. A. Gamonal é bolsista de pós-doutorado (CAPES – processo 88887.015648/2024-00). T. T. Torrent é bolsista de produtividade do CNPq (311241/2025-5). A rede ReiNVenta é apoiada pelo Instituto Nacional de Ciência e Tecnologia em Inteligência Artificial Responsável para Linguística Computacional, Tratamento e Disseminação da Informação (INCT-TILDIAR), financiado pelo CNPq, processo nº 408490/2024-1.

Glossário deste capítulo

- **Semântica de *Frames*:** Modelo teórico proposto por Charles Fillmore, segundo o qual o significado de uma palavra está associado a um esquema conceitual (*frame*) que organiza papéis semânticos e elementos relacionados a uma cena ou situação.
- ***Frame*:** Estrutura conceitual que representa uma situação recorrente, com seus participantes, objetos e relações, ativada pelo uso de determinadas palavras ou expressões.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

- **Elemento de *Frame* (EF):** Papel semântico associado a um *frame*, como Agente, Paciente, Causa etc.
- **Unidade Lexical (UL):** Combinação de um lema e um sentido específico que ativa um *frame*. Exemplo: em "Eu *comprei* a maçã da promoção", "comprar.v" evoca `Comércio_comprar`, já em "Os moradores *compraram* a briga contra o projeto de lei", o *frame* evocado é `Encontro_hostil`.
- **FrameNet Brasil (FN-Br):** Projeto que adapta e expande a FrameNet original (em inglês) para o português brasileiro, com adaptações teóricas e metodológicas, além de utilizar ferramentas computacionais próprias.
- **Relações *Frame*-a-*Frame:*** Conexões entre *frames* baseadas em relações, como Subframe, Herança, Perspectiva, Uso etc.
- **Relações Qualia Ternárias:** Conjunto de relações semânticas inspiradas na Teoria do Léxico Gerativo, proposta por James Pustejovsky e utilizada na FN-Br para modelar aspectos internos dos conceitos, como finalidade, composição e origem.
- **Anotação semântica baseada em *frames*:** Processo de rotulação de unidades lexicais, *frames* e elementos de *frame* em textos e outros modos com base na teoria da Semântica de *Frames*.
- **Anotação automática:** Análise feita por modelos computacionais supervisionados ou não supervisionados, que aplicam rótulos semânticos automaticamente a partir de dados previamente anotados.
- **Anotação manual:** Processo de anotação realizado por linguistas especialistas e/ou estudantes em formação em linguística, seguindo políticas rigorosas e validadas, com foco na interpretação precisa dos dados.
- **Padrão-ouro (*gold standard*):** *Dataset* anotado manualmente com alto grau de confiabilidade, usado como referência para treinamento e avaliação de sistemas automáticos.
- **Webtool:** Ferramenta web desenvolvida pela FN-Br para visualização e anotação de dados semânticos.
- **Lutma:** Ferramenta para não especialistas estruturada por meio de tutoriais com passo a passo para criação de *frames* semânticos.
- **LOME (*Large Ontology Multilingual Extraction*):** Modelo de rotulação semântica automática que opera com base na Semântica de *Frames*.
- ***Bounding box*:** Retângulo utilizado em visão computacional para delimitar objetos ou entidades visuais em imagens ou vídeos.
- **Similaridade semântica:** Medida computacional do grau de proximidade conceitual entre itens lexicais, expressões ou representações com base em sua estrutura semântica.
- ***Spread activation* (ativação propagada):** Técnica computacional que simula a ativação de nós conectados em uma rede semântica, permitindo capturar relações indiretas entre *frames*.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

- **Similaridade do cosseno:** Métrica de comparação entre dois vetores, usada para mensurar o ângulo entre eles e, assim, sua similaridade conceitual.
- **Audiodescrição (AD) - *fílmica***: Recurso de acessibilidade que consiste na narração de informações visuais relevantes para a experiência fílmica de produtos audiovisuais, voltado a pessoas cegas ou com baixa visão.
- **Tradução intersemiótica:** Tradução de significados entre sistemas semióticos diferentes, como de imagem para texto, ou de som para linguagem verbal.
- **Tradução Audiovisual Acessível:** Campo que estuda e desenvolve recursos de tradução que garantam acessibilidade a produtos audiovisuais, como audiodescrição, legendas, *closed caption* etc.
- **Violência baseada em gênero (VBG):** Qualquer forma de violência direcionada contra indivíduos com base em seu gênero, frequentemente associada a relações desiguais de poder.
- **DataToStopGBV:** Projeto que aplica anotação semântica baseada em *frames* e PLN a registros do SUS para apoiar o diagnóstico e a prevenção da violência baseada em gênero.
- ***Audition*:** *Dataset* multimodal anotado semanticamente, voltado à análise e geração de traduções acessíveis, como audiodescrição, com base na convergência entre modos comunicativos.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Capítulo 18

Redes complexas no processamento de língua natural do português

Vitor H. B. D. Santi
Lucas D. V. Figueiredo
Diego R. Amancio
Maria Cristina F. Oliveira
Osvaldo N. Oliveira Jr.

Publicado em: 16/04/2026

Este capítulo explora a aplicação de redes complexas no Processamento de Linguagem Natural (PLN) para o português. Redes complexas, estruturas compostas por nós (entidades) e arestas (conexões), servem de arcabouço para analisar relações linguísticas em tarefas como sumarização extrativa, análise de sentimentos e avaliação de qualidade textual, classificação de textos, detecção de notícias falsas, avaliação de traduções e diagnósticos médicos a partir de transcrições. São introduzidos conceitos fundamentais (como centralidade e coeficiente de aglomeração) e modelos de rede (aleatórias, livres de escala e mundo-pequeno). Um tópico central é a detecção de comunidades por algoritmos que identificam agrupamentos temáticos em textos. O capítulo destaca contribuições pioneiras de pesquisadores brasileiros, que utilizaram redes para tarefas desde tradução automática até diagnóstico médico. Para mapear a pesquisa em PLN para português, são empregadas redes de similaridade, identificando comunidades temáticas como inteligência artificial aplicada à educação, construção e utilização de *corpus* em tarefas de PLN, análise de sentimentos e desinformação. Conclui-se que, embora modelos de linguagem de larga escala (LLMs) apresentem desempenho superior em tarefas de PLN, as redes complexas continuam essenciais para tarefas de análise estrutural, interpretabilidade e visualização da literatura científica.

18.1 Introdução

Redes complexas são estruturas formadas por nós — que representam alguma entidade de interesse — e arestas — que estabelecem conexões entre pares de nós. Essas redes tipicamente descrevem sistemas com padrões de interconexão não triviais e oferecem um arcabouço para a análise de sistemas dinâmicos em diversos domínios, incluindo biologia, telecomunicações e ciências sociais (Estrada, 2011; Newman, 2010). Sua relevância se estende à área do Processamento de Linguagem Natural (PLN), em que o entendimento das relações entre palavras e conceitos é essencial para uma gama de tarefas como sumarização, análise de sentimento, modelagem de tópicos e recuperação de informação. No PLN, os

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

textos são frequentemente representados por métodos que focam no conteúdo, como o modelo *Bag-of-Words* (BoW), que desconsidera a ordem das palavras, ou por modelos de *embeddings* (como *Word2Vec* ou *BERT*), que capturam o significado semântico em um espaço vetorial. A abordagem de Redes Complexas (RC) é fundamentalmente diferente e complementar. Em vez de focar apenas na frequência ou na semântica de palavras isoladas, a RC modela explicitamente a estrutura relacional do texto. Ao representar palavras, sentenças ou conceitos como nós e suas conexões (sejam elas de coocorrência, adjacência ou relações sintáticas) como arestas, uma RC permite analisar a topologia do discurso. Essa perspectiva estrutural revela propriedades que outras representações não capturam diretamente. Com elas, é possível quantificar a coesão textual, identificar quais conceitos são mais centrais (ou "hubs") em uma argumentação e detectar a organização temática através da detecção de comunidades, como veremos na Seção 3. Portanto, as redes complexas oferecem uma "lupa" sobre a organização e a dinâmica do texto, indo além do conjunto de palavras que ele contém.

Medidas extraídas de redes complexas, como coeficientes de aglomeração e métricas de centralidade de nós, são fundamentais para entender a sua dinâmica e comportamento (OAEPublish, 2025). As redes podem também apresentar propriedades – como serem livre-de-escala ou pequeno-mundo – que impactam como as informações relacionadas às entidades se propagam e se estruturam no sistema (SpringerOpen Applied Network Science, 2017). Há grandes desafios para modelar redes, sobretudo porque muitos sistemas apresentam propriedades de autodissimilaridade em múltiplas escalas, para os quais a elaboração de modelos realistas requer esforço significativo. Ademais, a natureza evolutiva das redes complexas exige técnicas de análise para capturar as suas dinâmicas, especialmente em aplicações do mundo real, em que a disponibilidade de dados pode ser uma limitação (ScienceDirect, 2025a). Essa combinação de um sólido arcabouço teórico com a possibilidade de aplicações práticas ressalta a relevância das redes complexas para a pesquisa em PLN e seu potencial para ampliar a nossa compreensão das interações linguísticas. Também são assunto de debates as implicações do uso de redes complexas em várias áreas, por exemplo, a necessidade de uso ético de dados e potenciais vieses em interpretações algorítmicas. À medida que pesquisadores se aprofundam no estudo dessas redes, ganha cada vez mais importância a discussão sobre sua relevância e as responsabilidades associadas ao seu uso.

Neste capítulo, buscamos apresentar uma visão sobre usos de redes complexas no processamento da língua portuguesa. O conteúdo do capítulo está organizado da seguinte maneira: na Seção 18.2 introduzimos os conceitos e a terminologia básica em Redes Complexas. Na Seção 18.3 introduzimos o problema de detecção de comunidades em redes e alguns algoritmos clássicos. Na Seção 18.4 discutimos algumas aplicações de redes complexas em PLN para o português, e na Seção 18.5 recorremos às redes complexas para apresentar um panorama sobre as pesquisas em PLN no âmbito da Língua Portuguesa, não só as que envolvem redes complexas. As considerações finais são apresentadas na Seção 18.6.

18.2 Redes complexas

Como o objetivo deste capítulo é ilustrar o uso de redes complexas em PLN, apresentaremos algumas informações essenciais para a compreensão das aplicações, sem preocupação de tratar o tema de forma abrangente ou exaustiva. Assim, além de introduzir conceitos fundamentais, discutiremos as métricas das redes usadas em PLN, os modelos de redes

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

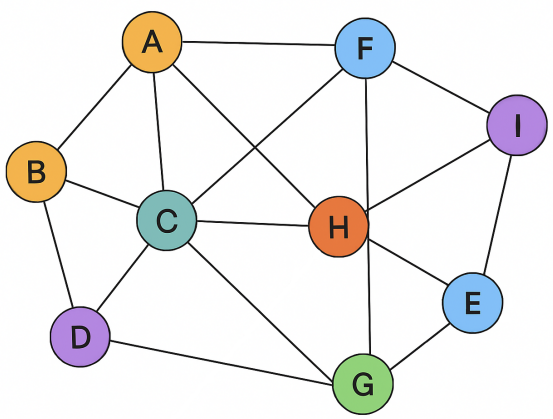
©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

e a dinâmica ou evolução temporal de redes. Uma introdução ao tema de redes complexas, que aborda conceitos e algumas aplicações, pode ser encontrada na referência (Appel; Junior, 2011).

18.2.1 Conceitos fundamentais

Os elementos fundamentais de uma rede são os **nós**, também denominados vértices, que representam as entidades de interesse, e as conexões entre nós, denominadas **arestas**. Formalmente, uma rede (ou grafo) pode ser representada como um par $G = (V, E)$, em que V é o conjunto de nós (ou vértices) e E é o conjunto de arestas.

Figura 18.1: Representação visual de uma rede não direcionada, em que os círculos representam os nós e as linhas representam as arestas.



A Figura 18.1 ilustra uma representação visual de uma rede. Cada **nó** $v \in V$ pode ser associado a dados que registram atributos específicos da entidade correspondente, representados como pares chave-valor. Por exemplo, um nó que representa uma pessoa pode incluir chaves como "nome", "idade", "localização", com seus valores correspondentes (Mousavi, 2024).

As **arestas** estabelecem as conexões entre as entidades. Cada aresta é um par ordenado ou não ordenado de nós, isto é, $e = (u, v)$ com $u, v \in V$. Quando o par é não ordenado, isto é, $(u, v) = (v, u)$, a rede é dita **não-direcionada**; quando o par é ordenado, a rede é **direcionada** e (u, v) representa uma conexão do nó u para o nó v. Além disso, as arestas podem ter pesos associados, definindo uma **rede ponderada**, em que cada aresta $e \in E$ tem um peso $w(e) \in \mathbb{R}$, representando a intensidade ou custo da relação.

Em uma rede social, uma aresta indicativa de uma relação de amizade pode ser bidirecional, enquanto que uma aresta representando uma relação de compra seria unidirecional – em ambos os casos, evidenciando o fluxo direcional da relação. Em redes sociais, um exemplo de relação unidirecional ocorre quando um usuário segue outro, mas esse segundo usuário não o segue de volta. Isso é comum em plataformas como o *X* ou o *Instagram*. A título de ilustração de uma representação textual simples, considere, por exemplo, a sentença "Redes complexas modelam relações entre elementos". Podemos representá-la como uma rede em que cada palavra é um nó (conjunto de nós V) e as arestas conectam as palavras que aparecem consecutivamente (conjunto de arestas E) (Figura 18.2):

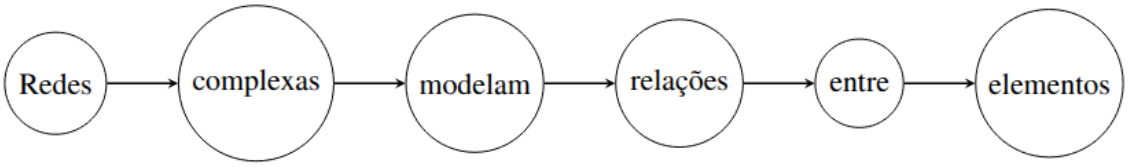
https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

$V = \{\text{Redes, complexas, modelam, relações, entre, elementos}\}$

$E = \{(\text{Redes, complexas}), (\text{complexas, modelam}), (\text{modelam, relações}), (\text{relações, entre}), (\text{entre, elementos})\}$

Figura 18.2: Representação de uma sentença como uma rede, em que cada palavra é um nó e as arestas conectam palavras consecutivas.



A importância de um nó na rede é tipicamente refletida pelo seu **grau**. O grau de um vértice v, denotado por $\deg(v)$, é definido como o número de arestas incidentes a v. Em redes direcionadas, distinguimos o grau de entrada ($\deg^-(v)$) e o grau de saída ($\deg^+(v)$), correspondendo, respectivamente, ao número de arestas que chegam e ao número de arestas que saem de v. Um grau alto tipicamente está associado a um nó mais central ou influente na rede.

A representação de entidades e conexões em termos de nós e arestas define a essência das redes complexas, permitindo modelar sistemas complexos do mundo real, como redes sociais e padrões de comunicação na linguagem (ScienceDirect, 2025b). Formalmente, o estudo de redes complexas utiliza a teoria dos grafos. Por exemplo, podemos definir um grafo $G = (V, E)$ em que V representa palavras, sentenças ou documentos, e E reflete relacionamentos entre eles, como coocorrência em um *corpus* ou relacionamentos sintáticos em uma sentença.

18.2.2 Propriedades

Várias propriedades exibidas pelas redes permitem entender a sua estrutura e comportamento, bem como analisar os mecanismos de propagação de informação na rede. **Medidas de centralidade** permitem quantificar e entender a importância de nós individuais. Além do **grau** (também conhecido como **Centralidade de Grau**), métricas como Centralidade de Intermediação e Centralidade de Proximidade fornecem indicações sobre como os nós estabelecem pontes ou estão estrategicamente posicionados dentro da rede. A **Centralidade de Intermediação** quantifica com que frequência um nó atua como uma ponte ao longo dos caminhos mais curtos entre outros nós – em PLN, por exemplo, isso poderia revelar palavras ou conceitos influentes em um texto. A **Centralidade de Proximidade** avalia quão rapidamente é possível alcançar todos os outros nós a partir de um dado nó, destacando os principais atores na distribuição de informações. A Figura 18.3 ilustra esses conceitos.

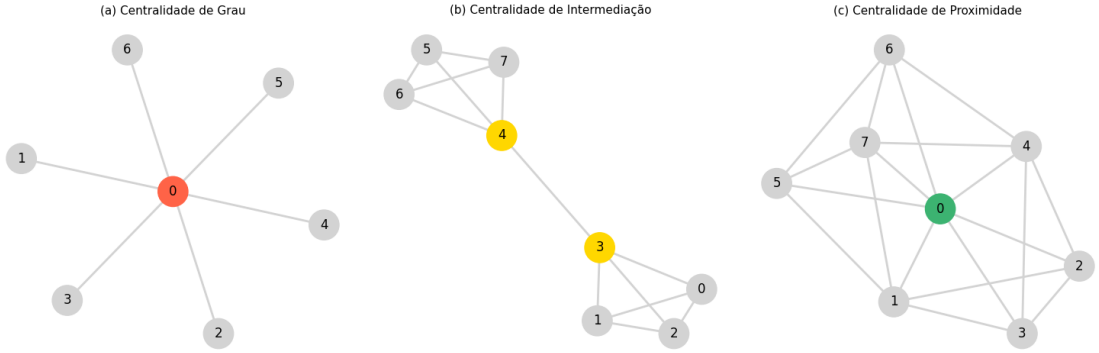
Para um grafo com N nós, as definições matemáticas de algumas dessas centralidades para um nó v são:

- **Centralidade de Grau (*Degree Centrality*):** em redes direcionadas, distingue-se entre o grau de entrada e o de saída. A forma normalizada é dada por:
  - Grau de Entrada (*In-Degree*): $C_D^{\text{in}}(v) = \frac{k_{\text{in}}(v)}{N-1}$, onde $k_{\text{in}}(v)$ é o número de arestas que chegam a v.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Figura 18.3: Representação visual para as medidas de centralidade mencionadas. (a) centralidade de grau. O nó 0 é considerado central, pois tem grau 6 enquanto os demais têm grau 1. (b) centralidade de intermediação. Os nós 3 e 4 são centrais para a conexão entre os dois grupos de nós mostrados na figura. (c) centralidade de proximidade. O nó 0 tem alta centralidade de proximidade, pois através dele os demais nós podem ser acessados mais facilmente. Ele obviamente também tem centralidade de grau, por ser mais conectado que os demais.



  - Grau de Saída (*Out-Degree*): $C_D^{\text{out}}(v) = \frac{k_{\text{out}}(v)}{N-1}$, onde $k_{\text{out}}(v)$ é o número de arestas que saem de v.

- **Centralidade de Intermediação (*Betweenness Centrality*):**

$$C_B(v) = \sum_{s \neq v \neq t} \frac{\sigma_{st}(v)}{\sigma_{st}}$$

  onde σ_{st} é o número total de caminhos mais curtos do nó s para o nó t, e $\sigma_{st}(v)$ é o número desses caminhos que passam por v.

- **Centralidade de Proximidade (*Closeness Centrality*):**

$$C_C(v) = \frac{N-1}{\sum_{u \neq v} d(v,u)}$$

  onde $d(v,u)$ é a distância do caminho mais curto entre os nós v e u.

Além de medidas de centralidade, várias estratégias foram propostas para a identificação de nós influentes na estrutura da rede, adaptadas a diferentes tipos de rede e cenários de aplicação. O **coeficiente de aglomeração** é uma medida de como os nós se agrupam localmente. Um alto coeficiente de aglomeração indica interconexões fortemente localizadas entre subgrupos de nós. No contexto de PLN, por exemplo, pode indicar a presença de conceitos ou palavras fortemente relacionados em um contexto particular ou tópico específico (Budel et al., 2023). Esta é uma propriedade particularmente relevante para análise semântica, em que é crucial entender os relacionamentos entre palavras ou frases ao implementar tarefas como desambiguação lexical ou reconhecimento de entidades nomeadas (Al-Rfou et al., 2014).

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

O **Coeficiente de Aglomeração Local** para um nó v é definido como:

$$C(v) = \frac{2E_v}{k_v(k_v - 1)}$$

onde k_v é o grau do nó v (o número de seus vizinhos) e E_v é o número de arestas existentes entre os vizinhos de v. O denominador $k_v(k_v - 1)/2$ representa o número máximo possível de arestas entre os vizinhos de v.

Uma revisão abrangente dessas estratégias as categoriza em abordagens tradicionais baseadas em topologia, para otimização, modelos de difusão de informação, métodos de aprendizado de máquina e técnicas mais recentes aplicáveis a redes dinâmicas (Chen et al., 2025). Ao contrário das medidas clássicas de centralidade, esses métodos consideram interações na rede ao longo do tempo, além das propriedades estáticas. Outra métrica relevante, denominada **Comprimento de Caminho Médio**, indica o número médio de passos ao longo dos caminhos mínimos entre todos os possíveis pares de nós. Em redes complexas, essa métrica tende a ter valores baixos, o que facilita a disseminação eficiente de informação na rede. O **Diâmetro**, por outro lado, é definido como o maior valor entre os comprimentos dos menores caminhos entre qualquer par de nós da rede.

As definições matemáticas para essas métricas são:

- **Comprimento de Caminho Médio (Average Path Length):**

$$L = \frac{1}{N(N-1)} \sum_{i \neq j} d(v_i, v_j)$$

  onde $d(v_i, v_j)$ é a distância do caminho mais curto entre os nós v_i e v_j.

- **Diâmetro (Diameter):**

$$D = \max_{i,j} d(v_i, v_j)$$

  que representa a maior distância de caminho mais curto entre qualquer par de nós (v_i, v_j) na rede.

18.2.3 Tipos de redes

As redes podem ser categorizadas com base em suas propriedades estruturais e comportamento, que de alguma forma refletem a natureza intrincada dos sistemas do mundo real. Muito conhecidas são as redes aleatórias, livres de escala e mundo-pequeno.

Redes aleatórias

Nesse tipo de rede, as conexões entre os nós são estabelecidas aleatoriamente. Pode-se citar, a título de exemplo, redes de sensores distribuídos aleatoriamente em ambientes físicos ou redes de coocorrência de palavras construídas a partir de textos embaralhados ou gerados aleatoriamente. A despeito de terem servido de base para estudos teóricos importantes na ciência de redes, elas não refletem o comportamento e os padrões de interconectividade observados em sistemas do mundo real. Ainda assim, a análise desse tipo de rede é importante para ilustrar os princípios fundamentais que governam redes de tipos mais complexos. O modelo canônico para redes aleatórias é o de **Erdős-Rényi (ER)**, $G(N, p)$. Sua formação se dá a partir de N nós, onde cada par de nós possível é

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

conectado por uma aresta com uma probabilidade p independente. A distribuição de grau $P(k)$ de uma rede aleatória segue uma distribuição binomial, que, para redes grandes e com baixa probabilidade de conexão, se aproxima de uma **distribuição de Poisson**:

$$P(k) \approx \frac{\lambda^k e^{-\lambda}}{k!}$$

onde λ é o grau médio da rede. A maioria dos nós tem um grau próximo da média. Em redes aleatórias do tipo ER, a distância típica entre vértices é relativamente pequena.

Redes livres de escala

Essas redes se caracterizam pelo fato de a distribuição de grau dos seus nós seguir uma lei de potência. Em outras palavras, um número reduzido de nós tem um altíssimo número de conexões (são os chamados "*hubs*"), enquanto a grande maioria dos demais é muito pouco conectada. Redes de citação entre artigos científicos e redes de coocorrência de palavras em grandes *corpora*, como a *Wikipedia* ou o *Google Books*, são exemplos de redes livres de escala (Steyvers; Tenenbaum, 2005). Em PLN, esse fenômeno está relacionado à **Lei de Zipf** (Zipf, 1949), um princípio fundamental da linguística. A Lei de Zipf postula que a frequência de uma palavra em um *corpus* é inversamente proporcional à sua posição em um *ranking* de frequência. Em uma rede de coocorrência de palavras, os termos mais frequentes (os *hubs*) naturalmente coocorrem com uma variedade muito maior do que outras palavras. Por outro lado, a maioria das palavras — aquelas de baixa frequência, que compõem a chamada "cauda longa" de Zipf — tende a apresentar poucas conexões.

A distribuição de grau em lei de potência observada nessas redes textuais, ilustrada na Figura 18.4, é, em grande parte, um reflexo da Lei de Zipf que governa a frequência das palavras. O surgimento desse tipo de rede pode ser explicado por mecanismos como crescimento e conexão preferencial, ilustrados no conhecido modelo Barabási-Albert (BA), proposto por Albert-László Barabási e Réka Albert em 1999 (Barabási; Albert, 1999). O padrão de distribuição de conexões segue uma **lei de potência**, em que a probabilidade $P(k)$ de um nó ter k conexões é dada por:

$$P(k) \sim k^{-\gamma}$$

em que γ é o expoente da distribuição, geralmente um valor entre 2 e 3. A formação pelo modelo **Barabási-Albert (BA)** se baseia no mecanismo de **conexão preferencial**: a probabilidade $\Pi(v_i)$ de um novo nó se conectar a um nó existente v_i é proporcional ao seu grau k_i:

$$\Pi(v_i) = \frac{k_i}{\sum_j k_j}$$

A inserção de novos nós com probabilidade de conexão proporcional ao grau dos nós já existentes resulta no fenômeno conhecido como "rico-fica-mais-rico", que induz à formação de *hubs*. Embora o modelo BA reproduza a característica de existência de *hubs*, algumas características de redes reais ainda não são representadas por este modelo. Uma delas é a possibilidade de surgimento de novas arestas entre vértices antigos ou o desaparecimento de arestas.

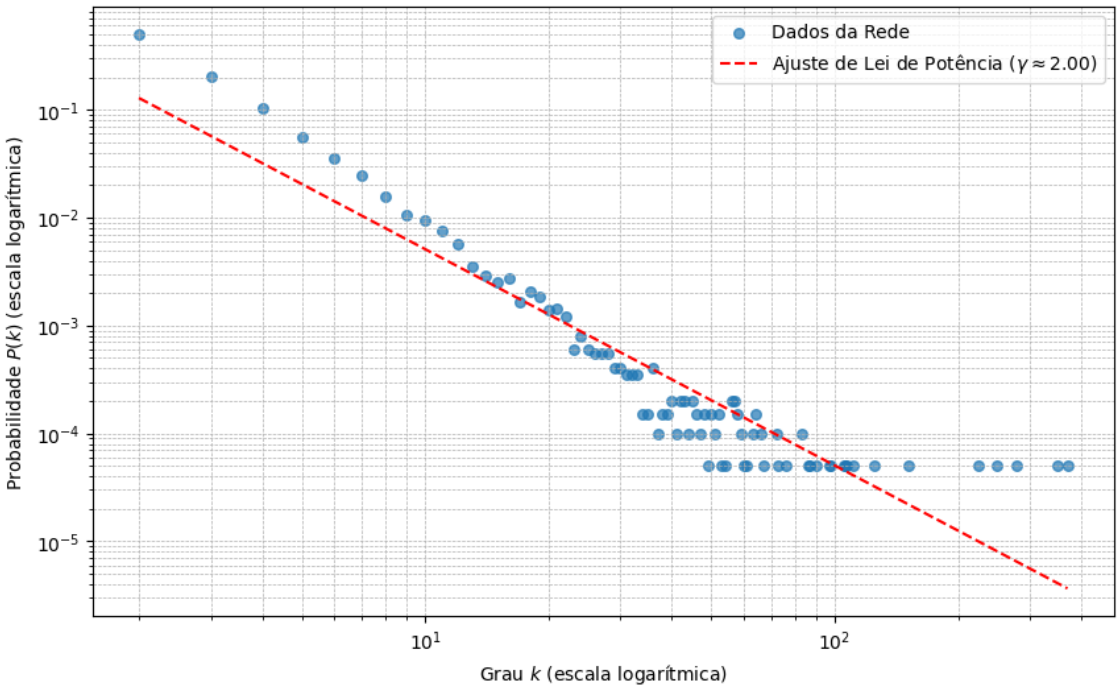
Redes mundo-pequeno

Essas redes são caracterizadas por comprimentos de caminho curtos e coeficientes de aglomeração altos, o que permite que a maioria dos nós seja acessível a partir de qualquer nó

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Figura 18.4: Distribuição de probabilidade para um grau k (eixo Y, logarítmico) versus o grau k (eixo X, logarítmico) em uma rede livre de escala. Os dados são relativos a uma rede sintética obtida usando um algoritmo que simula o modelo de Barabási-Albert. Foram gerados 20.000 nós, sendo que cada novo nó inserido na rede se conecta a dois nós pré-existentes. Os pontos seguem aproximadamente a linha tracejada vermelha indicativa da lei de potência, que é a "assinatura" visual dessa distribuição.



com um número pequeno de passos. Essa propriedade é frequentemente exemplificada pelo conceito dos "seis graus de separação". Segundo esse conceito, quaisquer duas pessoas no mundo estariam — em média — a seis conexões entre si. A rede de atores de Hollywood que atuaram juntos exibe propriedades de mundo pequeno, assim como a *WordNet* e redes sintáticas de línguas naturais (Steyvers; Tenenbaum, 2005).

O modelo mundo-pequeno, introduzido em 1998 por Duncan J. Watts e Steven Strogatz (Watts; Strogatz, 1998), ilustra como uma rede pode transicionar de um formato de grade regular para uma estrutura mais aleatória enquanto mantém suas características de mundo-pequeno. Isso ocorre por meio de uma reatribuição aleatória de arestas, o que enfatiza as conexões sem alterar significativamente a estrutura global. O modelo de **Watts-Strogatz (WS)** descreve a formação dessas redes a partir de uma rede regular (anel) em que cada aresta é religada aleatoriamente com uma probabilidade p. O padrão de distribuição não é definido por uma fórmula única de grau, mas pela coexistência de duas propriedades:

1. **Alto Coeficiente de Aglomeração (C):** Similar ao de uma rede regular e muito maior que o de uma rede aleatória ($C \gg C_{\text{aleatória}}$).

2. **Baixo Comprimento de Caminho Médio (L):** Similar ao de uma rede aleatória, crescendo logaritmicamente com o número de nós ($L \sim \log(N)$).

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Redes complexas possuem várias características definidoras, incluindo heterogeneidade na distribuição do grau dos nós, alta aglomeração e modularidade. A diversidade nas conexões leva à formação de grupos coesos (comunidades) na rede, com conexões mais esparsas mantidas entre diferentes grupos. Essa modularidade é crucial para entender como as redes operam e evoluem.

18.2.4 Dinâmica

O estudo da dinâmica de redes complexas permite determinar como elas evoluem ao longo do tempo e influenciam os vários processos que nelas ocorrem, o que é importante para entender fenômenos como a propagação de doenças e a difusão de informações. As redes não são estáticas; elas mudam e se adaptam à medida que novos nós e links são adicionados ou removidos. Essa dinâmica pode ser entendida utilizando diferentes modelos teóricos, permitindo aos pesquisadores prever estados futuros das redes e compreender os mecanismos subjacentes que impulsionam as mudanças. Esses modelos podem descrever como um estado ou informação se propaga de nó para nó. Dois exemplos clássicos são os modelos de epidemia, como o SIR (Suscetível-Infectado-Recuperado), e os modelos de difusão de informação, como o Modelo de Limiar Linear:

- **Modelo SIR (Suscetível-Infectado-Recuperado):** usado para modelar a propagação de doenças; nesse modelo, cada nó pode estar em um de três estados. A dinâmica pode ser descrita por um sistema de equações diferenciais (em uma aproximação de campo médio):

$$\begin{aligned}\frac{ds}{dt} &= -\beta\langle k\rangle s(t)i(t)\\ \frac{di}{dt} &= \beta\langle k\rangle s(t)i(t) - \delta i(t)\\ \frac{dr}{dt} &= \delta i(t)\end{aligned}$$

  Aqui, $s(t)$, $i(t)$ e $r(t)$ são as frações de nós *suscetíveis*, *infectados* e *recuperados*, respectivamente. β é a taxa de transmissão, δ é a taxa de recuperação e $\langle k\rangle$ é o grau médio da rede. A condição para uma epidemia ocorrer é que o número básico de reprodução, $R_0 = \frac{\beta\langle k\rangle}{\delta}$, seja maior do que 1.

- **Modelo de Limiar Linear (*Linear Threshold Model*):** usado para modelar a difusão de inovações ou comportamentos. Cada nó v possui um limiar de ativação $\theta_v \in [0, 1]$. Um nó inativo v torna-se ativo no instante $t+1$ se a soma das influências de seus vizinhos já ativos, ponderada por pesos b_{uv}, atinge seu limiar:

$$\sum_{u \in \text{vizinhos_ativos}(v)} b_{uv} \geq \theta_v$$

  onde $\sum_u b_{uv} \leq 1$. O processo continua até que nenhum novo nó possa ser ativado.

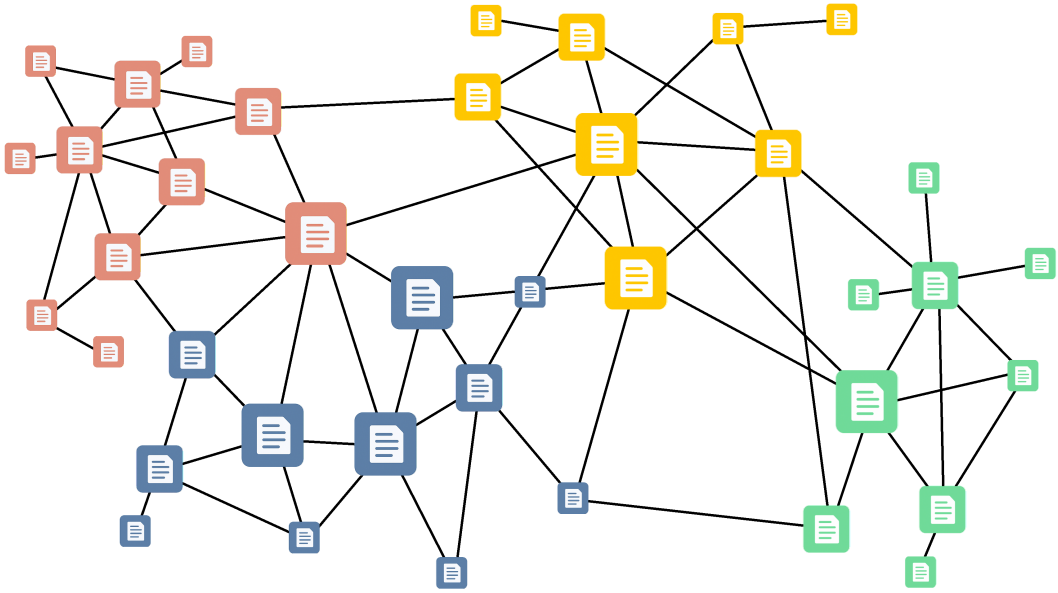
https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

18.3 Detecção de comunidades

No contexto do PLN, o conceito de "comunidade" estabelece um elo essencial entre a estrutura da rede e o significado linguístico. Na prática, uma comunidade identifica um agrupamento temático coeso ou um campo semântico em um *corpus*. De fato, em redes de linguagem, principalmente em redes semânticas (Steyvers; Tenenbaum, 2005), é comum observar grupos de nós fortemente conectados entre si e menos conectados com elementos externos ao grupo. No contexto de redes complexas, tais grupos são denominados comunidades. A Figura 18.5 ilustra uma rede hipotética e sua organização em comunidades, indicadas pelas diferentes cores. Em tarefas de processamento de linguagem e em diversas outras aplicações que envolvem a análise de redes, é comum a realização de uma etapa de identificação de comunidades.

Figura 18.5: Representação visual das comunidades detectadas em uma rede não direcionada.



Considere, por exemplo, uma rede construída a partir de um grande conjunto de notícias, em que os nós correspondem a palavras e as arestas conectam palavras que aparecem frequentemente juntas. Nessa rede, esperaríamos que palavras como "futebol", "campeonato", "jogador" e "partida" formassem um agrupamento muito denso, representando o tópico "Esportes" (ou especificamente "Futebol"). Da mesma forma, palavras como "governo", "eleição", "congresso" e "partido" formariam outra comunidade ("Política"). Identificar essas estruturas pode ser útil para revelar a organização temática latente do texto e constitui uma etapa crucial para tarefas como a modelagem de tópicos (em que comunidades podem indicar tópicos), a sumarização (ajudando a identificar os conceitos centrais de um texto) ou a desambiguação de palavras (entendendo os diferentes contextos em que uma palavra é usada).

Muitos algoritmos são usados para identificar comunidades em redes complexas, em que se busca maximizar a coesão entre os vértices de um mesmo grupo. Esses algoritmos são aplicados em vários domínios além do PLN, incluindo análise de redes sociais, biologia,

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

detecção de fraudes e telecomunicações. Métodos proeminentes incluem o *Algoritmo de Propagação de Rótulos* (LPA, *Label Propagation Algorithm*) (Raghavan et al., 2007); o algoritmo *Infomap* (Rosvall; Bergstrom, 2008); o algoritmo *Girvan-Newman* (Newman; Girvan, 2004), e o algoritmo *Louvain* (Blondel et al., 2008).

Eles variam significativamente quanto à abordagem (Palacio-Niño; Berzal, 2024) (Singh; Garg, 2025). O algoritmo de *Girvan-Newman* (Newman; Girvan, 2004), um dos pioneiros, opera pela remoção iterativa de arestas: ele calcula a centralidade de intermediação das arestas e remove aquelas com maior centralidade até que a rede seja segmentada em comunidades distintas. Essa abordagem enfatiza a importância da conectividade das arestas. O LPA também adota uma estratégia simples, atribuindo rótulos a nós com base no voto da maioria de seus vizinhos. Inicialmente, cada nó recebe um rótulo único e, em cada iteração, os rótulos dos nós são atualizados para refletir o rótulo mais frequente entre seus vizinhos. Esse processo iterativo continua até que os rótulos se estabilizem, resultando em comunidades distintas. O LPA é eficiente e escalável, adequado para grandes redes. No entanto, seus resultados podem ser sensíveis à atribuição dos rótulos iniciais e à ordem de processamento dos nós, muitas vezes exigindo múltiplas execuções para alcançar resultados consistentes. Uma abordagem de consenso entre partições distintas pode também ser necessária para melhorar a qualidade das partições (Lancichinetti; Fortunato, 2012).

Já o algoritmo *Infomap* adota princípios da teoria da informação e passeios aleatórios. Um passeio aleatório (*random walk*) é um processo estocástico que descreve a movimentação de um "agente" que se desloca de nó em nó em uma rede seguindo regras probabilísticas. O algoritmo busca minimizar o comprimento do caminho de um passeador aleatório através da rede, capturando o fluxo de informação. Ao conceituar a rede como um mapa, o *Infomap* identifica agrupamentos de nós que são frequentemente visitados pelo passeador aleatório. Este algoritmo é reconhecido por sua robustez e estabilidade, sendo uma opção confiável quando for essencial entender o fluxo de informação na rede.

O *Louvain* é um algoritmo heurístico introduzido por Blondel et al. (Blondel et al., 2008) para detecção de comunidades em grandes redes. O algoritmo otimiza uma métrica de modularidade, uma função de qualidade que mede quão bem uma partição de uma rede separa conexões densas intra-comunidade de conexões esparsas inter-comunidade. Ele opera em duas fases principais que são repetidas iterativamente: (i) Otimização local de modularidade, em que cada nó é inicialmente atribuído à sua própria comunidade, e nós são movidos para comunidades vizinhas se isso resultar em um ganho de modularidade; e (ii) Agregação de comunidades, a partir do momento que não há ganhos de modularidade adicional no nível local, cada comunidade detectada é colapsada em um único nó, produzindo uma "super-rede" menor. A primeira fase é então reaplicada. Esse processo hierárquico e multinível prossegue até que nenhuma melhoria na modularidade possa ser alcançada. O algoritmo é eficiente, mostrando-se adequado para redes muito grandes, com milhões de nós, e hierárquico, produzindo uma hierarquia de comunidades em diferentes níveis de resolução. Também é estocástico (i.e., não determinístico), uma vez que a ordem em que os nós são processados ao decidir mover um nó para uma comunidade vizinha afeta o resultado. Mais recentemente, uma extensão do *Louvain* foi criada de forma a garantir que as comunidades encontradas sejam comunidades conectadas, além de aumentar a velocidade do método (Traag et al., 2019).

A despeito dos avanços significativos, a detecção de comunidades segue com desafios em aberto. Uma das questões fundamentais é a ausência de métodos universalmente aplicáveis a diferentes estruturas de rede. Os algoritmos tradicionais geralmente pressupõem que as comunidades são grupos distintos e densos de nós, o que pode levar a imprecisões

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

em cenários em que as comunidades se sobrepõem, i.e., existem nós que pertencem a múltiplas comunidades. Além disso, é difícil avaliar objetivamente a qualidade do resultado, em vista da ausência de uma única métrica aplicável aos diferentes métodos. Métricas como o *Índice Rand Ajustado* (ARI), o *Índice Omega* e a *Modularidade* não capturam adequadamente a complexidade de estruturas de comunidades sobrepostas. Assim, muitas vezes é necessário adaptar os métodos de avaliação às características específicas das redes e comunidades analisadas. Outro desafio é identificar corretamente a granularidade (ou o nível de resolução) das comunidades detectadas, i.e., se estamos buscando identificar numerosas comunidades menores ou identificar um número menor de comunidades, mas maiores. A escolha afeta a interpretabilidade e a usabilidade da estrutura de comunidades obtida. Além disso, a análise das propriedades estruturais locais, que poderia aprimorar o processo de detecção, é muitas vezes limitada pela escalabilidade dos algoritmos existentes e pela ausência de dados de referência (*ground truth*) para validação.

A complexidade e a variabilidade inerentes às redes ressaltam a necessidade de abordagens customizadas, e pesquisas seguem na busca por ampliar a escalabilidade, o desempenho e a aplicabilidade dos algoritmos. Tendências incluem a integração de técnicas de aprendizado de máquina, o desenvolvimento de métodos de detecção de comunidades dinâmicas para análise em tempo real e algoritmos para lidar com comunidades sobrepostas.

18.4 Aplicações de redes complexas em PLN para o português

O desenvolvimento de aplicações de redes complexas em PLN tem tido participação significativa de pesquisadores brasileiros. A título de ilustração, dos cerca de 100 artigos indexados na *Web of Science* obtidos em uma busca expressa como ("natural language processing" and "complex network*"), realizada em setembro de 2025, aproximadamente um quarto são de autores brasileiros. Entretanto, essa busca não conseguiu capturar um conjunto representativo de artigos na área, como percebemos ao fazer um levantamento das contribuições de brasileiros da área. Possivelmente ocorreu o mesmo com autores de outros países. Aparentemente, a área está sub-representada na *Web of Science* e, portanto, o número real de artigos deve ser consideravelmente maior, provavelmente da ordem de algumas centenas. De todo modo, é notável a porcentagem de trabalhos brasileiros no tópico.

Em vista da grande flexibilidade das redes complexas como modelo de representação, sua diversidade de aplicações é enorme. A vantagem dessa abordagem é explorar a topologia da rede textual, que é complementar às análises de conteúdo. Por exemplo, em tarefas de sumarização extrativa, nos moldes das discutidas no Capítulo 5, a centralidade de um nó (representando uma sentença) pode indicar sua importância para o resumo, independentemente do peso semântico de suas palavras isoladas. Em análise de sentimentos, por exemplo, em textos de postagens em redes sociais (ver Capítulo 16), a forma como os conceitos positivos e negativos se agrupam na rede pode revelar a polaridade de um texto de opinião com mais robustez. Na avaliação de traduções (ver Capítulo 4), métricas de rede mostraram-se capazes de distinguir a qualidade e a fluidez de textos traduzidos por máquinas ou por humanos.

Apresentamos, nesta seção, um breve resumo de algumas contribuições de autores brasileiros, que foram pioneiros em demonstrar como essas métricas estruturais capturam aspectos complexos da qualidade textual, da simplificação e até mesmo do diagnóstico médico baseado em transcrições de fala.

Parcela significativa das contribuições descreve aplicações que abordam a classificação

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

de textos modelados como redes complexas. A avaliação da qualidade de redações é uma das aplicações de interesse (ver Capítulo 7). Um dos primeiros artigos a utilizar redes complexas em PLN (Antiqueira et al., 2007) identificou uma alta correlação de métricas de redes com a qualidade dos textos em redações de vestibular. As redações foram modeladas como redes de coocorrência em que os nós representam as palavras. As notas atribuídas por professores aos itens de coesão/coerência, correção gramatical e adequação ao tema diminuíram com o aumento dos valores de métricas como grau de saída, coeficiente de aglomeração e desvio da dependência linear no crescimento da rede. A maior correlação foi observada para o quesito coesão/coerência, indicando que as métricas da rede permitiam capturar esse aspecto qualitativo do texto.

Como um exemplo, considere o seguinte trecho de texto que apresenta boa coesão: *O pesquisador instalou gravadores na floresta. Os sons captados revelaram a presença de aves raras.* Nesse caso, uma rede de coocorrência de palavras revelaria conexões claras entre termos como *pesquisador*, *gravadores*, *floresta*, *sons* e *aves*, formando uma estrutura interligada com caminhos curtos entre os nós. A palavra *sons*, por exemplo, funcionaria como um elo entre a ação de gravar e a descoberta das aves, contribuindo para a coesão semântica do texto. Agora, observe um texto menos coeso: *O pesquisador instalou gravadores na floresta. A cidade estava movimentada com o festival de música.* Embora ambas as frases sejam gramaticalmente corretas, elas pertencem a contextos distintos e não compartilham elementos semânticos relevantes. A rede de coocorrência resultante apresentaria dois componentes desconexos, refletindo a ausência de continuidade temática. Métricas como o coeficiente de aglomeração reduzido ou um maior diâmetro da rede poderiam ser utilizadas para indicar essa quebra de coesão.

A topologia e a dinâmica de redes revelam características estruturais de um texto, o que permite avaliar questões linguísticas inerentes à tradução manual e automática (abordada no Capítulo 4). Amancio et al. (Amancio et al., 2011a) mostraram que é possível distinguir diferentes qualidades de traduções geradas por ferramentas de tradução automática de suas correspondentes manuais utilizando métricas como graus de entrada (ID) e de saída (OD), coeficiente de agrupamento (CC) e caminhos mínimos (SP). Os autores verificaram, por exemplo, que em redes de traduções automáticas o OD médio excede consistentemente os valores obtidos para traduções manuais, e que os valores de CC dos textos originais não são preservados em traduções manuais, mas o são em boas traduções automáticas. Isso provavelmente reflete os rearranjos textuais realizados por humanos durante a tradução manual.

Considere a frase original em inglês: *The biologist recorded bird songs at sunrise in the rainforest.* Essa sentença formaria uma rede de coocorrência coesa, conectando termos como *biologist*, *recorded*, *bird songs*, *sunrise*, e *rainforest*, todos relacionados ao contexto de pesquisa ecológica. Agora, compare duas possíveis traduções para o português:

- Tradução automática: O biólogo registrou sons de pássaros na manhã na floresta.
- Tradução humana: O biólogo gravou cantos de aves ao amanhecer na floresta tropical.

No exemplo, a tradução automática gerou estruturas mais genéricas e menos conectadas semanticamente. Termos como *registrou*, *sons*, *manhã* e *floresta* são corretos, mas formam uma rede com menor densidade e maior diâmetro, refletindo uma coesão temática mais fraca. Já a tradução humana preservou nuances do original, como *cantos de aves*

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

e *amanhecer*, e introduziu termos mais específicos, como *floresta tropical*, resultando em uma rede mais compacta e interligada.

Ao comparar diferentes redes, métricas como o coeficiente de aglomeração, a densidade de conexões e o número de componentes podem indicar maior fidelidade semântica e coesão na tradução humana. Esse tipo de análise é especialmente útil para avaliar sistemas de tradução automática anteriores à era dos modelos que empregam aprendizado de máquina, que frequentemente apresentavam limitações na preservação de contexto e fluidez textual. O estudo citado (Amancio et al., 2011a) mostrou que as redes de palavras desses dois tipos de texto, ou seja, gerados por traduções automáticas ou manuais, possuem "assinaturas" topológicas distintas. Por exemplo, métricas como o grau médio das palavras e o coeficiente de agrupamento diferem. Isso permitiu treinar classificadores para automaticamente distinguir traduções humanas de traduções automáticas de baixa qualidade.

Ressalte-se que o artigo em questão foi publicado há mais de 15 anos, quando o desempenho dos tradutores automáticos era muito inferior ao atual. Abordagem similar foi aplicada para avaliar a qualidade da tradução automática, inclusive comparando traduções do português para o espanhol e para o inglês (Amancio et al., 2011a). As métricas de redes foram usadas como entrada em algoritmos de aprendizado de máquina que conseguiram distinguir textos oriundos de tradução humana, tradução automática de alta qualidade e tradução automática de baixa qualidade. Um resultado importante foi que a acurácia na classificação aumentou quando foram consideradas métricas de diferentes níveis hierárquicos na rede. Concluiu-se que a possível captura de um contexto mais amplo com os níveis hierárquicos pode ser útil para aumentar a acurácia de classificação com aprendizado de máquina (Amancio et al., 2011a). É interessante fazer uma analogia com o que se faz hoje com aprendizado de máquina nos modelos de língua de larga escala (LLMs, de *large language models* em inglês), em que os *embeddings* e o mecanismo de atenção fornecem contexto amplo, como discutido na Seção 4.2.6.

Três artigos abordam a adequação de redes complexas para tarefas de sumarização extrativa em português (discutidas no Capítulo 5 deste livro). No primeiro, foram empregadas métricas de redes complexas para selecionar sentenças em um resumo extrativo (Antiqueira et al., 2009). A rede que representa um texto era constituída por nós correspondentes às sentenças, enquanto arestas conectavam sentenças que compartilhavam substantivos considerados significativos. Um conjunto de 14 sumarizadores foi produzido empregando diferentes métricas, como o grau dos nós e o comprimento de caminhos mínimos. Algumas versões apresentaram desempenho superior ao de sumarizadores que não empregavam conhecimento linguístico profundo. Em alguns casos, os resultados foram comparáveis aos dos sumarizadores de ponta baseados em recursos linguísticos de alto custo.

Para ilustrar a abordagem de sumarização baseada em redes de coocorrência semântica, considere o seguinte minitexto composto por três sentenças:

- **S1**: A inteligência artificial avança rapidamente no Brasil.
- **S2**: O avanço da inteligência artificial impacta diversos setores da sociedade.
- **S3**: O clima no Brasil é predominantemente tropical.

O modelo constrói uma rede em que os nós representam sentenças e as arestas conectam sentenças que compartilham conceitos ou termos relevantes (por exemplo, *inteligência artificial*, *Brasil*, *avanço*). Nesse cenário:

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

- **S1** compartilha os termos *inteligência artificial* e *Brasil* com **S2** e **S3**, respectivamente.
- **S2** conecta-se a**S1** por meio dos termos *inteligência artificial* e *avanço*.
- **S3** conecta-se a **S1** por meio do termo *Brasil*, mas não compartilha termos com **S2**.

A estrutura resultante é uma rede em que **S1** atua como um nó central, com maior grau de conectividade. De acordo com métricas de centralidade (como *centralidade de grau* ou *PageRank*), **S1** seria considerada a sentença mais representativa do conteúdo global do texto e, portanto, uma forte candidata à composição de um resumo extrativo. Esse tipo de abordagem é particularmente útil para sumarização automática, pois permite identificar sentenças que funcionam como "pontes semânticas" entre diferentes partes do texto, mesmo em *corpora* maiores e mais complexos.

Em outro trabalho (Tohalino; Amancio, 2018), redes também com sentenças representadas por nós foram usadas para identificar as sentenças mais relevantes para sumarização em vários documentos simultaneamente, em português e em inglês. No terceiro artigo (Amancio et al., 2012a), as redes eram formadas por coocorrência de palavras, ou seja, os nós eram constituídos por palavras em português. Mostrou-se que a incorporação de conhecimento linguístico foi capaz de melhorar, de maneira modesta, o desempenho do sumarizador automático. A conclusão principal foi quanto à utilidade de utilizar métricas de redes complexas para a tarefa de sumarização.

A simplificação de textos é essencial para garantir acessibilidade a leitores com habilidade de leitura limitada, o que tem sido objeto de pesquisa para o português do Brasil, como discutido no Capítulo 6 deste livro. Um desafio é identificar o nível de simplificação adequado para uma audiência específica. Em um trabalho com textos representados por redes de coocorrência, verificou-se que a regularidade topológica se correlaciona negativamente com a complexidade textual (Amancio et al., 2012b). Além disso, a distância entre conceitos, representados como nós, tende a diminuir em textos menos complexos. As métricas das redes complexas foram tratadas com técnicas multivariadas de reconhecimento de padrões, o que permitiu distinguir entre textos originais e suas versões simplificadas. Para cada texto original, duas versões simplificadas foram geradas manualmente com um número crescente de operações de simplificação. Como esperado, a distinção foi mais fácil para as versões fortemente simplificadas, em que as métricas mais relevantes foram força do nó, caminhos mínimos e diversidade. Além disso, a discriminação de textos complexos foi aprimorada com métricas hierárquicas mais altas da rede, apontando, assim, para a utilidade de considerar contextos mais amplos em torno dos conceitos.

A distinção de textos com opiniões positivas ou negativas sobre um tema (problema abordado no Capítulo 16) também pode ser feita classificando-se redes complexas que representam textos. No artigo de Amancio et al. (Amancio et al., 2011b), foram analisados pares de artigos de opinião publicados em um jornal de grande circulação, em que articulistas argumentavam contra ou a favor de um determinado tema. A distinção foi possível empregando várias métricas, incluindo graus, coeficiente de agrupamento, caminhos mínimos, eficiência global, centralidade de proximidade e acessibilidade.

A título de exemplo, considere duas resenhas curtas de um filme, uma positiva e outra negativa:

- Resenha positiva: O filme é ótimo. As atuações são ótimas e o roteiro também é ótimo.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

- Resenha negativa: O filme é lento. As atuações são previsíveis e o roteiro é muito fraco.

A rede de coocorrência no primeiro caso, da resenha positiva, seria pequena e altamente conectada. O termo *ótimo* aparece em todas as sentenças e se conecta diretamente a *filme*, *atuação* e roteiro, formando um nó central (*hub*) com grau elevado. Essa estrutura indica forte polarização positiva e alta coesão semântica, o que facilita a detecção automática do sentimento positivo. No caso da resenha negativa, o resultado seria uma rede mais esparsa, com termos como *lento*, *previsível* e *fraco* distribuídos entre diferentes aspectos do filme. Não há um único termo dominante que conecte todos os elementos, resultando em uma rede com vários nós periféricos e menor densidade. Essa dispersão semântica reflete uma crítica fragmentada e facilita a identificação do sentimento negativo. Ao comparar essas redes, métricas como grau médio, coeficiente de aglomeração, diâmetro da rede e centralidade de termos afetivos podem ser utilizadas para inferir automaticamente o sentimento predominante. Redes densas com *hubs* positivos tendem a indicar avaliações favoráveis, enquanto redes dispersas com múltiplos termos negativos sugerem críticas desfavoráveis.

O estudo (Amancio et al., 2011b) mostrou que essa diferença na estrutura da rede, capturada por métricas como o coeficiente de agrupamento e os caminhos mínimos, é suficiente para classificar automaticamente a polaridade do texto. O conjunto de dados multidimensional foi mapeado em um espaço bidimensional por meio da análise de componentes principais. A distinção foi quantificada utilizando algoritmos de aprendizado de máquina, com acerto de 84,4% na discriminação automática das opiniões negativas.

Há muitas aplicações de PLN em saúde, como discutido nos Capítulos 9 e 10 deste livro. O diagnóstico médico é essencialmente uma tarefa de classificação, razão pela qual atualmente se empregam diferentes algoritmos de aprendizado de máquina para a análise de imagens médicas e exames clínicos. Esse tipo de aplicação pode também ser feito com texto, como foi o caso da análise de textos transcritos a partir de falas de pacientes e voluntários para diagnosticar o comprometimento cognitivo leve (CCL), que normalmente antecede os sintomas de demências (Santos et al., 2017b). Como os textos eram curtos, as redes complexas foram enriquecidas utilizando-se de *embeddings* de palavras. Ao enriquecer os nós (palavras) com *embeddings*, o modelo consegue capturar nuances semânticas que, combinadas com as métricas da rede, ajudam o classificador a detectar o comprometimento cognitivo. A maior acurácia na classificação binária do CCL foi obtida com o algoritmo Máquina de Vetores de Suporte (*Support Vector Machine*)(Santos et al., 2017b).

18.5 Exemplo de aplicação: Panorama de PLN em língua portuguesa

Estabelecer um panorama para qualquer área de pesquisa exige hoje um grande esforço, em vista do enorme número de publicações. Uma estratégia introduzida há alguns anos (Silva et al., 2016) para obter tais panoramas consiste em combinar redes complexas com PLN, partindo-se de redes de citações ou de similaridade formadas com artigos extraídos de buscas em repositórios como a *Web of Science* ou o *OpenAlex*. Empregando algoritmos de identificação de comunidades, determinam-se as comunidades (agrupamentos) correspondentes a tópicos ou subtópicos da área sob estudo. A análise da rede, tanto em termos de seu formato e topologia quanto do conteúdo das comunidades, permite obter uma visão geral da área. Nesta seção, descrevemos os resultados de empregar essa estratégia para a

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

área de PLN em português. A metodologia é detalhada nas subseções 18.5.1 e 18.5.2, e os resultados são apresentados na subseção 18.5.3.

18.5.1 Redes de citações e similaridade

Uma rede de citações pode ser obtida a partir de um *corpus* de artigos, estabelecendo uma conexão entre dois artigos do *corpus* (representados como nós da rede) quando existe uma citação de um para o outro. Essa conexão é, em princípio, direcionada, i.e., sai do nó que representa o artigo que cita para o nó que representa o artigo citado. Entretanto, há situações em que se pode optar por ignorar o sentido das arestas, caso o objetivo seja simplesmente representar um relacionamento desse tipo entre os dois artigos.

Se o *corpus* é representativo de uma área de conhecimento, ou de um tópico de pesquisa, a rede de citações permite capturar informações sobre a estrutura do conhecimento nessa área ou tópico. É possível identificar artigos altamente influentes (muito citados), grupos temáticos, artigos que atuam como pontes entre diferentes grupos temáticos, e como ideias se propagaram ao longo do tempo. Artigos centrais muitas vezes são precursores de novas subáreas ou pontos de referência para a comunidade de pesquisa representada na rede. Já artigos que ligam grupos temáticos podem indicar trabalhos com alto grau de interdisciplinaridade, pelo potencial de aproximar comunidades de pesquisa que antes não dialogavam entre si. Esse tipo de análise permite responder perguntas como: quais artigos deram origem a um novo tema de pesquisa? Que áreas e subáreas estão crescendo rapidamente em número de citações? Existe uma fragmentação em comunidades?

As redes de citações evidenciam os relacionamentos entre os artigos, mas não capturam necessariamente o seu conteúdo semântico. Dois artigos podem abordar temas próximos ou relacionados sem que haja uma relação de citação entre eles. Para explorar relações de natureza semântica, pode-se recorrer às redes de similaridade textual. Nessa abordagem, uma conexão entre dois artigos representados como nós da rede é estabelecida quando seus conteúdos são considerados semanticamente próximos, ou seja, foi detectada alguma similaridade de conteúdo. Para quantificar essa proximidade semântica, utiliza-se uma representação vetorial conhecida como *embedding* (Pennington et al., 2014). Para mais informações sobre similaridade textual e *embeddings* sugerimos ao leitor consultar o Capítulo Modelos de linguagem deste livro.

Existem diversos modelos para gerar *embeddings*, que podem vetorizar desde palavras isoladas até textos completos. Na abordagem apresentada na Seção 18.5.3, cada resumo é representado por um vetor numérico usando o modelo *SentenceTransformer*, que pode mapear textos inteiros em um espaço multidimensional. A proximidade semântica dos textos é determinada pelo cálculo de similaridade do cosseno entre todos os pares de vetores. A expressão matemática para essa similaridade é dada por:

$$S_{cos}(A, B) = \frac{A \cdot B}{||A||\ ||B||}$$

onde $A \cdot B$ é o produto escalar entre os vetores A e B, e $||A||$ e $||B||$ indicam as respectivas normas dos vetores. Esse produto escalar varia entre -1 e 1; no entanto, no contexto de *embeddings* textuais, considera-se apenas o módulo. Assim, uma similaridade próxima de 0 indica que os dois textos são totalmente distintos, enquanto uma similaridade próxima a 1 revela que os textos são semanticamente próximos.

Para a construção da rede de similaridade nessa abordagem, considera-se ainda um limiar de similaridade, que estabelece o valor mínimo necessário para que uma conexão seja

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

adicionada à rede. Ou seja, utilizando um limiar de 70% ($S_{cos}(A, B) = 0.7$), apenas os pares de artigos cuja similaridade exceder este valor estarão conectados por uma aresta. Essa etapa evita que se crie uma rede excessivamente densa e garante que o modelo gerado ofereça uma visão estruturada da proximidade temática entre os artigos, servindo de base para a identificação de padrões na literatura. Artigos que compartilham temas centrais tendem a compor agrupamentos temáticos, revelando subáreas de estudo ou tópicos emergentes. Conexões entre agrupamentos podem indicar tópicos interdisciplinares ou tendências de convergência de temas de pesquisa. Essas redes permitem identificar artigos que abordam temas semelhantes, além de evidenciar padrões ocultos não evidenciados nas relações de citação. As redes de similaridade são então complementares às redes de citações. São particularmente úteis para pequenos grupos de artigos, como em uma área emergente na qual ainda não se identificam trabalhos centrais com alto número de citações.

18.5.2 Identificação de tópicos em redes de pesquisa

O método introduzido por Silva et al. (Silva et al., 2016) estabelece uma estratégia para identificar os principais temas (tópicos) associados a uma área de pesquisa, dada uma rede criada a partir de um *corpus* textual de artigos científicos representativo da área. A estratégia associa um algoritmo de detecção de comunidades na rede com a identificação dos tópicos abordados pelos artigos pertencentes às diferentes comunidades. Assim, comunidades são caracterizadas por conjuntos de termos (palavras-chave - unigramas e bigramas) extraídos do *corpus* a partir da quantificação da importância semântica dos termos, no contexto da sua ocorrência naquela comunidade. Em particular, os termos elencados são extraídos do título e resumo dos artigos que compõem a rede (seja ela de citação ou de similaridade), cujos conteúdos contêm informações suficientes para uma boa caracterização dos termos mais representativos.

A importância dos termos na rede é quantificada por meio de dois índices de frequência relativa. Inicialmente, é determinada a frequência de um termo dentro de sua própria comunidade e, em seguida, sua frequência no restante da rede. Por exemplo, para determinar a frequência de um termo (w) em uma comunidade (α), conta-se o total de artigos $n_\alpha(w)$ dessa comunidade que contêm o termo (w). Com isso, a frequência do termo relativa dentro da comunidade $F_\alpha^{in}(w)$ é:

$$F_\alpha^{in}(w) = \frac{n_\alpha(w)}{|\alpha|}$$

em que $|\alpha|$ é o número de artigos pertencentes à comunidade (α). Analogamente, definimos a frequência relativa fora da comunidade (α) como:

$$F_\alpha^{out}(w) = \sum_{\gamma \neq \alpha} \frac{n_\gamma(w)}{N - |\alpha|}$$

Neste caso, (γ) representa uma comunidade qualquer, diferente de (α), e N é o número total de artigos da rede. Ou seja, $F_\alpha^{out}(w)$ se refere à frequência relativa do termo (w) fora da comunidade (α).

Com as relações de frequência interna e externa à comunidade (α) estabelecidas, a importância de um termo $I(w)$ é definida como a maior diferença entre $F_\alpha^{in}(w)$ e $F_\alpha^{out}(w)$, ou seja:

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

$$I(w) = max_{\alpha}[F_{\alpha}^{in}(w) - F_{\alpha}^{out}(w)]$$

Os termos-chave são agrupados com base na distância topológica média entre os artigos que os contêm, ou seja, é utilizado o comprimento médio do caminho mais curto $\langle l \rangle_{uv}$ entre um par de termos-chave (u, v). O algoritmo inicialmente determina a menor distância (l_{ij}) entre cada par de artigos (i, j) na rede e então, para cada par de termos-chave (u, v), calcula a média dos menores caminhos entre todos os pares de resumos (A_i, A_j) dos artigos cujos termos-chave u e v estão presentes. A representação matemática desses conceitos é dada na equação abaixo:

$$\langle l \rangle_{uv} = \sum_{(u,v) \in (A_i \times A_j)} \frac{l_{ij}}{|(u, v) \in (A_i \times A_j)|}$$

Em termos gerais, significa que conjuntos de termos-chave são agrupados segundo a distância topológica média entre eles. Ressalte-se que os unigramas e bigramas são classificados de acordo com a mesma medida. Quando um bigrama possui alto $I(w)$, há grande probabilidade de que seus unigramas compostos também tenham alto $I(w)$, gerando redundância. Esse problema é resolvido com a remoção dos unigramas do conjunto de palavras-chave que fizerem parte de qualquer outro bigrama no conjunto. Essa prática proporciona uma maior prioridade a termos-chave (bigramas) mais específicos. Ao final, o método gera uma rede complexa composta por conjuntos densamente conectados (agrupamentos), que são caracterizados por seus termos-chave com maior índice de importância.

18.5.3 Pesquisa em PLN para o português

O desafio inicial consistiu em localizar as contribuições em PLN em português na literatura. Buscamos identificar artigos científicos publicados sobre o PLN para a língua portuguesa, escritos nas línguas portuguesa ou inglesa. Bases de dados como *Web of Science* e *OpenAlex* incluem textos em português, mas as estratégias de busca podem ser falhas, ou pelo menos mais difíceis de otimizar do que para textos em inglês. Há também o problema da menor visibilidade dos fóruns de publicação. Artigos em fóruns relevantes, como o PROPOR, podem não estar indexados. A baixa visibilidade representa, obviamente, uma limitação para o estudo proposto, assim como para a valorização das contribuições científicas de países lusófonos em PLN. Diante dessas limitações, não há como garantir que o panorama apresentado a seguir seja totalmente fiel ao panorama da literatura produzida para o PLN em português, ao contrário do que se pode afirmar para estudos anteriores em outros tópicos (Oliveira; Oliveira, 2025; Oliveira et al., 2025). Nestes últimos trabalhos não havia dúvidas de que os artigos mais relevantes para as respectivas áreas foram recuperados. Críticas sobre o panorama apresentado serão muito bem-vindas para garantir estudos mais precisos no futuro e para que a comunidade do PLN em português busque estratégias para ampliar a visibilidade dos trabalhos na área.

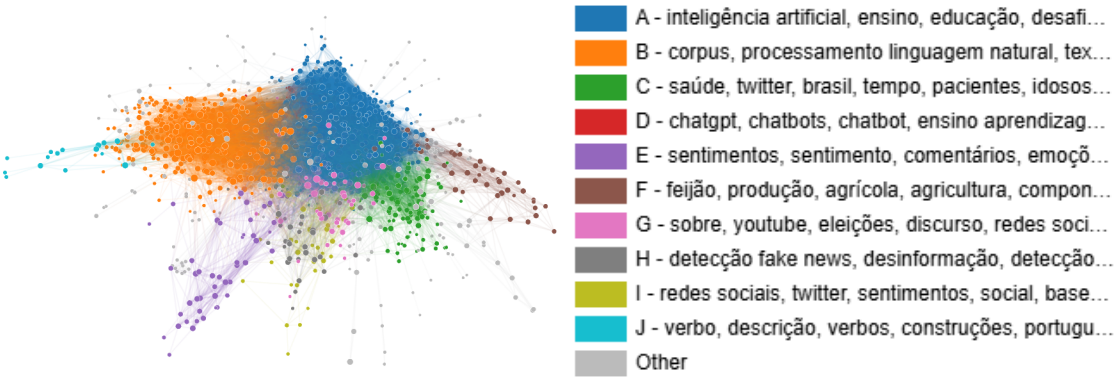
O número de artigos que abordam PLN para a língua portuguesa encontrados em diferentes buscas, na *Web of Science* e *OpenAlex*, foi sempre inferior a 5.000. Esse número reduzido inviabiliza a construção de redes de citações, que seriam esparsamente conectadas. A alternativa foi construir redes de similaridade, como descrito na Seção 18.5.1, para o que também houve algumas dificuldades. Usado em estudos anteriores, o SciBERT (Beltagy et al., 2019) foi treinado com artigos e textos científicos em inglês e mostrou-se inadequado para textos em português. Isso motivou a adoção do *SentenceTransformer*

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

multilíngue, um modelo BERT capaz de lidar com outras línguas. Apesar de não ser especializado em artigos científicos, foi a opção com melhores resultados. A busca no *OpenAlex* com os termos **"PLN" OR "NLP" OR "processamento de linguagem natural" OR "natural language processing"** retornou 3.303 artigos, dos quais 1.766 compõem a maior componente conectada da rede de similaridade, construída considerando um limiar de similaridade igual a 70%. A maior componente da rede resultante é exibida na Figura 18.6.

Figura 18.6: Rede de similaridade para publicações sobre Processamento de Linguagem Natural em português, mostrando as principais comunidades temáticas e suas inter-relações.



Link para visualização interativa da rede: Rede de Similaridade

O Quadro 18.1 lista as 10 maiores comunidades identificadas na rede da Figura 18.6, incluindo o número de artigos por comunidade e uma descrição gerada pelo ChatGPT a partir dos termos-chave associados a cada comunidade. Estas descrições foram geradas com um comando "*Faça uma descrição dos tópicos abordados pelo cluster X, com base nas palavras-chave que o caracterizam*". A maior comunidade (A, em azul, com 748 artigos) é dedicada à inteligência artificial, principalmente ao seu uso em diferentes domínios, como o jurídico e o educacional. Uma inspeção (sem valor estatístico) nos títulos dos artigos indica que muitos deles se referem a textos, alguns de fato envolvendo PLN, embora outros talvez não. Mencione-se aqui uma possível dificuldade de identificação de comunidades devido às ambiguidades de nomenclatura. O termo inteligência artificial (IA) é usado em contextos diversos, nem sempre com precisão na terminologia. É comum o uso do termo "aprendizado de máquina" de forma intercambiável com IA, como se fossem sinônimos, ignorando que aprendizado de máquina é subárea de IA. PLN é outra subárea de IA, mas suas comunidades de pesquisadores nem sempre se identificam com grupos de IA que atuam em outras subáreas. Parece, portanto, inevitável que a busca por trabalhos em PLN encontre artigos de IA, mas que não necessariamente abordam PLN. De qualquer forma, é surpreendente que quase metade dos artigos da rede estejam na comunidade A, em que o foco está em usos de IA - e não em PLN.

Já a segunda maior comunidade, B, em laranja, com 377 artigos, é de fato de artigos em PLN. Ela inclui artigos relacionados à construção e organização de *corpora*, classificação automatizada de textos e uso de linguística computacional para tarefas de análise e categorização. Artigos relacionados a aplicações em saúde aparecem na terceira comunidade, C, em verde, com 110 itens. Essa comunidade inclui trabalhos em temas variados, como análise sistemática da literatura em saúde e análise de sentimentos ou opinião relacionada

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

à saúde. Parte desses artigos pode abordar tópicos de PLN, mas provavelmente há outros que envolvem algum tipo de processamento de texto, não necessariamente PLN. As demais comunidades são consideravelmente menores, com algumas dezenas de artigos. São, em geral, bastante específicas e abordam temas em PLN, com uma possível exceção. Trata-se da comunidade F (em marrom, com 46 artigos), composta por artigos sobre agricultura. Dos poucos artigos efetivamente de PLN que podem ser identificados, destacam-se os de análise de literatura no tópico. A comunidade D (em vermelho, 53 artigos) inclui artigos sobre assistentes de conversação (*chatbots*), principalmente o ChatGPT. São, portanto, artigos recentes. A comunidade E (em roxo, com 47 artigos) é dedicada à análise de sentimentos. Eleições e política são os temas principais na comunidade G (rosa, com 41 artigos), com trabalhos de PLN principalmente dedicados à análise de texto em redes sociais. Outra comunidade típica de PLN é a H, em cinza-escuro, com 38 artigos que versam sobre detecção de notícias falsas ou desinformação. A comunidade I, em verde-claro, com 37 artigos, é vizinha à comunidade E na rede da Figura 18.6, pois também trata de análise de sentimentos, com foco no Twitter. A última comunidade, J, em verde-musgo, com 18 artigos, inclui estudos linguísticos, como análise semântico-discursiva utilizando PLN. Alguns poucos artigos que não foram categorizados como pertencendo a alguma comunidade aparecem em cinza-claro na Figura 18.6.

O formato da rede pode ser inspecionado cuidadosamente a partir do modelo tridimensional interativo acessado no link informado acima e também merece comentários. A rede é dominada por duas comunidades, A e B, que são bastante conectadas entre si. A comunidade J, de estudos linguísticos com e para PLN, é conectada apenas com a comunidade B, que engloba a maioria dos estudos em PLN. Pode-se afirmar, então, que essas duas são comunidades de pesquisadores de fato atuantes em PLN. O argumento se justifica porque o panorama que se depreende da rede da Figura 18.6 talvez permita distinguir trabalhos de grupos de pesquisadores que efetivamente atuam em PLN de trabalhos de pesquisadores não necessariamente identificados com essa área, mas que passaram a atuar em processamento de texto como consequência dos avanços tecnológicos em aprendizado de máquina e, mais recentemente, a IA gerativa. O que se observa é que as demais comunidades são sempre conectadas à comunidade A; seus artigos têm maior similaridade com aqueles classificados mais genericamente como IA do que com PLN. Apenas a comunidade E (em roxo), relacionada à análise de sentimentos, está próxima da comunidade B, embora ainda se mostre mais conectada à comunidade A.

Quadro 18.1: Comunidades na rede de PLN.

Comunidade	**Nós**	**Palavras-chaves**	**Descrição**
A	748	inteligência artificial, ensino, educação, desafios, aprendizado, tecnologias, científica, máquina, papel, ia	Aplicações de inteligência artificial no contexto educacional; tecnologias e estratégias para ensino, aprendizagem e desenvolvimento científico.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

B	377	*corpus*, processamento linguagem natural, textos, automática, classificação, linguística, português brasileiro, textual, língua, modelos	Processamento de linguagem natural e linguística computacional; análise automática de textos, classificação textual, *corpora* e modelos linguísticos em português brasileiro.
C	110	saúde, twitter, brasil, tempo, pacientes, idosos, in, pandemia covid-19, tratamento, revisão	Uso de redes sociais para monitoramento de saúde pública; análise da pandemia de COVID-19, impacto em pacientes e idosos, e revisão de tratamentos no Brasil.
D	53	chatgpt, chatbots, chatbot, ensino-aprendizagem, sobre, revisão, ferramenta, auxiliar, assistente, sistemática	Desenvolvimento e aplicação de chatbots e assistentes virtuais; ferramentas para ensino e aprendizagem e revisões sistemáticas de tecnologias educacionais.
E	47	sentimentos, sentimento, comentários, emoções, português, métricas, mídias sociais, mineração, rede, mapeamento	Análise de sentimentos em mídias sociais; mineração de dados textuais, métricas emocionais e mapeamento de interações em português.
F	46	feijão, produção, agrícola, agricultura, componentes, rendimento, produtividade, cultura soja, seleção, gesso	Produção e produtividade agrícola; técnicas de cultivo, seleção de componentes e otimização de rendimento em culturas como feijão e soja.
G	41	sobre, youtube, eleições, discurso, redes sociais, deputados, discursos, ódio, sentimento, twitter	Análise de discursos e sentimentos em redes sociais durante eleições; estudo de conteúdo em YouTube e Twitter, incluindo discursos de deputados e manifestações de ódio.
H	38	detecção fake news, desinformação, detecção automática, notícias falsas, jornalismo, utilizando, produção, covid-19 brasil, language, avaliação	Detecção automática de fake news e desinformação; análise de notícias falsas e produção jornalística durante a pandemia de COVID-19 no Brasil.
I	37	redes sociais, twitter, sentimentos, social, baseado, tweet sobre, eventos, classificação, identificação, caso	Análise de sentimentos e classificação de eventos em redes sociais; identificação de padrões sociais e comportamentais no Twitter.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

J	18	verbo, descrição, verbos, construções, português, linguagem natural, processamento, base, verbal, semântico-discursiva	Estudos linguísticos em português; análise de verbos, construções verbais e descrições semântico-discursivas utilizando processamento de linguagem natural.

Uma análise manual permite identificar os tópicos mais relevantes nas diferentes comunidades, mas é inviável para um estudo pormenorizado, em vista do número de artigos da rede (mais de 1.700). É possível, entretanto, buscar auxílio das ferramentas com modelos de língua de larga escala (LLMs, do inglês *large-language models*), que podem processar texto a uma velocidade muito superior à de um ser humano. Em setembro de 2025, não há garantias de que tais ferramentas forneçam análises ou resumos fiéis aos artigos sob estudo, principalmente porque a janela de contexto disponibilizada para um usuário pode não ser suficiente para cobrir todo o *corpus*. Há, entretanto, evidências empíricas de que essas ferramentas conseguem fornecer uma análise superficial adequada. Com essa expectativa, solicitamos ao ChatGPT uma expansão na descrição das comunidades da Figura 18.6.

Com os termos-chave de cada comunidade e suas respectivas descrições iniciais fornecidas pelo ChatGPT, foi solicitado um levantamento mais detalhado dos tópicos apontados por ele. Para isso, foi utilizado o seguinte comando "*Faça uma expansão para a descrição do cluster X, se aprofundando nos tópicos apresentados por ele de acordo com suas palavras-chave e a busca central utilizada para produzir a rede*". Os textos produzidos, individualmente para cada comunidade, apresentaram um aprofundamento dos tópicos descritos na coluna "Descrição" do Quadro 18.1. Os textos foram editados, por concisão, e suas análises e conclusões estão descritas a seguir.

Já comentamos a partir da análise manual que a maior comunidade, A, foca na aplicação de técnicas de IA e PLN na educação, revelando um movimento crescente de integração de tecnologias inteligentes em processos de ensino e aprendizagem, com destaque para sistemas de tutoria inteligente, plataformas adaptativas, ferramentas de correção automática de textos e *chatbots* educacionais. Os trabalhos também abordam de forma crítica os desafios pedagógicos e técnicos que acompanham essa transformação. Entre eles estão a dificuldade de adaptar modelos de PLN para diferentes contextos culturais, a necessidade de preparar professores para integrar essas ferramentas em suas práticas e os riscos de dependência tecnológica. Além disso, surgem preocupações sobre vieses algorítmicos, transparência na tomada de decisão das máquinas e privacidade dos dados estudantis, questões que exigem debates éticos e regulatórios. A literatura coberta nessa comunidade não se restringe às aplicações imediatas, mas também reflete sobre o impacto estrutural da adoção da IA na educação. Muitos artigos discutem como essas tecnologias estão redefinindo o papel do professor, não como substituto, mas como mediador em um ecossistema de aprendizagem híbrido, onde humanos e máquinas colaboram. Ao mesmo tempo, destacam-se análises empíricas sobre a eficácia dessas ferramentas na melhoria do desempenho acadêmico, no engajamento dos alunos e na inclusão de grupos tradicionalmente marginalizados no acesso ao ensino.

Como já mencionado, a comunidade B contém literatura fundamental de PLN, concentrando-se nos fundamentos linguísticos e computacionais com ênfase em *corpus*, classificação automática e análise textual. Os estudos desse grupo envolvem a construção, curadoria e utilização de *corpora*. Fica evidenciado o papel central dos *corpora* para

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

treinar, avaliar e comparar algoritmos de PLN em diferentes tarefas, como análise de sentimentos, categorização de documentos e reconhecimento de entidades nomeadas. São abordados também os desafios metodológicos relacionados à anotação, padronização e representatividade dos dados (questões abordadas no Capítulo Conjunto de dados, *dataset* e *corpus* deste livro), uma vez que a qualidade e a diversidade dos *corpus* impactam o desempenho dos modelos. Nesse sentido, surgem discussões sobre métodos automáticos e semiautomáticos de anotação, bem como sobre a integração de abordagens linguísticas tradicionais com técnicas de aprendizado de máquina. Além disso, chama-se a atenção para as particularidades do português, idioma menos contemplado na pesquisa internacional de PLN do que o inglês. Esse foco evidencia um esforço da comunidade científica em reduzir desigualdades linguísticas e promover avanços na adaptação de modelos globais a contextos locais. Os trabalhos também revelam preocupações quanto à dimensão sociotécnica do PLN, discutindo o risco de vieses reproduzidos por *corpora* desbalanceados ou pouco representativos. A literatura aponta ainda para a necessidade de ferramentas linguísticas acessíveis, que possam servir tanto à pesquisa acadêmica quanto a aplicações práticas em educação, saúde e setores industriais.

A comunidade C explora a aplicação do PLN em saúde, com ênfase em redes sociais como o Twitter e em contextos marcados pela pandemia da COVID-19. Os trabalhos desse grupo investigam como técnicas de PLN podem ser empregadas para monitorar, analisar e prever fenômenos relacionados à saúde pública, desde a identificação de sintomas relatados por pacientes até a avaliação de percepções sociais sobre políticas sanitárias. Essa vertente evidencia o papel estratégico do PLN na compreensão de grandes volumes de dados não estruturados, capazes de complementar fontes tradicionais de informação médica e epidemiológica. Os estudos também abordam a utilização do Twitter como fonte de dados em tempo real, permitindo a detecção de surtos, o acompanhamento de discussões sobre vacinas e tratamentos e a análise das emoções e preocupações expressas pela população. O caráter dinâmico e global das redes sociais potencializa o PLN como ferramenta de vigilância em saúde, mas também levanta desafios metodológicos, como a necessidade de filtrar ruídos, lidar com ambiguidade textual e diferenciar informações confiáveis de conteúdos desinformativos. Além disso, a pandemia de COVID-19 aparece em muitos artigos, em um período em que se empregou PLN para analisar notícias relacionadas e comentários em redes sociais. Outro ponto relevante é o impacto da análise automatizada na prática clínica e na formulação de políticas de saúde. Pesquisas avaliam o potencial de modelos de PLN para apoiar diagnósticos, identificar grupos de risco, prever demandas hospitalares e oferecer suporte a revisões sistemáticas de literatura biomédica. Paralelamente, surgem reflexões sobre os limites éticos e legais do uso de dados provenientes de redes sociais, sobretudo em relação à privacidade e ao consentimento dos usuários.

Os *chatbots* são o assunto central da comunidade D, principalmente o ChatGPT. A literatura agrupa estudos sobre o desenvolvimento de agentes conversacionais voltados para suporte pedagógico — desde *chatbots* que oferecem retorno imediato em tarefas escritas até assistentes que guiam o estudante passo a passo em resoluções de problemas. Muitos trabalhos relatam experimentos com plataformas educacionais integradas a *chatbots*, avaliando como esses agentes podem personalizar explicações, fornecer exercícios adaptativos e atuar como tutores suplementares em disciplinas diversas. Alguns trabalhos nessa comunidade combinam revisões sistemáticas com estudos empíricos para testar protótipos em sala de aula ou em ambientes controlados. As avaliações frequentemente medem ganhos de aprendizagem, engajamento, aceitabilidade por parte de alunos e professores, além de métricas de usabilidade e qualidade da conversação (coerência, relevância das respostas,

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

taxa de intervenção humana). Há também trabalhos técnicos que abordam arquiteturas (agentes baseados em regras vs. modelos gerativos), adaptação para domínios educacionais, técnicas de *prompting* e estratégias para integrar *chatbots* a sistemas de gestão de aprendizagem (para saber mais sobre o tema, sugerimos consultar o Capítulo ChatGPT, MariTalk e outros agentes de conversação deste livro).

A análise de sentimentos é o tópico dominante nas comunidades E e I, com destaque para dados provenientes de mídias sociais e comentários online em português. Os trabalhos desses grupos exploram como técnicas de PLN podem ser aplicadas para identificar, classificar e quantificar sentimentos positivos, negativos ou neutros, bem como emoções mais complexas, em grandes volumes de dados textuais. Essa linha de pesquisa é relevante para entender o comportamento dos usuários, monitorar percepções públicas sobre produtos, serviços ou eventos, fornecendo subsídios para áreas como publicidade, comunicação e políticas públicas. Os estudos evidenciam o desenvolvimento e a aplicação de métricas específicas para a análise de sentimentos, incluindo métodos baseados em léxicos, aprendizado de máquina e modelos de linguagem pré-treinados. Além disso, os trabalhos abordam questões metodológicas e práticas, incluindo desafios relacionados à ambiguidade textual, ironia, sarcasmo e diversidade linguística no português brasileiro.

A comunidade F pareceu ser a menos aderente ao PLN na análise manual, pois os títulos dos artigos normalmente não indicavam uso de PLN. Entretanto, segundo a descrição do *ChatGPT*, o PLN está presente na análise textual em contextos agrícolas, com ênfase em culturas como feijão e soja. Ainda segundo o ChatGPT, a comunidade contém literatura na interface entre tecnologia e prática agrícola, demonstrando como a análise de textos e a extração de conhecimento podem contribuir para avanços concretos na produção de alimentos. A comunidade G é dedicada ao estudo de eleições e políticas, concentrando-se na análise de discursos políticos, eventos eleitorais e interações em redes sociais, principalmente no Twitter e no *YouTube*. Há artigos explorando técnicas de PLN para identificar padrões de comunicação, detectar discursos de ódio, analisar sentimentos e compreender o impacto das mensagens veiculadas por políticos, influenciadores e cidadãos em contextos eleitorais. Essa abordagem permite mapear tanto a propagação de opiniões quanto a dinâmica de engajamento do público em diferentes plataformas digitais. Além disso, investigam estratégias de comunicação política, incluindo a construção de narrativas, o uso de linguagem persuasiva e a disseminação de mensagens polarizadoras.

Artigos sobre detecção de notícias falsas e desinformação predominam na comunidade H, em que há ênfase no uso de PLN para identificar padrões linguísticos, sinais de veracidade e estratégias de manipulação em notícias e conteúdos compartilhados em redes sociais. O objetivo central é desenvolver métodos que permitam avaliar a confiabilidade das informações, detectando conteúdos enganosos e prevenindo a disseminação de desinformação em larga escala. Os estudos nessa comunidade abordam múltiplas abordagens metodológicas, incluindo modelos baseados em aprendizado de máquina supervisionado e não supervisionado, análise de métricas linguísticas, redes semânticas e técnicas de classificação de textos. Além disso, a literatura enfatiza a importância de *corpus* específico para o português brasileiro, permitindo que os algoritmos lidem com variações linguísticas, regionalismos, gírias e construções sintáticas típicas do idioma. Muitas pesquisas também relacionam a detecção de notícias falsas sobre a pandemia de COVID-19, analisando a propagação de informações incorretas e seu impacto sobre comportamentos de saúde e percepção pública.

A pequena comunidade J concentra-se no estudo de verbos, construções verbais e aspectos semântico-discursivos do português. Há ênfase na aplicação de técnicas de PLN para

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

analisar padrões gramaticais, categorias verbais e relações semânticas, contribuindo para o desenvolvimento de modelos linguísticos mais precisos e adaptados às especificidades do idioma. Também aparecem pesquisas associadas à anotação e classificação de verbos, construção de bases de dados linguísticos e análise de estruturas sintáticas e semânticas em textos. Esses estudos permitem compreender como os verbos e suas construções funcionam em diferentes contextos comunicativos, servindo para tarefas de PLN como análise sintática, geração de linguagem natural e tradução automática.

Para fins de estudo e transparência, foi disponibilizado um arquivo complementar contendo uma lista dos artigos que compõem as dez maiores comunidades da rede de similaridade. O arquivo está organizado em abas que correspondem às comunidades identificadas; cada aba apresenta uma tabela com os títulos e respectivos links de acesso aos artigos no repositório de origem.

Link de acesso: Arquivo Complementar

18.6 Considerações finais

O sucesso do uso de redes complexas para o PLN parece ser indiscutível, a julgar pela gama de aplicações comprovadas. A expectativa foi confirmada de que é possível capturar características semânticas, sintáticas e mesmo pragmáticas ao representar um texto como uma rede. A literatura da área, entretanto, não é vasta e o número de artigos deve ser da ordem de algumas poucas centenas. Essa produção tem contribuição relevante de autores brasileiros, que estiveram entre os primeiros a aplicar conceitos e métodos de redes complexas em PLN. Embora os trabalhos dos autores brasileiros sejam majoritariamente focados no processamento de textos em inglês, há pelo menos algumas dezenas de artigos que abordam PLN para o português. Neste capítulo, apresentamos um breve resumo de algumas dessas contribuições, em sumarização, desambiguação lexical, tradução automática, análise de sentimentos e diagnóstico médico baseado em texto.

As perspectivas quanto aos usos de redes complexas para aplicações em PLN foram abordadas em um comentário (Amancio, 2014) que enfatizou a necessidade de maior rigor na interpretação dos fatores que contribuem para o sucesso dos modelos de rede. Por outro lado, é interessante fazer uma comparação ou analogia com modelos de língua de larga escala (LLMs), uma vez que muitas das tarefas de PLN realizadas com redes complexas podem atualmente ser feitas utilizando LLMs. Pode-se especular que as informações advindas das características estruturais e dinâmicas das redes representando textos sejam, de alguma maneira, inferidas por meio dos mecanismos de *embedding* e de atenção dos LLMs. A grande capacidade de processamento e a janela de contexto dos LLMs tendem a resultar em desempenho significativamente superior nas tarefas de PLN, em comparação ao uso de redes. De fato, em pelo menos um trabalho ainda não publicado de pesquisadores do *Núcleo Interinstitucional de Linguística Computacional* (NILC) mostrou-se que o uso de modelos de língua gera resultados consideravelmente melhores. Referimo-nos à capacidade de sistemas inteligentes de estimar a pontuação de redações de ENEM, em que modelos usando BERT ("transformer") mostraram-se mais eficientes do que o uso de métricas de redes complexas. A comparação não é necessariamente justa, pois o desempenho das redes não foi otimizado variando-se as métricas exaustivamente.

De todo modo, acreditamos que a maioria das tarefas de PLN será mais bem executada com LLMs, o que poderia indicar que não há muito futuro para o uso de redes complexas em PLN. Entretanto, deve-se considerar que as redes continuarão sendo adequadas para verificar modelos e estudar a importância de aspectos linguísticos nas diferentes tarefas.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

As informações estruturais capturadas pelos modelos podem contribuir para a melhor compreensão dos processos linguísticos envolvidos, ou seja, garantir mais interpretabilidade para os resultados do processamento de textos. Redes também serão importantes para arquiteturas baseadas em grafos e como fontes de informação em tarefas de visualização exploratória de itens da literatura científica, como no panorama apresentado neste capítulo para o PLN em português.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Referências

ABDIN, M. et al. **Phi-3 Technical Report: A Highly Capable Language Model Locally on Your Phone**., 2024. Disponível em: <https://arxiv.org/abs/2404.14219>

ABONIZIO, H. Q. et al. Language-Independent Fake News Detection: English, Portuguese, and Spanish Mutual Features. **Future Internet**, v. 12, n. 5, 2020.

ABREU, S. C. DE; VIEIRA, R. Relp: Portuguese open relation extraction. **KO KNOWLEDGE ORGANIZATION**, v. 44, n. 3, p. 163–177, 2017.

ACKEL, A. Abordagens digitais para estudos de Paleografia: desafios, atualidade, desdobramentos. **LaborHistórico**, v. 7, n. 3, p. 100–120, 2021.

AEJAS, B.; BELHI, A.; BOURAS, A. **Smart Contracts Auto-generation for Supply Chain Contexts**. (F. Noël et al., Eds.)Product Lifecycle Management. PLM in Transition Times: The Place of Humans and Transformative Technologies. **Anais**...Cham: Springer Nature Switzerland, 2023.

AGGARWAL, C. C. **Machine Learning for Text**. [s.l.] Springer, 2018.

AGICHTEIN, E.; GRAVANO, L. **Snowball: Extracting relations from large plain-text collections**. Proceedings of the fifth ACM conference on Digital libraries. **Anais**...2000.

AGNOLONI, T. et al. **Making Italian Parliamentary Records Machine-Actionable: the Construction of the ParlaMint-IT corpus**. Proceedings of the Workshop ParlaCLARIN III within the 13th Language Resources and Evaluation Conference. **Anais**...Marseille, France: European Language Resources Association, jun. 2022. Disponível em: <https://aclanthology.org/2022.parlaclarin-1.17>

AHA, D. W.; KIBLER, D.; ALBERT, M. K. Instance-based learning algorithms. **Machine Learning**, v. 6, n. 1, p. 37–66, 1 jan. 1991.

AI@META. Llama 3 Model Card. 2024.

AJAY, H. B.; TILLET, P.; PAGE, E. B. **Analysis of essays by computer (AEC-II)**. Storrs, CT: Univeristy of Connecticut, 1973.

AKBIK, A. et al. **Multilingual information extraction with PolyglotIE**. Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: System Demonstrations. **Anais**...2016. Disponível em: <https://aclanthology.org/C16-2056/>

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

ALBERTO, T. C.; LOCHTER, J. V.; ALMEIDA, T. A. Post or Block? Advances in Automatically Filtering Undesired Comments. **Journal of Intelligent and Robotic Systems**, v. 80, p. 245–259, 2015.

ALBUQUERQUE, G. et al. Applying event classification to reveal the Estado da Índia. **Proceedings of the International Conference on the Computational treatment of Portuguese, PROPOR**, a2024.

ALBUQUERQUE, H. et al. **UlyssesNERQ: Expanding Queries from Brazilian Portuguese Legislative Documents through Named Entity Recognition**. (P. Gamallo et al., Eds.)Proceedings of the 16th International Conference on Computational Processing of Portuguese - Vol. 1. **Anais**...Santiago de Compostela, Galicia/Spain: Association for Computational Lingustics, mar. b2024. Disponível em: <https://aclanthology.org/2024.propor-1.35>

ALBUQUERQUE, H. O. et al. **UlyssesNER-Br: A Corpus of Brazilian Legislative Documents for Named Entity Recognition**. (V. Pinheiro et al., Eds.)Computational Processing of the Portuguese Language. **Anais**...Cham: Springer International Publishing, 2022. Disponível em: <https://github.com/ulysses-camara/>

ALBUQUERQUE, H. O. et al. Named entity recognition: a survey for the portuguese language. **Procesamiento del Lenguaje Natural**, b2023.

ALBUQUERQUE, H. O. et al. **On the Assessment of Deep Learning Models for Named Entity Recognition of Brazilian Legal Documents**. (N. Moniz et al., Eds.)Progress in Artificial Intelligence. **Anais**...Cham: Springer Nature Switzerland, a2023.

ALBUQUERQUE, H. O. et al. **UlyssesLegalNER-Br: from Legislative to Legal, a comprehensive corpus of Brazilian legal documents for Named Entity Recognition**. 17th Computational Processing of the Portuguese Language (PROPOR 2026). **Anais**...2026.

ALEIXO, P.; PARDO, T. A. S. **Uma Ferramenta Semi-automática para Anotação de Córpus pela Teoria Discursiva Multidocumento CST**. [s.l.] Instituto de Ciências Matemáticas e de Computação, 2008.

ALIGULIYEV, R. M. et al. COSUM: Text summarization based on clustering and optimization. **Expert Systems**, v. 36, n. 1, p. e12340, 2019.

ALIKANIOTIS, D.; YANNAKOUDAKIS, H.; REI, M. **Automatic Text Scoring Using Neural Networks**. Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics. **Anais**...Association for Computational Linguistics, 2016.

ALLES, V. J. **Construção de um corpus para extrair entidades nomeadas do Diário Oficial da União utilizando aprendizado supervisionado**. mathesis—[s.l.] Master's thesis, Universidade Federal de Brasília, 2018.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

ALMEIDA, G. DE. **Translating the post-editor: an investigation of post-editing changes and correlations with professional experience across two Romance languages**. 2013. Disponível em: <https://api.semanticscholar.org/CorpusID:60255248>

ALMEIDA, P. G. R. Uma jornada para um Parlamento inteligente: Câmara dos Deputados do Brasil. **Red Información**, v. 24, 2021.

ALMEIDA, T. A. et al. Text normalization and semantic indexing to enhance Instant Messaging and SMS spam filtering. **Knowledge-Based Systems**, v. 108, p. 25–32, 2016.

ALMEIDA, T. A.; ALMEIDA, J.; YAMAKAMI, A. Spam filtering: how the dimensionality reduction affects the accuracy of Naive Bayes classifiers. **Journal of Internet Services and Applications**, v. 1, n. 3, p. 183–200, 2011.

ALMEIDA, T. A.; YAMAKAMI, A. Facing the spammers: A very effective approach to avoid junk e-mails. **Expert Systems with Applications**, v. 39, p. 6557–6561, 2012.

ALMOUZINI, S.; KHEMAKHEM, M.; ALAGEEL, A. Detecting Arabic Depressed Users from Twitter Data. **Procedia Computer Science**, v. 163, p. 257–265, 2019.

ALONSO, M. A. et al. Sentiment Analysis for Fake News Detection. **Electronics**, v. 10, n. 11, 2021.

AL-RFOU, R. et al. **Collective named entity disambiguation using graph-based ranking**. Proceedings of the 23rd International Conference on World Wide Web (WWW '14). **Anais**...ACM, 2014. Disponível em: <https://doi.org/10.1145/2566486.2568010>

ALUÍSIO, S. M. et al. Detecting Respiratory Insufficiency Via Voice Analysis: The SPIRA Project. Em: **Practical Machine Learning for Developing Countries on the Tenth International Conference on Learning Representations. Proceeding**. [s.l.] ICLR, 2022.

ALUÍSIO, S.; GASPERIN, C. **Fostering Digital Inclusion and Accessibility: The PorSimples project for Simplification of Portuguese Texts**. (T. Solorio, T. Pedersen, Eds.)Proceedings of the NAACL HLT 2010 Young Investigators Workshop on Computational Approaches to Languages of the Americas. **Anais**...Los Angeles, California: Association for Computational Linguistics, jun. 2010. Disponível em: <https://aclanthology.org/W10-1607>

ALVARES, R. V.; GARCIA, A. C. B.; FERRAZ, I. **STEMBR: A stemming algorithm for the Brazilian Portuguese language**. Portuguese conference on artificial intelligence. **Anais**...Springer, 2005.

AMANCIO, D. R. et al. Using metrics from complex networks to evaluate machine translation. **Physica A: Statistical Mechanics and its Applications**, v. 390, n. 1, p. 131–142, a2011.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

AMANCIO, D. R. et al. **Opinion Discrimination Using Complex Network Features**. Complex Networks: Second International Workshop on Complex Networks (CompleNet 2010). **Anais**...: Communications em Computer e Information Science.Springer, b2011. Disponível em: <https://doi.org/10.1007/978-3-642-25501-4_16>

AMANCIO, D. R. et al. Extractive summarization using complex networks and syntactic dependency. **Physica A: Statistical Mechanics and its Applications**, v. 391, n. 4, p. 1855–1864, a2012.

AMANCIO, D. R. et al. Complex networks analysis of language complexity. **Europhysics Letters (EPL)**, v. 100, n. 5, p. 58002, b2012.

AMANCIO, D. R. A perspective on the advancement of natural language processing tasks via topological analysis of complex networks. **Physics of Life Reviews**, v. 11, n. 4, p. 641–643, 2014.

AMARAL, D.; VIEIRA, R. Nerp-crf: uma ferramenta para o reconhecimento de entidades nomeadas por meio de conditional random fields. **Linguamática (Braga)**, 2014.

AMERICAN PSYCHIATRIC ASSOCIATION. **Diagnostic and Statistical Manual of Mental Disorders 5th edition**. Arlington, VA: American Psychiatric Association, 2013.

AMORIM, E.; CANÇADO, M.; VELOSO, A. **Automated Essay Scoring in the Presence of Biased Ratings**. Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. **Anais**...Association for Computational Linguistics, 2018.

AMORIM, E.; VELOSO, A. **A Multi-aspect Analysis of Automatic Essay Scoring for Brazilian Portuguese**. Proceedings of the Student Research Workshop at the 15th Conference of the European Chapter of the Association for Computational Linguistics. **Anais**...Valencia, Spain: Association for Computational Linguistics, abr. 2017.

ANANIADOU, S.; MCNAUGHT, J. **Text Mining for Biology And Biomedicine**. Norwood, MA, USA: Artech House, Inc., 2005.

ANCHIÊTA, R. T. et al. **PiLN IDPT 2021: Irony Detection in Portuguese Texts with Superficial Features and Embeddings**. Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2021) co-located with the Conference of the Spanish Society for Natural Language Processing (SEPLN 2021), XXXVII International Conference of the Spanish Society for Natural Language Processing., Málaga, Spain, September, 2021. **Anais**...2021.

ANDERSEN, P. M. et al. **Automatic extraction of facts from press releases to generate news stories**. Third Conference on Applied Natural Language Processing. **Anais**...1992.

ANDREW, J. J.; TANNIER, X. **Automatic Extraction of Entities and Relation from Legal Documents**. Proceedings of the Seventh Named Entities Workshop.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Anais...Melbourne, Australia: Association for Computational Linguistics, jul. 2018.

ANGELIDIS, I.; CHALKIDIS, I.; KOUBARAKIS, M. **Named entity recognition, linking and generation for greek legislation**. Legal Knowledge and Information Systems. **Anais**...IOS Press, 2018.

ANSARI, L.; JI, S. Ensemble hybrid learning methods for automated depression detection. **IEEE Transactions on computational Social Systems**, 2022.

ANTIQUEIRA, L. et al. Strong correlations between text quality and complex networks features. **Physica A: Statistical Mechanics and its Applications**, v. 373, p. 811–820, 2007.

ANTIQUEIRA, L. et al. A complex network approach to text summarization. **Information Sciences**, v. 179, n. 5, p. 584–599, 2009.

APPEL, A. P.; JUNIOR, E. R. H. Por dentro das redes complexas – detectando grupos e prevendo ligações. Em: VILAIN, P.; ROESLER, V. (Eds.). **Tópicos em Banco de Dados, Multimídia e Web: Minicursos do XXVI SBBD e do XVII WebMedia.** Porto Alegre: Sociedade Brasileira de Computação, 2011. p. 123–147.

ARAGÓN, M. E. et al. **Detecting Depression in Social Media using Fine-Grained Emotions**. NAACL-2019. **Anais**...Minneapolis, USA: Association for Computational Linguistics, 2019.

ARAUJO, P. H. L. DE et al. **LeNER-Br: A Dataset for Named Entity Recognition in Brazilian Legal Text**. (A. Villavicencio et al., Eds.)Computational Processing of the Portuguese Language. **Anais**...Cham: Springer International Publishing, 2018. Disponível em: <https://github.com/peluz/lener-br>

ARDILA, R. et al. **Common Voice: A Massively-Multilingual Speech Corpus**. (N. Calzolari et al., Eds.)Proceedings of the Twelfth Language Resources and Evaluation Conference. **Anais**...Marseille, France: European Language Resources Association, 2020. Disponível em: <https://aclanthology.org/2020.lrec-1.520/>

AREVALO, E. M.; FONTEYN, L. **MacBERTh: Development and Evaluation of a Historically Pre-trained Language Model for English (1450-1950)**. ICON Workshop on Natural Language Processing for Digital Humanities. **Anais**...2021.

ARFÉ, B.; MASON, L.; FAJARDO, I. Simplifying informational text structure for struggling readers. **Read Writ (2018) Volume 31, Issue 9**, p. 2191–2210, 2018.

ARORA, A. K. A. A. S. A. A. **Anxious Depression Prediction in Real-time Social Data**. International Conference on Advances in Engineering Science Management & Technology. **Anais**...Dehradun, India: 2019.

ASAHARA, M.; MATSUMOTO, Y. **Japanese named entity extraction with redundant morphological analysis**. Proceedings of the 2003 human language technology

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

conference of the North American chapter of the association for computational linguistics. **Anais**...2003.

ASCHBRENNER, J. A. N. A. A. B. A. J. T. A. K. A. Social Media and Mental Health: Benefits, Risks, and Opportunities for Research and Practice. **Journal of Technology in Behavioral Science**, v. 5, p. 245–257, 2020.

ASCHBRENNER, K. A. et al. A survey of online and mobile technology use at peer support agencies. **Psychiatric Quarterly**, p. 1–10, 2018.

ASSI, F. M. et al. **UFSCar's Team at ABSAPT 2022: Using Syntax, Semantics and Context for Solving the Tasks**. Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2022) co-located with the Conference of the Spanish Society for Natural Language Processing (SEPLN 2022), A Coruña, Spain, September 20, 2022. **Anais**...2022.

AVANÇO, L. V.; NUNES, M. DAS G. V. **Lexicon-Based Sentiment Analysis for Reviews of Products in Brazilian Portuguese**. Proceedings of the 2014 Brazilian Conference on Intelligent Systems. **Anais**...2014.

AVERINA, M.; LEVANOVA, O.; KASATKINA, N. **Named Entity Recognition for Russian Judicial Rulings Text**. 2022 32nd Conference of Open Innovations Association (FRUCT). **Anais**...2022.

AZIZ, W.; SPECIA, L. **Fully Automatic Compilation of a Portuguese-English Parallel Corpus for Statistical Machine Translation**. STIL 2011. **Anais**...Cuiabá, MT: 2011.

AZZIMONTI, M.; FERNANDES, M. Social media networks, fake news, and polarization. **European Journal of Political Economy**, v. 76, p. 102256, 2023.

BAADE, A.; PENG, P.; HARWATH, D. **MAE-AST: Masked Autoencoding Audio Spectrogram Transformer**. (H. Ko, J. H. L. Hansen, Eds.)Interspeech 2022, 23rd Annual Conference of the International Speech Communication Association, Incheon, Korea, 18-22 September 2022. **Anais**...ISCA, 2022. Disponível em: <https://doi.org/10.21437/Interspeech.2022-10961>

BABU, A. et al. XLS-R: Self-supervised cross-lingual speech representation learning at scale. **arXiv preprint arXiv:2111.09296**, 2021.

BACH, N. X. et al. **Reference Extraction from Vietnamese Legal Documents**. Proceedings of the 10th International Symposium on Information and Communication Technology. **Anais**...: SoICT '19.New York, NY, USA: Association for Computing Machinery, 2019.

BAEVSKI, A. et al. **wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations**., 2020. Disponível em: <https://arxiv.org/abs/2006.11477>

BAEZA-YATES, R. A.; RIBEIRO-NETO, B. A. Modern Information Retrieval-the con-

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

cepts and technology behind search. 2011.

BAEZA-YATES, R.; RIBEIRO-NETO, B. **Recuperação de Informação-: Conceitos e Tecnologia das Máquinas de Busca**. [s.l.] Bookman Editora, 2013.

BAHDANAU, D.; CHO, K.; BENGIO, Y. **Neural Machine Translation by Jointly Learning to Align and Translate**. (Y. Bengio, Y. LeCun, Eds.)3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings. **Anais**...San Diego, California.: 2015. Disponível em: <http://arxiv.org/abs/1409.0473>

BAI, J. et al. Qwen technical report. **arXiv preprint arXiv:2309.16609**, 2023.

BALAGE FILHO, P. P.; PARDO, T. A. S.; ALUÍSIO, S. M. **An Evaluation of the Brazilian Portuguese LIWC Dictionary for Sentiment Analysis**. Proceedings of the 9th Brazilian Symposium in Information and Human Language Technology. **Anais**...2013.

BALAGE FILHO, P. P.; PARDO, T. A. S.; NUNES, M. DAS G. V. **Sumarização automática de textos científicos: Estudo de caso com o sistema gistsumm**. [s.l.] Instituto de Ciências Matemáticas e de Computação da Universidade de São Paulo, 2007.

BALAMURALI, B. T. et al. Deep Neural Network-Based Respiratory Pathology Classification Using Cough Sounds. **Sensors**, v. 21, n. 16, 2021.

BANARESCU, L. et al. **Abstract Meaning Representation for Sembanking**. Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse. **Anais**...Sofia, Bulgaria: Association for Computational Linguistics, 2013. Disponível em: <http://aclweb.org/anthology/W13-2322>

BANERJEE, S.; LAVIE, A. **METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments**. (J. Goldstein et al., Eds.)Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization. **Anais**...Ann Arbor, Michigan: Association for Computational Linguistics, jun. 2005. Disponível em: <https://aclanthology.org/W05-0909>

BANKO, M. et al. **Open Information Extraction from the Web**. Proceedings of the 20th International Joint Conference on Artifical Intelligence. **Anais**...: IJCAI'07.San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2007. Disponível em: <http://citeseerx.ist.psu.edu/viewdoc/download;jsessionid=9909B5C03DA1A3CCFFF4263898B69100?doi=10.1.1.74.5174&rep=rep1&type=pdf>

BANZA, A. P. **A edição digital da História do Futuro, de António Vieira: arquivo e ferramentas**. Actas da Jornada de Humanidades Digitais do CIDEHUS (to appear). **Anais**...2022.

BARABÁSI, A.-L.; ALBERT, R. Emergence of scaling in random networks. **Science**, v. 286, n. 5439, p. 509–512, 1999.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

BARACHO, J.; LISBOA, L.; LOPES, R. **Levantamento e Análise Qualitativa de Bases de Dados de Fake News em Português**. Anais do VI Workshop sobre as Implicações da Computação na Sociedade. **Anais**...Porto Alegre, RS, Brasil: SBC, 2025. Disponível em: <https://sol.sbc.org.br/index.php/wics/article/view/35944>

BARBOSA DE LIMA, T. et al. Avaliação Automática de Redação: Uma revisão sistemática. **Revista Brasileira de Informática na Educação**, v. 31, p. 205–221, 2023.

BARBOSA, G. C. G.; GLAUBER, R.; CLARO, D. B. **Classificação de Relações Abertas Utilizando Features Independentes do Idioma**. Proceedings of the 4th Symposium on Knowledge Discovery, Mining and Learning (KDMiLe). **Anais**...SBC, 2016.

BARRAULT, L. et al. **Findings of the 2019 Conference on Machine Translation (WMT19)**. Proceedings of WMT. **Anais**...Florence, Italy: 2019.

BARRAULT, L. et al. **Findings of the 2020 Conference on Machine Translation (WMT20)**. Proceedings of the Fifth Conference on Machine Translation. **Anais**...Online: Association for Computational Linguistics, nov. 2020. Disponível em: <https://www.aclweb.org/anthology/2020.wmt-1.1>

BARRETT, M. J.; AGIC, Z.; SØGAARD, A. The Dundee Treebank. **Proceedings of the Fourteenth International Workshop on Treebanks and Linguistic Theories: TLT14**, p. 242–248, 2015.

BARRIERE, V.; FOURET, A. **May I Check Again? — A simple but efficient way to generate and use contextual dictionaries for Named Entity Recognition. Application to French Legal Texts.** Proceedings of the 22nd Nordic Conference on Computational Linguistics. **Anais**...Turku, Finland: Linköping University Electronic Press, 2019. Disponível em: <https://aclanthology.org/W19-6136>

BARROS, T. S. **Um modelo BERT para sumarização extrativa de textos em documentos da Polícia Federal**. mathesis—[s.l.] (Mestrado em Ciências da Computação) - Programa de Pós-Graduação em Ciência da Computação da Universidade Federal de Campina Grande, 2022.

BARZILAY, R.; ELHADAD, N.; MCKEOWN, K. **Sentence ordering in multidocument summarization**. Proceedings of the first international conference on Human language technology research. **Anais**...2001.

BASILE, V. et al. **We Need to Consider Disagreement in Evaluation**. (K. Church, M. Liberman, V. Kordoni, Eds.)Proceedings of the 1st Workshop on Benchmarking: Past, Present and Future. **Anais**...Online: Association for Computational Linguistics, ago. 2021. Disponível em: <https://aclanthology.org/2021.bppf-1.3/>

BATISTA, H. H. et al. **A comparative analysis of text embedding approach to extract named entities in Portuguese legal documents**. Anais do XVIII Encontro Nacional de Inteligência Artificial e Computacional. **Anais**...SBC, 2021.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

BAXENDALE, P. B. Machine-made index for technical literature—an experiment. **IBM Journal of research and development**, v. 2, n. 4, p. 354–361, 1958.

BELCAVELLO, F. et al. **Frame-Based Annotation of Multimodal Corpora: Tracking (A)Synchronies in Meaning Construction**. (T. T. Torrent et al., Eds.)Proceedings of the International FrameNet Workshop 2020: Towards a Global, Multilingual FrameNet. **Anais**...Marseille, France: European Language Resources Association, 2020. Disponível em: <https://aclanthology.org/2020.framenet-1.4>

BELCAVELLO, F. **FrameNet Annotation for Multimodal Corpora: devising a methodology for the semantic representation of text-image interactions in audiovisual productions**. Tese de Doutorado em Linguística—Juiz de Fora: Universidade Federal de Juiz de Fora, 2023.

BELCAVELLO, F. et al. **Frame2: A FrameNet-based Multimodal Dataset for Tackling Text-image Interactions in Video**. (N. Calzolari et al., Eds.)Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024). **Anais**...Torino, Italia: ELRA; ICCL, 2024. Disponível em: <https://aclanthology.org/2024.lrec-main.655/>

BELTAGY, I.; LO, K.; COHAN, A. **SciBERT: A Pretrained Language Model for Scientific Text**. Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). **Anais**...Association for Computational Linguistics, 2019. Disponível em: <https://aclanthology.org/D19-1371>

BENDER, E. M. **Linguistically Naïve != Language Independent: Why NLP Needs Linguistic Typology**. Proceedings of the EACL 2009 Workshop on the Interaction between Linguistics and Computational Linguistics: Virtuous, Vicious or Vacuous? **Anais**...Athens, Greece: Association for Computational Linguistics, mar. 2009. Disponível em: <https://www.aclweb.org/anthology/W09-0106>

BENGIO, Y.; COURVILLE, A.; VINCENT, P. Representation learning: A review and new perspectives. **IEEE transactions on pattern analysis and machine intelligence**, v. 35, n. 8, p. 1798–1828, 2013.

BERBER SARDINHA, T. **Linguística de Corpus**. Barueri: Manole, 2004.

BERTI, L. C. et al. Fundamental frequency related parameters in Brazilians with COVID-19. **The Journal of the Acoustical Society of America**, v. 153, n. 1, p. 576–585, 2023.

BERTI, L. C. et al. **Acoustic Characteristics of Voice and Speech in Post-COVID-19**. Healthcare. **Anais**...MDPI, 2025.

BHARDWAJ, S.; AGGARWAL, S.; MAUSAM, M. **CaRB: A crowdsourced benchmark for open IE**. Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Pro-

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

cessing (EMNLP-IJCNLP). **Anais**...2019.

BHATTACHARYA, D. et al. Coswara: A respiratory sounds and symptoms dataset for remote screening of SARS-CoV-2 infection. **Scientific data**, v. 10, n. 1, p. 397, 2023.

BIBER, D. Representativeness in Corpus Design. **Literary and Linguistic Computing**, v. 8, n. 4, p. 243–257, jan. 1993.

BIBER, D. Register: Overview. Em: BROWN, K. (Ed.). **Encyclopedia of Language & Linguistics (Second Edition)**. Second Edition ed. Oxford: Elsevier, 2006. p. 476–482.

BICK, E. **The Parsing System "Palavras": Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework**. tese de doutorado—[s.l.] Aarhus University Press, Denmark; University of Arhus, 2000.

BIKEL, D. M.; SCHWARTZ, R.; WEISCHEDEL, R. M. An algorithm that learns what's in a name. **Machine learning**, v. 34, p. 211–231, 1999.

BIRD, S. **NLTK: the natural language toolkit**. Proceedings of the COLING/ACL 2006 Interactive Presentation Sessions. **Anais**...2006.

BIRNBAUM, M. L. et al. Role of social media and the Internet in pathways to care for adolescents and young adults with psychotic disorders and nonpsychotic mood disorders. **Early Intervention in Psychiatry**, v. 11, n. 4, p. 290–295, 2017.

BITTENCOURT JR., J. A. S. **Avaliação automática de redação em língua portuguesa empregando redes neurais profundas**. mathesis—[s.l.] Universidade Federal de Goiás, 2020.

BITTENCOURT, M. M.; SILVA, R. M.; ALMEIDA, T. A. ML-MDLText: An efficient and lightweight multilabel text classifier with incremental learning. **APPLIED SOFT COMPUTING**, v. 96, p. 1–15, 2020.

BLEI, D. M.; MORENO, P. J. **Topic Segmentation with an Aspect Hidden Markov Model**. Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. **Anais**...New York, NY, USA: Association for Computing Machinery, 2001.

BLONDEL, V. D. et al. Fast unfolding of communities in large networks. **Journal of Statistical Mechanics: Theory and Experiment**, v. 2008, n. 10, p. P10008, 2008.

BOITO, M. Z. Simplificação lexical de substantivos e multiword expressions. **Salão de Iniciação Científica (26. : 2014 out. 20-24 : UFRGS, Porto Alegre, RS)**, 2014.

BOJANOWSKI, P. et al. Enriching Word Vectors with Subword Information. **Transactions of the Association for Computational Linguistics**, v. 5, p. 135–146, 2017.

BOJAR, O. et al. **Findings of the 2016 Conference on Machine Translation**.

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

Proceedings of the First Conference on Machine Translation. **Anais**...Berlin, Germany: Association for Computational Linguistics, 2016.

BOND JR., C. F.; DEPAULO, B. M. Accuracy of Deception Judgments. **Personality and Social Psychology Review**, v. 10, n. 3, p. 214–234, 2006.

BONIFACIO, L. H. et al. **A Study on the Impact of Intradomain Finetuning of Deep Language Models for Legal Named Entity Recognition in Portuguese**. (R. Cerri, R. C. Prati, Eds.)Intelligent Systems. **Anais**...Cham: Springer International Publishing, 2020.

BONIFACIO, L. H. et al. **mMARCO: A Multilingual Version of MS MARCO Passage Ranking Dataset**., 2021. Disponível em: <https://arxiv.org/abs/2108.13897>

BONIFACIO, L. H. N. **Modelos Profundos de Linguagem para Reconhecimento de Entidades Nomeadas em Domínio Jurídico**. mathesis—[s.l.] Master's thesis, Universidade Federal de Mato Grosso do Sul, 2020.

BORDINO, I. et al. Garnlp: a natural language processing pipeline for garnishment documents. **Information Systems Frontiers**, v. 23, p. 101–114, 2021.

BORIN, E.; DONATO, F. Financial Sustainability of Digitizing Cultural Heritage: The International Platform Europeana. **Journal of Risk and Financial Management**, v. 16, n. 10, p. 421, 2023.

BOWKER, L.; PEARSON, J. **Working with Specialized Language: A Practical Guide to Using Corpora**. 1. ed. London: Routledge, 2002.

BRANDT, M. B. **Modelagem da informação legislativa: arquitetura da informação para o processo legislativo brasileiro**. tese de doutorado—[s.l.] Universidade Estadual Paulista (Unesp), 2020.

BRAUN, H. I. Understanding Scoring Reliability: Experiments in Calibrating Essay Readers. **Journal of Educational Statistics**, v. 13, n. 1, p. 1–18, 1988.

BREIMAN, L. Random forests. **Machine learning**, v. 45, n. 1, p. 5–32, 2001.

BREITFELLER, L. et al. **Finding Microaggressions in the Wild: A Case for Locating Elusive Phenomena in Social Media Posts**. Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). **Anais**...2019.

BRICIU, A.; LUPEA, M. **Studying the language of mental illness in Romanian social media**. IEEE 14th International Conference on Intelligent Computer Communication and Processing (ICCP). **Anais**...2018.

BRIDGEMAN, B. Handbook of automated essay evaluation: Current applications and new directions. Em: SHERMIS, M. D.; BURSTEIN, J. (Eds.). [s.l.] Routledge/Taylor &

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Francis Group, 2013. p. 221–232.

BRIGHAM, E. O.; MORROW, R. The fast Fourier transform. **IEEE spectrum**, v. 4, n. 12, p. 63–70, 1967.

BRIN, S. **Extracting patterns and relations from the world wide web**. International workshop on the world wide web and databases. **Anais**...Springer, 1998.

BRITO, M. et al. **CDJUR-BR - Uma Coleção Dourada do Judiciário Brasileiro com Entidades Nomeadas Refinadas**. Anais do XIV Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana. **Anais**...Porto Alegre, RS, Brasil: SBC, 2023. Disponível em: <https://github.com/mauriciobritojr/CDJUR-BR>

BRITTO, H.; FINGER, M.; GALVES, C. Computational and linguistic aspects of the construction of The Tycho Brahe Parsed Corpus of Historical Portuguese. **Romanistische Korpuslinguistik, Korpora und gesprochene Sprache, Romance Corpus Linguistics, Corpora and Spoken Language, ScriptOralia, v. 126.**, 2002.

BROWN, C. et al. **Exploring automatic diagnosis of COVID-19 from crowdsourced respiratory sound data**. Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining. **Anais**...a2020.

BROWN, P. et al. **A statistical approach to language translation**. Proceedings of the 12th conference on Computational linguistics -. **Anais**...Budapest, Hungry: Association for Computational Linguistics, 1988. Disponível em: <http://portal.acm.org/citation.cfm?doid=991635.991651>. Acesso em: 10 jun. 2020

BROWN, T. et al. Language models are few-shot learners. **Advances in neural information processing systems**, v. 33, p. 1877–1901, b2020.

BROWN, T. B. et al. **Language Models are Few-Shot Learners**. (H. Larochelle et al., Eds.)Advances in Neural Information Processing Systems. **Anais**...Curran Associates, Inc., c2020. Disponível em: <https://proceedings.neurips.cc/paper/2020/hash/1457c0d6bfcb4967418bfb8ac142f64a-Abstract.html>

BRUM, H.; NUNES, M. DAS G. V. **Building a Sentiment Corpus of Tweets in Brazilian Portuguese**. (N. C. (Conference chair) et al., Eds.)Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018). **Anais**...Miyazaki, Japan: European Language Resources Association (ELRA), mar. 2018.

BRUNETTE, M. et al. Use of smartphones, computers and social media among people with SMI: opportunity for intervention. **Community Mental Health Journal**, p. 1–6, 2019.

BUCCI, S.; SCHWANNAUER, M.; BERRY, N. The digital revolution and its impact on mental health care. **Psychology and Psychotherapy: Theory, Research and Practice**, v. 92, n. 2, p. 277–297, 2019.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

BUCKLEY, C.; VOORHEES, E. M. **Evaluating evaluation measure stability**. ACM SIGIR Forum. **Anais**...ACM New York, NY, USA, 2017.

BUDEL, G. et al. Topological properties and organizing principles of semantic networks. **arXiv preprint arXiv:2304.12940**, 2023.

BUENO, R. O. et al. **Overview of the Task on Irony Detection in Spanish Variants**. Proceedings of the Iberian Languages Evaluation Forum co-located with 35th Conference of the Spanish Society for Natural Language Processing. **Anais**...2019.

BURDISSO, S. G.; ERRECALDE, M.; MONTES-Y-GÓMEZ, M. t-SS3: a text classifier with dynamic n-grams for early risk detection over text streams. **Pattern Recognition Letters**, v. 138, p. 130–137, 2020.

BURRISS, L. L. Attribution in network radio news: A cross-network analysis. **Journalism Quarterly**, v. 65, n. 3, p. 690–694, 1988.

BURSTEIN, J. **Opportunities for Natural Language Processing Research in Education**. (A. Gelbukh, Ed.)Computational Linguistics and Intelligent Text Processing. **Anais**...Berlin, Heidelberg: Springer Berlin Heidelberg, 2009.

CABRAL, B.; SOUZA, M.; CLARO, D. B. **PortNOIE: A Neural Framework for Open Information Extraction for the Portuguese Language**. International Conference on Computational Processing of the Portuguese Language. **Anais**...Springer, 2022.

CABRAL, L. et al. **FakeWhastApp.BR: NLP and Machine Learning Techniques for Misinformation Detection in Brazilian Portuguese WhatsApp Messages**. Proceedings of the 23rd International Conference on Enterprise Information Systems (ICEIS 2021) - Volume 1. **Anais**...2021.

CABRÉ, M. T. A Terminologia, uma disciplina em evolução: passado, presente e alguns elementos de futuro. **Debate Terminológico. ISSN: 1813-1867**, n. 01, 2005.

CABRERA-DIEGO, L. A.; GHEEWALA, A. **Jus Mundi at SemEval-2023 Task 6: Using a Frustratingly Easy Domain Adaption for a Legal Named Entity Recognition System**. Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023). **Anais**...Toronto, Canada: Association for Computational Linguistics, jul. 2023.

CAI, Y. et al. **Rank-Then-Score: Enhancing Large Language Models for Automated Essay Scoring**., 2025. Disponível em: <https://arxiv.org/abs/2504.05736>

CAMERON, H. F.; GONÇALVES, M. F.; QUARESMA, P. **Linguistic and orthographical classic Portuguese variants Challenges for NLP**. Proceedings of the 14th International Conference on the Computational Processing of Portuguese. **Anais**...2020.

CAMERON, H.; OLIVAL, F.; VIEIRA, R. Planear a normalização automática: tipologia de variação gráfica do corpus das Memórias Paroquiais (1758). **LaborHistórico**, v. 9, n.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

1, p. 52234, 2023.

CANAVILHAS, J. Webjornalismo: Da pirâmide invertida à pirâmide deitada. **Aprender**, n. 32, p. 58–65, 2012.

CANDIDO JUNIOR, A. **Análise bidirecional da língua na simplificação sintática em textos de português voltada à acessibilidade digital**. ICMC - USP São Carlos: Biblioteca Digital USP, 2013.

CANDIDO-JUNIOR, A.; OLIVEIRA, M. DE; ALUÍSIO, S. M. Simplifica: um Sistema Web de Autoria de Textos Simplificados. **Simpósio Brasileiro de Sistemas Multimídia e Web (Webmedia 2009) v.2**, p. 55–58, 2009.

CARDELLINO, C. et al. **Legal NERC with ontologies, Wikipedia and curriculum learning**. Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers. **Anais**...Valencia, Spain: Association for Computational Linguistics, abr. 2017. Disponível em: <https://aclanthology.org/E17-2041>

CARDOSO, E. F.; SILVA, R. M.; ALMEIDA, T. A. Towards automatic filtering of fake reviews. **Neurocomputing**, v. 309, p. 106–116, 2018.

CARDOSO, P. C. F. et al. **CSTNews-a discourse-annotated corpus for single and multi-document summarization of news texts in Brazilian Portuguese**. Proceedings of the 3rd RST Brazilian Meeting. **Anais**...2011.

CARDOSO, P. C. F. **Exploração de métodos de sumarização automática multidocumento com base em conhecimento semântico-discursivo**. tese de doutorado—[s.l.] (Doutorado em Ciências de Computação e Matemática Computacional) - Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, 2014.

CARL, M.; WAY, A. (EDS.). **Recent Advances in Example-Based Machine Translation**. [s.l.] Springer Netherlands, 2003.

CARMO, D. et al. PTT5: Pretraining and validating the T5 model on Brazilian Portuguese data. **CoRR**, v. abs/2008.09144, b2020.

CARMO, D. et al. PTT5: Pretraining and validating the T5 model on Brazilian Portuguese data. **arXiv preprint arXiv:2008.09144**, a2020.

CARNEIRO, F. C. A. D. F. A. F. J. N. A. V. Early Detection of Depression: Social Network Analysis and Random Forest Techniques. **J Med Internet Res**, v. 21, n. 6, p. e12554, 2019.

CARPINETO, C.; ROMANO, G. A survey of automatic query expansion in information retrieval. **Acm Computing Surveys (CSUR)**, v. 44, n. 1, p. 1–50, 2012.

CARROLL, J. et al. **Practical Simplification of English Newspaper Text to Assist**

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Aphasic Readers. In Proc. of AAAI-98 Workshop on Integrating Artificial Intelligence and Assistive Technology. **Anais**...1998.

CARVALHO, F.; SANTOS, G. DOS; GUEDES, G. P. **AffectPT-br: an Affective Lexicon based on LIWC 2015**. Proceedings of the 37th International Conference of the Chilean Computer Science Society. **Anais**...2018.

CARVALHO, P. et al. **Clues for Detecting Irony in User-Generated Contents: Oh...!! It's "so Easy";-)**. Proceedings of the 1st International CIKM Workshop on Topic-Sentiment Analysis for Mass Opinion. **Anais**...2009.

CARVALHO, P.; SILVA, M. J. **SentiLex-PT 02**. https://b2share.eudat.eu, 2017. Disponível em: <https://b2share.eudat.eu/records/93ab120efdaa4662baec6adee8e7585f>

CASANOVA, E. et al. **Transfer Learning and Data Augmentation Techniques to the COVID-19 Identification Tasks in ComParE 2021**. Proc. Interspeech 2021. **Anais**...a2021.

CASANOVA, E. et al. **Deep Learning against COVID-19: Respiratory Insufficiency Detection in Brazilian Portuguese Speech**. Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021. **Anais**...Online: Association for Computational Linguistics, ago. b2021.

CASELI, H. DE M. et al. Building a Brazilian Portuguese parallel corpus of original and simplified texts. **Advances in Computational Linguistics, Research in Computer Science (CICLing-2009)**, v. 41, p. 59–70, 2009.

CASELI, H. DE M.; INÁCIO, M. **NMT and PBSMT Error Analyses in English to Brazilian Portuguese Automatic Translations**. Proceedings of the Twelfth Language Resources and Evaluation Conference. **Anais**...Marseille, France: European Language Resources Association, 2020. Disponível em: <https://aclanthology.org/2020.lrec-1.446>

CÁSSIA ALVES, V. DE et al. College students-in-the-loop for their mental health: a case of AI and humans working together to support well-being. **Interaction Design and Architecture(s)**, n. 59, p. 79–94, 2024.

CASTANO, A.; CASACUBERTA, F. **A connectionist approach to machine translation**. 5th European Conference on Speech Communication and Technology (Eurospeech 1997). **Anais**...ISCA, set. 1997. Disponível em: <http://dx.doi.org/10.21437/eurospeech.1997-50>

CASTILHO, S. et al. **Does post-editing increase usability? A study with Brazilian Portuguese as Target Language**. Proceedings of the 17th annual conference of the European association for machine translation. **Anais**...2014.

CASTILHO, S. et al. **A comparative quality evaluation of PBSMT and NMT using professional translators**. Proceedings of Machine Translation Summit XVI: Research Track. **Anais**...a2017.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

CASTILHO, S. et al. Is Neural Machine Translation the New State of the Art? **The Prague Bulletin of Mathematical Linguistics**, v. 108, n. 1, p. 109–120, jun. b2017.

CASTILHO, S. et al. Approaches to Human and Machine Translation Quality Assessment. Em: **Translation Quality Assessment: From Principles to Practice**. Machine Translation: Technologies e Applications. [s.l.] Springer International Publishing, 2018. v. 1p. 9–38.

CASTILHO, S. et al. Editors' foreword to the special issue on human factors in neural machine translation. **Machine Translation**, v. 33, n. 1–2, p. 1–7, maio a2019.

CASTILHO, S. **On the Same Page? Comparing IAA in Sentence and Document Level Human MT Evaluation**. Proceedings of the Fifth Conference on Machine Translation. **Anais**...Association for Computational Linguistics, nov. 2020. Disponível em: <https://www.aclweb.org/anthology/2020.wmt-1.137>

CASTILHO, S. **Towards Document-Level Human MT Evaluation: On the Issues of Annotator Agreement, Effort and Misevaluation**. Proceedings of the Workshop on Human Evaluation of NLP Systems. **Anais**...Association for Computational Linguistics, abr. 2021. Disponível em: <https://www.aclweb.org/anthology/2021.humeval-1.4>

CASTILHO, S. **How Much Context Span is Enough? Examining Context-Related Issues for Document-level MT**. Proceedings of the Language Resources and Evaluation Conference. **Anais**...Marseille, France: European Language Resources Association, 2022. Disponível em: <https://aclanthology.org/2022.lrec-1.323>

CASTILHO, S. et al. **Translation Systems Care for Context? What About a GPT Model?** Proceedings of the 24th Annual Conference of the European Association for Machine Translation. **Anais**...Tampere, Finland: EAMT, 2023. Disponível em: <https://events.tuni.fi/uploads/2023/06/11678752-proceedings-eamt2023.pdf>

CASTILHO, S.; RESENDE, N. Post-Editese in Literary Translations. **Information**, v. 13, n. 2, p. 66, 2022.

CASTILHO, S.; RESENDE, N.; MITKOV, R. **What Influences the Features of Post-editese? A Preliminary Study**. Proceedings of the Human-Informed Translation and Interpreting Technology Workshop (HiT-IT 2019). **Anais**...Varna, Bulgaria: Incoma Ltd., Shoumen, Bulgaria, set. b2019. Disponível em: <https://aclanthology.org/W19-8703>

CASTRO, P. **Aprendizagem profunda para reconhecimento de entidades nomeadas em domínio jurídico**. mathesis—[s.l.] Master's thesis, Universidade Federal de Goiás, 2019.

CASTRO, P. V. Q. DE; SILVA, N. F. F. DA; SOARES, A. DA S. **Portuguese Named Entity Recognition Using LSTM-CRF**. (A. Villavicencio et al., Eds.)Proceedings of the 13th International Conference on the Computational Processing of the Portuguese Language. **Anais**...2018.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

ÇETINDAĞ, C.; YAZICIOĞLU, B.; KOÇ, A. Named-entity recognition in Turkish legal texts. **Natural Language Engineering**, p. 1–28, 2022.

CHAE, Y.; DAVIDSON, T. Large language models for text classification: from zero-shot learning to instruction-tuning. **Sociological Methods & Research**, p. 00491241251325243, 2025.

CHAKRABORTY, A. et al. **Stop Clickbait: Detecting and preventing clickbaits in online news media**. 2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM). **Anais**...2016.

CHALKIDIS, I. et al. **LEGAL-BERT: The Muppets straight out of Law School**. Findings of the Association for Computational Linguistics: EMNLP 2020. **Anais**...Online: Association for Computational Linguistics, nov. 2020.

CHALKIDIS, I. et al. **Regulatory Compliance through Doc2Doc Information Retrieval: A case study in EU/UK legislation where text similarity has limitations**. Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume. **Anais**...Online: Association for Computational Linguistics, abr. 2021.

CHALKIDIS, I.; ANDROUTSOPOULOS, I. A deep learning approach to contract element extraction. Em: **Legal knowledge and information systems**. [s.l.] IOS Press, 2017. p. 155–164.

CHALKIDIS, I.; ANDROUTSOPOULOS, I.; ALETRAS, N. **Neural Legal Judgment Prediction in English**. Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. **Anais**...Association for Computational Linguistics, 2019.

CHALKIDIS, I.; ANDROUTSOPOULOS, I.; MICHOS, A. **Extracting Contract Elements**. Proceedings of the 16th Edition of the International Conference on Articial Intelligence and Law. **Anais**...: ICAIL '17.New York, NY, USA: Association for Computing Machinery, 2017.

CHALL, J. S.; DALE, E. **Readability revisited: the new Dale-Chall readability formula**. [s.l.] Brookline Books, 1995.

CHALMERS, D. J. Syntactic transformations on distributed representations. **Connectionist Natural Language Processing: Readings from Connection Science**, p. 46–55, 1992.

CHANDRASEKAR, R.; DORAN, C.; SRINIVAS, B. Motivations and methods for text simplification. **Proceedings of the 16th International Conference on Computational Linguistics (COLING)**, p. 1041–1044, 1996.

CHARLES, A. C.; RUBACK, L.; OLIVEIRA, J. **Fakepedia Corpus: A Flexible Fake News Corpus in Portuguese**. Computational Processing of the Portuguese Language: 15th International Conference, PROPOR 2022, Fortaleza, Brazil, March 21–23,

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

2022, Proceedings. **Anais**...Berlin, Heidelberg: Springer-Verlag, 2022. Disponível em: <https://doi.org/10.1007/978-3-030-98305-5_4>

CHAVARRO, J. et al. **FakeTrueBR: Um corpus brasileiro de notícias falsas**. Anais da XVIII Escola Regional de Banco de Dados. **Anais**...Porto Alegre, RS, Brasil: SBC, 2023. Disponível em: <https://sol.sbc.org.br/index.php/erbd/article/view/24352>

CHEN, A.; CHEN, D. O. Simulation of a machine learning enabled learning health system for risk prediction using synthetic patient data. **Scientific Reports**, v. 12, n. 1, p. 17917, out. 2022.

CHEN, D. et al. Critical nodes identification in complex networks: a survey. **Complex Engineering Systems**, v. 5, n. 11, p. –, 2025.

CHEN, P. P. The Entity-Relationship Model - Toward a Unified View of Data. **ACM Trans. Database Syst.**, v. 1, n. 1, p. 9–36, 1976.

CHEN, T.; GUESTRIN, C. **XGBoost: A scalable tree boosting system**. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. **Anais**...ACM, 2016.

CHEN, Y. et al. **Joint Entity and Relation Extraction for Legal Documents with Legal Feature Enhancement**. Proceedings of the 28th International Conference on Computational Linguistics. **Anais**...Barcelona, Spain (Online): International Committee on Computational Linguistics, dez. 2020.

CHEN, Y.; CONROY, N. J.; RUBIN, V. L. **Misleading Online Content: Recognizing Clickbait as "False News"**. Proceedings of the 2015 ACM on Workshop on Multimodal Deception Detection. **Anais**...: WMDD '15.New York, NY, USA: Association for Computing Machinery, 2015. Disponível em: <https://doi.org/10.1145/2823465.2823467>

CHOUDHURY, M. D. et al. **Predicting Depression via Social Media**. International AAAI Conference on Web and Social Media (ICWSM). **Anais**...AAAI, 2013.

CHRISMAN, L. Learning recursive distributed representations for holistic computation. **Connection Science**, v. 3, n. 4, p. 345–366, 1991.

CIAMPAGLIA, G. L. et al. Computational fact checking from knowledge networks. **PloS one**, v. 10, n. 6, p. e0128193, 2015.

CIGNARELLA, A. T. et al. **Overview of the EVALITA 2018 Task on Irony Detection in Italian Tweets (IronITA)**. Proceedings of the Sixth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2018) co-located with the Fifth Italian Conference on Computational Linguistics (CLiC-it 2018). **Anais**...2018.

CLARK, K. et al. **ELECTRA: Pre-training Text Encoders as Discriminators Rather Than Generators**. 8th International Conference on Learning Representations,

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

ICLR 2020. **Anais**...Addis Ababa, Ethiopia: OpenReview.net, abr. 2020. Disponível em: <https://openreview.net/forum?id=r1xMH1BtvB>

CLARKE, D. J. B. et al. FAIRshake: Toolkit to Evaluate the FAIRness of Research Digital Resources. **Cell Systems**, v. 9, n. 5, p. 417–421, 2019.

CLEM, S. Post-Truth and Vices Opposed to Truth. v. 37, n. 2, p. 97–116, 2017.

COELHO, G. et al. **Information Extraction in the Legal Domain: Traditional Supervised Learning vs. ChatGPT**. INSTICC; SciTePress, 2024.

COELLO, J. M. A.; JUNQUEIRA, B. A. Automatic Analysis of Facebook Posts and Comments Written in Brazilian Portuguese. **International Journal for Innovation Education and Research**, 2019.

COHAN, A. et al. **SMHD: a Large-Scale Resource for Exploring Online Language Usage for Multiple Mental Health Conditions**. COLING-2018. **Anais**...Santa Fe, USA: Association for Computational Linguistics, 2018.

COHEN, J. P. et al. Problems in the deployment of machine-learned models in health care. **Cmaj**, v. 193, n. 35, p. E1391–E1394, 2021.

COLEMAN, M.; LIAU, T. L. A computer readability formula designed for machine scoring. **Journal of Applied Psychology**, v. 60, p. 283–284, 1975.

COLLOVINI, S. et al. **IberLEF 2019 Portuguese Named Entity Recognition and Relation Extraction Tasks**. Proceedings of the Iberian Languages Evaluation Forum co-located with 35th Conference of the Spanish Society for Natural Language Processing. **Anais**...2019. Disponível em: <http://ceur-ws.org/Vol-2421/NER/_Portuguese/_overview.pdf>

CONCEIÇÃO, M. C.; ZANOLA, M. T. **Terminologia e mediação linguística: métodos, práticas e atividades**. Universidade do Algarve Editora, 2020.

CONEGLIAN, C. S.; SANTAREM SEGUNDO, J. E. Europeana no Linked Open Data: conceitos de Web Semântica na dimensão aplicada das Humanidades Digitais. **Encontros Bibli: revista eletrônica de biblioteconomia e ciência da informação**, v. 22, n. 48, p. 88–99, 2017.

CONNEAU, A. et al. **Unsupervised cross-lingual representation learning at scale**. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. **Anais**...2020.

CONROY, N. K.; RUBIN, V. L.; CHEN, Y. Automatic deception detection: Methods for finding fake news. **Proceedings of the Association for Information Science and Technology**, v. 52, n. 1, p. 1–4, 2015.

CONSOLI, B. S. et al. **Embeddings for Named Entity Recognition in Geoscience**

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Portuguese Literature. Proceedings of The 12th Language Resources and Evaluation Conference. **Anais**...2020.

CONSORTIUM, L. D. ACE (Automatic Content Extraction) English Annotation Guidelines for Events. **Version**, n. 5.4.3, 2005.

COPPERSMITH, G. et al. **CLPsych 2015 Shared Task: Depression and PTSD on Twitter**. Second Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality. **Anais**...Denver, USA: Association for Computational Linguistics, 2015.

CORDEIRO, P. R.; PINHEIRO, V. **Um corpus de notıcias falsas do twitter e verificaçao automática de rumores em lıngua portuguesa**. Proceedings of the Symposium in Information and Human Language Technology. **Anais**...2019.

CORNU, G. **Linguistique juridique**. [s.l: s.n.].

CORRÊA, U. B. **Análise de sentimento baseada em aspectos usando aprendizado profundo: uma proposta aplicada à língua portuguesa**. tese de doutorado—[s.l.] Universidade Federal de Pelotas, 2021.

CORREIA, F. A. et al. Fine-grained legal entity annotation: A case study on the Brazilian Supreme Court. **Information Processing & Management**, v. 59, n. 1, p. 102794, 2022.

CORTES, C.; VAPNIK, V. Support-Vector Networks. **Machine Learning**, v. 20, n. 3, p. 273–297, set. 1995.

CORTES, E. et al. **An Empirical Comparison of Question Classification Methods for Question Answering Systems**. (N. Calzolari et al., Eds.)Proceedings of the Twelfth Language Resources and Evaluation Conference. **Anais**...Marseille, France: European Language Resources Association, 2020. Disponível em: <https://aclanthology.org/2020.lrec-1.665>

COSTA, A. et al. A linguistically motivated taxonomy for Machine Translation error analysis. **Machine Translation**, v. 29, n. 2, p. 127–161, 2015.

COSTA, A. D. **A tradução por máquina enriquecida semanticamente com frames e papéis qualia**. Ph.D. thesis—Juiz de Fora: Universidade Federal de Juiz de Fora, 2020.

COSTA, P. B. DA et al. **BERTabaporu: assessing a genre-specific language model for Portuguese NLP**. Recents Advances in Natural Language Processing (RANLP-2023). **Anais**...Varna, Bulgaria: 2023.

COSTA, R. et al. **Expanding UlyssesNER-Br Named Entity Recognition Corpus with Informal User-Generated Text**. (G. Marreiros et al., Eds.)Progress in Artificial Intelligence. **Anais**...Cham: Springer International Publishing, 2022. Disponível em: <https://github.com/ulysses-camara/>

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

COUTINHO, I.; MARTINS, B. Transformer-based models for ICD-10 coding of death certificates with Portuguese text. **Journal of Biomedical Informatics**, v. 136, p. 104232, 2022.

COUTO, J. M. M.; REIS, J. C. S.; BENEVENUTO, F. Can computer network attributes be useful for identifying low-credibility websites? A case study in Brazil. **Social Network Analysis and Mining**, v. 14, n. 1, p. 153, 2024.

COVER, T.; HART, P. Nearest neighbor pattern classification. **IEEE Transactions on Information Theory**, v. 13, n. 1, p. 21–27, 1967.

COWIE, J. R. **Automatic analysis of descriptive texts**. First Conference on Applied Natural Language Processing. **Anais**...1983.

COWIE, J.; LEHNERT, W. Information extraction. **Communications of the ACM**, v. 39, n. 1, p. 80–91, 1996.

COX, A. N. A. M., Simon J. D. AND Gonzalez-Beltran. Ten simple rules for making a vocabulary FAIR. **PLOS Computational Biology**, v. 17, n. 6, p. 1–15, jun. 2021.

COX, D. R. The regression analysis of binary sequences. **Journal of the Royal Statistical Society: Series B (Methodological)**, v. 20, n. 2, p. 215–232, 1958.

CRESTANI, F. et al. "Is this document relevant?... probably" a survey of probabilistic models in information retrieval. **ACM Computing Surveys (CSUR)**, v. 30, n. 4, p. 528–552, 1998.

CROFT, W. **Typology and Universals**. 2. ed. [s.l.] Cambridge University Press, 2002.

CROFT, W. B.; METZLER, D.; STROHMAN, T. **Search engines: Information retrieval in practice**. [s.l.] Addison-Wesley, 2010. v. 520

CSIKSZENTMIHALYI, M. **Flow: The Psychology of Optimal Experience**. [s.l.] Harper Perennial, 2008.

CUCCHIARELLI, A.; VELARDI, P. Unsupervised named entity recognition using syntactic and semantic contextual evidence. **Computational Linguistics**, v. 27, n. 1, p. 123–131, 2001.

CUI, H. et al. **Probabilistic query expansion using query logs**. Proceedings of the 11th international conference on World Wide Web. **Anais**...2002.

CUI, L.; WEI, F.; ZHOU, M. **Neural Open Information Extraction**. Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). **Anais**...2018.

CULOTTA, A.; MCCALLUM, A.; BETZ, J. **Integrating probabilistic extraction models and data mining to discover relations and patterns in text**. Procee-

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

dings of the Human Language Technology Conference of the NAACL, Main Conference. **Anais**...2006.

CUNHA, A. L. V. DA. **Coh-Metrix-Dementia: análise automática de distúrbios de linguagem nas demências utilizando Processamento de Línguas Naturais**. ICMC - USP São Carlos: Biblioteca Digital USP, 2015.

CUNHA, L. C. C. DA. **Um Corpus anotado de mensagens do WhatsApp em PT-BR para detecção automática de desinformação textual**. https://github.com/cabrau/FakeWhatsApp.Br, 2021.

DA SILVA JR., J. A. **Um avaliador automático de redações**. mathesis—[s.l.] Universidade Federal do Espírito Santo, 2021.

DADICO, C. M. **O Ódio Ancestral Como Elemento Constitutivo Do Estado Moderno e Seus Reflexos Na Compreensão dos Crimes De Ódio: Um Diálogo Entre o Direito Internacional e o Direito Brasileiro**. tese de doutorado—Porto Alegre, RS, Brazil: Programa de Pós-Grduação em Ciências Criminais da Escola de Direito da Pontifícia Universidade Católica do Rio Grande do Sul, 2020.

DAHL, Ö. **The Growth and Maintenance of Linguistic Complexity**. Amsterdam: John Benjamins Publishing Company, 2004. v. 71

DAI, E.; SUN, Y.; WANG, S. **Ginger cannot cure cancer: Battling fake health news with a comprehensive data repository**. Proceedings of the International AAAI Conference on Web and Social Media. **Anais**...Atlanta, USA: 2020.

DALE, R.; MAZUR, P. **Handling Conjunctions in Named Entities**. (A. Gelbukh, Ed.)Computational Linguistics and Intelligent Text Processing. **Anais**...Berlin, Heidelberg: Springer Berlin Heidelberg, 2007.

DALIANIS, H. Characteristics of Patient Records and Clinical Corpora. Em: **Clinical Text Mining: Secondary Use of Electronic Patient Records**. Cham: Springer International Publishing, 2018. p. 21–34.

DARJI, H.; MITROVIĆ, J.; GRANITZER, M. **German BERT Model for Legal Named Entity Recognition**. Proceedings of the 15th International Conference on Agents and Artificial Intelligence. **Anais**...SCITEPRESS - Science; Technology Publications, 2023.

DARPA (ED.). **Proceedings of the 3rd Message Understanding Conference (MUC-3)**. San Diego, EUA: Morgan Kaufmann, 1991.

DE OLIVEIRA, J. M.; ANTUNES, R. S.; DA COSTA, C. A. SOAP classifier for free-text clinical notes with domain-specific pre-trained language models. **Expert Systems with Applications**, v. 245, p. 123046, 2024.

DE PAIVA, V.; RADEMAKER, A.; MELO, G. DE. **OpenWordNet-PT: An Open**

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Brazilian Wordnet for Reasoning. Proceedings of COLING 2012: Demonstration Papers. **Anais**...2012.

DE SOUSA, S. C.; AZIZ, W.; SPECIA, L. **Assessing the post-editing effort for automatic and semi-automatic translations of DVD subtitles**. Proceedings of the International Conference Recent Advances in Natural Language Processing 2011. **Anais**...2011.

DE SOUZA, M. C. et al. Keywords attention for fake news detection using few positive labels. **Information Sciences**, v. 663, p. 120300, 2024.

DEERWESTER, S. et al. Indexing by latent semantic analysis. **Journal of the American society for information science**, v. 41, n. 6, p. 391–407, 1990.

DEJONG, G. Prediction and substantiation: A new approach to natural language processing. **Cognitive Science**, v. 3, n. 3, p. 251–273, 1979.

DEL CORRO, L.; GEMULLA, R. **Clausie: clause-based open information extraction**. Proceedings of the 22nd international conference on World Wide Web. **Anais**...: WWW '13.New York, NY, USA: ACM; ACM, 2013. Disponível em: <http://doi.acm.org/10.1145/2488388.2488420>

DELL'ORLETTA, F.; MONTEMAGNI, S.; VENTURI, G. Read-it: Assessing readability of italian texts with a view to text simplification. **Proceedings of the 2nd Workshop on Speech and Language Processing for Assistive Technologies**, p. 73–83, 2011.

DEMNER-FUSHMAN, D.; CHAPMAN, W. W.; MCDONALD, C. J. What can natural language processing do for clinical decision support? **J Biomed Inform**, v. 42, n. 5, p. 760–772, ago. 2009.

DESPOTOVIC, V. et al. Detection of COVID-19 from voice, cough and breathing patterns: Dataset and preliminary results. **Computers in Biology and Medicine**, v. 138, p. 104944, 2021.

DEVARAJU, A. et al. **FAIRsFAIR Data Object Assessment Metrics 0.5**. [s.l.] Research Data Alliance (RDA), out. 2020. Disponível em: <https://zenodo.org/record/6461229>.

DEVLIN, J. et al. **BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding**. (J. Burstein, C. Doran, T. Solorio, Eds.)Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019. **Anais**...Minneapolis, MN, USA: Association for Computational Linguistics, 2019. Disponível em: <https://doi.org/10.18653/v1/n19-1423>

DIAS, M. S. et al. **A qualitative analysis of a corpus of opinion summaries based on aspects**. Proceedings of the 1st Workshop on Tools and Resources for Automatically Processing Portuguese and Spanish. **Anais**...2014.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

DIAS-DA-SILVA, B. C.; MORALES, H. R. DE. A Construção de um Thesaurus Eletrônico para o Português. **Alfa**, 2003.

DIAZ, F.; MITRA, B.; CRASWELL, N. **Query Expansion with Locally-Trained Word Embeddings**. Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics. **Anais**...2016.

DODDINGTON, G. **Automatic Evaluation of Machine Translation Quality Using N-Gram Co-Occurrence Statistics**. Proceedings of the Second International Conference on Human Language Technology Research. **Anais**...: HLT '02.San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2002.

DODDINGTON, G. et al. **The Automatic Content Extraction (ACE) Program: Tasks, Data, and Evaluation**. (M. T. Lino et al., Eds.)Proceedings of LREC'2004, Fourth International Conference on Language resources and Evaluation (Lisboa, 26-28 May 2004). **Anais**...2004. Disponível em: <http://www.lrec-conf.org/proceedings/lrec2004/pdf/5.pdf>

DOHARE, S.; GUPTA, V.; KARNICK, H. **Unsupervised semantic abstractive summarization**. Proceedings of ACL 2018, Student Research Workshop. **Anais**...Melbourne, Australia: Association for Computational Linguistics (ACL), 2018.

DOHERTY, S. et al. **Mapping the industry I: Findings on translation technologies and quality assessment**. QTLaunchPad – Mapping the Industry I: Findings on Translation Technologies and Quality Assessment. **Anais**...GALA, 2013. Disponível em: <http://doras.dcu.ie/19474/1/Version_Participants_Final.pdf>. Acesso em: 11 nov. 2015

DOHERTY, S. et al. On Education and Training in Translation Quality Assessment. Em: MOORKENS, J. et al. (Eds.). **Translation Quality Assessment: From Principles to Practice**. Cham: Springer International Publishing, 2018. p. 95–106.

DONG, Z. et al. Exploring context window of large language models via decomposed positional vectors. **Advances in Neural Information Processing Systems**, v. 37, p. 10320–10347, 2024.

DORR, B. et al. Machine translation evaluation and optimization. Em: **Handbook of Natural Language Processing and Machine Translation: DARPA Global Autonomous Language Exploitation**. [s.l.] Springer, 2011. p. 745–843.

DOSOVITSKIY, A. et al. **An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale**. International Conference on Learning Representations. **Anais**...2021.

DOZIER, C. et al. Named Entity Recognition and Resolution in Legal Text. Em: FRANCESCONI, E. et al. (Eds.). **Semantic Processing of Legal Texts: Where the Language of Law Meets the Law of Language**. Berlin, Heidelberg: Springer Berlin Heidelberg, 2010. p. 27–43.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

DUBAY, W. Robert Gunning's Fog Readability Formula. **Plain Language At Work Newsletter**, v. 8, 2014.

DUBAY, W. H. **Smart Language: Readers, Readability, and the Grading of Text**. Costa Mesa, CA: Impact Information, 2007.

DUTRA, L. **Evaluating the Contribution of Framenet to Gender-Based Violence Identification: How semantic annotation can be used as a resource for identifying patterns of violence**. Dissertação (Mestrado em Tecnologia da Linguagem)— Gotemburgo: Universidade de Gotemburgo, 2024.

EARL, L. L. Experiments in automatic extracting and indexing. **Information Storage and Retrieval**, v. 6, n. 4, p. 313–330, 1970.

EDMUNDSON, H. P. New methods in automatic extracting. **Journal of the ACM (JACM)**, v. 16, n. 2, p. 264–285, 1969.

EFRON, B.; TIBSHIRANI, R. J. **An Introduction to the Bootstrap**. New York: Chapman & Hall/CRC, 1993.

EHRET, K. An Information-Theoretic View on Language Complexity and Register Variation: Compressing Naturalistic Corpus Data. **Corpus Linguistics and Linguistic Theory**, v. 17, n. 2, p. 383–410, out. 2021.

EHRET, K. et al. Measuring Language Complexity: Challenges and Opportunities. **Linguistics Vanguard**, v. 9, n. s1, p. 1–8, maio 2023.

EHRET, K.; SZMRECSANYI, B. An Information-Theoretic Approach to Assess Linguistic Complexity. Em: BAECHLER, R.; SEILER, G. (Eds.). **Complexity, Isolation, and Variation**. [s.l.] De Gruyter, 2016. p. 71–94.

EIGE. **European Institute for Gender Equality**. European Institute for Gender Equality, 2024. Disponível em: <https://eige.europa.eu/gender-based-violence/what-is-gender-based-violence>. Acesso em: 6 abr. 2024

EISENSTEIN, J. **Introduction to Natural Language Processing**. [s.l.] The MIT Press, 2019.

ELLIOT, N.; KLOBUCAR, A. Handbook of automated essay evaluation: Current applications and new directions. Em: SHERMIS, M. D.; BURSTEIN, J. (Eds.). [s.l.] Routledge/Taylor & Francis Group, 2013. p. 16–35.

ERDOĞAN, Y. E.; NARIN, A. COVID-19 detection with traditional and deep features on cough acoustic signals. **Computers in Biology and Medicine**, v. 136, p. 104765, 2021.

ERMAKOVA, L.; COSSU, J. V.; MOTHE, J. A survey on evaluation of summarization methods. **Information processing & management**, v. 56, n. 5, p. 1794–1814, 2019.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

ESTRADA, E. **The Structure of Complex Networks: Theory and Applications**. [s.l.] Oxford University Press, 2011.

ESTRELLA, P.; POPESCU-BELIS, A.; KING, M. The FEMTI guidelines for contextual MT evaluation: principles and resources. Em: WALTER DAELEMANS; VÉRONIQUE HOSTE (Eds.). **Evaluation of translation Technology**. Linguistica Antverpiensia new Series- themes em Translation Studies. [s.l: s.n.].

ETZIONI, O. et al. Unsupervised named-entity extraction from the web: An experimental study. **Artificial intelligence**, v. 165, n. 1, p. 91–134, 2005.

Euromatrix. Survey of Machine Translation Evaluation. [s.l.] Statistical; Hybrid Machine Translation Between All European Languages. Euromatrix, dez. 2007.

FACELI, K. et al. **Inteligência Artificial: Uma Abordagem de Aprendizado de Máquina**. 3. ed. Rio de Janeiro, Brasil: LTC, 2025.

FADER, A.; SODERLAND, S.; ETZIONI, O. **Identifying Relations for Open Information Extraction**. Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing. **Anais**...Edinburgh, Scotland, UK.: Association for Computational Linguistics, jul. 2011. Disponível em: <https://www.aclweb.org/anthology/D11-1142>

FAGHERAZZI, G. et al. Voice for health: the use of vocal biomarkers from research to clinical practice. **Digital biomarkers**, v. 5, n. 1, p. 78–88, 2021.

FAIR DATA MATURITY MODEL WORKING GROUP RDA. **FAIR Data Maturity Model. Specification and Guidelines**. Research Data Alliance; Zenodo, 2020. Disponível em: <https://doi.org/10.15497/rda00050>

FARHANGIAN, F.; CRUZ, R. M. O.; CAVALCANTI, G. D. C. Fake news detection: Taxonomy and comparative study. **Information Fusion**, v. 103, p. 102140, 2024.

FARIAS, D. S. et al. **Opinion-Meter: A Framework for Aspect-Based Sentiment Analysis**. Proceedings of the 22nd Brazilian Symposium on Multimedia and the Web. **Anais**...2016.

FARZINDAR, A.; INKPEN, D. **Natural Language Processing for Social Media**. Second edition ed. [s.l.] Morgan; Claypool, 2018.

FASEEH, M. et al. Hybrid approach to automated essay scoring: Integrating deep learning embeddings with handcrafted linguistic features for improved accuracy. **Mathematics**, v. 12, n. 21, p. 3416, 2024.

FAUSTINI, P. H. A.; COVÕES, T. F. Fake news detection in multiple platforms and languages. **Expert Systems with Applications**, v. 158, p. 113503, 2020.

FAUSTINI, P.; COVÕES, T. F. **Fake News Detection Using One-Class Classifica-**

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

tion. Proceedings of the 8th Brazilian Conference on Intelligent Systems (BRACIS'19). **Anais**...Salvador, BA, Brazil: IEEE, out. 2019.

FAWCETT, T. An introduction to ROC analysis. **Pattern recognition letters**, v. 27, n. 8, p. 861–874, 2006.

FEDERICO, M. et al. **Assessing the Impact of Translation Errors on Machine Translation Quality with Mixed-effects Models**. Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). **Anais**...Doha, Qatar: Association for Computational Linguistics, out. 2014. Disponível em: <https://aclanthology.org/D14-1172>

FELLBAUM, C. **WordNet: An Electronic Lexical Database**. [s.l.] The MIT Press, 1998.

FELTRIM, V. D. et al. **A Construção de uma Ferramenta de Auxílio à Escrita de Resumos Acadêmicos em Português**. Anais do Encontro Nacional de Inteligência Artificial (ENIA'2003). **Anais**...SBC, 2003.

FELTRIN, G. R.; VIANNA, D.; SILVA, A. DA. **Um Estudo Sobre Métricas de Avaliação para Sumarização de Acórdãos**. Anais do XXXVIII Simpósio Brasileiro de Bancos de Dados. **Anais**...SBC, 2023.

FENNELLY, O. et al. Use of standardized terminologies in clinical practice: A scoping review. **Int J Med Inform**, v. 149, p. 104431, fev. 2021.

FERNANDES, J. M.; WON, M.; MARTINS, B. Speechmaking and the Selectorate: Persuasion in Nonpreferential Electoral Systems. **Comparative Political Studies**, v. 53, n. 5, p. 667–699, a2020.

FERNANDES, W. P. D. et al. Appellate court modifications extraction for Portuguese. **Artificial Intelligence and Law**, v. 28, n. 3, p. 327–360, b2020.

FERNANDES-SVARTMAN, F. et al. **Temporal prosodic cues for COVID-19 in Brazilian Portuguese speakers**. Proc. Speech Prosody 2022. **Anais**...2022.

FERRÁNDEZ, Ó. et al. Tackling HAREM's portuguese named entity recognition task with spanish resources. **Reconhecimento de entidades mencionadas em português: Documentação e actas do HAREM, a primeira avaliação conjunta na área. Linguateca (http://www. linguateca. pt/aval_conjunta/LivroHAREM/Cap11-SantosCardoso2007-Ferrandezetal. pdf)**, 2007.

FERREIRA, A. C. et al. Padrões linguísticos para detecção de ironia em múltiplos idiomas. **Revista Gestão & Tecnologia**, 2017.

FERREIRA MELLO, R. et al. **Towards automated content analysis of rhetorical structure of written essays using sequential content-independent features in Portuguese**. (A. F. Wise, R. Martinez-Maldonado, I. Hilliger, Eds.)LAK22 Conference

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Proceedings. **Anais**...United States of America: Association for Computing Machinery (ACM), 2022.

FERREIRA, R. et al. Towards Automatic Content Analysis of Rhetorical Structure in Brazilian College Entrance Essays. Em: [s.l: s.n.]. p. 162–167.

FERREIRA-PAIVA, L. et al. **A survey of data augmentation for audio classification**. Congresso Brasileiro de Automática-CBA. **Anais**...2022.

FIAD, R. S. Reescrita, dialogismo e etnografia. **Linguagem em (Dis) curso**, v. 13, p. 463–480, 2013.

FIELDS, J.; CHOVANEC, K.; MADIRAJU, P. A survey of text classification with transformers: How wide? how large? how long? how accurate? how expensive? how safe? **IEEE Access**, v. 12, p. 6518–6531, 2024.

FILLMORE, C. J. Frame semantics. Em: KOREA, T. L. S. OF (Ed.). **Linguistics in the Morning Calm**. Seoul: Hanshin, 1982.

FILLMORE, C. J. Frames and the Semantics of Understanding. **Quaderni di Semantica**, v. 6, n. 2, p. 222–254, 1985.

FILLMORE, C. J.; JOHNSON, C. R.; PETRUCK, M. R. L. Background to FrameNet. **International Journal of Lexicography**, v. 16, n. 3, p. 235–250, 2003.

FINATTO, M. J. B. Projeto PorPopular, frequência de verbos em português e no jornal popular popular brasileiro. Em: UFMS/LABORATÓRIO DE EDIÇÃO DA FALE-UFMG, E. DA (Ed.). **As Ciências do Léxico: lexicologia, lexicografia, terminologia**. 1. ed. [s.l.] Aparecida Negri Isquerdo; Maria Cândida Trindade da Costa de Seabra, 2012. v. VIp. 227–244.

FINATTO, M. J. B. Humanidades digitais e estudos históricos do léxico. **Domínios de Lingu@gem**, v. 17, p. e1769, 2023.

FINATTO, M. J. B.; ESTEVES, F. F.; VILLAR, G. S. Construindo uma terminologia de raiz: textos legislativos sob exploração terminológica. **Revista Platô**, v. 5, n. 9, 2022.

FINATTO, M. J. B.; PARAGUASSU, L. B. Acessibilidade textual e terminológica. 2022.

FINATTO, M. J.; GONÇALVES, M. F.; LAZZARI, R. Léxico e terminologia em um novo gênero textual do século XVIII: o manual para enfermeiros. **In: Natalia Terrón Vinagre & Jenny Brumme (orgs.) Emergencia de nuevos géneros textuales y terminología en la historia de los lenguajes de especialidad.**, 2023.

FINE, K. Truthmaker semantics. **A Companion to the Philosophy of Language**, p. 556–577, 2017.

FINGER, M. Técnicas de otimização da precisão empregadas no etiquetador Tycho Brahe.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Proceedings of the International Conference on the Computational treatment of Portuguese, PROPOR, 2000.

FINGER, M. Inteligência Artificial e os rumos do processamento do português brasileiro. **Estudos Avançados**, v. 35, p. 51–72, 2021.

FINGER, M. et al. **SPIRA-BM: Biomarkers for Respiratory Conditions by Audio Analysis via Artificial Intelligence**. Anais do XXV Simpósio Brasileiro de Computação Aplicada à Saúde. **Anais**...Porto Alegre, RS, Brasil: SBC, 2025. Disponível em: <https://sol.sbc.org.br/index.php/sbcas/article/view/35566>

FIRDAUS SOLIHIN, R. F. A., Indra Budi; MAKARIM, E. Advancement of information extraction use in legal documents. **International Review of Law, Computers & Technology**, v. 35, n. 3, p. 322–351, 2021.

FIRTH, J. R. A synopsis of linguistic theory, 1930-1955. Em: FIRTH, J. R. (Ed.). **Studies in Linguistic Analysis**. Oxford: Blackwell, 1957.

FISCHER, M. et al. **Identifying Fake News in Brazilian Portuguese**. Natural Language Processing and Information Systems: 27th International Conference on Applications of Natural Language to Information Systems, NLDB 2022, Valencia, Spain, June 15–17, 2022, Proceedings. **Anais**...Berlin, Heidelberg: Springer-Verlag, 2022. Disponível em: <https://doi.org/10.1007/978-3-031-08473-7_10>

FLORES, F. N.; MOREIRA, V. P.; HEUSER, C. A. **Assessing the impact of stemming accuracy on information retrieval**. International Conference on Computational Processing of the Portuguese Language. **Anais**...Springer, 2010.

FLORIAN, R. et al. **Named entity recognition through classifier combination**. Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003. **Anais**...2003.

FONSECA, A.; SADAT, F.; MASSÉ, A. B. **Identifying portuguese multiword expressions using different classification algorithms-a comparative analysis**. Proceedings of the 4th International Workshop on Computational Terminology (Computerm). **Anais**...2014.

FONSECA, E. R. et al. **Automatically Grading Brazilian Student Essays**. (A. Villavicencio et al., Eds.)Computational Processing of the Portuguese Language. **Anais**...Springer International Publishing, 2018.

FONT LLITJÓS, A.; CARBONELL, J. G.; LAVIE, A. **A framework for interactive and automatic refinement of transfer-based machine translation**. Proceedings of the 10th EAMT Conference: Practical applications of machine translation. **Anais**...Budapest, Hungary: European Association for Machine Translation, 2005. Disponível em: <https://aclanthology.org/2005.eamt-1.13>

FORCADA, M. L.; ÑECO, R. P. **Recursive hetero-associative memories for trans-**

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

lation. International Work-Conference on Artificial Neural Networks. **Anais**...Springer, 1997.

FORNACIARI, T.; POESIO, M. Automatic deception detection in Italian court cases. **Artif. Intell. Law**, v. 21, n. 3, p. 303–340, set. 2013.

FORTUNA, P. et al. **A Hierarchically-Labeled Portuguese Hate Speech Dataset**. Proceedings of the Third Workshop on Abusive Language Online. **Anais**...2019.

FORTUNA, P.; NUNES, S. A survey on automatic detection of hate speech in text. **ACM Computing Surveys (CSUR)**, 2018.

FOVE. **Fove Eye Tracker**., 2018. Disponível em: <https://www.getfove.com/>

FREITAS ARAUJO, A. V. DE. **Informação, comunicação e documento**. Brasília, DF; Rio de Janeiro, RJ: CAPES: UAB; Departamento de Biblioteconomia, FACC/UFRJ, 2018. p. 96

FREITAS, B. L.; SILVA, R. M.; ALMEIDA, T. A. Gaussian Mixture Descriptors Learner. **KNOWLEDGE-BASED SYSTEMS**, v. 188, p. 105039, 2019.

FREITAS, C. et al. **Vampiro que brilha... rá! Desafios na anotação de opinião em um corpus de resenhas de livros**. Proceedings of XI Encontro de Linguística de Corpus. **Anais**...2012.

FREITAS, C. Sobre a construção de um léxico da afetividade para o processamento computacional do português. **Revista Brasileira de Linguística Aplicada**, 2013.

FREITAS, L. A. DE et al. **Pathways for irony detection in tweets**. Proceedings of the Symposium on Applied Computing (SAC). **Anais**...2014.

FREITAS, L. A. DE. **Feature-level sentiment analysis applied to brazilian portuguese reviews**. tese de doutorado—[s.l.] Pontifícia Universidade Católica do Rio Grande do Sul, 2015.

FREITAS, L. A. DE; SANTOS, L. DOS; DEON, D. Padrões linguísticos para detecção de ironia em múltiplos idiomas. **Revista Eletrônica de Iniciação Científica em Computação**, 2020.

FULLER, C. et al. **An Analysis of Text-Based Deception Detection Tools**. Proceedings of the Twelfth Americas Conference on Information Systems. **Anais**...2006.

GALHARDI, C. P. et al. Fato ou Fake? Uma análise da desinformação frente à pandemia da COVID-19 no Brasil. **Ciência & Saúde Coletiva**, v. 25, p. 4201–4210, out. 2020.

GAMALLO, P.; GARCIA, M. **Multilingual open information extraction**. (F. Pereira et al., Eds.)Portuguese Conference on Artificial Intelligence. **Anais**...Cham: Springer; Springer International Publishing, 2015. Disponível em: <https://doi.org/10.1007/978-3-

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

319-23485-4_72>

GAMALLO, P.; GARCIA, M.; FERNÁNDEZ-LANZA, S. **Dependency-based open information extraction**. Proceedings of the joint workshop on unsupervised and semi-supervised learning in NLP. **Anais**...: ROBUS-UNSUP '12.Stroudsburg, PA, USA: Association for Computational Linguistics; Association for Computational Linguistics, 2012. Disponível em: <http://dl.acm.org/citation.cfm?id=2389961.2389963>

GAMBHIR, M.; GUPTA, V. Recent automatic text summarization techniques: a survey. **Artificial Intelligence Review**, v. 47, p. 1–66, 2017.

GAMON, M. et al. Handbook of automated essay evaluation: Current applications and new directions. Em: SHERMIS, M. D.; BURSTEIN, J. (Eds.). [s.l.] Routledge/Taylor & Francis Group, 2013. p. 251–266.

GAMONAL, M. A. **Copa 2014 FrameNet Brasil: Diretrizes para a Constituição de um Dicionário Eletrônico Trilíngue a partir da Análise de Frames da Experiência Turística**. Dissertação de Mestrado em Linguística—[s.l.] Universidade Federal de Juiz de Fora, 2013.

GAMONAL, M. A. **Modelagem linguístico-computacional de metonímias na base de conhecimento multilíngue (m.knob) da FrameNet Brasil**. Tese de Doutorado—Juiz de Fora, Brasil: Universidade Federal de Juiz de Fora, 2017.

GAMONAL, M. A. et al. **Audition: A Frame-Annotated Multimodal Dataset for Accessible Audiovisual Content**. 2025.

GAO, M. et al. **Human-like Summarization Evaluation with ChatGPT**., 2023. Disponível em: <https://arxiv.org/abs/2304.02554>

GARBIN, C. A. S. et al. Desafios do profissional de saúde na notificação da violência: obrigatoriedade, efetivação e encaminhamento. **Ciência & Saúde Coletiva**, v. 20, p. 1879–1890, 2015.

GARCIA, E. A. S. et al. **RoBERTaLexPT: A Legal RoBERTa Model pretrained with deduplication for Portuguese**. (P. Gamallo et al., Eds.)Proceedings of the 16th International Conference on Computational Processing of Portuguese - Vol. 1. **Anais**...Santiago de Compostela, Galicia/Spain: Association for Computational Linguistics, mar. a2024. Disponível em: <https://aclanthology.org/2024.propor-1.38>

GARCIA, G. L. et al. **Text Summarization and Temporal Learning Models Applied to Portuguese Fake News Detection in a Novel Brazilian Corpus Dataset**. (P. Gamallo et al., Eds.)Proceedings of the 16th International Conference on Computational Processing of Portuguese - Vol. 1. **Anais**...Santiago de Compostela, Galicia/Spain: Association for Computational Linguistics, mar. b2024. Disponível em: <https://aclanthology.org/2024.propor-1.9>

GARCIA, G. L.; AFONSO, L. C.; PAPA, J. P. **FakeRecogna: a new brazilian corpus**

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

for fake news detection. International Conference on Computational Processing of the Portuguese Language. **Anais**...Springer, 2022.

GARCIA-MOLINA, H. et al. Challenges in data crowdsourcing. **IEEE Transactions on Knowledge and Data Engineering**, v. 28, n. 4, p. 901–911, 2016.

GARCIA-MORENO, C.; WATTS, C. Violence against women: an urgent public health priority. **Bulletin of the World Health Organization**, v. 89, p. 2–2, 2011.

GARIJO, D.; POVEDA-VILLALÓN, M. Best Practices for Implementing FAIR Vocabularies and Ontologies on the Web. **CoRR**, v. abs/2003.13084, 2020.

GAUY, M. M. et al. **Discriminant Audio Properties In Deep Learning Based Respiratory Insufficiency Detection In Brazilian Portuguese**. Artificial Intelligence in Medicine: 21st International Conference on Artificial Intelligence in Medicine, AIME 2023, Portorož, Slovenia, June 12–15, 2023, Proceedings. **Anais**...Berlin, Heidelberg: Springer-Verlag, 2023. Disponível em: <https://doi.org/10.1007/978-3-031-34344-5_32>

GAUY, M. M.; FINGER, M. Acoustic models of Brazilian Portuguese Speech based on Neural Transformers. **arXiv preprint arXiv:2312.09265**, 2023.

GAUY, M.; FINGER, M. **Audio MFCC-gram Transformers for respiratory insufficiency detection in COVID-19**. Anais do XIII Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana. **Anais**...Porto Alegre, RS, Brasil: SBC, 2021. Disponível em: <https://sol.sbc.org.br/index.php/stil/article/view/17793>

GAZZOLA, M.; LEAL, S. E.; ALUISIO, S. M. **Predição da Complexidade Textual de Recursos Educacionais Abertos em Português**. Proceedings of the Brazilian Symposium in Information and Human Language Technology. **Anais**...2019.

GEMMEKE, J. F. et al. **Audio set: An ontology and human-labeled dataset for audio events**. 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). **Anais**...IEEE, 2017.

GEORGE, J. F.; KEANE, B. T. **Deception Detection by Disinterested Third-Party Observers**. Proceedings of the Credibility Assessment and Information Quality in Government and Business Symposium, 39th Hawaii International Conference on System Sciences (HICSS). **Anais**...Kauai, HI: 2006.

GEURGAS, R.; TESSLER, L. R. Automatic detection of fake tweets about the COVID-19 Vaccine in Portuguese. **Social Network Analysis and Mining**, v. 14, n. 1, p. 55, 8 mar. 2024.

GHANEM, B. et al. **IDAT at FIRE2019: Overview of the Track on Irony Detection in Arabic Tweets**. Proceedings of the 11th Forum for Information Retrieval Evaluation. **Anais**...2019.

GHOSH, A. et al. **SemEval-2015 Task 11: Sentiment Analysis of Figurative**

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Language in Twitter. Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015). **Anais**...2015.

GIBBS, R. W.; COLSTON, H. L. **The Risks and Rewards of Ironic Communication**. Say not to say: new perspectives on miscommunication. **Anais**...2001. Disponível em: <https://api.semanticscholar.org/CorpusID:12510370>

GILLIN, N. **One-shot Prompt for Language Variety Identification**. Proceedings of the Eleventh Workshop on NLP for Similar Languages, Varieties, and Dialects (VarDial 2024). **Anais**...2024.

GLAUBER, R. et al. **Challenges of an Annotation Task for Open Information Extraction in Portuguese**. (A. Villavicencio et al., Eds.)Computational Processing of the Portuguese Language. **Anais**...Cham: Springer International Publishing, 2018.

GLAUBER, R.; CLARO, D. B. A systematic mapping study on open information extraction. **Expert Systems with Applications**, v. 112, p. 372–387, 2018.

GLAUBER, R.; CLARO, D. B.; OLIVEIRA, L. S. **Dependency Parser on Open Information Extraction for Portuguese Texts - DptOIE and DependentIE on IberLEF**. Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2019) co-located with 35th Conference of the Spanish Society for Natural Language Processing (SEPLN 2019). **Anais**...http://ceur-ws.org/Vol-2421/: CEUR Workshop Proceedings, a2019.

GLAUBER, R.; CLARO, D. B.; SENA, C. F. DE L. **Towards a Pragmatic Open Information Extraction for Portuguese Text - ICEIS17, InferPortOIE and PragmaticOIE on IberLEF**. Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2019) co-located with 35th Conference of the Spanish Society for Natural Language Processing (SEPLN 2019). **Anais**...http://ceur-ws.org/Vol-2421/: CEUR Workshop Proceedings, b2019.

GÔLO, M. P. S. et al. One-class learning for fake news detection through multimodal variational autoencoders. **Engineering Applications of Artificial Intelligence**, v. 122, p. 106088, 2023.

GÔLO, M. P. S. et al. **On the use of Large Language Models to Detect Brazilian Politics Fake News**. Proceedings of the 21st National Meeting on Artificial and Computational Intelligence (ENIAC'2024). **Anais**...Belém, PA, Brazil: Brazilian Computer Society, nov. 2024.

GOMES, J. R. S. et al. **Deep Learning Brasil at ABSAPT 2022: Portuguese Transformer Ensemble Approaches**. Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2022) co-located with the Conference of the Spanish Society for Natural Language Processing (SEPLN 2022), A Coruña, Spain, September 20, 2022. **Anais**...2022.

GONÇALO OLIVEIRA, H. et al. **Avaliação à medida no Segundo HAREM**. (C. Mota, D. Santos, Eds.)Desafios na avaliação conjunta do reconhecimento de entidades mencionadas: O Segundo HAREM. **Anais**...Linguateca, 2008.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

GONÇALVES, T. et al. Clinical Screening Prediction in the Portuguese National Health Service: Data Analysis, Machine Learning Models, Explainability and Meta-Evaluation. **Future Internet**, v. 15, n. 1, p. 26, 2023.

GONÇALVES, T.; QUARESMA, P. **A preliminary approach to the multilabel classification problem of Portuguese juridical documents**. Portuguese Conference on Artificial Intelligence. **Anais**...Springer, 2003.

GORDEEFF, E. M. Avaliação sobre Animação e Cinema de vida real: semelhanças e diferenças. **Diálogo com a Economia Criativa**, v. 8, n. 24, p. 50–63, 2023.

GORMAN, K.; BEDRICK, S. **We Need to Talk about Standard Splits**. (A. Korhonen, D. Traum, L. Màrquez, Eds.)Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. **Anais**...Florence, Italy: Association for Computational Linguistics, jul. 2019. Disponível em: <https://aclanthology.org/P19-1267/>

GOU, J. et al. Knowledge distillation: A survey. **International journal of computer vision**, v. 129, n. 6, p. 1789–1819, 2021.

GOURISARIA, M. K. et al. Comparative analysis of audio classification with MFCC and STFT features using machine learning techniques. **Discover Internet of Things**, v. 4, n. 1, p. 1, 2024.

GRAESSER, A. C. et al. Coh-Metrix: Analysis of text on cohesion and language. **Behavior Research Methods, Instruments, n Computer - Springer**, p. 193–202, 2004.

GRAESSER, A. C.; MCNAMARA, D. S.; KULIKOWICH, J. M. Coh-Metrix: Providing Multilevel Analyses of Text Characteristics. **Educational Researcher Vol. 40, N. 5**, p. 223–234, 2011.

GRAHAM, Y. et al. **Is all that Glitters in Machine Translation Quality Estimation really Gold?** Proceedings of COLING 2016: Technical Papers. **Anais**...Osaka, Japan: The COLING 2016 Organizing Committee, dez. 2016. Disponível em: <https://www.aclweb.org/anthology/C16-1294>

GRATTAFIORI, A. et al. **The Llama 3 Herd of Models**., 2024. Disponível em: <https://arxiv.org/abs/2407.21783>

GRAVES, A.; MOHAMED, A.; HINTON, G. **Speech recognition with deep recurrent neural networks**. 2013 IEEE International Conference on Acoustics, Speech and Signal Processing. **Anais**...2013.

GREENBERG, J. H. Some universals of grammar with particular reference to the order of meaningful elements. Em: GREENBERG, J. H. (Ed.). **Universals of Grammar**. 2. ed. Cambridge, Mass.: MIT Press, 1966. p. 73–113.

GRICE, H. P. Logic and Conversation. Em: **Syntax and Semantics: Vol. 3: Speech Acts**. [s.l.] Academic Press, 1975.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

GRISHMAN, R.; SUNDHEIM, B. **Message Understanding Conference- 6: A Brief History**. COLING 1996 Volume 1: The 16th International Conference on Computational Linguistics. **Anais**...1996. Disponível em: <https://aclanthology.org/C96-1079>

GRUBER, A.; WEISS, Y.; ROSEN-ZVI, M. **Hidden Topic Markov Models**. Proceedings of the Eleventh International Conference on Artificial Intelligence and Statistics. **Anais**...: Proceedings of Machine Learning Research.San Juan, Puerto Rico: PMLR, mar. 2007.

GRUPPI, M.; HORNE, B. D.; ADALI, S. NELA-GT-2019: A Large Multi-Labelled News Dataset for The Study of Misinformation in News Articles. **CoRR**, v. abs/2003.08444, p. 1–5, 2020.

GRUPPI, M.; HORNE, B. D.; ADALI, S. NELA-GT-2020: A Large Multi-Labelled News Dataset for The Study of Misinformation in News Articles. **CoRR**, v. abs/2102.04567, p. 1–6, 2021.

GU, J. et al. **A Survey on LLM-as-a-Judge**., 2025. Disponível em: <https://arxiv.org/abs/2411.15594>

GUARINO, N.; GUIZZARDI, G. **We need to Discuss the Relationship: Revisiting Relationships as Modeling Constructs**. Proceedings of the 27th International Conference on Advanced Information Systems Engineering (CAISE 2015). **Anais**...Springer-Verlag, 2015.

GUIMARÃES, G. M. C. et al. Legal Document Segmentation and Labeling Through Named Entity Recognition Approaches. **Journal of Information and Data Management**, v. 15, n. 1, 2024.

GUIMARÃES, J. A. C.; SANTOS, J. C. G. A ementa jurisprudencial como resumo informativo em um domínio especializado: aspectos estruturais. **Brazilian Journal of Information Science: research trends**, v. 10, n. 3, 2016.

GUIMARÃES, S. S. et al. **Characterizing Toxicity on Facebook Comments in Brazil**. Proceedings of the Brazilian Symposium on Multimedia and the Web. **Anais**...2020.

GUIZZARDI, G. Ontology, Ontologies and the "I" of FAIR. **Data Int.**, v. 2, n. 1-2, p. 181–191, 2020.

GULDEN, C. et al. Extractive summarization of clinical trial descriptions. **International Journal of Medical Informatics**, v. 129, p. 114–121, 2019.

GUMIEL, Y. B. et al. Temporal Relation Extraction in Clinical Texts: A Systematic Review. v. 54, n. 7, set. 2021.

GUPTA, S.; GUPTA, S. Abstractive summarization: An overview of the state of the art. **Expert Systems with Applications**, v. 121, p. 49–65, 2019.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

HABIBI, M. et al. Deep learning with word embeddings improves biomedical named entity recognition. **Bioinformatics**, v. 33, n. 14, p. i37–i48, 2017.

HAENDCHEN FILHO, A. et al. **An approach to evaluate adherence to the theme and the argumentative structure of essays**. International Conference on Knowledge-Based Intelligent Information & Engineering Systems. **Anais**...2018.

HAENDCHEN FILHO, A. et al. Imbalanced Learning Techniques for Improving the Performance of Statistical Models in Automated Essay Scoring. **Procedia Computer Science**, v. 159, p. 764–773, jan. 2019.

HAENLEIN, M.; KAPLAN, A. A brief history of artificial intelligence: On the past, present, and future of artificial intelligence. **California management review**, v. 61, n. 4, p. 5–14, 2019.

HAILU, T. T.; YU, J.; FANTAYE, T. G. A framework for word embedding based automatic text summarization and evaluation. **Information**, v. 11, n. 2, p. 78, 2020.

HAKUTA, K. Handbook of Automated Essay Evaluation: Current Applications and New Directions. Em: SHERMIS, M. D.; BURSTEIN, J. (Eds.). [s.l.] Routledge/Taylor & Francis Group, 2013. p. 347–353.

HAN, J. et al. **Exploring automatic COVID-19 diagnosis via voice and symptoms from crowdsourced data**. ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). **Anais**...IEEE, 2021.

HANLEY, J. A.; MCNEIL, B. J. The meaning and use of the area under a receiver operating characteristic (ROC) curve. **Radiology**, v. 143, n. 1, p. 29–36, 1982.

HARTMANN, N. S. et al. **Portuguese word embeddings: Evaluating on word analogies and natural language tasks**. Proceedings of Symposium in Information and Human Language Technology. **Anais**...[S.l.: s.n.]: 2017.

HARTMANN, N. S.; ALUÍSIO, S. M. Adaptação Lexical Automática em Textos Informativos do Português Brasileiro para o Ensino Fundamental. **Linguamática**, v. 12, n. 2, p. 3–27, dez. 2020.

HARTMANN PEIXOTO, F. Projeto Victor: relato do desenvolvimento da Inteligência Artificial na Repercussão Geral do Supremo Tribunal Federal. **Revista Brasileira de Inteligência Artificial e Direito - RBIAD**, v. 1, n. 1, p. 1–22, 2020.

HASEGAWA, T.; SEKINE, S.; GRISHMAN, R. **Discovering relations among named entities from large corpora**. Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (acl-04). **Anais**...2004.

HASSAN, A.; SHAHIN, I.; ALSABEK, M. B. **COVID-19 detection system using recurrent neural networks**. 2020 International conference on communications, computing, cybersecurity, and informatics (CCCI). **Anais**...IEEE, 2020.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

HASSAN, H. et al. Achieving Human Parity on Automatic Chinese to English News Translation. **arXiv preprint 1803.05567**, 2018.

HAUCH, V. et al. **Linguistic Cues to Deception Assessed by Computer Programs: A Meta-analysis**. Proceedings of the Workshop on Computational Approaches to Deception Detection. **Anais**...2012.

HE, K. et al. **Masked autoencoders are scalable vision learners**. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. **Anais**...2022.

HEARST, M. A. **Automatic acquisition of hyponyms from large text corpora**. Proceedings of the 14th conference on Computational linguistics-Volume 2. **Anais**...Association for Computational Linguistics, 1992.

HEE, C. V.; LEFEVER, E.; HOSTE, V. **SemEval-2018 Task 3: Irony Detection in English Tweets**. Proceedings of the 12th International Workshop on Semantic Evaluation. **Anais**...2018.

HEINRICH, T.; MARCHI, F. **TeamUFPR at ABSAPT 2022: Aspect Extraction with CRF and BERT**. Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2022) co-located with the Conference of the Spanish Society for Natural Language Processing (SEPLN 2022), A Coruña, Spain, September 20, 2022. **Anais**...2022.

HEUSDEN, R. VAN; KAMPS, J.; MARX, M. Neural Coreference Resolution for Dutch Parliamentary Documents with the DutchParliament Dataset. **Data**, v. 8, n. 2, 2023.

HITZLER, A. H. Y. A. M. S. M. A. G. B. A. W. R. A. A. S. A. A. H. M. A. K. T. A. J. M. M. A. A. M. A. J. P. A. P. Multimodal mental health analysis in social media. **PLOS ONE**, v. 15, n. 4, p. 1–27, 2020.

HOCHREITER, S.; SCHMIDHUBER, J. Long Short-Term Memory. **Neural Computation**, v. 9, n. 8, p. 1735–1780, nov. 1997.

HOCKETT, C. F. **A Course in Modern Linguistics**. [s.l.] Macmillan, 1958.

HOFMANN, T. **Probabilistic Latent Semantic Indexing**. Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '99). **Anais**...New York, NY, USA: Association for Computing Machinery, 1999.

HOVY, E.; KING, M.; POPESCU-BELIS, A. **An introduction to MT evaluation**. Proceedings of Machine Translation Evaluation: Human Evaluators meet Automated Metrics. Workshop at the LREC 2002 Conference. Las Palmas, Spain. **Anais**...2002.

HSU, W.-N. et al. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. **IEEE/ACM Transactions on Audio, Speech, and Language Processing**, v. 29, p. 3451–3460, 2021.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

HU, M.; LIU, B. **Mining Opinion Features in Customer Reviews**. Proceedings of the 19th National Conference on Artifical Intelligence. **Anais**...2004.

HUA, Y.; DENG, Z.; MCKEOWN, K. **Improving Long Dialogue Summarization with Semantic Graph Representation**. (A. Rogers, J. Boyd-Graber, N. Okazaki, Eds.)Findings of the Association for Computational Linguistics: ACL 2023. **Anais**...Toronto, Canada: Association for Computational Linguistics, jul. 2023. Disponível em: <https://aclanthology.org/2023.findings-acl.871/>

HUANG, P.-Y. et al. Masked autoencoders that listen. **Advances in Neural Information Processing Systems**, v. 35, p. 28708–28720, 2022.

HUSSAIN, A. S.; THOMAS, A. **Large Language Models for Judicial Entity Extraction: A Comparative Study**., 2024. Disponível em: <https://arxiv.org/abs/2407.05786>

HUTCHINS, J. **Towards a definition of example-based machine translation**., Proceedings of Second Workshop on Example-Based Machine Translation; **Anais**...2005.

HUTCHINS, W. Machine Translation: A Concise History. **Journal of Translation Studies: Special Issue on The Teaching of Computer-aided Translation**, v. 13, p. 1–2, 2010.

HUTCHINS, W. J. Machine translation over fifty years. **Histoire, Epistemologie, Langage**, v. XXII, n. 1, p. 7–31, 2001.

IFTIKHAR, A.; UL QOUNAIN JAFFRY, S. W.; MALIK, M. K. Information Mining From Criminal Judgments of Lahore High Court. **IEEE Access**, v. 7, p. 59539–59547, 2019.

IMOTIONS. **Eye Tracking - The Complete Pocket Guide**. [s.l.] www.imotions.com, 2017.

INÁCIO, M. L.; CASELI, H. DE M. **Word Embeddings at Post-Editing**. (P. Quaresma et al., Eds.)Computational Processing of the Portuguese Language. **Anais**...Cham: Springer International Publishing, 2020.

INDYK, P.; MOTWANI, R. **Approximate nearest neighbors: towards removing the curse of dimensionality**. Proceedings of the Thirtieth Annual ACM Symposium on Theory of Computing (STOC). **Anais**...Dallas, Texas, USA: ACM, 1998.

IOFFE, S.; SZEGEDY, C. **Batch normalization: Accelerating deep network training by reducing internal covariate shift**. International conference on machine learning. **Anais**...PMLR, 2015.

IPM. INAF Brasil 2018: Indicador de Alfabetismo Funcional - Resultados Preliminares. **Instituto Paulo Montenegro**, 2018.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

JACOBS, R. A. et al. Adaptive mixtures of local experts. **Neural computation**, v. 3, n. 1, p. 79–87, 1991.

JACOBSEN, A. et al. **FAIR principles: interpretations and implementation considerations**. **Data intelligence**MIT Press One Rogers Street, Cambridge, MA 02142-1209, USA journals-info …, 2020.

JAHAN, M. S.; OUSSALAH, M. A systematic review of hate speech automatic detection using natural language processing. **Neurocomputing**, 2023.

JARGAS, A. M. **Expressões Regulares - 5a edição: Uma Abordagem Divertida**. [s.l.] Novatec Editora, 2016.

JÄRVELIN, K.; KEKÄLÄINEN, J. Cumulated gain-based evaluation of IR techniques. **ACM Transactions on Information Systems (TOIS)**, v. 20, n. 4, p. 422–446, 2002.

JERONIMO, C. et al. Characterization of Fake News Based on Subjectivity Lexicons. **Journal of Data Intelligence**, v. 1, p. 419–441, dez. 2020.

JIANG, A. Q. et al. **Mistral 7B.**, 2023. Disponível em: <https://arxiv.org/abs/2310.06825>

JIANG, S. et al. Multi-Ontology Refined Embeddings (MORE): A hybrid multi-ontology and corpus-based semantic representation model for biomedical concepts. **Journal of Biomedical Informatics**, v. 111, p. 103581, 2020.

JIANG, S. et al. **Irony Detection in the Portuguese Language using BERT**. Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2021) co-located with the Conference of the Spanish Society for Natural Language Processing (SEPLN 2021), XXXVII International Conference of the Spanish Society for Natural Language Processing., Málaga, Spain, September, 2021. **Anais**...2021.

JOHNSTONE, M. A.-M. A. T. In an Absolute State: Elevated Use of Absolutist Words Is a Marker Specific to Anxiety, Depression, and Suicidal Ideation. **Clinical Psychological Science**, v. 6, n. 4, p. 529–542, 2018.

JONES, K. H. et al. Toward the Development of Data Governance Standards for Using Clinical Free-Text Data in Health Research: Position Paper. **J Med Internet Res**, v. 22, n. 6, p. e16760, jun. 2020.

JONES, K. S. What might be in a summary? **Information retrieval**, v. 93, n. 1, p. 9–26, 1993.

JONNALAGADDA, S.; GONZALEZ, G. Biosimplify: an open source sentence simplification engine to improve recall in automatic biomedical information extraction. **AMIA Annual Symposium Proceedings**, p. 351–356, 2010.

JULIÃO, A. **Algoritmo do Google: Veja o impacto que tem no SEO**. Disponível

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

em: <https://blog.ajestrategia.com.br/algoritmo-do-google-veja-o-impacto-que-tem-no-seo/>.

JUOLA, P. Measuring Linguistic Complexity: The Morphological Tier. **Journal of Quantitative Linguistics**, v. 5, n. 3, p. 206–213, dez. 1998.

JUOLA, P. Assessing Linguistic Complexity. Em: MIESTAMO, M.; SINNEMÄKI, K.; KARLSSON, F. (Eds.). **Studies in Language Companion Series**. Amsterdam: John Benjamins Publishing Company, 2008. v. 94p. 89–108.

JURAFSKY, D.; MARTIN, J. H. **Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition**. 3rd. ed. USA: Prentice Hall PTR, 2023.

JURAFSKY, D.; MARTIN, J. H. Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition. Em: 3rd. ed. [s.l.] Stanford University, 2026.

JUSTIÇA - CNJ, C. N. DE. **Conselho Nacional de Justiça — Justiça em Números**. https://www.cnj.jus.br/pesquisas-judiciarias/justica-em-numeros/, maio 24DC.

KAHLE, P. et al. **Transkribus-a service platform for transcription, recognition and retrieval of historical documents**. 2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR). **Anais**...IEEE, 2017.

KAMBHATLA, N. **Combining lexical, syntactic, and semantic features with maximum entropy models for information extraction**. Proceedings of the ACL interactive poster and demonstration sessions. **Anais**...2004.

KAPETANIDIS, P. et al. Respiratory diseases diagnosis using audio analysis and artificial intelligence: a systematic review. **Sensors**, v. 24, n. 4, p. 1173, 2024.

KAPOOR, A. et al. **HLDC: Hindi Legal Documents Corpus**. Findings of the Association for Computational Linguistics: ACL 2022. **Anais**...Association for Computational Linguistics, 2022.

KARELITZ, J. L.; MICHAEL, V. C.; PERKINS, K. A. Analysis of agreement between expired-air carbon monoxide monitors. **Journal of smoking cessation**, v. 2, n. 12, p. 105–112, 2017.

KATCHAPAKIRIN, K. et al. **Facebook Social Media for Depression Detection in the Thai Community**. 15th International Joint Conference on Computer Science and Software Engineering (JCSSE). **Anais**...2018.

KHAYRALLAH, H.; KOEHN, P. **On the Impact of Various Types of Noise on Neural Machine Translation**. Proceedings of the 2nd Workshop on Neural Machine Translation and Generation. **Anais**...Melbourne, Australia: Association for Computational Linguistics, jul. 2018. Disponível em: <https://aclanthology.org/W18-2709>

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

KILGARRIFF, A.; AL., ET. PtTenTen: A Corpus for Portuguese Lexicography. Em: **Working with Portuguese Corpora**. [s.l: s.n.]. p. 111–130.

KIM, Y. **Convolutional Neural Networks for Sentence Classification**. Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). **Anais**...2014.

KINCAID, J. P. et al. Derivation of new readability formulas (automated readability index, fog count, and flesch reading ease formula) for Navy enlisted personnel. **Research Branch Report**, p. 8–75, 1975.

KIND, L. et al. Subnotificação e (in) visibilidade da violência contra mulheres na atenção primária à saúde. **Cadernos de Saúde Pública**, v. 29, p. 1805–1815, 2013.

KOEHN, P. et al. **Moses: Open Source Toolkit for Statistical Machine Translation**. Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions. **Anais**...Prague, Czech Republic: Association for Computational Linguistics, jun. 2007. Disponível em: <https://aclanthology.org/P07-2045>

KOEHN, P. **Statistical Machine Translation**. [s.l.] Cambridge University Press, 2009.

KOEHN, P. **Neural Machine Translation**. [s.l.] Cambridge University Press, 2020.

KOEHN, P.; OCH, F. J.; MARCU, D. **Statistical phrase-based translation**. Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - NAACL '03. **Anais**...Association for Computational Linguistics, 2003. Disponível em: <http://dx.doi.org/10.3115/1073445.1073462>

KOHAVI, R. **A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection**. Proceedings of the 14th International Joint Conference on Artificial Intelligence (IJCAI). **Anais**...1995.

KOHO, M. et al. WarSampo Knowledge Graph: Finland in the Second World War as Linked Open Data. **Semantic Web – Interoperability, Usability, Applicability**, v. 12, n. 2, p. 265–278, 2021.

KOJIMA, T. et al. Large language models are zero-shot reasoners. **Advances in neural information processing systems**, v. 35, p. 22199–22213, 2022.

KOLECK, T. A. et al. Natural language processing of symptoms documented in free-text narratives of electronic health records: a systematic review. **J Am Med Inform Assoc**, v. 26, n. 4, p. 364–379, abr. 2019.

KONG, Q. et al. **Cross-Task Learning for Audio Tagging, Sound Event Detection and Spatial Localization: DCASE 2019 Baseline Systems**. [s.l.] DCASE2019 Challenge, 2019.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

KONG, Q. et al. Panns: Large-scale pretrained audio neural networks for audio pattern recognition. **IEEE/ACM Transactions on Audio, Speech, and Language Processing**, v. 28, p. 2880–2894, 2020.

KONSTANTINOVA, N. **Review of relation extraction methods: What is new out there?** Analysis of Images, Social Networks and Texts: Third International Conference, AIST 2014, Yekaterinburg, Russia, April 10-12, 2014, Revised Selected Papers 3. **Anais**...Springer, 2014.

KORNILOVA, A.; EIDELMAN, V. **BillSum: A Corpus for Automatic Summarization of US Legislation**. Proceedings of the 2nd Workshop on New Frontiers in Summarization. **Anais**...Association for Computational Linguistics, 2019.

KRINGS, H. P. **Repairing Texts: Empirical Investigations of Machine Translation Post-editing Processes**. [s.l.] Kent State University Press, 2001.

KULKARNI, M. et al. **Towards a Unified Multi-Domain Multilingual Named Entity Recognition Model**. Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics. **Anais**...Dubrovnik, Croatia: Association for Computational Linguistics, 2023. Disponível em: <https://aclanthology.org/2023.eacl-main.161>

KUMAR, D. et al. **Understanding the Behaviors of Toxic Accounts on Reddit**. Proceedings of the ACM Web Conference 2023. **Anais**...2023.

KUZI, S.; SHTOK, A.; KURLAND, O. **Query expansion using word embeddings**. Proceedings of the 25th ACM international on conference on information and knowledge management. **Anais**...2016.

LAFFERTY, J. D.; MCCALLUM, A.; PEREIRA, F. C. N. **Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data**. Proceedings of the Eighteenth International Conference on Machine Learning. **Anais**...: ICML '01.San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2001. Disponível em: <https://dl.acm.org/doi/abs/10.5555/645530.655813>

LAGE-FREITAS, A. et al. **Predicting Brazilian court decisions**. arXiv, 2019. Disponível em: <https://arxiv.org/abs/1905.10348>

LAGUARTA, J.; HUETO, F.; SUBIRANA, B. COVID-19 artificial intelligence diagnosis using only cough recordings. **IEEE Open Journal of Engineering in Medicine and Biology**, v. 1, p. 275–281, 2020.

LAMPLE, G. et al. **Neural Architectures for Named Entity Recognition**. (K. Knight, A. Nenkova, O. Rambow, Eds.)Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. **Anais**...San Diego, California: Association for Computational Linguistics, jun. 2016.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

LANCICHINETTI, A.; FORTUNATO, S. Consensus clustering in complex networks. **Scientific reports**, v. 2, n. 1, p. 336, 2012.

LANDRY, V. et al. Audio-based digital biomarkers in diagnosing and managing respiratory diseases: a systematic review and bibliometric analysis. **European Respiratory Review**, v. 34, n. 176, 2025.

LÄUBLI, S. et al. A set of recommendations for assessing human–machine parity in language translation. **Journal of Artificial Intelligence Research**, v. 67, p. 653–672, 2020.

LÄUBLI, S.; SENNRICH, R.; VOLK, M. **Has Machine Translation Achieved Human Parity? A Case for Document-level Evaluation**. Proceedings of EMNLP. **Anais**...Brussels, Belgium: 2018.

LAZER, D. M. J. et al. The science of fake news. **Science**, v. 359, n. 6380, p. 1094–1096, 2018.

LAZZARI, R. R.; FINATTO, M. J. B. Exame do vocabulário médico no Português no século XVIII: contribuições da lexicometria para o desenho de um dicionário histórico. **Mandinga-Revista de Estudos Linguísticos (ISSN: 2526-3455)**, v. 7, n. 1, p. 102–123, 2023.

LE, Q.; MIKOLOV, T. **Distributed representations of sentences and documents**. International conference on machine learning. **Anais**...PMLR, 2014.

LEACOCK, C. et al. **Automated Grammatical Error Detection for Language Learners**. [s.l.] Morgan; Claypool Publishers, 2010.

LEAL, S. E. et al. **Avaliação automática da complexidade de sentenças do português brasileiro para o domínio rural**. Symposium in Information and Human Language Technology - STIL. **Anais**...SBC, 2019.

LEAL, S. E. et al. **Using Eye-tracking Data to Predict the Readability of Brazilian Portuguese Sentences in Single-task, Multi-task and Sequential Transfer Learning Approaches**. Proceedings of the 28th International Conference on Computational Linguistics. **Anais**...Barcelona, Spain (Online): International Committee on Computational Linguistics, dez. 2020. Disponível em: <https://www.aclweb.org/anthology/2020.coling-main.512>

LEAL, S. E. **Predição da complexidade sentencial do português brasileiro escrito, usando métricas linguísticas, psicolinguísticas e de rastreamento ocular**. tese de doutorado—[s.l.] Universidade de São Paulo, 2021.

LEAL, S. E. et al. RastrOS Project: Natural Language Processing contributions to the development of an eye-tracking corpus with predictability norms for Brazilian Portuguese. **Language Resources and Evaluation**, p. 1333–1372, 2022.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

LEAL, S. E. et al. NILC-Metrix: assessing the complexity of written and spoken language in Brazilian Portuguese. **Language Resources and Evaluation**, 2023.

LEAL, S. E.; DURAN, M. S.; ALUÍSIO, S. M. **A Nontrivial Sentence Corpus for the Task of Sentence Readability Assessment in Portuguese**. Proceedings of the 27th International Conference on Computational Linguistics. **Anais**...Association for Computational Linguistics, ago. 2018.

LÉCHELLE, W.; GOTTI, F.; LANGLAIS, P. WiRe57: A Fine-Grained Benchmark for Open Information Extraction. **arXiv preprint arXiv:1809.08962**, 2018.

LEE, J. et al. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. **Bioinformatics**, v. 36, n. 4, p. 1234–1240, set. 2019.

LEE, S. et al. A Survey on Evaluation Metrics for Machine Translation. **Mathematics**, v. 11, n. 4, 2023.

LEHNERT, W.; SUNDHEIM, B. A performance evaluation of text-analysis technologies. **AI magazine**, v. 12, n. 3, p. 81–81, 1991.

LEITÃO, M. M.; RIBEIRO, A. J. C.; MAIA, M. Penalidade do Nome Repetido e Rastreamento Ocular em Português Brasileiro. **Revista LinguíStica**, v. v8 n2, 2012.

LEITE, H. et al. **WRITEME: uma Ferramenta de Auxílio à Escrita de READMEs Baseada em Dados Abertos**. Anais do XVII Congresso Latino-Americano de Software Livre e Tecnologias Abertas. **Anais**...Porto Alegre, RS, Brasil: SBC, 2020.

LEITNER, E.; REHM, G.; MORENO-SCHNEIDER, J. **Fine-Grained Named Entity Recognition in Legal Documents**. (M. Acosta et al., Eds.)Semantic Systems. The Power of AI and Knowledge Graphs. **Anais**...Cham: Springer International Publishing, 2019.

LESK, M. **The seven ages of information retrieval**., 1995. Disponível em: <https://archive.ifla.org/VI/5/op/udtop5/udt-op5.pdf>

LÉTOFFÉ, O.; HUANG, X.; MARQUES-SILVA, J. **Towards trustable SHAP scores**. Proceedings of the AAAI Conference on Artificial Intelligence. **Anais**...2025.

LEWIS, M. et al. **BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension**. (D. Jurafsky et al., Eds.)Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020. **Anais**...Association for Computational Linguistics, 2020. Disponível em: <https://doi.org/10.18653/v1/2020.acl-main.703>

LI, Q.; JI, H. **Incremental joint extraction of entity mentions and relations**. Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). **Anais**...2014.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

LI, Y. et al. **MAGE: Machine-generated Text Detection in the Wild**. (L.-W. Ku, A. Martins, V. Srikumar, Eds.)Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). **Anais**...Bangkok, Thailand: Association for Computational Linguistics, ago. 2024.

LIKERT, R. **A Technique for the Measurement of Attitudes**. [s.l.] Archives of Psychology, 1932.

LIMA, A. DA S.; BORGES, V. R. Training and evaluating Named Entity Recognition Models using a Legal Corpus of publications from Government Gazettes. 2022.

LIMA, J. P.; COSTA, J. A.; ARAÚJO, D. C. **Comparison of Feature Extraction Methods for Brazilian Legal Documents Clustering**. 2021 IEEE Latin American Conference on Computational Intelligence (LA-CCI). **Anais**...IEEE, 2021. Disponível em: <https://doi.org/10.1109/LA-CCI48322.2021.9769839>

LIMA, T. B. DE et al. Avaliação Automática de Redação: Uma revisáo sistemática. **Revista Brasileira de Informática na Educação**, v. 31, p. 205--221, maio 2023.

LIN, C. et al. SenseMood: Depression Detection on Social Media. Em: **2020 International Conference on Multimedia Retrieval**. New York, USA: Association for Computing Machinery, 2020a. p. 407–411.

LIN, C.-Y. **ROUGE: A Package for Automatic Evaluation of Summaries**. Text Summarization Branches Out. **Anais**...Barcelona, Spain: Association for Computational Linguistics, jul. 2004. Disponível em: <https://aclanthology.org/W04-1013>

LIN, H.; NG, V. Abstractive Summarization: A Survey of the State of the Art. **Proceedings of the AAAI Conference on Artificial Intelligence**, v. 33, n. 01, p. 9815–9822, 2019.

LIN, J.; NOGUEIRA, R.; YATES, A. Pretrained Transformers for Text Ranking: BERT and Beyond. **arXiv preprint arXiv:2010.06467**, b2020.

LIU, A. T. et al. **Mockingjay: Unsupervised speech representation learning with deep bidirectional transformer encoders**. ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). **Anais**...IEEE, a2020.

LIU, A. T.; LI, S.-W.; LEE, H. Tera: Self-supervised learning of transformer encoder representation for speech. **arXiv preprint arXiv:2007.06028**, b2020.

LIU, B. Sentiment Analysis and Opinion Mining. **Synthesis Lectures on Human Language Technologies**, 2012.

LIU, T.; YAO, J.-G.; LIN, C.-Y. **Towards improving neural named entity recognition with gazetteers**. Proceedings of the 57th annual meeting of the association for computational linguistics. **Anais**...a2019.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

LIU, X. et al. **Evaluating the Factuality of Large Language Models using Large-Scale Knowledge Graphs.**, 2024. Disponível em: <https://arxiv.org/abs/2404.00942>

LIU, Y. et al. **RoBERTa: A Robustly Optimized BERT Pretraining Approach.**, c2019. Disponível em: <https://arxiv.org/abs/1907.11692>

LIU, Y. et al. Roberta: A robustly optimized bert pretraining approach. **arXiv preprint arXiv:1907.11692**, b2019.

LIU, Y.; HEARNE, J.; CONRAD, B. **Recognizing proper names in ur iii texts through supervised learning**. The Twenty-Ninth International Flairs Conference. **Anais**...2016.

LIU, Y.; LAPATA, M. Text summarization with pretrained encoders. **arXiv preprint arXiv:1908.08345**, 2019.

LIU, Z. et al. De-identification of clinical notes via recurrent neural network and conditional random field. **J Biomed Inform**, v. 75S, p. S34–S42, jun. 2017.

LIU, Z. et al. **A Robustly Optimized BERT Pre-Training Approach with Post-Training**. Chinese Computational Linguistics: 20th China National Conference, CCL 2021, Hohhot, China, August 13–15, 2021, Proceedings. **Anais**...Berlin, Heidelberg: Springer-Verlag, 2021. Disponível em: <https://doi.org/10.1007/978-3-030-84186-7_31>

LIVESO. **O que é BERT? - O mais recente algoritmo da Google**. Disponível em: <https://liveseo.com.br/seo/o-que-e-bert-o-mais-recente-algoritmo-da-google/#:~:text=Bem%2C%20o%20BERT%2C%20de%20maneira,respostas%20possíveis%20para%20seus%20usuários>.

LO, C. **YiSi - a Unified Semantic MT Quality Evaluation and Estimation Metric for Languages with Different Levels of Available Resources**. Proceedings of the Fourth Conference on Machine Translation, WMT 2019, Florence, Italy, August 1-2, 2019 - Volume 2: Shared Task Papers, Day 1. **Anais**...2019. Disponível em: <https://doi.org/10.18653/v1/w19-5358>

LO, C.; WU, D. **MEANT: An inexpensive, high-accuracy, semi-automatic metric for evaluating translation utility based on semantic roles**. The 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, Proceedings of the Conference, 19-24 June, 2011, Portland, Oregon, USA. **Anais**...2011. Disponível em: <https://aclanthology.org/P11-1023/>

LO, S. L. et al. Multilingual Sentiment Analysis: From Formal to Informal and Scarce Resource Languages. **Artificial Intelligence Review**, 2017.

LOCATELLI, M. S. et al. **Examining the Behavior of LLM Architectures within the Framework of Standardized National Exams in Brazil**. Proceedings of the 2024 AAAI/ACM Conference on AI, Ethics, and Society. **Anais**...AAAI Press, 2025.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

LOMMEL, A.; MELBY, A. **Tutorial: MQM-DQF: A Good Marriage (Translation Quality for the 21st Century)**. Proceedings of the 13th Conference of the Association for Machine Translation in the Americas (Volume 2: User Track). **Anais**...Boston, MA: Association for Machine Translation in the Americas, mar. 2018. Disponível em: <https://aclanthology.org/W18-1925>

LOOI, Z. Q. et al. **MeLoDicA AI-Machine Learning Based Detection of Asthma via Vocal Audio Analysis**. 2024 IEEE Conference on Artificial Intelligence (CAI). **Anais**...IEEE, 2024.

LOPES, C. DE S. et al. Trend in the prevalence of depressive symptoms in Brazil: results from the Brazilian National Health Survey 2013 and 2019. **Cad Saude Publica**, 6 maio 2022.

LOPES, F.; TEIXEIRA, C.; GONÇALO OLIVEIRA, H. **Contributions to Clinical Named Entity Recognition in Portuguese**. Proceedings of the 18th BioNLP Workshop and Shared Task. **Anais**...Florence, Italy: Association for Computational Linguistics, ago. 2019. Disponível em: <https://www.aclweb.org/anthology/W19-5024>

LÓPEZ, R. et al. **A qualitative analysis of a corpus of opinion summaries based on aspects**. Proceedings of the 9th Linguistic Annotation Workshop. **Anais**...2015.

LOPEZ-GAZPIO, I. et al. Interpretable semantic textual similarity: Finding and explaining differences between sentences. **Knowledge-Based Systems**, v. 119, p. 186–199, 2017.

LORÈ, F. et al. An AI framework to support decisions on GDPR compliance. **Journal of Intelligent Information Systems**, p. 1–28, 2023.

LOSADA, D. E.; CRESTANI, F. **A Test Collection for Research on Depression and Language Use**. Experimental IR Meets Multilinguality, Multimodality, and Interaction. **Anais**...Cham: Springer, 2016.

LOSADA, D. E.; CRESTANI, F.; PARAPAR, J. **eRISK 2017: CLEF lab on early risk prediction on the internet: experimental foundations**. Lecture Notes in Computer Science vol 10456. **Anais**...Cham: Springer, 2017.

LOSADA, D. E.; CRESTANI, F.; PARAPAR, J. **Overview of eRisk: Early Risk Prediction on the Internet**. Lecture Notes in Computer Science vol 11018. **Anais**...Cham: Springer, 2018.

LOSADA, D. E.; CRESTANI, F.; PARAPAR, J. **Overview of eRisk 2019 Early Risk Prediction on the Internet**. Lecture Notes in Computer Science vol 11696. **Anais**...2019.

LOUIS, A.; HIGGINS, D. **Off-topic essay detection using short prompt texts**. Proceedings of the NAACL HLT 2010 Fifth Workshop on Innovative Use of NLP for Building Educational Applications. **Anais**...Los Angeles, California: Association for Computatio-

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

nal Linguistics, jun. 2010.

LOUIS, A.; NENKOVA, A. Automatically assessing machine summary content without a gold standard. **Computational Linguistics**, v. 39, n. 2, p. 267–300, 2013.

LOVEYS, K. et al. **Small but Mighty: Affective Micropatterns for Quantifying Mental Health from Social Media Language**. Fourth Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality. **Anais**...Vancouver, Canada: Association for Computational Linguistics, 2017.

LOVINS, J. B. Development of a stemming algorithm. **Mech. Transl. Comput. Linguistics**, v. 11, n. 1-2, p. 22–31, 1968.

LUCAS, J. et al. **Fighting Fire with Fire: The Dual Role of LLMs in Crafting and Detecting Elusive Disinformation**. (H. Bouamor, J. Pino, K. Bali, Eds.)Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. **Anais**...Singapore: Association for Computational Linguistics, dez. 2023. Disponível em: <https://aclanthology.org/2023.emnlp-main.883>

LUHN, H. P. The automatic creation of literature abstracts. **IBM Journal of research and development**, v. 2, n. 2, p. 159–165, 1958.

LUNDBERG, S. M.; LEE, S.-I. A unified approach to interpreting model predictions. **Advances in neural information processing systems**, v. 30, 2017.

MA, Q. et al. **Blend: a Novel Combined MT Metric Based on Direct Assessment - CASICT-DCU submission to WMT17 Metrics Task**. Proceedings of the Second Conference on Machine Translation, WMT 2017, Copenhagen, Denmark, September 7-8, 2017. **Anais**...2017. Disponível em: <https://doi.org/10.18653/v1/w17-4768>

MACDONALD, C.; TONELLOTTO, N. **Declarative Experimentation in Information Retrieval using PyTerrier**. Proceedings of ICTIR 2020. **Anais**...2020.

MACHADO, A. A. A. et al. **Personalitatem Lexicon: um léxico em português brasileiro para mineração de traços de personalidade em textos**. Proceedings of the Brazilian Symposium on Computers in Education. **Anais**...2015.

MACHADO, M. T.; PARDO, T. A. S. **NILC at ABSAPT 2022: Aspect Extraction for Portuguese**. Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2022) co-located with the Conference of the Spanish Society for Natural Language Processing (SEPLN 2022), A Coruña, Spain, September 20, 2022. **Anais**...2022.

MACHADO, M. T.; PARDO, T. A. S.; RUIZ, E. E. S. **Creating a portuguese context sensitive lexicon for sentiment analysis**. Proceedings of the 13th international conference on computational processing of the Portuguese Language (PROPOR). **Anais**...2018.

MACIEL, A. M. B. **Para o reconhecimento da especificidade do termo jurídico**. mathesis—[s.l.] Universidade Federal do Rio Grande do Sul, RS, 2001.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

MACOHIN, A.; CARNEIRO, J. V. V. Web Crawling e Web Scraping em sites de tribunais: publicidade processual e proteção de dados pessoais nas experiências europeia e brasileira. Em: WACHOWICZ, M. (Ed.). **Proteção de Dados Pessoais em Perspectiva: LGPD e RGPD na Ótica do Direito Comparado**. Curitiba: Gedai, UFPR, 2020.

MAIA, D. F. et al. **UlyssesSD-Br: Stance Detection in Brazilian Political Polls**. (G. Marreiros et al., Eds.)Progress in Artificial Intelligence. **Anais**...Cham: Springer International Publishing, 2022. Disponível em: <https://github.com/Dyonnatan/UlyssesSD-Br>

MAIA, M.; LEMLE, M.; FRANÇA, A. I. Efeito stroop e rastreamento ocular no processamento de palavras. **Ciências e Cognição 2007**, v. 12, p. 02–17, 2007.

MALENCHINI, F. M. et al. **Um Benchmark para Sistemas de Extração de Informação Aberta em Português**. Proceedings of theSymposium in Information and Human Language Technology (STIL 2019). **Anais**...Salvador, Bahia: SBC, out. 2019.

MALIK, V. et al. **ILDC for CJPE: Indian Legal Documents Corpus for Court Judgment Prediction and Explanation**. Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). **Anais**...Association for Computational Linguistics, 2021.

MALOUF, R. **A Comparison of Algorithms for Maximum Entropy Parameter Estimation**. COLING-02: The 6th Conference on Natural Language Learning 2002 (CoNLL-2002). **Anais**...2002. Disponível em: <https://aclanthology.org/W02-2018>

MALTA, D. et al. Prevalence of sexual violence in school children in Brazil: National School Health Survey 2019. **Population Medicine**, v. 5, abr. 2023.

MANDAL, A. et al. Unsupervised approaches for measuring textual similarity between legal court case reports. **Artificial Intelligence and Law**, v. 29, n. 3, p. 417–451, 2021.

MANI, I. Automatic Summarization. **John Benjamins Publishing Company**, v. 2, p. 399–408, 2001.

MANI, I.; MAYBURY, M. T. **Advances in automatic text summarization**. [s.l.] MIT press, 1999.

MANN, P.; MATSUSHIMA, E. H.; PAES, A. **Detecting Depression from Social Media Data as a Multiple-Instance Learning Task**. 10th International Conference on Affective Computing and Intelligent Interaction (ACII). **Anais**...2022.

MANN, P.; PAES, A.; MATSUSHIMA, E. H. **See and Read: Detecting Depression Symptoms in Higher Education Students Using Multimodal Social Media Data**. Proceedings of the International AAAI Conference on Web and Social Media. **Anais**...2020.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

MANN, W.; THOMPSON, S. Rethorical Structure Theory: Toward a functional theory of text organization. **Text**, v. 8, p. 243–281, jan. 1988.

MANNING, C. D.; RAGHAVAN, P.; SCHÜTZE, H. **Introduction to Information Retrieval**. New York, NY, USA: Cambridge University Press, 2008a.

MANNING, C. D.; SCHÜTZE, H. Foundations of Statistical Natural Language Processing. 1999.

MANNING, C. D.; SCHÜTZE, H.; RAGHAVAN, P. **Introduction to information retrieval**. [s.l.] Cambridge University Press Cambridge, 2008b.

MARCUSCHI, L. A. **Produção textual, análise de gêneros e compreensão**. [s.l.] Parábola Ed., 2008.

MARGARIDO, P. R. A. et al. **Automatic Summarization for Text Simplification: Evaluating Text Understanding by Poor Readers**. Companion Proceedings of the XIV Brazilian Symposium on Multimedia and the Web. **Anais**...: WebMedia '08.New York, NY, USA: ACM, 2008. Disponível em: <http://doi.acm.org/10.1145/1809980.1810057>

MARIE, B.; FUJITA, A.; RUBINO, R. Scientific Credibility of Machine Translation Research: A Meta-Evaluation of 769 Papers. **arXiv:2106.15195 [cs]**, jun. 2021.

MARINHO, J. et al. **Automated Essay Scoring: An approach based on ENEM competencies**. Anais do XIX Encontro Nacional de Inteligência Artificial e Computacional. **Anais**...SBC, a2022.

MARINHO, J.; ANCHIÊTA, R.; MOURA, R. Essay-BR: a Brazilian Corpus to Automatic Essay Scoring Task. **Journal of Information and Data Management**, v. 13, n. 1, p. 65–76, b2022.

MARON, M. E. Automatic indexing: an experimental inquiry. **Journal of the ACM (JACM)**, v. 8, n. 3, p. 404–417, 1961.

MARTINS, D. B. DE J. **Pós-edição automática de textos traduzidos automaticamente de inglês para português do Brasil**. Mestrado—São Carlos: Universidade Federal de São Carlos, 2014.

MARTINS, D. B. DE J.; CASELI, H. DE M. Automatic machine translation error identification. **Machine Translation**, v. 29, n. 1, p. 1–24, 2015.

MARTINS, R. T. et al. **An interlingua aiming at communication on the Web: How language-independent can it be?** NAACL-ANLP 2000 Workshop: Applied Interlinguas: Practical Applications of Interlingual Approaches to NLP. **Anais**...2000. Disponível em: <https://aclanthology.org/W00-0204>

MARTINS, T. B. F. et al. **Readability Formulas Applied to Textbooks in Brazilian**

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

Portuguese. [s.l.] ICMSC-USP, 1996.

MASCARENHAS, M. D. M. et al. Análise das notificações de violência por parceiro íntimo contra mulheres, Brasil, 2011-2017. **Revista Brasileira de Epidemiologia**, v. 23, p. e200007.SUPL.1, 2020.

MATHEUS GAUY, M. et al. Contrasting deep learning audio models for direct respiratory insufficiency detection versus blood oxygen saturation estimation. **Intelligence-Based Medicine**, v. 13, p. 100331, 2026.

MATOS, A. et al. **Accent classification is challenging but pre-training helps: a case study with novel brazilian portuguese datasets**. Proceedings of the 16th International Conference on Computational Processing of Portuguese-Vol. 1. **Anais**...2024.

MATTEI, L. D. et al. ATE ABSITA@ EVALITA2020: Overview of the Aspect Term Extraction and Aspect-based Sentiment Analysis Task. **Proceedings of the 7th Evaluation Campaign of Natural Language Processing and Speech tools for Italian (EVALITA 2020)**, 2020.

MATTHIESSEN, M. C. M. I. Applying systemic functional linguistics in healthcare contexts. **Text and Talk**, v. 33, n. 4-5, p. 437–447, 19 ago. 2013.

MATTHIESSEN, M. C. M. I.; TERUYA, K.; WU, C. Multilingual studies as a multidimensional space of interconnected language studies. Em: **Meaning in context : strategies for implementing intelligent applications of language studies**. [s.l.] Continuum, 2008. p. 146–221.

MAX, A. **Writing for Language-Impaired Readers**. In: Gelbukh A. (eds) Computational Linguistics and Intelligent Text Processing. CICLing 2006. Lecture Notes in Computer Science, vol 3878. **Anais**...Springer, Berlin, Heidelberg, 2006.

MAXWELL, K. T.; SCHAFER, B. Concept and context in legal information retrieval. Em: **Legal Knowledge and Information Systems**. [s.l.] IOS Press, 2008. p. 63–72.

MAYER, R. E. Elaboration techniques that increase the meaningfulness of technical text: An experimental test of the learning strategy hypothesis. **Journal of Educational Psychology**, v. 72, n. 6, p. 770–784, 1980.

MAYFIELD, E.; BLACK, A. W. **Should You Fine-Tune BERT for Automated Essay Scoring?** Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications. **Anais**...Association for Computational Linguistics, jul. 2020.

MAZIERO, E. G.; PARDO, T. A. S.; ALUÍSIO, S. M. Ferramenta de Análise Automática de Inteligibilidade de Córpus (AIC). **NILC - ICMC-USP**, 2008.

MCCALLUM, A.; LI, W. **Early results for named entity recognition with conditional random fields, feature induction and web-enhanced lexicons**. Proceedings

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

of the seventh conference on Natural language learning at HLT-NAACL 2003-Volume 4. **Anais**...2003.

MCNAMARA, D. S. et al. **Coh-Metrix Common Core T.E.R.A. version 1.0.**, 2013. Disponível em: <http://www.commoncoretera.com/>

MCNAMARA, D. S. et al. **Automated Evaluation of Text and Discourse with Coh-Metrix**. 1a. ed. [s.l.] Cambridge University Press, 2014.

MCWHORTER, J. H. The Worlds Simplest Grammars Are Creole Grammars. **Linguistic Typology**, v. 5, n. 2-3, jan. 2001.

MENDES, A. R.; CASELI, H. M. **Identifying Fine-grained Depression Signs in Social Media Posts**. Proceedings of the 2024 joint international conference on computational linguistics, language resources and evaluation (LREC 2024). **Anais**...2024.

MEYER, C. F. et al. The world wide web as linguistic corpus. Em: **Corpus Analysis**. [s.l.] Brill Rodopi, 2003. p. 241–254.

MIKOLOV, T. et al. **Distributed Representations of Words and Phrases and their Compositionality**. (C. J. Burges et al., Eds.)Advances in Neural Information Processing Systems. **Anais**...Curran Associates, Inc., b2013. Disponível em: <https://proceedings.neurips.cc/paper_files/paper/2013/file/9aa42b31882ec039965f3c4923ce901b-Paper.pdf>

MIKOLOV, T. et al. **Efficient Estimation of Word Representations in Vector Space**. Proceedings of the International Conference on Learning Representations (ICLR 2013) - Workshop Track. **Anais**...Scottsdale, Arizona, USA: a2013.

MILLER, G. A. WordNet: A Lexical Database for English. **Communications of the ACM**, v. Vol. 38, No. 11, p. 39–41, 1995.

MINAEE, S. et al. Deep learning–based text classification: a comprehensive review. **ACM computing surveys (CSUR)**, v. 54, n. 3, p. 1–40, 2021.

MIWA, M.; BANSAL, M. **End-to-End Relation Extraction using LSTMs on Sequences and Tree Structures**. Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). **Anais**...Association for Computational Linguistics, 2016.

MOHAN, S. et al. **The Impact of Toxic Language on the Health of Reddit Communities**. Proceedings of the Canadian Conference on AI. **Anais**...2017.

MOLLAS, I. et al. ETHOS: a multi-label hate speech detection dataset. **Complex & Intelligent Systems**, 2022.

MONTEIRO, R. A. et al. **Contributions to the Study of Fake News in Portuguese: New Corpus and Automatic Detection Results**. Proceedings of the 13th international conference on computational processing of the Portuguese Language. **Anais**...Canela,

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

Rio Grande do Sul, Brazil: Springer International Publishing, set. 2018.

MONTORO, A. F. **Curso de Teoria Geral do Direito - Aula 2: A linguagem do direito: semântica, sintática e pragmática**. Disponível em: <http://www.dialdata.com.br/ilam/aula2>.

MOORKENS, J. et al. Correlations of perceived post-editing effort with measurements of actual effort. **Machine Translation**, v. 29, n. 3/4, p. 267–284, 2015.

MOORKENS, J. Under pressure: translation in times of austerity. **Perspectives**, v. 25, n. 3, p. 464–477, fev. 2017.

MORENO, G. C. DE L. et al. ALT: um software para analise de legibilidade de textos em Lingua Portuguesa. **preprint**, 2022.

MORENO, J.; BRESSAN, G. **FACTCK.BR: a new dataset to study fake news**. : WebMedia '19.New York, NY, USA: Association for Computing Machinery, 2019. Disponível em: <https://doi.org/10.1145/3323503.3361698>

MORENO SCHNEIDER, J. et al. Lynx: A knowledge-based AI service platform for content processing, enrichment and analysis for the legal domain. **Information Systems**, v. 106, p. 101966, 2022.

MOTA, C. R3M, uma participação minimalista no Segundo HAREM. **quot; In Cristina Mota; Diana Santos (ed) Desafios na avaliação conjunta do reconhecimento de entidades mencionadas: O Segundo HAREM Linguateca 2008**, 2008.

MOTA, C. C. et al. **Reconhecimento de entidades nomeadas em documentos jurídicos em português utilizando redes neurais**. Encontro Nacional de Inteligência Artificial e Computacional (ENIAC). **Anais**...SBC, 2021.

MOTA, C.; SANTOS, D. (EDS.). **Desafios na avaliação conjunta do reconhecimento de entidades mencionadas: O Segundo HAREM**. [s.l.] Linguateca, 2008.

MOTA, C.; SANTOS, D.; RANCHHOD, E. Avaliação de reconhecimento de entidades mencionadas: princípio de HAREM. **Avaliação conjunta: um novo paradigma no processamento computacional da língua portuguesa**, p. 161–175, 2007.

MOTTA, E. Sentenças Judiciais e Acessibilidade Textual e Terminológica. **Domínios de Lingu@gem**, v. 15, n. 3, p. 761–813, 2021.

MOTTA, E. **SENTENÇAS JUDICIAIS E LINGUAGEM SIMPLES: um encontro possível e necessário**. mathesis—[s.l.] Universidade Federal do Rio Grande do Sul, RS, 2022.

MOUAWAD, P.; DUBNOV, T.; DUBNOV, S. Robust detection of COVID-19 in cough sounds: using recurrence dynamics and variable Markov model. **SN Computer Science**, v. 2, n. 1, p. 34, 2021.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

MOUJAHID, H. et al. Combining CNN and Grad-Cam for COVID-19 Disease Prediction and Visual Explanation. **Intelligent Automation & Soft Computing**, v. 32, n. 2, 2022.

MOUSAVI, E. **Complex Network Series: Part 1 — An Overview**., 2024. Disponível em: <https://medium.com/@ebimsv/complex-network-series-part-1-an-overview-ddd9e0b6ec3a>

MURPHY, K. P. **Machine learning: a probabilistic perspective**. [s.l.] MIT press, 2012.

NADEAU, D. **Semi-Supervised Named Entity Recognition: Learning to Recognize 100 Entity Types with Little Supervision**. tese de doutorado—[s.l.] University of Ottawa, 2007.

NAGAO, M. A Framework of a Mechanical Translation between Japanese and English by Analogy Principle. Em: NIRENBURG, S.; SOMERS, H. L.; WILKS, Y. A. (Eds.). **Readings in Machine Translation**. [s.l.] The MIT Press, 1984.

NAIR, S. S.; JEEVEN, V. A brief overview of metadata formats. **DESIDOC Journal of Library & Information Technology**, v. 24, n. 4, 2004.

NAIR, V.; HINTON, G. E. **Rectified linear units improve restricted boltzmann machines**. Icml. **Anais**...2010.

NARDE, W. **Análise de notícias falsas em rede social: uma abordagem utilizando transferência de aprendizagem e Transformers**. https://www.monografias.ufop.br/bitstream/35400000/3122/6/MONOGRAFIA_AnáliseNotíciasFalsas.pdf, 2021.

NASAR, Z.; JAFFRY, S. W.; MALIK, M. K. Named entity recognition and relation extraction: State-of-the-art. **ACM Computing Surveys (CSUR)**, v. 54, n. 1, p. 1–39, 2021.

NASCIMENTO, D. N. C. R. DO. **Sumarização de artigos científicos em português no domínio da saúde**. mathesis—[s.l.] (Mestrado em Informática) - Programa de Pós-Graduação em Informática da PUC-Rio, Rio de Janeiro, 2023.

NASCIMENTO, G. et al. **Hate speech detection using brazilian imageboards**. Proceedings of the 25th Brazillian Symposium on Multimedia and the Web. **Anais**...2019.

NASCIMENTO, R. DA S. et al. **Identificando Sinais de Comportamento Depressivo em Redes Sociais**. Anais do VII Brazilian Workshop on Social Network Analysis and Mining. **Anais**...Porto Alegre, Brazil: SBC, 2018.

NATH, N.; LEE, S.-H.; LEE, I. NEAR: Named Entity and Attribute Recognition of Clinical Concepts. **J. of Biomedical Informatics**, v. 130, n. C, jun. 2022.

NECO, R. P.; FORCADA, M. L. **Asynchronous translations with recurrent**

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

neural nets. Proceedings of International Conference on Neural Networks (ICNN'97). **Anais**...1997.

NETO, J. R. C. S. A. V. S.; FALEIROS, T. DE P. **Deep Active-Self Learning Applied to Named Entity Recognition**. (A. Britto, K. Valdivia Delgado, Eds.)Intelligent Systems. **Anais**...Cham: Springer International Publishing, 2021. Disponível em: <https://avio11.github.io/resources/aposentadoria/aposentadoria.html>

NEURALMIND. **NeuralMind disponibiliza modelo BERT do Google em português. Neuralmind blog**. Disponível em: <https://neuralmind.ai/2020/01/26/neuralmind-disponibiliza-modelo-bert-inteligencia-artificial-do-google-em-portugues/>.

NEWELL, E. et al. **Assessing the Verifiability of Attributions in News Text**. (G. Kondrak, T. Watanabe, Eds.)Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers). **Anais**...Taipei, Taiwan: Asian Federation of Natural Language Processing, nov. 2017. Disponível em: <https://aclanthology.org/I17-1076>

NEWMAN, M. E. J. **Networks: An Introduction**. Oxford: Oxford University Press, 2010.

NEWMAN, M. E. J.; GIRVAN, M. Finding and evaluating community structure in networks. **Physical Review E**, v. 69, n. 2, p. 026113, 2004.

NEWMAN, N. et al. **Reuters institute digital news report 2020**. [s.l.] Report of the Reuters Institute for the Study of Journalism, 2020.

NGUYEN, D. B.; THEOBALD, M.; WEIKUM, G. J-NERD: joint named entity recognition and disambiguation with rich linguistic features. **Transactions of the Association for Computational Linguistics**, v. 4, p. 215–229, 2016.

NICHOLS, J. **Linguistic Diversity in Space and Time**. [s.l.] University of Chicago Press, 1998.

NIIZUMI, D. et al. **Masked spectrogram modeling using masked autoencoders for learning general-purpose audio representation**. HEAR: Holistic Evaluation of Audio Representations. **Anais**...PMLR, 2022.

NIIZUMI, D. et al. **Masked modeling duo: Learning representations by encouraging both networks to model the input**. ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). **Anais**...IEEE, 2023.

NIIZUMI, D. et al. Towards Pre-training an Effective Respiratory Audio Foundation Model. **arXiv preprint arXiv:2505.15307**, 2025.

NOGUEIRA, R. et al. Document expansion by query prediction. **arXiv preprint arXiv:1904.08375**, 2019.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

NUNES, M. DAS G. V. et al. O uso de interlíngua para comunicação via Internet: a decodificação UNL-português. **Revista Tecnologia da Informação**, v. 3, n. 1, p. 49–55, 2003.

NUNES, R. **LegiSubject-Br-Keywords**. https://huggingface.co/datasets/ronunes/LegiSubject-Br-Keywords, 2024.

NUNES, R. O. et al. **Out of Sesame Street: A Study of Portuguese Legal Named Entity Recognition Through In-Context Learning**. INSTICC; SciTePress, a2024.

NUNES, R. O. et al. **A Named Entity Recognition Approach for Portuguese Legislative Texts Using Self-Learning**. (P. Gamallo et al., Eds.)Proceedings of the 16th International Conference on Computational Processing of Portuguese - Vol. 1. **Anais**...Santiago de Compostela, Galicia/Spain: Association for Computational Linguistics, mar. b2024. Disponível em: <https://aclanthology.org/2024.propor-1.30>

NUNES, R. O. **Data contamination in specialized named entity recognition corpora**. Dissertação de Mestrado—Porto Alegre, Brasil: Universidade Federal do Rio Grande do Sul, 2025.

O'BRIEN, S. Towards predicting post-editing productivity. **Machine translation**, v. 25, p. 197–215, 2011.

O'BRIEN, S. et al. **Dynamic Quality Evaluation Framework**. [s.l.] TAUS Labs Report. The Translation Automation User Society-TAUS, 2011.

OAEPUBLISH. **CES 2025 Article**., 2025. Disponível em: <https://www.oaepublish.com/articles/ces.2025.34>

OCH, F. J.; NEY, H. The Alignment Template Approach to Statistical Machine Translation. **Computational Linguistics**, v. 30, n. 4, p. 417–449, dez. 2004.

ÖHMAN, A. et al. The public health turn on violence against women: analysing Swedish healthcare law, public health and gender-equality policies. **BMC Public Health**, v. 20, p. 1–12, 2020.

OKANO, E. Y. et al. **Fake News Detection on Fake.Br Using Hierarchical Attention Networks**. (P. Quaresma et al., Eds.)Computational Processing of the Portuguese Language. **Anais**...Cham: Springer International Publishing, 2020.

OKANO, E. Y.; RUIZ, E. E. S. **Using linguistic cues to detect fake news on the brazilian portuguese parallel corpus Fake.br**. Proceedings of the Symposium in Information and Human Language Technology. **Anais**...Brazilian Computer Society, 2019.

OKSANEN, A. et al. Semantic Finlex: Transforming, Publishing, and Using Finnish Legislation and Case Law As Linked Open Data on the Web. Em: PERUGINELLI, G.; FARO, S. (Eds.). **Knowledge of the Law in the Big Data Age**. Frontiers em Artificial Intelligence e Applications. [s.l.] IOS Press, 2019b. v. 317p. 212–228.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

OKSANEN, A. et al. **ANOPPI: A Pseudonymization Service for Finnish Court Documents.** JURIX. **Anais**...a2019.

OKSANEN, A. et al. **An Anonymization Tool for Open Data Publication of Legal Documents**. Joint Proceedings of ISWC2022 Workshops. **Anais**...CEUR-WS. org, 2022. Disponível em: <https://ceur-ws.org/Vol-3257/>

OLIVAL, F.; CAMERON, H.; VIEIRA, R. **As Memórias Paroquiais: do manuscrito ao digital**. Actas da Jornada de Humanidades Digitais do CIDEHUS (to appear). **Anais**...2022.

OLIVEIRA, A. et al. **Toxic Text Classification in Portuguese: Is LLaMA 3.1 8B All You Need?** Proceedings of the 15th Brazilian Symposium in Information and Human Language Technology. **Anais**...a2024.

OLIVEIRA, I. L. Uma mentira repetida mil vezes se transforma em verdade? Reflexões sobre as dinâmicas discursivas e seus efeitos na saúde. Em: **Desinformação o mal do século: Distorções, inverdades, fake news: a democracia ameaçada**. [s.l: s.n.]. p. 299–315.

OLIVEIRA, L. E. S. et al. SemClinBr - a multi-institutional and multi-specialty semantically annotated corpus for Portuguese clinical NLP tasks. **Journal of Biomedical Semantics**, v. 13, n. 1, a2022.

OLIVEIRA, L. F. A. DE et al. **Challenges In Annotating A Treebank Of Clinical Narratives In Brazilian Portuguese**. Computational Processing of the Portuguese Language: 15th International Conference, PROPOR 2022, Fortaleza, Brazil, March 21–23, 2022, Proceedings. **Anais**...Berlin, Heidelberg: Springer-Verlag, b2022. Disponível em: <https://doi.org/10.1007/978-3-030-98305-5_9>

OLIVEIRA, L. F. DE; OLIVEIRA, O. N., Jr. The Landscape of Wearable Sensors and Automated Literature Analysis with Large-Language Models. **ACS Omega**, v. 10, n. 37, p. 42127–42134, 2025.

OLIVEIRA, L. M. DE; MOURA, M. F. DE; OLIVEIRA, O. N., Jr. Using network analysis and large-language models to obtain a landscape of the literature on dressing materials for wound healing: The predominance of chitosan and other biomacromolecules. **International Journal of Biological Macromolecules**, v. 306.3, p. 141565, 2025.

OLIVEIRA, L.; CLARO, D.; SOUZA, M. DptOIE: a Portuguese open information extraction based on dependency analysis. **Artificial Intelligence Review**, v. 56, p. 1–32, dez. c2022.

OLIVEIRA, N. et al. Processamento de Linguagem Natural para Identificação de Notícias Falsas em Redes Sociais: Ferramentas, Tendências e Desafios. Em: [s.l.] SBC, 2020.

OLIVEIRA, R. L. DE; MARTINS, J. T.; PARABONI, I. Mental health prediction from social media connections. **New Review of Hypermedia and Multimedia**, b2024.

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

OLIVEIRA, R. L. DE; PARABONI, I. **A Bag-of-Users approach to mental health prediction from social media data**. 16th International Conference on Computational Processing of Portuguese (PROPOR 2024). **Anais**...Santiago de Compostela, Spain: 2024.

OLIVEIRA, V. et al. Combining prompt-based language models and weak supervision for labeling named entity recognition on legal documents. **Artificial Intelligence and Law**, p. 1–21, fev. c2024.

OMOHUNDRO, S. M. **Five balltree construction algorithms**. Berkeley, CA, USA: International Computer Science Institute, 1989.

ONAGA, T.; FUJITA, M.; YOSHINOBU, K. **Japanese Legal Bar Problem Solver Focusing on Person Names**. Proceedings of the Tenth International Competition on Legal Information Extraction/Entailment (COLIEE 2023). **Anais**...2023. Disponível em: <https://sites.ualberta.ca/~rabelo/COLIEE2023>

ORENGO, V. M.; BURIOL, L. S.; COELHO, A. R. **A study on the use of stemming for monolingual ad-hoc Portuguese information retrieval**. Workshop of the Cross-Language Evaluation Forum for European Languages. **Anais**...Springer, 2006.

ORENGO, V. M.; HUYCK, C. **A Stemming Algorithmm for the Portuguese Language**. Proceedings Eighth Symposium on String Processing and Information Retrieval. **Anais**...IEEE Computer Society, 2001.

ORLANDIC, L.; TEIJEIRO, T.; ATIENZA, D. The COUGHVID crowdsourcing dataset, a corpus for the study of large-scale cough analysis algorithms. **Scientific Data**, v. 8, n. 1, p. 156, 2021.

OSTENDORFF, M. et al. **Evaluating document representations for content-based legal literature recommendations**. Proceedings of the Eighteenth International Conference on Artificial Intelligence and Law. **Anais**...2021. Disponível em: <https://doi.org/10.1145/3462757.3466073>

OTT, M. et al. **Finding Deceptive Opinion Spam by Any Stretch of the Imagination**. (D. Lin, Y. Matsumoto, R. Mihalcea, Eds.)Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies. **Anais**...Portland, Oregon, USA: Association for Computational Linguistics, jun. 2011. Disponível em: <https://aclanthology.org/P11-1032>

OUYANG, L. et al. Training language models to follow instructions with human feedback. **Advances in Neural Information Processing Systems (NeurIPS)**, v. 35, p. 27730–27744, 2022.

PACK, A.; BARRETT, A.; ESCALANTE, J. Large language models and automated essay scoring of English language learner writing: Insights into validity and reliability. **Computers and Education: Artificial Intelligence**, v. 6, p. 100234, 2024.

PAETZOLD, G. H.; SPECIA, L. **Unsupervised Lexical Simplification for Non-**

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

native Speakers. Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence. **Anais**...: AAAI'16.Phoenix, Arizona: AAAI Press, a2016. Disponível em: <http://dl.acm.org/citation.cfm?id=3016387.3016433>

PAETZOLD, G.; SPECIA, L. **Inferring Psycholinguistic Properties of Words**. NAACL HLT 2016, The 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, San Diego California, USA, June 12-17, 2016. **Anais**...b2016. Disponível em: <http://aclweb.org/anthology/N/N16/N16-1050.pdf>

PAETZOLD, G.; SPECIA, L. **Understanding the Lexical Simplification Needs of Non-Native Speakers of English**. (Y. Matsumoto, R. Prasad, Eds.)Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers. **Anais**...Osaka, Japan: The COLING 2016 Organizing Committee, dez. c2016. Disponível em: <https://aclanthology.org/C16-1069>

PAETZOLD, G.; SPECIA, L. **Lexical Simplification with Neural Ranking**. Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers. **Anais**...Valencia, Spain: Association for Computational Linguistics, abr. 2017. Disponível em: <http://www.aclweb.org/anthology/E17-2006>

PAIOLA, P. H. **Sumarização abstrativa de textos em português utilizando aprendizado de máquina**. mathesis—[s.l.] (Mestrado em Ciências da Computação) - Programa de Pós-Graduação em Ciência da Computação, Universidade Estadual Paulista "Júlio de Mesquita Filho", 2022.

PAIOLA, P. H. et al. **RecognaSumm: A Novel Brazilian Summarization Dataset**. (P. Gamallo et al., Eds.)Proceedings of the 16th International Conference on Computational Processing of Portuguese - Vol. 1. **Anais**...Santiago de Compostela, Galicia/Spain: Association for Computational Lingustics, mar. 2024. Disponível em: <https://aclanthology.org/2024.propor-1.63>

PAIOLA, P. H.; ROSA, G. H. DE; PAPA, J. P. **Deep Learning-Based Abstractive Summarization for Brazilian Portuguese Texts**. Intelligent Systems: 11th Brazilian Conference, BRACIS 2022, Campinas, Brazil, November 28 – December 1, 2022, Proceedings, Part II. **Anais**...Berlin, Heidelberg: Springer-Verlag, 2022. Disponível em: <https://doi.org/10.1007/978-3-031-21689-3_34>

PAIS, V. et al. **Named Entity Recognition in the Romanian Legal Domain**. Proceedings of the Natural Legal Language Processing Workshop 2021. **Anais**...Punta Cana, Dominican Republic: Association for Computational Linguistics, nov. 2021.

PAIS, V. et al. LegalNERo: A linked corpus for named entity recognition in the Romanian legal domain. **Semantic Web journal**, 2024.

PALACIO-NIÑO, J.-O.; BERZAL, F. **Enhancing Community Detection in Networks: A Comparative Analysis of Local Metrics and Hierarchical Algo-**

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

rithms. arXiv preprint, 2024. Disponível em: <https://arxiv.org/html/2408.09072v2>

PANWAR, H. et al. A deep learning and grad-CAM based color visualization approach for fast detection of COVID-19 cases using chest X-ray and CT-Scan images. **Chaos, Solitons & Fractals**, v. 140, p. 110190, 2020.

PAPA, J. P.; FALCÃO, A. X.; SUZUKI, C. T. N. Supervised pattern classification based on optimum-path forest. **International Journal of Imaging Systems and Technology**, v. 19, n. 2, p. 120–131, 2009.

PAPINENI, K. et al. **BLEU: A Method for Automatic Evaluation of Machine Translation**. Proceedings of the 40th Annual Meeting on Association for Computational Linguistics. **Anais**...: ACL '02.USA: Association for Computational Linguistics, 2002. Disponível em: <https://doi.org/10.3115/1073083.1073135>

PARAGUASSU, L. et al. **MedSimples: An Automated Simplification Tool for Promoting Health Literacy in Brazil**. DHandNLP@PROPOR. **Anais**...2020. Disponível em: <https://api.semanticscholar.org/CorpusID:218910691>

PARDO, T. A. S. **Gistsumm: Um sumarizador automático baseado na ideia principal de textos**. [s.l.] Série de Relatórios do Núcleo Interinstitucional de Linguística Computacional, Universidade de São Paulo, 2002.

PARDO, T. A. S.; RINO, L. H. M. **TeMário: Um corpus para sumarização automática de textos**. [s.l.] Série de Relatórios Técnicos da Universidade de São Carlos, 2003.

PARIDA, S.; MOTLICEK, P. **Abstract text summarization: A low resource challenge**. Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). **Anais**...2019.

PARK, S. et al. Activities on Facebook Reveal the Depressive State of Users. **J Med Internet Res**, v. 15, n. 10, p. e217, 2013.

PAROUBEK, P.; CHAUDIRON, S.; HIRSCHMAN, L. **Principles of Evaluation in Natural Language Processing**. Traitement Automatique des Langues, Volume 48, Numéro 1 : Principes de l'évaluation en Traitement Automatique des Langues [Principles of Evaluation in Natural Language Processing]. **Anais**...France: ATALA (Association pour le Traitement Automatique des Langues), 2007. Disponível em: <https://aclanthology.org/2007.tal-1.1>

PASQUALINI, B. **Corpop : um corpus de referência do português popular escrito do Brasil**. UFRGS - Porto Alegre - RS: Instituto de Letras - UFRGS, 2018.

PASQUALOTTI, P. R. **WordNet Affect BR – uma base de expressões de emoção em Português**. [s.l.] Novas Edições Acadêmicas, 2015.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

PELLE, R. P. DE; MOREIRA, V. **Offensive Comments in the Brazilian Web: a dataset and baseline results**. Anais do VI Brazilian Workshop on Social Network Analysis and Mining. **Anais**...2017.

PENNEBAKER, J. W. et al. The development and psychometric properties of LIWC2015. **The University of Texas at Austin**, 2015.

PENNEBAKER, J. W.; FRANCIS, M. E.; BOOTH, R. J. **Linguistic Inquiry and Word Count**. [s.l.] Lawerence Erlbaum Associates, 2001.

PENNINGTON, J.; SOCHER, R.; MANNING, C. **GloVe: Global Vectors for Word Representation**. Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). **Anais**...Doha, Qatar: Association for Computational Linguistics, out. 2014. Disponível em: <https://aclanthology.org/D14-1162>

PEREIRA, D. A. A Survey of Sentiment Analysis in the Portuguese Language. **Artificial Intelligence Review**, 2021.

PEREIRA, V.; PINHEIRO, V. **Report - um sistema de extração de informações aberta para língua portuguesa**. Anais do X Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana. **Anais**...SBC, 2015.

PÉREZ-ROSAS, V.; MIHALCEA, R. **Cross-cultural Deception Detection**. Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics. **Anais**...Baltimore, MD, USA: Association for Computational Linguistics, 2014.

PERSING, I.; NG, V. **Modeling Prompt Adherence in Student Essays**. Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics. **Anais**...Baltimore, Maryland: Association for Computational Linguistics, jun. 2014.

PETERS, M. E. et al. **Semi-supervised sequence tagging with bidirectional language models**. Proc. of ACL-2017. **Anais**...Vancouver, Canada: Association for Computational Linguistics, 2017.

PETERS, M. E. et al. **Deep Contextualized Word Representations**. (M. A. Walker, H. Ji, A. Stent, Eds.)Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2018, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 1 (Long Papers). **Anais**...Association for Computational Linguistics, 2018. Disponível em: <https://doi.org/10.18653/v1/n18-1202>

PETRI, M. J. C. **Manual de Linguagem Jurídica**. 3rd. ed. São Paulo: Saraiva, 2017.

PILÁN, I. et al. The text anonymization benchmark (tab): A dedicated corpus and evaluation framework for text anonymization. **Computational Linguistics**, v. 48, n. 4, p. 1053–1101, 2022.

PINKAS, G. et al. SARS-CoV-2 Detection From Voice. **IEEE Open Journal of Engi-**

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

neering in Medicine and Biology, v. 1, p. 268–274, 2020.

PINTO, I. V. et al. Fatores associados ao óbito de mulheres com notificação de violência por parceiro íntimo no Brasil. **Ciência & Saúde Coletiva**, v. 26, n. 3, p. 975–985, mar. 2021.

PINTO, L.; MELGAR, A. **A classification model for Portuguese documents in the juridical domain**. 2016 11th Iberian Conference on Information Systems and Technologies (CISTI). **Anais**...IEEE, 2016.

PIRES, R. et al. **Sabiá: Portuguese Large Language Models**. (M. C. Naldi, R. A. C. Bianchi, Eds.)Intelligent Systems. **Anais**...Cham: Springer Nature Switzerland, 2023.

PIRES, V. B.; GUERREIRO, D.; et al. **Portuguese Fake News Classification with BERT models**. Encontro Nacional de Inteligência Artificial e Computacional (ENIAC). **Anais**...SBC, 2024.

PIRES, V.; SILVA, D. G. E. **Portuguese Fake News Classification with BERT models**. Anais do XXI Encontro Nacional de Inteligência Artificial e Computacional. **Anais**...Porto Alegre, RS, Brasil: SBC, 2024. Disponível em: <https://sol.sbc.org.br/index.php/eniac/article/view/33848>

PIRINA, I.; ÇÖLTEKIN, ÇAĞRI. **Identifying Depression on Reddit: The Effect of Training Data**. Proceedings of the 2018 EMNLP Workshop SMM4H: The 3rd Social Media Mining for Health Applications Workshop & Shared Task. **Anais**...2018.

PIROVANI, J. P. C. **CRF+ LG: uma abordagem híbrida para o reconhecimento de entidades nomeadas em português**. tese de doutorado—[s.l.] Universidade Federal do Espírito Santo, Vitória (Brasil), 2019.

PITLER, E.; LOUIS, A.; NENKOVA, A. **Automatic evaluation of linguistic quality in multi-document summarization**. Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics. **Anais**...2010.

POLO, F. M. et al. **LegalNLP – Natural Language Processing methods for the Brazilian Legal Language**., 2021. Disponível em: <https://arxiv.org/abs/2110.15709>

PONTES, L. B. L.; OLIVEIRA, H. T. A. DE; ASSIS BOLDT, F. DE. **Avaliação de Modelos Neurais para Sumarização de Código-fonte**. Anais do XLIX Seminário Integrado de Software e Hardware. **Anais**...SBC, 2022.

PONTIKI, M. et al. **SemEval-2014 Task 4: Aspect Based Sentiment Analysis**. Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014). **Anais**...Association for Computational Linguistics, 2014. Disponível em: <https://aclanthology.org/S14-2004/>

PONTIKI, M. et al. **SemEval-2015 Task 12: Aspect Based Sentiment Analysis**. Proceedings of the 9th International Workshop on Semantic Evaluation. **Anais**...2015.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

PONTIKI, M. et al. **SemEval-2016 Task 5: Aspect Based Sentiment Analysis**. Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016). **Anais**...2016.

POPOVIC, M.; BURCHARDT, A. **From Human to Automatic Error Classification for Machine Translation Output**. Proceedings of the 15th Conference of the European Association for Machine Translation. **Anais**...Leuven, Belgium: 2011. Disponível em: <https://aclanthology.org/2011.eamt-1.36.pdf>

POPOVIĆ, M. **chrF: character n-gram F-score for automatic MT evaluation**. Proceedings of the Tenth Workshop on Statistical Machine Translation. **Anais**...Lisbon, Portugal: Association for Computational Linguistics, set. 2015. Disponível em: <https://aclanthology.org/W15-3049>

PORTER, M. F. An algorithm for suffix stripping. **Program**, v. 14, n. 3, p. 130–137, 1980.

PRADEEP, R. et al. H2oloo at trec 2020: When all you got is a hammer... deep learning, health misinformation, and precision medicine. **Corpus**, v. 5, n. d3, p. d2, 2020.

PUSTEJOVSKY, J. **The Generative Lexicon**. Cambridge, USA: MIT Press, 1995.

QIU, H. et al. **AMRFact: Enhancing Summarization Factuality Evaluation with AMR-Driven Negative Samples Generation**. (K. Duh, H. Gomez, S. Bethard, Eds.)Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers). **Anais**...Mexico City, Mexico: Association for Computational Linguistics, jun. 2024. Disponível em: <https://aclanthology.org/2024.naacl-long.33/>

QIU, Q. et al. BiLSTM-CRF for geological named entity recognition from the geoscience literature. **Earth Science Informatics**, v. 12, n. 4, p. 565–579, 2019.

QUARESMA, P.; FINATTO, M. J. B. **Information Extraction from Historical Texts: a Case Study.** DHandNLP@ PROPOR. **Anais**...2020.

QUARESMA, P.; GONÇALVES, T. Using Linguistic Information and Machine Learning Techniques to Identify Entities from Juridical Documents. Em: FRANCESCONI, E. et al. (Eds.). **Semantic Processing of Legal Texts: Where the Language of Law Meets the Law of Language**. Berlin, Heidelberg: Springer Berlin Heidelberg, 2010. p. 44–59.

QUINLAN, J. R. **C4.5: Programs for Machine Learning**. San Mateo, CA: Morgan Kaufmann Publishers, 1993.

RADEV, D. R. **A common theory of information fusion from multiple text sources step one: cross-document structure**. 1st SIGdial workshop on Discourse and Dialogue. **Anais**...2000.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

RAFFEL, C. et al. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. **Journal of Machine Learning Research**, v. 21, n. 140, p. 1–67, 2020.

RAGHAVAN, U. N.; ALBERT, R.; KUMARA, S. Near linear time algorithm to detect community structures in large-scale networks. **Physical Review E**, v. 76, n. 3, p. 036106, 2007.

RAMISCH, R. **Caracterização de desvios sintáticos em redações de estudantes do ensino médio: subsídios para o processamento automático das línguas naturais**. mathesis—[s.l.] Universidade Federal de São Carlos, 2020.

RANA, M. S. et al. Deepfake Detection: A Systematic Literature Review. **IEEE Access**, v. 10, p. 25494–25513, 2022.

RAU, L. F. **Extracting company names from text**. Proceedings the Seventh IEEE Conference on Artificial Intelligence Application. **Anais**...IEEE Computer Society, 1991.

RAYNER, K. Eye Movements in Reading and Information Processing: 20 Years of Research. **Psychological Bulletin - APA, vol. 124 n. 3**, p. 372–422, 1998.

REAL, L.; FONSECA, E.; GONÇALO OLIVEIRA, H. **The ASSIN 2 Shared Task: A Quick Overview**. Computational Processing of the Portuguese Language: 14th International Conference, PROPOR 2020, Evora, Portugal, March 2–4, 2020, Proceedings. **Anais**...Berlin, Heidelberg: Springer-Verlag, 2020. Disponível em: <https://doi.org/10.1007/978-3-030-41505-1_39>

REAL, L.; OSHIRO, M.; MAFRA, A. **B2W-Reviews01-An open product reviews corpus**. the Proceedings of the XII Symposium in Information and Human Language Technology. **Anais**...2019.

RECUERO, R. **Redes Sociais na Internet**. [s.l.] Ciber Cultura, 2009.

REI, R. et al. **COMET: A Neural Framework for MT Evaluation**. Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). **Anais**...Online: Association for Computational Linguistics, nov. 2020. Disponível em: <https://aclanthology.org/2020.emnlp-main.213>

REIS, G. B. Predição da Complexidade Textual de Notícias Jornalísticas usando uma Plataforma Crowdsourcing. **Monografia Conclusão Curso - USP**, 2017.

RESENDE, G. et al. **(Mis)Information Dissemination in WhatsApp: Gathering, Analyzing and Countermeasures**. Proceedings of the World Wide Web Conference. **Anais**...2019.

REYES, A.; ROSSO, P.; BUSCALDI, D. From Humor Recognition to Irony Detection: The Figurative Language of Social Media. **Data & Knowledge Engineering**, 2012.

RIBEIRO, A. S. **O projecto MONSOON: perspectivas digitais da Índia por-**

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

tuguesa. Actas da Jornada de Humanidades Digitais do CIDEHUS (to appear). **Anais**...2022.

RIBEIRO, M. T.; SINGH, S.; GUESTRIN, C. **" Why should i trust you?" Explaining the predictions of any classifier**. Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining. **Anais**...2016.

RICARTE NETO, F. A. et al. **Team PiLN at ABSAPT 2022: Lexical and BERT Strategies for Aspect-Based Sentiment Analysis in Portuguese**. Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2022) co-located with the Conference of the Spanish Society for Natural Language Processing (SEPLN 2022), A Coruña, Spain, September 20, 2022. **Anais**...2022.

RICKARD, E. M. S. A. M. L. K. A. B. D. F. A. N. S. Predicting Depression From Language-Based Emotion Dynamics: Longitudinal Analysis of Facebook and Twitter Status Updates. **Journal of Medical Internet Research**, v. 20, n. 5, p. e168, 2018.

RIJSBERGEN, C. JOOST. VAN. **Information Retrieval**. [s.l.] Butterworths, 1979.

RILOFF, E. et al. **Automatically constructing a dictionary for information extraction tasks**. AAAI. **Anais**...Citeseer, 1993.

RILOFF, E.; JONES, R.; et al. **Learning dictionaries for information extraction by multi-level bootstrapping**. AAAI/IAAI. **Anais**...1999.

RINO, L. H. M.; PARDO, T. A. S. **A Sumarização Automática de textos: principais características e metodologias**. Anais do XXIII Congresso da Sociedade Brasileira de Computação. **Anais**...2003.

RO, Y.; LEE, Y.; KANG, P. **Multi^2OIE: Multilingual Open Information Extraction Based on Multi-Head Attention with BERT**. Findings of the Association for Computational Linguistics: EMNLP 2020. **Anais**...Online: Association for Computational Linguistics, nov. 2020. Disponível em: <https://aclanthology.org/2020.findings-emnlp.99>

ROARK, B.; CHARNIAK, E. Noun-phrase co-occurrence statistics for semi-automatic semantic lexicon construction. **arXiv preprint cs/0008026**, 2000.

ROBERTSON, S. E.; SPÄRCK JONES, K. Relevance weighting of search terms. **Journal of the American Society for Information science**, v. 27, n. 3, p. 129–146, 1976.

ROBERTSON, S. E.; WALKER, S. **Okapi/keenbow at trec-8**. TREC. **Anais**...Citeseer, 1999. Disponível em: <https://trec.nist.gov/pubs/trec8/papers/okapi.pdf>

ROBOTTI, C. et al. Machine Learning-based Voice Assessment for the Detection of Positive and Recovered COVID-19 Patients. **Journal of Voice**, 2021.

ROCCHIO-JR, J. J. Relevance feedback in information retrieval. **The SMART retrieval**

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

system: experiments in automatic document processing, 1971.

RODRÍGUEZ, M. M.; BEZERRA, B. L. D. Processamento de linguagem natural para reconhecimento de entidades nomeadas em textos jurídicos de atos administrativos (portarias). **Revista de Engenharia e Pesquisa Aplicada**, v. 5, n. 1, p. 67–77, 2020.

ROGERS, A.; KOVALEVA, O.; RUMSHISKY, A. A primer in BERTology: What we know about how BERT works. **Transactions of the Association for Computational Linguistics**, v. 8, p. 842–866, 2021.

ROSA, G. M. et al. Yes, bm25 is a strong baseline for legal case retrieval. **arXiv preprint arXiv:2105.05686**, b2021.

ROSA, G. M. et al. A cost-benefit analysis of cross-lingual transfer methods. **arXiv preprint arXiv:2105.06813**, a2021.

ROSVALL, M.; BERGSTROM, C. T. Maps of random walks on complex networks reveal community structure. **Proceedings of the National Academy of Sciences**, v. 105, n. 4, p. 1118–1123, 2008.

ROTH, D.; YIH, W. Global inference for entity and relation identification via a linear programming formulation. **Introduction to statistical relational learning**, p. 553–580, 2007.

ROY, K.; GARG, T.; PALIT, V. **Knowledge Graph Guided Semantic Evaluation of Language Models For User Trust**. 2023 IEEE Conference on Artificial Intelligence (CAI). **Anais**...2023.

RUBIN, V. L. TALIP Perspectives, Guest Editorial Commentary: Pragmatic and Cultural Considerations for Deception Detection in Asian Languages. v. 13, n. 2, p. 10:1–10:8, 2014.

RUBIN, V. L.; CHEN, Y.; CONROY, N. J. Deception detection for news: Three types of fakes. **Proceedings of the Association for Information Science and Technology**, v. 52, n. 1, p. 1–4, 2015.

RUBIN, V. L.; CONROY, N. J. Challenges in automated deception detection in computer-mediated communication. **Proceedings of the American Society for Information Science and Technology**, v. 48, n. 1, p. 1–4, 2011.

RUPPENHOFER, J. et al. **FrameNet II: Extended Theory and Practice**. [s.l: s.n.].

RUSSELL, M. A. **Mineração de Dados da Web Social**. Primeira edição ed. São Paulo: O'Reilly Novatec, 2011.

RUSSELL-ROSE, T.; CHAMBERLAIN, J.; AZZOPARDI, L. Information retrieval in the workplace: A comparison of professional search practices. **Information Processing & Management**, v. 54, n. 6, p. 1042–1057, 2018.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

SACRAMENTO, A. DA S. B.; SOUZA, M. **Joint Event Extraction with Contextualized Word Embeddings for the Portuguese Language**. Brazilian Conference on Intelligent Systems. **Anais**...Springer, 2021.

SAGER, N. Natural language information formatting: the automatic conversion of texts to a structured data base. Em: **Advances in computers**. [s.l.] Elsevier, 1978. v. 17p. 89–162.

SAGER, N.; FRIEDMAN, C.; LYMAN, M. S. **Medical language processing: computer management of narrative data**. [s.l.] Addison-Wesley Longman Publishing Co., Inc., 1987.

SAINZ, O. et al. NLP Evaluation in trouble: On the Need to Measure LLM Data Contamination for each Benchmark. **Findings of the Association for Computational Linguistics: EMNLP 2023**, 2023.

SAKIYAMA, K. M. **Geração Automática de Verbetações para Recuperação de Informações no Domínio Jurídico Brasileiro**. mathesis—[s.l.] Instituto de Ciências Matemáticas e de Computação - Universidade de São Paulo, 2023.

SALES ALMEIDA, T. et al. Sabiá-2: A New Generation of Portuguese Large Language Models. **arXiv e-prints**, p. arXiv–2403, 2024.

SALMINEN, J. et al. Creating and detecting fake reviews of online products. **Journal of Retailing and Consumer Services**, v. 64, p. 102771, 2022.

SALTON, G.; MCGILL, M. J. **Introduction to Modern Information Retrieval**. [s.l.] McGraw-Hill, 1983.

SALVI, C. et al. Going Viral: How Fear, Socio-Cognitive Polarization and Problem-Solving Influence Fake News Detection and Proliferation During COVID-19 Pandemic. **Frontiers in Communication**, v. 5, p. 127, 2021.

SAMY, D. Reconocimiento y clasificación de entidades nombradas en textos legalesen español. **Procesamiento del lenguaje natural**, v. 67, p. 103–114, 2021.

SANDERSON, M. et al. Test collection based evaluation of information retrieval systems. **Foundations and Trends® in Information Retrieval**, v. 4, n. 4, p. 247–375, 2010.

SANG, E. F. T. K.; DE MEULDER, F. **Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition**. Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003. **Anais**...2003. Disponível em: <https://aclanthology.org/W03-0419>

SANSONE, C.; SPERLÍ, G. Legal Information Retrieval systems: State-of-the-art and open issues. **Information Systems**, v. 106, p. 101967, 2022.

SANTANA, B. S. **A computational-linguistic-based approach to support the**

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

analysis of the discursive configuration of violence on social media. tese de doutorado—[s.l.] Universidade Federal do Rio Grande do Sul, 2023.

SANTOS, C. N. DOS; GUIMARÃES, V. **Boosting Named Entity Recognition with Neural Character Embeddings**. (X. Duan et al., Eds.)Proceedings of the 5th Named Entity Workshop. **Anais**...Association for Computational Linguistics, 2015.

SANTOS, D.; CARDOSO, N. **Breve introdução ao HAREM**. (D. Santos, N. Cardoso, Eds.)Reconhecimento de entidades mencionadas em português: Documentação e actas do HAREM, a primeira avaliação conjunta na área. **Anais**...Linguateca, a2007. Disponível em: <http://www.linguateca.pt/LivroHAREM/>

SANTOS, D.; CARDOSO, N. **A golden resource for named entity recognition in portuguese**. Proceeding of the 7th International conference on the computational processing of portuguese. **Anais**...Springer, b2007.

SANTOS, D.; CARDOSO, N.; SECO, N. **Avaliação no HAREM: Métodos e medidas**. (D. Santos, N. Cardoso, Eds.)Reconhecimento de entidades mencionadas em português: Documentação e actas do HAREM, a primeira avaliação conjunta na área. **Anais**...Linguateca, 2007.

SANTOS, D.; ROCHA, P. **The key to the first CLEF with Portuguese: Topics, questions and answers in CHAVE**. Workshop of the Cross-Language Evaluation Forum for European Languages. **Anais**...b2004.

SANTOS, D.; ROCHA, P. CHAVE: topics and questions on the Portuguese participation in CLEF. a2004.

SANTOS, H. D. P. DOS et al. **Fall Detection in EHR using Word Embeddings and Deep Learning**. 2019 IEEE 19th International Conference on Bioinformatics and Bioengineering (BIBE). **Anais**...a2019.

SANTOS, H. D. P. D.; ULBRICH, A. H. D. P. S.; VIEIRA, R. **Evaluation of a Prescription Outlier Detection System in Hospital's Pharmacy Services**. 2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). **Anais**...IEEE, a2021.

SANTOS, J. et al. **Assessing the Impact of Contextual Embeddings for Portuguese Named Entity Recognition**. Proceedings of the 8th Brazilian Conference on Intelligent Systems. **Anais**...b2019.

SANTOS, J. et al. **De-identification of clinical notes using contextualized language models and a token classifier**. Brazilian Conference on Intelligent Systems. **Anais**...Springer, b2021.

SANTOS, J. et al. Named entity recognition specialised for Portuguese 18th-century History research. **Proceedings of the International Conference on the Computational treatment of Portuguese, PROPOR**, a2024.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

SANTOS, J.; SANTOS, H. D. P. DOS; VIEIRA, R. **Fall Detection in Clinical Notes using Language Models and Token Classifier**. (A. G. S. de Herrera et al., Eds.)Proceedings of the 33rd IEEE International Symposium on Computer-Based Medical Systems. **Anais**...a2020.

SANTOS, L. B. DOS et al. A Lightweight Regression Method to Infer Psycholinguistic Properties for Brazilian Portuguese. **International Conference on Text, Speech, and Dialogue**, p. 281–289, a2017.

SANTOS, L. B. DOS et al. **Enriching Complex Networks with Word Embeddings for Detecting Mild Cognitive Impairment from Speech Transcripts**. Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). **Anais**...Vancouver, Canada: Association for Computational Linguistics, jul. b2017. Disponível em: <https://aclanthology.org/P17-1118/>

SANTOS, R. et al. **Measuring the Impact of Readability Features in Fake News Detection**. (N. Calzolari et al., Eds.)Proceedings of the Twelfth Language Resources and Evaluation Conference. **Anais**...Marseille, France: European Language Resources Association, b2020. Disponível em: <https://aclanthology.org/2020.lrec-1.176>

SANTOS, R. L. DE S. **Detecção Automática de Notícias Falsas em Português**. Ph.D. Thesis—São Carlos, Brazil: Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, 2022.

SANTOS, R. L. DE S.; PARDO, T. A. S. **Fact-Checking for Portuguese: Knowledge Graph and Google Search-Based Methods**. Proceedings of the 14th International Conference on the Computational Processing of Portuguese (PROPOR). **Anais**...: Lecture Notes em Artificial Intelligence (LNAI).Évora, Portugal: Springer, 2020.

SANTOS, R. L. DE S.; PARDO, T. A. S. **Structural Characterization and Graph-based Detection of Fake News in Portuguese**. Proceedings of the XIV Symposium in Information and Human Language (STIL). **Anais**...2021.

SANTOS, W. R. DOS; FUNABASHI, A. M. M.; PARABONI, I. **Searching Brazilian Twitter for signs of mental health issues**. 12th International Conference on Language Resources and Evaluation (LREC-2020). **Anais**...Marseille, France: ELRA, c2020.

SANTOS, W. R. DOS; OLIVEIRA, R. L. DE; PARABONI, I. SetembroBR: a social media corpus for depression and anxiety disorder prediction. **Language Resources and Evaluation**, v. 58, n. 1, p. 273–300, b2024.

SANTOS, W. R. DOS; PARABONI, I. **Predição de transtorno depressivo em redes sociais: BERT supervisionado ou ChatGPT zero-shot?** XIV Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana (STIL-2023). **Anais**...Porto Alegre, Brasil: SBC, 2023. Disponível em: <https://sol.sbc.org.br/index.php/stil/article/view/25433>

SANTOS, W. R. DOS; PARABONI, I. **Prompt-based mental health screening**

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

from social media text. Brazilian Workshop on Social Network Analysis and Mining (BraSNAM-2024). **Anais**...Brasilia, DF: 2024.

SANTOS, W. R. DOS; YOON, S.; PARABONI, I. Mental Health Prediction from Social Media Text Using Mixture of Experts. **IEEE Latin America Transactions**, v. 21, n. 6, p. 723–729, 2023.

SAQUETE, E. et al. Fighting Post-truth using Natural Language Processing: A Review and Open Challenges. **Expert Systems with Applications**, v. 141, p. 112943, 2019.

SARACEVIC, T. Ciência da informação: origem, evolução e relações. **Perspectivas em ciência da informação**, v. 1, n. 1, 1996.

SARAH HICKEY. **Nimdzi 100 - Language Services Industry Market Report 2020.pdf**. [s.l: s.n.].

SARDINHA, T. B. Lingüística de Corpus: histórico e problemática. **DELTA: Documentação de Estudos em Lingüística Teórica e Aplicada**, v. 16, n. 2, p. 323–367, 2000.

SARMENTO, M. A.; OLIVEIRA, H. T. A. DE. **Sumarização Automática de Artigos de Notícias em Português: Da Extração à Abstração com Abordagens Clássicas e Modelos de Neurais**. Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana (STIL). **Anais**...SBC, 2024.

SCAO, T. L. et al. BLOOM: A 176B-Parameter Open-Access Multilingual Language Model. **CoRR**, v. abs/2211.05100, 2022.

SCARTON, C. et al. Simplifica: a tool for authoring simplified texts in Brazilian Portuguese guided by readability assessments. **Proceedings of the 2010 Conference of the North American Chapter of the Association for Computational Linguistics - Human Language Technologies**, p. 41–44, 2010.

SCARTON, C. E.; ALUÍSIO, S. M. Análise da Inteligibilidade de textos via ferramentas de Processamento de Língua Natural: adaptando as métricas do Coh-Metrix para o Português. **Linguamática**, v. 2, n. 1, p. 45–61, 2010.

SCARTON, C.; SPECIA, L. Learning Simplifications for Specific Target Audiences. **Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Short Papers)**, p. 712–718, 2018.

SCHANK, R. C. et al. **MARGIE: Memory Analysis Response Generation, and Inference on English.** IJCAI. **Anais**...1973.

SCHMITZ, M. et al. **Open language learning for information extraction**. Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning. **Anais**...: EMNLP-CoNLL '12.Stroudsburg, PA, USA: Association for Computational Linguistics; Association for Computational

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Linguistics, 2012. Disponível em: <http://dl.acm.org/citation.cfm?id=2390948.2391009>

SCHNEIDER, E. T. R. et al. **BioBERTpt - A Portuguese Neural Language Model for Clinical Named Entity Recognition**. (A. Rumshisky et al., Eds.)Proceedings of the 3rd Clinical Natural Language Processing Workshop. **Anais**...Online: Association for Computational Linguistics, nov. 2020. Disponível em: <https://aclanthology.org/2020.clinicalnlp-1.7>

SCHNEIDER, E. T. R. et al. **A GPT-2 Language Model for Biomedical Texts in Portuguese**. 2021 IEEE 34th International Symposium on Computer-Based Medical Systems (CBMS). **Anais**...2021.

SCHNEIDER, E. T. R. et al. **CardioBERTpt: Transformer-based Models for Cardiology Language Representation in Portuguese**. 2023 IEEE 36th International Symposium on Computer-Based Medical Systems (CBMS). **Anais**...2023.

SCHRAAGEN, M. et al. Evaluation of Named Entity Recognition in Dutch online criminal complaints. **Computational Linguistics in the Netherlands Journal**, v. 7, p. 3–16, 2017.

SCHRIML, L. M. et al. Disease Ontology: a backbone for disease semantic integration. **Nucleic acids research**, v. 40, n. D1, p. D940–D946, 2012.

SCHUBERT, G.; FREITAS, L. A. DE. **A Construção de um Corpus para Detecção de Ironia e Sarcasmo em Português**. Anais do XVII Encontro Nacional de Inteligência Artificial e Computacional. **Anais**...2020.

SCHULLER, B. W. et al. **The INTERSPEECH 2021 Computational Paralinguistics Challenge: COVID-19 Cough, COVID-19 Speech, Escalation & Primates**. Interspeech 2021. **Anais**...2021.

SCHUSTER, M.; PALIWAL, K. K. Bidirectional recurrent neural networks. **IEEE transactions on Signal Processing**, v. 45, n. 11, p. 2673–2681, 1997.

SCHWARTZ, R. et al. Green ai. **Communications of the ACM**, v. 63, n. 12, p. 54–63, 2020.

SCIENCEDIRECT. **Complex Networks - an overview**., b2025. Disponível em: <https://www.sciencedirect.com/topics/computer-science/complex-networks>

SCIENCEDIRECT. **Complex Networks Article**., a2025. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0966692325001619>

SEBASTIANI, F. Machine learning in automated text categorization. **ACM computing surveys (CSUR)**, v. 34, n. 1, p. 1–47, 2002.

SEKINE, S. **Description of the Japanese NE system used for MET-2**. Seventh

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Message Understanding Conference (MUC-7): Proceedings of a Conference Held in Fairfax, Virginia, April 29-May 1, 1998. **Anais**...1998.

SELLAM, T.; DAS, D.; PARIKH, A. P. **BLEURT: Learning Robust Metrics for Text Generation**. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020. **Anais**...2020. Disponível em: <https://doi.org/10.18653/v1/2020.acl-main.704>

SELVARAJU, R. R. et al. **Grad-cam: Visual explanations from deep networks via gradient-based localization**. Proceedings of the IEEE international conference on computer vision. **Anais**...2017.

SEMENOV, A. et al. **Discerning Depression Propensity Among Participants of Suicide and Depression-Related Groups of Vk.com**. Analysis of Images, Social Networks and Texts. **Anais**...Cham: Springer International Publishing, 2015.

SENA, C. F. L.; CLARO, D. B. InferPortOIE: A Portuguese Open Information Extraction system with inferences. **Natural Language Engineering**, v. 25, n. 2, p. 287–306, 2019.

SENA, C. F. L.; CLARO, D. B. PragmaticOIE: a pragmatic open information extraction for Portuguese language. **Knowl. Inf. Syst.**, v. 62, n. 9, p. 3811–3836, 2020.

SENA, C. F. L.; GLAUBER, R.; CLARO, D. B. **Inference Approach to Enhance a Portuguese Open Information Extraction**. Proceedings of the 19th International Conference on Enterprise Information Systems - Volume 3: ICEIS. **Anais**...INSTICC; SciTePress, 2017.

SENNRICH, R.; HADDOW, B.; BIRCH, A. **Improving Neural Machine Translation Models with Monolingual Data**. Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL 2016). **Anais**...2016. Disponível em: <https://arxiv.org/abs/1511.06709>

SERRAS, F. et al. **Analysing and Validating Language Complexity Metrics Across South American Indigenous Languages**. (T. Kuribayashi et al., Eds.)Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics. **Anais**...Bangkok, Thailand: Association for Computational Linguistics, ago. 2024. Disponível em: <https://aclanthology.org/2024.cmcl-1.13/>

SHAMMA, S. A. et al. **Information Extraction from Arabic Law Documents**. 2020 IEEE 14th International Conference on Application of Information and Communication Technologies (AICT). **Anais**...2020.

SHAOWEI, Z. et al. Survey of Supervised Joint Entity Relation Extraction Methods. **Journal of Frontiers of Computer Science & Technology**, v. 16, n. 4, 2022.

SHARDLOW, M. A Survey of Automated Text Simplification. **International Journal of Advanced Computer Science and Applications(IJACSA), Special Issue on Natural Language Processing 2014**, v. 4, n. 1, 2014.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

SHEIKHALISHAHI, S. et al. Natural Language Processing of Clinical Notes on Chronic Diseases: Systematic Review. **JMIR Med Inform**, v. 7, n. 2, p. e12239, abr. 2019.

SHEN, G. et al. **Depression Detection via Harvesting Social Media: A Multimodal Dictionary Learning Solution**. 26th International Joint Conference on Artificial Intelligence, IJCAI-17. **Anais**...2017.

SHEN, J. H.; RUDZICZ, F. **Detecting Anxiety on Reddit**. Fourth Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality. **Anais**...Vancouver, Canada: Association for Computational Linguistics, 2017.

SHEN, T. et al. **Cross-Domain Depression Detection via Harvesting Social Media**. Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18. **Anais**...International Joint Conferences on Artificial Intelligence Organization, 2018.

SHERMIS, M. D.; BURSTEIN, J. **Handbook of Automated Essay Evaluation: Current Applications and New Directions**. [s.l.] Routledge/Taylor & Francis Group, 2013.

SHIBATA, T.; MIYAMURA, Y. **LCES: Zero-shot Automated Essay Scoring via Pairwise Comparisons Using Large Language Models**., 2025. Disponível em: <https://arxiv.org/abs/2505.08498>

SHICKEL, B. et al. Deep EHR: A Survey of Recent Advances in Deep Learning Techniques for Electronic Health Record (EHR) Analysis. **IEEE J Biomed Health Inform**, v. 22, n. 5, p. 1589–1604, out. 2017.

SHIMANAKA, H.; KAJIWARA, T.; KOMACHI, M. Machine Translation Evaluation with BERT Regressor. **arXiv**, v. abs/1907.12679, 2019.

SHRESTHA, A.; SPEZZANO, F. **Detecting Depressed Users in Online Forums**. 2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM). **Anais**...2019.

SHTERIONOV, D. et al. Human versus Automatic Quality Evaluation of NMT and PBSMT. **Machine Translation**, v. 32, n. 3, p. 217–235, 2018.

SHUJA, J. et al. COVID-19 open source data sets: a comprehensive survey. **Applied Intelligence**, v. 51, n. 3, p. 1296–1325, 2021.

SIDDHARTHAN, A. Syntactic Simplifcation and Text Cohesion. **Research on Language and Computation - Springer**, 2006.

SILLA, C. N.; FREITAS, A. A. A survey of hierarchical classification across different application domains. **Data Mining and Knowledge Discovery**, v. 22, n. 1-2, p. 31–72, 2011.

SILVA, A. P. DA et al. Risco de queda relacionado a medicamentos em hospitais: abor-

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

dagem de aprendizado de máquina. **Acta Paulista de Enfermagem**, v. 36, a2023.

SILVA, D. P. P. DA et al. Interpretability analysis of deep models for COVID-19 detection. **Artificial Intelligence in Health**, v. 1, n. 3, p. 114–126, 2024.

SILVA, F. DA et al. **Towards analysis on textual inference at ASSIN-2 dataset**. Proceedings of the 14th Brazilian Symposium in Information and Human Language Technology. **Anais**...b2023.

SILVA, F. L. V. DA et al. ABSAPT 2022 at IberLEF: Overview of the Task on Aspect-Based Sentiment Analysis in Portuguese. **Procesamiento del Lenguaje Natural**, v. 69, p. 199–205, a2022.

SILVA, F. N. et al. Using network science and text analytics to produce surveys in a scientific topic. **Journal of Informetrics**, v. 10, n. 2, p. 487–502, 2016.

SILVA, F. R. A. DA. **Detecção de Ironia e Sarcasmo em Língua Portuguesa: uma abordagem utilizando Deep Learning**. https://github.com/fabio-ricardo/deteccao-ironia, 2018.

SILVA, I. A. L. DA et al. Translation, post-editing and directionality. **Translation in transition: Between cognition, computing and technology**, p. 107–134, a2017.

SILVA, J. F. F. **Estratégias para sumarização de documentos**. mathesis—[s.l.] (Mestrado em Engenharia informática) - Faculdade de Ciências e Tecnologia da Universidade Nova de Lisboa, 2022.

SILVA, M. J.; CARVALHO, P.; SARMENTO, L. **Building a Sentiment Lexicon for Social Judgement Mining**. Proceedings of the 10th International Conference on Computational Processing of the Portuguese Language. **Anais**...2012.

SILVA, N. F. F. DA et al. **Evaluating Topic Models in Portuguese Political Comments About Bills from Brazil's Chamber of Deputies**. (A. Britto, K. Valdivia Delgado, Eds.)Intelligent Systems. **Anais**...Cham: Springer International Publishing, 2021.

SILVA, N. L. DA; DI FELIPPO, A. **Descrição e Análise do Fenômeno da Contradição para a Sumarização Automática Multidocumento**. [s.l.] Série de Relatórios Técnicos do Núcleo Interinstitucional de Linguística Computacional, 2014.

SILVA, R. M. et al. Towards Automatically Filtering Fake News in Portuguese. **Expert Systems with Applications**, v. 146, p. 1–48, b2020.

SILVA, R. M. et al. Towards Automatically Filtering Fake News in Portuguese. **Expert Systems With Applications**, v. 146, p. 1–14, a2020.

SILVA, R. M. et al. Fake News Detection in Portuguese Under Large Language Model-Generated Content. **Journal of the Brazilian Computer Society**, p. 1–18, 2025.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

SILVA, R. M.; ALMEIDA, T. A. **How concept drift can impair the classification of fake news**. Proceedings of the 9th Symposium on Knowledge Discovery, Mining and Learning (KDMiLe'21). **Anais**...2021.

SILVA, R. M.; ALMEIDA, T. A.; YAMAKAMI, A. MDLText: An Efficient and Lightweight Text Classifier. **Knowledge-Based Systems**, v. 118, p. 152–164, b2017.

SILVA, R. M.; PIRES, P. R.; ALMEIDA, T. A. Incremental learning for fake news detection. **Journal of Information and Data Management – JIDM**, v. 13, p. 566–579, b2022.

SINGH, D.; GARG, R. A survey of Community Detection algorithms and its comparative performance analysis. **Computer Science Review**, v. 58, 2025.

SIQUEIRA, F. et al. Ulysses Tesemõ: a new large corpus for Brazilian legal and governmental domain. **Language Resources and Evaluation**, p. 1–20, jul. 2024.

SJÖHOLM, J. **Probability as readability: A new machine learning approach to readability assessment for written Swedish**. [s.l.] LiU Electronic Press, 2012.

Slator 2019 Language Industry Market Report. p. 33, 2019.

SLEIMI, A. et al. An automated framework for the extraction of semantic legal metadata from legal texts. **Empirical Software Engineering**, v. 26, p. 1–50, 2021.

SMIRNOVA, A.; CUDRÉ-MAUROUX, P. Relation extraction using distant supervision: A survey. **ACM Computing Surveys (CSUR)**, v. 51, n. 5, p. 1–35, 2018.

SMITH, K. S. **On Integrating Discourse in Machine Translation**. Proceedings of the Third Workshop on Discourse in Machine Translation. **Anais**...2017.

SMYWIŃSKI-POHL, A. et al. **Automatic Extraction of Amendments from Polish Statutory Law**. Proceedings of the Eighteenth International Conference on Artificial Intelligence and Law. **Anais**...: ICAIL '21.New York, NY, USA: Association for Computing Machinery, 2021.

SNOVER, M. G. et al. **A Study of Translation Edit Rate with Targeted Human Annotation**. Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers, AMTA 2006, Cambridge, Massachusetts, USA, August 8-12, 2006. **Anais**...2006. Disponível em: <https://aclanthology.org/2006.amta-papers.25/>

SOARES, M. O que é letramento? **Presença Pedagógica Volume 2, n. 10**, p. 15–25, 1996.

SOBAHI, N. et al. Explainable COVID-19 detection using fractal dimension and vision transformer with Grad-CAM on cough sounds. **Biocybernetics and Biomedical Engineering**, v. 42, n. 3, p. 1066–1080, 2022.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

SOCHER, R. et al. **Semantic compositionality through recursive matrix-vector spaces**. Proceedings of the 2012 joint conference on empirical methods in natural language processing and computational natural language learning. **Anais**...2012.

SODERLAND, S. et al. **CRYSTAL inducing a conceptual dictionary**. Proceedings of the 14th international joint conference on Artificial intelligence-Volume 2. **Anais**...1995.

SOLORIO, T. MALINCHE: A NER system for Portuguese that reuses knowledge from Spanish. **Reconhecimento de entidades mencionadas em português: Documentação e atas do HAREM, a primeira avaliação conjunta na área, Capítulo**, v. 10, p. 123–136, 2007.

SONG, H. et al. **Feature Attention Network: Interpretable Depression Detection from Social Media**. 32nd Pacific Asia Conference on Language, Information and Computation. **Anais**...Hong Kong: Association for Computational Linguistics, 2018.

SOUSA, A. et al. **Cross-Lingual Annotation Projection for Argument Mining in Portuguese**. (G. Marreiros et al., Eds.)Progress in Artificial Intelligence. **Anais**...Springer International Publishing, 2021.

SOUSA, M. C. P. DE. O Corpus Tycho Brahe: contribuições para as humanidades digitais no Brasil. **Filologia e linguística portuguesa**, v. 16, n. esp., p. 53–93, 2014.

SOUSA, R. F. DE; BRUM, H. B.; NUNES, M. DAS G. V. **A bunch of helpfulness and sentiment corpora in brazilian portuguese**. Proceedings of Symposium in Information and Human Language Technology. **Anais**...2019.

SOUZA, E. et al. An Information Retrieval Pipeline for Legislative Documents from the Brazilian Chamber of Deputies. Em: **Legal Knowledge and Information Systems**. [s.l.] IOS Press, 2021a. p. 119–126.

SOUZA, E. N. P. DE; CLARO, D. B.; GLAUBER, R. A Similarity Grammatical Structures Based Method for Improving Open Information Systems. **j-jucs**, v. 24, n. 1, p. 43–69, 28 jan. 2018.

SOUZA, E. N. P.; CLARO, D. B. Extração de Relações utilizando Features Diferenciadas para Português. **Linguamática**, v. 6, n. 2, p. 57–65, 2014.

SOUZA, F. D.; FILHO, J. B. DE O. E. S. **BERT for sentiment analysis: Pre-trained and fine-tuned alternatives**. International Conference on Computational Processing of the Portuguese Language. **Anais**...Springer, 2022.

SOUZA, F.; NOGUEIRA, R.; LOTUFO, R. **BERTimbau: pretrained BERT models for Brazilian Portuguese**. (R. Cerri, R. C. Prati, Eds.)Proceedings of the 2020 Brazilian Conference on Intelligent Systems. **Anais**...Springer International Publishing, a2020.

SOUZA FREIRE, P. M.; MATIAS DA SILVA, F. R.; GOLDSCHMIDT, R. R. Fake news detection based on explicit and implicit signals of a hybrid crowd: An approach inspired

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

in meta-learning. **Expert Systems with Applications**, v. 183, p. 115414, 2021.

SOUZA, J. W. DA C. **Aprofundamento da caracterização linguístico-computacional da complementaridade em um corpus jornalístico multidocumento**. tese de doutorado—[s.l.] (Doutorado em Linguística) - Programa de Pós-Graduação em Linguística, Universidade Federal de São Carlos, 2019.

SOUZA, J. W. DA C.; FELIPPO, A. D. Caracterização da complementaridade temporal: subsídios para sumarização automática multidocumento. **Alfa: Revista de Linguística (São José do Rio Preto)**, v. 62, p. 125–150, 2018.

SOUZA, J. W. DA C.; FELIPPO, A. D.; PARDO, T. A. S. **Investigação da Identificação da Redundância na Sumarização Multidocumento**. Anais do III Student Workshop on Information and Human Language Technology. **Anais**...a2011.

SOUZA, M. et al. **Construction of a Portuguese Opinion Lexicon from multiple resources**. Proceedings of the 8th Brazilian Symposium in Information and Human Language Technology. **Anais**...b2011.

SOUZA, V.; NOBRE, J.; BECKER, K. **Characterization of Anxiety, Depression, and their Comorbidity from Texts of Social Networks**. Anais do XXXV Simpósio Brasileiro de Bancos de Dados. **Anais**...Porto Alegre, Brazil: SBC, b2020.

SOUZA, V.; NOBRE, J.; BECKER, K. A Deep Learning Ensemble to Classify Anxiety, Depression, and their Comorbidity from Texts of Social Networks. **Journal of Information and Data Management**, v. 12, n. 3, p. 306–325, b2021.

SPÄRCK JONES, K. Report on the need for and provision of an 'ideal' information retrieval test collection. **Computer Laboratory**, 1975.

SPÄRCK JONES, K.; WALKER, S.; ROBERTSON, S. E. A probabilistic model of information retrieval: development and comparative experiments. **Information processing & management**, v. 36, n. 6, p. 809–840, 2000.

SPARCK-JONES, K. **Automatic Summarizing: Factors and Directions. In Mani, I. And Maybury, M., editors, Advances in Automatic Text Summarization**. MIT Press, 1998.

SPEER, R.; CHIN, J.; HAVASI, C. **ConceptNet 5.5: an open multilingual graph of general knowledge**. Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence. **Anais**...: AAAI'17.San Francisco, California, USA: AAAI Press, 2017.

SPEICHER, T. et al. **Potential for discrimination in online targeted advertising**. Proceedings of the Conference on Fairness, Accountability and Transparency. **Anais**...ACM, 2018.

SPRINGEROPEN APPLIED NETWORK SCIENCE. **Small-world and scale-free properties in networks**., 2017. Disponível em: <https://appliednetsci.springeropen.c

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

om/articles/10.1007/s41109-017-0054-z>

SRIVASTAVA, N. et al. Dropout: a simple way to prevent neural networks from overfitting. **The journal of machine learning research**, v. 15, n. 1, p. 1929–1958, 2014.

SRIVASTAVA, S.; SHARMA, G. **OmniVec2 - A Novel Transformer Based Network for Large Scale Multimodal and Multitask Learning**. 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). **Anais**...2024.

STANOJEVIC, M.; SIMA'AN, K. **BEER: BEtter Evaluation as Ranking**. Proceedings of the Ninth Workshop on Statistical Machine Translation, WMT@ACL 2014, June 26-27, 2014, Baltimore, Maryland, USA. **Anais**...2014. Disponível em: <https://doi.org/10.3115/v1/w14-3354>

STANOVSKY, G. et al. **Supervised open information extraction**. Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers). **Anais**...2018.

STEYVERS, M.; TENENBAUM, J. B. The large-scale structure of semantic networks: statistical analyses and a model of semantic growth. **Cognitive science**, v. 29, n. 1, p. 41–78, 2005.

STJ. **Supremo Tribunal de Justiça — Composição**. https://www.stj.jus.br/sites/portalp/Institucional/Composicao, maio 24DC.

STRIEN, D. VAN et al. **Assessing the impact of OCR quality on downstream NLP tasks**. ICAART 2020 - Proceedings of the 12th International Conference on Agents and Artificial Intelligence. **Anais**...2020.

SU, J.; CARDIE, C.; NAKOV, P. **Adapting Fake News Detection to the Era of Large Language Models**. Proceedings of the Findings of the Association for Computational Linguistics: NAACL 2024. **Anais**...2024.

SU, K.-Y.; WU, M.-W.; CHANG, J.-S. **A new quantitative quality measure for machine translation systems**. Proceedings of the 14th conference on Computational linguistics -. **Anais**...Association for Computational Linguistics, 1992. Disponível em: <http://dx.doi.org/10.3115/992133.992137>

SUN, C.; EMONET, V.; DUMONTIER, M. **A Comprehensive Comparison of Automated FAIRness Evaluation Tools**. (K. Wolstencroft et al., Eds.)13th International Conference on Semantic Web Applications and Tools for Health Care and Life Sciences, SWAT4HCLS 2022, Virtual Event, Leiden, The Netherlands, January 10th to 14th, 2022. **Anais**...: CEUR Workshop Proceedings.CEUR-WS.org, 2022. Disponível em: <http://ceur-ws.org/Vol-3127/paper-6.pdf>

SUTSKEVER, I.; VINYALS, O.; LE, Q. V. **Sequence to Sequence Learning with Neural Networks**. (Z. Ghahramani et al., Eds.)Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014,

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

December 8-13 2014, Montreal, Quebec, Canada. **Anais**...2014. Disponível em: <https://proceedings.neurips.cc/paper/2014/hash/a14ac55a4f27472c5d894ec1c3c743d2-Abstract.html>

SWEET, P. L. Every bone of my body: Domestic violence and the diagnostic body. **Social Science & Medicine**, v. 122, p. 44–52, 2014.

TAKAMATSU, S.; SATO, I.; NAKAGAWA, H. **Reducing wrong labels in distant supervision for relation extraction**. Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). **Anais**...2012.

TAN, K. L.; LEE, C. P.; LIM, K. M. A Survey of Sentiment Analysis: Approaches, Datasets, and Future Research. **Applied Sciences**, 2023.

TANDOC JR., E. C.; LIM, Z. W.; LING, R. Defining "Fake News". **Digital Journalism**, v. 6, n. 2, p. 137–153, 2018.

TAUS. **TAUS - The Translation Industry in 2022 Report**., 2020. Disponível em: <https://info.taus.net/translation-industry-2022-report-download>. Acesso em: 19 ago. 2020

TAVARES, L. L.; SILVA, R. M.; ALMEIDA, T. A. **Towards Automatically Creating Large Labeled Datasets for Training Question Domain Classifiers**. 2018 International Joint Conference on Neural Networks (IJCNN). **Anais**...2018. Disponível em: <https://doi.org/10.1109/ijcnn.2018.8489124>

TEAM, G. et al. **Gemma: Open Models Based on Gemini Research and Technology**., 2024. Disponível em: <https://arxiv.org/abs/2403.08295>

TEIXEIRA, E. N.; FONSECA, M. C. M.; SOARES, M. E. Resolução do pronome nulo em Português Brasileiro: Evidência de movimentação ocular. **VEREDAS: Sintaxe das Línguas Brasileiras**, v. 18, 2014.

TEIXEIRA, F. et al. Classification of control/pathologic subjects with support vector machines. **Procedia computer science**, v. 138, p. 272–279, 2018.

TEIXEIRA, S. C. S. B.; MARENGO, S. M. D. A.; FINATTO, M. J. B. Construindo fichas terminológicas para estudos sócio-históricos. **Revista Diálogos**, v. 10, n. 3, p. 261–279, 2022.

TEIXEIRA, S. H.; ZAMORA, M. H. Pensando a interseccionalidade a partir da vida e morte de Marielle Franco. **Dignidade Re-Vista**, 2019.

THAKAR, H.; BHATT, B. Fake News Detection: Recent Trends and Challenges. **Social Network Analysis and Mining**, v. 14, 2024.

THOMAS, C. et al. Automatic Detection and Rating of Dementia of Alzheimer Type through Lexical Analysis of Spontaneous Speech. **Proceedings of the IEEE Interna-**

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

tional Conference on Mechatronics and Automation, p. 1569–1574, 2005.

THORNE, J.; VLACHOS, A. **Automated Fact Checking: Task Formulations, Methods and Future Directions**. (E. M. Bender, L. Derczynski, P. Isabelle, Eds.)Proceedings of the 27th International Conference on Computational Linguistics. **Anais**...Santa Fe, New Mexico, USA: Association for Computational Linguistics, ago. 2018.

TIRRELL, L. Toxic Speech: Inoculations and Antidotes. **The Southern Journal of Philosophy**, 2018.

TOHALINO, J. V.; AMANCIO, D. R. Extractive multi-document summarization using multilayer networks. **Physica A: Statistical Mechanics and its Applications**, v. 503, p. 526–539, 2018.

TOLLES, J.; MEURER, W. J. Logistic Regression: Relating Patient Characteristics to Outcomes. **JAMA**, v. 316, n. 5, p. 533–534, ago. 2016.

TORAL, A. et al. **Attaining the Unattainable? Reassessing Claims of Human Parity in Neural Machine Translation**. Proceedings of WMT. **Anais**...Brussels, Belgium: 2018.

TORRENT, T. T. et al. Representing Context in FrameNet: A Multidimensional, Multimodal Approach. **Frontiers in Psychology**, v. Volume 13 - 2022, 2022.

TORRENT, T. T. et al. A flexible tool for a qualia-enriched FrameNet: the FrameNet Brasil WebTool. **Language Resources and Evaluation**, jan. 2024.

TORRENT, T. T.; CONEGLIAN, A. V. L. **Caligrama: Revista de Estudos Românicos**, v. 30, n. 1, p. 8–22, 2025.

TRAAG, V. A.; WALTMAN, L.; ECK, N. J. VAN. From Louvain to Leiden: guaranteeing well-connected communities. **Scientific Reports**, v. 9, n. 1, p. 5233, 2019.

TRAJANO, D.; BORDINI, R. H.; VIEIRA, R. OLID-BR: offensive language identification dataset for Brazilian Portuguese. **Language Resources and Evaluation**, 2023.

TRIFU, R. et al. Linguistic indicators of language in major depressive disorder (MDD). An evidence based research. **Journal of Evidence-Based Psychotherapies**, v. 17, p. 105–128, mar. 2017.

TROTZEK, M.; KOITKA, S.; FRIEDRICH, C. M. Utilizing neural networks and linguistic metadata for early detection of depression indications in text sequences. **IEEE Transactions on Knowledge and Data Engineering**, 2018.

TSOUMAKAS, G.; KATAKIS, I. Multi-label classification: An overview. **International Journal of Data Warehousing and Mining**, v. 3, n. 3, p. 1–13, 2007.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

TSUGAWA, S. et al. **Recognizing Depression from Twitter Activity**. 33rd Annual ACM Conference on Human Factors in Computing Systems. **Anais**...New York, USA: Association for Computing Machinery, 2015.

TURCHIOE, M. R. et al. Systematic review of current natural language processing methods and applications in cardiology. **Heart**, v. 108, n. 12, p. 909–916, 2022.

UCHIDA, H.; ZHU, M.; DELLA SENTA, T. A gift for a millennium. **IAS/UNU, Tokyo**, 1999.

USZKOREIT, H.; LOMMEL, A. **Multidimensional Quality Metrics: A New Unified Paradigm for Human and Machine Translation Quality Assessment**. [s.l: s.n.].

VAJJALA, S.; MEURERS, D. Readability-based Sentence Ranking for Evaluating Text Simplification. **CoRR**, v. abs/1603.06009, 2016.

VAN DER LAAN, R. H. **Tesauro e terminologia: uma inter-relação lógica**. Tese (Doutorado em Letras)—Porto Alegre: Universidade Federal do Rio Grande do Sul; Instituto de Letras, Programa de Pós-Graduação em Letras, 2002.

VARELA, P. J.; JUSTINO, E. J.; OLIVEIRA, L. E. **Identificação de Autoria de Textos através do uso de Classes Linguísticas da Língua Portuguesa (Authorship Identification Using Linguistic Classes for Portuguese)[in Portuguese**]. Proceedings of the 8th Brazilian Symposium in Information and Human Language Technology. **Anais**...2011.

VARGAS, D. F.; VAN DER LANN, R. H. A contribuição da terminologia na construção de linguagens documentárias como os tesauros. **Biblos**, v. 25, n. 1, p. 21–34, 2011.

VARGAS, F. et al. **HateBR: A Large Expert Annotated Corpus of Brazilian Instagram Comments for Offensive Language and Hate Speech Detection**. Proceedings of the Thirteenth Language Resources and Evaluation Conference. **Anais**...b2022.

VARGAS, F. et al. **HateBR: A Large Expert Annotated Corpus of Brazilian Instagram Comments for Offensive Language and Hate Speech Detection**. Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022). **Anais**...Marseille, France: European Language Resources Association, a2022. Disponível em: <https://aclanthology.org/2022.lrec-1.777>

VARGAS, F. et al. **Rhetorical Structure Approach for Online Deception Detection: A Survey**. (N. Calzolari et al., Eds.)Proceedings of the Thirteenth Language Resources and Evaluation Conference. **Anais**...Marseille, France: European Language Resources Association, jun. c2022. Disponível em: <https://aclanthology.org/2022.lrec-1.635>

VARGAS, F. et al. **Predicting Sentence-Level Factuality of News and Bias of Media Outlets**. (R. Mitkov, G. Angelova, Eds.)Proceedings of the 14th International

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Conference on Recent Advances in Natural Language Processing. **Anais**...Varna, Bulgaria: INCOMA Ltd., Shoumen, Bulgaria, set. 2023. Disponível em: <https://aclanthology.org/2023.ranlp-1.127>

VARGAS, F. et al. **Improving Explainable Fact-Checking via Sentence-Level Factual Reasoning**. (M. Schlichtkrull et al., Eds.)Proceedings of the Seventh Fact Extraction and VERification Workshop (FEVER). **Anais**...Miami, Florida, USA: Association for Computational Linguistics, nov. 2024. Disponível em: <https://aclanthology.org/2024.fever-1.23/>

VARGAS, F. A.; SANTOS, R. S. S. D.; ROCHA, P. R. **Identifying Fine-Grained Opinion and Classifying Polarity on Coronavirus Pandemic**. Proceedings of the Brazilian Conference on Intelligent Systems. **Anais**...2020.

VARGAS, F.; PARDO, T.; BENEVENUTO, F. **Socially Responsible and Explainable Automated Fact-Checking and Hate Speech Detection**. Anais do XXXVIII Concurso de Teses e Dissertações. **Anais**...Porto Alegre, RS, Brasil: SBC, 2025. Disponível em: <https://sol.sbc.org.br/index.php/ctd/article/view/36355>

VAROL, O. et al. **Online human-bot interactions: Detection, estimation, and characterization**. Proceedings of the International AAAI Conference on Web and Social Media. **Anais**...AAAI Press, 2017.

VASWANI, A. et al. **Attention is All you Need**. (I. Guyon et al., Eds.)Advances in Neural Information Processing Systems. **Anais**...Curran Associates, Inc., b2017. Disponível em: <https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html>

VASWANI, A. et al. Attention is all you need. **Advances in neural information processing systems**, v. 30, a2017.

VIEIRA, C.; SOUZA, J.; CAVALCANTI, G. **Detecção de Fake News em Português: Análise Comparativa entre Métodos de Representação em Português, Inglês e Multilíngues**. Anais do XIV Brazilian Workshop on Social Network Analysis and Mining. **Anais**...Porto Alegre, RS, Brasil: SBC, 2025. Disponível em: <https://sol.sbc.org.br/index.php/brasnam/article/view/36378>

VIEIRA, J. M. M. **The Brazilian Portuguese eye tracking corpus with a predictability study focusing on lexical and partial prediction**. mathesis—Universidade Federal do Ceará, Biblioteca Universitária: Federal University of Ceará (UFC), 2020.

VIEIRA, R. et al. Enriching the 1758 Portuguese Parish Memories (Alentejo) with Named Entities. **Journal of Open Humanities Data**, v. 7, p. 20, 2021.

VILAR, D. et al. **Error Analysis of Statistical Machine Translation Output**. Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06). **Anais**...Genoa, Italy: European Language Resources Association (ELRA), 2006. Disponível em: <http://www.lrec-conf.org/proceedings/lrec2006/pdf/413_pdf.pd

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

f>

VILLAR, G. S.; FINATTO, M. J. B. Acessibilidade textual e terminológica: novos glossários sobre oncologia para a ferramenta MedSimples. **Mandinga-Revista de Estudos Linguísticos (ISSN: 2526-3455)**, v. 7, n. 2, p. 23–42, 2023.

VIMAL, B. et al. **MFCC Based Audio Classification Using Machine Learning**. 2021 12th International Conference on Computing Communication and Networking Technologies (ICCCNT). **Anais**...2021.

VIRIDIANO, M. et al. **Framed Multi30K: A Frame-Based Multimodal-Multilingual Dataset**. (N. Calzolari et al., Eds.)Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024). **Anais**...Torino, Italia: ELRA; ICCL, 2024. Disponível em: <https://aclanthology.org/2024.lrec-main.656/>

VITÓRIO, D. et al. **Ulysses-RFSQ: A Novel Method to Improve Legal Information Retrieval Based on Relevance Feedback**. (J. C. Xavier-Junior, R. A. Rios, Eds.)Intelligent Systems. **Anais**...Cham: Springer International Publishing, 2022.

VITÓRIO, D. et al. Building a relevance feedback corpus for legal information retrieval in the real-case scenario of the Brazilian Chamber of Deputies. **Language Resources and Evaluation**, 2024.

VLACHOS, A.; RIEDEL, S. **Fact checking: Task definition and dataset construction**. Proceedings of the ACL 2014 Workshop on Language Technologies and Computational Social Science. **Anais**...Association for Computational Linguistics, 2014.

VOSOUGHI, S.; ROY, D.; ARAL, S. The spread of true and false news online. **Science**, v. 359, n. 6380, p. 1146–1151, 2018.

WAGNER FILHO, J. A. et al. **The brWaC Corpus: A New Open Resource for Brazilian Portuguese**. (N. Calzolari et al., Eds.)Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018). **Anais**...Miyazaki, Japan: European Language Resources Association (ELRA), 2018. Disponível em: <https://aclanthology.org/L18-1686>

WANDERLEY, M. G. et al. **A Moving Target: Detecting Concept Drift in Brazilian Portuguese Fake News**. Proceedings of the 16th Symposium in Information and Human Language Technology (STIL'2025). **Anais**...Fortaleza, CE, Brazil: SBC, out. 2025.

WANG, L. et al. **Relation classification via multi-level attention cnns**. Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). **Anais**...2016.

WANG, S. et al. **Want To Reduce Labeling Cost? GPT-3 Can Help**. (M.-F. Moens et al., Eds.)Findings of the Association for Computational Linguistics: EMNLP 2021.

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

Anais...Punta Cana, Dominican Republic: Association for Computational Linguistics, nov. 2021.

WANG, S. et al. **GPT-NER: Named Entity Recognition via Large Language Models**., 2023. Disponível em: <https://arxiv.org/abs/2304.10428>

WANG, W. Y. **"Liar, Liar Pants on Fire": A New Benchmark Dataset for Fake News Detection**. Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). **Anais**...Vancouver, Canada: Association for Computational Linguistics, jul. 2017.

WARDLE, C.; DERAKHSHAN, H. **Information Disorder: Toward an Interdisciplinary Framework for Research and Policy Making**. [s.l.] Council of Europe, 2017. Disponível em: <https://rm.coe.int/information-disorder-toward-an-interdisciplinary-framework-for-researc/168076277c>.

WASSERMAN, S.; FAUST, K. **Social network analysis: Methods and applications**. [s.l.] Cambridge university press, 1994.

WATANABE, W. M. et al. **Facilita: helping the reading of texts available on the web**. XV Brazilian Symposium on Multimedia and the Web, WebMedia '09, Fortaleza, Ceará, Brazil, October 5-7, 2009. **Anais**...a2009. Disponível em: <http://doi.acm.org/10.1145/1858477.1858516>

WATANABE, W. M. et al. **Facilita: reading assistance for low-literacy readers**. Proceedings of the 27th Annual International Conference on Design of Communication, SIGDOC 2009, Bloomington, Indiana, USA, October 5-7, 2009. **Anais**...b2009. Disponível em: <http://doi.acm.org/10.1145/1621995.1622002>

WATASE, T. et al. Severity Classification Using Dynamic Time Warping–Based Voice Biomarkers for Patients With COVID-19: Feasibility Cross-Sectional Study. **JMIR Biomedical Engineering**, v. 8, n. 1, p. e50924, 2023.

WATTS, D. J.; STROGATZ, S. H. Collective dynamics of "small-world" networks. **Nature**, v. 393, n. 6684, p. 440–442, 1998.

WAY, A. Quality Expectations of Machine Translation. Em: MOORKENS, J. et al. (Eds.). **Translation Quality Assessment: From Principles to Practice**. Cham: Springer International Publishing, 2018. p. 159–178.

WAY, A.; FORCADA, M. L. Editors' foreword to the invited issue on SMT and NMT. **Machine Translation**, v. 32, n. 3, p. 191–194, set. 2018.

WEI, J. et al. Emergent abilities of large language models. **arXiv preprint arXiv:2206.07682**, 2022.

WEI, X. et al. **ChatIE: Zero-Shot Information Extraction via Chatting with ChatGPT**., 2024. Disponível em: <https://arxiv.org/abs/2302.10205>

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

WHO. **Comprehensive mental health action plan 2013–2030**. [s.l.] World Health Organization; World Health Organization, 2021.

WHO. **World Health Organization**. World Health Organization, 2024. Disponível em: <https://www.who.int/health-topics/violence-against-women#tab=tab_2>. Acesso em: 6 abr. 2024

WILKINSON, M.; DUMONTIER, M.; AALBERSBERG, ET AL. The FAIR Guiding Principles for scientific data management and stewardship. **Scientific data**, v. 3, n. 1, p. 1–9, 2016.

WILKINSON, M.; DUMONTIER, M.; SANSONE, ET AL. Evaluating FAIR maturity through a scalable, automated, community-governed framework. **Sc. Data**, v. 6, n. 1, p. 1–12, 2019.

WIVES, L. K. **Técnicas de Recuperação de Informações Com Ênfase em Informações Textuais**. tese de doutorado—[s.l.] Universidade Federal do Rio Grande do Sul, 1997.

WOLFRAM, W. Variation and Language: Overview. Em: BROWN, K. (Ed.). **Encyclopedia of Language & Linguistics (Second Edition)**. Second Edition ed. Oxford: Elsevier, 2006. p. 333–341.

WOLINSKI, F.; VICHOT, F.; DILLET, B. **Automatic processing proper names in texts**. Proc. Conference on European Chapter of the Association for Computational Linguistics. **Anais**...EACL, 1995.

WU, H. et al. SemEHR: A general-purpose semantic search system to surface semantic data from clinical notes for tailored care, trial recruitment, and clinical research. **J Am Med Inform Assoc**, v. 25, n. 5, p. 530–537, 2018.

WU, Y. et al. Google's neural machine translation system: Bridging the gap between human and machine translation. **arXiv preprint arXiv:1609.08144**, 2016.

XAVIER, C. C.; LIMA, V. L. S. DE; SOUZA, M. Open information extraction based on lexical semantics. **Journal of the Brazilian Computer Society**, v. 21, n. 1, p. 1–14, 2015.

XAVIER, R. C. **Português no Direito: Linguagem Forense**. Rio de Janeiro: Forense, 2002. p. 1

XIA, P. et al. **LOME: Large Ontology Multilingual Extraction**. (D. Gkatzia, D. Seddah, Eds.)Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations. **Anais**...Online: Association for Computational Linguistics, abr. 2021. Disponível em: <https://aclanthology.org/2021.eacl-demos.19>

XIA, T.; HAN, J.; MASCOLO, C. Exploring machine learning for audio-based respiratory

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

condition screening: A concise review of databases, methods, and open issues. **Experimental Biology and Medicine**, v. 247, n. 22, p. 2053–2061, 2022.

XIAO, C. et al. **Human-AI Collaborative Essay Scoring: A Dual-Process Framework with LLMs**. Proceedings of the 15th International Learning Analytics and Knowledge Conference. **Anais**...: LAK '25.New York, NY, USA: Association for Computing Machinery, 2025.

XU, S. et al. Automated Lightweight Model for Asthma Detection Using Respiratory and Cough Sound Signals. **Diagnostics**, v. 15, n. 9, 2025.

XU, W.; CALLISON-BURCH, C.; NAPOLES, C. Problems in Current Text Simplification Research: New Data Can Help. **Transactions of the Association for Computational Linguistics**, v. 3, p. 283–297, 2015.

YADAV, S. et al. **Identifying Depressive Symptoms from Tweets: Figurative Language Enabled Multitask Learning Framework**. 28th International Conference on Computational Linguistics. **Anais**...Barcelona, Spain (Online): International Committee on Computational Linguistics, 2020.

YAN, M. Y.; GUSTAD, L. T.; NYTRØ, Ø. Sepsis prediction, early detection, and identification using clinical text for machine learning: a systematic review. **J Am Med Inform Assoc**, v. 29, n. 3, p. 559–575, jan. 2022.

YANG, H. et al. Clinical Trial Classification of SNS24 Calls with Neural Networks. **Future Internet**, v. 14, n. 5, p. 130, 2022.

YANG, K.-C. et al. Scalable and generalizable social bot detection through data selection. **Proceedings of the AAAI Conference on Artificial Intelligence**, v. 34, n. 01, p. 1096–1103, a2020.

YANG, P.; FANG, H.; LIN, J. **Anserini: Enabling the use of lucene for information retrieval research**. Proceedings of the 40th international ACM SIGIR conference on research and development in information retrieval. **Anais**...2017.

YANG, R. et al. **Enhancing automated essay scoring performance via fine-tuning pre-trained language models with combination of regression and ranking**. Association for Computational Linguistics (ACL), b2020.

YASUNAGA, M. et al. **Deep Bidirectional Language-Knowledge Graph Pretraining**., 2022. Disponível em: <https://arxiv.org/abs/2210.09338>

YATES, A.; COHAN, A.; GOHARIAN, N. **Depression and Self-Harm Risk Assessment in Online Forums**. Conference on Empirical Methods in Natural Language Processing. **Anais**...Copenhagen, Denmark: Association for Computational Linguistics, 2017.

YAZDAVAR, A. H. et al. **Semi-Supervised Approach to Monitoring Clinical De-**

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

pressive Symptoms in Social Media. IEEE/ACM International Conference on Advances in Social Network Analysis and Mining. **Anais**...2017.

YU, X.; LAM, W. **Jointly identifying entities and extracting relations in encyclopedia text via a graphical model approach**. Coling 2010: Posters. **Anais**...2010.

YUAN, W.; NEUBIG, G.; LIU, P. **BARTScore: Evaluating Generated Text as Text Generation**. Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual. **Anais**...a2021. Disponível em: <https://proceedings.neurips.cc/paper/2021/hash/e4d2b6e6fdeca3e60e0f1a62fee3d9dd-Abstract.html>

YUAN, Y. et al. **A relation-specific attention network for joint entity and relation extraction**. International joint conference on artificial intelligence. **Anais**...International Joint Conference on Artificial Intelligence, b2021.

YUE, Z. et al. **Evidence-Driven Retrieval Augmented Response Generation for Online Misinformation**. (K. Duh, H. Gomez, S. Bethard, Eds.)Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers). **Anais**...Mexico City, Mexico: Association for Computational Linguistics, jun. 2024. Disponível em: <https://aclanthology.org/2024.naacl-long.313>

ZANUZ, L.; RIGO, S. J. **Fostering Judiciary Applications with New Fine-Tuned Models for Legal Named Entity Recognition in Portuguese**. (V. Pinheiro et al., Eds.)Computational Processing of the Portuguese Language. **Anais**...Cham: Springer International Publishing, 2022.

ZAROCOSTAS, J. How to fight an infodemic. **The lancet**, v. 395, n. 10225, p. 676, 2020.

ZELENINA, M. **Eye Tracking for NLP**. **SlideShare**, 2015. Disponível em: <https://www.slideshare.net/mariezelenina/presentation-2-47610828>

ZELENKO, D.; AONE, C.; RICHARDELLA, A. Kernel methods for relation extraction. **Journal of machine learning research**, v. 3, n. Feb, p. 1083–1106, 2003.

ZENG, D. et al. **Relation classification via convolutional deep neural network**. Proceedings of COLING 2014, the 25th international conference on computational linguistics: technical papers. **Anais**...2014.

ZHANG, A. et al. **Dive into Deep Learning**. [s.l.] Cambridge University Press, 2023.

ZHANG, H. **The Optimality of Naive Bayes**. Proceedings of the Seventeenth International Florida Artificial Intelligence Research Society Conference. **Anais**...2004.

ZHANG, S. et al. Opt: Open pre-trained transformer language models. **arXiv preprint arXiv:2205.01068**, 2022.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

ZHANG, S. X. et al. Predictors of Depression and Anxiety Symptoms in Brazil during COVID-19. **Int J Environ Res Public Health**, v. 18, n. 13, 30 jun. 2021.

ZHANG, S.; DUH, K.; VAN DURME, B. **Mt/ie: Cross-lingual open information extraction with neural sequence-to-sequence models**. Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers. **Anais**...2017.

ZHANG, T. et al. **BERTScore: Evaluating Text Generation with BERT**. 8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020. **Anais**...OpenReview.net, 2020. Disponível em: <https://openreview.net/forum?id=SkeHuCVFDr>

ZHANG, Y. et al. **A Comprehensive Survey on Process-Oriented Automatic Text Summarization with Exploration of LLM-Based Methods**., 2025. Disponível em: <https://arxiv.org/abs/2403.02901>

ZHAO, S.; GRISHMAN, R. **Extracting relations with integrated information using kernel methods**. Proceedings of the 43rd annual meeting of the association for computational linguistics (acl'05). **Anais**...2005.

ZHONG, H. et al. **How Does NLP Benefit Legal System: A Summary of Legal Artificial Intelligence**. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. **Anais**...Association for Computational Linguistics, 2020.

ZHOU, L. et al. A comparison of classification methods for predicting deception in computer-mediated communication. **Journal of Management Information Systems**, v. 20, n. 4, p. 139–165, 2004.

ZHOU, L. An empirical investigation of deception behavior in instant messaging. **IEEE Transactions on Professional Communication**, v. 48, n. 2, p. 147–160, 2005.

ZHUANG, F. et al. A comprehensive survey on transfer learning. **Proceedings of the IEEE**, v. 109, n. 1, p. 43–76, 2020.

ZILIO, L.; FINATTO, M. J.; VIEIRA, R. **Named Entity Recognition Applied to Portuguese Texts from the XVIII Century**. (C. Trojahn et al., Eds.)Proceedings of the Second Workshop on Digital Humanities and Natural Language Processing (2nd DHandNLP 2022) co-located with International Conference on the Computational Processing of Portuguese (PROPOR 2022), Virtual Event, Fortaleza, Brazil, 21st March, 2022. **Anais**...: CEUR Workshop Proceedings.CEUR-WS.org, 2022. Disponível em: <http://ceur-ws.org/Vol-3128/paper10.pdf>

ZILIO, L.; LAZZARI, R. R.; FINATTO, M. J. B. NLP for historical Portuguese: Analysing 18th-century medical texts. **Proceedings of the International Conference on the Computational treatment of Portuguese, PROPOR**, 2024.

ZIPF, G. K. **Human behavior and the principle of least effort: an introduction**

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

to human ecology. [s.l.] Addison-Wesley Press, 1949.

ZOBEL, J. **How reliable are the results of large-scale information retrieval experiments?** Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval. **Anais**...ACM, 1998.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Apêndice A

Considerações prévias à sumarização automática

Apêndice do Capítulo 5

Jackson Wilke da Cruz Souza
Paula Christina Figueira Cardoso
Crysttian Arantes Paixão

Publicado em: 13/03/2024

Ao analisarmos os métodos e sistemas de sumarização, torna-se necessário estabelecer e detalhar alguns procedimentos importantes para trabalhar com sumarização de textos. Muitas vezes, o usuário de sistemas de SA tem em mente o que deseja realizar com o seu texto, porém nem sempre se atenta aos detalhes que podem influenciar no processamento, análise e resultados.

A) Formalidades e variações da língua

O primeiro detalhe é conhecer a qualidade do texto que se deseja fazer a sumarização. A origem do texto determina a qualidade e o padrão da grafia textual. Um texto pode vir de diferentes fontes para compor um *corpus* ou corpora; Por exemplo, em textos de redes sociais, nos quais o processo de escrita não segue, necessariamente a variedade, é formal da língua e possui limitação de caracteres, dependendo do público que gerou as mensagens, demandará uma atenção especial. Uma base de texto que siga uma padronização, de preferência normalizado, reduzirá em muito o tratamento das inconsistências, otimizando a análise (cf. Capítulo 2).

Problemas envolvendo acentuação são os mais comuns, desconsiderando o problema de grafia. Vamos ao exemplo do uso da palavra "não". Em mensagens do Twitter, um usuário pode escrevê-la de várias formas para transmitir a mensagem, como "n", "não" e "nao", sendo que todas essas variações correspondem ao "não". Para resolver esse problema, o uso de dicionário seria uma opção para corrigir as palavras (corretor ortográfico), como existem nos editores de textos tradicionais.

B) Obtenção dos textos-fonte

Dependendo da origem dos textos, alguns problemas podem ser encontrados. No caso das redes sociais, quando se possui uma interface de programação de aplicações - API, o texto é coletado de forma automatizada. Porém, os problemas com a falta de padronização ainda podem persistir.

Alguns pesquisadores realizam estudos envolvendo apenas textos que possuem uma normalização, exemplo os que analisam trabalhos de conclusão de curso - TCC, dissertações e teses. Existem ainda os que utilizam artigos, de revistas e jornais, para realizarem suas pesquisas.

Nesses textos, por conta da formalização textual, muitos problemas são minimizados. Entretanto, eles ainda podem ocorrer. Um exemplo é quando a base é composta por

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

textos que seguem a formação dos programas Word da Microsoft, Write do OpenOffice, documentos do Google e em PDF (*Portable Document Format*). Esses formatos utilizam uma linguagem para a formatação dos textos que não é visível ao usuário, como *Extensible Markup Language* - XML do docx do Word, que tem que ser tratada no momento da importação para análise, pois apenas o texto deve compor o *corpus*. Para visualizar a fonte de um arquivo docx, sugerimos que o arquivo docx analisado seja aberto em um editor de texto simples, como o bloco de notas do Windows. No caso de arquivos no formato PDF, a conversão também é necessária. Existem ainda textos que são oriundos de páginas da internet que são capturadas, nos quais a linguagem que estrutura o site na maioria das vezes, é a Linguagem de Marcação de HiperTexto - HTML.

Esse exemplo é fácil de se constatar. Abra uma página da internet, selecione o texto e cole no Word. Você notará que a formatação das palavras também foi mantida. A pergunta é: como? A resposta é simples: junto com o texto desejado, você copiou um texto de marcação que irá compor o seu documento; logo, este deve ser tratado antes de realizar a análise.

Uma forma de contornar esse problema é copiar o texto da internet, colar em um editor de textos simples e, em seguida, copiar novamente o texto do editor para o Word. Dessa forma,a linguagem invisível ao olhar humano é removida. Porém, trata-se de um serviço manual e dependente da quantidade de textos, o que o torna inviável. Assim, em algum momento nesse processo, será necessário utilizar recursos computacionais para o processamento textual. Em suma, independente da fonte, deve-se realizar uma limpeza dos arquivos, o que usualmente recebe o nome de pré-processamento, deixando apenas o texto a ser analisado.

Uma sugestão é que os textos a serem analisados, na medida do possível, estejam em formato de texto puro, como os criados em blocos de notas. Recomenda-se sempre que, após a importação ou carregamento de uma base textual, uma inspeção visual seja realizada em uma amostra do texto para a verificação de possíveis problemas.

C) Fator cultural

Dependendo da região do país, existem fatores culturais que impactam diretamente a forma de expressão, como o uso de gírias. Portanto, em alguns casos, como desejamos monitorar o que está repercutindo, devemos padronizar essa informação. Para isso, torna-se necessário o uso de léxicos de gírias. O léxico serve como uma orientação na troca das palavras de mesmo significado para se estabelecer um padrão. Basicamente, o léxico funcionaria como um "corretor ortográfico", estabelecendo o padrão textual para análise. Como exemplo, temos o uso dos termos "pão de sal", "pão", "pãozinho", "cassetinho", entre outros, que referem-se a mesma coisa.

O desafio imposto é a obtenção desses léxicos. Na maioria dos casos, o usuário deverá criá-lo, do zero ou a partir de algum outro que contemple os termos que deseja, pois dificilmente encontrará, salvo exceções, um léxico específico para sua aplicação.

D) Emojis

Outro fato relevante em textos de redes sociais é o uso de emojis, ideogramas e smileys. O usuário se vale de tais recursos em suas mensagens para transmitir uma informação relevante, o que em determinada parte do texto pode dificultar o processo de sumarização, demandando configurações especiais. No processo de SA, os emojis serão considerados como palavras e, dependendo da análise, deverá ser utilizado um dicionário de sinônimos para descrevê-los.

Além disso, consideremos outros sistemas, nos quais o usuário de certas redes sociais pode se expressar diretamente com outros elementos multimodais (como vídeos e fotos).

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

A interação dos usuários também leva em conta o regionalismo, que torna necessário a adaptação de um léxico para tentar minimizar o efeito de idiossincrasias em função da informalidade textual. As temáticas abordadas devem ser consideradas para determinar se o texto pode compor um sumário.

E) Expressões regulares

Aos leitores interessados em análise de textos, recomendamos um estudo, pelo menos introdutório, sobre expressões regulares. Como referência, indicamos o livro de Expressões Regulares (Jargas, 2016). O livro apresenta de forma simples os fundamentos de expressões e suas aplicações. Basicamente, uma expressão regular é utilizada para representar um padrão, o qual é muito utilizado dentro das linguagens de programação que podem ser executadas no processamento textual.

Um exemplo é o caso da palavra "não" com diversas ocorrências que podem ser representadas pela expressão regular "[Nn][aã]?o?", referindo-se às variações "Não", "Nao", "N", "não", "nao" e "n". Grande parte desse pré-processamento depende do uso de funções da linguagem de programação, que recebem como parâmetro uma expressão regular para realizar as operações. O uso dessas funções otimiza o processamento textual, permitindo padronizar o texto, reduzindo problemas e viabilizando as análises.

F) Pré-processamento

Até o momento, relatamos o problema envolvendo a criação da base de dados. Porém, antes de analisar o texto, devem ser feitas, como capitalização, acentuação, símbolos de pontuação e *stopwords*, por exemplo

Algumas perguntas podem surgir ao se pensar que uma palavra escrita com letras maiúsculas e minúsculas é diferente. Como exemplo, temos: o computador diferencia "Carro" e "carro"? ou existe um tipo de caixa para otimizar o processamento? Você já sabe a resposta! A forma com a qual são armazenadas é o que as diferencia. Lembre-se sempre que o computador não possui discernimento entre o significado da palavra, mas sim de como ela é representada. Nesse caso, letras maiúsculas e minúsculas possuem codificações diferentes. O mesmo ocorre com a acentuação.

Já para os símbolos de pontuação, em algum determinado momento, o usuário deve escolher qual o caractere que servirá de delimitador de palavras (*tokens,* em PLN). Em alguns idiomas, esse delimitador é o espaço em branco, como na frase "O dia estava lindo, mas pode chover no período da tarde". Facilmente, você consegue identificar as palavras que compõem essa oração. Porém o computador, dependendo da forma como for programado, pode considerar que a palavra "lindo" não esteja presente na frase, por conta da pontuação. Nesse caso "lindo," com destaque para a presença da vírgula, seria a palavra considerada. A vírgula também é processada como um caractere, logo pode influenciar a análise em decorrência que "lindo" é diferente de "lindo,". Isso ocorre muito em nuvens de palavras quando não se atenta pela remoção desses caracteres de pontuação. Logo, uma palavra com determinada frequência poderá ocorrer um número maior de vezes no texto, enviesando a análise.

Mais um detalhe a ser analisado é a forma da escrita que constitui o texto. Apesar de ocorrer em menor frequência, temos alguns itens a serem sempre analisados. O primeiro é o uso de separação silábica, pois a palavra "casa" é diferente de "ca-sa" quando inserida em um texto e processada pelo computador, pois o "-" é um símbolo.

Dependendo da situação, o texto ainda tem de ser pré-processado para eliminar palavras sem valor informacional, comumente denominadas *stopwords*. No caso de SA, nem sempre existe a eliminação dessa "classe" de palavras, mas vale chamar a atenção. Muitas vezes, essa lista de palavras a serem removidas consta de um léxico já padronizado.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Porém, em determinados domínios e públicos torna-se necessário atualizá-lo ou adaptá-lo. Como exemplo, considere um estudo que envolva a palavra-chave "COVID" a ser coletada em mensagens no Twitter. As mensagens que serão coletadas terão a palavra "COVID" em seu texto, sendo que não há adição de informação no seu uso, podendo ser removida do corpus. Todos os textos, de forma geral, terão essa palavra e dependendo da análise, ela é dispensável, ou seja, é interessante removê-la.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Apêndice B

Exemplo de sumarização abstrativa

Apêndice do Capítulo 5

Pedro Henrique Paiola
Gabriel Lino Garcia
João Vitor Mariano Correia
João Renato Ribeiro Manesco
Danilo Samuel Jodas
Douglas Rodrigues
João Paulo Papa

Publicado em: 16/04/2026

Este apêndice apresenta os exemplos de sumarização omitidos do texto do capítulo de sumarização automática.

Quadro B.1: Sumários gerados por cada modelo

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Modelo	Texto
Documento1 Problema em sonda japonesa	Leia mais: Segundo a empresa, quando a sonda atingiu uma altitude de cerca de 100 km, ela ativou corretamente a sequência de descida. No entanto, o software estimou erroneamente que a sua altitude era zero quando ela ainda pairava a cerca de 5 km do chão. Ou seja, a Hakuto-R pensava que já havia pousado, mas seguiu descendo lentamente a uma velocidade inferior a 1 m/s. Por fim, seu sistema de propulsão ficou sem combustível e, apesar de nesse momento a equipe não conseguir mais contato com a espaçonave, acredita-se que ela caiu em queda livre em solo lunar. Confusão é a razão mais provável para esse erro do software. Enquanto voava para o local de pouso, a espaçonave passou pela borda de uma cratera. Lá, o sensor obteve uma leitura de altitude de 3 km e, aparentemente, isso foi maior do que o valor de altitude estimado anteriormente pela equipe que projetou a missão. Por isso, o software da sonda pensou erroneamente que o sensor relatou uma medida anormal e rejeitou-a. Isso ocorreu porque o programa tinha a capacidade de desconsiderar valores anormais de altitude no módulo de pouso como medida de segurança para caso houvesse um problema com o sensor. Com isso, infelizmente, o software falhou ao incorporar o verdadeiro terreno lunar na rota de pouso da espaçonave. Se a aterrisagem da sonda japonesa tivesse sido bem sucedida, essa seria a primeira missão privada e o primeiro veículo construído no Japão a pousar na Lua. Apesar do fracasso, a ispace usará os dados dessa missão para melhorar suas próximas espaçonaves com destino ao satélite natural, com lançamentos previstos em 2024 e 2025. Layse Ventura é jornalista formada pela Uerj e mestre em Engenharia e Gestão do Conhecimento pela Ufsc. Acumula mais de 10 anos de experiência como repórter, copywriter e analista de SEO.
Sumário de referência	Cratera confundiu sonda japonesa que caiu na Lua, revela ispace. Essa seria a primeira missão privada e o primeiro veículo construído no Japão a pousar na Lua.
PTT5-Summ:	Ispace tenta pousar na Lua com erro de software. O software da ispace tentou incorporar o terreno lunar na rota de pouso da espaçonave.
Qwen3-8B:	A sonda lunar Hakuto-R, desenvolvida pela empresa ispace, falhou em sua missão de pouso na Lua devido a um erro no software. Quando atingiu cerca de 100 km de altitude, o sistema ativou a sequência de descida, mas estimou incorretamente a altitude como zero, mesmo que a sonda estivesse a 5 km do solo. Como resultado, a sonda continuou descendo lentamente, sem combustível, e caiu em queda livre. O erro provavelmente ocorreu porque o sensor de altitude capturou uma leitura de 3 km em uma cratera, que foi rejeitada pelo software como anormal, levando a uma falha na navegação. Apesar do fracasso, os dados da missão serão usados para melhorar futuras missões, previstas para 2024 e 2025.
InternLM-Chatbode-7B:	A sonda japonesa Hakuto-R, que foi lançada para a Lua em 2021, caiu em queda livre após um erro de software que estimou erroneamente que a sonda havia pousado. O software estimou que a sonda tinha pousado quando ela ainda estava voando a uma distância de 5 km do chão. O sistema de propulsão da sonda ficou sem combustível e ela caiu em queda livre. A ispace, a empresa que operou a missão, usará os dados dessa missão para melhorar suas próximas espaçonaves com destino ao satélite natural, com lançamentos previstos em 2024 e 2025.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Sabiá-7B:	A missão Hakuto-R, da empresa ispace, foi a primeira missão privada a pousar na Lua. A sonda japonesa, que foi lançada em 2019, tinha como objetivo realizar um pouso suave na superfície lunar. No entanto, o pouso não foi bem sucedido, e a sonda acabou caindo em uma cratera. Apesar do fracasso, a ispace usará os dados dessa missão para melhorar suas próximas espaçonaves com destino ao satélite natural, com lançamentos previstos em 2024 e 2025. Fonte: https://www.tecmundo.com.br/ciencia/229069-missao-hakuto-r-falha-na-lua-e-ispace-vai-usar-dados-para-melhorar-missao-de-2024.htm Texto 2: A missão Hakuto-R, da empresa ispace, foi a primeira missão privada

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Sobre as/os autoras/es

Helena de Medeiros Caseli

Helena Caseli é formada em Ciência da Computação pela UFU, com pós-graduação no ICMC/USP. É professora titular na UFSCar, onde atua desde 2008 e coordena o LALIC (Laboratório de Linguística e Inteligência Computacional). Suas áreas de interesse incluem aprendizado de máquina, tradução automática e PLN aplicado a redes sociais em domínios como política e saúde mental. Foi a idealizadora, é uma das organizadoras deste livro e colaborou com os Capítulos O que é PLN?, Sequência de caracteres e palavras, E agora, PLN?, 4 e 10.
Email: helenacaseli@ufscar.br

Maria das Graças Volpe Nunes

Maria das Graças Volpe Nunes é formada em Ciência da Computação pela UFSCar, mestre pela USP e doutora pela PUC-Rio. Foi docente no ICMC-USP, São Carlos, de 1981 a 2013, onde agora é professora sênior. Atualmente é membro do C4AI, da USP, IBM e FAPESP. É uma das fundadoras do NILC, onde atuou em várias áreas do PLN para português: revisão ortográfica e gramatical, tradução automática, sumarização, análise de sentimentos, entre outras. É uma das organizadoras deste livro e colaborou com os Capítulos O que é PLN?, Questões éticas em IA e PLN e E agora, PLN?.
Email: gracan@icmc.usp.br

Adriana Silvina Pagano

Adriana Pagano é formada em Letras pela Universidad Nacional de La Plata. Fez seu mestrado no Programa de Inglês e Literatura Correspondente da UFSC e seu doutorado no Programa de Estudos Linguísticos da UFMG. É professora titular da Faculdade de Letras da UFMG desde 1995. Suas áreas de interesse incluem tradução, produção textual multilíngue e modelagem sistêmico-funcional da linguagem. Neste livro, colaborou com os Capítulos O que é PLN?, A ordem e a função das palavras em uma sentença, Ferramentas e recursos para o processamento sintático e 9.
Email: apagano@ufmg.br

Adriano L. I. Oliveira

Adriano Oliveira é engenheiro eletrônico (1993), mestre e doutor em Computação pela UFPE (1997 e 2004). Foi Professor Visitante da École de Technologie Supérieure (ÉTS, Université du Québec, Canadá), 2018-2019. É Professor da UFPE e bolsista de produtividade em pesquisa do CNPq, nível 1D. Neste livro, colaborou com o Capítulo 13.
Email: alio@cin.ufpe.br

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Aline Macohin

Aline Macohin é formada em Direito e Computação, com Mestrado em Computação Aplicada pela UTFPR e Doutorado em Direito pela UFPR. É advogada, analista de sistemas no poder público há 13 anos, membro da Comissão de Direito Digital e Proteção de dados da OAB/PR e professora de pós-graduação na PUCPR. Autora do site https://www.iaresponsavel.com.br. Suas áreas de interesse incluem aprendizado de máquina, PLN, governança de inteligência artificial e proteção de dados. Neste livro, colaborou com o Capítulo 12.
Email: aline.macohin@gmail.com

Amanda Pontes Rassi

Amanda Rassi é bacharel e mestre em Linguística pela Universidade Federal de Goiás (UFG) e doutora na mesma área pela Universidade Federal de São Carlos (UFSCar). Tem experiência com correção automática de redações, tendo atuado na Redação Nota 1000 e atualmente na Somos Educação. Suas áreas de interesse incluem sintaxe, semântica, lexicologia, lexicografia e PLN. Neste livro, colaborou com os Capítulos Sequência de caracteres e palavras, A ordem e a função das palavras em uma sentença e 7.
Email: amanda.rassi@somoseducacao.com.br

Ana Paula Banza

Ana Paula Banza é Licenciada em Línguas e Literaturas Modernas pela Universidade de Lisboa (Portugal); Doutora em Linguística pela Universidade de Évora (Portugal), onde lecciona, no Departamento de Linguística e Literaturas, desde 1991, sendo, atualmente, professora associada com agregação. Desenvolve investigação nas áreas da Filologia, Crítica Textual, Linguística Histórica/História da Língua e Variação linguística, centrando-se os seus interesses na edição crítica de textos literários e não literários, com destaque para a obra do Padre António Vieira, e no estudo dos fenómenos de mudança e de variação na língua portuguesa. Neste livro, colaborou com o Capítulo 15.
Email: anabanza@uevora.pt

Ana Sofia Ribeiro

Ana Sofia Ribeiro é formada em História e doutorada em História pela Universidade do Porto. É investigadora auxiliar do CIDEHUS, Universidade de Évora, Portugal. As suas áreas de interesse estão ligadas aos métodos de utilização de PLN aos textos históricos e como os resultados de técnicas como o reconhecimento de entidades nomeadas (NER) ou grafos de conhecimento (*knowledge graphs*) podem ser usados como variáveis da investigação em história. Neste livro, colaborou com o Capítulo 15.
Email: asvribeiro@uevora.pt

André Carlos Ponce de Leon Ferreira de Carvalho

André Carlos de Carvalho é professor titular do ICMC-USP e Vice-Diretor do Centro de Inteligência Artificial e Aprendizado de Máquina (CIAAM) da USP. Seus principais

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

interesses de pesquisa são Aprendizado de Máquina, Mineração de Dados e Ciência de Dados, com aplicações em várias áreas. Escreveu vários livros, entre eles *Inteligência Artificial: Uma abordagem de Aprendizado de Máquina*, publicado pelo GrupoGen em 2011 e prêmio Jabuti 2012, e *A General Introduction to Data Analytics*, publicado pela Wiley, em 2018. Neste livro, colaborou com o Capítulo 13.
Email: andre@icmc.usp.br

Beatriz Raposo de Medeiros

Beatriz Raposo de Medeiros é professora associada de Fonética da Universidade de São Paulo (USP), onde coordena o Laboratório de Fonética e Linguagem Lafalin. É doutora em Linguística pela Universidade Estadual de Campinas (Unicamp). Realizou pós-doutorado no Laboratoire Langue Parole, na Université Aix-Marseille em 2007. Em 2023 recebeu a BPE, bolsa de pesquisa da FAPESP, em período como professora visitante da University of Southern California, para estudo da laringe e sua relação mais estreita com capacidades entoacionais humanas. Seu trabalho de pesquisa abrange comparações entre fala e canto, tratando aspectos como a inteligibilidade, a organização temporal, a sincronização e a estabilidade da frequência fundamental. Neste livro, colaborou com o Capítulo 11.
Email: biarm@usp.br

Brenda Salenave Santana

Brenda Salenave Santana é formada em Ciência da Computação pela Universidade Federal de Santa Maria (UFSM), com mestrado e doutorado em Computação pela Universidade Federal do Rio Grande do Sul (UFRGS). É docente na Universidade Federal de Pelotas (UFPel) desde 2023. Suas áreas de interesse incluem aprendizado de máquina e detecção de discursos de ódio. Neste livro, colaborou com o Capítulo 16.
Email: bssalenave@inf.ufpel.edu.br

Cassia Trojahn dos Santos

Cassia Trojahn dos Santos é formada em Ciência da Computação pela Unisc, mestre pela Unisinos e doutora pela Universidade de Evora (Portugal), com pós-doutorado pelo INRIA (França). Foi docente do Departamento de Informática e Matemática na Universidade de Grenoble (França) entre 2012 e 2024. Atualmente, é docente na Universidade de Grenoble Alpes, associada ao centro de pesquisa INRIA Alpes. Suas áreas de interesse incluem representação de conhecimento, alinhamento de ontologias, e dados FAIR. Neste livro, colaborou com o Capítulo 15.
Email: cassia.ts@gmail.com

Celso Ricardo Fernandes de Carvalho

Celso Ricardo Fernandes de Carvalho é formado em Fisioterapia pela UFSCar. Atualmente é Professor Associado da Faculdade de Medicina da Universidade de São Paulo (FMUSP). Tem interesse pelas áreas Processamento de Linguagem Natural e Inteligência Artificial em pessoas com doenças respiratórias. Neste livro, colaborou com o Capítulo 11.
Email: cscarval@usp.br

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Claudia Maria Cabral Moro Barra

Claudia Moro é engenheira de computação pela PUCPR, com mestrado (UNICAMP) e doutorado (USP) em engenharia elétrica/biomédica. É professora titular da PUCPR onde atua desde 1995, vinculada ao Programa de Pós-graduação em Tecnologia em Saúde (desde 2003), atuando nas áreas de PLN em textos clínicos, inteligência artificial em saúde e, interoperabilidade entre sistemas de informações em saúde. Neste livro, colaborou com o Capítulo 9.
Email: claudia.moro@gmail.com

Crysttian Arantes Paixão

Crysttian Arantes Paixão é formado em Ciência da Computação pela Universidade Federal de Lavras (UFLA) e Estatística pela Anhembi Morumbi, com pós-graduação em Administração de Redes Linux, Modelagem Computacional e Estatística na UFLA e Pós-Doutorado em Modelagem Computacional pela Fundação Getulio Vargas. É docente na Universidade Federal da Bahia (UFBA) desde 2022. Suas áreas de interesse incluem Estatística, Modelagem Computacional e Inteligência Artificial. Neste livro, colaborou com o Capítulo 5.
Email: crysttian@gmail.com

Daniela Barreiro Claro

Daniela Barreiro Claro é professora titular da Universidade Federal da Bahia (UFBA), fez seu Mestrado em Ciências da Computação pela Universidade Federal de Santa Catarina (UFSC) e o seu Doutorado na Université d'Angers (França). Em 2009, ela fundou o Grupo de Pesquisa FORMAS na área de Semântica e PLN. Suas principais áreas de pesquisa incluem extração de informação, aprendizado de máquina e aprendizado multimodal. Neste livro, colaborou com os Capítulos Representação vetorial e semântica distribucional e 3.
Email: dclaro@ufba.br

Danilo Samuel Jodas

Danilo Jodas é formado em Ciência da Computação pelo Centro Universitário do Norte Paulista (UNORP), com pós-graduação em processamento de imagens e aprendizado de máquina pela UNESP e FEUP (Universidade do Porto). É pesquisador de pós-doutorado na UNESP de Bauru desde 2019, atuando em pesquisas de aprendizado de máquina e PLN. Neste livro, colaborou com os Capítulos 5 e Treinamento de Grandes Modelos de Linguagem na prática.
Email: danilo.jodas@unesp.br

Diego Raphael Amancio

Diego Raphael Amancio é doutor em Física Computacional pela USP (2013), tendo realizado estágios de pós-doutorado na própria USP (2014) e na Indiana University, nos Estados Unidos (2018). Atualmente, é professor associado (livre-docente) e pesquisador do ICMC/USP. Possui experiência na área de Ciência da Computação, com ênfase em

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Redes Complexas, Cienciometria, Aprendizado de Máquina e PLN. Seus interesses de pesquisa incluem a modelagem de sistemas reais por meio de grafos, com destaque para a representação e análise da linguagem humana. Neste livro, colaborou com o Capítulo 18.
Email: diego@icmc.usp.br

Douglas Rodrigues

Douglas Rodrigues é formado em Informática para Gestão de Negócios pela FATEC/Botucatu, Mestrado (UNESP/Bauru) e Doutorado (UFSCar/São Carlos) em Ciências da Computação. Atualmente, atua como pós-doutorando e pesquisador no Recogna/UNESP. Suas áreas de interesse incluem aprendizado de máquina e otimização matemática. Neste livro, colaborou com os Capítulos 5 e Treinamento de Grandes Modelos de Linguagem na prática.
Email: d.rodrigues@unesp.br

Elisa Terumi Rubel Schneider

Elisa Terumi é formada em Informática pela Universidade Tecnológica Federal do Paraná (UTFPR), com especialização em Desenvolvimento Web (PUCPR), mestrado em Bioinformática (UFPR) e doutorado em Informática atuando na linha de pesquisa de Inteligência Artificial (PUCPR). Tem interesse em PLN aplicado na área de saúde. Neste livro, colaborou com os Capítulos Ferramentas e recursos para o processamento sintático e 9.
Email: lisa.terumi@gmail.com

Ellen Souza

Ellen Souza é formada em Ciência da Computação com doutorado em inteligência artificial pelo CIn/UFPE e pós-doutoranda na área de processamento de textos legislativos no ICMC-USP desde 2020. É docente do curso de Bacharelado em Sistemas de Informação da UFRPE desde 2009. Suas áreas de interesse incluem aprendizado de máquina, recuperação de informação e reconhecimento de entidades nomeadas. Neste livro, colaborou com o Capítulo 13.
Email: ellen.ramos@ufrpe.br

Evelin Amorim

Evelin Amorim é formada em Ciência da Computação pela Universidade Federal do Espírito Santo (UFES), com doutorado em Ciência da Computação pela Universidade Federal de Minas Gerais (UFMG). É pesquisadora no INESC TEC, em Portugal, desde 2020. Suas áreas de interesse incluem correção automática de redações e extração de informação. Neste livro, colaborou com o Capítulo 1.
Email: evelin.amorim@gmail.com

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Fátima Farrica

Fátima Farrica é doutorada em História pela Universidade de Évora (Portugal), com pós-graduação em Ciências Documentais: Arquivologia. É bolseira de investigação do CIDEHUS-Universidade de Évora (Portugal). As suas áreas de interesse incluem a investigação na área da História, projetos de arquivística, paleografia e tratamento documental. Neste livro, colaborou com o Capítulo 15.
Email: ffarrica@uevora.pt

Felipe Ribas Serras

Felipe Ribas Serras é Bacharel em Física pelo Instituto de Física da Universidade de São Paulo (USP) e Mestre em Ciência da Computação pelo Instituto e Matemática e Estatística da USP. Suas áreas de interesse incluem complexidade de linguagem e abordagens computacionais à variação linguística e tipologia. Neste livro, colaborou com os capítulos 6, O papel dos dados no pré-treinamento de Grandes Modelos de Linguagem.
Email: frserras@ime.usp.br

Fernanda Olival

Fernanda Olival é doutorada em História pela Universidade de Évora (Portugal), onde leciona desde 1991. É também diretora do CIDEHUS (Portugal). Trabalha sobre História Social e Política do período Moderno. Interessa-se por Processamento de Linguagem Natural (PLN) como ferramenta para anotar, gerir e recuperar informação, a partir de fontes históricas. Neste livro, colaborou com o Capítulo 15.
Email: mfo@uevora.pt

Flaviane Romani Fernandes Svartman

Flaviane Svartman possui graduação e doutorado em Linguística pela Unicamp, com período de estágio de doutorado na Faculdade de Letras da Universidade de Lisboa e pós-doutorado também pela Unicamp. Atua como docente na área de Filologia e Língua Portuguesa na Faculdade de Filosofia, Letras e Ciências Humanas da USP desde 2010. Suas investigações concernem ao estudo da fonologia e da fonética da língua portuguesa, com especial interesse na prosódia, na interface sintaxe-fonologia, na comparação entre variedades africanas, brasileiras e europeias do português e no PLN. Neste livro, colaborou com os Capítulos Recursos para o processamento de fala e 11.
Email: flaviane@gmail.com

Gabriel Lino Garcia

Gabriel Lino Garcia é graduado em Análise e Desenvolvimento de Sistemas pela FATEC e mestre em Inteligência Artificial (IA) e Processamento de Linguagem Natural (PLN) pela UNESP, onde faz doutorado focado na aplicação de modelos de linguagem na área médica. Seus interesses incluem aprendizado de máquina, grandes modelos de linguagem, aprendizado multimodal e aplicações médicas. Neste livro, colaborou com os Capítulos 5 e Treinamento de Grandes Modelos de Linguagem na prática.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Email: linaoglino@gmail.com

Helena Freire Cameron

Helena Freire Cameron é doutora em Linguística pela Universidade de Aveiro (Portugal) e pós-doutora em Humanidades Digitais e PLN pela Universidade de Évora (Portugal). Integrou vários projetos de investigação em Linguística Computacional (EUROTRA, DIGRAMA, Corpus Lexicográfico do Português, entre outros). É investigadora do CIDEHUS-Universidade de Évora (Portugal), onde integra um grupo multidisciplinar de Humanidades Digitais, envolvendo História, Linguística e Computação-PLN. Nos anos mais recentes tem-se dedicado ao estudo da variação ortográfica no século XVIII, do ponto de vista linguístico e computacional-PLN. Neste livro, colaborou com o Capítulo 15.
Email: helenac@ipportalegre.pt

Hidelberg O. Albuquerque

Hidelberg Albuquerque é formado em Análise e Desenvolvimento de Sistemas pelo Instituto Federal da Paraíba (IFPB), com pós-graduação pelo PPGI/UFPB, e atualmente doutorando em Ciência da Computação no Cin/UFPE. É docente na Unidade Acadêmica de Serra Talhada (UAST)/UFRPE desde 2015. Suas áreas de interesse incluem aprendizado de máquina, reconhecimento de entidade nomeada (NER), sumarização e recuperação de informação. Neste livro, colaborou com o Capítulo 13.
Email: hidelberg.albuquerque@ufrpe.br

Ivandré Paraboni

Ivandré Paraboni é doutor em Ciência da Computação pela University of Brighton (Inglaterra, 2003), com estágio de pós-doutorado junto à University of Aberdeen (Escócia, 2012). É professor associado (livre docente) e pesquisador da USP/EACH. Possui experiência na área de Ciência da Computação, atuando principalmente nas áreas de PLN e IA. Interesses de pesquisa abrangem o processamento da língua humana em um sentido amplo do termo, incluindo métodos computacionais derivados das Ciências Cognitivas e aplicações práticas de classificação de documentos extraídos da WEB. Neste livro, colaborou com os Capítulos Geração de linguagem natural e 10.
Email: ivandre@usp.br

Jackson Wilke da Cruz Souza

Jackson Souza é bacharel (2012), mestre (2015) e doutor (2019) em Linguística pela Universidade Federal de São Carlos (UFSCar). Atualmente, é professor da Universidade Federal da Bahia (UFBA), atuando junto ao Departamento de Ciência, Tecnologia e Inovação e ao Programa de Pós-Graduação em Língua e Cultura da UFBA. Tem interesse nas áreas de análise de discurso, sumarização automática e análise e descrição linguística. Neste livro, colaborou com os Capítulos Modelos discursivos e 5.
Email: jackcruzsouza@gmail.com

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

João Paulo Papa

João Paulo Papa é formado em Sistemas de Informação pela UNESP, com pós-graduação em processamento de imagens e aprendizado de máquina pela UFSCar e Unicamp, respectivamente. Possui pós-doutorados na Unicamp, Harvard e MIT. É professor titular da Unesp onde atua desde 2009 e suas áreas de interesse incluem visão computacional, processamento de imagens, aprendizado de máquina e aprendizado multimodal. Neste livro, colaborou com os Capítulos 5 e Treinamento de Grandes Modelos de Linguagem na prática.
Email: joao.papa@unesp.br

João Renato Ribeiro Manesco

João Renato Ribeiro Manesco é doutorando em Ciência da Computação pela UNESP, onde também concluiu mestrado e graduação. Durante o mestrado, realizou estágio de pesquisa no Media Integration and Communication Center da Universidade de Florença, Itália. Seus interesses de pesquisa incluem adaptação de domínio, modelos de linguagem generativos e multimodais, e aplicações em contexto médico. Neste livro, colaborou com os Capítulos 5 e Treinamento de Grandes Modelos de Linguagem na prática.
Email: joao.r.manesco@unesp.br

João Vitor Mariano Correia

João Vitor Mariano Correia é bacharel em Ciência da Computação pela UNESP, onde atualmente cursa o mestrado no Programa de Pós-Graduação em Ciência da Computação. Seus principais interesses de pesquisa estão na área de Processamento de Linguagem Natural (PLN), com foco em representações de conhecimento. Neste livro, colaborou com os Capítulos 5 e Treinamento de Grandes Modelos de Linguagem na prática.
Email: mariano.correia@unesp.br

Joaquim Santos

Joaquim Santos fez licenciatura em Matemática na Universidade Regional do Cariri (URCA), é mestre em Ciência da Computação pela PUCRS e atualmente é estudante de doutorado no programa de Computação Aplicada da UNISINOS. Tem trabalhado intensamente com Processamento de Linguagem Natural (PLN) em problemas de extração de informação, modelos de linguagem e bases de conhecimento. Neste livro, colaborou com o Capítulo 3.
Email: netojoaquim@edu.unisinos.br

Kenzo Sakiyama

Kenzo Sakiyama é formado em Engenharia de Computação pela Universidade Federal do Mato Grosso do Sul (UFMS), com mestrado em Ciência da Computação pelo ICMC/USP. Atualmente possui um doutorado em andamento também no ICMC/USP. Suas áreas de interesse incluem redes neurais profundas, mineração de texto em redes sociais e grafos. Neste livro, colaborou com o Capítulo 14.

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Email: kenzosakiyama@usp.br

Larissa Cristina Berti

Larissa Berti é formada em Fonoaudiologia pela UNESP de Marília, com doutorado em Linguística pela Unicamp. É docente do Departamento de Fonoaudiologia da UNESP desde 2009. Sua área de pesquisa é produção e percepção da fala, além do uso de métodos instrumentais na análise de fala. Neste livro, colaborou com o Capítulo 11.
Email: larissa.berti@unesp.br

Larissa A. de Freitas

Larissa Freitas é doutora em Ciência da Computação pela PUCRS (2015). Atualmente é professa da Universidade Federal de Pelotas (UFPel). Suas áreas de interesse incluem análise de sentimento, ontologias, *chatbots*, detecção de sarcasmo/ironia, detecção de discurso de ódio e detecção de notícias falsas. Neste livro, colaborou com o Capítulo 16.
Email: larissa@inf.ufpel.edu.br

Lilian Mie Mukai Cintho

Lilian Mie Mukai Cintho é formada em Enfermagem e Obstetrícia pela Universidade Federal de São Carlos (UFSCar), com pós-graduação em Tecnologia em Saúde na PUCPR. É docente colaboradora na Universidade Estadual de Ponta Grossa (UEPG). Suas áreas de interesse incluem, gerenciamento de programas de promoção à saúde, enfermagem médico cirúrgica e informática em saúde. Neste livro, colaborou com o Capítulo 9.
Email: miemukai@hotmail.com

Lívia Vicente Dutra

Lívia Vicente Dutra é formada em Letras pela Universidade Federal de Juiz de Fora (UFJF) e mestre em Tecnologia Linguística pela Universidade de Gotemburgo, onde atualmente é doutoranda no programa de Multilinguismo. É também pesquisadora do Laboratório FrameNet Brasil. Seus interesses de pesquisa incluem Semântica de Frames e Processamento de Linguagem Natural (PLN). Neste livro, colaborou com o Capítulo 17.
Email: livia.vdutralet@gmail.com

Lucas Daniel Volpiano Figueiredo

Lucas Daniel Volpiano Figueiredo é graduando em Sistemas de Informação pelo ICMC-USP. Foi estagiário de Inteligência Artificial no IFSC-USP por oito meses, atuando com implementação e pesquisa. Suas áreas de interesse incluem: PLN, aprendizado de máquina e inteligência artificial. Neste livro, colaborou com o Capítulo 18.
Email: volpi@usp.br

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Marcelo Finger

Marcelo Finger é professor titular de Ciência da Copmputação no IME-USP, Pesquisador Principal do Centro de Inteligência Artificial USP-IBM-FAPESP C4AI, com interesses em Geração de Recursos de Processamento de Linguagem Natural (PLN), uso de PLN para a área de Saúde, estudo e interpretabilidade de redes neurais. Neste livro, colaborou com os Capítulos 6, 11 e O papel dos dados no pré-treinamento de Grandes Modelos de Linguagem.
Email: mfinger@ime.usp.br

Marcelo Matheus Gauy

Marcelo Gauy é Bacharel em Matemática pela Universidade de São Paulo (USP), com mestrado na área de Combinatória pela USP e doutorado em Inteligência Artificial pelo ETH Zuerich, na Suíça. É docente na UNESP desde 2025. Suas áreas de interesse incluem aprendizado profundo, computação bioinspirada, processamento de áudio, combinatória e processos estocásticos. Neste livro, colaborou com o Capítulo 11.
Email: marcelo.gauy@unesp.br

Marcelo Gomes de Queiroz

Marcelo Queiroz é doutor em Ciência da Computação pelo IME-USP e docente do mesmo departamento desde 2002. Tem interesse em processamento de fala e voz e aplicações em medicina e linguística. Neste livro, colaborou com o Capítulo 11.
Email: mqz@ime.usp.br

Maria Cristina Ferreira de Oliveira

Cristina Oliveira é formada em Ciência da Computação pelo ICMC/USP (1985), com pós-graduação em Electronic Engineering pela University of Wales, Bangor, no Reino Unido (1990). É docente do ICMC, no Departamento de Ciências da Computação, desde 1986, como Professora Titular desde 2008. Atua em Ciência da Computação, em projetos de pesquisa que atualmente se concentram nos seguintes temas: visualização de informação, mineração visual de dados, análise visual de dados e aplicações. Neste livro, colaborou com o Capítulo 18.
Email: cristina@icmc.usp.br

Maria José Bocorny Finatto

Maria José Bocorny Finatto é linguista especializada em Estudos do Léxico e Terminologia. Investiga a história das terminologias médicas e modos de aplicar técnicas da Linguagem Simples para maior acessibilidade das informações escritas para diferentes perfis de pessoas no Brasil. Análise textual com recursos PLN é seu foco. Orientadora de mestrado e doutorado junto à Universidade Federal do Rio Grande do Sul (UFRGS). Neste livro, colaborou com os Capítulos Sequência de caracteres e palavras, 12 e 15.
Email: mariafinatto@gmail.com

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Mariane de Carvalho Pinto

Mariane Carvalho é graduada em Letras pela Universidade Federal de Juiz de Fora (UFJF), onde cursa o mestrado em Linguística Cognitiva e Computacional. Seu trabalho de pesquisa envolve a geração de roteiros de audiodescrição por Inteligência Artificial (IA), com base na Semântica de Frames e na anotação multimodal. É vinculada ao projeto FrameNet Brasil. Neste livro, colaborou com o Capítulo 17.
Email: mariane.carvalho@estudante.ufjf.br

Marlo Souza

Marlo Souza é formado em Ciência da Computação pela Unifacs e em Matemática pela Universidade Federal da Bahia (UFBA), com pós-graduação em PLN e Lógica pela PUC-RS e UFRGS, respectivamente. Atualmente, é professor adjunto na UFBA e co-líder do grupo de pesquisa FORMAS. Tem interesse em pesquisas nas áreas de semântica computacional, extração de informação e modelos de representação de conhecimento em PLN. Neste livro, colaborou com o Capítulo 3.
Email: msouza1@ufba.br

Matheus Cerqueira

Matheus Cerqueira é Técnico em Informática pelo Instituto Federal do Paraná (IFPR) de Jacarezinho-PR e graduando em Ciências de Computação pelo ICMC/USP desde 2020. Atualmente, também é estagiário em Engenharia de Software na Google. Suas áreas de interesse incluem Aprendizado de Máquina e, mais especificamente, Processamento de Linguagem Natural (PLN) e Visão Computacional. Neste livro, colaborou com o Capítulo 13.
Email: matheus.crqra@gmail.com

Maucha Andrade Gamonal

Maucha Gamonal atua como residente de pós-doutorado na Universidade Federal de Juiz de Fora (UFJF). É pesquisadora vinculada ao Projeto FrameNet Brasil desde 2009. Suas áreas de interesse incluem Linguística Cognitiva, Semântica de Frames e Linguística Computacional. É curiosa sobre a produção de sentido na linguagem humana e gosta de plantas, crianças, ar fresco e de seu golden retriever Luke. Neste livro, colaborou com o Capítulo 17.
Email: maucha.andrade@visitante.ufjf.br

Nádia Félix Felipe da Silva

Nádia Félix é formanda em Ciência da Computação pela Universidade Federal de Goiás (UFG), com pós-graduação em Inteligência Computacional no ICMC/USP. É docente na UFG desde 2016. Suas áreas de interesse incluem Aprendizado de Máquina, Processamento de Linguagem Natural (PLN) e Aprendizado Multimodal. Neste livro, colaborou com o Capítulo 13.
Email: nadia.felix@ufg.br

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Osvaldo Novais de Oliveira Jr.

Osvaldo Novais de Oliveira Jr. é físico, professor do Instituto de Física da Universidade de São Paulo (USP) de São Carlos, e um dos membros fundadores do Núcleo Interinstitucional de Linguística Computacional (NILC). Suas áreas de interesse são a ciência dos materiais, além de PLN. Neste livro, colaborou com o Capítulo 18.
Email: chu@ifsc.usp.br

Paula Christina Figueira Cardoso

Paula Figueira Cardoso tem graduação em Informática pelo ILES de Santarém, mestrado em Engenharia Elétrica pela Universidade Federal do Pará (UFPA) e doutorado pelo Instituto de Ciências Matemáticas e de Computação (ICMC) – USP/São Carlos. Atualmente é docente na UFPA. Suas áreas de interesse incluem geração de textos, sumarização automática, tecnologias educacionais e pesquisas aplicadas na área de interface homem-máquina. Neste livro, colaborou com os Capítulos Modelos discursivos e 5.
Email: paulastm@gmail.com

Pedro Henrique Paiola

Pedro Henrique Paiola é mestre e doutorando em Ciência da Computação pela UNESP, onde também atua como professor substituto desde 2022. É pesquisador no projeto RADIAR, desenvolvido em parceria com a Petrobras. Suas áreas de interesse incluem sumarização automática, modelos de linguagem e avaliação automática de texto. Neste livro, colaborou com os Capítulos 5 e Treinamento de Grandes Modelos de Linguagem na prática.
Email: pedro.paiola@unesp.br

Priscilla de Abreu Lopes

Priscilla de Abreu Lopes é formada em Ciência da Computação pela Universidade Federal de São Carlos (UFSCar), com pós-graduação em inteligência artificial no PPGCC/UFSCar. Suas áreas de interesse incluem Aprendizado de Máquina (AM) e Processamento de Linguagem Natural (PLN). Neste livro, colaborou com o Capítulo 7.
Email: alopes.priscilla@gmail.com

Rafael Oleques Nunes

Rafael Oleques Nunes é mestre e doutorando em Ciência da Computação no Instituto de Informática da UFRGS, bacharel em Ciência da Computação pela mesma instituição e técnico em Informática pelo IFSUL Campus Charqueadas. Pesquisador em PLN, concentra suas investigações em Reconhecimento de Entidades Nomeadas, qualidade de dados e uso de LLMs para criação e validação automática de dados em domínios especializados e de baixos recursos, com aplicações em Direito, Geologia e Saúde. Possui grande interesse nas interseções entre tecnologia e humanidades, explorando o potencial do PLN na análise de textos literários e históricos. Neste livro, colaborou com o Capítulo 13.
Email: rafael.nunes@inf.ufrgs.br

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Raphael Montanari

Raphael Montanari é formado em Ciências de Computação pela Universidade de São Paulo (USP), com pós-graduação em Inteligência Artificial (IA) e robótica no ICMC/USP. É servidor na USP desde 2013. Suas áreas de interesse incluem aprendizado de máquina, visão computacional e aplicações de Processamento de Linguagem Natural (PLN) para robótica. Neste livro, colaborou com o Capítulo 14.
Email: rmontanari@usp.br

Renata Vieira

Renata Vieira é PhD em Informática e Ciências Cognitivas pela Universidade de Edimburgo, Escócia. É pesquisadora da área de Processamento de Linguagem Natural (PLN) atuando em diversas áreas de aplicação. É investigadora principal convidada do Centro de História, Culturas e Sociedades da Universidade de Évora (Portugal) e conselheira científica do Instituto de Inteligência Artificial na Saúde. Neste livro, colaborou com os Capítulos Resolução de correferência, Perguntas e Respostas, 3 e 15.
Email: renata.vieira@gmail.com

Renato Moraes Silva

Renato Moraes Silva é graduado em Licenciatura Plena em Informática pela UFMT (2010) e Mestre (2013) e Doutor (2017) em Engenharia Elétrica pela Unicamp. Foi pesquisador de pós-doutorado pela UFSCar/Sorocaba (2017-2021) e professor e pesquisador no Centro Universitário Facens (2021-2024). Atualmente, é docente no ICMC/USP. Sua pesquisa é voltada para o PLN, com ênfase em aplicações como análise de sentimentos, detecção de spam, classificação de mensagens instantâneas e identificação de *fake news*. Além disso, desenvolveu pesquisas que aplicam aprendizado de máquina a problemas de visão computacional e estudos que utilizam análises de dados no contexto do setor financeiro. Neste livro, colaborou com o Capítulo 8.
Email: renatoms@icmc.usp.br

Rodrigo Nogueira

Rodrigo Nogueira é formado em Engenharia Elétrica pela UNICAMP e possui doutorado em Ciência da Computação pela Universidade de Nova York (NYU). Ao longo de sua carreira, Nogueira contribuiu para os campos de Recuperação de Informação e PLN através da criação de modelos como BERTimbau, doc2query, monoT5, e mais recentemente, os modelos Sabiá 1, 2 e 3, que são LLMs especializados no Brasil. Nogueira é o fundador e CEO da Maritaca AI, uma empresa especializada no desenvolvimento de LLMs especializados no Brasil. Neste livro, colaborou com o Capítulo 14.
Email: rfn@unicamp.br

Roney Lira de Sales Santos

Roney Santos é formado em Ciência da Computação pela Universidade Federal do Piauí (UFPI) em 2014, com mestrado em Processamento de Linguagem Natural (PLN) na UFPI

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

e doutorado em PLN no ICMC/USP. É docente no ICTI/UFBA desde 2023. Suas áreas de interesse incluem inteligência arttificial, aprendizado de máquina, PLN e detecção de conteúdo enganoso. Neste livro, colaborou com o Capítulo 8.
Email: roneysantos@ufba.br

Roseli A. F. Romero

Roseli A. F. Romero fez graduação e Mestrado no ICMC-USP e Doutorado na FEEC-UNICAMP. É professora titular do ICMC-USP desde 2013 e trabalha na área de inteligência artificial há 30 anos. Suas áreas de interesse incluem aprendizado de máquina, aprendizado profundo, com aplicação nas áreas de visão computacional, educação, meio ambiente, robótica e direito. Neste livro, colaborou com o Capítulo 14.
Email: rafrance@icmc.usp.br

Roseval Donisete Malaquias Junior

Roseval Donisete Malaquias Junior é bacharel em Ciência da Computação pela Universidade Federal de São Carlos (UFSCar) e mestrando em Ciência da Computação e Matemática Computacional no ICMC/USP. É pesquisador na Maritaca AI desde 2024. Suas áreas de interesse incluem aprendizado de máquina, grandes modelos de linguagem e pré-treinamento contínuo. Neste livro, colaborou com o Capítulo 14.
Email: roseval@usp.br

Sandra Maria Aluísio

Sandra Maria Aluísio é formada em Ciência da Computação pela UFSCar, com Doutorado em IA e PLN, pela USP. Foi docente de 1988 a 2018, no ICMC/USP, onde hoje atua no Programa Professor Sênior. As áreas de interesse incluem linguística de *corpus*, simplificação textual, recursos semânticos, detecção automática da estrutura do discurso científico e ferramentas de suporte à escrita, análise automática de distúrbios de linguagem, e processamento de fala. Neste livro, colaborou com os Capítulos Recursos para o processamento de fala e 6.
Email: sandra@icmc.usp.br

Sheila Castilho

Sheila Castilho é licenciada em Letras pela UNIOESTE. Possui mestrado em PLN pela Universidade do Algarve (PT) e pela University of Wolverhampton (UK). Completou seu doutorado pela Dublin City University em 2016. Atualmente, ela é Assistant Professor na Dublin City University. Seus interesses incluem tradução automática, avaliação automática e humana, tecnologias da tradução, e PLN para tradução. Neste livro, colaborou com o Capítulo 4.
Email: sheila.castilho@dcu.ie

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Sidney Evaldo Leal

Sidney Leal é pesquisador e cientista de dados no Venturus; pós-doc em Processamento de Fala (Tarsila/C4AI/USP); doutor em Ciências da Computação na área de Inteligência Artificial; MBA em Gestão da Tecnologia da Informação; bacharel em Ciência da Computação e técnico em Processamento de Dados. Atua na indústria desde 1995 em projetos de tecnologia de diversos portes para governos, bancos e multinacionais. Áreas de interesse: aprendizagem de máquina, processamento de linguagem natural, reconhecimento e síntese de fala, complexidade/simplificação de textos e internet das coisas. Neste livro, colaborou com o Capítulo 6.
Email: sidleal@gmail.com

Thiago Alexandre Salgueiro Pardo

Thiago Alexandre Salgueiro Pardo é formado em Ciências da Computação, com foco em pesquisa em PLN. Fez graduação e Mestrado na UFSCar e doutorado e pós-doutorado no ICMC/USP. É professor titular no ICMC, onde atua desde 2006. Seus interesses incluem análise de sentimentos, detecção de notícias falsas, sumarização de textos e modelos linguístico-computacionais de representação textual e métodos automáticos de análise. Neste livro, colaborou com o Capítulo 8.
Email: taspardo@icmc.usp.br

Tiago Agostinho de Almeida

Tiago Almeida é professor associado da UFSCar e especialista em ciência de dados, aprendizado de máquina e processamento de linguagem natural. Possui experiência em ensino e pesquisa em Inteligência Computacional, Aprendizado de Máquina e Reconhecimento de Padrões. É autor de um livro premiado de aprendizado de máquina e +100 artigos científicos. Neste livro, colaborou com o Capítulo 1.
Email: talmeida@ufscar.br

Tiago Timponi Torrent

Tiago Torrent é doutor em Linguística pela Universidade Federal do Rio de Janeiro (UFRJ), tendo feito mestrado e graduação em Letras na Universidade Federal de Juiz de Fora (UFJF). É professor do Programa de Pós-Graduação em Linguística e lidera o laboratório FrameNet Brasil da UFJF desde 2010. Seus interesses de pesquisa são modelos de língua aplicados à saúde digital e ao processamento computacional da multimodalidade. Neste livro, colaborou com o Capítulo 17.
Email: tiago.torrent@ufjf.br

Vitor Hugo Bridi Dei Santi

Vitor Hugo Bridi Dei Santi é formado em Física Computacional pela Universidade de São Paulo (USP). Atualmente, é mestrando no Programa de Pós-Graduação em Física Computacional do IFSC-USP. Suas áreas de interesse incluem redes complexas, simulações

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/


©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

computacionais e análise de literatura científica por técnicas de PLN e IA. Neste livro, colaborou com o Capítulo 18.
Email: vitor.santi@usp.br

Viviane Pereira Moreira

Viviane P. Moreira é doutora em computação na Middlesex University (2004), Inglaterra, mestre em ciência da computação na Universidade Federal do Rio Grande do Sil (UFRGS, 1999) e bacharel em análise de sistemas na UCPel (1996). É professora na UFRGS e orientadora do Programa de Pós-Graduação em Computação. Sua pesquisa concentra-se em recuperação de informação, Processamento de Linguagem Natural (PLN) e mineração de dados. Neste livro, colaborou com o Capítulo 2.
Email: viviane@inf.ufrgs.br

Yohan Bonescki Gumiel

Yohan Bonescki Gumiel é formado em Engenharia de Controle e Automação, com uma carreira dedicada à tecnologia e inovação. Ele fez doutorado na área de PLN voltada a textos clínicos, um campo de estudo que demonstra sua paixão por unir tecnologia e saúde. Atualmente, Yohan ocupa o cargo de cientista de dados na A3data e é pós-doutorando na UFMG, onde continua a aprofundar sua pesquisa e conhecimento em áreas críticas. Suas áreas de interesse incluem aprendizado de máquina, PLN e aplicações voltadas à saúde. Neste livro, colaborou com o Capítulo 9.
Email: yohan.gumiel@gmail.com

https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/

©2026 Direitos autorais deste livro a seus autores. Uso permitido sob licença Creative Commons Atribuição-NãoComercial-SemDerivações 3.0 Brasil (CC BY-NC-ND 3.0 BR).

Uma iniciativa



brasileiraspln.com

Apoio





