

PROCESSAMENTO DE LINGUAGEM NATURAL

Conceitos, Técnicas
e Aplicações em
Português

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

4ª Edição | 2025

Volume 1

*Este livro é o produto de um sonho coletivo do grupo
Brasileiras em Processamento de Linguagem Natural (BPLN)*

Processamento de Linguagem Natural

Conceitos, Técnicas e Aplicações em Português

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

4^a Edição – Volume 1
Novembro 2025



<https://braspln.com/livro-pln/4a-edicao/volume1/>

Conteúdo

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 Os Paradigmas de PLN	9
1.4 Vale a pena lembrar	10
 II Fala	 11
2 Texto ou fala?	12
2.1 Histórico e panorama da área	12
2.2 Aspectos teóricos fundamentais	14
2.2.1 A estrutura da língua falada	15
2.2.2 Reconhecimento de fala	24
2.2.3 Síntese de fala	35
2.2.4 Considerações finais	37
 3 Recursos para o processamento de fala	 39
3.1 Introdução	39
3.2 Recursos para síntese de fala	42
3.3 Recursos para segmentação prosódica	45
3.4 Recursos para reconhecimento automático de fala	48
3.5 Recursos para Reconhecimento de Emoções	50
3.6 Recursos para predição de pontuação no cenário de ASR	53
3.6.1 Ilustrando o uso do ASR Whisper na predição da pontuação em português	54
3.6.2 Descrição do Corpus MuPe e seu <i>dataset</i> de teste	56
3.6.3 <i>Corpora</i> usados nos trabalhos de Predição da Pontuação para ASRs	57
3.7 Considerações finais	60
Agradecimentos	60



III	Palavras	61
4	Sequência de caracteres e palavras	62
4.1	Conceitos básicos da morfologia	64
4.1.1	Morfema	64
4.1.2	<i>Token</i> e <i>Type</i>	67
4.1.3	Lexema, Lexia e Lema	67
4.1.4	Léxico e Gramática	68
4.1.5	Léxico comum e Léxico especializado	69
4.1.6	Palavras funcionais e palavras lexicais	70
4.1.7	Processos de formação das palavras	70
4.1.8	Morfologia e morfossintaxe	71
4.2	O processamento morfológico em PLN	71
4.2.1	Sentenciação	72
4.2.2	Tokenização	74
4.2.3	Tokenização em Subpalavras	75
4.2.4	Normalização	76
4.2.5	PoS <i>tagging</i>	77
4.2.6	Anotação de atributos morfológicos	79
4.3	Ferramentas e recursos para o processamento morfológico	80
4.3.1	Ferramentas para o português	80
4.3.2	Recursos para o português	84
4.3.3	PortiLexicon-UD: um recurso para processamento léxico em português	85
4.4	Considerações finais	91
5	Expressões multipalavras	94
5.1	Introdução	94
5.2	Hora de dar nome aos bois	95
5.2.1	As pedras fundamentais	95
5.2.2	Definições de MWE: uma pedra no nosso sapato	96
5.2.3	Separando o joio do trigo	98
5.2.4	MWEs para dar e vender	99
5.2.5	Um osso duro de roer	101
5.3	Tarefas: botando a mão na massa	102
5.4	Recursos: uma joia rara	104
5.4.1	Léxicos de MWEs	104
5.4.2	<i>Corpora</i> de MWEs	107
5.4.3	<i>Corpora</i> de MWEs em português	108
5.5	Avaliação: colocando os pingos nos 'i's	110
5.6	Até onde chegamos e para onde vamos	111
	Criação e manutenção de recursos	111
	Diálogo entre projetos e equipes	112
	Diversidade linguística	112
	Descrição aprimorada e ampliada de MWEs	113
	Processamento cognitivo de MWEs	113
5.7	Ao infinito e além	113
	Agradecimentos	114



IV	Estrutura	115
6	A ordem e a função das palavras em uma sentença	116
6.1	Introdução	116
6.2	Reflexões Iniciais	118
6.3	Noções Básicas de Sintaxe	119
6.4	Tipos de representação	121
6.4.1	Colchetes	121
6.4.2	Árvores	122
6.4.3	Setas	123
6.4.4	Parênteses	123
6.4.5	Indentação	124
6.5	Sintaxe de constituição	124
6.5.1	Possibilidades de interpretação e ambiguidades sintáticas	126
6.6	Sintaxe de dependência	128
6.6.1	Núcleo e dependente	129
6.6.2	A representação da sintaxe de dependência	130
6.6.3	Projetos de anotação multilingue: Universal Dependencies	133
6.7	Qual é melhor: constituição ou dependência?	137
6.8	Fronteiras da sintaxe	138
6.8.1	Sintaxe e Morfologia	138
6.8.2	Sintaxe e Semântica	139
6.8.3	Sintaxe e Pragmática	140
6.8.4	Sintaxe e Discurso	140
6.9	Considerações finais	141
7	Ferramentas e recursos para o processamento sintático	142
7.1	Introdução	142
7.2	Tipos de <i>parsing</i>	142
7.3	Recursos e ferramentas para o português	144
7.3.1	<i>Corpora</i>	145
7.3.2	<i>Parsers</i>	146
7.4	Visualização, anotação e edição de <i>treebanks</i>	156
7.4.1	Árvores de constituição	156
7.4.2	Árvores de dependência	157
7.4.3	Anotação de <i>corpus</i> em múltiplos níveis	158
7.5	Considerações finais	160
V	Significado	161
8	E o significado?	162
9	Semântica com técnicas simbólicas	169
9.1	Bases de Conhecimento Semântico	171
9.1.1	<i>Wordnets</i>	172
9.1.2	FrameNet	176
9.1.3	ConceptNet	181
9.2	Considerações finais	187



10 Semântica distribucional	189
10.1 Semântica Vetorial	190
10.1.1 Matriz termo-documento	191
10.1.2 Matriz termo-contexto	193
10.1.3 Calculando a similaridade entre vetores semânticos	194
10.2 Vetores esparsos	195
10.2.1 Atribuindo pesos aos termos da matriz termo-documento com TF-IDF	196
10.2.2 Atribuindo pesos aos termos da matriz termo-contexto com PMI . .	197
10.2.3 Reduzindo a dimensionalidade com LSA	199
10.3 Vetores densos estáticos	201
10.3.1 Word2Vec	202
10.3.2 Fasttext	206
10.3.3 Glove	207
10.4 Considerações finais	210
 VI Discurso	 211
11 Modelos discursivos	212
11.1 Introdução	212
11.2 Modelos de relações discursivas	214
11.2.1 O modelo RST (Mann e Thompson, 1988)	217
11.2.2 CST: Cross-document Structure Theory, 2000	219
11.3 Recursos e aplicações para o português brasileiro	222
11.4 Considerações finais	223
 12 Resolução de correferência	 225
12.1 Introdução	225
12.2 Resolução de Correferência	225
12.2.1 Referentes	226
12.2.2 Relações Semânticas Referenciais	228
12.2.3 Correferência, Anáfora e Catáfora	229
12.2.4 Referências Endofóricas e Exofóricas	231
12.2.5 Correferência, coerência e coesão	231
12.3 Abordagens Computacionais para Resolução de Correferência	231
12.3.1 Modelos Baseados em Regras	232
12.3.2 Modelos Baseados em Aprendizado de Máquina	232
12.3.3 Entity-Mention	235
12.3.4 Mention-Ranking	236
12.3.5 Antecedent-Tree	237
12.3.6 Modelos Voltados à Língua Portuguesa	238
12.4 Avaliação da Tarefa de Resolução de Correferência	239
12.4.1 Métricas de Avaliação	240
12.5 Aplicações	243
 VII Geração e Interação	 245
13 Geração de linguagem natural	246



13.1	Introdução	246
13.1.1	Geração de Linguagem Natural	248
13.1.2	Seleção de Conteúdo e Realização Textual	249
13.1.3	Avaliação de sistemas de GLN	249
13.1.4	Arquiteturas de GLN	250
13.2	Arquitetura Baseada em Templates	250
13.2.1	Aplicações	251
13.2.2	Exemplo – Sistema Rosie	251
13.2.3	Limitações	252
13.3	Arquitetura <i>pipeline</i>	253
13.3.1	Aplicações	255
13.3.2	Exemplo	255
13.3.3	Limitações	260
13.4	Arquitetura Ponta-a-Ponta	260
13.4.1	Aplicações	262
13.4.2	Exemplo – sistema CoronaReporter	263
13.4.3	Limitações	264
13.5	Arquiteturas Pré-treinadas	265
13.5.1	Aplicações	266
13.5.2	Exemplo	266
13.5.3	Limitações	267
13.6	Conclusão	268
14	Perguntas e Respostas	270
14.1	Introdução	270
14.2	Classificação de Sistemas de Perguntas e Respostas	272
14.2.1	Tipo de pergunta	272
14.2.2	Tipo de Fonte de Conhecimento	274
14.2.3	Domínio de conhecimento	278
14.3	Abordagens	279
14.3.1	Abordagem modular com documentos	280
14.3.2	Abordagem modular com grafo de conhecimento	284
14.3.3	Abordagem <i>end-to-end</i>	286
14.4	Métodos de avaliação	288
14.5	Outras tarefas relacionadas a PR	290
14.5.1	Comunidades de Perguntas e Respostas - <i>Community QA</i>	290
14.5.2	PR de múltiplas etapas <i>Multihopping QA</i>	291
14.5.3	PR com dados visuais	291
14.6	Sistemas de PR para o Português	291
14.7	Considerações finais	292
15	Diálogo e Interatividade	294
15.1	Introdução: Dialogando	294
15.2	Conceitos: O que forma um diálogo?	296
15.3	Fenômenos: O que ocorre em uma conversa?	299
15.3.1	Incrementalidade	299
15.3.2	Troca de Turnos e Estrutura Sequencial	301
15.3.3	Disfluências e Reparos	302



15.3.4	Construção de Base Comum e de Referências	305
15.3.5	Sinais Multimodais	306
15.3.6	Análises de diálogo em português	307
15.4	Modelos e Tarefas: Implementando um processo complexo	308
15.4.1	Modelos	308
15.4.2	Tarefas	311
15.5	Avaliação: Verificando a qualidade de um sistema	312
	Considerações éticas	315
15.6	Recursos: Trabalhando com dados e sistemas interativos	315
15.6.1	Atuando de Forma Responsável	316
15.6.2	<i>Corpora</i>	316
15.6.3	Ferramentas	320
15.6.4	Para Saber Mais	321
15.7	Conclusão: Mantendo o valor humano em foco	321
15.8	Exercícios	323
	Agradecimentos	325

VIII Dados 326

16 Conjunto de dados, *dataset* e *corpus* 327

16.1	Introdução	327
16.1.1	<i>Dataset</i> ou <i>corpus</i> anotado?	328
16.1.2	Anotação linguística	330
16.2	<i>Datasets</i> pra quê?	332
16.2.1	Sobre a importância dos dados	334
16.3	Características de um bom <i>dataset</i> linguístico	335
16.4	Por onde começar?	337
16.4.1	Definição do problema ou tarefa	337
16.4.2	Conjunto de etiquetas e instruções: o esquema de anotação	337
16.4.3	Escolha do <i>corpus</i>	339
16.4.4	Codificação	340
16.4.5	Anotação: sabedoria de especialistas ou sabedoria da multidão?	342
16.4.6	Ferramentas de anotação	345
16.4.7	Estratégias de anotação	345
16.4.8	Formas de avaliação	346
16.5	Procedimentos e estratégias de anotação e revisão	346
16.5.1	Palavras, regras e padrões linguísticos na construção de um <i>corpus</i> padrão ouro	346
16.5.2	Revisão de anotação	350
16.5.3	Até onde precisamos rever?	351
16.6	Como avaliar a qualidade do <i>dataset</i> ?	352
16.6.1	Concordância entre-anotadores	353
16.6.2	Quando a discordância não é erro: divergências legítimas	355
16.6.3	Avaliação intrínseca	356
16.6.4	Avaliação extrínseca	358
16.7	Em resumo...	360
	Agradecimentos	361



17	Aprendizado Transdutivo em PLN	362
17.1	Introdução	362
17.2	E o que seria esse tal de aprendizado transdutivo?	363
17.2.1	Transdução como forma de aprendizado	365
17.3	Aplicações do Aprendizado Transdutivo em PLN	367
17.3.1	Métodos baseados em propagação de rótulos	367
17.3.2	Adaptação de modelos supervisionados	370
17.3.3	Avaliação dos dados anotados	375
17.4	Benefícios e Limitações	375
IX	Avaliação	377
18	Avaliação de tecnologias de linguagem	378
18.1	Introdução: Avaliando	378
18.2	Contexto: Por onde começar?	380
18.2.1	Um pouco de história	380
18.2.2	Formalização de problemas e de modelos	382
18.2.3	Caracterização de tarefas de PLN	383
18.3	Paradigmas: Tipos de avaliação	385
	Avaliação Conjunta e <i>Leaderboards</i>	388
18.4	Procedimentos: Como avaliar?	390
18.4.1	Hipóteses e Experimentos	390
18.4.2	Referências	391
18.4.3	Mensurações	392
18.4.4	Partição de dados	393
18.4.5	Validação Cruzada	393
18.4.6	Comparação de Performance	394
18.4.7	Testes de Significância Estatística	395
18.4.8	<i>Baselines</i>	396
18.4.9	Ablação e Substituição	397
18.4.10	Análise de Dados	398
18.4.11	Análise de Erro	398
18.4.12	Avaliação Humana	399
18.5	Métricas: Medindo a performance	400
18.5.1	Usando métricas de forma responsável	401
18.5.2	Métricas comuns em PLN	402
18.6	Uso Responsável e Boas Práticas	403
18.6.1	Como reportar resultados?	403
18.6.2	Como tornar os resultados reprodutíveis?	405
18.6.3	Como testar e documentar um sistema?	407
18.6.4	Como organizar experimentos?	408
18.6.5	Atuando com ética	408
18.7	Para Concluir	409
18.8	Exercícios	410
	Agradecimentos	412
19	Avaliação conjunta em português	413



19.1	Apresentação	413
19.2	Avaliação conjunta: o que é e para que serve	414
19.2.1	Características e ingredientes de uma avaliação conjunta	416
19.3	Por que participar de uma avaliação conjunta	417
19.4	Como implementar uma avaliação conjunta	418
19.4.1	Cartografia do problema e da tarefa	418
19.4.2	Criação dos recursos	419
19.4.3	Medidas de avaliação	420
19.5	Como avaliar uma avaliação conjunta	420
19.5.1	Críticas e limitações	420
19.6	Morfolimpiadas: análise morfológica (2003-2004)	422
19.7	Primeiro e Segundo HAREM: avaliação de entidades, relações e expressões temporais (2005 e 2007-2008)	423
19.7.1	Reconhecimento de entidades mencionadas	423
19.7.2	Identificação, classificação e normalização de expressões temporais	424
19.7.3	ReReLEM - relações entre entidades	425
19.8	CLEF: Recuperação de Informação e QA (2003-2009)	426
19.8.1	CLEF clássico	426
19.8.2	Resposta automática a perguntas (QA@CLEF)	427
19.8.3	GeoCLEF	427
19.8.4	GikiP: Informação geográfica (2008)	428
19.8.5	GikiCLEF	428
19.9	Págico: RI complexa sobre culturas de língua portuguesa (2012)	428
19.10	ASSIN e ASSIN 2: similaridade semântica e inferência (2016 e 2019)	430
19.11	IberLEF 2019: reconhecimento de entidades e de relações	431
19.12	IDPT: Identificação de ironia (2021)	432
19.13	ABSAPT: Avaliação de mineração de opiniões por aspecto (2022)	432
19.14	DIP: Desafio de Identificação de Personagens (2022-2023)	433
19.15	Resumo das avaliações conjuntas	433
19.16	Considerações finais e lições aprendidas	434
	Agradecimentos	435
X	Desafios e Perspectivas	436
20	Questões éticas em IA e PLN	437
20.1	Ética em IA	437
20.2	Ética em PLN	443
20.3	Modelos de língua como fonte de conhecimento?	444
21	E agora, PLN?	446
21.1	Desafios e perspectivas para o PLN-Português	446
21.2	Há limites para o PLN?	449
	Referências	451



Apêndices	533
A Alguns pressupostos para o processamento de fala	533
A.1 Alguns pressupostos: estatística, probabilidade, teoria da informação	533
A.1.1 Probabilidade	534
A.1.2 Variáveis aleatórias	534
A.1.3 Média e Variância	535
A.1.4 Covariância e Correlação	535
A.1.5 Vetores aleatórios e distribuições multivariadas	536
A.1.6 Algumas distribuições úteis	536
A.1.7 Teoria da Estimção e Teste de Significância	537
A.1.8 Teoria da Informação, Entropia e Informação Mútua	538
B Diretrizes de avaliação	540
B.1 Diretrizes de avaliação para um modelo de cinco estágios	540
C EAGLES	544
C.1 EAGLES	544
Sobre as/os autoras/es	546
Helena de Medeiros Caseli	546
Maria das Graças Volpe Nunes	546
Adriana Silvina Pagano	546
Aline Aver Vanin	546
Aline Villavicencio	547
Amanda Pontes Rassi	547
Ana Clara Souza Pagano	547
Brielen Madureira	547
Camila de Araújo Azevedo	548
Carlos Ramisch	548
Cláudia Freitas	548
Daniela Barreiro Claro	548
Dante Augusto Barone	548
Diana Santos	549
Eduardo Cortes	549
Elisa Terumi Rubel Schneider	549
Eloize Rossi Marques Seno	549
Evandro Fonseca	550
Heliana Mello	550
Ivandr� Paraboni	550
Jackson Wilke da Cruz Souza	550
Jessica Rodrigues	551
Laila Mota	551
Lucelene Lopes	551
Maria Jos� Bocorny Finatto	551
Mariza Ferro	551
Nataly Leopoldina Patti da Silva	552
Norton Trevisan Roman	552
Paula Christina Figueira Cardoso	552



Conteúdo

Priscila Osório Côrtes	552
Renata Vieira	553
Renata Ramisch	553
Roana Rodrigues	553
Tayane Soares	553
Thiago Castro Ferreira	553
Valéria de Paiva	554
Vlândia Pinheiro	554



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 O que é PLN?, 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 (Texto ou fala?), e técnicas, recursos e aplicações (Recursos para o processamento de fala). 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 Sequência de caracteres e palavras 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 ao leitor interessa o processamento de expressões idiomáticas, o Capítulo Expressões multipalavras 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 A ordem e a função das palavras em uma sentença fornece toda a conceitualização de sintaxe, os principais tipos de análise, suas diferenças, vantagens e desvantagens. O Capítulo Ferramentas e recursos para o processamento sintático 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, seja 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 E o significado? introduz toda a complexidade da semântica da língua. Os Capítulos Semântica com técnicas simbólicas e Semântica distribucional 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 livro. No Capítulo Modelos discursivos, o leitor encontra os principais modelos discursivos para PLN; no Capítulo Resolução de correferência, 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 Geração de linguagem natural, e cobre, nos Capítulos Perguntas e Respostas e Diálogo e Interatividade, 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 Conjunto de dados, *dataset* e *corpus* aborda tudo o que está envolvido na construção e na anotação de *datasets* ou *corpus*, bem como seu papel no treinamento de algoritmos de aprendizado de máquina. E o Capítulo Aprendizado Transdutivo em PLN 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 Avaliação de tecnologias de linguagem que apresenta um panorama dos métodos de avaliação comumente usados para medir, analisar e comparar o desempenho de sistemas de PLN, e o Capítulo Avaliação conjunta em português, 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.

Antes de finalizar o volume, a **Parte 10 (Desafios e Perspectivas)** discute algumas questões éticas (Questões éticas em IA e PLN) que a IA, em geral, e o PLN, em particular, têm provocado, pela forma como novas tecnologias têm sido criadas e usadas recentemente. Finalmente, o último capítulo (E agora, PLN?) 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, 2025. v. 1. Disponível em: <https://brasileiraspln.com/livro-pln/4a-edicao/volume1>.

Note 1: Bibref



```
@Book{BPLN_livro_4ed_vol1:2025,  
  title      = {Processamento de Linguagem Natural: Conceitos,  
                Técnicas e Aplicações em Português},  
  editor     = {Caseli, H. M. and Nunes, M. G. V.},  
  year       = {2025},  
  publisher  = {BPLN},  
  edition    = {4},  
  volume     = {1},  
  isbn       = {XXX-XX-XX-XXXXX-X},  
  url        = {https://brasileiraspln.com/livro-pln/4a-edicao/volume1/}  
}
```



Parte I

Introdução



Capítulo 1

O que é PLN?

Helena de Medeiros Caseli
Maria das Graças Volpe Nunes
Adriana Pagano

Publicado em: 26/09/2023

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 (p. ex. palavras e suas classes morfológicas e gramaticais, seus traços semânticos ou ontológicos etc.). Já interpretação se refere à tentativa de apreender significados

¹Embora 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.



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.

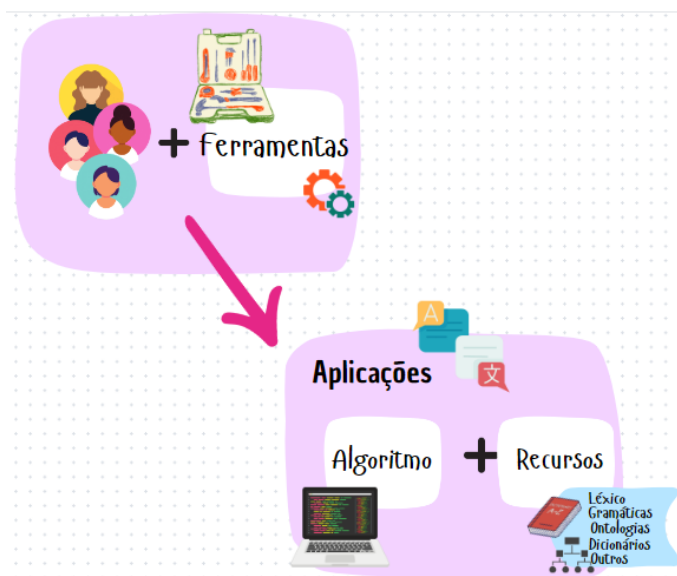
Em NLG, por outro lado, o objetivo é a geração de linguagem natural. Um exemplo de NLG é a geração de respostas ao usuário dos *chatbots*. Para o sistema, isso significa decidir o que responder e como apresentar essa resposta ao usuário. Atualmente, o ChatGPT² é o exemplo de maior sucesso: é capaz de 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

²<https://chat.openai.com/>

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 a criação de recursos em línguas de comunidades tecnologicamente menos desenvolvidas; um sumário 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 (<i>input</i>) em linguagem natural e a transformam produzindo um determinado resultado	- tradutor automático - corretor ortográfico ou gramatical - assistentes virtuais/<i>chatbots</i> - sumário 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 <i>fake news</i>
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 - <i>corpus</i> (<i>datasets</i> 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 <i>tokens</i> (tokenizador), sentenças, parágrafos - <i>stemmers</i> (extratores de raiz de uma palavra (e.g. “corr” de “correr”)) - lematizadores (e.g. “correr” de “corri”) - etiquetadores morfosintáticos (PoS <i>taggers</i>) para classe de palavras (verbo, substantivo, adjetivo, artigo, preposição, advérbio etc.) - etiquetadores semânticos, analisadores sintáticos parciais (<i>chunkers</i>) e completos (<i>parsers</i>) - concordanciadores - interfaces de anotação de <i>corpus</i> - componentes em kits de ferramentas (<i>toolkits</i>), como o NLTK³

Neste livro iremos aumentar gradativamente a complexidade do tratamento da língua no PLN, com foco no português brasileiro. Antes de iniciar esta trajetória, a A língua

³<https://www.nltk.org>



apresenta nosso objeto de pesquisa, a língua. Em seguida, a Os Paradigmas de PLN introduz os principais paradigmas do PLN, que serão retomados em diversos momentos neste livro. Por fim, a Vale a pena lembrar destaca os principais pontos apresentados no capítulo.

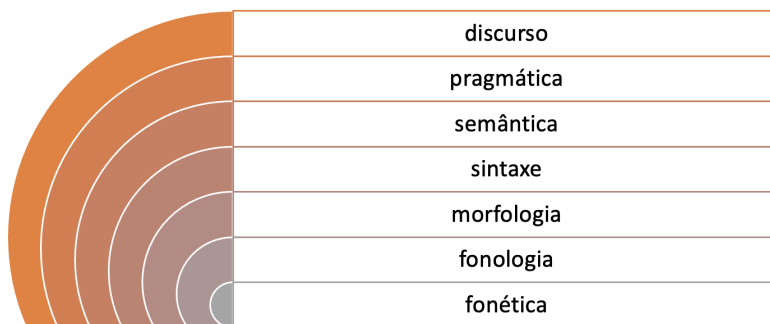
1.2 A língua

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 escrita⁴, 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

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

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. 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 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. A 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. 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 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.

1.4 Vale a pena lembrar

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 “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.

Parte II

Fala



Capítulo 2

Texto ou fala?

Camila de Araújo Azevedo
Heliana Ribeiro de Mello
Priscila Osório Côrtes

Publicado em: 26/09/2023

 <https://brasileiraspln.com/livro-pln/4a-edicao/volume1/>

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 fala¹ 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 morfemas². 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

¹Ressaltamos 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.

²Cf. 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>.



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 (Síntese de fala). 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 (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 fala³ 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 Recursos para o processamento de fala.

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**

³Para saber mais sobre *datasets* e *corpora*, sugere-se a leitura do Conjunto de dados, *dataset* e *corpus*.

⁴<https://github.com/nile-nlp/CORAA> e <https://sites.google.com/view/tarsila-c4ai/coraa-versions?pli=1>



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



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

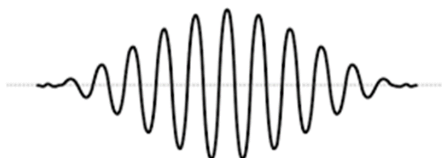
2.2.1 A estrutura da língua falada

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 Fala⁵

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.

⁵Recomendamos o PRAAT como software para análise da fala: <https://www.fon.hum.uva.nl/praat/manual/Intro.html>

⁶Toda a seção sobre produção da fala é orientada pelo livro de Huang et al. (2001).

⁷Cf. O site Fonética e Fonologia: <https://fonologia.org/>

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 /a λ u/;
- **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 Articuladores, 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(f_0). 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.



Uma vez que a onda glotal é periódica, consistindo na frequência fundamental (f_0) e em um número de harmônicos (múltiplos integrais de f_0), 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 cm³. 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.



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.



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).



- **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 [pe] 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.



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 (Articuladores).

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”/ 'beɫ/ e “bola”/ 'boɫ/ 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) (Articuladores). 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



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 fonemas⁸ – 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 “cê-sasessonspasansavas”)

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

⁸Consultar a tabela da Associação Internacional de Fonética disponível em <https://www.internationalphoneticassociation.org/content/full-ipa-chart> e no Alguns pressupostos para o processamento de fala.



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” (Taxa de articulação e Coarticulação) – 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.



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 compreensão⁹ 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.

⁹A 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.

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 também¹⁰ 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.

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



Portanto, é preciso mostrar à rede neural uma parcela significativa de áudios com esses tipos de ruído¹¹.

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 horas¹⁴. 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 Etapas adicionais.

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 [têbêj]¹⁶ corresponderia a “tambem”, e, por conseguinte, a saída do modelo conteria o *typo*

¹¹É 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>

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

¹⁴O 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.

¹⁵Na 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.



“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 desenvolvida¹⁶.

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 treinamento¹⁷. As técnicas de normalização são as mesmas utilizadas em processamento de texto (Sequência de caracteres e palavras), 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 distorcidos¹⁸ 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 aprendizagem. 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 A estrutura da língua falada, 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

¹⁶Vale 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.

¹⁷Há 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.

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

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



pelas cordas 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 assumem²⁰. 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 estacionário²². Por isso, para representar um sinal com duração de vários segundos ou até minutos, utiliza-se o método de janelamento²³. 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 anterior²⁴. 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

²⁰Em 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.

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

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

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

²⁴Segundo 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).



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 ruído²⁶, 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 Janelamento, 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-gramas²⁷ e assumiam independência entre as probabilidades de ocorrência dos fones, e, por isso, não eram eficazes em processar informações de longa distância²⁸. Hoje, as arquiteturas do tipo *encoder-decoder* são as mais utilizadas em ASR.

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

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

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

²⁸Outra 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



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 <u>sessão</u>

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 falada²⁹. 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.

²⁹Alcain 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.

³⁰RNN 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.

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 (Sequência de caracteres e palavras). Mas não devem ser confundidos com morfemas! Diferentemente dos morfemas, as *wordpieces* não carregam nenhum significado necessariamente³¹. 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 (Semântica distribucional) do

³¹Apenas 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/>.



processamento de texto, a ideia do wav2vec é obter representações vetoriais diretamente a partir do áudio 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 prever 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

³³Isso 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.

quinhentos mil reais”, que pode ser transformado em “R\$ 2.500.000” ou “R\$ 2500000”, a depender da convenção adotada.

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 (Sequência de caracteres e palavras). Por exemplo, uma sentença como “Mande um beijo para a Juliana” reconhecida por um ASR como “Mande um beijo pra Juliana” seria muito menos 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 **para 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”.

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 Métricas de avaliação, 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 (Sequência de caracteres e palavras) 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.



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
km ²	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”



Links/URLs	“www.google.com”
Porcentagem	“12,8%” “21,3%”
Unidades de medida	“3cm” “8,515 mi Km ² ” “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



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 Coleta de dados, 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 Recursos para o processamento de fala 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.



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: 29/09/2025

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 1.4². 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 Corpus³.

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 Librivox⁴, 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 Corpus⁵ e

¹<https://commonvoice.mozilla.org/pt/>

²O *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>



o LibriTTS⁶, 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 TaRSila⁷ do Center for Artificial Intelligence⁸ 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 predição 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, conseqüentemente, 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/>



Fala⁹ (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 Whisper¹⁰ da OpenAI, por exemplo, foi treinado dessa forma. O uso de pares áudio-transcrição para treinamento de reconhedores 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-U¹¹. 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>



Começamos apresentando, na Recursos para síntese de fala, 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 Recursos para segmentação prosódica, apresentamos o *corpus* CORAA NURC-SP, 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 Recursos para reconhecimento automático de fala, apresentamos o *corpus* CORAA ASR versão 1.1, composto por quatro *corpora* disponíveis na literatura, que foram validados para a tarefa de ASR, e uma coleção de TeD Talks, totalizando aproximadamente 290 horas. Na 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 Recursos para predição de pontuação no cenário de ASR, apresentamos o *corpus* do Museu da Pessoa (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. O futuro *corpus*, após ser anotado e anonimizado para as várias tarefas em estudo, será denominado CORAA MuPe. Na Recursos para predição de pontuação no cenário de ASR apresentamos o *dataset* de teste do CORAA MuPe, com aproximadamente 17 horas, e que foi usado para a avaliação da tarefa de predição de pontuação do ASR Whisper da OpenAI¹². Finalizamos o capítulo com a apresentação dos recursos futuros que serão criados ou expandidos a partir dos já descritos neste capítulo (Considerações finais).

3.2 Recursos para síntese de fala

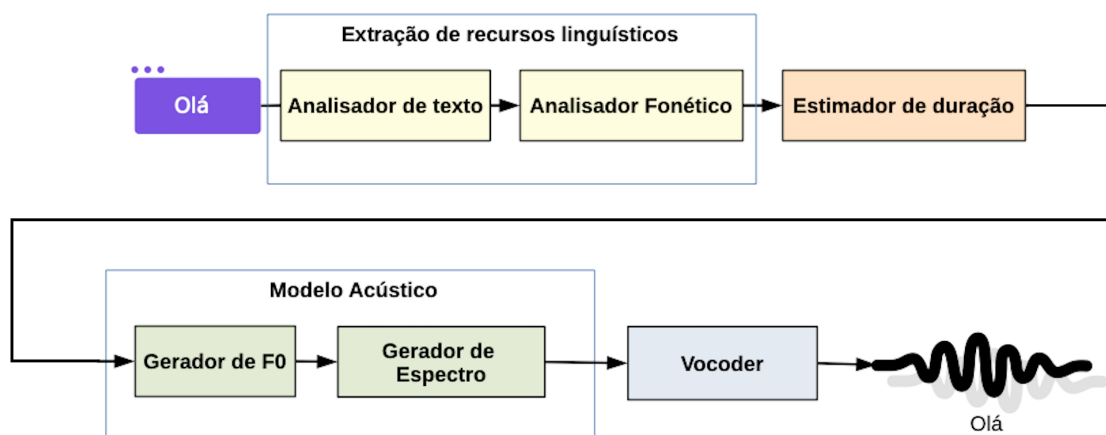
Os sistemas de síntese de fala receberam muita atenção a partir de 2017 devido ao grande avanço proporcionado pela aplicação do *deep learning* (Goodfellow et al., 2016) nessa área, que permitiu a popularização e aprimoramento de assistentes virtuais, como Apple Siri (Gruber, 2009), Amazon Alexa (Purinton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

¹²<https://openai.com/research/whisper>



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)

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 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 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-corpus¹³, 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

¹³<https://github.com/gunthercox/chatbot-corpus/>



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 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.

<i>Corpora</i>	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 vem sendo preparado tanto para viabilizar estudos linguísticos quanto o processamento computacional, contará com ≈ 334 horas de fala transcrita, das



quais pelo menos 40 horas serão prosodicamente anotadas.

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 (Whisper¹⁵), um alinhador forçado áudio-transcrição baseado em síntese de fala (aeneas¹⁶) 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

¹⁴<https://nurc.ffmpeg.usp.br/>

¹⁵<https://github.com/openai/whisper/>

¹⁶<https://www.readbeyond.it/aeneas/>



Mínimo (Gris et al., 2022), o modelo Whisper foi escolhido, pois, além de realizar 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 Clarin¹⁷ 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:

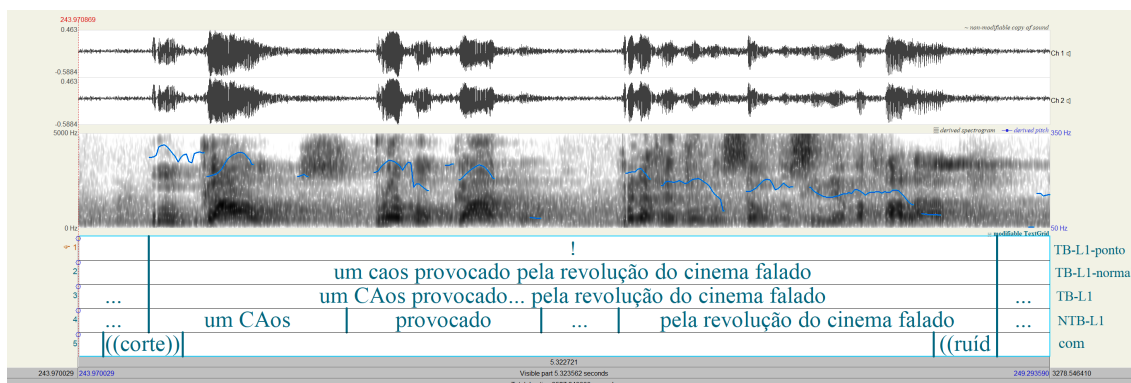
- 2 camadas (TB-, NTB-) nas quais a fala de cada locutor (-L1, -L2) e documentador (-DOC1, -DOC2) é segmentada em unidades prosódicas e transcrita de acordo com normas adaptadas do Projeto NURC.
- 1 camada (LA) para a fala transcrita e segmentada de qualquer locutor aleatório.
- 1 camada para comentários (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 (Conjunto de dados, *dataset* e *corpus*) 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

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



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



seguida, a ortografia foi verificada e o texto foi normalizado, a fim de tornar o conjunto de dados adequado para o processamento 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 portuguesa¹⁸. 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.

<i>Corpora</i>	Horas	Áudios	Falantes	Licença	Lançamento
CORAA ASR 1.1	289	402.000	1.700	CC BY-NC-ND	2022
Mozilla Common Voice 13.0	197	130.000	3.100	CC-0	2023
MultiLingual LibriSpeech	285	40.000	68	CC BY	2020
MultiLingual TeDx Corpus	164	93.000	-	CC BY-NC-ND	2021
Spotify Podcast Dataset	7.600	123.000	-	Proprietária	2022

CORAA ASR¹⁹ (Candido Junior et al., 2022) é um *corpus* para reconhecimento automático de fala que contém também fala espontânea, um tópico pouco pesquisado em

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

¹⁹<https://github.com/nile-nlp/CORAA>

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 Electronic 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* Common Voice²⁰ (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.

O *corpus* MultiLingual LibriSpeech²¹ (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 Recursos para síntese de fala. 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 Recursos para síntese de fala 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 Corpus²² (Salesky et al., 2021) foi proposto para permitir pesquisas nas áreas de reconhecimento automático da fala e tradução da fala para texto²³. 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.

²⁰<https://commonvoice.mozilla.org/pt>

²¹<http://www.openslr.org/94/>

²²<http://www.openslr.org/100>

²³Na tradução de fala para texto, o idioma entre o áudio original e a transcrição são diferentes.



O *corpus* Spotify²⁴ (Clifton et al., 2020) foi lançado primeiramente para a língua inglesa. Em 2022, a empresa lançou uma nova versão incorporando o português (Tanaka et al., 2022), oferecendo diversos áudios para a língua portuguesa provenientes principalmente de podcasts disponíveis na plataforma. Ao todo, 76 mil horas de áudios foram disponibilizadas a partir de 123 mil episódios de shows da plataforma. As transcrições foram geradas automaticamente e estão sujeitas a erros de transcrição. Apesar da licença livre para uso acadêmico, até o momento de escrita deste texto, o *corpus* ainda não estava totalmente disponível para uso. Pesquisadores interessados em acessar os áudios devem entrar com um pedido de acesso na página web dos organizadores.

Existem outras bases para tarefa de ASR que também valem a pena ser citadas. Entre elas, o Multilingual Spoken Corpus²⁵ (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 pesquisador²⁶.

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 Populi²⁷ (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.

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

²⁴<https://podcastsdataset.byspotify.com/>

²⁵<https://mlcommons.org/en/multilingual-spoken-words/>

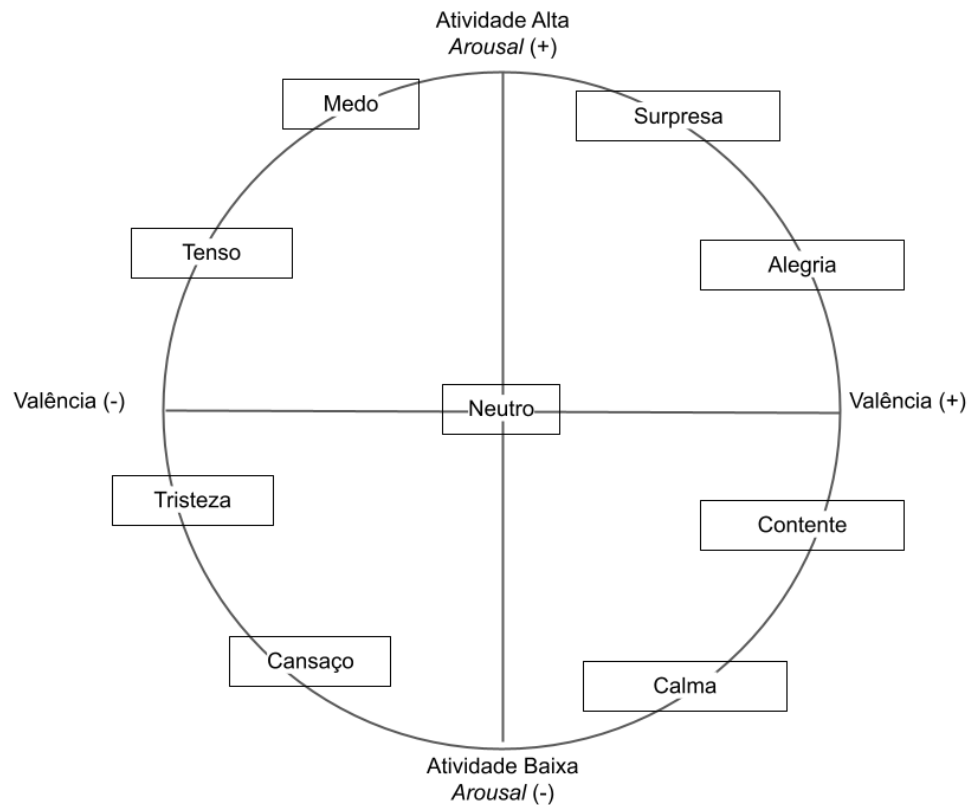
²⁶<https://igormq.github.io/datasets/>

²⁷<https://github.com/facebookresearch/voxpathuli>



virtuais mais empáticos, o que ajuda a desenvolver técnicas mais eficientes para interação humano-computador (Wani et al., 2021).

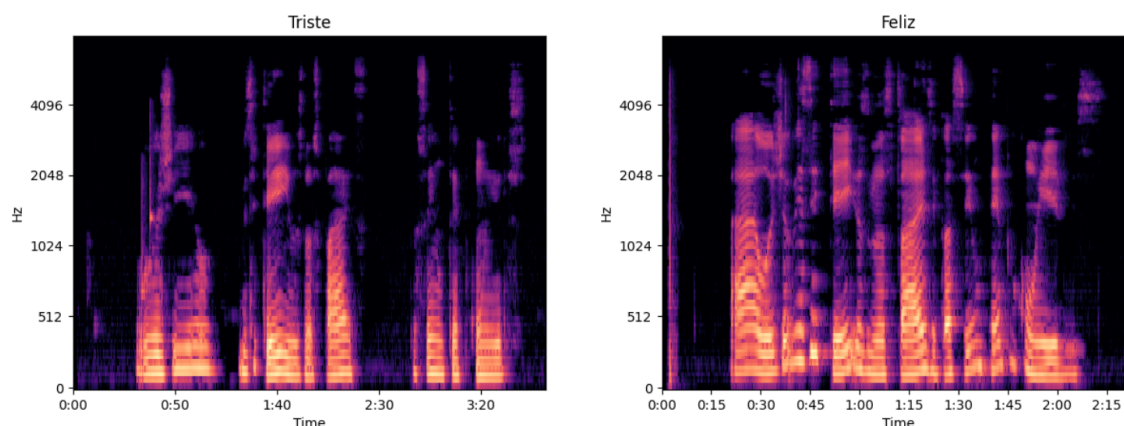
Figura 3.4: Modelo Circumplexo de Russell (simplificado) para representar as emoções em um espaço bidimensional, com eixos de valência (positivo/negativo) e intensidade (ativa/passiva).



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 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.

O segundo desafio está relacionado à disponibilidade de *corpora* anotados para a tarefa

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.



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 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 publicamente²⁸. 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

²⁸CORAA-SER v1: <https://github.com/rmarcacini/ser-coraa-pt-br>

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 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 MuPe²⁹ 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 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 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

²⁹https://museudapessoa.org/historia-detalle/?id=7136&download_integra_text_pdf/



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?
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 Corpus CORAA MuPe está atualmente em fase de processamento. Ele 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 inicial do convênio é o estudo e desenvolvimento de modelos de ASR, de métodos de segmentação automática de transcrição e modelagem de tópicos baseada nas transcrições de vídeos.

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 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 publicamente³⁰,

³⁰<https://github.com/nile-nlp/asr-punctuation-evaluation>



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 $\pm$ 169.44	138.80 $\pm$ 325.46	105.11 $\pm$ 240.72
# Sentenças	4.100	4.640	8.740
Tamanho Médio da Sentença	17.16 $\pm$ 27.47	14.57 $\pm$ 14.64	15.79 $\pm$ 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

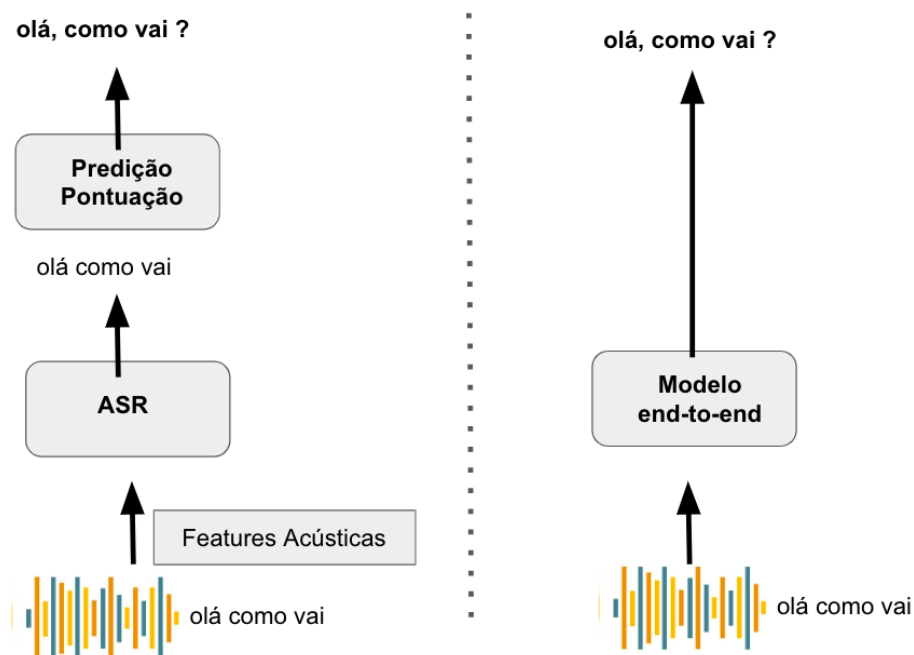
3.6.3 Corpora usados nos trabalhos de Predição da Pontuação para ASRs

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).

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

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)

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 wav2vec³¹.

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 MuPe³². 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).

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

³¹<https://github.com/pytorch/fairseq/blob/master/examples/wav2vec>

³²<https://github.com/nile-nlp/asr-punctuation-evaluation>

Trabalhos	Língua(s)	Test Set	Duração/Tam.
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 Face³³. 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* multilíngue³⁴ (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 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).

³³<https://huggingface.co/docs/transformers/index>

³⁴MuST-C inclui gravações de áudio de TED Talks em inglês alinhadas no nível da frase com suas transcrições e traduções manuais.



3.7 Considerações finais

Apresentamos neste capítulo os recursos de processamento de fala que foram criados nos três anos iniciais do projeto TaRSila. Dois grandes *corpora* estão ainda em fase de processamento para serem lançados em 2024: (i) o Corpus CORAA NURC-SP e o Corpus CORAA MuPe. Além dos estudos previstos no convênio com o MuPe, uma tarefa futura será a compilação de um *dataset* para modelagem e teste de sistemas TTS, nos moldes do LibriTTS, citado na Recursos para síntese de fala, já que as histórias de vida cedidas foram gravadas em estúdio. A diferença entre o LibriTTS e o *corpus* a ser criado a partir do Corpus MuPe está no gênero dos textos: as histórias de vida do MuPe são exemplos da fala espontânea, guiada por entrevista, e o LibriTTS contém fala lida.

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 na 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.



Parte III

Palavras



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

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 (Texto ou fala?), por exemplo, considera o fonema como a menor unidade sonora e distintiva de uma língua. Se tomarmos o exemplo do que diferencia as palavras “sábua” (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.



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 pontuação¹. 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 (Extração de Informação) e tradução (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 hífen como em “guarda-chuva”.

No mesmo sentido, dependendo dos propósitos da tarefa, um modelo de aprendizado de máquina pode precisar definir expressões multipalavras (Expressões multipalavras) (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”, “ciência de da-

¹Considerando 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.



dos” e “euamopuzzles”. 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 “sábã”, “sabã” e “sabiã”, lembrando que estamos no território dos sons e não da escrita. Para saber mais sobre o processamento da fala, sugere-se a leitura de Texto ou fala? e Recursos para o processamento de fala.

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 “experimental”. 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;
- (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”, “**im**ortal”, “**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: “ativamente”, “imaginação”, “crescimento”).

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: “atleta”, “colega”, “dentista”), “e” (ex: “agente”, “recorrente”, “alegre”) ou “o” (ex: “pássaro”, “crocodilo”, “dezembro”); e
2. **verbais**, que indicam as 3 conjugações verbais: “a”, para verbos da primeira conjugação (ex: “andar”, “passear”, “falar”), “e” ou “o”, para verbos da segunda conjugação (ex: “escrever”, “ler”, “fazer”, “pôr”, “repor”, “compor”) ou “i”, para verbos da terceira conjugação (ex: “agir”, “assumir”, “partir”).

As vogais temáticas não devem ser confundidas com as desinências nominais, por exemplo, que indicam gênero em “menino” e “menina”. 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.



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] ; morfológicamente, 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 **vocabulo**. 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 Sarmiento (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

²Para 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.



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:



- 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 (Tradução Automática), os léxicos são bilíngues (ou multilíngues) especificando não apenas as palavras que compõem os léxicos dos vários idiomas,

³https://www.planalto.gov.br/ccivil_03/leis/l8069.htm

⁴Fonte: <https://www.biologianet.com/biologia-celular/dna.htm>



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”.



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 Sequência de caracteres e palavras. 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



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 (sentencição), 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 morfosintática 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 Sentencição

A sentencição (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.



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 NLTK⁵ (Bird; Loper, 2004).
- **Abordagens baseadas em aprendizado de máquina supervisionado**, ou seja, modelos computacionais treinados sobre conjuntos anotados (*gold standard*, veja Conjunto de dados, *dataset* e *corpus*) 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 sentencição 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.

⁵<https://www.nltk.org>



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 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 português⁶ 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 Sequência de caracteres e palavras.

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

⁶<https://portulanclarin.net/workbench/lx-tokenizer/>



do processo de tokenização é, no entanto, menor que a do processo de sentencição, 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.

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 texto⁷.

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 Semântica distribucional. 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

⁷Geralmente, 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.

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 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 youtuberis 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

⁸Vale 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.



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.

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 sofisticções 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 (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 realiza a etiquetagem morfossintática) pode usar apenas as etiquetas de granularidade grossa (em inglês, *coarse tags* ou UPoS), que correspondem às 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* LXUTagger⁹.

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

DET NOUN ADJ VERB ADV PUNCT
<p> <s> O gato preto correu rapidamente . </s> </p>

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, resumizando 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 (Conjunto de dados, *dataset* e *corpus*), 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;

⁹<https://portulanclarin.net/workbench/lx-utagger/>

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* $T=\{t_1, t_2, t_3... t_n\}$ para cada palavra em uma sentença $W=\{w_1, w_2, w_3...w_n\}$ 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*, 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

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/>



Figura 4.3: Exemplo de anotação de *features* morfológicas

	Gender=Masc Number=Sing	Gender=Masc Number=Sing	Gender=Masc Number=Sing	Mood=Ind Number=Sing Person=3 Tense=Past		
	O	GATO	PRETO	CORRER	RAPIDAMENTE	.
	DET	NOUN	ADJ	VERB	ADV	PUNCT
<p>	O	gato	preto	correu	rapidamente	.
<s>						</s> </p>

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 (Conjunto de dados, *dataset* e *corpus*). 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 NLTK¹¹ e spaCy¹², são módulos mais completos que implementam submódulos e funcionalidades

¹¹<https://www.nltk.org>

¹²<https://spacy.io/>



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 (Expressões multipalavras), 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 Snowball¹⁷ 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.

¹³<https://www.nltk.org/api/nltk.tokenize.html>

¹⁴<https://spacy.io/api/tokenizer>

¹⁵<https://spacy.io/api/lemmatizer>

¹⁶<https://www.inf.ufrgs.br/~viviane/rspl/index.htm>

¹⁷<https://snowballstem.org/>



Sentenciadores

- **Punkt** é uma ferramenta para a segmentação de sentenças em português, ou seja, é um sentenciador. O NLTK inclui um modelo do Punkt¹⁸, 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.
- No spaCy, a sentencição 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: FreeLing²¹, NLTK²², OpenNLP²³, NLPyPort²⁴, PolyGlott²⁵, spaCy²⁶, StanfordNLP²⁷, TreeTagger²⁸ e LinguaKit²⁹. 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:

¹⁸https://www.nltk.org/howto/portuguese_en.html#sentence-segmentation

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/PolyGlott/>

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>



- 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 Dependencies³⁰ (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 transformers³¹. 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 geral³².
- O **tagger do UDPipe**³³ consiste em vários modelos MorphoDiTa³⁴ 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

³⁰<https://universaldependencies.org>

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>



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 Conjunto de dados, *dataset* e *corpus* aborda *corpus* e *datasets*. Ao final da seção, exploraremos 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 Ferramentas e recursos para o processamento sintático.

- 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 Porttinari-base para ilustrar o contraste com a anotação

³⁷<http://nilc.icmc.usp.br/macmorpho/>

³⁸Dados da última versão (v3), disponível em <http://nilc.icmc.usp.br/macmorpho/>.

³⁹<https://www.linguatca.pt/Floresta/>

⁴⁰<https://sites.google.com/icmc.usp.br/poetisa/resources-and-tools>



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.

Nos repositórios do NILC⁴², POeTiSA⁴³ e Linguatca⁴⁴ 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 Conjunto de dados, *dataset* e *corpus* 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 atual⁴⁶, 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”.

⁴¹<https://github.com/LR-POR/MorphoBr>

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.linguatca.pt>

⁴⁵Para um estudo mais aprofundado do projeto Universal Dependencies, ver Sequência de caracteres e palavras.

⁴⁶O PortiLexicon-UD está disponível no endereço: <https://portilexicon.icmc.usp.br/> e lá também pode ter seus arquivos de dados baixados.

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 participio como “cansado” e “cedido”.
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), sím-



bolos (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

Português	PoS UD	Palavras
Advérbios	ADV	Todos os advérbios
Preposições	ADP	Todas as preposições
Conjunções	CCONJ	Todas as conjunções coordenativas
Coordenativas		
Conjunções	SCONJ	Todas as conjunções subordinativas
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 participípios

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 varie-

⁴⁷Tupla 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”.

dade 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

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
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 distribuição 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;
- 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.

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	<i>feats</i> ou <i>features</i> ou atributos morfológicos	em linguística
Morfossintaxe	palavra	<i>part-of-speech</i> 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 TeP⁴⁸), redes semânticas (e.g. WordNetBr⁴⁹,

⁴⁸<http://www.nilc.icmc.usp.br/tep2/>

⁴⁹<http://www.nilc.icmc.usp.br/wordnetbr/>



VerbNetBr⁵⁰, PropBankBr⁵¹), léxicos, dicionários comuns, dicionários especializados, ontologias, glossários e tantos outros recursos construídos para o português.

⁵⁰<http://143.107.183.175:21380/portlex/index.php/en/projects/verbnbringl>

⁵¹<http://143.107.183.175:21380/portlex/index.php/en/projects/propbankbringl>



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

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 dito¹. 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 Hora de dar nome aos bois 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 e consensual. Para completar a fundamentação teórica, trazemos

¹Como 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*.



uma discussão sobre três características importantes das MWEs: ambiguidade, variabilidade e arbitrariedade.

Após as conceituações, descrevemos na Tarefas: botando a mão na massa 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 Recursos: uma joia rara 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 Avaliação: colocando os pingos nos 'i's, 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 (Até onde chegamos e para onde vamos). 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 Ao infinito e além. 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* (*Datasets* pra quê?) para avaliar sistemas de identificação de MWEs².

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 Sequência de caracteres e palavras 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*

²O 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.

multpalavras já está consolidado, e não pretendemos a essa altura do campeonato erguer 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 compostos³. 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 páginas⁴!

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.

³Veja a história da Rhababerbarbara: <https://www.youtube.com/watch?v=IFoyspFAKnM>

⁴Disponível em: <https://parsemefr.lis-lab.fr/parseme-st-guidelines/>

⁵Para saber mais sobre definições de MWEs, consulte o estudo de Seretan (2011, p. 182–184)



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 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 (E o significado?), 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 (A ordem e a função das palavras em uma sentença), mas o PARSEME se baseia na sintaxe de dependências do Universal Dependencies (Projetos de anotação multilíngue: Universal Dependencies). 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.

⁶A 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).



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 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 expressão⁷.

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 multpalavras 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 acaso⁸. 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 (PLN na Saúde) ou do direito (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,

⁷Essa 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).

⁸Veja Evert (2009) para uma discussão detalhada.

⁹Algumas MWEs são metáforas “congeladas”, embora sua etimologia não seja sempre óbvia, por exemplo, “**lágrimas de crocodilo**” (https://pt.wikipedia.org/wiki/Lágrimas_de_crocodilo).



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 das expressões multipalavras, elas já dão bastante pano pra manga¹⁰. Assim, parece razoável delegar o tratamento de entidades nomeadas e termos a outras comunidades de pesquisa, por exemplo, em extração de informações (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* (Recursos: uma joia rara). 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.

MWEs verbais

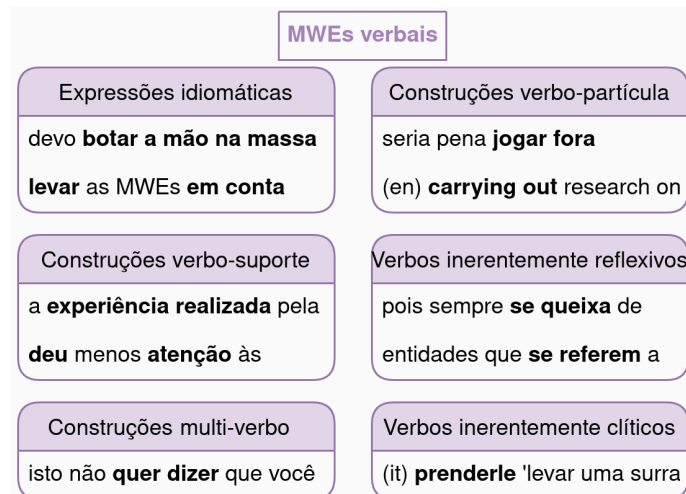
Uma categorização de MWE única, operacional e válida em todas as línguas, abrangendo todos os fenômenos que correspondem à definição do Quadro 5.1, é uma questão de pesquisa aberta e ambiciosa. No entanto, um passo importante nessa direção é a tipologia do PARSEME para MWEs verbais, que abrange um grande 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

¹⁰Em português brasileiro, a troca de *pra* por *para*, ainda que possível, provavelmente geraria algum estranhamento.

¹¹Para uma discussão aprofundada sobre MWEs vs. entidades nomeadas, consulte Candito et al. (2021).



Figura 5.1: Taxonomia das MWEs verbais proposta pelo PARSEME



Fonte: Imagem adaptada e traduzida de (Ramisch, 2023).

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 5.1¹². Falaremos mais sobre essas categorias na *Corpora* de MWEs em português, 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* STREUSLE¹³ 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 Projetos de anotação multilingue: Universal Dependencies. 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

¹²Veja detalhes e exemplos em português em <http://parsemefr.lis-lab.fr/parseme-st-guidelines/>.

¹³Disponível em: <https://github.com/nert-nlp/streusle>

como especificadores do significado ou da função sintática das três categorias principais. A tipologia proposta por Ramisch (2023) estende essas noções às expressões multipalavras.

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 arbitrariedade¹⁴. 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 desambiguação contextual estatística usando aprendizado de máquina (Pasquer et al., 2020). No entanto, seria necessário aumentar consideravelmente a cobertura dos léxicos de MWEs (Recursos: uma joia rara) 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 substituíbilidade** (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 idiosincrasia semântica, já que o resultado geralmente é inaceitável ou agramatical, ou produz uma mudança de significado inesperada (por exemplo, “**bater**

¹⁴Detalhes 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).



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 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 identificação¹⁵.

5.3 Tarefas: botando a mão na massa

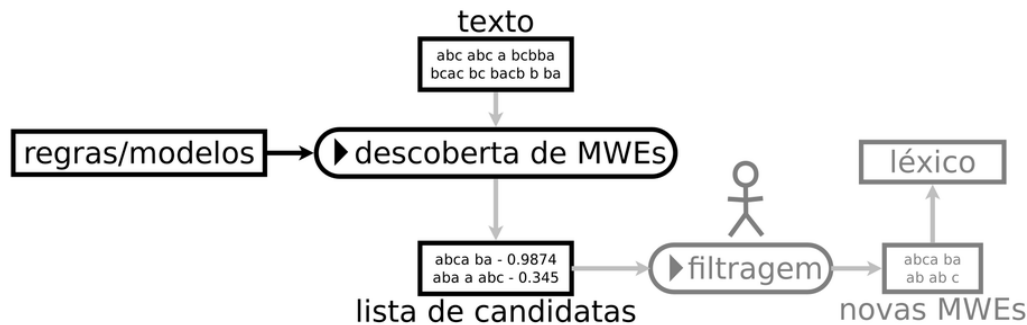
Uma das principais confusões terminológicas da área diz respeito à nomenclatura e à definição das tarefas computacionais relativas às MWEs. O que chamamos de *processamento de MWEs* tem sido chamado na literatura de identificação (Tsvetkov; Wintner, 2011), extração (Tsvetkov; Wintner, 2012), aquisição (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

¹⁵Ainda 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).

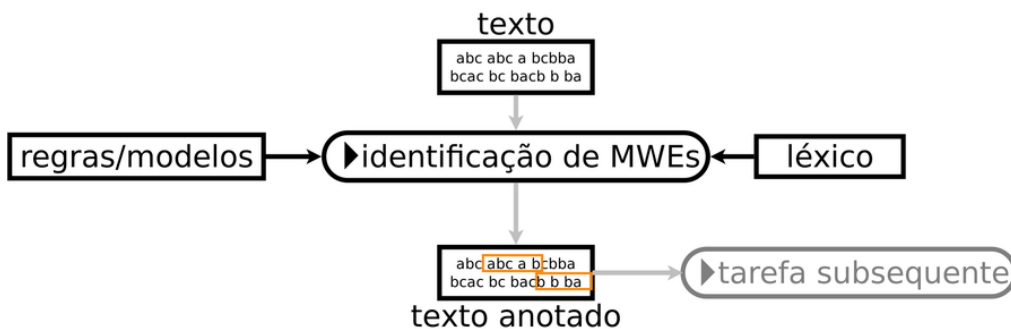


Figura 5.2: Descoberta de MWEs: regras ou modelos extraem do texto uma lista de candidatas. A lista é então filtrada por especialistas, que adicionam novas entradas ao léxico.



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).

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 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 (Avaliação: colocando os pingos nos 'i's').

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 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 *Datasets* de composicionalidade.

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 (Goding 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 (Léxicos de MWEs) e *corpora* (*Corpora* de MWEs). 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 (Extração de Informação). Formas mais sofisticadas de léxicos de MWEs podem incluir informações sobre a categoria, a função morfosintática (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) possam 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-PB¹⁶ (Muniz, 2004), ou como parte da ferramenta de análise sintática do português europeu STRING¹⁷ (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 Losnegard 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 $\pm$ desvio padrão dos mais polêmicos (acima) aos mais consensuais (abaixo).

	MWE	SUB	ADJ	SUB-ADJ
Polêmica+	pavio curto	1.6 $\pm$ 1.8	1.1 $\pm$ 1.9	1.9 $\pm$ 2.3
	sexto sentido	4.0 $\pm$ 1.4	2.5 $\pm$ 2.1	2.8 $\pm$ 2.2
	gelo-seco	3.2 $\pm$ 1.6	3.2 $\pm$ 1.8	3.0 $\pm$ 2.1
	mau-olhado	1.8 $\pm$ 1.2	4.2 $\pm$ 1.5	2.3 $\pm$ 2.1
	câmara fria	3.6 $\pm$ 2.2	5.0 $\pm$ 0.0	3.4 $\pm$ 2.1
	poção mágica	3.3 $\pm$ 2.3	3.8 $\pm$ 1.5	3.4 $\pm$ 2.1
	estrela cadente	1.7 $\pm$ 2.1	4.0 $\pm$ 1.2	1.7 $\pm$ 2.1
Consenso+	farinha integral	5.0 $\pm$ 0.0	4.5 $\pm$ 0.6	5.0 $\pm$ 0.0
	gato-pingado	0.0 $\pm$ 0.0	0.3 $\pm$ 0.5	0.0 $\pm$ 0.0
	núcleo atômico	5.0 $\pm$ 0.0	4.4 $\pm$ 1.8	5.0 $\pm$ 0.0
	pão-duro	0.0 $\pm$ 0.0	1.0 $\pm$ 1.7	0.0 $\pm$ 0.0
	sentença judicial	5.0 $\pm$ 0.0	5.0 $\pm$ 0.0	5.0 $\pm$ 0.0
	tartaruga-marinha	5.0 $\pm$ 0.0	5.0 $\pm$ 0.0	5.0 $\pm$ 0.0
	vôo internacional	5.0 $\pm$ 0.0	5.0 $\pm$ 0.0	5.0 $\pm$ 0.0

¹⁶Disponível em: <http://www.nilc.icmc.usp.br/nilc/projects/unitex-pb/web/>

¹⁷Disponível em: <https://string.hlt.inesc-id.pt/w/index.php/Dictionaries>

¹⁸Por exemplo, eles foram úteis ao escrever este capítulo, para encontrar exemplos de MWEs.



Fonte: Exemplos de (Ramisch et al., 2016b).

Uma informação particularmente interessante para nós é o grau de composicionalidade das MWEs. Como dito na Tarefas: botando a mão na massa, fazer uma predição automática do grau de composicionalidade de um grupo de palavras constitui uma sub tarefa importante na descoberta 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 de avaliar a robustez dos resultados obtidos nos experimentos. Esse conjunto de dados foi estendido com paráfrases (Wilkins et al., 2017) e anotações de grau de composicionalidade em contexto de sentenças (Garcia et al., 2021; Tayyar Madabushi et al., 2022).

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* (*Corpora*), 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 (Eryigit 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 anotação¹⁹. 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.

¹⁹Disponível em: <https://parsemefr.lis-lab.fr/parseme-st-guidelines/1.0/>



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 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 Avaliação: colocando os pingos nos 'i's.

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, 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.

²⁰As categorias foram definidas na MWEs para dar e vender.



5.5 Avaliação: colocando os pingos nos 'i's

Como descrito na Tarefas: botando a mão na massa, podemos dividir as tarefas de processamento de expressões multpalavras em dois grupos: descoberta de MWEs e identificação de MWEs. Cada tipo 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 (Léxicos de MWEs), 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 MWEs²¹. 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 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 (MWEs para dar e vender) 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)

Gold: tomar₁ uma decisão₁ a fim de₂ segmentar frases em itens lexicais₃
Sistema: tomar₁ uma decisão₁ a fim₂ de segmentar frases₃ em itens lexicais₄

Medidas baseadas em MWEs:

MWEs_corretas = 1 $P = 1/4$ $R = 1/3$ $F = 2/7 \approx 0.28$

Medidas baseadas em tokens:

Tokens_corretos = 5 $P = 5/7$ $R = 5/7$ $F = 5/7 \approx 0.71$

Fonte: Adaptado de <https://gitlab.com/parseme/mwesinanutshell/>

As medidas de avaliação baseadas em MWEs recompensam apenas combinações com-

²¹Veja mais sobre Avaliação de sistemas de PLN no Avaliação de tecnologias de linguagem.



pletas, 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. Portanto, 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 exemplo²².

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 (Taslimipour 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 multpalavras. 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 verbais²³. 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 *Datasets* pra quê?. Em primeiro lugar, a supervisão humana em geral é muito mais eficaz do que aumentar

²²Para 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*.

²³Se 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 satisfeitos pelos avanços nesse quesito.



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 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 reprodutibilidade. 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 MWEs²⁴. 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

²⁴Esse é um dos objetivos do projeto UniDive: <https://unidive.lisn.upsaclay.fr/>



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* (Ética em PLN).

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 (*Datasets* de composicionalidade), ou ainda a vinculação das unidades textuais anotadas às unidades lexicais correspondentes, como entradas em uma Wordnet (*Wordnets*) ou *frames* semânticos (FrameNet).

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 vieses 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 con-



ferências em linguística computacional²⁵. 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 Linguistics, 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 área²⁸. 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 idiomas³¹.
- 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 idiomatidade em contexto³³.

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 Idiomatity 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.

²⁵Disponível em: <https://multiword.org>

²⁶Disponível em: <https://aclanthology.org/>

²⁷Última edição no momento da escrita do capítulo: <http://europhras.com/2022/mumttt-2022-2/>

²⁸Disponível em: <https://langsci-press.org/catalog/series/pmwe>

²⁹Disponível em: <http://parseme.eu>

³⁰Disponível em: https://clarino.uib.no/iness/page?page-id=MWEs_in_Parseme

³¹Disponível em: <https://gitlab.com/parseme/corpora/-/wikis/home>

³²Disponível em: <https://dimsum16.github.io>

³³Disponível em: <https://sites.google.com/view/semEval2022task2-idiomaticity>



Parte IV

Estrutura



Capítulo 6

A ordem e a função das palavras em uma sentença

Sintaxe

Adriana S Pagano
Amanda Rassi
Ana Clara S Pagano

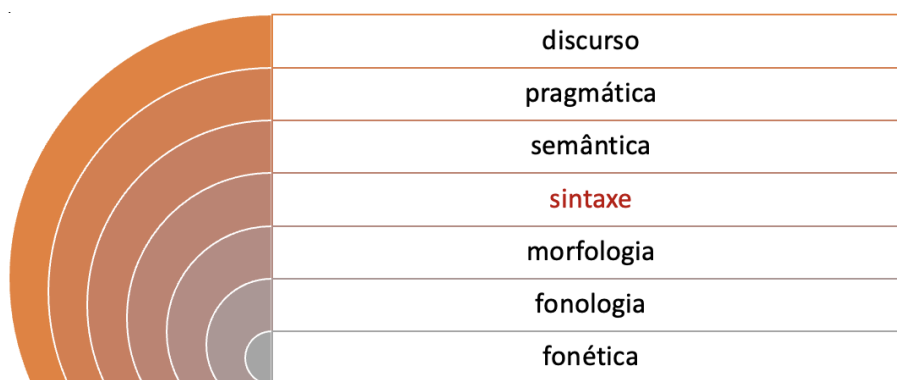
Publicado em: 26/09/2023

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 O que é PLN?), 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

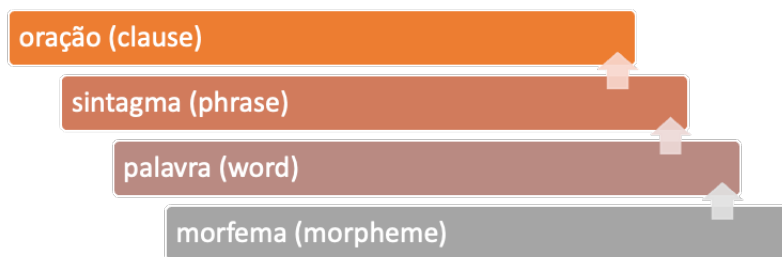


Ainda em relação ao sistema linguístico, além de sua organização em estratos, temos uma segunda forma de organização: a escala de ordens (em inglês, *rank scale*), que organiza



diferentes unidades de análise em níveis hierárquicos. A escala de ordens¹ 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 constituição e a sintaxe de dependência. Por último, abordaremos as interseções entre a sintaxe e os demais estratos do sistema linguístico ao final deste capítulo, na Fronteiras da sintaxe.

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 Ferramentas e recursos para o processamento sintático, em PLN a análise sintática automática é realizada por meio de softwares denominados *parsers*. Também

¹Cabe 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*”.

²Em 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).

veremos alguns desafios que a análise em constituintes 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 Ferramentas e recursos para o processamento sintático, 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”.



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”.

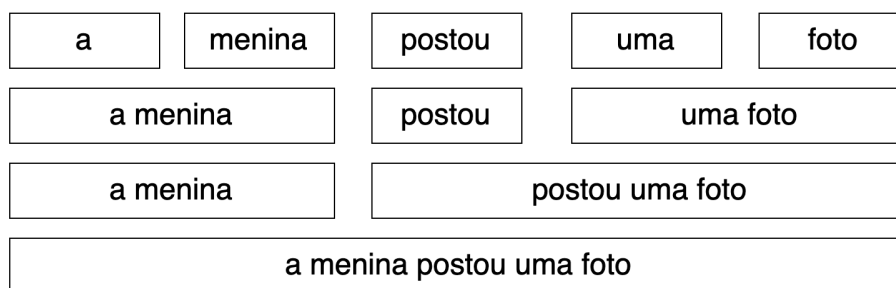
Com estas reflexões iniciais, passamos agora a revisar alguns conceitos básicos de sintaxe.

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

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

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 constituição 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.



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]]]]][.]]

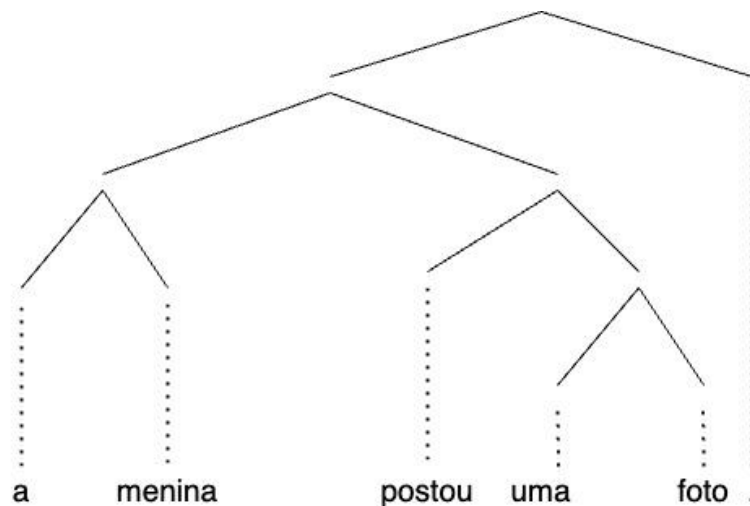
```

Este tipo de representação, como veremos neste capítulo, pode ser feito tanto para sintaxe de constituintes 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



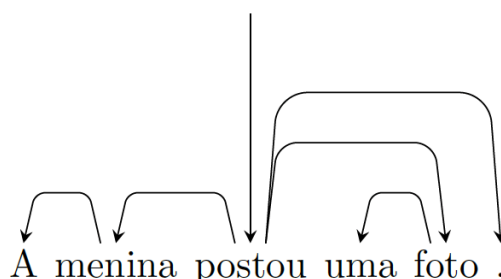
Na notação com estruturas arbóreas, um diagrama chamado de árvore representa graficamente a estrutura de uma sentença como uma hierarquia, sendo os **nós** (em inglês, *nodes*), elementos discretos (ou elementos finais) em um grafo (ou representação simbólica) e as **arestas** (em inglês, *edges*) linhas 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 constituição 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 Sintaxe de dependência.

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.



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

1	postou
2	menina
3	a
4	foto
5	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 constituintes e o *parsing* de dependência.

6.5 Sintaxe de constituintes

Na sintaxe de constituintes, 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 constituintes, 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,

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.

Vejamus 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 árvore³ na Figura 6.9.

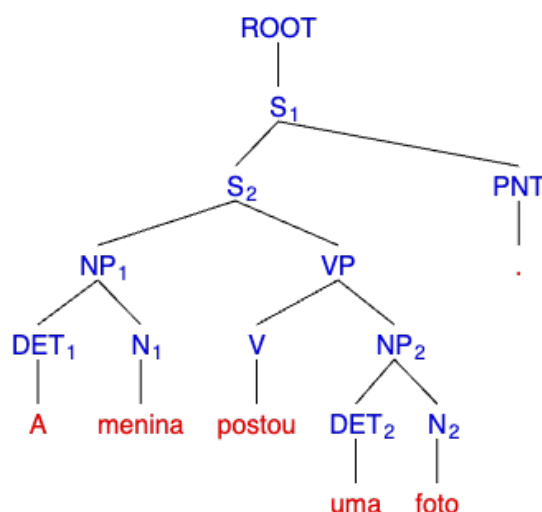
A árvore pode ser lida, de cima para baixo, da seguinte forma: O nó denominado **ROOT** indica a raiz da árvore. O nó **S1** representa a sentença toda como um único constituinte; é o nó que domina todos os outros nós da árvore. No caso de uma sentença grafológica (unidade delimitada por um sinal de 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 constituição, 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

³As árvores de constituição 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/>.



Figura 6.9: Representação de sintaxe de constituição 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.

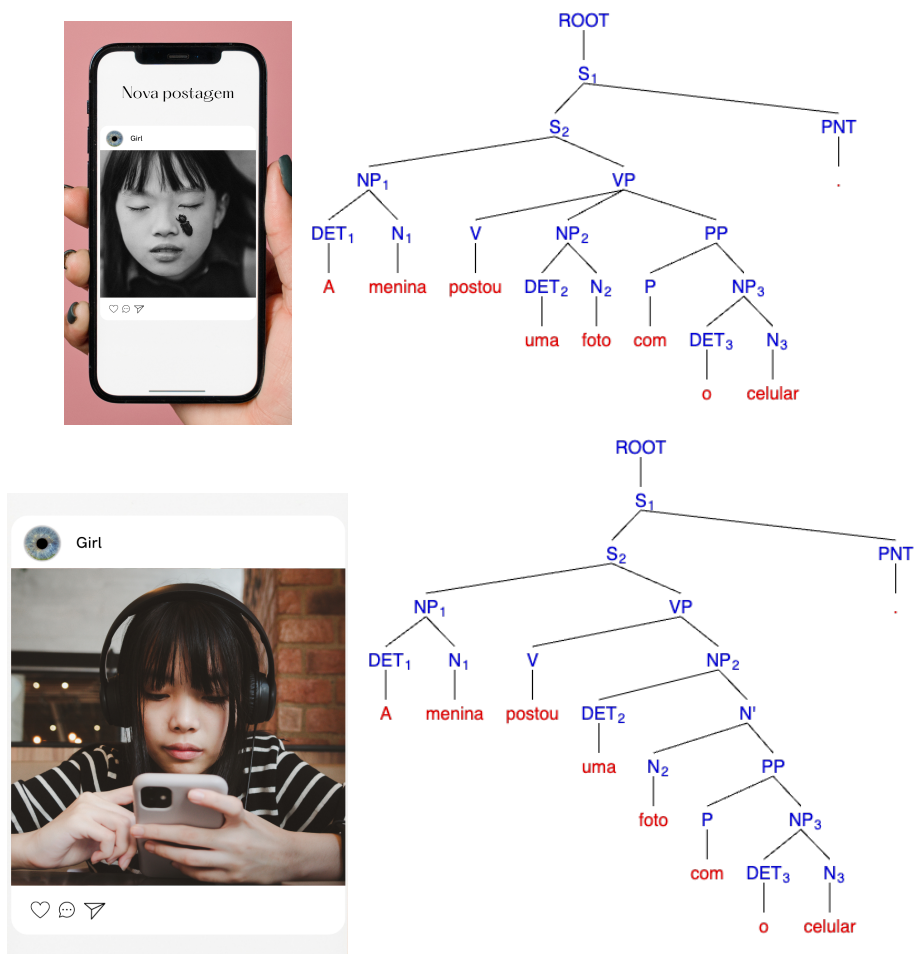
A menina postou uma foto com o celular.

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 6.10⁴.

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.

⁴Imagens geradas com o aplicação <https://gencraft.com/> e modificadas pelas autoras.

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

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 constituintes, 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 Núcleo e dependente. 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



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.

⁵Entendemos que todos esses termos são sinônimos, mas adotaremos neste capítulo o termo “núcleo”.

⁶Vale ressaltar que estes termos também são usados em inglês: “*parent*”, “*child*”, “*grandparent*”, “*grand-child*” e “*sibling*”.

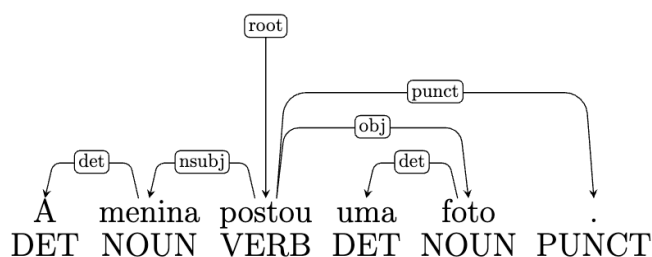


6.6.2 A representação da sintaxe de dependência

Assim como na sintaxe de constituição, 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 morfo-sintática, 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 Projetos de anotação multilingue: Universal Dependencies.

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”).

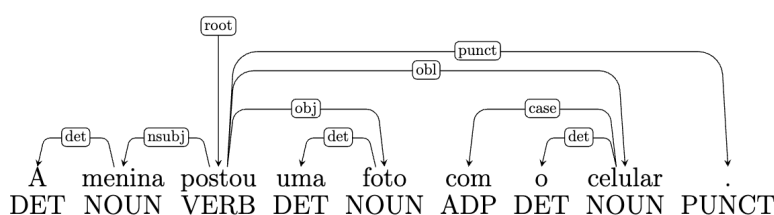
Em uma árvore de dependência, algumas classes de palavras podem ser núcleo; outras, não. Mas cada palavra é dependente de uma outra. Isso significa que, na representação por diagrama com setas, em toda palavra deve chegar alguma (e exclusivamente uma) seta. A única exceção é para a raiz, na qual também 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 Núcleo e dependente, o ponto de partida da seta representa o

núcleo, e 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.

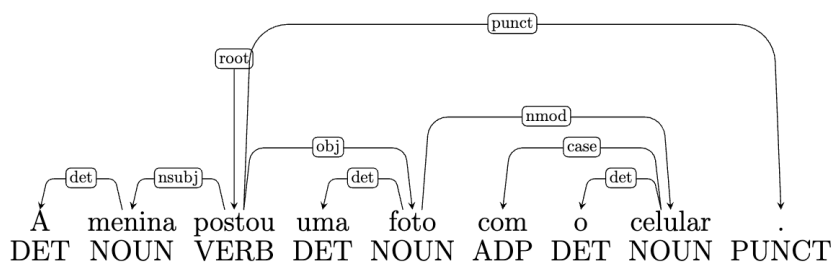
Na Figura 6.12, temos a representação da sentença “A menina postou uma foto com o celular”, aqui anotada em sintaxe de dependência. Nela, “com o celular” é interpretado como Instrumento e como uma função dependente do verbo “postou”.

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.

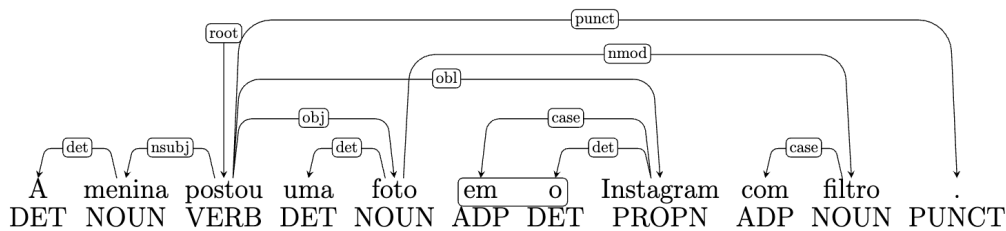
Exemplo 6.8.

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

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,

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 (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 multilíngue: 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 línguas⁷. 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.

⁷O *parser* Stanford também dispõe de um conversor que transforma uma árvore de constituição 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>.

Outra iniciativa relevante foi o **Intersect interlândia** (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 Intersect, 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 básicas⁹ 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.

⁸Tradução 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)*”.

⁹As premissas podem ser consultadas em <https://universaldependencies.org/introduction.html>.



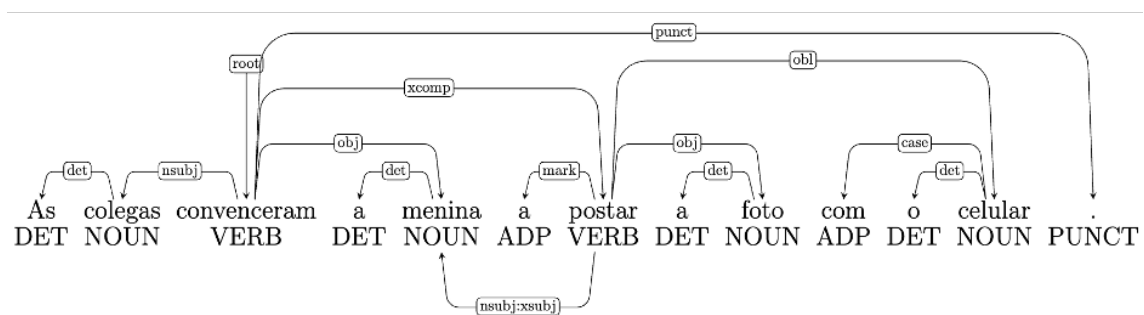
6. A proposta das UD precisa apoiar bem as tarefas de Compreensão de Linguagem Natural (O que é PLN?) subsequentes ao *parsing* sintático (extração de relações, compreensão de textos, tradução automática, dentre outras).

Além da representação básica por dependência já explicitada ao longo da Sintaxe de dependência, 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 *nsbj*, o sujeito ("as colegas") do verbo principal ("convenceram"). Nas relações *enhanced*, podemos indicar, por meio da relação *nsbj: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 Sintaxe de constituição, 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".



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 UD¹⁰

	Nominals	Clauses	Modifier words	Function words
Core arguments	nsubj obj iobj	csbj ccom xcomp		
Non-core dependents	obl vocative expl dislocated	advel	admod discourse	aux cop mark
Nominal dependents	nmod appos nummod	acl	amod	det clf case
Coordination	MWE	Loose	Special	Other
conj	fixed	list	orphan	punt
cc	flat	parataxis	goeswith	root
	compound		rearandum	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

¹⁰Referência: <https://universaldependencies.org/u/dep/all.html>.



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**).

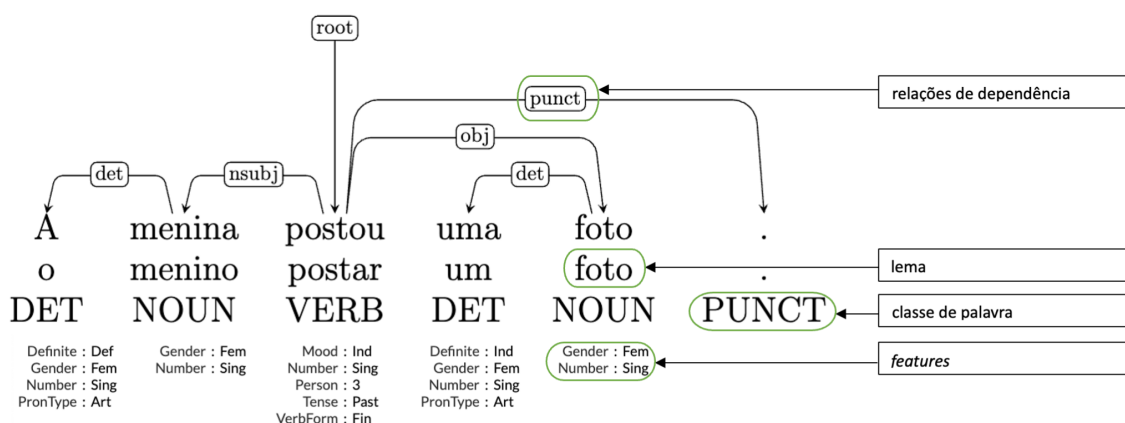
Também podem ser relações que correspondem a funções adicionais àquelas básicas (*non-core dependents of clausal predicates*), tais como adjuntos (**obl**), orações adverbiais (**advcl**), vocativos (**vocative**) etc. Essas relações se aplicam tanto no nível da oração como do sintagma, como é o caso 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) (Expressões multipalavras), 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 UD pode ser observada na Figura 6.18.

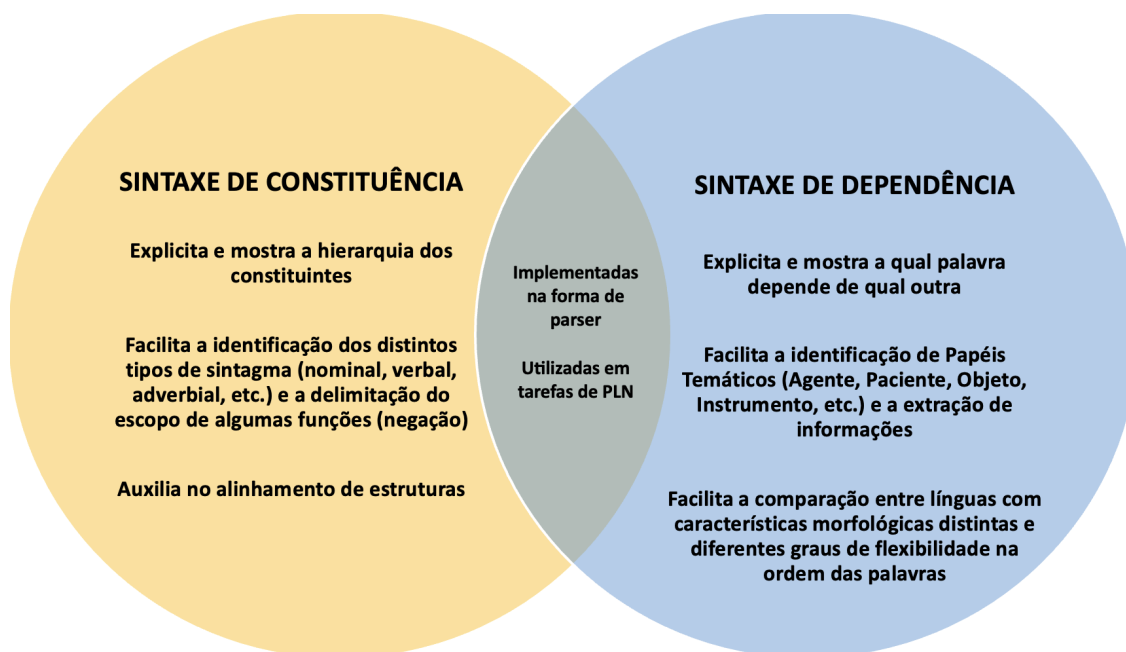
Figura 6.18: Visualização de todas as informações anotadas para uma sentença no projeto UD



6.7 Qual é melhor: constituintes 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.

Figura 6.19: Vantagens e desvantagens da sintaxe de constituição e de dependência



6.8 Fronteiras da sintaxe

Conforme apontado na Introdução 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 morfológicamente 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: “médico” e “médica”. Há, porém, substantivos chamados “epícenos”, 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

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 “incapaz” ou “normal” x “anormal” 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 constituintes nos permite identificar dois constituintes maiores: “João” e “deu um pulo”, este último



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 (Modelos discursivos) 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



discursivas, tais como RST (*Rhetorical 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**, **constituente**, **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 (Ferramentas e recursos para o processamento sintático) 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.

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

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 A ordem e a função das palavras em uma sentença, 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 A ordem e a função das palavras em uma sentença, 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 constituição e *parsers* de dependência.



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).

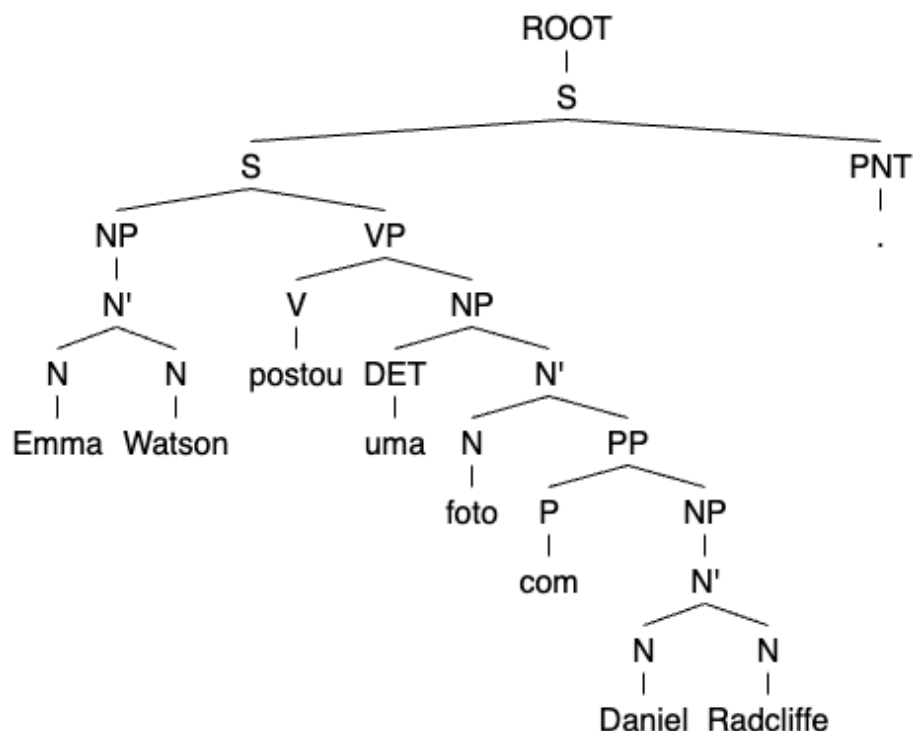
Tanto o *parsing* de constituintes como o *parsing* de dependência podem ser executados de forma completa ou parcial. Tomando como exemplo o *parsing* de constituintes, uma análise completa ou *deep* extrai todos os agrupamentos e as relações sintáticas em uma sentença. Por exemplo, dada a sentença:

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.

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

```

1  !pip install spacy
2  !python -m spacy download pt_core_news_sm
3  import spacy
4  nlp = spacy.load("pt_core_news_sm")
5  sentence = "Emma Watson postou uma foto com Daniel Radcliffe."
6  doc = nlp(sentence)
7  for chunk in doc.noun_chunks:
8      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 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.



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 Bosque¹, amplamente utilizado para treinar modelos de análise sintática (Veja Conjunto de dados, *dataset* e *corpus*).

O *corpus* Bosque é parte de um *corpus* maior, chamado Floresta Sintá(c)tica², 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 Bosque³ foi convertida para o formato UD (Universal Dependencies), apresentado no A ordem e a função das palavras em uma sentença, 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 – CINTIL⁴, 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-UDep⁵, é disponibilizada com anotações no padrão UD.

Mais recentemente, o *corpus* PetroGold⁶ 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 Artificial⁷ (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 POeTiSA⁸ desenvolve o *treebank* Porttinari⁹, um *corpus* multi-gênero de textos em português brasileiro anotados de acordo com o padrão UD. Inclui textos jornalísticos do *corpus* da Folha de São Paulo/Kaggle, o *corpus* MAC-MORPHO¹⁰ de textos jornalísticos, o *corpus* DANTE¹¹ (*Dependency-ANalised corpora of TwEets*), integrado por tweets da Bolsa de Valores, B2W-reviews01¹², 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 disponível¹³.

Uma iniciativa também em andamento é o *corpus* Veredas¹⁴, desenvolvido na Faculdade de Letras da UFMG, que visa à construção de *treebanks* de textos anotados de acordo com

¹<https://www.linguateca.pt/Floresta/corpus.html>

²<https://www.linguateca.pt/Floresta/principal.html>

³https://universaldependencies.org/treebanks/pt_bosque/index.html

⁴<http://cintil.ul.pt/>

⁵https://universaldependencies.org/treebanks/pt_cintil/index.html

⁶https://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>

¹⁰Mac-Morpho: <http://www.nilc.icmc.usp.br/macmorpho/>

¹¹Brazilian Stock Market Tweets with Emotions|Kaggle: <https://www.kaggle.com/datasets/fernandojvda-silva/stock-tweets-ptbr-emotions>

¹²<https://opencor.gitlab.io/corpora/real19b2wreviews01/>

¹³Corpus Porttinari: <https://sites.google.com/icmc.usp.br/poetisa/porttinari>

¹⁴<http://www.letras.ufmg.br/veredas/>



o padrão das UD. Inclui amostras de uma variedade de textos em inglês, espanhol e português 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 Curupira¹⁵, Donatus¹⁶ e PassPort¹⁷. 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 NLTK¹⁹ 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 spaCy²⁰, Stanza²¹ 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/>

¹⁷PassPort (A Dependency Parsing Model for Portuguese | SpringerLink): https://link.springer.com/chapter/10.1007/978-3-319-99722-3_48

¹⁸No 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.

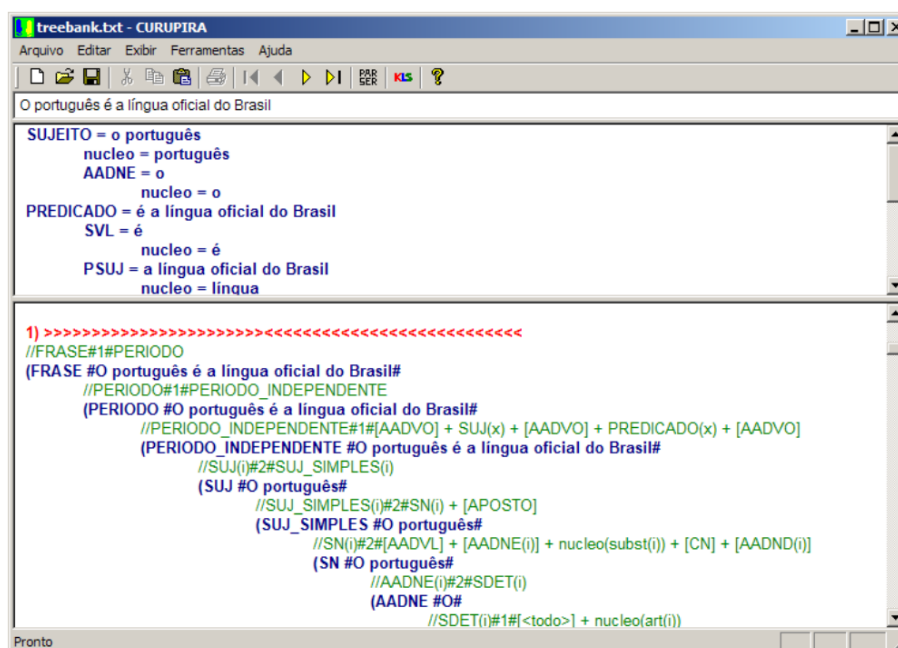
¹⁹NLTK :: *Natural Language Toolkit* <https://www.nltk.org>

²⁰<https://spacy.io/>

²¹<https://stanfordnlp.github.io/stanza/index.html>



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.

Stanza²² 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 *Regex 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

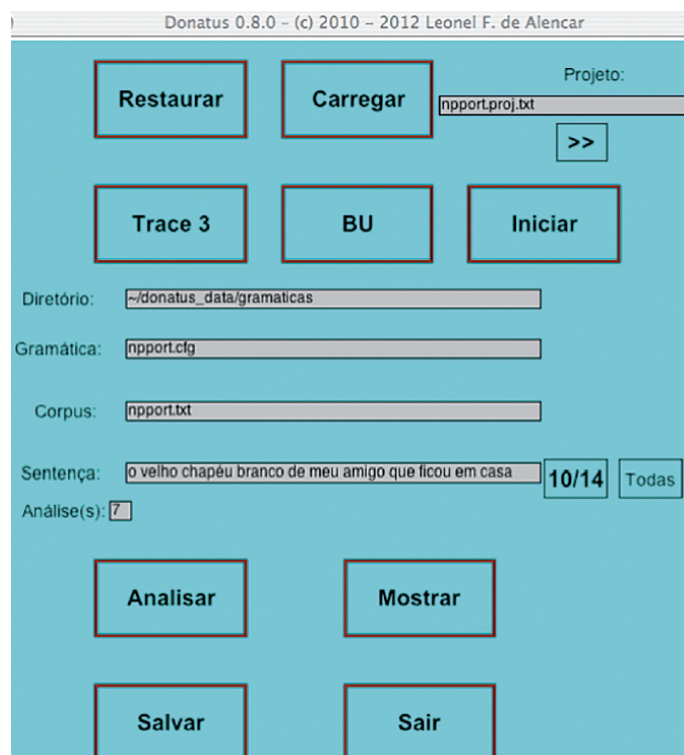
A seguir apresentaremos algumas das ferramentas web disponíveis para *parsing* de texto em português, permitindo executar a análise sintática e obter como saída arquivos com distintos formatos, incluindo a visualização das árvores sintáticas. Todas as ferramentas apresentadas são de acesso livre e gratuito.

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

²²Cabe 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.

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 online²⁴. 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).

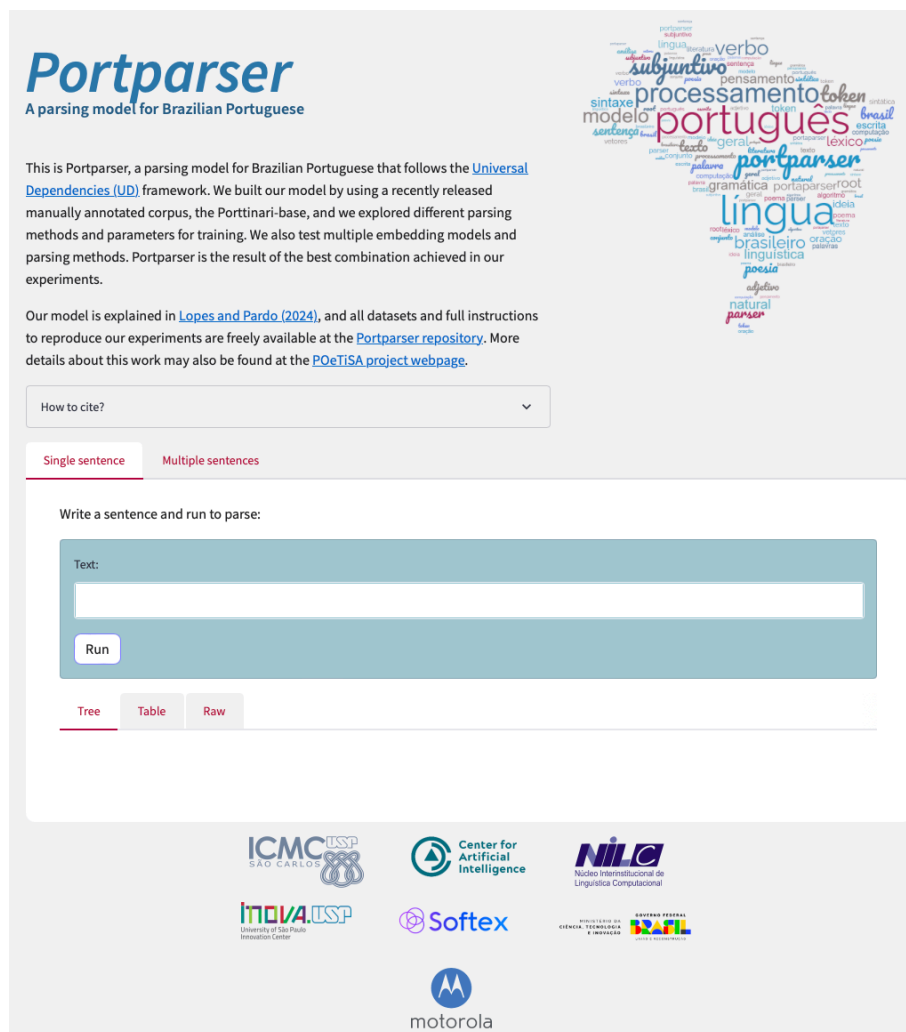
O Portparser aceita entrada de sentenças isoladas (Single sentence) ou upload de arquivo com diversas sentenças (Multiple sentences). No caso de entrada de um arquivo com diversas sentenças, o Portparser oferece a possibilidade de executar a segmentação sentencial do texto ou considerar que o texto já 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/>

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.

Figura 7.7: Execução do Portparser para a sentença “Gostei da variedade e da qualidade.” visualizando a árvore de dependência gerada.

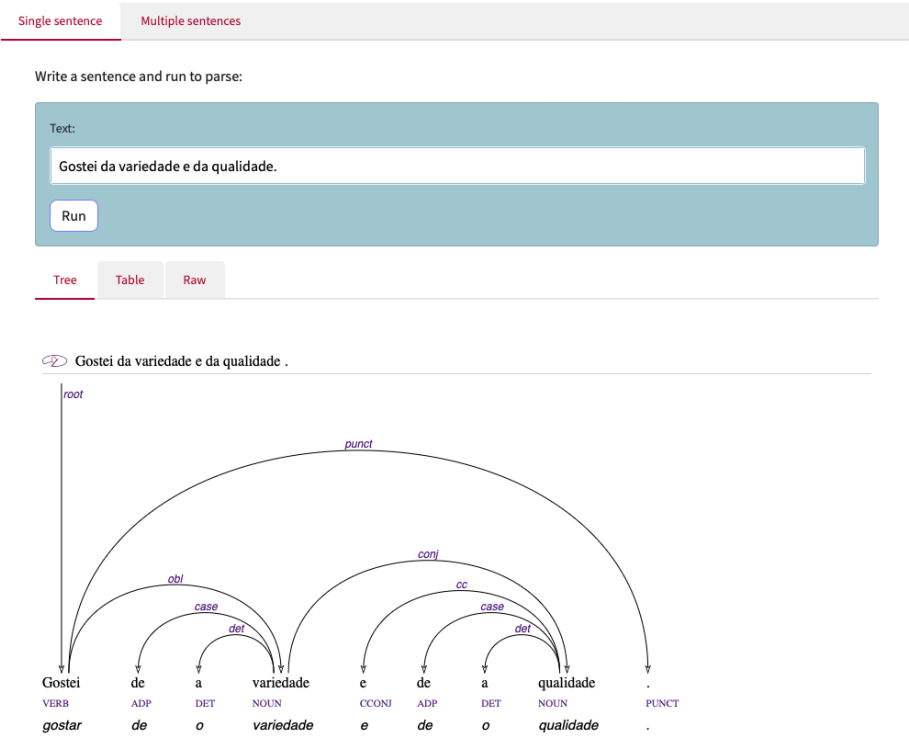


Figura 7.8: Execução do Portparser para a sentença “Gostei da variedade e da qualidade.” visualizando os campos do formato CoNLL-U gerado em forma de tabela.

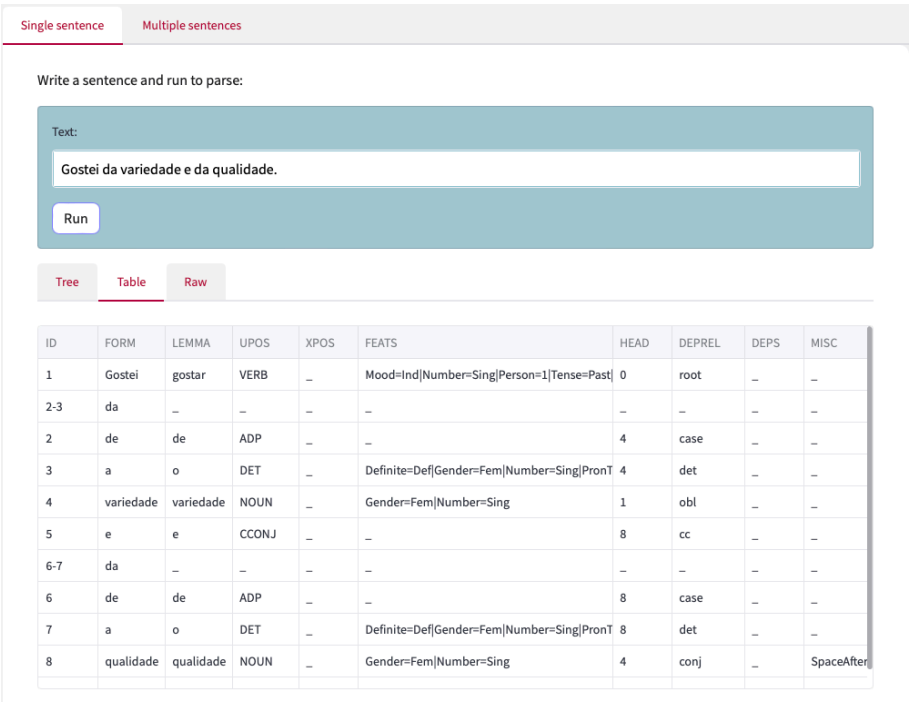


Figura 7.9: Execução do Portparser para a sentença “Gostei da variedade e da qualidade.” visualizando os campos do formato CoNLL-U gerado em forma de texto puro.

Single sentenceMultiple sentences

Write a sentence and run to parse:

Text:

Gostei da variedade e da qualidade.

Run

TreeTableRaw

```
# newdoc id = /tmp/tmp706ile4n/21082024_1154156515253_input.conllu
# newpar
# sent_id = S0001
# text = Gostei da variedade e da qualidade.
1      Gostei  gostar  VERB      -      Mood=Ind|Number=Sing|Person=1|Tense=Past|VerbForm=Fin  0      root
2-3    da      -      -      -      -      -      -      -
2      de      de      ADP      -      -      4      case      -
3      a      o      DET      -      Definite=Def|Gender=Fem|Number=Sing|PronType=Art  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|PronType=Art  8      det
8      qualidade  qualidade  NOUN      -      Gender=Fem|Number=Sing  4      conj      -
9      .      .      PUNCT     -      -      1      punct     -
```

O Parser LX

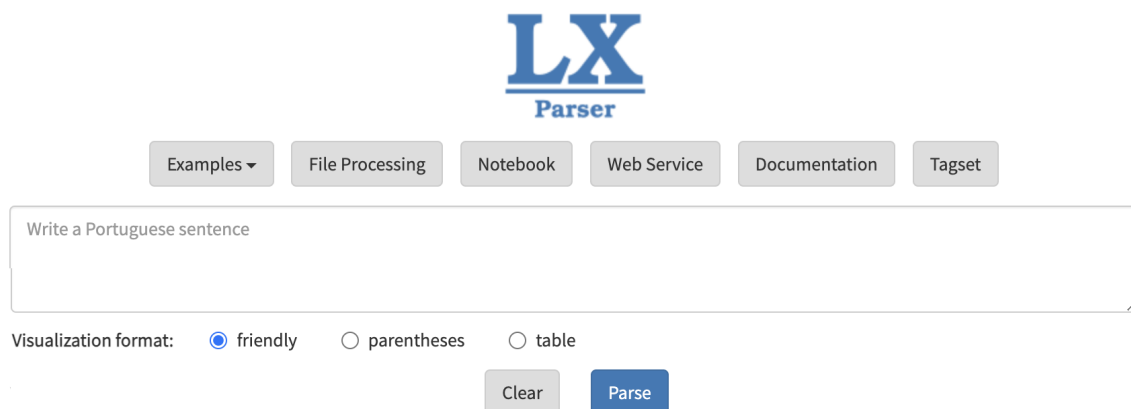
O Parser LX, ferramenta integrante do PORTULAN CLARIN²⁵, é parte integrante de um portal de acesso a infraestrutura de tecnologia linguística no escopo do projeto internacional CLARIN ERIC²⁶.

O Parser LX é disponibilizado tanto para *parsing* de constituição²⁷ como de dependência, este último em duas versões: LX-DepParser²⁸ e LX-UDParser²⁹

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 constituição.

Figura 7.10: Captura de tela mostrando a interface do *parser* de constituição 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 constituição 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 A ordem e a função das palavras em uma sentença.

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 A ordem e a função das palavras em uma sentença.

Para a sentença “Emma Watson postou uma foto com Daniel Radcliffe”, o *parser* gera a árvore de constituição 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/>



Figura 7.11: Diagrama de árvore de constituição gerado pelo *parser* LX.

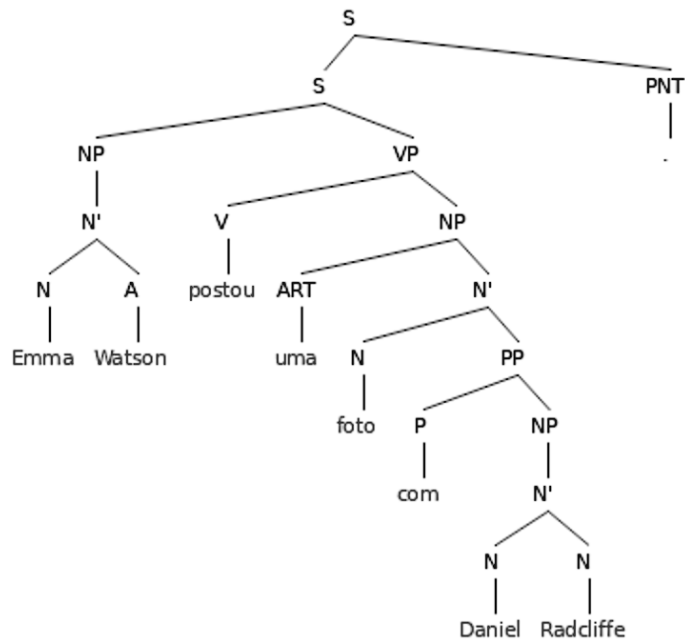
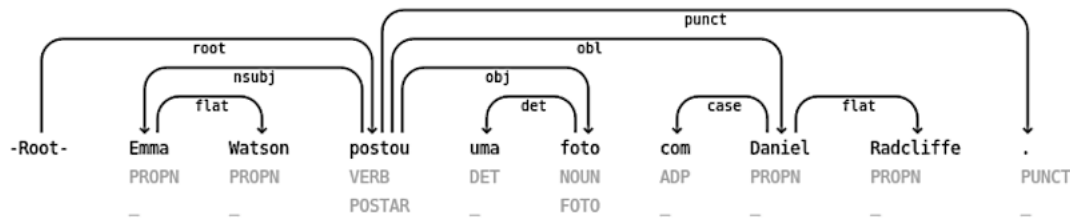


Figura 7.12: Captura de tela mostrando a interface do *parser* de dependência LX.



Figura 7.13: Diagrama com setas gerado pelo Parser LX.

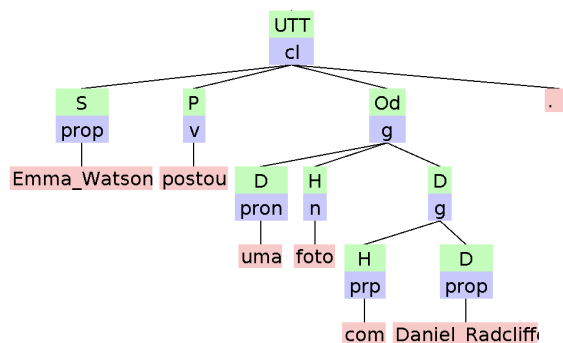


O Parser VISL³⁰

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 constituinte 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 constituinte apresentada na Figura 7.14.

Figura 7.14: Diagrama de árvore de constituinte 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/>

Figura 7.15: Captura de tela mostrando a interface do UDPipe.

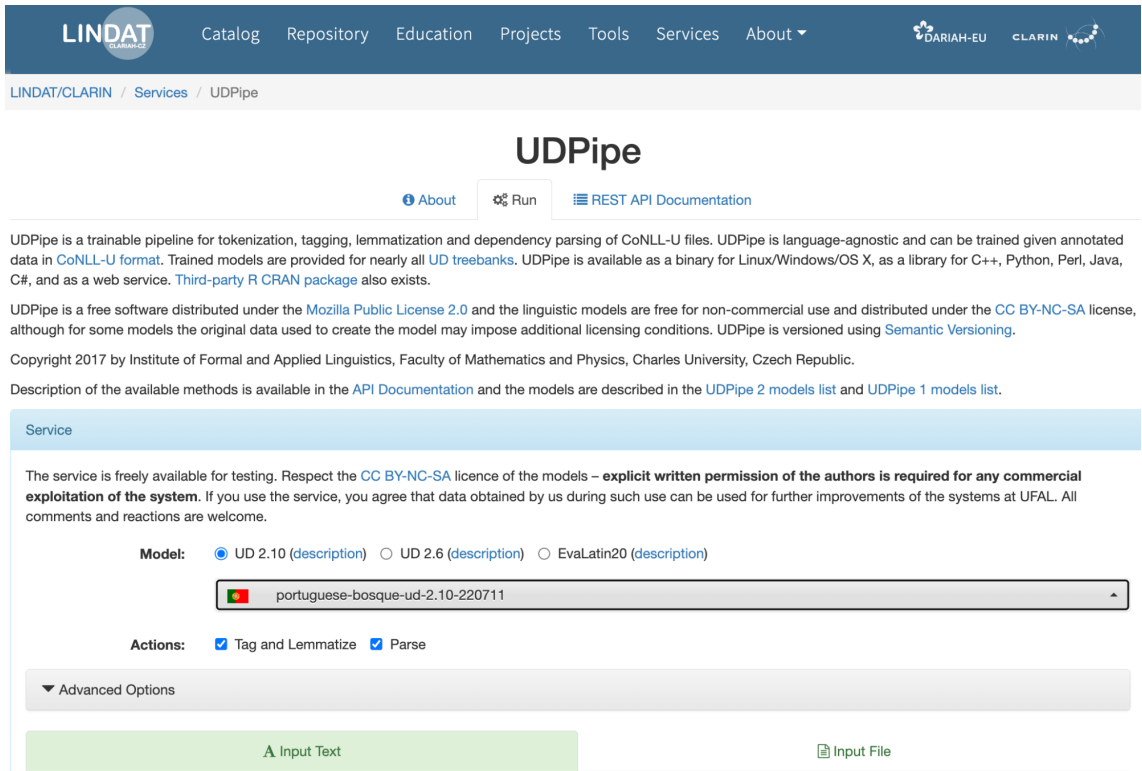
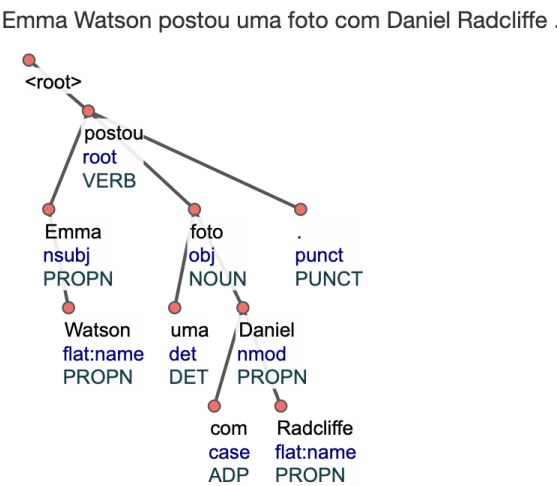


Figura 7.16: Diagrama de árvore de dependência gerado pelo UDPipe.



7.4 Visualização, anotação e edição de *treebanks*

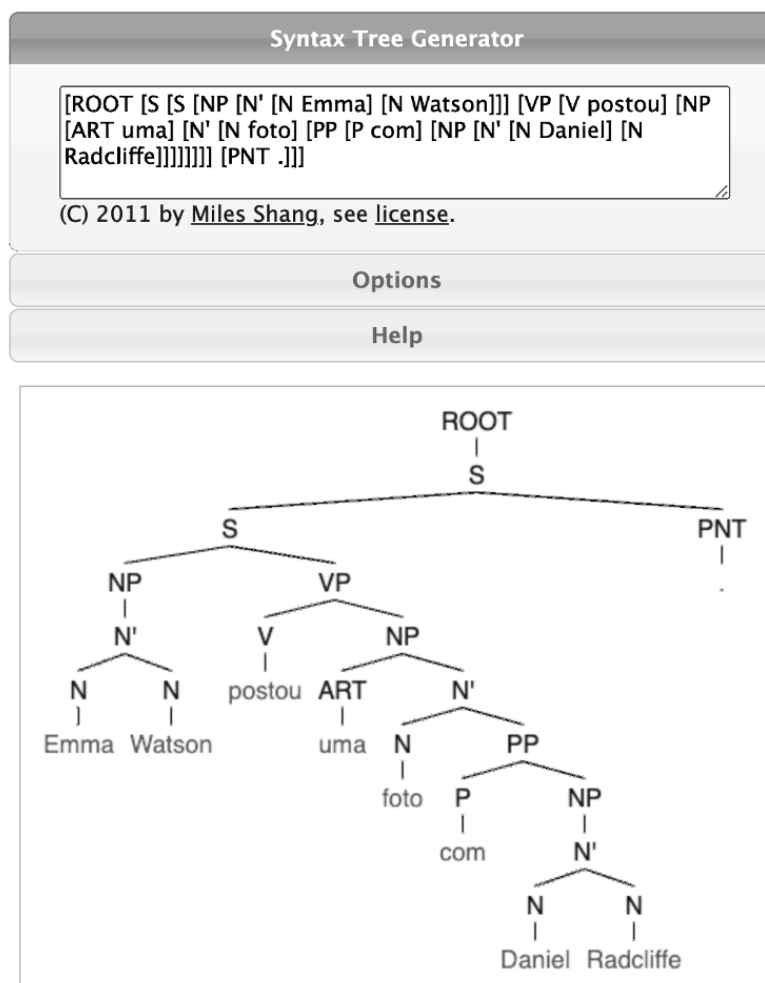
Há diversas ferramentas que permitem tanto a visualização de árvores de constituição e dependência, como a anotação de sentenças e a edição de sentenças já anotadas manualmente ou automaticamente.

7.4.1 Árvores de constituição

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 constituição 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/>

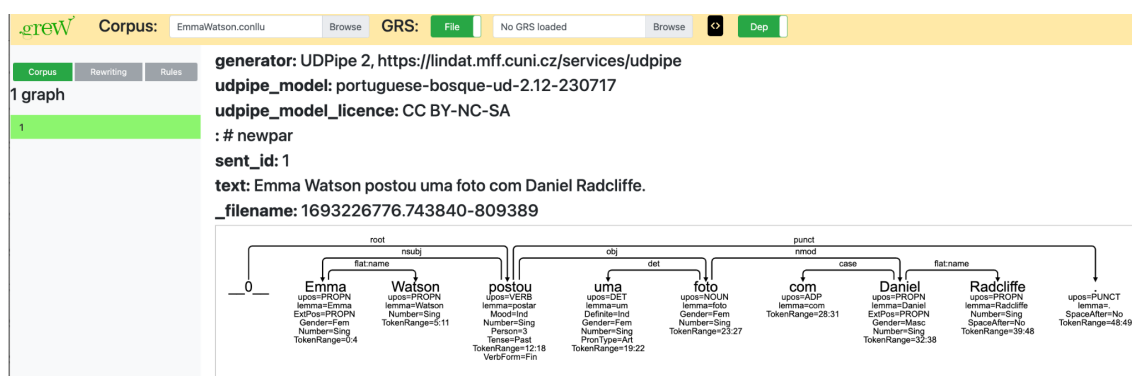
- RSyntaxTree (escrita na linguagem de programação Ruby)³³

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 ArboratorGrew³⁶, que possui uma versão customizada no Brasil pela equipe do ICMC da USP: Arborator-Grew-NILC³⁷. 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/>

³⁴https://universaldependencies.org/conllu_viewer.html

³⁵<https://weblicht.sfs.uni-tuebingen.de/Tundra/>

³⁶<https://arboratorgrew.elizia.net/#/>

³⁷<https://arborator.icmc.usp.br/#/>

Figura 7.19: Captura de tela da ferramenta Grew Match.

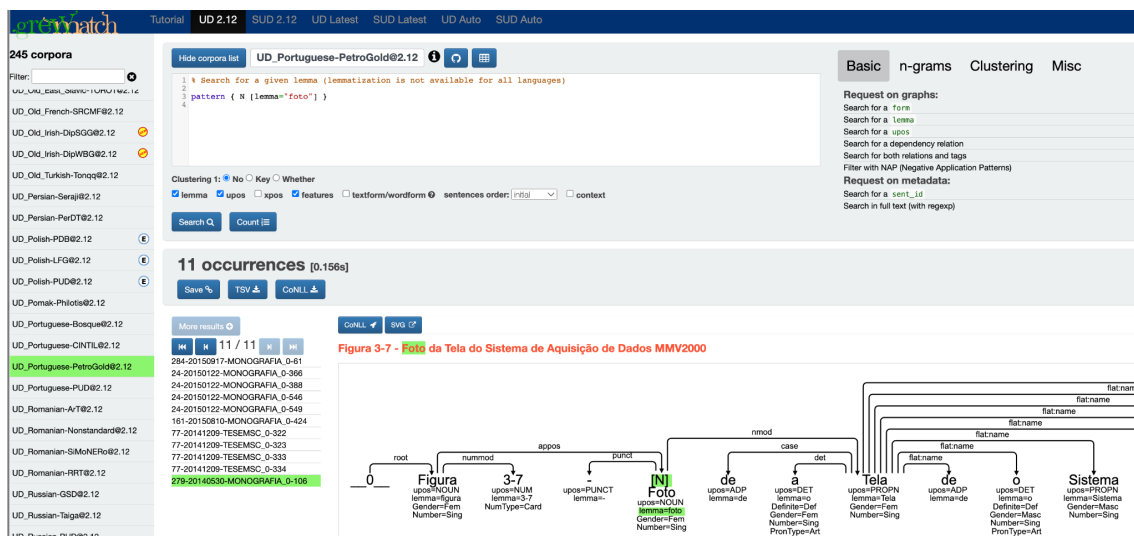


Figura 3-7 - Foto da Tela do Sistema de Aquisição de Dados MMV2000

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 atualizando a lista de ferramentas disponíveis, a qual pode ser consultada no site do projeto³⁸.

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 é INCEpTION³⁹, 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/>



Figura 7.20: Captura de tela no momento de edição de etiqueta de PoS na ferramenta Arborator-Grew-NILC.

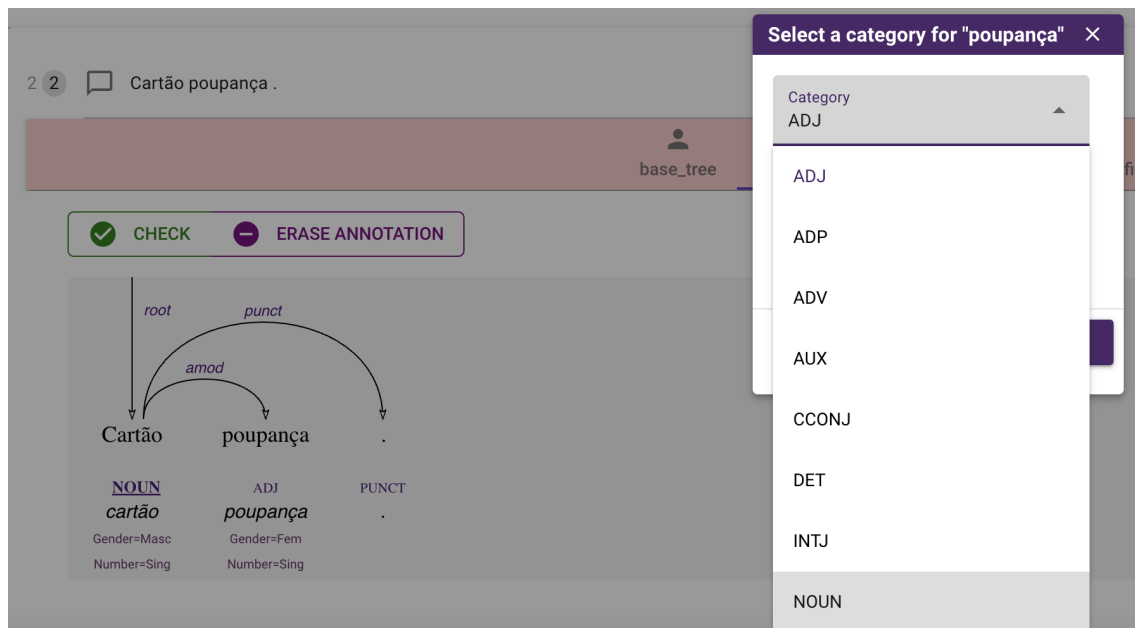
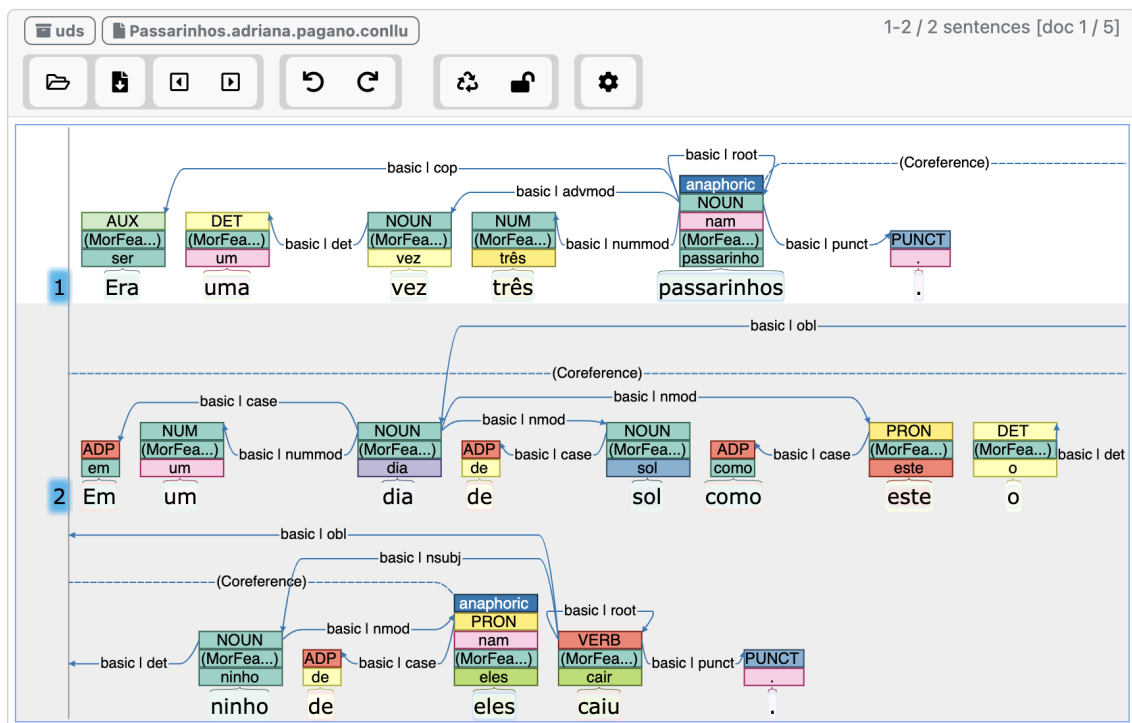


Figura 7.21: Captura de tela com anotação em múltiplos níveis na ferramenta INCEP-TION.



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.



Parte V

Significado



Capítulo 8

E o significado?

Cláudia Freitas

Publicado em: 26/09/2023

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 Semântica com técnicas simbólicas e Semântica distribucional.

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 mundo¹. 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

¹Devo 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.



uma onda da que se lhe segue de imediato e que parece às vezes suplantá-la ou acrescentar-se 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 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 Semântica com técnicas simbólicas) ou usando representações distribuídas (veja Semântica distribucional), 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 numeradas), faz parecer que os significados das palavras se organizam “naturalmente” assim². No PLN, reconhecemos esta maneira de lidar com o significado em recursos como *word-nets*, 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 Semântica com técnicas simbólicas.

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 compreensão** (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 Semântica distribucional)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 vimos⁷. 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

²Por trás de dicionários estão lexicógrafos e decisões editoriais.

³<http://wordnet.princeton.edu/>

⁴Dentre 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.

⁵Mas mesmo estas poderiam ser aceitas se assim fosse previamente estipulado.

⁶Mas não se alinha à busca do algoritmo capaz de fornecer a representação distribuída “correta”, ou “verdadeira”.

⁷É 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”.



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 (ou entidade) externa que determina o significado da palavra. Vetores produzidos a partir de conjuntos de dados diferentes irão levar a representações diferentes⁸.

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 Kilgarrieff (1997); Kilgarrieff (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 Conjunto de dados, *dataset* e *corpus*), 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*, *verbnet*s e *framenet*s) 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 online⁹, aquela adequada ao contexto (se a anotação usasse o inventário de uma *wordnet* como fonte, o inventário de significados poderia ser diferente¹⁰).

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.*

⁸Mas 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>

¹⁰Aqui é possível consultar a OpenWordNet-PT para os significados de “trabalho”: https://www.openwordnet-pt.org/search?search_field=all&term=trabalho



7. Grande esforço; TRABALHÃO; TRABALHEIRA
8. Exercício para treino: *A professora passou muito trabalho para casa.*
9. Ação contínua de uma força da natureza e seu efeito: *O trabalho do vento resulta na erosão eólica.*
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 Conjunto de dados, *dataset* e *corpus*), 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 previsto¹¹. 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”?

¹¹A 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.

¹²Dicionários mudam definição de família <https://www.metropoles.com/vida-e-estilo/comportamento/dicionario-houaiss-muda-significado-da-palavra-familia>



Os próximos capítulos aprofundam cada uma dessas maneiras de trabalhar com o significado no PLN.



<https://brasileiraspln.com/livro-pln/4a-edicao/volume1/>



Capítulo 9

Semântica com técnicas simbólicas

Eloize Rossi Marques Seno
Valéria de Paiva
Vlândia Pinheiro

Publicado em: 26/09/2023

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 (Sequência de caracteres e palavras). 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.

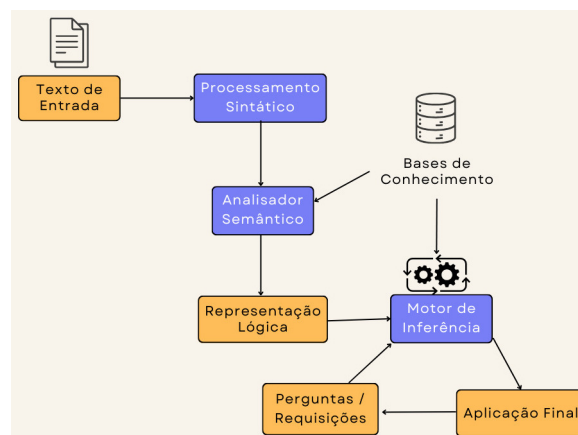
É pertinente fazer uma observação neste ponto para uma definição de base de conhecimento. Uma “base de conhecimento” refere-se a um repositório centralizado, processável por máquina, que contém informações, dados, regras e procedimentos que são usados para capturar, representar e armazenar conhecimento 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 OWL¹), formalismos para troca de dados entre sistemas (e.g. o formato JSON²), 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>



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-

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 (Sequência de caracteres e palavras). 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 NLTK⁴ e spaCy⁵. Dentre as *wordnets*, temos a original e a mais proeminente – a WordNet de Princeton ou PWN⁶ (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/>



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 ConceptNet⁷ (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 9.1⁸, 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, discorreremos 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/>

⁸Esse 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>.



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.

The screenshot shows the 'WordNet Search - 3.1' interface. At the top, there's a search bar with 'murder' entered and a 'Search WordNet' button. Below the search bar, there are 'Display Options' with a dropdown menu set to '(Select option to change)' and a 'Change' button. A key is provided: 'Key: "S:" = Show Synset (semantic) relations, "W:" = Show Word (lexical) relations'. Below this, the 'Display options for sense: (gloss) "an example sentence"' are listed. The results are categorized under 'Noun' and 'Verb'. Under 'Noun', there is one entry: 'S: (n) murder, slaying, execution (unlawful premeditated killing of a human being by a human being)'. Under 'Verb', there are two main entries. The first is 'S: (v) murder, slay, hit, dispatch, bump off, off, polish off, remove (kill intentionally and with premeditation) "The mafia boss ordered his enemies murdered"', which has several sub-entries: 'direct troponym / full troponym' (including 'S: (v) burke' and 'S: (v) execute'), 'direct hypernym / inherited hypernym / sister term' (including 'S: (v) kill'), 'derivationally related form', and 'sentence frame'. The second main entry is 'S: (v) mangle, mutilate, murder (alter so as to make unrecognizable) "The tourists murdered the French language"'.

⁹<http://wordnet.princeton.edu/>

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 – Expressões multipalavras). 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 Linguateca¹¹. O NILC¹² tem uma coleção de recursos listados no portal PortLex¹³, 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-PT¹⁴ (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/>



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 multilínguas (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 him")
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 multilínguas tais como Open Multilingual WordNet (OMW)¹⁵ (Bond; Foster, 2013) ou somente em português? Somente alternativas alinhadas a PWN ou o alinhamento¹⁶ 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/>

¹⁶O 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”.

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 Bases de Conhecimento Semântico. 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 Disambiguation* – WSD). Após o exemplo, são listadas algumas afirmações (especificamente de **hiperonímia**) 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 Berkeley¹⁷, 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/>



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* `Committing_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* `Committing_crime`¹⁸.

Figura 9.4: Descrição do *frame* `Committing_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] Semantic Type: Sentient	The individual that commits a Crime . 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* `Committing_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:

¹⁸A descrição completa do *frame* está disponível em <https://framenet.icsi.berkeley.edu/frameIndex>.

- [*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 ao *frame*. Cada palavra polissêmica¹⁹ 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* *Committing_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-

¹⁹Uma 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.



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 lexicais²⁰.

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 SemLink²¹, 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 FLAME²³, relativos aos domínios de

²⁰Mais 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/>



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-Br²⁴, adaptado do *frame* *Committing_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

²⁴Ver descrição completa em <https://webtool.framenetbr.ufjf.br/index.php/webtool/report/frame/main>



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 Bases de Conhecimento Semântico. 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

ConceptNet²⁵ (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 “*studeny*” 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/>



- Cozinhar é motivada por você está com fome (*Cook is motivated 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, conseqüentemente, 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 ConceptNet²⁶ 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 Wikcionário²⁷, 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;

²⁶Sua última versão é a ConceptNet 5.8 e a documentação completa está disponível em <https://github.com/commonsense/conceptnet5/wiki>

²⁷Disponível em https://pt.wiktionary.org/wiki/Wikcion%C3%A1rio:P%C3%A1gina_principal

²⁸<https://github.com/commonsense/conceptnet5/wiki/Edges>



- *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 naturais³¹ 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 REST³² 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).

²⁹A lista completa das relações expressas na ConceptNet está disponível em <https://github.com/commonsense/conceptnet5/wiki/Relations>

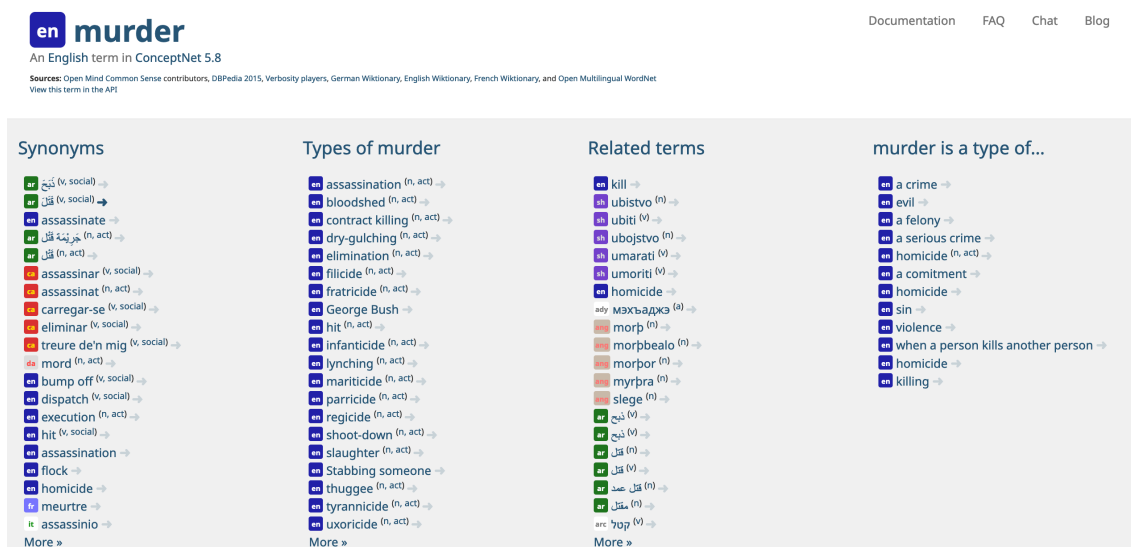
³⁰Pode ser nulo porque nem todas as afirmações foram derivadas de entrada em linguagem natural

³¹As bases de afirmações da ConceptNet para todas as linguagens naturais suportadas pode ser consultada em <https://github.com/commonsense/conceptnet5/wiki/Languages>

³²Acessível em api.conceptnet.io. A documentação completa da API da Conceptnet pode ser acessada em <https://github.com/commonsense/conceptnet5/wiki/API>



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.;

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;



- 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)

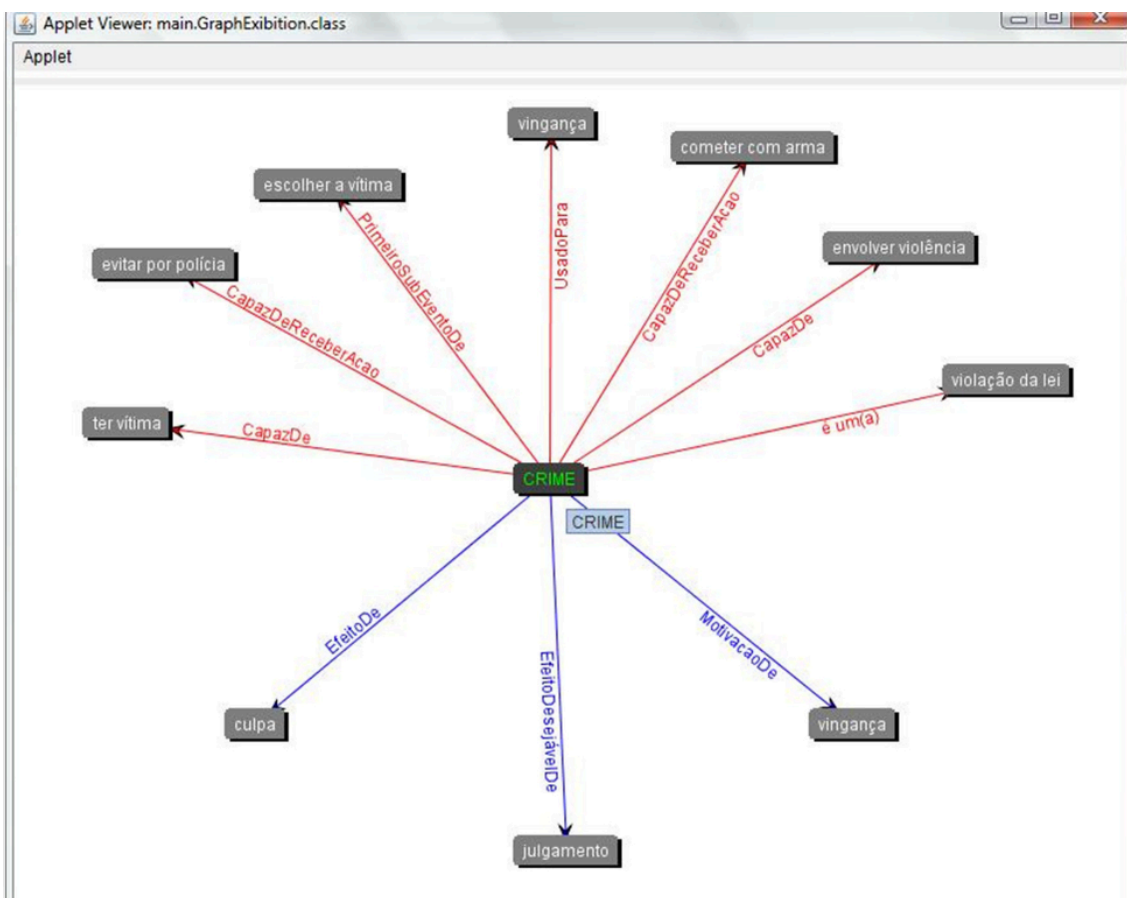
Afirmações (*edges*) do termo *balear* (“*shoot*”, em inglês):

- “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:



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 “baleiar” 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 “baleiar” 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

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.



Capítulo 10

Semântica distribucional

Eloize Rossi Marques Seno

Daniela Claro

Laila Mota

Jessica Rodrigues

Publicado em: 26/09/2023

É relativamente fácil para nós, seres humanos, visualizarmos um texto e, a partir de uma simples leitura, extrairmos dele determinados tipos de informação. Por exemplo, ao ler o texto “Ser feliz sem motivo é a forma mais autêntica de felicidade.” podemos reconhecer o sentido das palavras e o significado do texto formado pela junção de todas as palavras. Diferente dos humanos, os algoritmos computacionais não conseguem processar símbolos/palavras. Ao invés disso, eles requerem uma representação numérica de um documento ou texto a ser processado, para que consigam realizar suas operações.

A semântica distribucional tem sido atualmente a principal abordagem de representação do significado lexical adotada nas mais diversas tarefas do processamento de linguagem natural. Nessa abordagem, os itens lexicais (palavras) são representados por meio de vetores de valores reais, conhecidos por **vetores semânticos**, que codificam o significado das palavras a partir de sua distribuição em textos.

A semântica distribucional 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 seguir¹:

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 semânticos podem ser aprendidos automaticamente a partir de textos, sem que haja supervisão de um humano

¹Sentenças 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”.



(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).

Os MSD são, frequentemente, classificados como **vetores esparsos** e **vetores densos**. Por exemplo, o modelo TF-IDF, amplamente adotado em tarefas que envolvem a comparação de similaridade entre documentos, como a detecção de plágio, a inferência textual e a recuperação de informações, é um exemplo clássico de vetor esparsos. Nesse modelo, o significado de uma palavra é representado por meio de uma função simples calculada com base na frequência da palavra em uma coleção de documentos, conforme será visto. Atribuindo pesos aos termos da matriz termo-documento com TF-IDF. Como muitas palavras nunca ocorrem em alguns documentos, frequentemente, esse modelo leva a vetores muito grandes e esparsos, ou seja, com muitos zeros. Por outro lado, os modelos da família Word2Vec (Mikolov et al., 2013b) são considerados vetores densos (não esparsos), onde as entradas são números reais que representam propriedades semânticas úteis (conforme será abordado na Word2Vec, ao invés de contagens quase zero).

A Figura 10.1 ilustra a estrutura geral dos Modelos Semânticos Distribucionais. Neste capítulo serão abordados apenas os **vetores esparsos** e os **vetores densos estáticos**. Mais especificamente, iniciaremos apresentando os modelos esparsos TF-IDF e PMI (Vetores esparsos) e depois apresentaremos os principais vetores densos estáticos como o Word2Vec, o GloVe e o FastText (Vetores densos estáticos). Os vetores densos dinâmicos, por sua vez, como ELMO, BERT e GPT serão abordados no Modelos de linguagem. Antes, porém, de apresentar os MSD, introduziremos alguns conceitos fundamentais da semântica vetorial e apresentaremos a **similaridade do cosseno** (Semântica Vetorial), uma maneira padrão de usar os vetores semânticos para calcular a similaridade entre palavras, sentenças e documentos, que é uma ferramenta fundamental em aplicações práticas como a sumarização automática e a recuperação de informações.

10.1 Semântica Vetorial

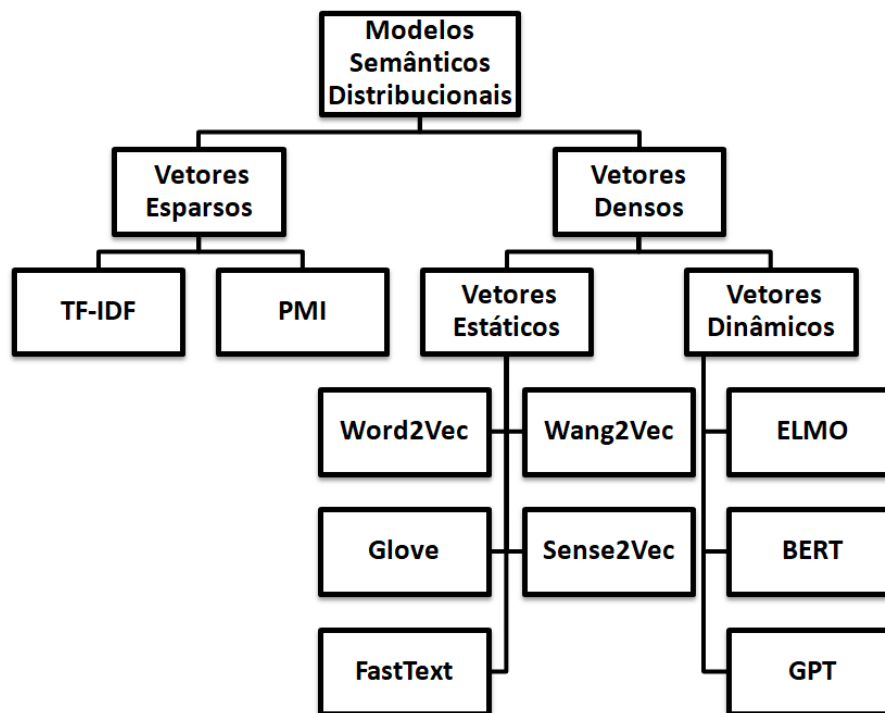
As primeiras investigações no campo da semântica vetorial, também conhecida por **métodos distribucionais**, tiveram início na década de 1950, impulsionadas pela convicção de que o significado de uma palavra pode ser definido a partir da sua distribuição nos contextos linguísticos em que ela ocorre, 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. Portanto, os métodos distribucionais se definem como uma representação vetorial que retrata o significado de uma palavra a partir da distribuição das palavras que formam o seu contexto (Jurafsky; Martin, 2023). Por exemplo, a Figura 10.2 representa o espaço vetorial semântico da palavra “ensino”². Palavras como “educação”, “estudantes”, “professores” e “alunos” são alguns exemplos de palavras que compartilham esse mesmo espaço semântico.

A representação vetorial semântica, ou simplesmente **vetores semânticos**, é um padrão de representação muito usual em PLN, que pode retratar vários aspectos do significado das palavras, como a similaridade (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 claramente relacionados, uma vez que futebol se joga com uma bola), entre outros

²<http://nilc.icmc.usp.br/nilc/index.php/repositorio-de-word-embeddings-do-nilc>



Figura 10.1: Ilustração dos Modelos Semânticos Distribucionais.



aspectos.

A ideia dos vetores semânticos é, então, representar cada palavra como um ponto em um **espaço vetorial multidimensional**, construído a partir da distribuição de suas palavras vizinhas.

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 até documentos completos como vetores em um espaço multidimensional.

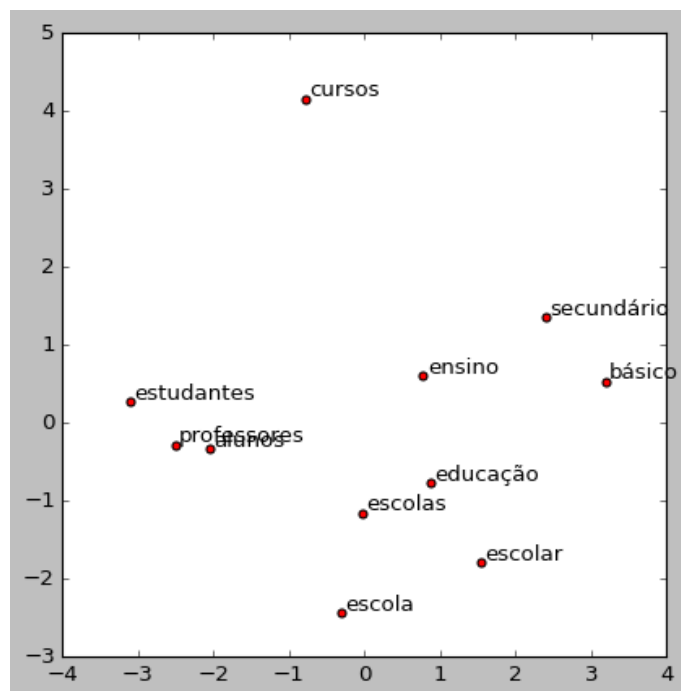
Geralmente, os vetores semânticos são representados por meio de uma **matriz de coocorrência** (ou distribuição de coocorrência), que retrata a frequência de coocorrência das palavras. As representações matriciais mais comuns são a **matriz termo-documento**, onde cada dimensão (vetor) da matriz representa 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 documentos³ 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,

³O termo “documento” é usado nesta seção de forma genérica, podendo se referir a sentenças, parágrafos ou documentos completos.

Figura 10.2: Espaço vetorial semântico da palavra “ensino” (representação gerada com o modelo GloVe (Glove), disponibilizado pelo NILC).



por sua vez, representa a frequência de uma palavra em particular em um documento em particular, ou seja, quantas vezes aquela palavra ocorreu naquele documento.

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, representam pontos em um espaço tridimensional. A matriz foi computada a partir de três textos de divulgação científica extraídos da revista online Pesquisa FAPESP⁴, da seção de Tecnologia, edições 308 e 310 de 2021. Os textos se referem aos seguintes temas: energias renováveis⁵ (coluna 1), risco de escassez de energia fornecida pelas hidrelétricas brasileiras⁶ (coluna 2) e veículos elétricos movidos a etanol⁷ (coluna 3). A matriz da Tabela 10.1 representa apenas um subconjunto das palavras que 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

⁴<https://revistapesquisa.fapesp.br/>

⁵<https://revistapesquisa.fapesp.br/a-forca-das-renovaveis/>

⁶<https://revistapesquisa.fapesp.br/sob-o-risco-da-escassez/>

⁷<https://revistapesquisa.fapesp.br/eletricos-movidos-a-etanol/>



	energias renováveis	escassez de energia	veículos elétricos
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 Reduzindo a dimensionalidade com LSA.

10.1.2 Matriz termo-contexto

Os vetores semânticos também podem ser usados para representar o significado de **palavras** e não apenas de documentos. Para tanto, ao invés de usarmos uma matriz de **termo-documento** (como visto na Matriz termo-documento), 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 ocorrência 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 Matriz termo-documento. 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.

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 do Word2Vec, Glove e Fastext (Vetores densos estáticos), e não aos vetores esparsos como o TF-IDF e o PMI (Vetores esparsos).

10.1.3 Calculando a similaridade entre vetores semânticos

Uma tarefa bastante comum do processamento de linguagem natural consiste em calcular a similaridade entre vetores de documentos ou vetores de palavras, seja para estabelecer uma métrica de semelhança entre dois textos ou para se ter uma medida de equivalência entre duas palavras. Para tanto, faz-se necessário o emprego de alguma medida de similaridade entre vetores.

A **medida do Cosseno**, também conhecida por **distância do Cosseno**, é sem dúvida uma das mais clássicas da área de PLN. Essa medida 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 ortogonais⁸, 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:

$$\langle x, y \rangle = x \cdot y = \|x\| \|y\| \cos \theta$$

$$\cos \theta = \frac{x \cdot y}{\|x\| \|y\|}$$

onde $\langle x, y \rangle$ 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:

$$x \cdot y = \sum_{i=1}^n x_i y_i = x_1 y_1 + x_2 y_2 + x_3 y_3 + \dots + x_n y_n \quad (10.1)$$

$$\|v\| = \sqrt{\sum_{i=1}^n v_i^2} \quad (10.2)$$

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.

⁸Dois vetores são ortogonais se o Produto Escalar entre eles é nulo (zero), conforme veremos na Equação 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 :

$$\text{Cos}(a, b) = \frac{2}{4} = 0,5$$

A similaridade entre os vetores a e b é diretamente proporcional ao cosseno do ângulo entre os dois vetores, 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 principal razão que faz com que essa medida seja tão frequente na Semântica Distribucional e muito usada para calcular a similaridade entre vetores de palavras.

10.2 Vetores esparsos

Vimos nas Seções Matriz termo-documento e Matriz termo-contexto 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) 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 Atribuindo pesos aos termos da matriz termo-documento com TF-IDF e Atribuindo pesos aos termos da matriz termo-contexto com PMI, respectivamente. Em seguida, na Reduzindo a dimensionalidade com LSA, apresentamos o LSA (do inglês, *Latent Semantic Analysis*), um modelo muito adotado em PLN com o objetivo de reduzir a dimensionalidade de um espaço multidimensional criado com o uso do TF-IDF ou PMI.

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.

$$tf(t, d) = \frac{\text{numero\_ocorrencias}(t, d)}{\text{total\_termos}(d)} \quad (10.3)$$

$$idf(t) = \log \left(\frac{N}{\text{total\_documentos}(t)} \right) \quad (10.4)$$

Assim, o TF-IDF de um termo t de um documento d é dado pelo produto entre o valor TF (definido pela Equação 10.3) e o valor de IDF (definido pela Equação 10.4), conforme a Equação 10.5:

$$tf-idf(t, d) = tf(t, d) * idf(t) \quad (10.5)$$

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 são agrupados juntos.
- 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. Dito de outra forma, 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.

$$pmi(x, y) = \log_2 \frac{p(x, y)}{p(x)p(y)} = \log_2 \frac{p(x|y)}{p(x)} = \log_2 \frac{p(y|x)}{p(y)} \quad (10.6)$$

onde:

$$p(t) = \frac{\text{frequencia}(t)}{\text{total\_termos}}$$

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 dois eventos independentes ocorrerem é dada pelo produto das probabilidades dos dois eventos⁹.

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.

⁹Os dois últimos termos da Equação 10.6 são considerados equivalentes pelo Teorema de Bayes.



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:

$$ppmi(x, y) = \max \left(\log_2 \frac{p(x, y)}{p(x)p(y)}, 0 \right) \quad (10.7)$$

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 frequentemente 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

Os modelos TF-IDF e PMI/PPMI apresentados nas Seções Atribuindo pesos aos termos da matriz termo-documento com TF-IDF e Atribuindo pesos aos termos da matriz termo-contexto com PMI frequentemente geram matrizes esparsas contendo muitas células com valor nulo (zero), 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, pode ser aplicada para reduzir a dimensionalidade de um espaço multidimensional. O objetivo principal do 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 LSA pode ser aplicado tanto em matrizes de termo-documento como em matrizes de termo-contexto (ou palavra-palavra). A ideia por trás desse modelo 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, o LSA utiliza 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 T , S e D^T , conforme segue:

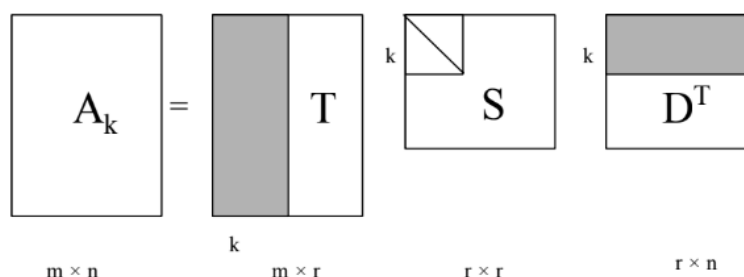
$$A = TSD^T$$

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 10.3).

Na Figura 10.3, 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 ($A(\min(m, n))$).



Figura 10.3: Decomposição em Valores Singulares



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** (Calculando a similaridade entre vetores semânticos) 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 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 o LSA seja uma técnica poderosa para análise de texto e redução de dimensionalidade, o modelo também 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, o modelo também não captura a estrutura sintática da linguagem.

Outra limitação importante é que o 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. O 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. Modelos mais recentes, como os baseados em redes neurais, podem lidar melhor com essas nuances.

Embora o 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, ele não é capaz de capturar nuances semânticas mais complexas, como a ambiguidade.

Em resumo, o 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 Vetores esparsos como representar uma palavra por meio de um vetor esparsos 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 Vetores esparsos. 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 modelo 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.

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 (Word2Vec), ou, ainda, usando representação estatística da matriz de coocorrência de termos, como é o caso do modelo Glove (Glove).

Os vetores de *embeddings* podem ser **estáticos** ou **dinâmicos**. Os *embeddings* **estáticos** permanecem fixos uma vez aprendidos, ou seja, eles não podem ser ajustados ou modificados para uma tarefa específica. Ao contrário desses, os *embeddings* **dinâmicos** podem ser ajustados em tarefas específicas, se adaptando às nuances específicas da tarefa e ao contexto atual. A escolha entre essas abordagens depende das necessidades



da aplicação, do domínio e das características das tarefas em que os *embeddings* serão utilizados.

Nesta seção o foco será apenas na descrição dos ***embeddings* estáticos**, mais especificamente, nos modelos Word2Vec (Word2Vec), Fasttext (Fasttext) e GloVe (Glove). Os ***embeddings* dinâmicos** como os modelos ELMo, BERT e GPT são abordados no 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, 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.



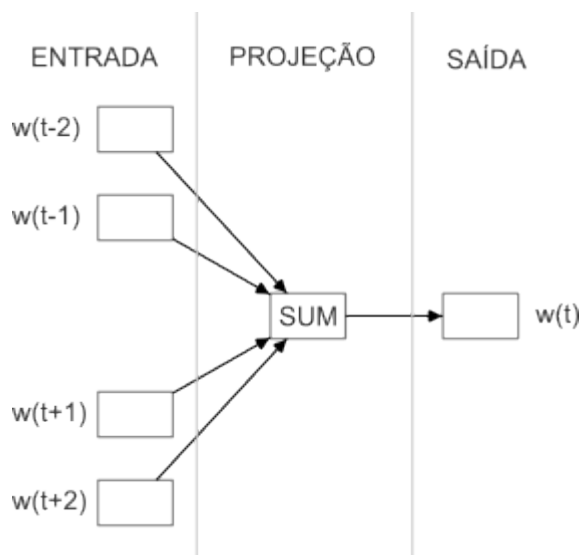
As subseções CBOW e Skip-gram explicam as arquiteturas CBOW e Skip-gram, respectivamente.

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 10.4). 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.

Figura 10.4: Arquitetura CBOW. Prediz a palavra-alvo com base no contexto.



Fonte: Adaptado de (Mikolov et al., 2013b)

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.

$$Q = N \times D + D \times \log_2(V) \quad (10.8)$$



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.

$$\text{softmax}(v_i) = \frac{\exp(v_i)}{\sum_{j=1}^{|V|} \exp(v_j)} \quad (10.9)$$

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 10.5. 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.

$$Q = C \times (D + D \times \log_2(V)) \quad (10.10)$$

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 (Calculando a similaridade entre vetores semânticos).

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).

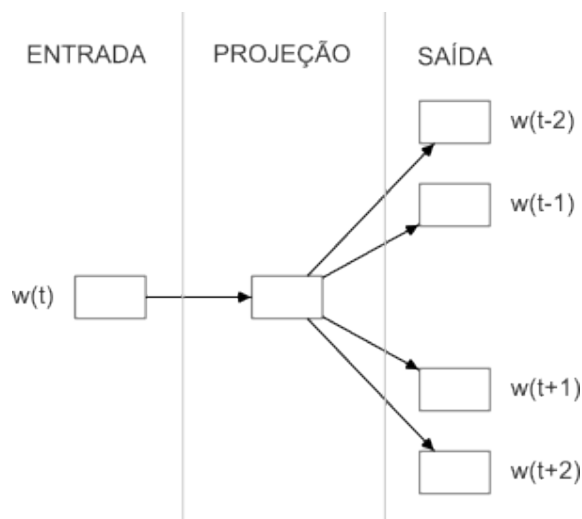
10.3.1.3 Treinamento do Word2Vec

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



Figura 10.5: Arquitetura Skip-gram. Prediz o contexto com base na palavra-alvo.



Fonte: Adaptado de (Mikolov et al., 2013b)

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 semânticas 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 morfológicamente 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

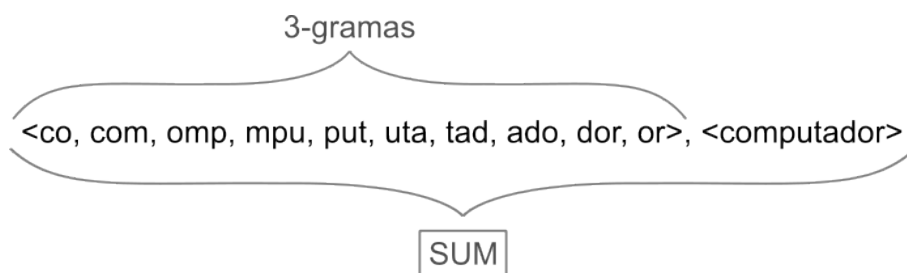
O Fasttext é uma extensão do modelo Word2Vec criada pelo grupo *Facebook AI Research* (FAIR), que amplia o conceito do Skip-gram, no qual cada palavra é representada como uma combinação de n -gramas de caracteres. O modelo leva em consideração não apenas as palavras individuais, mas também as subpalavras (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 10.6 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.

Figura 10.6: Exemplo de representação de trigramas adotados pelo modelo FastText.



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.

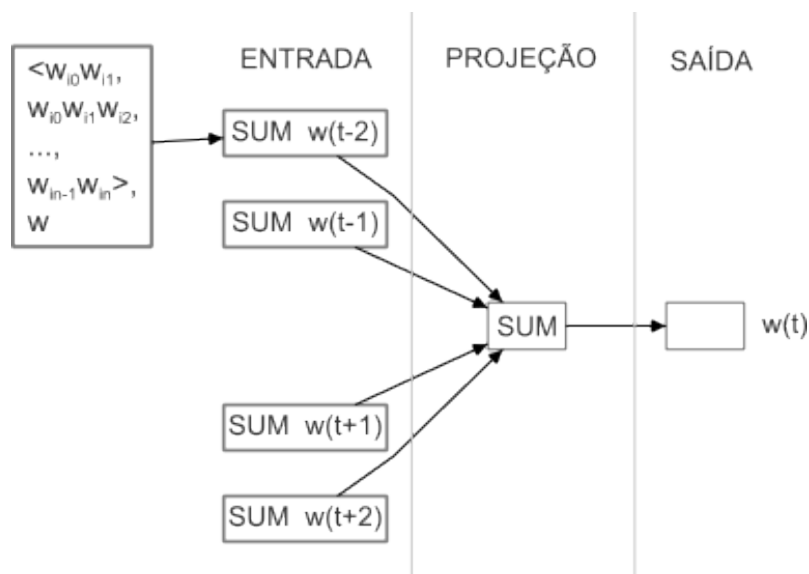
Em termos arquiteturais, o FastText também possui duas arquiteturas assim como o Word2Vec: CBOW – *Continuous Bag-of-Words* e Skip-gram.

A Figura 10.7 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

Figura 10.7: Arquitetura FastText-CBOW.



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.

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 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 de-

pendentes 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).

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:

$$F(w_i, w_j, \tilde{w}_k) = \frac{P_{ik}}{P_{jk}} \quad (10.11)$$

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

$$F(w_i - w_j, \tilde{w}_k) = \frac{P_{ik}}{P_{jk}}. \quad (10.12)$$



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:

$$J = \sum_{i,j=1}^V f(X_{ij})(w_i^T \tilde{w}_j + b_i + \tilde{b}_j - \log X_{ij})^2 \quad (10.13)$$

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 $x \rightarrow 0$ rápido o suficiente dado que o $\lim_{x \rightarrow 0} f(x) \log^2 x$ 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 $f(x)$:

$$f(x) = \begin{cases} (x/x_{max})^\alpha & \text{if } x < x_{max} \\ 1 & \text{c.c.} \end{cases} \quad (10.14)$$

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/PMML, 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* apresentem algumas vantagens em relação aos modelos 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 modelos estáticos, ou seja, que não são atualizados durante o treinamento da tarefa na qual são empregados, eles podem não se adaptar bem a tarefas específicas, especialmente aquelas que exigem informações contextuais específicas da tarefa como é o caso da análise de sentimentos, resolução de ambiguidades e 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* dinâmicos têm sido estudadas na literatura recentemente. Ao contrário dos vetores estáticos, os vetores dinâmicos 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* dinâmicos são o assunto do Modelos de linguagem.



Parte VI

Discurso



Capítulo 11

Modelos discursivos

Paula Christina Figueira Cardoso
Jackson Wilke da Cruz Souza
Roana Rodrigues

Publicado em: 26/09/2023

11.1 Introdução

No Dicionário Houaiss¹, 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

¹Dicionário Houaiss, disponível em: <https://houaiss.uol.com.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 coesão²: 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 *decausa* 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 Modelos de relações discursivas, 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 O modelo RST (Mann e Thompson, 1988) e na CST: Cross-document Structure Theory, 2000, 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*

²Os conceitos de coesão são definidos e discutidos no Resolução de correferência.



Structure Theory (CST). Na Recursos e aplicações para o português brasileiro, apresentamos os principais recursos disponíveis e aplicações em PLN que utilizaram modelos discursivos para sua constituição e/ou realização. Em Considerações Finais (Considerações finais), 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.



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)).

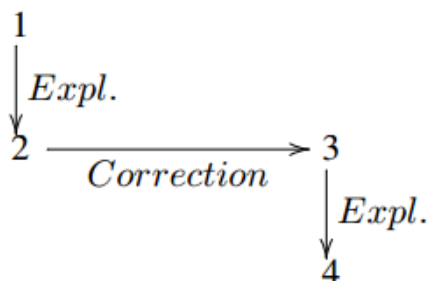
Esta teoria foi bastante explorada para modelar diálogos, pois permite representar contra-argumentação, um fenômeno pouco tratado em outros modelos discursivos (Afan-tenos; Asher, 2014; Asher et al., 2016; Badene et al., 2019; Li et al., 2020). Portanto, a teoria SDRT possui mecanismos que podem ser 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-**

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) 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 RST³. 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 da **fábrica** à nova filosofia.

³A ser detalhada na próxima seção.



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, Interpretation, 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

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*

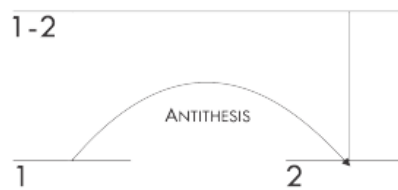
Contrast	Restrições sobre Ns: Não mais do que dois Ns; as situações nos Ns são (a) compreendidas como similares em vários aspectos, (b) compreendidas como diferentes em vários aspectos e (c) comparadas em relação a uma ou mais dessas diferenças Efeito no receptor: O leitor reconhece as similaridades 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.

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.



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.

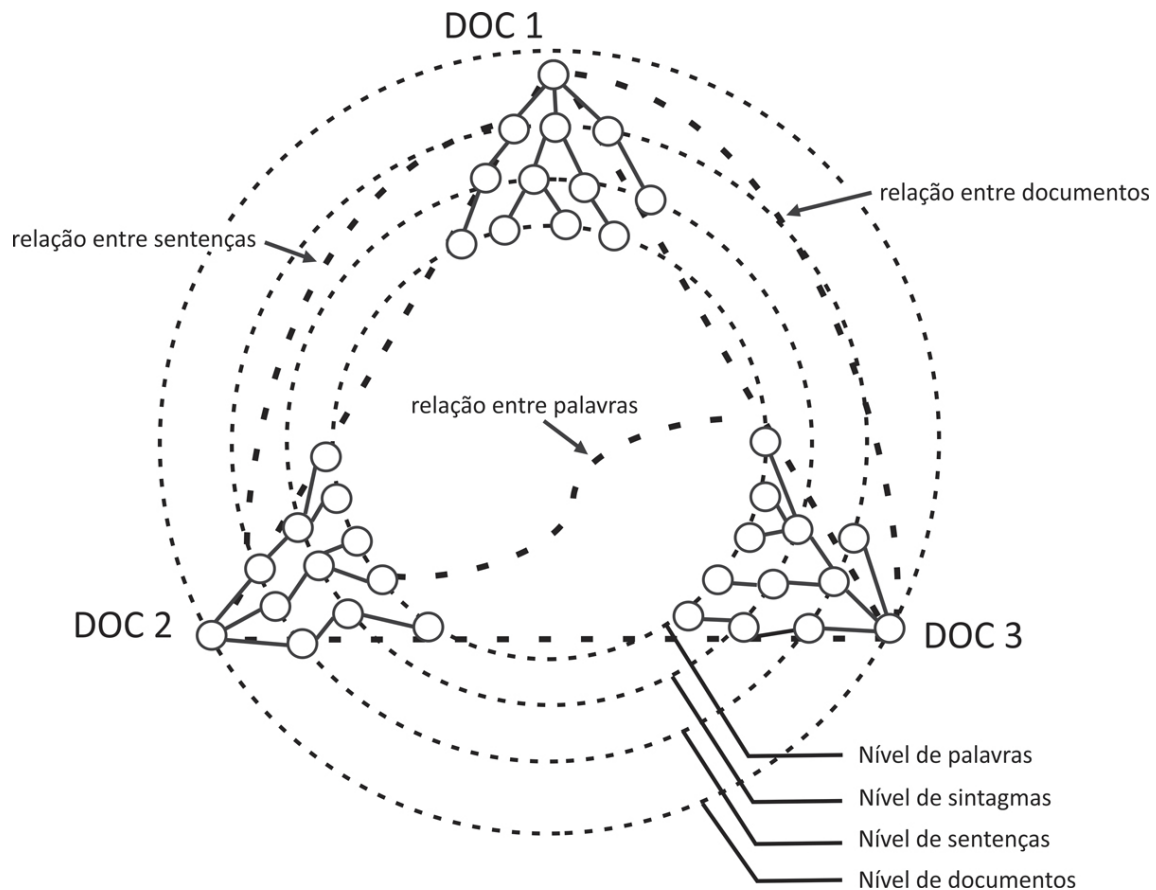
De acordo com Radev (2000), as relações CST podem ocorrer entre diferentes unidades informativas, tais como, palavras, sintagmas, sentenças, parágrafos e documentos, formando um grafo, como ilustrado na Figura 11.4.

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:



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.

Quadro 11.4. Relações CST

Attribution	Identity
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-it⁴ (Collovini et al., 2007; Fonseca et al., 2016a) e CSTNews⁵ (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>



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* DiZer⁶ (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 CSTTool⁷ (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 CSTParser⁸ (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 discusso 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, conseqüentemente, 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/>



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 infundáveis 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 (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, conseqüentemente, sem os modelos discursivos aqui apresentados?

Capítulo 12

Resolução de correferência

Evandro Fonseca
Aline Aver Vanin
Renata Vieira

Publicado em: 26/09/2023

 <https://brasileiraspln.com/livro-pln/4a-edicao/volume1/>

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 predicções, 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



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 ...

¹O Reconhecimento de Entidades Nomeadas no Domínio Legal apresenta técnicas para o tratamento de entidades nomeadas no domínio do Direito.



12.2.1.2 Sintagmas Nominais

São unidades formadas por uma ou mais palavras que, juntas, desempenham uma função sintática específica na frase (Sequência de caracteres e palavras). 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.



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



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”:



- 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 apostro, 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 apostro 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



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,



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 CoNLL² (*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/>



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 : $P_n = \{(m_q, m_p), (m_q, m_o) \text{ e } (m_q, m_n)\}$ (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.

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.

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

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

I_n	<i>Ranking</i>
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).



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 <i>threshold</i>
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 (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*.



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 Tep2.0³ (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 (Sarmiento 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 discurso⁴. 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>

⁴Considera-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.



Figura 12.1: Algoritmo de agrupamento proposto por Fonseca (2018)

```

1: enquanto (tamanho de  $M > 0$ ) faça
2:   int  $j \leftarrow 0$ ;
3:   int[ ]  $S$ ;
4:   para cada  $i \in C$  faça
5:     se  $M_0$  tem relação com  $C_i$  então
6:        $S_j \leftarrow C_i$ 
7:        $j \leftarrow j + +$ 
8:     fim se
9:   fim para
10:  se  $j > 0$  então
11:    int  $k \leftarrow \text{CritérioDeAgrupamento}(M_0, S, C)$ 
12:     $C_k \leftarrow M_0$ 
13:  senão
14:     $C \leftarrow \text{criaNovaCadeia}(M_0)$ 
15:  fim se
16:   $M \leftarrow \text{Remove}(M, 0)$ 
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 realizados⁶. 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

⁵Notemos que, para cada iteração, M_0 muda.

⁶Veja mais sobre Avaliação de sistemas de PLN no Avaliação de tecnologias de linguagem.

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:

$$Abrangência = \frac{\sum_{i=1}^{N_k} \sum_{j=1}^{N_k} \frac{\|K_i \cap R_j\|^2}{K_i}}{\sum_{i=1}^{N_k} K_i}$$



$$Precisão = \frac{\sum_{i=1}^{N_k} \sum_{j=1}^{N_k} \frac{\|K_i \cap R_j\|^2}{R_j}}{\sum_{i=1}^{N_k} R_j}$$

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:

$$Abrangência = \frac{1}{12} * [\frac{5}{5} + \frac{5}{5} + \frac{5}{5} + \frac{5}{5} + \frac{5}{5} + \frac{2}{2} + \frac{2}{2} + \frac{5}{7} + \frac{5}{7} + \frac{5}{7} + \frac{5}{7} + \frac{5}{7}] = 1 \text{ (100\%)}$$

$$Precisão = \frac{1}{12} * [\frac{5}{5} + \frac{5}{5} + \frac{5}{5} + \frac{5}{5} + \frac{5}{5} + \frac{2}{7} + \frac{2}{7} + \frac{5}{7} + \frac{5}{7} + \frac{5}{7} + \frac{5}{7} + \frac{5}{7}] = \frac{16}{21} = 0,76 \text{ (76\%)}$$

12.4.1.3 CEAF:

Baseada no alinhamento de menções e entidades, possui duas variações: $CEAF_m (\Phi_3)$ e $CEAF_e (\Phi_4)$.

$$\Phi_3(K, R) = \|K \cap R\|$$

$$\Phi_4(K, R) = \frac{2\|K \cap R\|}{\|K\| + \|R\|}$$

$$Abrangência = \frac{\Phi_x}{\sum_{i=1} \|K_i\|}$$

$$Precisão = \frac{\Phi_x}{\sum_{i=1} \|R_i\|}$$

Por exemplo, dadas as cadeias de referência:

- $C_{K1} = \{A, B, C, D, E\};$
- $C_{K2} = \{F, G\}.$

E as cadeias preditas pelo modelo:

- $C_{R1} = \{A, B, C, D, E\}.$



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:

$$\text{CEAF}_m: \text{Abrangência} = \frac{5}{7} = 0,71 \text{ (71\%)} \text{ e } \text{Precisão} = \frac{5}{5} = 1 \text{ (100\%)}$$

Para CEAF_e , dado que $\Phi_4 = \frac{2 * 5}{5 + 5} = 1$, temos:

$$\text{Abrangência} = \frac{1}{2} = 0,5 \text{ (50\%)} \text{ e } \text{Precisão} = \frac{1}{1} = 1 \text{ (100\%)}$$

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; 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.

$$\text{Abrangência}_C = \frac{\|C_k \cap C_r\|}{C_k}$$

$$\text{Precisão}_C = \frac{\|C_k \cap C_r\|}{C_r}$$

$$\text{Abrangência}_N = \frac{\|N_k \cap N_r\|}{N_k}$$

$$\text{Precisão}_N = \frac{\|N_k \cap N_r\|}{N_r}$$

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:

$$\text{BLANC}_{\text{Precisao}} = \frac{\text{Precisão}_C + \text{Precisão}_N}{2}$$

$$\text{BLANC}_{\text{Abrangencia}} = \frac{\text{Abrangência}_C + \text{Abrangência}_N}{2}$$

Por exemplo: dados os seguintes links de correferência:

- $C_{K1} = \{A-B, B-C, C-D, D-E\}$;
- $C_{K2} = \{F-G\}$.



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, B³ 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 (Resolução de correferência) 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). 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 vegetaissaúdáveis) 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...



Parte VII

Geração e Interação



Capítulo 13

Geração de linguagem natural

Thiago Castro Ferreira
Ivandr  Paraboni

Publicado em: 20/11/2024

13.1 Introdu  o

Em um dos seus contos mais famosos, o escritor Argentino Jorge Luis Borges introduz a Biblioteca de Babel. Arquitetada de forma hexagonal¹, a biblioteca descrita armazena todos os textos gerados a partir da combina  o de 23 caracteres do alfabeto romano: A-Z, espa  o, v rgula e ponto, como na Figura 13.1.

Figura 13.1: Vers o do primeiro par grafo deste cap tulo, encontrado no projeto “The Library of Babel”.

Page 46 of 410

2hodvv8rkmt00
y27i3v2bsm9zs
9wevvguqvvr0
...-w4-s1-v03

em um dos seus contos mais famosos, o escritor argentino jorge luis borges intro
duz a biblioteca de babel. arquitetada de forma hexagonal, a biblioteca descrita
armazena todos os textos gerados a partir da combinacao de vinte e tres caract
eres do alfabeto romano az, espaco, virgula e ponto, como na figura um.

Single Page

Anglshize

Bookmarkable

Download

Back to Portal

O conto sugere o cen rio (dist pico?) onde todos os textos j  escritos, assim como todos que ainda ser o escritos, como jornais, livros, este pr prio texto e o conto do autor, est o presentes na biblioteca universal. Apesar de alertar que esta pode ser a quest o 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 atrav s de um algoritmo simples (mas pouco elegante e eficiente) como o representado no algoritmo da Figura 13.2², capaz de gerar (em um tempo infinito) tal biblioteca com livros de at  50 mil caracteres.

¹A figura geom trica sem curvas mais pr xima de um c rculo.

²Pseudoc digo gerado por GPT-4 Turbo e traduzido e revisado pelos autores.



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:    $caracteres \leftarrow ['a', 'b', \dots, 'z', ',', '.', ' ']$   $\triangleright$  lista com os 23 caracteres
3:    $total\_livros \leftarrow \infty$   $\triangleright$  Definir número de livros como infinito
4:    $livro\_tamanho \leftarrow 50,000$   $\triangleright$  Definir tamanho de livros como 50,000
   caracteres
5:   for  $livro \leftarrow 1$  to  $total\_livros$  do
6:      $texto \leftarrow$  string vazia
7:     for  $i \leftarrow 1$  to  $livro\_tamanho$  do
8:        $char\_indice \leftarrow \text{RANDOM}(1, |caracteres|)$   $\triangleright$  Definindo um caractere
       aleatório
9:        $texto \leftarrow texto + caracteres[char\_indice]$ 
10:    end for
11:    PRINT(texto)  $\triangleright$  Imprimindo o livro gerado
12:  end for
13: end procedure

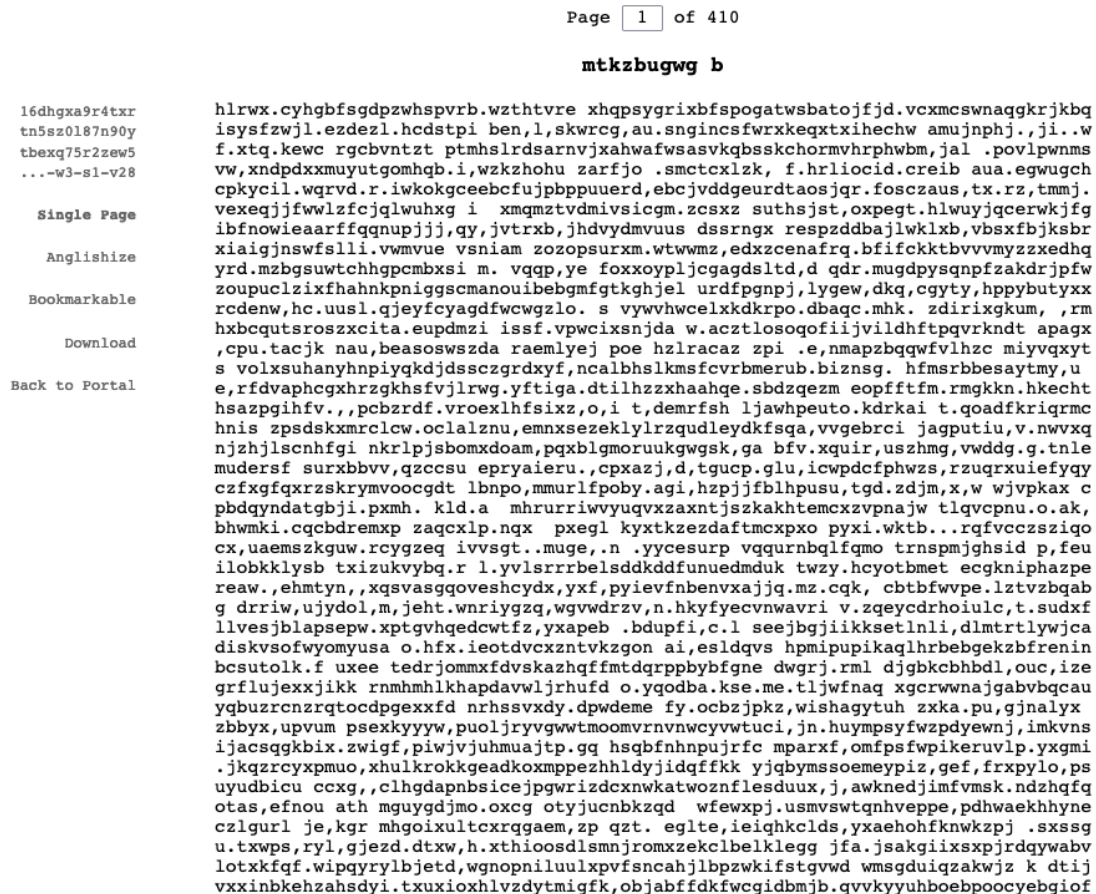
```

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>



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. Yngve⁴, o qual apresenta o primeiro programa de computador para produção de sentenças gramaticais em inglês⁵. O objetivo do autor era avaliar a adequação da gramática, gerando sentenças sem necessariamente ter um objetivo comunicativo ou significado.

⁴Professor de Linguística da Universidade de Chicago e do Instituto de Tecnologia de Massachusetss (MIT).

⁵As sentenças foram automaticamente geradas a partir do livro infantil chamado Engineer Small and the Little Train.



Como expresseo 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 expresseo 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 expresseo. 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 expresseo (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,



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 (*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



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.

- Robotox⁶ 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.
- colaborabot⁸ é um sistema de linguagem natural criado pelo grupo Colaboradados⁹ 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.
- Rui¹⁰ 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.
- Rosie¹², 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 13.4¹³ 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>

⁷Parceria entre a Agência Pública e *Repórter Brasil*.

⁸https://x.com/colabora_bot

⁹<https://colaboradados.github.io>

¹⁰<https://x.com/ruiarbot>

¹¹<https://x.com/elasnocongresso>

¹²<https://x.com/RosieDaSerenata>

¹³Extraída de <https://x.com/RosieDaSerenata>.



Figura 13.4: Publicações geradas pelo sistema Rosie.



O *template* consiste de três *tags*, TÍTULO, 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

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:

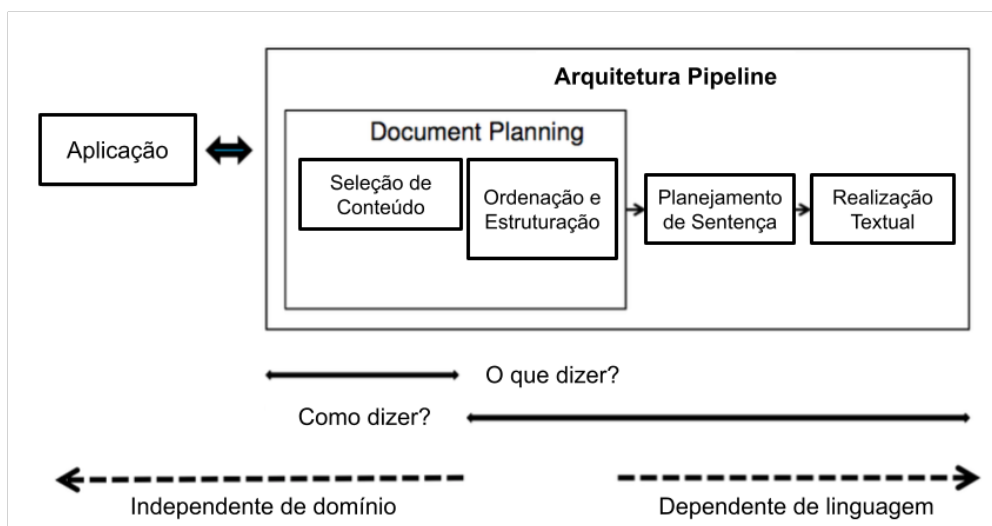
Gasto suspeito de **Dep. Antonio (PF)**. Você pode me ajudar a verificar? O parlamentar possui **1** alertas no mês.

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.

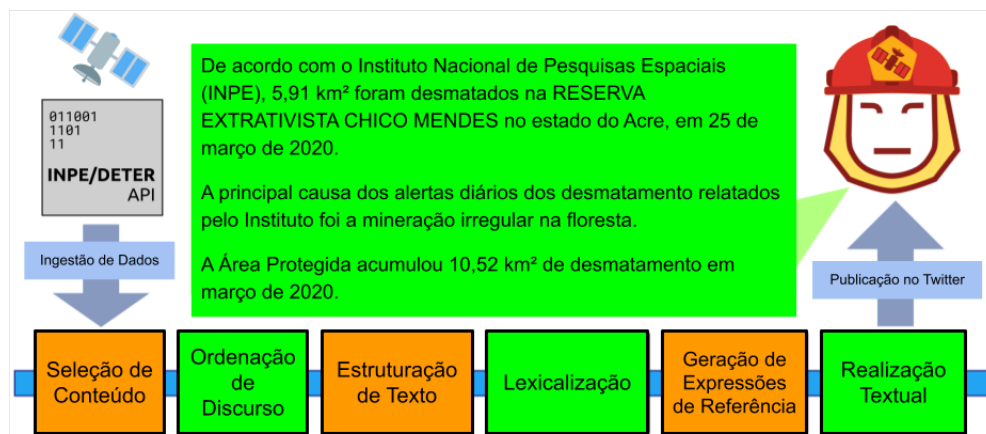
Figura 13.5: Arquitetura *pipeline* de Reiter; Dale (2000).



Fonte: (Santhanam; Shaikh, 2019)

O Planejamento do Documento contempla a seleção do conteúdo e sua estruturação na forma 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 Discurso 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.

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:

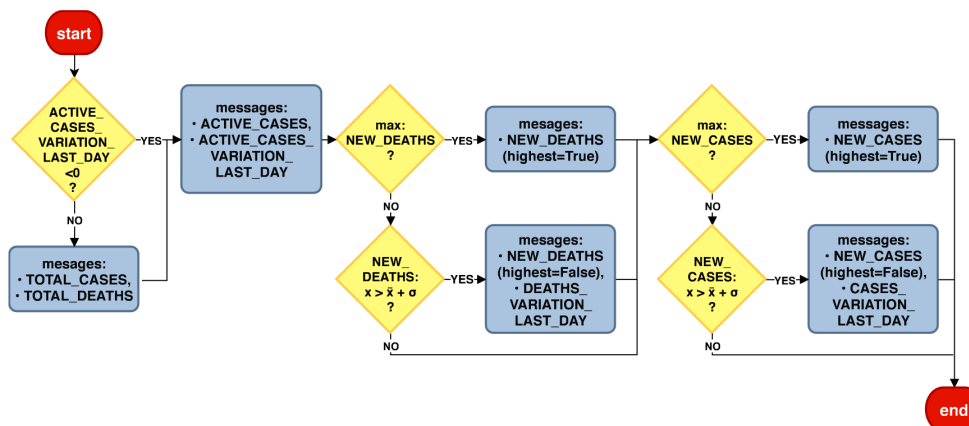
- **ACTIVE_CASES**: número de casos ativos de COVID-19 no Brasil no dia atual;
- **TOTAL_CASES**: número total de casos de COVID-19 no Brasil;
- **NEW_CASES**: número de novos casos de COVID-19 no Brasil no dia atual;
- **ACTIVE_CASES_VARIATION_LAST_DAY**: variação de casos ativos de COVID-19 no Brasil 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



(`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



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



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
```



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).



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

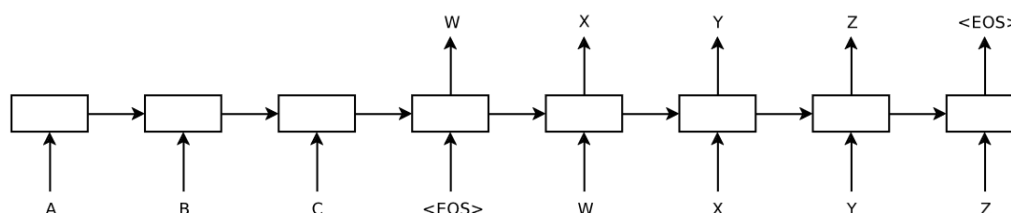


popularizou-se a partir do avanço dos modelos de aprendizado profundo, principal ferramenta utilizada na geração de linguagem natural ponta-a-ponta. Estas arquiteturas são discutidas a seguir, fazendo uso de alguns conceitos que também são discutidos na Modelos de Linguagem Neurais Modernos, como o uso 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.

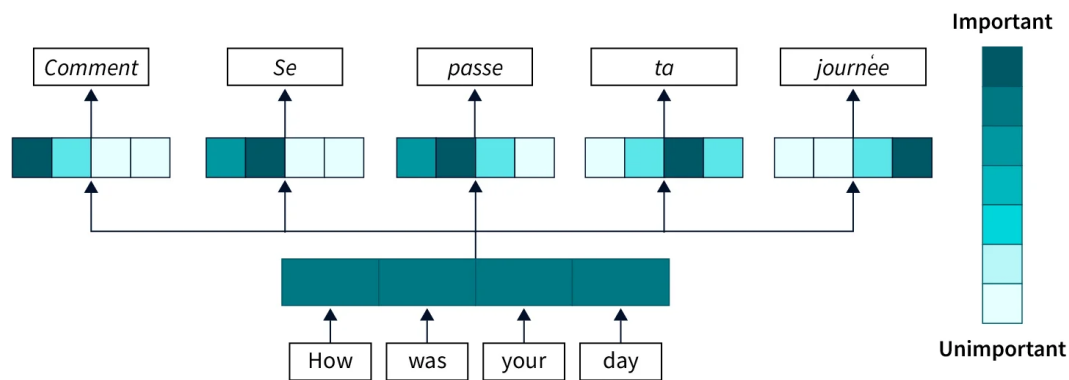


Fonte: (Sutskever et al., 2014)

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.

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 antecessora¹⁴.

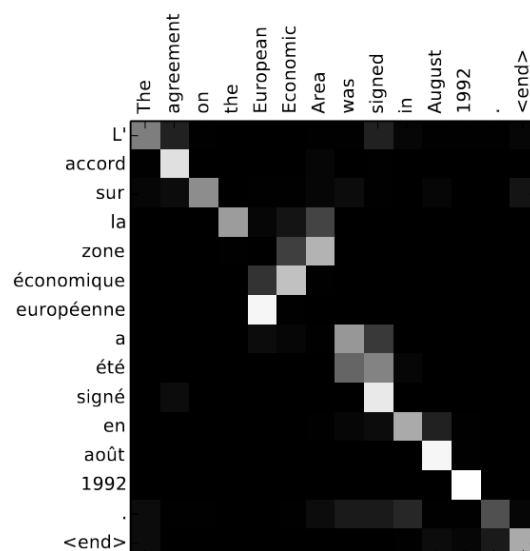
13.4.1 Aplicações

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),

¹⁴Apesar 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.

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

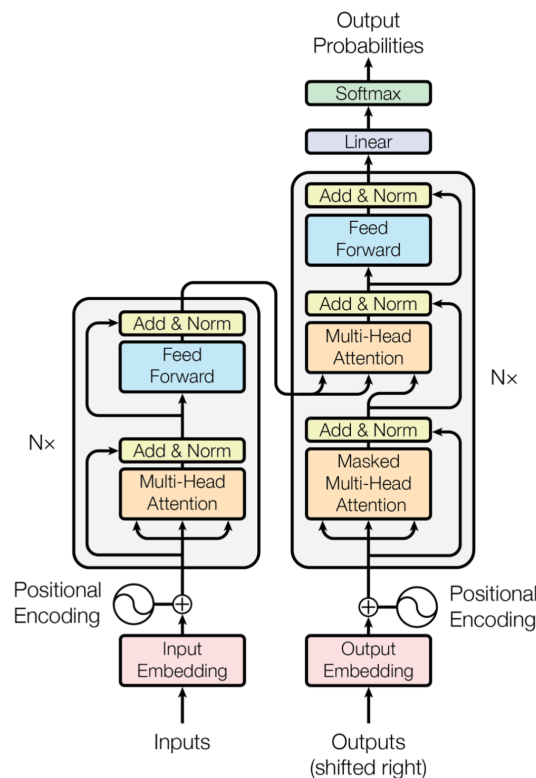
Como exemplificado na Exemplo, o sistema CoronaReporter (Campos et al., 2020) utiliza uma arquitetura *pipeline* com múltiplas representações intermediárias entre os números de COVID-19 no Brasil e o relatório textual final sobre o assunto. Se este mesmo robô-jornalista fosse construído de acordo com 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 Exemplo 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)
```



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

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 Exemplo – sistema CoronaReporter. 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 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 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 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 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



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.

SYSTEM

Você receberá um resumo dos casos de COVID-19 no Brasil, apresentado de maneira estruturada. Sua tarefa é verbalizar o conteúdo dessas informações de forma clara, fluente e precisa em Português Brasileiro, sem acrescentar ou omitir nenhum dado.

USER

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)

ASSISTANT

Nas últimas 24 horas, houve uma alta de 6% na variação de óbitos por COVID-19 no Brasil. O número de novos casos confirmados foi de 15.305, o maior registrado até agora. No mesmo período, foram contabilizadas 824 novas mortes, mas este não foi o maior número de novas mortes registrado até hoje. No total, o Brasil acumula 218.223 casos confirmados de COVID-19 e 14.817 óbitos desde o início da pandemia.

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.

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.

USER

Quantas medalhas o Brasil ganhou nas Olimpíadas de Paris em 2024?

ASSISTANT

Não tenho acesso a eventos ou informações atuais, pois minha última atualização foi em outubro de 2023. Portanto, não tenho como fornecer o número de medalhas que o Brasil ganhou nas Olimpíadas de Paris em 2024. Sugiro consultar fontes atualizadas, como notícias online, para obter essa informação.



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 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 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, conseqüentemente, 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.



Capítulo 14

Perguntas e Respostas

Eduardo G. Cortes

Renata Vieira

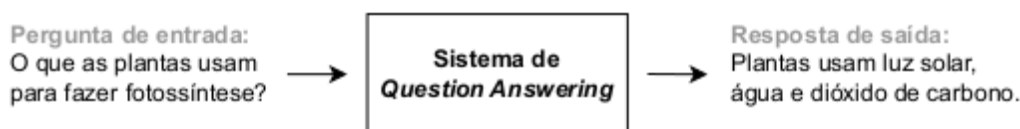
Dante A. C. Barone

Publicado em: 13/03/2024

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 (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.



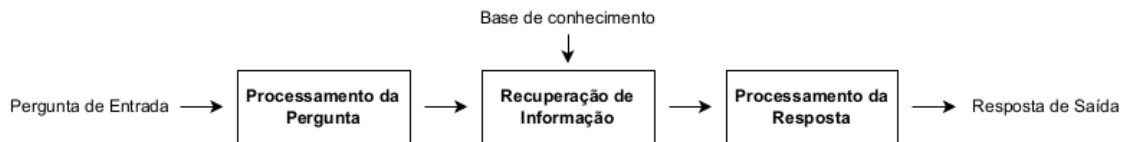
Neste capítulo, abordamos a área de PR no contexto de sistemas especialistas focados em entregar informações diretas sobre o que foi perguntado. Por outro lado, existem os sistemas *chatbots*, conhecidos também como agentes de conversação, que visam manter um diálogo contínuo e engajante com o usuário/a. 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 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



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 artificiais (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).

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.



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	ABREVIÇÃO
Símbolo	Grupo	Exp
Técnica	Indivíduo	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

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.



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 (Conjunto de dados, *dataset* e *corpus*). 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, extraíndo 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



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  $6\text{CO}_2 + 6\text{H}_2\text{O} + \text{luz solar} \rightarrow \text{C}_6\text{H}_{12}\text{O}_6 + 6\text{O}_2$ .
</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';
```

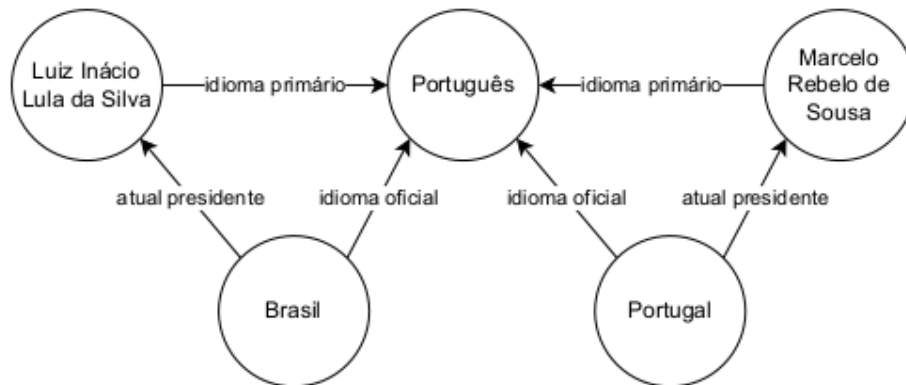
Esta consulta instrui o banco de dados a selecionar o valor da coluna “especialização” na tabela “medicos” onde a coluna “nome” é igual a “Maria da Silva”. A execução dessa consulta resultaria na resposta desejada pelo sistema.

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



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 (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

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



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 Classificação de Sistemas de Perguntas e Respostas.

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-



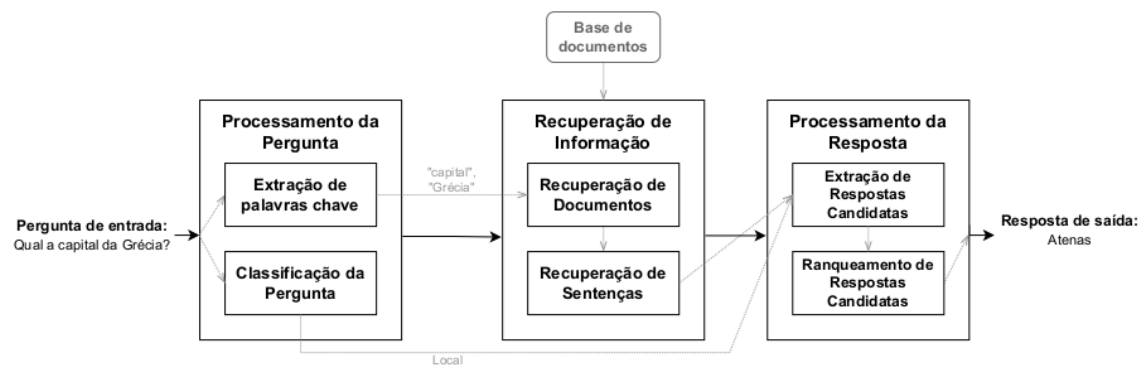
tagens das abordagens modular e *end-to-end*, considerando suas implicações em termos de flexibilidade, complexidade de implementação e capacidade de atender às necessidades específicas do projeto.

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

palavras mais significativas. Por exemplo, na pergunta “Qual a capital da Grécia?”, podemos dividi-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



Figura 14.7: Exemplo ilustrativo de sentenças com suas entidades identificadas, onde somente as sentenças com entidades do tipo “Pessoa” são selecionadas.

X A Revolução Francesa começou em 1789.	Nome
X A Cruz Vermelha presta assistência humanitária em todo o mundo.	Local
✓ Lucas e Lara foram ao cinema assistir ao novo filme de ação.	Data
X Paris é conhecida como a cidade do amor.	Organização
X A ONU foi fundada após a segunda guerra mundial.	
✓ Maria está planejando sua viagem para a França no próximo mês.	
X O Monte Everest é a montanha mais alta da terra.	
✓ O professor Carlos explicou a teoria de maneira muito clara.	
X A Praça da Liberdade fica no centro de Belo Horizonte.	
X O homem pisou na lua pela primeira vez em 20 de julho de 1969.	

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 (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.

```
# 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 (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



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

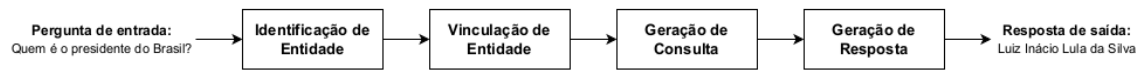
Nesta abordagem, consideramos um sistema de PR que utiliza grafos de conhecimento como fonte principal para recuperar respostas de perguntas factuais sobre um domínio restrito. Grafos de conhecimento são estruturas que armazenam informações de forma semântica, representando entidades como nós e relações entre elas como arestas, como mostrado no exemplo da Figura 14.5, apresentada na Dados estruturados. 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



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-

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.

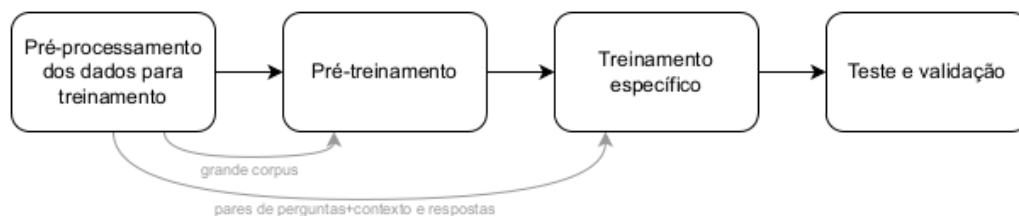
14.3.3 Abordagem *end-to-end*

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 Abordagem modular com documentos.

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.

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 *Bilingual Evaluation Under-study* (BLEU)) e BERTScore, são comumente utilizadas para avaliar a qualidade



das respostas do modelo. Os métodos de avaliação serão vistos em mais detalhes na Métodos de avaliação.

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 Avaliação de tecnologias de linguagem 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 saída do sistema com respostas de referências 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%.

$$Precisão = \frac{\text{Número de Respostas Corretas}}{\text{Número Total de Respostas Fornecidas}} \quad (14.1)$$

- **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%.

$$Revocação = \frac{\text{Número de Respostas Corretas}}{\text{Número Total de Respostas Corretas Possíveis}} \quad (14.2)$$

- **F1-Score:** considera a precisão e a revocação juntas, através de sua média harmônica.

$$F1-Score = 2 \times \frac{\text{Precisão} \times \text{Revocação}}{\text{Precisão} + \text{Revocação}} \quad (14.3)$$

- **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%.

$$Acurácia = \frac{\text{Número de Respostas Corretas}}{\text{Número Total de Respostas}} \quad (14.4)$$

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).



- **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 prever 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



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 uma 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.



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.



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.



Capítulo 15

Diálogo e Interatividade

Brielen Madureira

Publicado em: 13/03/2024

 <https://brasileiraspln.com/livro-pln/4a-edicao/volume1/>

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 dialogando¹. 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

¹Este 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>.



(“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,



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 Conceitos: O que forma um diálogo? e Fenômenos: O que ocorre em uma conversa?), bem como de tipos de modelos e tarefas (Modelos e Tarefas: Implementando um processo complexo), 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 (Avaliação: Verificando a qualidade de um sistema). 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 (Recursos: Trabalhando com dados e sistemas interativos). 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”,

²Em 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.



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 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). Começamos, 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

³Fonte: Dicionário de Oxford. Disponível em https://www.oed.com/dictionary/dialogue_n?tab=etymology#6924392.

⁴Usaremos “fala” de modo genérico para a produção linguística, englobando também língua de sinais.

⁵Não 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).



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 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.

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-dizer ou propor uma continuação antes do enunciado acabar, a ouvinte precisou já estar processando o enunciado 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 aeo ah / as pessoas querem saber de você e querem-s assim / querem saber e 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 //

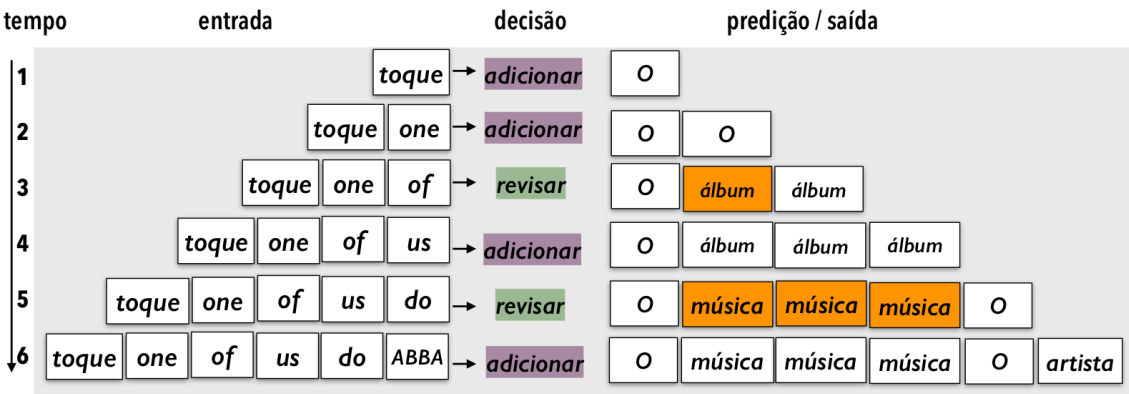
Fonte: CRPC, arquivo Surpresas_da_Fotografia

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



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)

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.

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.

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 / **ch** / eu fiquei ~~em-eia~~ hospedada / numa outra cidade vizinha / se não me engano / ai 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-ge

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

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 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 Texto ou fala?, 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



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

exemplos exibidos na Figura 15.6:

- **pausa preenchida:** sons como “hmm”, “ééé”, 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;
- **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



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 / **ch** / que eles falam / a analogia / que / por exemplo / se tem uma lei que / **ch** / ~~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 / **ch** / eu acho que isso você acaba / **ch** / pegando com o tempo / não é //

Fonte: CRPC, arquivo O_Mundo_do_Direito

(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.

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

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 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)).

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 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.

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 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 instrumento⁶, 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 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

⁶Também conhecida pela máxima “se você tem um martelo, tudo se parece um prego”. Fonte: https://en.wikipedia.org/wiki/Law_of_the_instrument.



(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), 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).

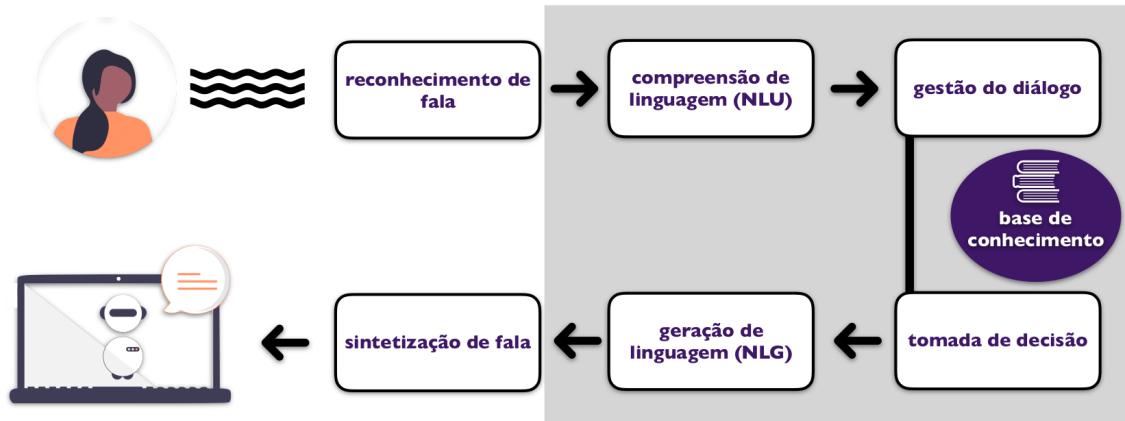
Agora, **o que um sistema de diálogo precisa ter?** Vamos retomar os elementos básicos do formato elementar de uma conversa, expostos na Conceitos: O que forma um diálogo?. Momento a momento, um participante precisa (i) captar o que está sendo dito, (ii) compreendê-lo, (iii) atualizar o estado da situação 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.

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

Atualmente, na era do aprendizado de máquina, a modularização tem sido substituída por tentativas de propor 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 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 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 pode-

mos 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, extraíndo 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 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.
- **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 (Perguntas e Respostas), 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.

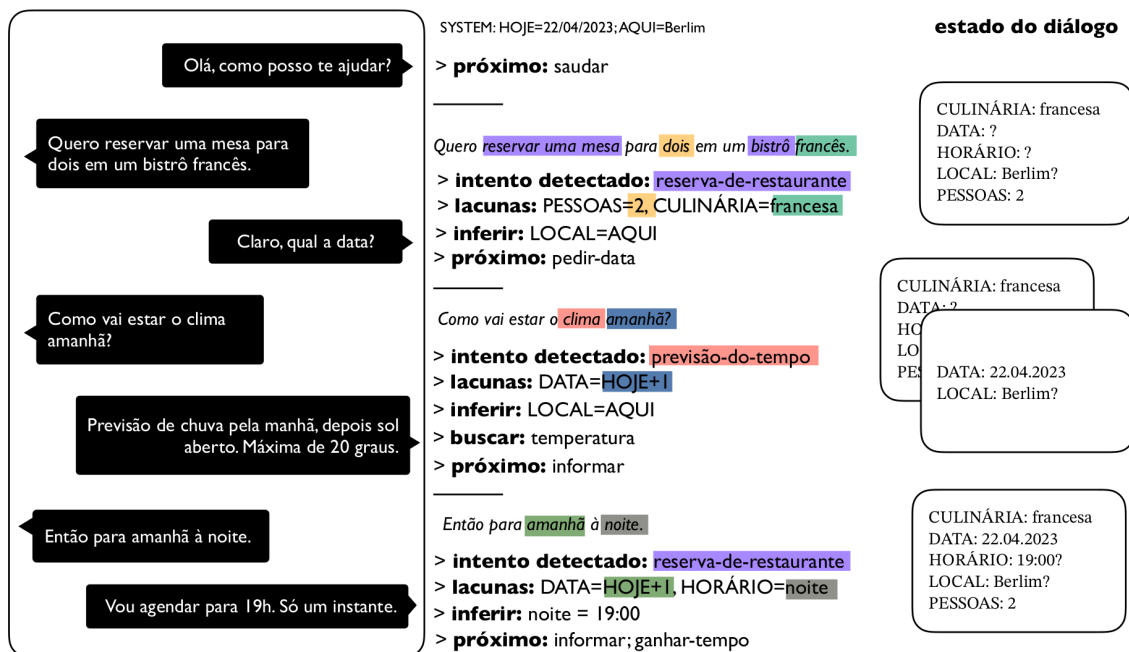
15.5 Avaliação: Verificando a qualidade de um sistema

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



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)

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”.

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 respeito 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 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 Avaliação de tecnologias de linguagem, 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 problema⁷.

Há duas formas de se efetuar avaliação humana: interativa, na qual a pessoa avalia uma

⁷Se já é difícil avaliar um modelo, como modelar e avaliar uma simulação de um humano interagindo com esse modelo?

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.

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 Questões éticas em IA e PLN 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 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 levandade.**

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 Conceitos: O que forma um diálogo?. Como exposto no Conjunto de dados, *dataset* e *corpus*, **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 (Paulada 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 diálogos 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 Avaliação de tecnologias de linguagem). 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 Conceitos: O que forma um diálogo?. Há recursos compostos de dados assíncronos, 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 fariam 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 inserção⁸. Tentar produzi-los artificialmente de forma explícita acaba resultando em dados menos autênticos. Há formas mais sutis de induzi-los implicitamente a acontecer.

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 humano¹⁰.

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 ParlAI¹¹ 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

⁸Esse dilema é conhecido como esparsidade de dados (*data sparsity*), usual em fenômenos linguísticos.

⁹https://en.wikipedia.org/wiki/Wizard_of_Oz_experiment

¹⁰Obviamente, nesses casos os participantes deve ser informados sobre isso após o experimento.

¹¹<https://parlai.ai/about/>



antiga) disponibilizada pela Linguateca¹².

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)	<i>corpus</i> 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 coletânea de diversos <i>corpora</i>	300 mil palavras
C-ORAL-BRASIL	Raso et al. (2015)	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 <i>tokens</i>
FakeWhatsApp.Br	Cabral et al. (2021)	<i>dataset</i> 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)	<i>dataset</i> 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)	<i>dataset</i> de mensagens de texto instantâneas sobre COVID-19	1.390
–	Sanches et al. (2022)	traduções de <i>dataset</i> com diálogos orientados a tarefas e coletânea de <i>posts</i> em fóruns da internet	milhões

¹²https://www.linguateca.pt/corpora_info.html



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, também 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 ou 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 (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: <i>chatbot</i> para auxiliar o combate à COVID-19	Fernandes et al. (2021)
AstroBot: <i>chatbot</i> para ensino e aprendizagem de física	Dantas et al. (2019)
BLAB: <i>chatbot</i> para disseminação de conhecimento sobre o território marítimo brasileiro	Matos et al. (2023)
CAERS: agente de conversação para intervenção pedagógica	Rossi et al. (2021)
<i>chatbot</i> para alocação de leitos em hospitais	Engelmann et al. (2021)

	Trabalho
<i>chatbot</i> 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: <i>chatbot</i> baseado em <i>software</i> livre	Lacerda; Aguiar (2019)
Guardião: <i>chatbot</i> para idosos	Ferreira et al. (2019)
Mordomo virtual	Coelho et al. (2008)
NL-SIIUE: agente recuperador de informações acadêmicas	Quintano; Rodrigues (2003)
Plantão Coronavírus: <i>chatbot</i> para detecção de sintomas da COVID	Coelho Da Silva et al. (2023)
Renan: <i>chatbot</i> para criação semi-automática de ontologias	Azevedo (2015)
TOB-SST: <i>chatbot</i> para suporte de educação sobre testes de <i>software</i>	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 encontrados nas referências. Para encontrar mais artigos científicos, recomendamos as seguintes fontes como ponto de partida¹³:

- 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/iwds/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.

¹³Devemos, claro, sempre manter nosso olhar crítico e questionador: não é porque um artigo foi publicado que tudo nele está necessariamente “certo”, validado ou justificado.



Ainda cabe uma pergunta, e talvez 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.

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 justificativa 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

¹⁴Por 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/>



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 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 Fenômenos: O que ocorre em uma conversa? e clique no *link* correspondente para acessar o exemplar completo. Anote todos os fenômenos discutidos na Fenômenos: O que ocorre em uma conversa? 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 Fenômenos: O que ocorre em uma conversa? não estão presentes? Quais parecem ocorrer mais ou menos raramente do que no exercício anterior?
3. Acesse o *site* do *corpus* NURC¹⁶, clique no ícone “Acessar página com o material” e depois em um título de sua escolha. Inspeção 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

¹⁵Visão 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/>



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 Conceitos: O que forma um diálogo? e considere também quais aspectos de multimodalidade estão envolvidos.

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 Modelos e Tarefas: Implementando um processo complexo.
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) 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.

¹⁷https://pt.wikipedia.org/wiki/Teoria_do_Garden_Path



Parte VIII

Dados



Capítulo 16

Conjunto de dados, *dataset* e *corpus*

Cláudia Freitas

Publicado em: 26/09/2023

Atualizado em: 29/09/2025



<https://brasileiraspln.com/livro-pln/4a-edicao/volume1/>

16.1 Introdução

Dados linguísticos são centrais no PLN, sobretudo nos paradigmas estatístico, neural e híbrido (Os Paradigmas de PLN). 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 dados¹.

Apesar da relativa popularização das abordagens neurais e dos LLMs 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;

¹Este 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.



- 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 distribuimos, ou classificamos, ou rotulamos) as palavras do texto conforme sua classe morfosintática, no segundo, podemos dizer que organizamos (ou distribuimos, 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 (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 Semântica distribucional. 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 lexicográficos², 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

²Lexicografia é a área que se dedica a fazer dicionários.



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 WinoBias³ (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 ódio⁴ (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>

⁴Como 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.



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 (Resolução de correferência). 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:

Analizamos 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 *datasets* – *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>



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 Codificação 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 Anotação: sabedoria de especialistas ou sabedoria da multidão? 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.

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 exemplos⁷ continuam necessários**, com a vantagem de agora os algoritmos necessitarem de uma quantidade menor deles, graças ao ajuste fino (Modelos de linguagem). Já se sabe que quanto mais cuidado na preparação dos dados, melhor o desempenho dos modelos – melhor a qualidade das predições⁸. 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.

⁷Bons exemplos são exemplos representativos, variados e analisados de maneira consistente, como veremos na Características de um bom *dataset* linguístico.

⁸Os 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

⁹Quando 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.



Por fim, é preciso 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 ?@sec-cap-modelos-avaliacao 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 **avancar 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 tarefa¹⁰. 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 Avaliação conjunta em português).

Voltando um pouco no tempo para exemplificar, foi a criação de um *corpus* padrão ouro (chamado “Coleção Dourada”) no âmbito da avaliação conjunta HAREM¹¹, 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.

¹⁰Já um *corpus* anotado padrão ouro pode ter sido criado com a intenção de estudar um determinado fenômeno linguístico.

¹¹<https://www.linguateca.pt/HAREM/>.

¹²http://propor2016.di.fc.ul.pt/?page_id=381 e <https://sites.google.com/view/assin2/>.



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 (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 (Semântica distribucional), 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 (Anotação: sabedoria de especialistas ou sabedoria da multidão?), 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 diferente¹³. Os dados são sempre um registro do que foi dito em algum tempo/espço, 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

¹³<https://theconversation.com/ai-free-from-bias-and-ideology-is-a-fantasy-humans-cant-organise-data-without-distorting-reality-262523>



o objetivo de eliminar vieses indesejados, ao estilo do WinoBias, que artificialmente (e indiretamente) constroem relações igualitárias de gênero, raça ou etnia, seriam capazes de produzir linguagem e, conseqüentemente, 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 conterà, invariavelmente, muito mais avaliações negativas do que positivas. Se desejamos um *dataset* para reconhecimento de fala (Recursos para o processamento de fala), 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 todos¹⁴.

¹⁴De 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.



- **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 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 enviados, 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 morfosintática 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 Anotação: sabedoria de especialistas ou sabedoria da multidão?.

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

- 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 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 morfosintática.

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*; 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.

- Nunca_{ADV}[0] sofri_V[-] tanto_{ADV}[0] para_{PREP}[0] ler_V[0] um_{ART}[0] livro_{SUBST}[0] [frase negativa]
- Sério, sofri pra terminar o livro, o livro ainda consegue ser pior que os filmes. [frase negativa]
- 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* Bosque¹⁵, e da coleção dourada do HAREM¹⁶, 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

¹⁵<https://www.linguateca.pt/Floresta/BibliaFlorestal/>.

¹⁶https://www.linguateca.pt/aval_conjunta/HAREM/directivas.html.



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.

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 impossível 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 (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 mostrou¹⁷. 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

¹⁷Embora, 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”.

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 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 (Sequência de caracteres e palavras), 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;
- 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 morfo-sintáticas, e informações não preenchidas são marcadas com o caractere especial “_”.

A Figura 16.3 mostra uma frase no formato CoNLL-U¹⁸.

Figura 16.3: Anotação no formato CoNLL-U para dependências sintáticas.

```
# sent_id = 1
# text = Ana conheceu a Serra da Mantiqueira ontem.
1  Ana  Ana  PROP  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  PROP  Gender=Fem|Number=Sing 2  obj  _  _
5-6  da  de  ADP  _  _  _  _
6  a  o  DET  Definite=Def|Gender=Fem|Number=Sing|PronType=Art 7  det  _  _
7  Mantiqueira  Mantiqueira  PROP  Number=Sing 4  nmod  _  _
8  ontem  ontem  ADV  _  _  _  _
9  .  .  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* HateXplain¹⁹, 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 selecio-

¹⁸Mais 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>.

nada 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.

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 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 Conjunto de etiquetas e instruções: o esquema de anotação. 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 Questões éticas em IA e PLN) até a crítica de que o que se chama de inteligência artificial é o fruto de milhares de trabalhadores invisíveis e precarizados²⁰.

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 Conjunto de etiquetas e instruções: o esquema de anotação). 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

²⁰A 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/>.

²¹O 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>.



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 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 exige 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 lei²², 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 Procedimentos e estratégias de anotação e revisão.

²²Veja-se esta página do Ministério Público Federal dedicada ao tema: <https://respeitediferenca.mpf.m.p.br/www/discurso-odio.html>.



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, 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

- BRAT²³: para anotação de entidades mencionadas, correferência, dependências sintáticas, entre outras.
- Label Studio²⁴: para anotação de entidades mencionadas, perguntas e respostas, análise de sentimento, entre outras.
- Inception²⁵: para diversos tipos de anotação semântica e discursiva.
- WebAnno²⁶: para anotação morfológica, sintática e semântica.
- Arborator²⁷: para anotar e buscar dependências sintáticas em formato UD.
- ET: dois ambientes integrados para buscar, revisar e avaliar dependências sintáticas em formato UD: Interrogatório²⁸, para buscar e revisar anotações e Julgamento²⁹, 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

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>



é revista por pessoas³⁰. Este será o tema da Procedimentos e estratégias de anotação e revisão.

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 Como avaliar a qualidade do *dataset*?

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 (Sequência de caracteres e palavras e (Freitas, 2022)). A língua portuguesa dispõe de alguns, para diferentes tarefas³¹.

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 Emocionário³² contém uma lista de palavras que, em algum contexto, descrevem algum tipo de emoção e/ou sentimento da língua portuguesa³³. 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.

³⁰Já quando falamos de um *corpus* anotado, além dessas duas maneiras, existe a possibilidade de uma anotação 100% automática.

³¹Ver GitHub do Brasileiras em PLN: <https://github.com/brasileiras-pln>

³²<https://www.linguatca.pt/acesso/corpos/dicemocoos.iso.txt>

³³Para mais informações, veja <https://www.linguatca.pt/Gramateca/Emocionario.html>.



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.

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 país	moto caminhão avião asteróide	Maria

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 morfosintática (codificada nos atributos *lemma* e *deprel*, que indica a função sintática). Ou seja, a regra diz³⁴:

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

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 regras³⁶.

³⁴O exemplo é baseado na utilização da ferramenta Interrogatório <https://github.com/alvelvis/Interrogatorio>.

³⁵“Controlado” porque sabemos onde começa e onde termina, isto é, é um conjunto finito de enunciados linguísticos.

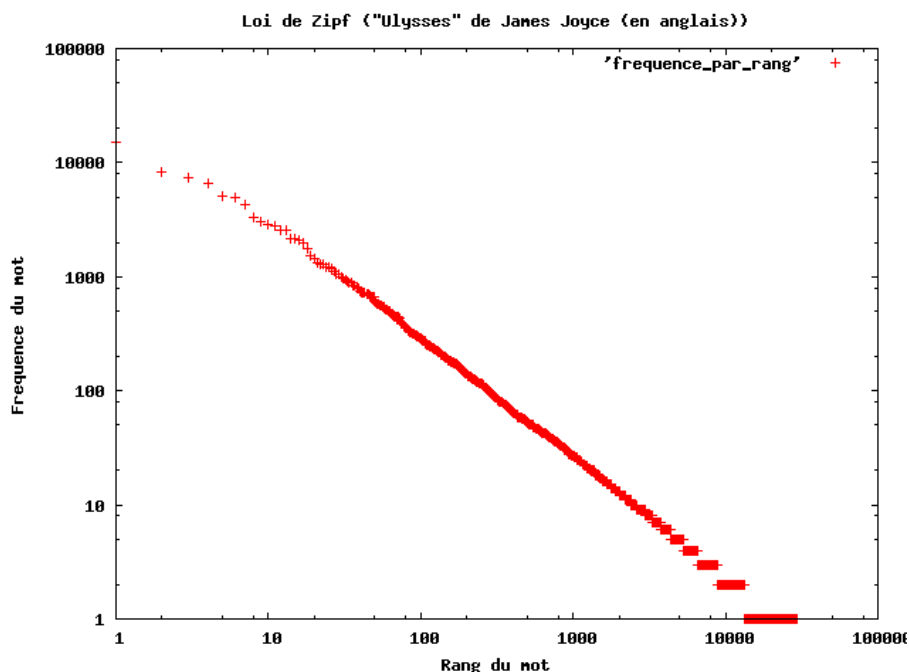
³⁶E não custa lembrar que os anos iniciais do PLN foram dominados pela crença de que o processamento linguístico poderia ser inteiramente resolvido por meio de regras linguísticas.

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.

Figura 16.5: Distribuição das palavras do livro *Ulisses*, de James Joyce.

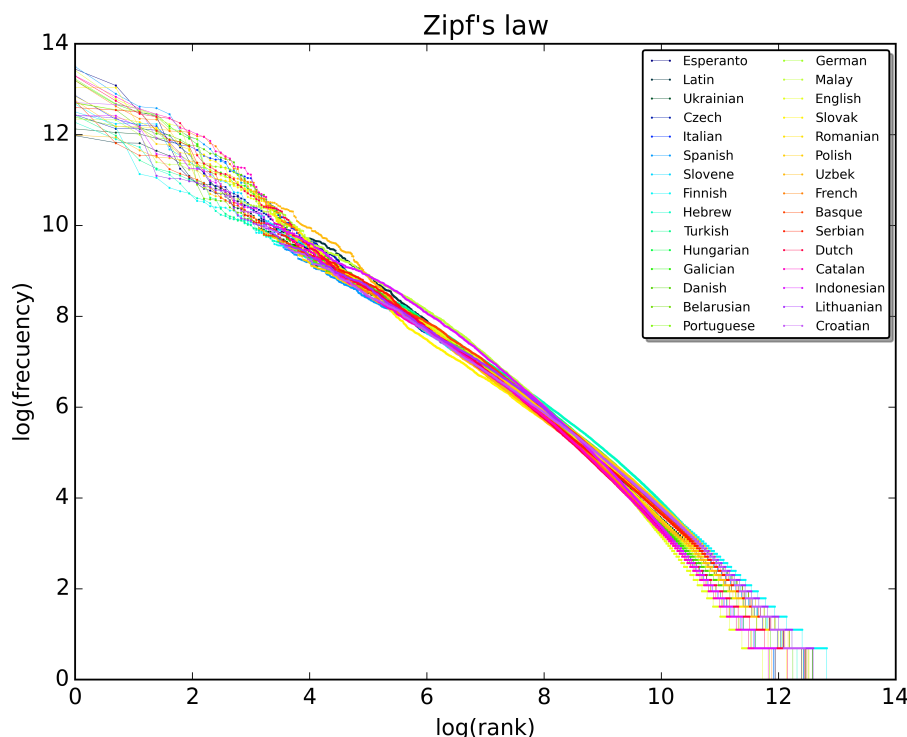


Fonte: Wikipédia

Quando ficamos cientes desta propriedade distribucional, conseguimos entender melhor



Figura 16.6: Distribuição das 10 milhões de palavras mais frequentes em 30 wikipédias.



Fonte: Wikipédia

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 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 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.

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 v3³⁷, cada rodada de revisões foi acompanhada de uma rodada de “avaliação de aprendizagem” (uma *avaliação intrínseca*, como veremos na Como avaliar a qualidade do *dataset*?) com o objetivo de verificar o impacto das melhorias linguísticas na aprendizagem automática. Os resultados estão na Tabela 16.1³⁸:

Tabela 16.1: Evolução de um *corpus* comparando quantidade de revisões e melhoria no aprendizado de máquina.

	<i>Tokens</i> 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 *Datasets* pra quê?, 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 qualidade³⁹.

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 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.

³⁷<https://petroles.puc-rio.ai/files/Corpora/petrogold-v3.zip>

³⁸O trabalho está detalhadamente descrito em (Souza, 2023) e (Souza; Freitas, 2023).

³⁹Veja mais sobre Avaliação de sistemas de PLN no Avaliação de tecnologias de linguagem.



- 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.

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 (Quando a discordância não é erro: divergências legítimas)).

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 (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 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 (E o significado?), 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”)

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.

⁴⁰<https://nlperspectives.di.unito.it/perspectivist-approaches-to-nlp/>



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 (Modelos de linguagem).

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 ve-



mos 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 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 Até onde precisamos rever?.

A segunda vantagem desta abordagem já foi esboçada nos parágrafos anteriores: uma avaliação intrínseca de *datasets* permite verificar o resultado de experimentações na clas-

sificação 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 (Sequência de caracteres e palavras). 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 melhorar o desempenho em outras tarefas, como extração de informação⁴¹ (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 sintáticas⁴² (Sequência de caracteres e palavras), e prepararam *datasets* anotados conforme as três abordagens gramaticais: as chamadas dependências sintáticas de Stanford, as dependências sintáticas

⁴¹Para os estudos linguísticos a anotação sintática pode ser um fim, quando o interesse está em estudar aspectos sintáticos da língua.

⁴²Importante notar que a tarefa não pressupõe a existência de uma anotação sintática anterior.



do projeto Universal Dependencies e dependências sintáticas usadas nas tarefas CONLL⁴³ (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 obtém 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; Övrelid (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. MacMorpho⁴⁴ e Bosque, os mais antigos, são os que contém mais variações, e para o Bosque-UD⁴⁵ 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 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)

⁴³Considerando 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.

⁴⁴O *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>.

⁴⁵O Bosque anotado com a abordagem UD está disponível em https://universaldependencies.org/treebanks/pt_bosque/index.html. Demais versões estão disponíveis em <https://www.linguatca.pt/Floresta/levantamento.html>

⁴⁶A preparação do *corpus* está descrita em (Freitas; Souza, 2021) e o *corpus* está disponível em <https://github.com/alvelvis/desocultando-sujeitos>

ATRIBUTO	VARIAÇÃO	CORPUS	REFERÊNCIA
Sintaxe	PALAVRAS/Floresta	Bosque	(Freitas et al., 2008b)
	Sintá(c)tica		
	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. (*Datasets* pra quê?)
- *Datasets* bons são consistentes (boa avaliação intrínseca), bem documentados, de tamanho adequado e com dados adequados. (Características de um bom *dataset* linguístico)
- A anotação de um *dataset* não precisa ser 100% perfeita para funcionar no aprendizado de máquina. (Até onde precisamos rever?)
- A preparação de um *dataset* se beneficia se a revisão/anotação é feita de maneira sistemática. (Procedimentos e estratégias de anotação e revisão)
- Avaliação intrínseca não significa, necessariamente, melhor adequação à tarefa, que é medida pela avaliação extrínseca. (Como avaliar a qualidade do *dataset*?)

Um bom projeto cuidadoso de anotação, por sua vez, deverá levar em conta (Por onde começar?):

- Clareza quanto ao fenômeno que será anotado (que se reflete em um bom esquema de anotação);
- 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 (Como avaliar a qualidade do *dataset*?):

- 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!).



Capítulo 17

Aprendizado Transdutivo em PLN

Nataly Leopoldina Patti da Silva
Norton Trevisan Roman Publicado em: 29/09/2025

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* (Conjunto de dados, *dataset* e *corpus*). 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 Ferramentas e recursos para o processamento sintático)? 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!).



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.

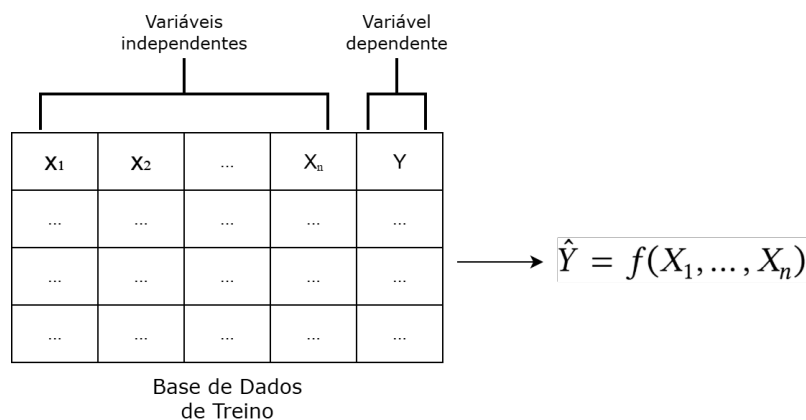
Que fazer? Se sua sorte não estiver tão ruim, é possível que existam ferramentas para anotação automática do *corpus*, como as descritas no Capítulo Ferramentas e recursos para o processamento sintático. Restaria então analisar uma amostra dessa anotação para se obter uma estimativa do desempenho da 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 (veja (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 $\{X_1, \dots, X_n\}$ em conjunção com determinados valores para uma variável Y , induzimos uma relação

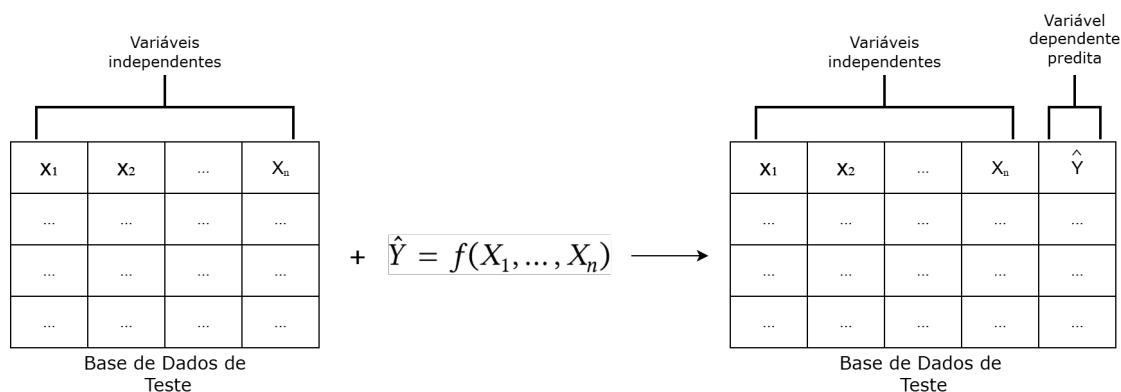
$\hat{Y} = f(X_1, \dots, X_n)$ entre elas¹, conforme ilustra a Figura 17.2.

Figura 17.2: Processo de Indução. A partir de um conjunto de observações para um grupo de variáveis preditoras ($X_1, \dots, X_n$) 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 $\hat{Y} = f(X_1, \dots, X_n)$, e do conhecimento do valor de $\{X_1, \dots, X_n\}$, deduzimos o valor de Y ou, no caso, apresentamos uma estimativa $\hat{Y}$ para Y (Figura 17.3).

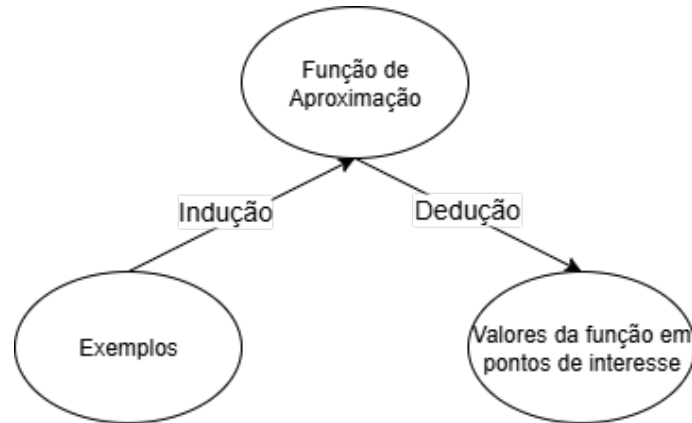
Figura 17.3: Processo de Dedução. A partir de um conjunto de observações para um grupo de variáveis preditoras ($X_1, \dots, X_n$), correspondentes a exemplos de dados não rotulados, em conjunção com uma relação matemática entre elas e a variável alvo desejada ($\hat{Y} = f(X_1, \dots, X_n)$), o modelo 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 $\hat{Y} = f(X_1, \dots, X_n)$ entre eles. Para verificar a plausibilidade de nossa indução, aplicamos $\hat{Y}$ aos valores de $\{X_1, \dots, X_n\}$ 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.

¹Usamos a notação $\hat{Y}$ para denotar que $\hat{Y}$ é uma estimativa do valor de Y .

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).

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 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 Vapnik² (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.

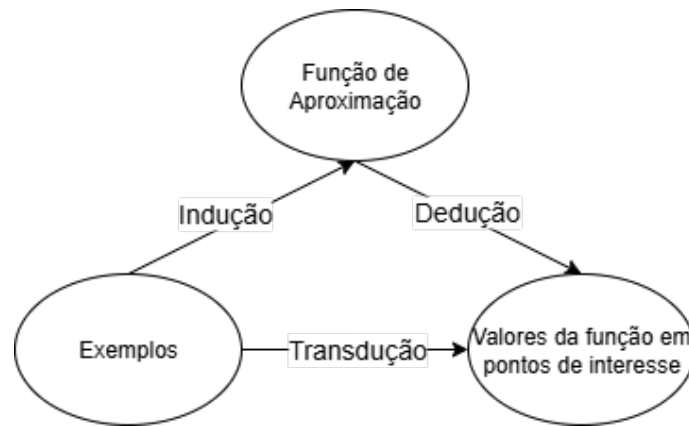
Formalmente³, o aprendizado transdutivo parte de um conjunto de m instâncias:

$$S = \{s_1, s_2, \dots, s_m\}$$

²Também conhecido por seu papel fundamental no desenvolvimento das Máquinas de Vetores de Suporte (SVMs).

³Veja (Chapelle et al., 2006) e (Zhu et al., 2009).

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).

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 $s_i \in S$:

$$X = (\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_m)$$

A partir de X , definem-se dois subconjuntos, $X_{\text{treino}} = (\mathbf{x}_{tr_1}, \mathbf{x}_{tr_2}, \dots, \mathbf{x}_{tr_r})$ e $X_{\text{teste}} = (\mathbf{x}_{ts_1}, \mathbf{x}_{ts_2}, \dots, \mathbf{x}_{ts_n})$, 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 $X = X_{\text{treino}} \cup X_{\text{teste}}$.

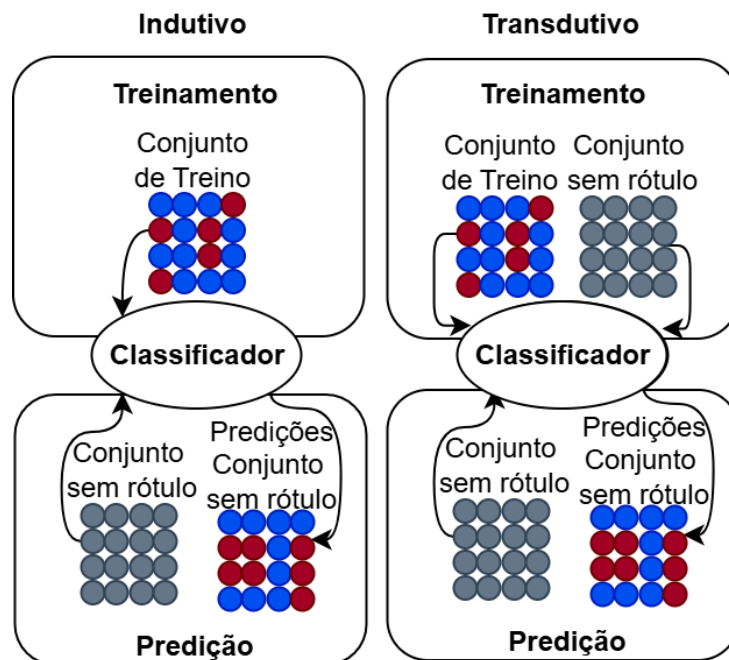
Ao conjunto de treino X_{treino} associamos seus rótulos $Y_{\text{treino}} = (y_{tr_1}, y_{tr_2}, \dots, y_{tr_r})$ correspondentes, tais que cada rótulo $y_{tr_i} \in Y_{\text{treino}}$, onde $1 \leq i \leq r$ está associado a seu respectivo $\mathbf{x}_{tr_i} \in X_{\text{treino}}, 1 \leq i \leq r$. Durante o processo de aprendizado, o método transdutivo utiliza esses conjuntos para inferir os rótulos $\hat{Y}_{\text{teste}} = (\hat{y}_{ts_1}, \hat{y}_{ts_2}, \dots, \hat{y}_{ts_n})$ 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.

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.



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.

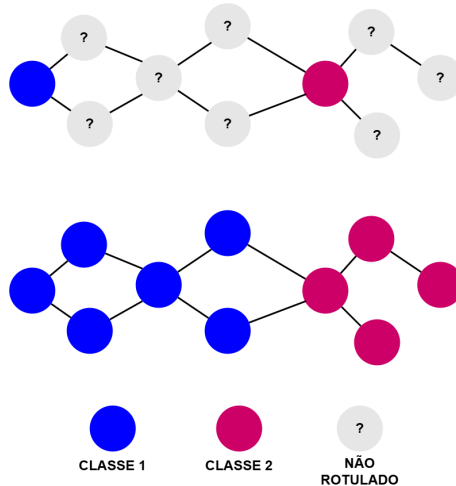
17.3.1 Métodos baseados em propagação de rótulos

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 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 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 $V = 1, \dots, n$ 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 $W_{i,j}$.

- 1: Calcular a matriz de afinidade $\mathbf{W}$ por meio da função de Kernel Gaussiana com largura σ , em que $\mathbf{W}_{ij} = e^{-\frac{\|x_i - x_j\|^2}{2\sigma^2}}$
- 2: Calcular a matriz diagonal de graus $\mathbf{D}$, em que $\mathbf{D}_{ii} \leftarrow \sum_j W_{ij}$
- 3: Inicializar $\hat{Y}^{(0)} \leftarrow (y_1, \dots, y_l, 0, 0, \dots, 0)$
- 4: **enquanto** não convergir para $\hat{Y}^{(\infty)}$ **faça**
- 5: $\hat{Y}^{(t+1)} \leftarrow \mathbf{D}^{-1} \mathbf{W} \hat{Y}^{(t)}$
- 6: $\hat{Y}_l^{(t+1)} \leftarrow Y_l$
- 7: **fim enquanto**
- 8: 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.

Algoritmo 17.2: Técnica de Espalhamento de Rótulos (Chapelle et al., 2006)

Entrada: Grafo $g = (V, E)$ com nós $V = (1, \dots, n)$ 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 $W_{i,j}$.

- 1: 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 $\mathbf{W}_{ii} \leftarrow 0$)
- 2: Calcular a matriz diagonal de graus $\mathbf{D}$ por $\mathbf{D}_{ii} \leftarrow \sum_j W_{ij}$
- 3: Calcular o Laplaciano normalizado do grafo $\mathbf{L} \leftarrow \mathbf{D}^{-1/2} \mathbf{W} \mathbf{D}^{-1/2}$
- 4: Inicializar $\hat{Y}^{(0)} \leftarrow (y_1, \dots, y_l, 0, 0, \dots, 0)$
- 5: Escolher um parâmetro $\alpha \in [0, 1]$
- 6: **repita**
- 7: $\hat{Y}^{(t+1)} \leftarrow \alpha \mathbf{L} \hat{Y}^{(t)} + (1 - \alpha) \hat{Y}^{(0)}$
- 8: **até que** convergir para $\hat{Y}^{(\infty)}$
- 9: 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 *Label Propagation* ou *Label Spreading*, 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 (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.

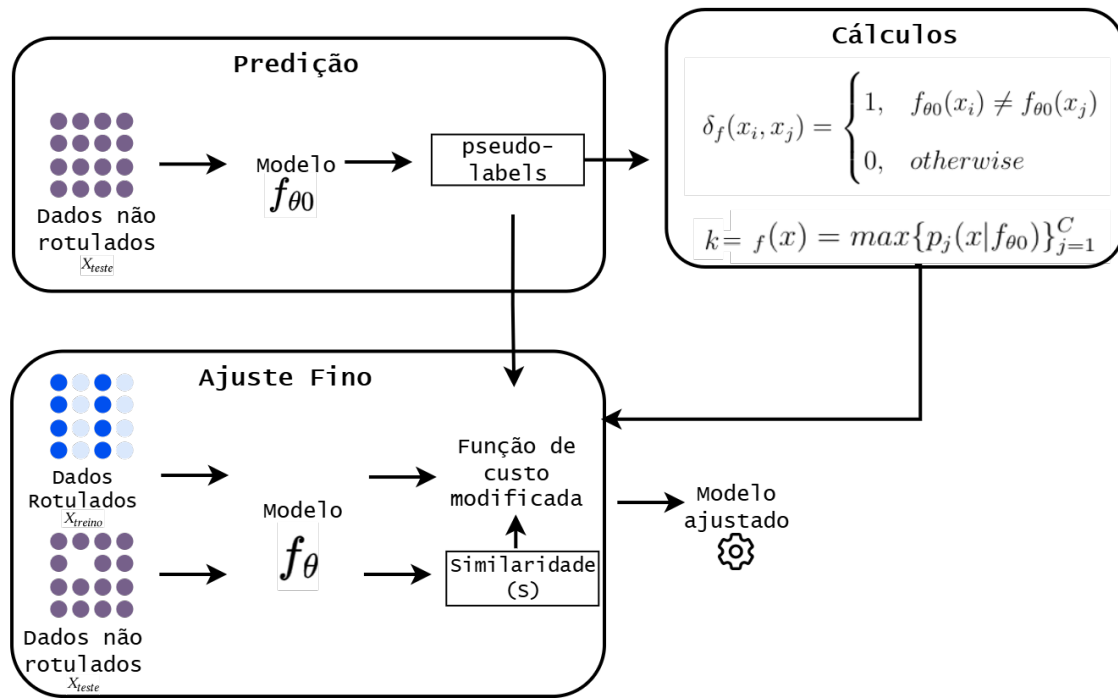
Inspirado no método TransBoost (Belhasin et al., 2022), é possível modificar a função de custo do processo de ajuste fino de um modelo pré-treinado para que ela leve em consideração não apenas os exemplos rotulados, mas também os não rotulados. Essa abordagem permite então ao modelo explorar a estrutura 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.

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_{θ_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, três pontos principais são considerados para o cálculo do termo transdutivo na função de custo:



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_{θ_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_{θ_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_{θ_0} nessa probabilidade dada. Essas informações serão então utilizadas durante o ajuste fino transdutivo para adicionar maior penalidade ou não à 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 .

```

1: para  $epoch \leftarrow 1$  até  $E$  faça
2:   // Amostrar lotes rotulados e não rotulados de forma cíclica
3:   para  $S'_r \triangleq (X'_{treino}, Y'_{treino}) \subset S_r, X'_{teste} \subset X_{teste}$  até  $\max\{\lceil \frac{R}{R'} \rceil, \lceil \frac{N}{N'} \rceil\}$ 
     faça
4:     // Preparar pares aleatórios de teste
5:      $X'_{teste} \leftarrow$  gerar uma permutação aleatória de  $X'_{teste}$ 
6:     // Aproximar a perda do TransBoost
7:      $L_{TransBoost} \leftarrow 0$ 
8:     para  $i \leftarrow 0$  até  $N'$  faça
9:        $L_{TransBoost} \leftarrow L_{TransBoost} + \kappa(x_i)\kappa(x'_i)\delta(x_i, x'_i)\mathcal{S}(x_i, x'_i)$ 
10:    fim para
11:     $L_{TransBoost} \leftarrow \begin{cases} \frac{L_{TransBoost}}{\sum_{i=1}^{N'} \delta(x_i, x'_i)} & \text{se } \sum_{i=1}^{N'} \delta(x_i, x'_i) \neq 0 \\ 0 & \text{caso contrário} \end{cases}$ 
12:    // Aplicar uma etapa de otimização
13:     $\theta \leftarrow$  otimizar  $\left[ \frac{1}{R'} \sum_{i=1}^{R'} \ell(f_\theta(x_i), y_i) + \lambda \cdot L_{TransBoost} \right]$ 
14:  fim para
15: 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 $X_{treino} = \{(x_1, y_1), \dots, (x_r, y_r)\}$ e de um conjunto finito de dados não rotulados $X_{teste} = \{x_{r+1}, \dots, x_{r+n}\}$. Este objetivo é dado por:

$$L(X_{treino}, Y_{treino}, X_{teste} | f_\theta, S, \delta, k) = \frac{1}{R} \sum_{n=1}^R \underbrace{\ell(f_\theta(x_i), y_i)}_{\text{perda indutiva}} + \lambda \cdot \underbrace{L_{Transdutivo}(X_{teste} | f_\theta, S, \delta, k)}_{\text{perda\_transdutiva}}$$

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_{θ_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_{θ_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_{θ_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:

$$L_{\text{indutiva}} = \frac{1}{2} [-\log(0,8) - \log(0,7)] \approx \frac{1}{2}(0,223 + 0,357) = 0,29$$

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_n . Seu funcionamento é dado por:

$$l_{\text{Transdutivo}}(X_{\text{teste}}|f_\theta, S, \delta, k) = \frac{1}{N_\delta} \sum_{1 \leq i \leq N} k(x_i)k(x_j)\delta(x_i, x_j)S(x_i, x_j)$$

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 $\delta : X \times X \rightarrow \{0, 1\}$, obtida da seguinte maneira:

$$\delta_f(x_i, x_j) = \begin{cases} 1, & f_{\theta_0}(x_i) \neq f_{\theta_0}(x_j) \\ 0, & \text{caso contrário} \end{cases}$$

onde f_{θ_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_{θ_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_{θ_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.

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 o mesmo pseudo-rótulo 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).

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.

Parte IX

Avaliação



Capítulo 18

Avaliação de tecnologias de linguagem

Brielen Madureira

Publicado em: 13/03/2024

Atualizado em: 20/11/2024



<https://brasileiraspln.com/livro-pln/4a-edicao/volume1/>

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



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



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 Diretrizes de avaliação. 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 (Tradução Automática) e recuperação de informação (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 habitabilidade, 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

¹O 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.



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 porções 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 EAGLES, 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 Paradigmas: Tipos de avaliação.

Mais recentemente, podemos citar a série SemEval³ 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 Avaliação conjunta em português. Há também *workshops* internacionais dedicados exclusivamente a esse tema, como o Eval4NLP⁴, sobre avaliação em geral, e o HumEval⁵, sobre avaliação por humanos. A conferência LREC⁶ também tem um forte

²https://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/>



foco nesse tópico, e 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 matematicamente 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



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 $L = \{L_1, \dots, L_M\}$ 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 (Sequência de caracteres e palavras) e reconhecimento de entidades nomeadas (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, conseqüentemente, 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, $V = \{w_1, \dots, w_K\}$ 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



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 substituí-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 (A ordem e a função das palavras em uma sentença 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 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 extraíndo 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.

⁷Não 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.

⁸Conceitos baseados em (Köhn, 2018).



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	<i>input</i>	<i>output</i>	exemplos
estimar uma pontuação ou probabilidade classificação	qualquer	um número	similaridade de textos, análise de sentimento
	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 <i>input</i>	anotação de <i>PoS</i> , reconhecimento de entidades nomeadas
<i>sequence-to-sequence</i>	uma sequência com n elementos	uma sequência com m elementos	tradução automática, sumarização
predição estruturada	qualquer	uma estrutura	extração de relações, <i>parsing</i>
agrupamento (<i>clustering</i>)	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-



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 *Datasets* pra quê?). 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.

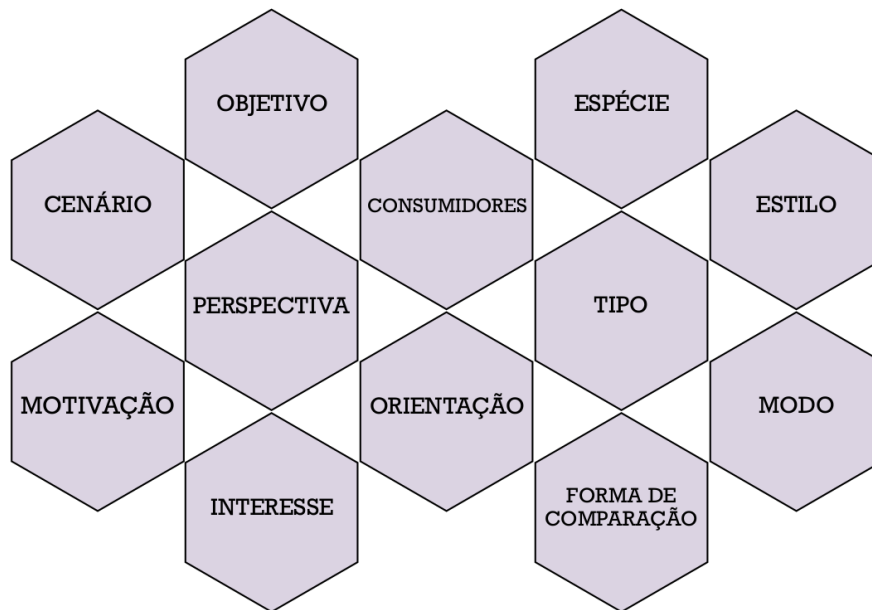
Boa parte das avaliações em PLN pressupõe a existência de uma referência “padrão ouro” (*gold standard*), ou seja, a resposta considerada correta. Um procedimento costumeiro é comparar a resposta do sistema com a referência e julgar, com base em uma métrica, quão próximas ou parecidas elas são. Nem sempre 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 (E o significado?).

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



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.

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,



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

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 Avaliação conjunta em português.

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,

⁹Para 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.



demonstramos a vantagem daquele componente para nosso sistema. Se não ocorre, argumentamos sobre os possíveis motivos para nossa hipótese estar errada. Talvez estejamos 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 reproduzíveis 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 (Perguntas e Respostas), 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”,

¹⁰Na 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.

¹¹Novamente, 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).



que não está errado, apesar de ser uma construção diferente que não estava entre nossas referências.

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 ou suficientes 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 Como avaliar a qualidade do *dataset*?).

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

¹²<http://pdai.info/>



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 mais às métricas que usamos, pois algumas delas são enviesadas. Trataremos de métricas em mais detalhes na Métricas: Medindo a performance.

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

¹³Ou seja, um “representante substituto”.

¹⁴https://pt.wikipedia.org/wiki/Validação_cruzada



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 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 estimativa.

Há algumas formas diferentes de se realizar esse procedimento:

- **holdout:** a partição fixa discutida no item anterior, trata-se 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).

18.4.6 Comparação de Performance

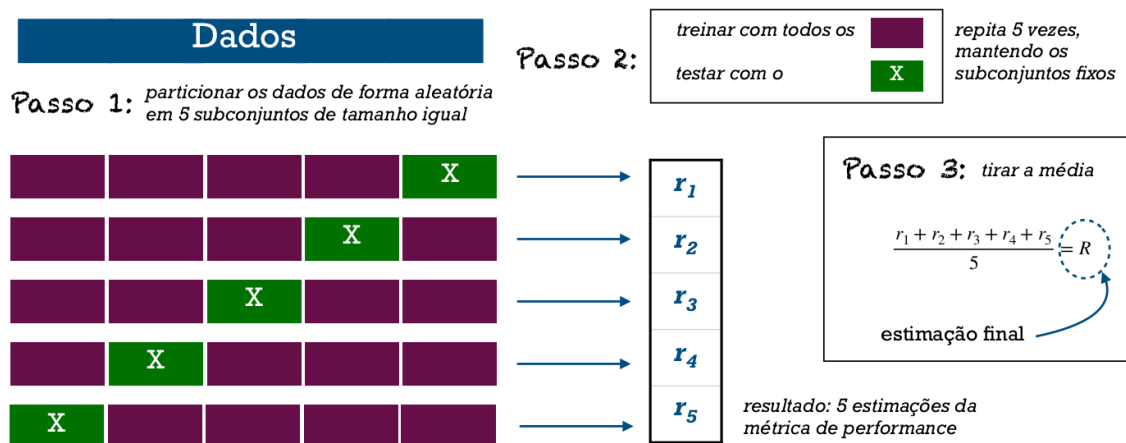
Assim como o valor numérico de uma métrica sozinho não é muito informativo, o desempenho de um sistema também faz mais sentido se ocorrer em relação a outros sistemas ou às referências chamadas *baselines*. Há quatro principais *baselines* com as quais podemos comparar um sistema. A primeira é saber 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



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.



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 Concordância entre-anotadores). 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 Uso Responsável e Boas Práticas.

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 , 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 amostras dos dados e estimam um p-valor que indica a probabilidade de observarmos 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 significativa só deve ser usada na descrição se resultados de 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 previsões 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 previsão 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. prever que todos os textos são sobre mudanças climáticas
2. prever 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

¹⁵Esta seção se baseia em <https://blog.ml.cmu.edu/2020/08/31/3-baselines/> e <https://ehudreiter.com/2018/08/30/use-proper-baselines/>.



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 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 sub-componentes 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.



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 Anotação: sabedoria de especialistas ou sabedoria da multidão?). 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 Conjunto de dados, *dataset* e *corpus*, (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.

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. Vide Conjunto de dados, *dataset* e *corpus* 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

¹⁶Ver, por exemplo, as recomendações em <https://naacl2018.wordpress.com/2017/12/19/putting-the-linguistics-in-computational-linguistics/>.

¹⁷Para detalhes da taxonomia em tradução automática, ver a Métricas Humanas.

¹⁸https://pt.wikipedia.org/wiki/Escala_Likert



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 funciona 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).

¹⁹<https://humeval.github.io/>



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 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 *Bilingual Evaluation Under-study* (BLEU)) para avaliação de geração de linguagem natural em tradução automática (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 BLEU²⁰.

²⁰Esse 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



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, 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 Contexto: Por onde começar?.

Quadro 18.3. Algumas métricas usualmente utilizadas na avaliação de sistemas de PLN.

pequenas variações no BLEU não têm sentido na realidade, em <https://ehudreiter.com/2020/07/28/small-differences-in-bleu-are-meaningless/>.

²¹https://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>



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
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)?
- 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-las? 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!).

²⁴Recomendação 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.

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).

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.

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 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.

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,



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

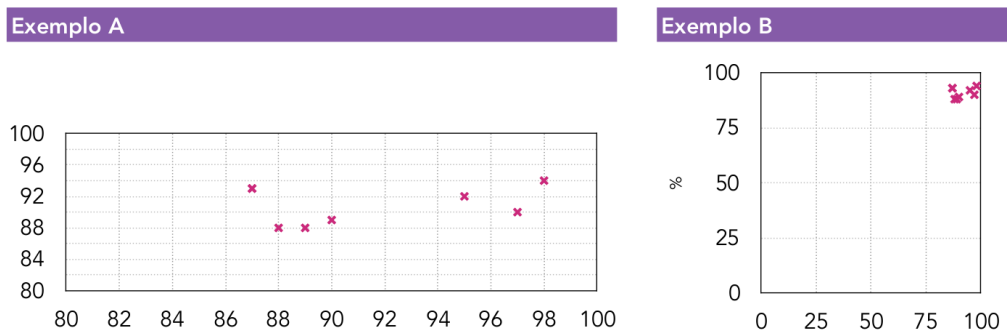
Tabela 1: Performance dos três modelos avaliados. Para todas as métricas, quanto maior, melhor. Valores em negrito representam o melhor resultado em cada métrica e tarefa. O modelo 2 foi superior aos demais nas métricas 1 e 2, enquanto o modelo 1 teve melhores resultados na métrica 3, na qual os modelos 2 e 3 estão abaixo da *baseline*.

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 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.

²⁵<http://coling2018.org/slowly-growing-offspring-zigglebottom-anno-2017-guest-post/>

Figura 18.6: Dois exemplos contendo 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 das outras, pois temos o panorama completo de mínimo e máximo da porcentagem e a proporção dos eixos é igual.



- **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 convenceremos 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

²⁶Ver 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/>).

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).

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 Questões éticas em IA e PLN e E agora, PLN?, 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 ética²⁷ 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.

²⁷Disponível em <https://www.acm.org/code-of-ethics>.



- 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 convenceremos de que a avaliação é parte fundamental de qualquer tecnologia de linguagem, especialmente na era 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.
- 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?

²⁸Por 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>.



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*
 - 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.
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/>



Capítulo 19

Avaliação conjunta em português

Diana Santos
Cláudia Freitas

Publicado em: 20/11/2024

 <https://brasileiraspln.com/livro-pln/4a-edicao/volume1/>

19.1 Apresentação

Avaliação conjunta¹ 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 época². A Linguateca³, 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.

Vejam 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 despendar, 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

¹Este capítulo é fortemente inspirado em Santos (2007a) e Santos (2021).

²Diana Santos é a coordenadora da Linguateca, e Cláudia Freitas integrou a equipe de 2007 a 2023.

³www.linguateca.pt



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

⁴Para uma descrição da Linguatca em fases distintas da sua existência, ver Santos (2000) e Santos (2009). Tendo o financiamento acabado em 2011, a Linguatca transformou-se num projeto voluntário.



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 (Avaliação de tecnologias de linguagem):

- **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.



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.



- 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 (CLEF: Recuperação de Informação e QA (2003-2009)) ou do IberLEF (IberLEF 2019: reconhecimento de entidades e de relações).

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.

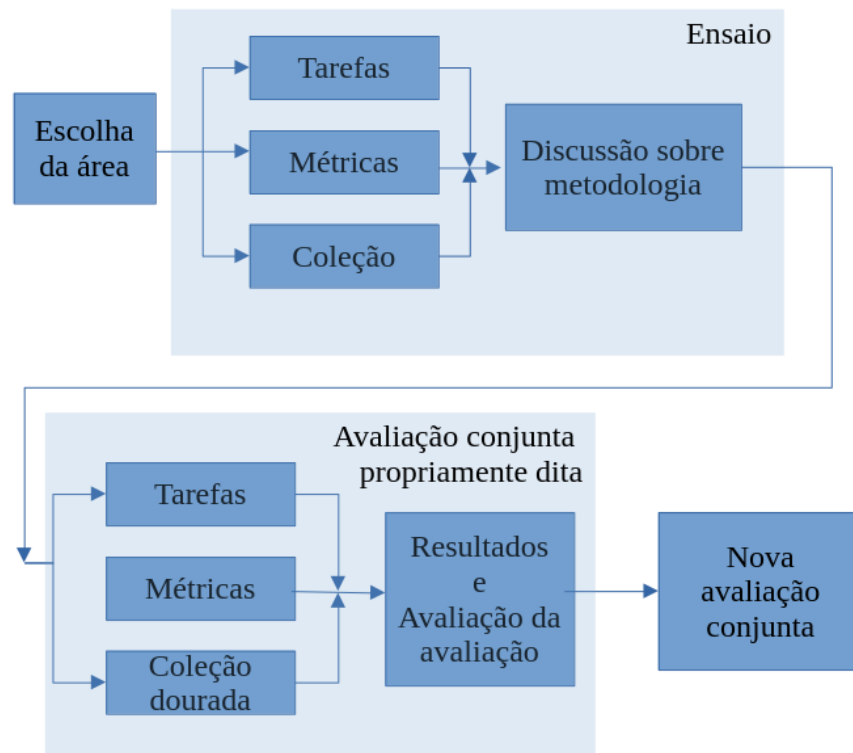
É também preciso sublinhar que a existência de uma comunidade de pessoas que se reconhecem num mesmo objetivo e com as quais é possível discutir ideias e observar as consequências de opções diferentes, levando a uma fertilização cruzada e a um reconhecimento público, é extremamente importante para transformar 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.

⁵O 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.



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 avaliações⁶, 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**

⁶Como pode ser observado dada a proliferação de *shared tasks* multilíngues.

(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 (Primeiro e Segundo HAREM: avaliação de entidades, relações e expressões temporais (2005 e 2007-2008)) 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 Resumo das avaliações conjuntas, é preciso sempre compreender o contexto e a definição da tarefa subjacente.

Para isso, nunca é demais insistir numa documentação detalhada das opções tomadas, para que o reuso possa ser feito de maneira informada e consciente.

⁷Esta 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.



19.4.3 Medidas de avaliação

As medidas por que se pauta uma avaliação conjunta são a sua pedra de toque. Ver também a Métricas: Medindo a performance.

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



ser hora de mudar radicalmente o cenário da avaliação. Hirschman (1998), ao descrever o MUC⁸, 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 Senseval¹⁰, na esteira do que já Kilgarrieff (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 TREC¹¹ e o CLEF (descrito na CLEF: Recuperação de Informação e QA (2003-2009)) 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

⁸MUC, *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.

⁹Foi o que aconteceu no TREC QA, uma avaliação conjunta de respostas a perguntas, o que foi amargamente discutido na comunidade de PLN da altura.

¹⁰O 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>).

¹¹O 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).



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 Testes de Significância Estatística.

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-ouro¹²: 200 formas que, por sua vez, correspondiam a 345 análises diferentes. Os sistemas participantes deveriam então retornar todos os três materiais de teste morfológicamente 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. radicalização¹³

¹²Na Linguatca 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...

¹³Esta tarefa é chamada em inglês *stemming* e é usada em recuperação de informação; é parecida com lematização mas não pressupõe análise linguística.



Além disso, talvez o resultado mais interessante das Morfolimpiadas 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 Morfolimpiadas levaram a um grande aprendizado sobre como avaliar, cooperativamente, uma determinada aplicação.

Todos os resultados, dados e programas (em Perl) usados para organizar as Morfolimpiadas foram divulgados na página da Linguatca dedicada a esta avaliação - onde ainda podem ser encontrados¹⁴.

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.linguatca.pt/Morfolimpiadas/>

¹⁵Notamos que o termo “entidades nomeadas” também tem sido utilizado em português.

¹⁶No 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).



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 brasileira (classe PESSOA), ou, ainda, à organização política/Estado (classe ORGANIZAÇÃO). 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 Linguatca 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 Linguatca.

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).



Diagrama de uma rede semântica sobre o tema "EM" (Ente Morante). O nó central "EM" está conectado a nove nós principais, que por sua vez se ramificam em sub-nós secundários.

- TEMPO** (nó principal):
 - INTERVALO
 - HORA
 - DATA
 - TEMPO_CALEND
 - FREQUENCIA
 - GENERICO
 - DURACAO
 - CICLICO
 - PERIODO
- VALOR** (nó principal):
 - QUANTIDADE
 - CLASSIFICACAO
 - MOEDA
- OBRA** (nó principal):
 - ARTE
 - REPRODUZIDA
 - PLANO
- LOCAL** (nó principal):
 - FISICO** (nó secundário):
 - AGUACURSO
 - ILHA
 - PLANETA
 - REGIAO
 - RELEVO
 - AGUAMASSA
 - VIRTUAL** (nó secundário):
 - SITIO
 - COMSOCIAL
 - OBRA
 - PAIS
 - REGIAO
 - DIVISAO
 - RUA
 - CONSTRUCAO
- ORGANIZACAO** (nó principal):
 - INSTITUICAO
 - EMPRESA
 - ADMINISTRACAO
 - SUB
- PESSOA** (nó principal):
 - MEMBRO
 - CARGO
 - INDIVIDUAL
 - POVO
 - GRUPOIND
 - GRUPOMEMBRO
 - GRUPOCARGO
- ABSTRACAO** (nó principal):
 - VARIADO
 - IDEIA
 - NOME
 - ESCOLA
 - NOME
 - PLANO
- COISA** (nó principal):
 - SUBSTANCIA
 - OBJECTO
 - CLASSE
 - MEMBROCLASSE
- ACONTECIMENTO** (nó principal):
 - ORGANIZADO
 - EVENTO
 - EFEMERIDE

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.

Exemplo 19.1.

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.



Todo o material das duas edições do HAREM, incluindo programas de avaliação e corpora anotados, estão disponíveis¹⁷.

19.8 CLEF: Recuperação de Informação e QA (2003-2009)

Em 2003, a Linguatca 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 Linguatca 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 Linguatca 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 CHAVE¹⁸, para que os sistemas pudessem usá-los para treinamento.

¹⁷https://www.linguatca.pt/aval_conjunta/HAREM/

¹⁸<https://www.linguatca.pt/CHAVE/>



19.8.2 Resposta automática a perguntas (QA@CLEF)

A tarefa de resposta a perguntas¹⁹ 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 trilha²⁰ 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 quintupla 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 CHAVE²¹.

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 CHAVE²².

¹⁹No CLEF, o nome em inglês é *question answering*, QA.

²⁰Uma trilha (*track*) é uma tarefa específica dentro de uma avaliação conjunta maior.

²¹<https://www.linguateca.pt/CHAVE/>

²²<https://www.linguateca.pt/CHAVE/>



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 GikiP²³.

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 GIRA²⁵.

Das avaliações organizadas pela Linguatca, 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 Linguatca, 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.linguatca.pt/GikiP/>

²⁴Há dois padrões de escrita da língua norueguesa, cada um com a sua wikipédia, e daí a disparidade entre 9 e 10.

²⁵<https://www.linguatca.pt/GikiCLEF/GIRA/>



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 Linguatca 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.

²⁶Pseudo, porque não sabemos todas as respostas que se encontram na Wikipédia. Só conhecemos as que são certas e forem encontradas pelo conjunto de sistemas.

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 Cartola²⁷, 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 ASSIN²⁸.

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 técnica 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 jornalístico²⁹.

²⁷<https://www.linguateca.pt/Cartola>

²⁸<http://nilc.icmc.usp.br/assin/>

²⁹Seria interessante avaliar se a causa da maior simplicidade vinha sobretudo da diferença entre as línguas, ou da diferença entre os gêneros.



- 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 ASSIN2³⁰. 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/>



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 2019³¹.

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 públicos³², 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 IDPT2021³⁴.

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 SemEval³⁵, 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 ABSAPT2022³⁶.

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>



19.14 DIP: Desafio de Identificação de Personagens (2022-2023)

No âmbito do recente interesse pelas Humanidades Digitais, e mais especificamente pela leitura distante³⁷, a Linguatca, 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 DIP³⁸.

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
Morfolimpiadas	2003-4	7	https://www.linguatca.pt/Morfolimpiadas/
Primeiro HAREM	2004-6	10	https://www.linguatca.pt/primeiroHAREM/
Segundo HAREM	2007-8	10	https://www.linguatca.pt/HAREM/
CLEF RI	2003-7	34	https://www.linguatca.pt/CHAVE
QA@CLEF	2003-9	23	https://www.linguatca.pt/CHAVE
GeoCLEF	2005-8	10	https://www.linguatca.pt/CHAVE
GikiP	2008	3	https://www.linguatca.pt/GikiP/
GikiCLEF	2009	8	https://www.linguatca.pt/GikiCLEF/
Págico	2012	7	https://www.linguatca.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

³⁷Uma abordagem computacional dos estudos literários, iniciada por Moretti (2013) e chamada *distant reading* em inglês.

³⁸<https://www.linguatca.pt/DIP/>



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; Kilgarrieff (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 ConLL³⁹.

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, **avancar o estado da arte e da técnica** nessa área.

³⁹<https://www.conll.org/>



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.



Parte X

Desafios e Perspectivas



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: 20/11/2024

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 (2024), 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 Conjunto de dados, *dataset* e *corpus*, 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 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 2024, que são as redes neurais artificiais) é que aquilo que aprendem não serem recuperáveis 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 chamado *Deep Learning* (Goodfellow et al., 2016), que nada mais é do que um modelo especial de redes neurais. Desempenham muito bem, é verdade, mas não conseguimos entender como o fazem. Desse modo, a impossibilidade de se justificar, aliada à presença cada vez mais sentida 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 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 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 o foco tem sido evitar excessos regulatórios, caminhando para cultura da autorregulação, em que princípios são declarados e segui-los é responsabilidade das empresas. Porém, para isso há medidas concretas e rígidas sendo estabelecidas e se caminha para impor restrições e exigência de deveres para as grandes empresas. 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 primeira Lei da Inteligência Artificial codificada (*Artificial Intelligence Act*), uma lei baseada no risco, ou seja, quanto maior o risco de danos à sociedade, mais rigorosa será a lei. Embora prevista para entrar em vigor apenas em 2026, algumas proibições se antecipam: fica proibido usar sistemas para pontuação social, policiamento preditivo e coleta indiscriminada de imagens faciais disponíveis na internet ou capturadas por câmeras de circuito interno. Por exemplo, o uso de vigilância biométrica pelo governo se restringirá a casos que envolvam crimes, terrorismo e busca de suspeitos de crimes graves¹.

No Brasil, o mais recente projeto de lei (PL 2338/2023) encontra-se, em 2024, em tramitação na Câmara dos Deputados e no Senado Federal. O projeto estabelece “normas gerais de caráter nacional para o desenvolvimento, implementação e uso responsável de sistemas de inteligência artificial (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”². 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.

¹<https://artificialintelligenceact.eu/ai-act-explorer/>

²<https://legis.senado.leg.br/sdleg-getter/documento?dm=9347622&ts=1720062232582&disposition=inline>



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 [cap-ética-3].

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 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 reconhe-



cer 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 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.

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 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 em Modelos de linguagem. 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 no 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 algoritmos³. 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 (Conjunto de dados, *dataset* e *corpus*). 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 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.

³<https://www.facebook.com/help/contact/510058597920541>



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.



Capítulo 21

E agora, PLN?

Maria das Graças Volpe Nunes

Helena de Medeiros Caseli

Publicado em: 26/09/2023

Atualizado em: 29/09/2025

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 merece¹.

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 PROPOR², e mantêm vínculos acadêmicos há várias déca-

¹<https://www.publico.pt/2023/02/09/opiniao/opiniao/lingua-portuguesa-tecnologia-futuro-2038078>

²CE-PLN. PROPOR (*International Conference on Computational Processing of Portuguese Language*). Disponível em: <https://sites.google.com/view/ce-pln/eventos/propor>.



das. No Brasil, os recursos de PLN compartilhados pela comunidade distribuem-se pelos centros de pesquisa, sendo dois exemplos o NILC³ e o C4AI⁴. Em Portugal, dois importantes repositórios de recursos e ferramentas para português europeu e brasileiro são a Linguateca⁵ e o Portulan Clarin⁶.

Em países extensos como o Brasil, onde há uma grande variedade linguística, a exemplo das diferentes línguas indígenas faladas em território nacional⁷, 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 pesquisadores⁸. Já o *corpus* Carolina, do Centro de IA, C4AI⁹, 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 TaRSila¹⁰ 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, CORAA¹¹, 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 Hugging Faece.

No mesmo C4AI, o projeto PROINDL¹², 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>

⁴https://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.ufgrs.br/pln/wiki/index.php?title=BrWaC>

⁹<https://sites.usp.br/corpuscarolina/>

¹⁰TaRSila. Disponível em: <https://sites.google.com/view/tarsila-c4ai>.

¹¹<https://sites.google.com/view/tarsila-c4ai/coraa-versions>

¹²<https://conexoesoriginarias.github.io/>



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), Albertina¹³ e o Bode¹⁴). 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 estudo¹⁵ 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 TPUs¹⁶) 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 plano¹⁷ 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

¹³Família de modelos treinados para as variantes europeia 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>

¹⁶TPUs (Unidades de Processamento de Tensor) são aceleradores de treinamento e geração de modelos de machine learning.

¹⁷https://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



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 (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



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.



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.



- 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. Frequência de ocorrência dos fonemas 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 corpus 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.; SCHIÖNNING, 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



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



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>>

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**. Proceedings 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: MOTA, C.; SANTOS, D. (Eds.). **Desafios na avaliação conjunta do reconhecimento de entidades mencionadas.** [s.l.] Linguatca, 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 Translation 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 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. **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.

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 As-



sociation for Computational Linguistics. **Anais...Online: Association for Computational Linguistics**, jul. 2020. Disponível em: <<https://aclanthology.org/2020.acl-main.463>>

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 Linguística 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 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 Aarhus, 2000.

BICK, E. **A dependency-based approach to anaphora annotation**. Proceedings of the 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**, 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.
- 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 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/1457c0d6bfc4967418bfb8ac142f64a-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 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 Irecconf (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/i>>



ndex.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/proceedings/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 – Estudos. 141–292.

CASTILHO, A. T. DE. Gramática do Português Brasileiro: fundamentos, perspectivas. **Cadernos de Linguística**, v. 2, n. 1, p. e252, abr. 2021.

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 in 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 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 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.
- CLIFTON, A. et al. **100,000 podcasts: A spoken English document corpus**. Proceedings of the 28th International Conference on Computational Linguistics. **Anais...**2020.
- 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>
- 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



ces 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.

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 Linguistics, 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 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 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, 2016. 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/pdf2006/6.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/lrec2>>



004/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: Introducing 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>

ERYİĞİT, 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 Linguistics, nov. 2020. Disponível em: <<https://aclanthology.org/2020.emnlp-main.393>>

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 Linguística 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**. Proceedings 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 Idiomatcity 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.

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 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.] Linguatca, 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.
- 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.linguatca.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>>
- 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, 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.

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 (OASIs).Dagstuhl, Germany: Schloss Dagstuhl – Leibniz-Zentrum für Informatik, 2019. Disponível em: <<https://drops.dagstuhl.de/entities/document/10.4230/OASIs.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. 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.; TASLIMIPPOOR, 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**. 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>>

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. **Com-**



putational 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 (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 (TREC's): 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://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**, 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 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...: NLPPIR '24**. New York, NY, USA: Association for Computing Machinery, 2025. Disponível em: <<https://>>



[//doi.org/10.1145/3711542.3711570](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 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.

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, 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 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>>

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://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>

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.

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 inter-subjetividade 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 Linguistics, 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://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 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, 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 (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 Subjec-**



tive 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 de referência do português brasileiro falado informal.**, no prelo no 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: Firenze 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.). **Desafios na avaliação conjunta do reconhecimento de entidades mencionadas**. [s.l.] Linguatca, 2008. p. 55–75.
- MOTA, C. et al. **Págico: Evaluating Wikipedia-based information retrieval in Portuguese**. (N. Calzolari et al., Eds.) Proceedings of the Eighth 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 Linguística**, 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.

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.

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 Linguas 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 Algorithm 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.

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.

PĂIȘ, V.; TUFİȘ, 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>>

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.ea-cl-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 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>

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 unrestricted 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 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 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 trans-**



late 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 Linguas 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.

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**, 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 avaliação 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 interação 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>>
- ROTHER, 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. **Calidoscópico**, 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, 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/SantosSPROPOR2000.pdf>>

SANTOS, D. **Evaluation in natural language processing.**, c2007. Disponível em: <<http://www.linguateca.pt/Diana/download/EvaluationESSLI07.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. 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.

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://nlsreiter.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, 2019b/2019b.

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.

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 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 Balance-**



adas. 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 sumário automático de estruturas RST**. 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 Berkeley, 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.

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 Universal Dependencies**. 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 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 continuação a partir da linguística sistêmico-funcional. **EntreLetras**, v. 13, n. 1, p. 158–183, 2022.

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 sumariação 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 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.

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 Summit. 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.

TANAKA, E. et al. Cem Mil Podcasts: A Spoken Portuguese Document Corpus. **arXiv preprint arXiv:2209.11871**, 2022.

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.

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. **Éléments 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, 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 Linguísticos**, 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.

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: 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 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 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.

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 hierarchical prosody model**. 2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). **Anais...**2011. 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. **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.

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.



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

A.1 Alguns pressupostos:¹ estatística², 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.

¹Esta seção é amplamente baseada nas discussões em (Gries, 2019; Jurafsky; Martin, 2023)

²Recomendamos 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>)



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:

$$P(X = x) = P(s \mid X(s) = x)$$



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 $x_1, x_2, \dots, x_n$, ou no máximo, uma sequência infinita de valores diferentes $x_1, x_2, \dots$. 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 :

$$P_X(x) = P(X = x)$$

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



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.



- 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



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.



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.



Apêndice B

Diretrizes de avaliação

Apêndice do Capítulo 18

Brielen Madureira

Publicado em: 13/03/2024

 <https://brasileiraspln.com/livro-pln/4a-edicao/volume1/>

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 significativa? 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.?



- 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?



CRITÉRIOS PARA AVALIAÇÃO DO DESIGN DO EXPERIMENTO

1. Quantos exemplos podem ser demonstrados?
 - Eles são qualitativamente diferentes?
 - Esses exemplos ilustram todas as 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 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 previsões 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?



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?



Apêndice C

EAGLES

Apêndice do Capítulo 18

Brielen Madureira

Publicado em: 13/03/2024

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 papéis 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

¹Disponível em <https://www.issco.unige.ch/en/research/projects/eagles/ewg99/7steps.html>



- 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.



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?, 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 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 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 Resolução de correferência.

Email: alinevanin@ufcspa.edu.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 Expressões multipalavras.

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 Sequência de caracteres e palavras, A ordem e a função das palavras em uma sentença 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 A ordem e a função das palavras em uma sentença e Ferramentas e recursos para o processamento sintático.

Email: anapagano.ufmg@gmail.com

Brielen Madureira

Brielen Madureira é formada em Matemática Aplicada e Computacional pela 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 da Universidade de Leipzig (Alemanha), na área de análise de dados para discursos sobre o clima. Neste livro, colaborou com os Capítulos Avaliação de tecnologias de linguagem e Diálogo e Interatividade.

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 Texto ou fala?.

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 Expressões multipalavras. Email: carlos.ramisch@lis-lab.fr

Cláudia Freitas

Cláudia Freitas é formada em Letras pela PUC-Rio, com pós-graduação em Estudos da Linguagem pela mesma instituição. Foi docente da PUC-Rio de 2002 a 2023, e desde 2023 é pesquisadora do ICMC/USP, vinculada ao C4AI. Suas áreas de interesse incluem construção de *datasets/corpora* anotados para o aprendizado de máquina, extração de informação (em suas várias formas), e representações linguísticas para o aprendizado de máquina. Neste livro, colaborou com os Capítulos E o significado?, Conjunto de dados, *dataset* e *corpus*, Avaliação conjunta em português 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 Semântica distribucional e Extração de Informação.

Email: dclaro@ufba.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 Perguntas e Respostas.

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 Avaliação conjunta em português.

Email: dianamspsantos@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 Perguntas e Respostas.

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 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 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 Semântica com técnicas simbólicas e Semântica distribucional.



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 Resolução de correferência.

Email: metalmorphy@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 Texto ou fala?.

Email: hmello@ufmg.br

Ivandr  Paraboni

Ivandr  Paraboni   doutor em Ci ncia da Computa  o pela University of Brighton (Inglaterra, 2003), com est gio de p s-doutorado junto   University of Aberdeen (Esc cia, 2012).   professor associado (livre docente) e pesquisador da USP/EACH. Possui experi ncia na  rea de Ci ncia da Computa  o, atuando principalmente nas  reas de PLN e IA. Interesses de pesquisa abrangem o processamento da l ngua humana em um sentido amplo do termo, incluindo m todos computacionais derivados das Ci ncias Cognitivas e aplica  es pr ticas de classifica  o de documentos extra dos da WEB. Neste livro, colaborou com os Cap tulos Gera  o de linguagem natural e Detec  o de Transtornos de Sa de Mental a partir de Texto.

Email: ivandre@usp.br

Jackson Wilke da Cruz Souza

Jackson Souza   bacharel (2012), mestre (2015) e doutor (2019) em Lingu stica pela Universidade Federal de S o Carlos (UFSCar). Atualmente,   professor da Universidade Federal da Bahia (UFBA), atuando junto ao Departamento de Ci ncia, Tecnologia e Inova  o e ao Programa de P s-Gradua  o em L ngua e Cultura da UFBA. Tem interesse nas  reas de an lise de discurso, sumariza  o autom tica e an lise e descri  o lingu stica. Neste livro, colaborou com os Cap tulos Modelos discursivos e Sumariza  o Autom tica.

Email: jackcruzsouza@gmail.com



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 os Capítulos Semântica distribucional e Modelos de linguagem.
Email: jessica.rodrigues@oii.ox.ac.uk

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 Semântica distribucional.
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 Sequência de caracteres e palavras e Ferramentas e recursos para o processamento sintático.
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 Sequência de caracteres e palavras, PLN no Direito e PLN e Humanidades Digitais.
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 Questões éticas em IA e PLN.

Email: mariza@ic.uff.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 especial de doutorado na USP e atua como Cientista de Dados no Instituto de Pesquisas SiDi, contribuindo com projetos de PLN, especialmente no Samsung Search e na Bixby. Suas áreas de interesse incluem PLN, aprendizado de máquina e recuperação de informação. Neste livro, colaborou com o Capítulo Aprendizado Transdutivo em PLN.

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 Aprendizado Transdutivo em PLN.

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 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 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 Texto ou fala?.

Email: pritico@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 Resolução de correferência e PLN e Humanidades Digitais.

Email: renata.vieira@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 Expressões multipalavras.

Email: renata.ramisch@gmail.com

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 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 Modelos discursivos.

Email: roana@academico.ufs.br

Tayane Soares

Tayane Soares é Cientista de Dados na empresa Blip, formada em Letras pela UFMG e atualmente estudante de mestrado em Linguística Aplicada pelo PosLin-UFMG. Seu trabalho de pesquisa tem 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, especialmente no que diz respeito a aprendizado de máquina, tradução automática e questões éticas em IA e PLN. Neste livro, colaborou com o Capítulo Questões éticas em IA e PLN.

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 interesses em Engenharia de Software, Processamento de Linguagem Natural, Aprendizado



de Máquina, Modelos de Linguagem e Agentes. Neste livro, colaborou com o Capítulo Geração de linguagem natural.

Email: thiago.castro.ferreira@gmail.com

Valéria de Paiva

Valeria de Paiva é formada em Matemática pela PUC-Rio, com pós-graduação em Álgebra 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 Semântica com técnicas simbólicas.

Email: valeria@topos.institute

Vlória Pinheiro

Vlória 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 Semântica com técnicas simbólicas.

Email: vladiacelia@gmail.com



Uma iniciativa



brasileiraspln.com

Apoio

