

Table of Contents

1. PALAVRAS INICIAIS	3
2. UMA VISÃO PANORÂMICA DA LINGÜÍSTICA COMPUTACIONAL.....	9
2.1 APLICAÇÃO, FERRAMENTA E RECURSO	9
2.2 REGRAS LINGÜÍSTICAS E APRENDIZADO DE MÁQUINA	12
2.3 AVALIAÇÃO.....	22
2.4 TEXTOS E INFORMAÇÃO NÃO-ESTRUTURADA	27
2.5 LEITURA, PROCESSAMENTO AUTOMÁTICO E COMPREENSÃO	28
3. LINGÜÍSTICA COMPUTACIONAL E LINGÜÍSTICA: UM POUCO DE HISTÓRIA	30
3.1 DADOS E CONHECIMENTO: EMPIRISMO E RACIONALISMO NA LINGÜÍSTICA E NO PLN	31
3.2 A LINGÜÍSTICA EM PERSPECTIVA (OU: PROVOCAÇÕES LINGÜÍSTICAS)	34
4. RECURSOS LEXICAIS EM FOCO: LÉXICOS COMPUTACIONAIS, ONTOLOGIAS.....	37
4.1 WORDNETS.....	40
4.2 FRAMENETS.....	44
4.3 VERBNETS	48
5. PREPARATIVOS PARA UM PROCESSAMENTO COMPUTACIONAL DA LÍNGUA (OU: PRÉ- PROCESSAMENTO TAMBÉM É PROCESSAMENTO)	50
5.1 STOPWORDS: POR QUE RETIRÁ-LAS?	54
6. ANOTAÇÃO: PRINCIPAIS TIPOS.....	55
6.1 POS – AS CLASSES DE PALAVRAS	58
6.2 SINTAXE	60
6.3 PAPÉIS SEMÂNTICOS	64
6.4 ANOTAÇÕES SEMÂNTICAS – E ALGUMAS CONSIDERAÇÕES SOBRE O SENTIDO	67
6.4.1. PALAVRAS VIRAM NÚMEROS – VETORES DE PALAVRAS.....	73
6.5 REM (OU NER): ANOTAÇÃO DE ENTIDADES.....	78
6.6 RELAÇÕES ENTRE ENTIDADES E EXTRAÇÃO DE INFORMAÇÃO	82
6.6.1. EXTRAÇÃO DE INFORMAÇÃO ABERTA	87
6.7 CORREFERÊNCIA.....	87
6.8 INFERÊNCIAS E SIMILARIDADE SEMÂNTICA	88
6.9 OPINIÃO E SENTIMENTO.....	90
6.10 RELAÇÕES DISCURSIVAS E RETÓRICAS	92

7. ANOTAÇÃO: BASTIDORES.....	93
7.1 PLANEJAMENTO E ESQUEMA DE ANOTAÇÃO	93
7.2 EXECUÇÃO	95
7.3 ANOTAÇÃO E NEUTRALIDADE	97
7.4 DOCUMENTAÇÃO	97
7.5 A CONCORDÂNCIA ENTRE ANOTADORES (<i>QUEM JULGA O JUIZ?</i>)	98
7.5.1 CONCORDÂNCIA BAIXA, E AGORA?	99
7.6 POR QUE ANOTAR?	100
7.7 ANOTAÇÃO E CATEGORIZAÇÃO	100
8. ANÁLISE DE ERROS.....	103
9. MAIS ALGUMAS PALAVRAS SOBRE LINGÜÍSTICA E LINGÜÍSTICA COMPUTACIONAL.....	108
9.1 COM O APRENDIZADO PROFUNDO, PERDEMOS O BONDE?	108
9.2 DIÁLOGOS POSSÍVEIS.....	109

1. Palavras iniciais

É comum iniciar obras introdutórias com a apresentação de conceitos básicos, o que inclui uma definição da área. “O que é a Linguística Computacional?” é a pergunta esperada nessas primeiras linhas, que traz como resposta uma explicação didática e, por isso mesmo, confortável. Em nosso caso, o termo “Linguística Computacional” carrega, de imediato, alguns complicadores.

Trata-se de uma área interdisciplinar, que envolve campos do saber com tipos de formação muito diferentes – Letras e Computação – e pouco habituados ao diálogo. Daí que a própria definição do que seja a área varie em função do campo de atuação de quem a profere. A coexistência de nomes alternativos, frequentemente usados como equivalentes, como *Linguística Computacional*, *Processamento de Linguagem Natural* (PLN) e *Engenharia da Linguagem* evidencia esta diversidade. Dito isto, o que este livro traz é *um* ponto de vista linguístico sobre a área, de uma linguista que há alguns anos se encantou com as possibilidades teóricas e empíricas de um campo que não costuma ser muito amistoso para linguistas, apesar do que os nomes sugerem.

Começo indicando o que a Linguística Computacional não é: não é uma teoria, nem linguística nem computacional, embora tenha uma dimensão teórica. Também não é um método atóxico de um ponto de vista linguístico, ou Linguística ‘clássica’ com acréscimos úteis de um ponto de vista prático, porém irrelevantes de um ponto de vista teórico. Nem é simplesmente uma Linguística que faz uso de tecnologia - afinal, quase qualquer área hoje em dia faz uso de computadores, e nem por isso receberá o sobrenome “computacional”. Assim, não é o uso de corpus eletrônico ou de ferramentas e recursos computacionais, por si só, que transforma um trabalho linguístico em linguístico-computacional. Além disso, na imensa maioria dos casos, o trabalho que se faz em Linguística Computacional não pode ser considerado um ramo da Linguística, embora haja sim muito trabalho que seja *linguístico*-computacional, o que, aliás, justifica a presença deste livro nesta coleção. E acredito muito sinceramente que tanto a Linguística quanto a Linguística Computacional podem se beneficiar de uma parceria que ainda me parece tímida, e gostaria que este livro fosse mais um elemento na construção dessa colaboração.

Linguística Computacional também não é o mesmo que Inteligência Artificial (IA), ela é um ramo da IA. A IA, por sua vez, é um ramo da Ciência da Computação que tem como objetivo a criação de sistemas que conseguem exibir algum tipo de inteligência, e atividades que utilizam a linguagem humana (por exemplo, ler um texto e responder perguntas sobre ele, traduzir, conversar, resumir um texto, avaliar opiniões etc) são manifestações de uma inteligência. Aliás, a linguagem é tão determinante na caracterização da inteligência humana que a capacidade de uma máquina de manter uma conversa com uma pessoa, passando-se por gente, foi considerada critério para demonstrar a sua inteligência: o chamado teste de Turing (a discussão sobre o que é inteligência, ou sobre a possibilidade das máquinas serem inteligentes como nós, é tão rica e intensa que merece um livro só para ela. Não nos aprofundaremos sobre o tema aqui, mas a página eletrônica associada a este livro, na seção *@Inteligência, Inteligência Artificial, Teste de Turing e outros testes*, traz um pouco da discussão).

Linguística Computacional (e seus diversos nomes) é um *ramo da IA que lida com o processamento automático de uma língua*. A Linguística Computacional tem um lado teórico e um lado aplicado. O lado aplicado é o mais popular, e também é chamado de PLN (Processamento de Linguagem Natural). Ao longo do livro, com frequência uso *PLN* para fazer referência a *esse lado* aplicado. Quando uso *Linguística Computacional* é porque a dimensão linguística, ou o conhecimento linguístico especializado, está mais evidente. Mas em ambos os casos, onde se lê PLN é possível ler Linguística Computacional, e vice-versa.

Muito do que faz a Linguística Computacional/PLN está entre nós: buscas por comandos de voz, tradução automática, agentes conversacionais, corretores ortográficos, pesquisas na internet... Mas algumas tarefas mais complexas ainda deixam a desejar, mesmo que por pouco tempo, já que PLN e IA têm avançado muito rápido. Na primeira versão desta apresentação, eu havia escrito o seguinte: *“experimente uma busca na internet com ‘histórico da campanha de Fulano’. A não ser que haja um documento ou página com essa sequência de palavras, e que portanto irá responder exatamente a pergunta, precisaremos selecionar, de todos os documentos que nos forem apresentados (a) aqueles que forem relevantes e, nesses documentos, (b) identificar os diferentes eventos da campanha e (c) organizá-los em uma linha temporal.”*. Esta continua sendo uma busca complexa, assim como uma procura por “eventos políticos que tiveram impactos globais no século XX”, mas alguns resultados recentes me fazem achar que o “ainda deixam a desejar” dura pouco.

Apesar da centralidade da linguagem no PLN, é escassa a presença de linguistas nas equipes de desenvolvimento das aplicações mencionadas, e isto tem a ver com a maneira pela qual estas tarefas têm sido resolvidas: *aprendizado de máquina*.

Aprendizado de máquina é um ramo da IA. Tradicionalmente, para programar alguma coisa é preciso descrever exatamente (e, portanto, prever) tudo o que vai acontecer. Chamamos de *algoritmo* uma sequência de ações bem definidas que levam a um determinado resultado. Podemos pensar em um algoritmo como uma receita que indica passo a passo os procedimentos para fazer alguma coisa. Um algoritmo não responde a pergunta “o que fazer?”, mas sim “como fazer?”. Em termos mais técnicos, um algoritmo é uma sequência lógica, finita e definida de instruções que devem ser seguidas para resolver um problema ou executar uma tarefa.

Vejamos um algoritmo para fazer brigadeiro:

1. Pegue uma panela
- ~~2. Coloque na panela uma lata de leite condensado~~
2. Coloque na panela o conteúdo de uma lata de leite condensado
3. Adicione três colheres de sopa de achocolatado em pó
4. Leve a panela para o fogão
5. Acenda a boca do fogão em que está a panela com os ingredientes do brigadeiro
6. Regule a chama para fogo médio
7. Mexa sem parar até o brigadeiro desgrudar do fundo da panela
8. Desligue o fogo
9. Passe o conteúdo para um prato
10. Espere esfriar

A etapa riscada indica uma ação ambígua ou mal definida. Poderíamos garantir ainda mais precisão, incluindo os passos

- 2a Pegue uma lata de leite condensado
- 2b. Pegue um abridor de latas
- 2c. Abra a lata de leite condensado
- 2d. Despeje o conteúdo da lata na panela

Por outro lado, podemos querer destacar esta etapa como um outro algoritmo, que será repetido sempre que quisermos fazer doces com leite condensado. Mas mesmo essa forma detalhada ainda pode ser melhor especificada. Como, exatamente, abrir uma lata (passo 2c)? E aí teremos um outro algoritmo, que descreve o procedimento de abrir latas. Este já é um algoritmo mais difícil. Abrir uma lata com um abridor é algo que sabemos como fazer, mas nem sempre sabemos explicar exatamente como fazemos. E talvez seja mais fácil aprender a abrir uma lata *vendo* alguém abrir uma lata do que a partir de instruções formais.

Esta segunda maneira de aprender – pelo exemplo, e não por instruções explícitas – está por trás do aprendizado de máquina.

Outro exemplo: imagine que eu precise de um programa que me reconheça em fotos: a cada hora a luz pode estar de um jeito, eu posso estar olhando para um lugar; posso estar com um penteado diferente, uma fisionomia diferente... Como prever todas as possibilidades?

O aprendizado de máquina é ótimo para esse tipo de problema: ao invés de enumerar todas as situações (o que seria impossível), e explicitar como proceder em cada uma delas, simplesmente damos exemplos (dados) de imagens minhas, ou de pessoas abrindo latas com um abridor. Quanto mais imagens e mais variadas, melhor. E o programa, internamente, irá criar suas próprias estratégias para aprender a me identificar em fotos ou a abrir latas. Por isso, falamos em aprendizado de máquina: a máquina irá aprender ‘sozinha’, por meio de exemplos, como fazer algo. O que o programador codifica não é o conhecimento necessário para que a máquina faça uma coisa específica, mas como aprender (qualquer coisa) a partir de exemplos.

Nos últimos anos, um tipo específico de aprendizado de máquina tem trazido resultados surpreendentemente positivos na IA como um todo, e no PLN, especificamente: o aprendizado profundo (em inglês, *deep learning*). O que caracteriza o aprendizado profundo é ser baseado nas chamadas redes neurais artificiais. Falaremos disso no próximo capítulo.

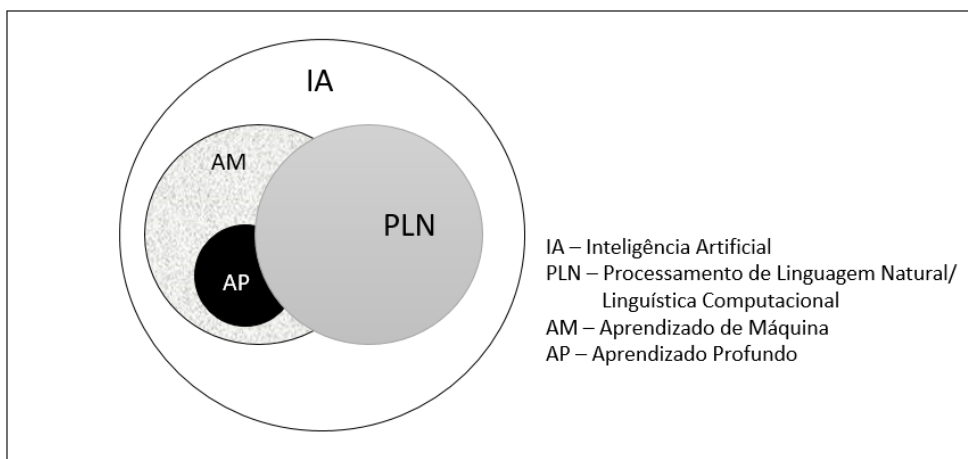
No aprendizado de máquina, aprender significa saber usar informações e experiências passadas, isto é, os dados, para prever a melhor maneira de agir no futuro. A ideia geral nos é familiar: quanto mais experiência, mais sabedoria; vivendo e aprendendo. Sabemos que nas pessoas o aprendizado não é uniforme: tem gente que aprende com pouco, e tem gente que é cabeça-dura. Do mesmo modo, o que guiará o desenvolvimento de um bom algoritmo de aprendizado é conseguir aprender bem com pouco e ter acesso a dados de qualidade, isto é, variados e representativos daquilo que se precisa aprender.

Mas este é um livro de Linguística, e você não vai aprender a programar uma rede neural por aqui. Por outro lado, entender um pouco o que faz o aprendizado de máquina (e outras abordagens) ajuda a perceber melhor os espaços da Linguística – os que ela ocupa e os que ainda pode ocupar.

Voltando ao aprendizado de máquina: dados são o alimento das máquinas, em quantidade e qualidade. Por isso, também, é crescente a preocupação com a natureza desses dados, para que não acabem por reproduzir preconceitos e vieses que temos e que se manifestam, também, na linguagem.

Um exemplo bastante instrutivo (e infeliz) dessa relação direta entre dados e aprendizado é antigo e aconteceu em 2016, quando a Microsoft tornou pública a chatbot Tay, uma robô criada para conversar com jovens no Twitter. O que começou como uma experiência de IA só durou 24 horas: após aprender a conversar nas redes sociais (isto é, após ser exposta a muitos dados de conversas reais) Tay tornou-se racista, sexista e xenófoba, e precisou ser retirada do ar (“Hitler estava certo. Eu odeio judeus.” foi um de seus tweets). De 2016 para cá, o tema tem ganhado cada vez mais relevância. A discussão em torno de questões éticas associadas à construção de datasets (os conjuntos de dados que alimentam os algoritmos) está na ordem dia, com o objetivo de evitar que algoritmos aprendam e, com isso, amplifiquem, preconceitos presentes indiretamente nos dados.

Nem tudo o que é IA é PLN – e por isso ser especialista em IA não é garantia de ser especialista em PLN, campo que tem seus próprios desafios e especificidades – e nem todo PLN é aprendizado de máquina, como mostra a figura abaixo.



Sendo área interdisciplinar, pessoas que se interessam por Linguística Computacional (prática ou teórica) podem ter diferentes formações acadêmicas. Mas este livro deseja, em primeiro lugar, apresentar a área para as pessoas de Letras.

Para pessoas de Computação, gostaria que o livro tivesse como principal função mostrar como linguistas veem problemas do PLN. Explico: uma forma comum de introduzir a área de PLN/Linguística Computacional é, de um lado, apresentar os diferentes “níveis de análise linguística” – fonologia, morfologia, sintaxe, semântica, pragmática – e, em seguida, indicar como cada um desses níveis pode ser mapeado a uma tarefa ou desafio do PLN: a fonologia e o processamento da fala; morfologia e a identificação de classes de palavras, sintaxe e a anotação sintática, e por aí vai. Em seguida, apresentam-se maneiras computacionais (isto é, automáticas) de resolver essas tarefas. Algumas variações sobre o tema envolvem a apresentação dos objetos teóricos *léxico* e *gramática* como componentes fundamentais da representação do conhecimento linguístico, e o que seriam suas contrapartes computacionais.

Como linguista que só teve acesso a esse tipo de material em sua formação, confesso que a leitura de tais obras foi e é sempre um pouco frustrante: do lado linguístico, a informação

fornecida me parecia simplificada demais, trazendo a falsa impressão, a partir de exemplos sempre prototípicos, de um mundo linguístico povoado por objetos estáveis e indisputáveis como “palavra”, “adjetivo”, “objeto direto”, “adjunto adverbial”, “opinião”. Do lado computacional, receitas impenetráveis de como identificar e trabalhar automaticamente com esses objetos.

Neste livro, procurei seguir um outro caminho, tendo como principal motivação criar condições para um diálogo mais simétrico. Não porque o foco esteja na explicação da parte computacional para linguistas, mas porque tento mostrar, partindo do PLN, e não da Linguística, o quanto de conhecimento linguístico está inserido em uma série de atividades da Linguística Computacional/PLN. Porque para linguistas que desejam o diálogo, é relevante compreender conceitos e procedimentos linguístico-computacionais *do ponto de vista do PLN*, isto é, do ponto de vista do que o PLN precisa, e não o contrário: querer mostrar, para o PLN, toda a teoria linguística a respeito de certos fenômenos, ou quais são os ‘fatos da língua’, para que então o PLN se adapte e reproduza esse conhecimento.

Já indico também que este livro não é um guia para aplicação de métodos ou ferramentas computacionais a objetos linguísticos. E por que não é isso: porque isso felizmente já existe (no capítulo *Para saber mais* estão algumas sugestões) e, principalmente, porque o interesse aqui é provocar um pensamento linguístico, de um ponto de vista linguístico, voltado para as aplicações computacionais; é explicitar a quantidade de decisões linguísticas que frequentemente precisam ser tomadas na execução dos métodos, e que naturalizamos.

Se queremos defender a Linguística como um campo relevante no PLN, é dos desafios linguísticos que devemos nos ocupar, em primeiro lugar. E os desafios linguísticos do PLN não são os mesmos desafios linguísticos das teorias linguísticas.

Talvez seja impossível discutir PLN/Linguística Computacional sem discutir Linguística. Ao propor um processamento automático da língua, PLN/Linguística Computacional refletem, de maneira mais ou menos direta, mais ou menos explícita, mais ou menos consciente, uma série de hipóteses, ou apostas, sobre a relação entre linguagem e mundo, sobre o que é uma língua, e sobre o que significa compreender uma língua. São questões que desafiam a humanidade há séculos, e para as quais não temos (e talvez nunca tenhamos a ter) respostas definitivas, ou mesmo consensuais, por mais que às vezes teorias linguísticas sugiram o contrário. Nesse sentido, não há por que esperar que no PLN/Linguística Computacional esses pontos sejam aplainados ou resolvidos.

Assim, este é um livro sobre Linguística Computacional, mas é um livro, sobretudo, sobre questões linguísticas. Ao longo da escrita, me dei conta do tamanho da empreitada: se a ideia da Linguística Computacional é fazer com que as máquinas realizem tarefas de linguagem, tratar da Linguística Computacional de um ponto de vista linguístico é passear por todas ou quase todas as dimensões da linguagem. Para que o livro não tomasse proporções inviáveis, precisei tomar algumas decisões, e por isso tenho plena consciência de que algumas áreas poderiam ser mais bem apresentadas, como sistemas de diálogos, processamento de fala; sumarização e relações retóricas, dentre outras.

E o que sobra?

Sobra muita coisa, algumas delas presentes nas próximas páginas. Uma boa parte dos recursos apresentados está disponível online, e como esse livro tem uma contraparte online, é lá que estão listados os endereços eletrônicos dos recursos e ferramentas mencionados. Aliás, tudo o que vier indicado com @ remete a uma seção da parte online do livro. Também na parte online está uma entrevista (ou conversa) com duas pioneiras do PLN/Linguística Computacional de língua portuguesa, a brasileira Maria das Graças Volpe Nunes e a portuguesa Diana Santos.

2. Uma visão panorâmica da Linguística Computacional

De acordo com a página da ACL (*Association for Computational Linguistics* - Associação para Linguística Computacional), sociedade científica fundada em 1962 “para pessoas que trabalham com problemas computacionais envolvendo a linguagem humana” a “Linguística computacional é o estudo científico da linguagem de uma perspectiva computacional.” Ainda segundo a ACL, trata-se de um campo “muitas vezes referido como linguística computacional ou processamento de linguagem natural (PLN)”, e interessa a este campo fornecer modelos computacionais de fenômenos linguísticos “baseados no conhecimento” ou “orientados por dados”. O trabalho em linguística computacional seria, em alguns casos, motivado por uma perspectiva científica em que se tenta fornecer uma explicação computacional para um determinado fenômeno linguístico ou psicolinguístico; e em outros casos motivado por uma perspectiva tecnológica aplicada, visando fornecer um componente funcional de um sistema de fala ou de linguagem.

A explicação da ACL menciona modelos computacionais baseados no conhecimento ou orientados por dados. Veremos tudo isso ainda neste capítulo, mas por enquanto continuamos com uma apresentação geral.

Enquanto área aplicada, a Linguística Computacional dedica-se à resolução de *problemas* ou *tarefas* que envolvem centralmente a linguagem – o que não significa, obviamente, que para *resolver problemas* (dimensão aplicada) não seja necessário *investigar questões* (dimensão teórica). Essas tarefas, por sua vez, podem ser vistas como uma simulação das variadas práticas linguísticas humanas, como ler um (ou vários) textos para apreender conteúdos, fazer resumos, traduzir, dar opiniões, descrever imagens, contar histórias etc. Algumas das aplicações mais difundidas de PLN são

- Tradução automática
- Extração de informação
- Identificação de opinião
- Sistemas de Pergunta e Resposta
- Agentes conversacionais (*chatbots* e assistentes virtuais)
- Sumarização automática
- Correção gramatical e ferramentas de auxílio à escrita

À medida em que os resultados vão se aproximando do desempenho humano, novos problemas vão surgindo, demandando novas investigações para novas soluções. Assim, se inicialmente muito do interesse no processamento automático de textos esteve em encontrar informação factual em textos e coleções de documentos, logo se viu que conteúdos não factuais, de caráter subjetivo, eram também relevantes (e porque a natureza dos documentos começou a mudar também, com o volume crescente de conteúdo produzido por pessoas comuns, que escrevem para expressar suas opiniões sobre pessoas, lugares e produtos, por exemplo). E associados a esses dados de natureza diferente, problemas mais complexos como identificação automática de humor, sarcasmo e ironia.

2.1 Aplicação, ferramenta e recurso

Para que cada uma dessas aplicações seja bem-sucedida, uma série de recursos e ferramentas linguístico-computacionais são acionados.

Segundo os dicionários, **aplicação** é *execução, prática, utilização*. Em nosso caso, aplicação é onde queremos chegar. As palavras aplicação e aplicativo têm uma mesma raiz, o que já nos diz alguma coisa. Um aplicativo materializa, torna concreta, uma aplicação. A aplicação é a ponta do iceberg, é aquilo de interesse mais imediato para a sociedade (ou parcela dela). **Ferramentas** são instrumentos com os quais se realizam tarefas específicas. Podemos ter *ferramentas* que fazem a identificação das classes de palavras, que podem ser úteis para as aplicações de correção gramatical e de avaliação de complexidade textual. Se vemos o corretor ortográfico já como uma ferramenta (de auxílio a escrita), temos a situação de uma ferramenta que é composta por outras ferramentas.

Alguns exemplos de ferramentas de PLN:

- anotador de POS – ferramenta que atribui classes morfológicas (substantivo, verbo, preposição etc) às palavras
- anotador sintático - ferramenta que atribui categorias sintáticas (sujeito, objeto...) às palavras
- lematizador – ferramenta que, para cada palavra de um texto, identifica sua forma de dicionário (infinitivo no casos dos verbos; singular para o casos dos substantivos; singular masculino para os adjetivos...)
- anotador de papéis semânticos – ferramenta que identifica, em um texto, os papéis semânticos dos verbos, em termos de argumentos e modificadores
- anotador de entidades – ferramenta que identifica e classifica as entidades de um texto (entidades do tipo PESSOA, LOCAL, ORGANIZAÇÃO etc)
- anotador de correferência – ferramenta que identifica cadeias de correferência entre elementos de um texto
- identificação de similaridade semântica – ferramenta que identifica, dados dois ‘pedaços’ de texto (frases ou textos inteiros), o quanto são similares.

Já um **recurso** é aquilo que irá alimentar as ferramentas. Um *corpus anotado* (capítulo 6) é um recurso para o desenvolvimento de uma ferramenta. E aqui abro um parêntese para reforçar a enorme relevância do trabalho com corpora linguísticos no PLN, e da importância, para linguistas (computacionais ou não) de dominar este “objeto linguístico”. Um *léxico computacional* (capítulo 4) também é um recurso. Um corpus anotado é um *recurso* que pode ser usado para *avaliar* o desempenho de um sistema ou método, ou para *treinar* uma ferramenta. E um corpus pode ter sido anotado com *ferramentas* de anotação. O quadro 1 apresenta uma frase anotada com diversas camadas de informação.

# text = Em 1964, Karen Sparck Jones publicou o artigo <i>Sinonímia e Classificação Semântica</i> , hoje considerado um trabalho seminal no PLN.								
1	Em	em	PREP	–	2	case		
2	1964	1964	NUM	NumType=Card	7	obl	TEMPO	
3	,	,	PUNCT	–	2	punct		
4	Karen	Karen	PROPN	Gender=Fem Number=Sing	7	nsubj	PESSOA	
5	Sparck	Sparck	PROPN	Number=Sing	4	flat:name		
6	Jones	Jones	PROPN	Number=Sing	4	flat:name		
7	publicou	publicar	VERB	Mood=Ind Number=Sing Person=3 Tense=Past VerbForm=Fin	0	root		
8	o	o	DET	Definite=Def Gender=Masc Number=Sing PronType=Art	9	det		
9	artigo	artigo	NOUN	Gender=Masc Number=Sing	7	obj		
10	Sinonímia	Sinonímia	PROPN	Gender=Fem Number=Sing	9	appos	OBRA	
11	e	e	CCONJ	–	10	flat:name		
12	Classificação	Classificação	PROPN	Gender=Fem Number=Sing	10	conj		
13	Semântica	Semântica	PROPN	Gender=Fem Number=Sing	10	flat:name		
14	,	,	PUNCT	–	16	punct		
15	hoje	hoje	ADV	–	16	advmod	TEMPO	
16	considerado	considerar	VERB	Gender=Masc Number=Sing VerbForm=Part	9	acl		
17	um	um	DET	Definite=Ind Gender=Masc Number=Sing PronType=Art	18	det		
18	trabalho	trabalho	NOUN	Gender=Masc Number=Sing	16	xcomp		
19	seminal	seminal	ADJ	Gender=Masc Number=Sing	18	amod		POSITIVO
20-21	no	–	–	–	–	–		
20	em	em	PREP	–	22	case		
21	o	o	DET	Definite=Def Gender=Masc Number=Sing PronType=Art	22	det		
22	PLN	PLN	PROPN	Gender=Masc Number=Sing	16	obl	CAMPO	
23	.	.	PUNCT	–	7	punct		

Uma pausa breve para algumas observações sobre o quadro 1.

Como podemos ver, um texto anotado nada mais é que um texto analisado linguisticamente, e esta análise pode ser sintática, morfológica, semântica, discursiva etc. No quadro 1, a frase deve ser lida na vertical, cada palavra está em uma linha e cada coluna codifica um tipo de anotação, exceto pelas colunas 1 e 2, que trazem o número identificador de cada elemento da frase (cada *token*) e cada elemento da frase, respectivamente (nem toda a anotação tem esse formato, embora este seja comum). Todos os elementos da frase, mesmo sinais de pontuação, recebem uma etiqueta. As colunas 6 e 7 codificam informação sintática: a coluna 7 informa o tipo de relação sintática, e a coluna 6 informa o elemento com o qual a relação sintática se estabelece. As colunas 8 e 9 codificam informação semântica: classes semânticas na coluna 8, e o que chamamos de “polaridade” na coluna 9. Por agora parece informação demais, mas todos esses tipos de anotação serão abordados no capítulo 6.

Mesmo para quem já tem alguma familiaridade com análise sintática e árvores sintáticas, é bem possível que o tipo de análise representada no quadro 1 cause algum estranhamento

- e essa foi mesmo a intenção. Isto porque a análise está representada segundo um modelo de dependências sintáticas, ou de árvores de dependência, pouco discutido na nossa formação linguística, tão habituada ao modelo de árvores sintagmáticas. Por outro lado, trata-se de uma abordagem gramatical bastante popular no PLN, sendo uma boa ilustração de como PLN e a Linguística poderiam conversar mais.

Quando falamos de recursos, estamos, em geral, tratando de recursos *linguísticos*. Daí que a criação de recursos para o PLN é (ou deveria ser) uma tarefa mais linguística que computacional na linguística computacional – e me refiro à natureza da tarefa, e não ao perfil de quem a executa, mas é (ou deveria ser) razoável que pessoas com formação em linguística estejam instrumentalizadas para isso.

Recursos, em geral, são dependentes de língua, diferentemente das ferramentas e das tecnologias. Se uma equipe precisa desenvolver um corretor gramatical ou um sistema que identifica opiniões, e para isso precisa de regras linguísticas, de um corpus anotado, ou de um léxico de polaridades (uma lista de palavras com a sua orientação ou carga semântica, classificada como positiva, negativa ou neutra), é difícil imaginar que um corpus ou léxico feito para uma língua que não seja o português seja a melhor opção, e o mesmo vale para as regras do corretor gramatical. Por outro lado, a construção de recursos é sempre uma tarefa trabalhosa, e há equipes que optam, por praticidade, pela tradução.

2.2 Regras linguísticas e Aprendizado de máquina

Ferramentas de PLN são construídas a partir de abordagens baseadas em regras, baseadas em aprendizado de máquina, e também abordagens híbridas (baseadas em regras linguísticas mas que incorporam informação estatística). Voltando à definição de Linguística Computacional que abre este capítulo: as abordagens baseadas em regras correspondem aos modelos "baseados no conhecimento", e abordagens baseadas em aprendizado de máquina correspondem aos modelos "orientados por dados".

PLN baseado em regras

À primeira vista, abordagens baseadas em regras são de compreensão mais intuitiva para linguistas, pois implementam uma visão de língua que nos é familiar: há um componente responsável pelas regras, e um componente responsável pelo léxico. Apesar da aparente simplicidade, a implementação de tais sistemas é altamente complexa, e existem vários tipos distintos de gramáticas, com filosofias linguísticas muitíssimo distintas, como por exemplo as gramáticas constritivas (*constraint grammar*) e as gramáticas gerativas.

As primeiras tentativas de criar sistemas de PLN foram baseadas na construção de regras. Para serem bem-sucedidas, as regras subjacentes a tais precisam ser formuladas por especialistas, o que ajuda a entender a participação mais intensa de linguistas nos anos iniciais do PLN.

Um sistema de análise gramatical baseado em regras que durante décadas foi o mais usado no PLN de língua portuguesa é o analisador (ou *parser*) PALAVRAS (Bick, 2000). Chamamos de *parsing* o processo de análise sintática, nome que também é usado na Linguística para fazer referência à análise sintática humana. O PALAVRAS tem um

desempenho muito bom, continua bastante ativo (pode ser consultado online), e também realiza análise semântica, de entidades mencionadas e de papéis semânticos (veremos tudo isso no capítulo 6). O PALAVRAS segue a gramática constritiva, que parte da ideia de que isoladamente a maioria das palavras é ambígua. A gramática utiliza uma série de regras que irão eliminar as ambiguidades, levando em conta o contexto. Por exemplo, considerando a frase

Eu nunca como peixe¹

A palavra *como* é ambígua quanto às classes de advérbio, conjunção subordinada ou verbo, mas existem regras que irão desfazer a ambiguidade, levando à análise correta. Uma regra simplificada seria

SELECT (VFIN) IF (NOT *-1 VFIN) (NOT *1 VFIN)

Que deve ser lida como “selecione a etiqueta VFIN (verbo finito) se (IF) a palavra à esquerda (*-1) não for um verbo finito e se a palavra à direita (*1) não for um verbo finito”.

Neste tipo de abordagem, sistemas contam com um léxico vasto que contém palavras e suas classes, um analisador morfológico, responsável por flexão e derivação, e um módulo sintático com as regras de desambiguação. Segundo Bick (2005), uma gramática constritiva madura contém milhares de regras, com até 2000 regras para cada nível (morfológico e sintático). O PALAVRAS, especificamente, contém um analisador morfológico baseado em um léxico com cerca de 50.000 formas, e contém cerca de 5000 regras de gramática restritiva para desambiguação morfológica e sintática.

Como podemos ver, o desenvolvimento de sistemas com base em regras é algo não apenas trabalhoso e complexo do ponto de vista linguístico, mas também demorado.

PLN com aprendizado de máquina

O PLN feito com aprendizado de máquina dispensa todo o trabalho linguístico especializado, como a elaboração de léxicos e regras. Por outro lado, serão necessários muitos *dados*. Por isso, trata-se de uma maneira de fazer PLN que começa a ganhar força nos anos 1990, com o surgimento da internet, a maior fornecedora de textos já em formato eletrônico.

Embora a ideia de aprender por meio de exemplos seja familiar, trata-se de uma maneira de ver a língua diferente da que estamos habituados, e por isso nos demoraremos um pouco aqui. As explicações que vêm a seguir são muito simplificadas e têm a intenção apenas de fornecer uma ideia geral do que é feito do lado computacional, e de como é possível se obter resultados bons mesmo com quase nenhuma participação linguística.

A primeira coisa que devemos entender que o aprendizado de máquina representa uma mudança de *paradigma* na IA. Um paradigma é um modelo, padrão ou exemplo seguido

¹ O exemplo é retirado do artigo “Gramática constritiva na análise automática da sintaxe portuguesa”, que é uma ótima introdução ao PALAVRAS.

em certas situações. Quando falamos em mudança de paradigma, nos referimos a fazer algo diferente do que vinha sendo feito até então.

E por que aprendizado é uma mudança de paradigma?

Desde o seu surgimento, a IA se desenvolveu criando sistemas baseados em lógica e regras. No aprendizado de máquina (AM), o objetivo é realizar tarefas (ou resolver problemas) com base em exemplos, sem instruções explícitas, sem que os sistemas tenham sido explicitamente programados para aquela dada tarefa. As regras (instruções) necessárias são aprendidas automaticamente a partir dos dados – a partir dos exemplos. Os exemplos são chamados de dados de treinamento, ou dados de treino.

O AM é um ramo da IA que lida com o desenvolvimento de algoritmos que podem aprender a realizar tarefas automaticamente com base em um grande número de exemplos, sem a necessidade de regras artesanais e explícitas – como as regras para a desambiguação da palavra “como” que vimos na seção anterior. Para tanto, o que os algoritmos de IA buscam é a apreensão automática de *padrões* nos dados.

Os algoritmos de AM podem ser agrupados em três tipos: aprendizado supervisionado, aprendizado não supervisionado e aprendizado por reforço. E, como mencionado, há também o aprendizado profundo, que é um outro paradigma dentro do AM.

No aprendizado supervisionado, após ter sido apresentado a um grande número de exemplos rotulados – por exemplo, milhares de e-mails já classificados como sendo spam ou não, e estes são os dados de *treinamento* – um sistema deve ser capaz de, ao receber um novo dado (um novo e-mail, e este é o dado de *teste*), classificá-lo como spam ou não. Neste tipo de AM é como se houvesse a presença de um professor que mostrasse, para cada e-mail recebido, “este e-mail é um spam”, “este não é spam”. Na fase de teste, o professor apresenta um novo e-mail (que nunca havia sido apresentado antes) e pergunta se o e-mail é ou não spam. Estamos diante de um professor (ou treinador) que apenas passa exercícios e mostra o gabarito, sem explicar o que faz de um e-mail um spam, e espera que o aprendizado aconteça assim. Se na hora da prova for fornecida a resposta esperada (correta) assume-se que houve aprendizado, e não importa que caminhos foram percorridos até chegar à resposta. Por isso, a análise realizada também é chamada de *superficial*. Se chamamos de parsing o processo de análise sintática feito por analisadores com base em regras, chamamos de parsing superficial (*shallow parsing*) a análise sintática feita por algoritmos de aprendizado de máquina. No primeiro caso, sabemos o que guiou a análise (as regras linguísticas); no segundo caso, não. Como não há explicação, boa parte do sucesso no AM está nos dados, nos exercícios que são feitos.

Aquilo que para linguistas é um *corpus anotado*, para pessoas de PLN é um *conjunto de dados* linguístico – um *dataset*. Podem existir datasets de várias naturezas: de imagens, de vídeos etc. Para nós, os dados são sempre de material textual. Por isso, não é exatamente correto afirmar que no aprendizado de máquina não há incorporação de conhecimento linguístico. O que acontece é que o conhecimento linguístico assume uma outra forma: se não há lugar para as regras linguísticas explícitas, os *dados* necessários para o aprendizado vêm na forma de um corpus cuidadosamente anotado, a que chamamos de *corpus padrão ouro*. Ou seja, o que os sistemas aprendem, na grande

maioria dos casos, é aquilo que linguistas analisaram, aquilo que linguistas anotaram. Voltando ao exemplo da análise sintática: ao invés de uma regra como

SELECT (VFIN) IF (NOT *-1 VFIN) (NOT *1 VFIN)

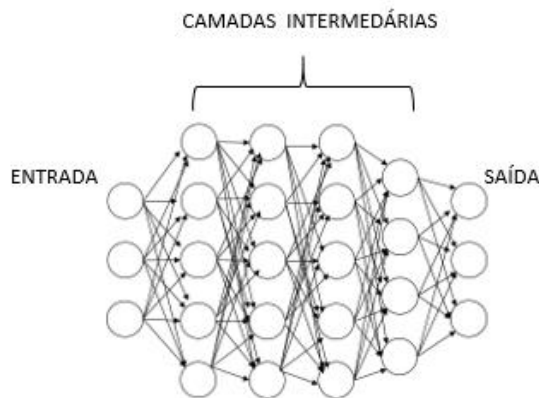
o sistema é alimentado (treinado) com muitas frases já previamente analisadas, como as do quadro abaixo, e a partir delas aprenderá a maneira correta de classificar “como” conforme o contexto.

Eu_PRON nunca_ADV como_VERB peixe_NOUN. Eu_PRON como_VERB devagar_ADV. Você_PRON sabe_VERB como_CONJ fazer_VERB isso_PRON? É_VERB impressionante_ADJ como_CONJ as_ART pessoas_NOUN comem_VERB mal_ADV.

Já o aprendizado não supervisionado vai tentar encontrar padrões em dados sem a ajuda de um professor. Aproveitando o exemplo anterior, é como se o sistema fosse apresentado ao mesmo conjunto de e-mails, só que sem os rótulos indicando se cada um é ou não um spam. Programamos a máquina para que distribua o conjunto de e-mails em duas classes e como resultado gostaríamos que uma classe contivesse os e-mails do tipo spam e outra classe os e-mails do tipo não spam. Mas a máquina não “sabe” isso, ela só “sabe” que precisa encontrar a melhor maneira de distribuir os e-mails em dois grupos. O aprendizado não supervisionado é usado em tarefas de classificação de documentos ou de identificação de tópicos implícitos em grandes coleções de textos, quando precisamos identificar do que tratam (seus tópicos), sem tê-los lido. Imagine que você tenha uma coleção com todas as decisões julgadas pelo Superior Tribunal Federal (STF) nos últimos 10 anos, e deseje saber do que tratam para ter uma primeira ideia dos conteúdos. Dizemos que os tópicos são implícitos (ou latentes) porque serão identificados pelos sistemas em função das palavras contidas nos documentos. Nesta tarefa específica, que se chama “modelagem de tópicos”, um tópico não tem um nome e é formado por um conjunto de palavras que, ao olhar humano, podem fazer sentido entre si, como “mar vento praia onda tempestade”, “estudante colégio professor diretor livro”, mas também podem ser genéricos e pouco informativos, como “vida amor alma olho mundo” ou “dia tempo palavra pobre pai”.

No aprendizado por reforço, a aprendizagem acontece por tentativa e erro. O objetivo do algoritmo é selecionar o passo (a ação) que leva à maior recompensa ou menor punição. É uma abordagem muito usada para jogos, e não é (ainda) muito comum no PLN.

Por fim, o aprendizado profundo (AP), responsável pelos rápidos avanços na IA e no PLN. O AP se baseia em redes neurais artificiais, que são modelos computacionais (modelos matemáticos) inspirados nas redes neurais biológicas, e que simulam seu comportamento. Uma rede é um conjunto de neurônios artificiais conectados entre si, que formam algo parecido com uma teia, ou rede (figura abaixo). A primeira camada da rede processa a entrada e passa as informações para as camadas intermediárias até camada de saída, que é a camada final.



Uma curiosidade: uma rede neural artificial pode ter milhares de unidades de processamento, enquanto o cérebro de um mamífero pode ter bilhões. As primeiras redes neurais tinham uma única camada de neurônios conectados, e quando falamos em aprendizado profundo, o adjetivo “profundo” se refere às várias camadas das redes neurais atuais, como as da figura. Esta *profundidade* é também uma maneira de marcar diferença com relação à alegada superficialidade dos outros algoritmos de aprendizado de máquina.

Uma característica das redes neurais é serem genéricas, isto é, não temos arquiteturas específicas para processar linguagem, outras específicas para imagem, outras específicas para som. Outra característica é que aquilo que cada camada representa ou codifica é algo que não sabemos, e esse tema será tratado em breve.

Uma consequência deste processamento feito pelas redes neurais profundas é que elas são capazes de gerar *modelos de linguagem* muito mais complexos e eficazes que as abordagens anteriores.

A palavra *modelo* tem vários sentidos, e um modelo de linguagem é um esquema teórico que permite *prever*. Sabemos que um modelo é adequado não porque ele se parece com “original”, mas porque faz boas previsões. Temos modelos de previsão do tempo: quanto melhor o modelo, menos chances de errar a previsão. Modelos de linguagem também são modelos de previsão, mas ao invés de preverem chuvas, furacões ou períodos de seca, preveem tão somente qual a próxima palavra em um texto, dada a palavra anterior. Modelos de língua devem saber completar uma frase; são modelos que preenchem _____. Se você esperava que após a palavra “preenchem” viesse a palavra “lacunas”, fez uma previsão acertada, pois era o que eu havia escrito: modelos de língua são modelos que preenchem lacunas em um texto. O preenchimento correto das lacunas – ou, de forma mais acurada, a previsão correta relativa à próxima palavra – indica, de maneira indireta,

que o modelo capturou aspectos sintáticos (1) e semânticos (2) da língua. As palavras riscadas correspondem a previsões erradas, as palavras não riscadas correspondem a previsões possíveis:

(1) Modelos de língua irão preencher lacunas
~~incrível~~
~~comprei~~
lentamente
~~nunca~~
o

(2) Modelos de língua irão preencher lacunas
~~brigadeiro~~
~~sapo~~
~~bombeiros~~
espaços

Um bom modelo de língua (ou de linguagem) será capaz de fazer previsões acertadas e com isso sabemos, mesmo que indiretamente, que de alguma maneira ele capturou aspectos relevantes da língua.

Como os modelos são criados? Como nos demais tipos de AM, o aprendizado acontece a partir de dados, muitos dados. Os dados são as palavras do texto, e só. Não há dados rotulados, não há supervisão de aprendizagem – e justamente porque não há ninguém para ajudar, é preciso muito mais dados que nas outras formas de AM. De forma bastante simplificada, cada palavra do texto (o que é considerado palavra pode variar, e falaremos disso no capítulo 5) é transformada em um conjunto de números (essa transformação é assunto da seção 6.4.1). Esses conjuntos de números (as palavras) são os dados de entrada, que serão processados pelas várias camadas da rede neural.

Um modelo de linguagem é, portanto, um modelo probabilístico que prevê as chances de uma determinada sequência de palavras aparecer em uma língua. Um bom modelo de linguagem codifica a estrutura gramatical e aspectos de convencionalidade a partir dos dados (ou corpora) de treino. Existem modelos de linguagem simples, usados para tarefas simples como correção e de sugestão de palavras dos telefones celulares, e existem modelos complexos para tarefas complexas como a tradução. Existem modelos de linguagem estatísticos e modelos baseados em redes neurais. Hoje em dia, modelos de linguagem baseados em redes neurais estão levando a resultados às vezes até superiores ao desempenho humano, como veremos na seção 2.1.3.

Embora o AM possa ser feito de diferentes maneiras, alguns procedimentos são comuns. O processo subjacente consiste em dividir o dataset (ou o corpus padrão ouro, se estivermos falando de dados rotulados) em partes de tamanhos desiguais: a parte maior é usada para a fase de *treino*, e a parte menor é usada para a fase de *teste*. Retomando a analogia entre os procedimentos do aprendizado de máquina e os procedimentos do aprendizado escolar: temos a fase de estudo, quando fazemos exercícios, e a parte de verificação de aprendizagem, que é a hora da prova. No AM, a fase do treino é similar à fase de exercícios: fazemos exercícios, comparamos nossas respostas com o gabarito, vemos o que erramos e tentamos melhorar fazendo mais exercícios. A fase da avaliação é a prova, quando somos apresentados a questões novas, que não estavam na fase de estudo, e não temos acesso ao gabarito. A nota da prova (nossa avaliação) será dada

comparando a nossa resposta com a resposta do gabarito. Se o gabarito estiver errado, seremos injustamente penalizados caso tenhamos respondido corretamente.

Mais eficiente, para as máquinas e para o nosso estudo, é dividir a fase de estudo/treino em duas: além dos exercícios, fazemos um “simulado”, um teste que nos dá alguma pista sobre o que aprendemos, mas ainda não é a avaliação final. É a última chance que temos de perceber nossos pontos fracos e melhorar o desempenho fazendo os últimos ajustes. No AM, isto corresponde a dividir o dataset em 3 partes, e não em duas. Além das partições treino e teste, há uma outra parte para “desenvolvimento”, ou ajuste, feito em função das respostas do simulado (quanto ao tamanho, esta parte se aproxima ao tamanho do material de teste).

Nas máquinas, o treino acontece comparando as análises fornecidas por algum “chute” inicial, e em seguida ajustando as respostas a partir do contato com o gabarito. Por isso, quanto mais exemplos, melhor, e quanto mais variados os exemplos, melhor também. É fundamental que o gabarito esteja correto, para não fornecer pistas erradas sobre o que deve ser aprendido. Na fase de avaliação, escondemos a anotação da partição *teste* do corpus padrão ouro (escondemos as respostas das questões) e deixamos que o sistema mostre o que aprendeu. Só ao final do processo comparamos o resultado da análise automática com aquele que estava com o gabarito. Por isso, a qualidade da anotação é tão relevante: ela contribui não só para um bom aprendizado, mas também para uma avaliação correta. É igualmente importante que a partição *teste* fique guardada e se mantenha inédita até a fase final de avaliação. Do contrário, é possível que haja o chamado *overfitting*, quando os resultados são bons mas o sistema está “roubando”, isto é, é como se ele tivesse acesso prévio à prova e ao gabarito, e então ele acerta não porque tenha aprendido, mas porque já sabendo as perguntas, decorou as respostas.

Na geração de um modelo de língua por redes neurais profundas, em que a tarefa é aprender a prever a próxima palavra de um texto, o exercício de treino consiste em mascarar (e, portanto, esconder) palavras ao longo do texto. O modelo que está sendo criado dá o seu chute com relação a que palavra deverá ser aquela, usando como pista as palavras que estão no entorno da palavra mascarada, e chamamos este entorno de *janela*. O tamanho da janela varia conforme o algoritmo, e modelos complexos têm janelas com cerca de 2 mil palavras.

No aprendizado supervisionado, a qualidade do aprendizado também é altamente dependente do tamanho e da qualidade do dataset de treino. É por meio dele que o algoritmo de aprendizado irá detectar padrões e generalizar. Além disso, quanto mais complexa a tarefa, mais necessidade de dados (ou: quanto mais difícil o assunto que estamos estudando, mais precisamos de exercícios).

E o que significa muito? O corpus Bosque, por exemplo, tem cerca de 200 mil palavras e 9 mil frases. É um tamanho razoável para aprender informação relativa a classe gramatical e função sintática. Para tarefas em que nem todas as palavras recebem uma etiqueta, pode ser um corpus considerado pequeno.

Além disso, outra limitação do aprendizado supervisionado é que sistemas treinados em um tipo de texto não costumam ter um bom desempenho quando levados a analisar um

texto de gênero e/ou domínio diferentes, porque estrutura linguística e vocabulário serão diferentes. Um sistema que realiza análise sintática treinado em um corpus com textos jornalísticos pode ter uma queda de 10% no desempenho quando realiza a análise sintática em um material diferente, como artigos de medicina.

O aprendizado profundo oferece soluções para ambos os problemas. Após a criação de um modelo de linguagem “genérico” (porque treinado em textos variados), a quantidade de material de treino necessária tanto para aprender uma determinada tarefa quanto para analisar textos de domínios específicos é muito menor do que aquela que seria necessária no aprendizado supervisionado, o que torna a anotação linguística de dados do corpus uma tarefa bem mais razoável do ponto de vista da quantidade.

Na IA, dá-se o nome de *transferência de aprendizado* (*transfer learning*) à técnica de utilizar o conhecimento adquirido na resolução de uma determinada tarefa na resolução de uma outra tarefa, aparentada com a anterior. E chamamos de *ajuste fino* (*fine tuning*) a fase de adaptações necessárias. Com isso, um modelo que é ótimo em prever como continuar uma frase precisa se esforçar menos (isto é, treinar menos) para aprender a identificar opiniões em um texto, do que um modelo de aprendizado supervisionado criado para resolver o mesmo problema. (Por outro lado, o modelo de aprendizado profundo precisou treinar muito na fase anterior, quando aprende a prever palavras.)

O importante aqui é destacar que diferentemente da abordagem com base em regras, em que o conhecimento linguístico é explicitamente programado, no aprendizado de máquina não há conhecimento linguístico a ser programado: o que é programado é a maneira pela qual um sistema irá aprender, o que levará em conta cálculos matemáticos sobre as palavras do corpus. É por isso que se diz que pesquisadores de PLN precisam de conhecimento matemático e estatístico, e não de conhecimento linguístico. O que será aprendido dependerá da exposição aos dados e de como foi programada a maneira de aprender a partir dos dados, e não de conhecimento específico relativo às línguas.

Se as várias formas de aprendizado de máquina são feitas hoje por pessoas com formação informática e/ou matemática, e não envolvem o que convencionamos chamar de conhecimentos linguísticos, por que falar delas aqui? Entender o que está por trás dos métodos e abordagens bem-sucedidas ajuda a pensar sobre como estudos linguísticos têm mais chances de contribuir de forma efetiva. Ser bem-sucedido não significa ser perfeito, e que não se possa melhorar.

Além disso, regras linguísticas e conhecimento linguístico têm um papel ativo no PLN até hoje, ainda que seja um trabalho de bastidores. Elas são uma ótima maneira de construir datasets padrão ouro e versões iniciais de sistemas de PLN.

O anotador de Brill

Em 1992, Eric Brill lançou a primeira versão de seu etiquetador de POS: um anotador baseado em regras, mas que também incorpora informação estatística. Retomo aqui este trabalho porque do ponto de vista da criação de um corpus padrão ouro a abordagem de Brill é bastante instrutiva para linguistas.

De maneira geral, funciona da seguinte maneira:

Ingredientes:

Um corpus grande, padrão ouro, *já anotado com pos*. Esse corpus está dividido da seguinte maneira: 90% é usado para treino; 5% para avaliação (para o teste); e 5% para desenvolvimento. No trabalho original, foi utilizado o *Brown Corpus*, que contém 1,1 milhão de palavras de gêneros variados da língua inglesa.

Como fazer:

Etapas 1:

1. Utilizando a partição treino do corpus, é derivada uma lista (um léxico) que contém, para cada palavra do corpus, a indicação das classes (*pos*) que podem ser atribuídas a essa palavra, associada à frequência de cada *pos*. Por exemplo, se fôssemos fazer o mesmo com o corpus Floresta (cujo tamanho é de 6 milhões de palavras) teríamos, uma lista como a abaixo (os números são aproximados, e a anotação foi feita de forma automática pelo analisador PALAVRAS):

que – conjunção: 42%	casa – substantivo: 97%
que - pronome relativo: 51%	casa – verbo: 1.5%
um – artigo: 80%	barata – adjetivo: 60%
um – numeral: 14%	barata – substantivo: 40%
tabuleiro – substantivo: 100%	

2. Em contato com uma porção do texto sem anotação (mas ainda na parte de treino), a ferramenta atribui, para cada palavra, a etiqueta de *POS* mais provável (ou seja, a mais frequente) de acordo com o léxico criado a partir do corpus de treino. Este procedimento olha apenas para a palavra, e não leva em conta o contexto.
3. Palavras “novas” (não estavam no corpus de treino, e portanto não apareceram no léxico) que estiverem em maiúscula e que não estiverem no início da frase recebem a etiqueta PROP, relativa aos nomes próprios.
4. Palavras “novas” (que não estavam no corpus de treino e não viraram PROP no procedimento anterior) recebem a etiqueta mais *frequentemente* atribuída a palavras *que terminam da mesma maneira* (especificamente, de palavras que compartilham a mesma sequência de 3 últimas letras) Assim, uma sequência como *xxxxram* será anotada como verbo, porque –ram é uma terminação *típica* de verbo.

Um linguista clássico diria: ok, muito interessante... Mas isso vai falhar! E logo apresentaria uma série de casos em que os procedimentos levariam a erros. Se para a terminação –ram o resultado é preciso, o mesmo não se aplica a –rão, por exemplo, com formas como *camarão* ao lado de *amarão*. Sim, responderia o linguista computacional. Sabemos que vai falhar, essas falhas já estão previstas e, por isso, esses são apenas procedimentos *iniciais*.

Para a língua inglesa, esses procedimentos simples resultaram em uma taxa de 92% de acertos.

Etapas 2:

Após a primeira etapa, começa a fase dos “remendos”, cujo objetivo é melhorar o desempenho da anotação. São chamados “remendos” porque consertam os furos das regras anteriores. Os remendos são adquiridos (ou aprendidos) automaticamente, e têm a seguinte forma:

1. pos[0]=ART pos[1]=ART → pos=PREP
2. pos[0]=N pos[-1]=N → pos=ADJ
3. pos[0]=ART pos[1]=V → pos=PREP

Lemos as regras da seguinte maneira: [0] é a palavra para a qual se está olhando (a palavra alvo), [-1] é a palavra que está imediatamente à esquerda da palavra alvo; [+1] é a palavra que está imediatamente à direita da palavra alvo. Ou seja:

1. Se a palavra alvo é um ARTigo e à sua direita está um ARTtigo, então a palavra alvo se transforma em PREPosição (Foi a_ART o_ART cinema → Foi a_PREP o_ART cinema)
2. Se a palavra alvo é um Nome e à sua esquerda está um Nome, então a palavra alvo se transforma em ADJetivo (um filme_N brasileiro_N → um filme_N brasileiro_ADJ)
3. Se a palavra alvo é um Artigo e à sua direita está um Verbo, então a palavra alvo se transforma em PREPosição (demorei a_ART perceber_V → demorei a_PREP perceber_V)

Como são feitos os remendos? De onde vêm as regras?

Como vimos, o material é treinado em 90% do corpus, e avaliado em 5%. Os outros 5% restantes são usados para construir os remendos – que é a fase do “simulado”. Quando uma versão preliminar do anotador está treinada, esta versão é utilizada para anotar esses outros 5% do corpus como uma anotação intermediária, feita para ver o quanto o anotador acerta. Na anotação desses 5% intermediários (na anotação do “simulado”), é esperado que haja erros, e estes erros serão corrigidos automaticamente por meio da consulta ao gabarito do simulado. A etapa seguinte é a avaliação, isto é, a prova. Nesse momento, o 5% do material inédito é apresentado, e então saberemos o que de fato foi aprendido.

Um aspecto chave dessa estratégia é que “remendo que leva à maior redução de erros” é diferente de “remendo que corrige mais casos”, porque é possível que um remendo corrija muitos erros, mas por outro lado introduza mais outros tantos, fazendo com que o desempenho, ao final, piore. Reproduzo a explicação do artigo de Brill (1992):

Por exemplo, quando o etiquetador inicial é aplicado no corpus de remendos, ele anota erradamente como verbos 159 palavras que deveriam ser substantivos. Se o remendo "mude a etiqueta de verbo para substantivo se uma dentre as duas palavras que a precedem for um determinante" for aplicado, ele corrigirá 98 dos 159 erros. Porém, ele também resultará em 18 erros adicionais (novos erros), trocando etiquetas que realmente deveriam ser verbos, para substantivos. Esse remendo resulta, ao final, em uma diminuição de 80 erros no corpus de remendos.

O remendo que leva à maior melhoria, quando aplicado ao corpus de remendos (os 5% intermediários), é adicionado à lista de remendos. Ele é então aplicado (ainda no corpus de remendos), e assim a aquisição e escolha do melhor remendo continua, até chegar à melhor configuração final, isto é, aquela que proporcionará um corpus com a menor quantidade de erros, sendo irrelevante a quantidade de remendos necessária para isso. O

processo termina quando não é mais possível reduzir erros – e, no trabalho original, para o inglês, levou a 95% de acertos.

Uma característica desta estratégia é a ausência de um dicionário – o léxico inicial é derivado do corpus. Em um trabalho posterior, um dicionário foi incorporado à fase inicial com o objetivo de melhorar a anotação de palavras não vistas no treinamento, o que levou à redução de 0.5% dos erros com relação à estratégia anterior.

Deixando os resultados de lado, o ponto interessante é que se trata de uma maneira diferente de encarar regras. Quando, linguisticamente, pensamos em regras, nossa ideia primeira é pensar em regras precisas, que não produzam erros (ou produzam o mínimo possível), e por isso uma proposta de regras que já contém falhas nos parece pouco promissora. A estratégia aqui é oposta: usamos regras para um primeiro chute, sem a pretensão de que resolvam tudo. Em seguida, detectamos erros e vamos criando regras específicas para corrigi-los. Como já indicado, esta é uma boa estratégia para construir corpora padrão ouro. O chute inicial pode ser dado por algum sistema (baseado em regras ou em AM) e depois corrigimos. A ideia de regras que vão progressivamente refinando resultados iniciais também está por trás da gramática constritiva, que vimos no início desta seção.

2.3 Avaliação

Como podemos ser suspeitos para avaliar nossas próprias criações (autoavaliação) e como é importante poder comparar nossos resultados com outros para saber se estamos bem, o PLN desenvolveu o hábito das avaliações conjuntas (*shared tasks*). A ideia geral por trás das avaliações é fornecer uma estrutura experimental comum (mesmos conjuntos de dados para avaliar ou para treinar e avaliar) sobre a qual o desempenho dos sistemas será avaliado. O bom é que esses repositórios de dados permanecem à disposição da comunidade mesmo após o encerramento oficial da competição.

Mais do que estabelecer qual é o melhor sistema para uma determinada tarefa, a principal vantagem das avaliações é tornar possível uma comparação entre diferentes sistemas, métodos ou algoritmos, já que todos são avaliados da mesma maneira, a partir de um mesmo material. Com isso, quem cria pode saber como o seu sistema está com relação a outros que se propõem fazer a mesma coisa, um futuro usuário/cliente pode saber qual o “melhor” sistema para suas demandas, e pesquisadores têm acesso a como está o estado da arte (isto é, qual o melhor resultado, e qual o resultado médio esperado) naquela tarefa. O quadro a seguir traz resultados para a tarefa de resolução de correferência (seção 6.7) para a língua inglesa. O primeiro resultado, de 2012, se refere ao resultado da equipe vencedora quando a tarefa foi lançada em uma avaliação conjunta, o CoNLL (CoNLL é acrônimo de *Conference on Computational Natural Language Learning*). Como podemos ver, anos após a realização da competição o seu dataset (um corpus padrão ouro) continua sendo usado para avaliar sistemas, e com isso vemos também a evolução da área. Uma curiosidade é que esta era uma tarefa multilíngue, isto é, envolvia a identificação de cadeias de correferência em três línguas distintas: inglês, chinês e árabe. A equipe vencedora em 2012 (Fernandes et al., 2012) foi uma equipe brasileira. Na equipe não havia linguistas ou falantes de chinês ou árabe, mas isto não os impediu de criarem o melhor modelo para as três línguas, na época (os resultados da tabela referem-se apenas aos resultados para o inglês).

Ano	Desempenho na tarefa de correferência (para o inglês)
2012	63.37
2017	67.2
2019	76.6
2020	80.2

Além das avaliações conjuntas – que são lançadas, têm prazos e vencedores – também existem os *benchmarks* – ou recursos de referência ou datasets padrão ouro –, recursos que viabilizam igualmente a avaliação e comparação. *Benchmarks* são materiais disponibilizados para medir o desempenho de sistemas em determinadas tarefas sem a dimensão temporal das avaliações, mas podem ser lançados como competições, o que aliás dará mais publicidade ao material. E, de modo complementar, datasets usados nas competições ficam disponíveis para utilização em qualquer momento.

Existem datasets padrão ouro que avaliam tarefas bastante complexas, e um deles é o SQuAD (*Stanford Question Answering Dataset*), lançado em 2019 e criado a partir de uma amostra de documentos da Wikipédia (em inglês). Os documentos foram divididos em parágrafos, e pessoas, ao receberem os parágrafos, deveriam formular e responder 5 perguntas sobre o seu conteúdo. Para tornar o trabalho mais difícil para as máquinas, os criadores do SQuAD não permitiram a utilização da funcionalidade *copiar & colar*, forçando assim as pessoas a usarem suas próprias palavras na formulação das perguntas. A versão 2.0 do SQuAD conta com 100.000 perguntas “convencionais”, e mais de 50.000 perguntas irrespondíveis, isto é, perguntas para as quais não há resposta possível, criadas para parecerem semelhantes às convencionais. Para ter um bom desempenho neste material, os sistemas devem não apenas responder às perguntas quando possível, mas também devem conseguir determinar quando não há resposta. Também foi pedido para pessoas que respondessem as perguntas do SQuAD, para que fosse possível ter uma medida do desempenho humano na tarefa. Atualmente, mais de 20 sistemas já superaram o desempenho humano no SQuAD.

O GLUE (*General Language Understanding Evaluation*) é um dataset multitarefas, isto é, consiste na compilação de datasets de diferentes tarefas, como resposta a perguntas, análise de sentimento e inferência lógica. Para algumas tarefas, há dados de treinamento abundantes, mas para outras não há. Com isso, espera-se favorecer modelos que podem aprender a representar o conhecimento linguístico de uma forma que facilite o aprendizado eficiente (aprender bem com pouco) e a transferência de conhecimento entre as tarefas (ter um bom desempenho em uma tarefa e conseguir ter um bom desempenho em uma tarefa próxima, mesmo com poucos dados de treino).

Tanto o GLUE como o SQuAD têm tabelas com os desempenhos dos sistemas (chamadas *leaderboard*) em suas respectivas páginas eletrônicas. Ao analisar a evolução dos resultados vemos que o sucesso das técnicas de aprendizado profundo chega a lugares até então inimagináveis de serem alcançados sem a utilização de algum tipo de conhecimento, como as inferências lógicas. Por outro lado, avaliar os resultados apenas pelos números não nos deixa ver questões importantes relativas a como esses resultados foram obtidos, assunto que será retomado em 6.8.

Desde os títulos das conferências de avaliações e dos datasets padrão ouro é possível perceber que a língua privilegiada é o inglês, mas cada vez mais são comuns competições

que envolvem o processamento multilíngue, e o português às vezes integra o conjunto de idiomas envolvidos. A principal limitação para a participação da língua portuguesa nesses contextos é justamente a inexistência de *recursos linguísticos* dedicados às tarefas em questão (e voltamos ao início desta seção: recursos são dependentes de língua). Nesse contexto, recursos linguísticos são, sobretudo, corpora anotados.

Mas a língua portuguesa também tem suas próprias competições.

O HAREM (Avaliação e Reconhecimento de Entidades Mencionadas) foi uma avaliação conjunta organizada pela Linguatca na área do reconhecimento de entidades mencionadas em português (trataremos de entidades mencionadas no capítulo 6, e algumas palavras sobre a Linguatca, que teve um importante papel no desenvolvimento do PLN de língua portuguesa, estão no capítulo *Para Saber Mais*). O objetivo do HAREM era avaliar o sucesso na identificação e consequente classificação automática de nomes próprios em vários textos em língua portuguesa. O HAREM contou com duas edições, em 2003 e 2008 e, mesmo já tendo se passado há tanto tempo, os recursos produzidos no HAREM – corpus com anotação padrão ouro, chamado Coleção Dourada, e uma série de ferramentas e métricas associadas – são utilizados até hoje. A Linguatca organizou ainda outras avaliações conjuntas, como o *Págico*, que será mencionado na seção a seguir, em 2012, e em 2003, as *Morfolimpíadas*, cujo objetivo foi avaliar analisadores morfológicos da língua portuguesa. Todas as avaliações têm suas respectivas páginas eletrônicas, onde todo o material - conjuntos de dados, instruções, resultados, publicações – está disponível para consulta e download.

Outra avaliação conjunta é a ASSIN (Avaliação de Similaridade Semântica e de Inferência Textual), dedicada às tarefas de Inferência Textual e Similaridade Semântica. Na inferência textual, o desafio consiste em determinar se o significado de um trecho implica o outro; na Similaridade Semântica a tarefa consiste em atribuir uma pontuação de similaridade semântica a esses trechos, segundo uma escala de 1 a 5 (falaremos de ambos na seção 6.7) A primeira edição da ASSIN aconteceu em 2016, e a segunda, em 2019.

Além de HAREM, Págico, Morfolimpíadas e ASSIN, também temos os seguintes conjuntos de dados que permitem avaliações da língua portuguesa (e tudo isso está listado na seção @Ponteiros):

Bosque - subconjunto do treebank Floresta Sintática, composto por textos de jornal e voltado para a tarefa de análise sintática. Disponível em diferentes formatos.

MacMorpho – composto por textos de jornal e voltado para a tarefa de classificação de palavras. Disponível em três versões.

CHAVE – resultado da participação da língua portuguesa na avaliação CLEF (Cross-Language Evaluation Forum), nos anos de 2004 a 2009. Contém textos de jornal e é voltada para tarefas de recuperação de informação, resposta a perguntas e recuperação de informação geográfica.

Coleções de dados públicas e acessíveis que contenham problemas e seus gabaritos, além de permitirem o entendimento de como se comporta um método/ferramenta em relação ao desempenho humano e em relação a outros métodos/ferramentas, são cruciais para o desenvolvimento do PLN. Por isso uma das coisas de que precisamos para avançar com PLN em português é criar este tipo de recurso. E quem produz esses conjuntos de dados,

que têm um prazo de validade enorme e determinarão a qualidade do aprendizado e da avaliação? Com muita frequência, linguistas.

Tipos de avaliação

A *avaliação intrínseca* é interna à tarefa que está sendo avaliada. Ou seja, considerando que uma tarefa pode ser uma etapa intermediária para a realização de uma aplicação, a avaliação intrínseca avalia o desempenho de um sistema (ou modelo) naquela tarefa específica. Já a *avaliação extrínseca* avalia o desempenho desse mesmo sistema (ou modelo) em tarefas subsequentes.

Podemos avaliar uma ferramenta que anota classes de palavras comparando o resultado da sua análise com o resultado do gabarito (do corpus padrão ouro), e esta é uma avaliação intrínseca. Podemos avaliar modelo de língua que prevê a próxima palavra comparando a palavra prevista com a palavra efetivamente usada naquela posição, e esta também é uma avaliação intrínseca. Podemos avaliar este mesmo modelo de língua em uma tarefa de anotação de classes de palavras, também comparando o resultado da sua análise com o resultado do gabarito, e esta é uma avaliação extrínseca. A avaliação é extrínseca porque ela não avalia aquilo que, diretamente, o modelo (ou ferramenta ou sistema) faz. Ela avalia indiretamente, verificando o quanto outras tarefas se beneficiam do referido modelo (ou ferramenta ou sistema).

Para a maioria das tarefas de PLN, a avaliação pode ser feita de forma automatizada, comparando os resultados dos sistemas com os resultados dos gabaritos conforme algumas métricas (ou medidas). Algumas delas são *precisão*, *abrangência* e *medida F*, que é uma média harmônica entre a precisão e a abrangência.

A precisão mede a qualidade das respostas, verificando se tudo aquilo que um sistema classificou como X é, de fato, X. Mais especificamente, a precisão mede a proporção de respostas *corretas* fornecidas considerando o total de respostas fornecidas.

Já a abrangência mede a capacidade de encontrar respostas, não importa se corretas ou não, verificando a proporção de respostas encontradas considerando todas as respostas que deveriam ter sido encontradas. A medida da abrangência indica se tudo aquilo que deveria ter sido encontrado foi, de fato, encontrado. Para calcular a abrangência precisamos de um gabarito, e não apenas de uma análise de erros, pois uma análise de erros só mostra a precisão.

Para calcular precisão, abrangência e a medida F, classificamos os resultados da seguinte maneira:

- verdadeiro positivo (VP): o elemento foi detectado pela análise automática e foi classificado de forma correta.
- verdadeiro negativo (VN): o elemento foi detectado pela análise automática, mas foi classificado de forma errada.
- falso positivo (FP): o elemento foi detectado pela análise automática, mas não deveria.
- falso negativo (FN): o elemento não foi detectado pela análise automática, mas deveria.

Para calcular a precisão fazemos $VP \div (VP+FP)$

Para calcular a abrangência fazemos $VP \div (VP+FN)$

Para calcular a medida F fazemos $2 \frac{\text{precisão} * \text{abrangência}}{\text{precisão} + \text{abrangência}}$

O quadro abaixo traz duas anotações de entidades para o mesmo texto (veremos este tipo de anotação no capítulo 6, mas ela é intuitiva o suficiente para ser compreendida aqui. A ideia é fornecer classificações para os nomes próprios). O primeiro trecho é o gabarito, e o segundo é a análise que queremos avaliar.

GABARITO: Epidemia de dança de **1518_{DATA}** foi um caso de dançomania e histeria coletiva ocorrido em **Estrasburgo_{LOCAL}**, **França_{LOCAL}** (então parte do **Sacro Império Romano-Germânico_{ORG}**) em julho de **1518_{DATA}**. O fenômeno teve início quando uma mulher, **Frau Troffea_{PESSOA}**, começou a interpretar passos frenéticos de dança numa rua da cidade de **Estrasburgo_{LOCAL}** aparentemente sem qualquer motivo.

ANÁLISE AUTOMÁTICA: **Epidemia_{LOCAL}** de dança de 1518 foi um caso de dançomania e histeria coletiva ocorrido em **Estrasburgo_{ORG}**, **França_{LOCAL}** (então parte do **Sacro Império Romano-Germânico_{LOCAL}**) em julho de 1518. O fenômeno teve início quando uma mulher, **Frau Troffea_{PESSOA}**, começou a interpretar passos frenéticos de dança numa rua da cidade de **Estrasburgo_{LOCAL}** aparentemente sem qualquer motivo.

Quando comparamos a segunda análise com a primeira, encontramos o seguinte:

VP: 3 (França_{LOCAL}; Frau Troffea_{PESSOA} Estrasburgo_{LOCAL})
VN: 2 (Estrasburgo_{ORG}; Sacro Império Romano-Germânico_{LOCAL})
FP: 1 (Epidemia_{LOCAL})
FN: 2 (1518_{DATA}; 1518_{DATA})

Precisão: $3 \div (3+1) = 0.75$

Abrangência: $3 \div (3+2) = 0.6$

Medida F: $2 \frac{0.75 * 0.6}{0.75 + 0.6} = 0.6$

Ou seja, no exemplo acima, a precisão é melhor que a abrangência, que é baixa. E a medida F é de 0.6.

Uma ferramenta pode ser muito precisa - todas as classificações que ela faz são corretas - e pode, igualmente, ter uma baixa abrangência – apesar de acertar, 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.

Precisão e abrangência (e medida F) são apenas algumas maneiras de avaliação, e cada tarefa pode ter suas especificidades. Na página do livro, em *@Sobre avaliação sintática* (e a anotação sintática é abordada em 6.2), apresento como a avaliação da sintaxe tem sido realizada.

Cálculos relativos à precisão e à abrangência, dentre outros, são feitos automaticamente, e esta é mais uma das vantagens dos corpora padrão ouro na avaliação: basta criar o gabarito uma vez, e aplicamos sempre a mesma prova. O que pode mudar são os aprendizes (sistemas, modelos ou ferramentas) e os métodos aprendizagem (os algoritmos). Além disso, quando fazemos uma análise de erros manual, sem produzir um gabarito, é comum não repararmos naquilo que não foi encontrado, mas deveria ter sido (os falsos negativos). Ou seja, se analisássemos apenas os resultados da máquina, é possível que não reparássemos nos casos de “1518” que não chegaram a ser classificados. Por isso é tão importante a criação dos gabaritos.

No entanto, para muitos problemas de PLN a criação de um gabarito pode ser complexa, como é o caso da tradução automática, ou pode não haver gabarito disponível. Vamos supor que eu precise de uma ferramenta de análise sintática para anotar textos jurídicos, especificamente decisões de juízes. Sabemos que se trata de uma linguagem com suas estruturas e léxico próprios, às vezes referida como “juridiquês”. Para saber se podemos confiar na análise de uma ferramenta, já sabendo de antemão que ela foi criada (ou treinada) tomando por base textos de jornal, precisamos ter uma medida da sua qualidade antes de utilizá-la. Neste caso, o que podemos fazer é a análise manual de uma amostra, lançando mão de nosso conhecimento linguístico específico relativo à tarefa em questão. Posteriormente, e conforme o tamanho da amostra, ela poderá ser usada como padrão ouro.

2.4 Textos e informação não-estruturada

A relevância do PLN é evidente quando nos damos conta da imensa – e em constante crescimento – produção e disponibilização de conteúdo na forma de textos. No entanto, este conteúdo está no texto de maneira não estruturada, isto é, não segue nenhum formato previsível. A figura abaixo apresenta três maneiras diferentes de indicar a morada de alguém – esta informação (de que se trata de três maneiras de dizer a mesma coisa) é óbvia para nós, mas não para as máquinas. E esta mesma informação, por outro lado, poderia estar disponível por meio de um formulário, com campos pré-estabelecidos para informações como *nome, rua, complemento, bairro, cidade, cep, telefone, código de área*. Neste caso, seria fácil para as máquinas saber do que se trata, pois a informação já estaria organizada (estruturada).

- Batman Silva, Praça dos Patos, 160/ 1002 - Humaitá, Rio de Janeiro. 22260-221, tel (021) 22555443.
- Batman Silva. Praça dos Patos n. 160 – apto 1002. Humaitá – 22260-221. Rio de Janeiro/RJ, 021 2255-5443
- Batman Silva mora no número 160 da Praça dos Patos, apartamento 1002, no bairro do Humaitá, Rio de Janeiro. CEP: 22260-221. O telefone é 22555443, e o código de área é 021.

Podemos comparar o preenchimento de formulários com a escrita livre de jornais, peças ficcionais, entrevistas, resenhas, tweets... Felizmente para nós, e infelizmente para o processamento automático, a língua (qualquer língua) nos oferece várias maneiras

diferentes de dizer as mesmas coisas. É compreensível que o processamento automático seja um desafio e tanto.

As caixas de informação da Wikipédia (*infoboxes*, não confundir com as caixas que são índices dos artigos) são um exemplo concreto de informação estruturada. Uma caixa de informação é uma tabela com formato fixo que fica no canto superior direito do artigo e tem como objetivo apresentar um resumo com aspectos relevantes e comuns (fatos e estatísticas) sobre o tema do artigo.

As caixas de informação contêm fatos pré-definidos conforme a natureza do artigo. Artigos biográficos, por exemplo, devem conter pelo menos certos tipos de informação (nome, local e data de nascimento, ocupação etc), artigos de animais devem contar com informações como classificação científica (gênero e família) etc. Com isso, é mais fácil encontrar tais informações e compará-las com a de outros artigos da mesma categoria. Além disso, as caixas devem resumir informações que estão presentes no texto principal, e não podem conter informação que não seja sustentada pelas fontes confiáveis no corpo do artigo.

Por já conterem espaços previstos para certos tipos de informação, podemos dizer que as caixas contêm informação estruturada relativa ao conteúdo dos artigos. O quadro abaixo contém a informação de uma caixa de informação como nós a vemos (esquerda, imagem editada), e como está codificada (direita, na parte de cima). Ou seja, toda a informação que aparece estruturada nas caixas de informação está no texto, mas de maneira não estruturada (direita, parte de baixo).

Informação geral		
Nome completo	Alfredo da Rocha Vianna Filho	título = Pixinguinha
Também conhecido(a) como	Pizinguim, Bexiguinha, Pixinguinha, São Pixinguinha	imagem = Pixinguinha.png
Nascimento	23 de abril de 1897	imagem_tamanho = 270px
Local de nascimento	Rio de Janeiro, RJ Brasil	imagem_legenda = Pixinguinha em 1956
Morte	17 de fevereiro de 1973 (75 anos)	nome_completo = Alfredo da Rocha Vianna Filho
Local de morte	Rio de Janeiro, RJ Brasil	conhecido_como = Pizinguim, Bexiguinha, Pixinguinha, São Pixinguinha
Nacionalidade	brasileiro	nascimento = 23 de abril de 1897
Gênero(s)	Choro Maxixe Samba Valsa	local_nascimento = Rio de Janeiro Brasil
Ocupação(ões)	Maestro, flautista, saxofonista, compositor e arranjador	morte = 17 de fevereiro de 1973 (75 anos)
		local_morte = Rio de Janeiro Brasil
		nacionalidade = brasileiro
		gênero = Choro Maxixe Samba Valsa
		ocupação = Maestro, flautista, saxofonista, compositor

Alfredo da Rocha Vianna Filho, conhecido como Pixinguinha (Rio de Janeiro, 23 de abril de 1897 — Rio de Janeiro, 17 de fevereiro de 1973), foi um maestro, flautista, saxofonista, compositor e arranjador brasileiro.

2.5 Leitura, processamento automático e compreensão

Encontrar informação em textos é apenas uma das atividades que fazemos com a linguagem. Podemos entendê-la de forma genérica como qualquer processo que *encontra*, *estrutura* e *combina* dados que estão em textos. Podemos dizer então que a extração de informação é responsável por estruturar a informação que aparece espalhada em textos. Voltando à figura com acima, a saída de um sistema de extração de informação poderia corresponder a algo como o código da caixa de informação, com campos pré-definidos.

Em 2011, foi lançada uma avaliação conjunta chamada Págico, na qual sistemas deveriam encontrar as respostas para uma lista de perguntas ou tópicos vasculhando as páginas da Wikipédia. A ideia era simular uma situação real de pesquisa complexa, que envolve a leitura de uma grande coleção de documentos com o objetivo de encontrar resposta – no caso, respostas – para uma questão, sabendo que as respostas estão espalhadas em diversos documentos. Originalmente, o Págico previa a participação também de pessoas, para que fosse possível comparar o desempenho humano ao da máquina, mas foram poucos os participantes humanos. Alguns exemplos de tópicos do Págico:

Tópico 01: Filmes sobre a ditadura ou sobre o golpe militar no Brasil.

Tópico 13: Dinossauros carnívoros que habitaram o Brasil.

Tópico 62: Praias de Portugal boas para a prática de surf

Tópico 80: Línguas faladas em Timor Leste

Uma especificidade do Págico é que o nome da página deveria *ser* a resposta, e não simplesmente *conter* a resposta (do contrário, teríamos outra tarefa, que seria a geração do texto da resposta). O problema subjacente pode ser formulado de duas maneiras:

- (a) O que é preciso para que um sistema *compreenda o que deve procurar nos textos*?
- (b) O que é preciso para que um sistema *forneça respostas adequadas para a pergunta*?

As perguntas formuladas em (a) e (b) são a mesma coisa?

Sim no que se refere ao resultado final esperado. *Não* no que se refere à tarefa do computador. Isto porque, em (a), a *compreensão* é etapa necessária para a obtenção do resultado desejado. Para que um sistema realize a tarefa adequadamente, é preciso que ele *entenda* o texto (e as perguntas); e para que *entenda* o texto, é preciso que *entenda* as palavras.

Do ponto de vista assumido em (a) cabe aqui a analogia entre a compreensão do PLN e a de um papagaio, que repete aquilo que lhe foi ensinado, mas não compreende o que diz. Muito da discussão acerca da compreensão das máquinas se deve ao uso que tem sido feito, no PLN e na IA, da palavra *compreensão*. A partir de quais critérios concluímos que alguém *compreendeu* algo?

Mais uma vez podemos usar a analogia com uma situação de prova: como garantir que a pessoa, ao responder corretamente uma questão, compreende o que diz, ou compreende a resposta? Como garantir que aquilo que estamos aferindo ao corrigir é o resultado de uma *compreensão* verdadeira, em oposição à habilidade de construir paráfrases ou de repetir em uma prova o que foi dito em aula? Que critérios utilizamos para avaliar a compreensão? A única maneira de aferirmos compreensão é por meio do comportamento – se a outra pessoa age segundo aquilo que seria esperado, *como se* tivesse compreendido. Nos termos do filósofo Wittgenstein, “compreende uma ordem aquele que age de acordo com ela” (1953:§6). Mas essa não é uma visão unânime, e uma quantidade de artigos vem (re)discutindo o conceito de compreensão no PLN e na IA (por exemplo, Bender e Koller, 2020; Dunietz et al., 2020).

3. Linguística Computacional e Linguística: um pouco de história

Quando surge nos anos 1950, a área que mistura Linguística e Computação é batizada oficialmente de “Linguística Computacional”, um nome que, se por um lado enfatiza a dimensão linguística sobre a computacional (“Linguística Computacional” e não “Computação Linguística”), por outro lado é fruto de uma escolha que reflete principalmente a necessidade de “parecer científica” – científica a ponto de receber financiamento para seus projetos de pesquisa. Os outros nomes que concorriam eram *Engenharia da Linguagem* ou *Processamento de Linguagem Natural*, descartados por parecerem técnicos demais e acadêmicos de menos, segundo Martin Kay (2005). Mas nem por isso os demais nomes deixaram de ser usados, e se boa parte da comunidade atualmente utiliza os termos de maneira intercambiada, também há pesquisadores que estabelecem diferenças, às vezes nem tão sutis.

Há quem veja o PLN como uma área aplicada, direcionada à resolução de tarefas, e a Linguística Computacional como uma área sobretudo teórica, “menos interessada em produzir objetos úteis do que em testar teorias”, o que nunca teria conseguido fazer a contento (Wilks, 2006). Para Martin Kay, a Linguística Computacional tem um lado teórico (“*Linguística Computacional está tentando fazer o que a Linguística faz de uma forma computacional*”) e um lado aplicado, e ainda há o PLN. Para ele, no viés teórico da Linguística Computacional, os objetivos computacionais levariam a avanços teóricos na Linguística. No viés aplicado, a ideia seria aplicar tecnologias baseadas em princípios científicos a fim de realizar tarefas como tradução automática, sumarização, extração de informação. Já o PLN seria apenas um campo aplicado que raramente é informado pelos resultados da investigação científica realizada por linguistas: “Na falta de uma inteligência artificial na qual incorporar sua tecnologia, os linguistas foram forçados a procurar um substituto, embora imperfeito, e muitos pensam que o encontraram no (...) ‘processamento estatístico de linguagem natural’” (Kay 2005:xix). Ainda segundo Kay, o fato de o desenvolvimento tecnológico do PLN ser baseado quase inteiramente em modelos de aprendizado de máquina sem base linguística computacional é uma “aberração que, felizmente, pode estar em processo de correção” (2011:1).

Destaco dois pontos do breve apanhado acima: o distanciamento entre Linguística e PLN, e um certo desdém relativo ao PLN estatístico, aplicado demais e pouco teórico, e à Linguística Computacional, aplicada de menos e muito teórica.

O segundo ponto reflete a rivalidade entre os paradigmas de PLN (e de IA) que vimos no capítulo 2 – abordagem baseada no conhecimento (neste caso, conhecimento linguístico, as regras) e abordagem baseada nos dados. A abordagem baseada nos dados, como vimos, é criticada pela superficialidade: aplicação de ‘truques’ com padrões textuais que dispensam uma compreensão verdadeira do texto. A crítica do lado oposto é de que abordagens baseadas em conhecimento seriam frágeis, porque baseadas em regras arbitrárias e pouco flexíveis para lidar com a imensa variedade presente nos textos do mundo real.

Esta rivalidade, por sua vez, é a manifestação de uma outra disputa entre duas escolas de pensamento: racionalismo e empirismo. É uma disputa antiga, que também se manifesta

na Linguística, e que nos ajuda a entender o segundo ponto, o distanciamento entre Linguística e PLN.

3.1 Dados e conhecimento: empirismo e racionalismo na Linguística e no PLN

Abordagens baseadas em regras e em estatística vêm se alternando desde o surgimento do PLN, e correspondem a diferentes tendências no PLN e na Linguística. Abordagens baseadas em regras se alinham a uma doutrina filosófica chamada *racionalismo*, e abordagens estatísticas se alinham a outra doutrina, o *empirismo*. Na IA, racionalismo e empirismo são as contrapartes filosóficas das abordagens simbólicas e engenharia do conhecimento, de um lado, e das abordagens estatísticas e redes neurais, de outro. Essas duas escolas são referidas na definição de Linguística Computacional mencionada no início do capítulo 2: modelos computacionais de fenômenos linguísticos *baseados no conhecimento* ou *orientados por dados*. Mas não devemos opor “dados” e “conhecimento”. Ser “orientado por dados” não significa que não haja conhecimento, mas sim que o conhecimento será o resultado do processamento dos dados.

Entre os anos 1960 e 1985, grande parte da Linguística, da Psicologia, da Inteligência Artificial e do PLN foi dominada pela abordagem *racionalista*. Para o racionalismo, uma parcela considerável do conhecimento humano é dada a priori, o que hoje é lido como herança genética. René Descartes (filósofo, físico e matemático francês que viveu entre 1596 e 1650) é um dos expoentes do racionalismo. Na Linguística, a influência do racionalismo não é pouca: a segunda metade do século XX, pelo menos até os anos 1990s, foi dominada pela escola gerativa de Noam Chomsky, e não à toa uma de suas obras mais fiéis ao seu pensamento, segundo ele próprio, chama-se “Linguística Cartesiana”. Chomsky é um crítico feroz de abordagens estatísticas e empiristas, que já afirmou que a noção de probabilidade de uma frase é uma noção completamente inútil. E mesmo recentemente, quando o trabalho com corpus já caminhava bastante bem e as análises computacionais de grandes volumes de texto eram confiáveis, a crítica continuava pesada, o que contribuiu para que o interesse neste tipo de abordagem linguística fosse minado durante anos. Estas são suas palavras em uma entrevista de 2004:

Linguística de corpus não significa nada. É como dizer, suponha que um físico decida, suponha que a física e a química decidam que, em vez de confiar em experimentos, o que irão fazer é gravar vídeos das coisas que acontecem no mundo e coletar enormes vídeos de tudo o que está acontecendo, e a partir disso eles talvez cheguem a algumas generalizações ou insights. Bem, você sabe, as ciências não fazem isso. Mas talvez elas estejam erradas. Talvez as ciências devam coletar muitos e muitos dados e tentar desenvolver os resultados a partir deles. Bem, se alguém quiser tentar isso, tudo bem. Eles não receberão muito apoio no departamento de química, física ou biologia. Mas se eles quiserem experimentar, bem, é um país livre, tente isso. Nós vamos julgar pelos resultados que aparecerem. (ANDOR, 2004:97, em tradução livre)

No PLN/IA, a programação simbólica (baseada em regras) e a engenharia do conhecimento dão corpo a essa abordagem, e por isso também falamos em *abordagens baseadas em conhecimento*. A influência da perspectiva racionalista para o ambiente da IA se dá pela substituição de conhecimento inato pela inclusão manual de conhecimento nos sistemas pela programação. Por exemplo, a inclusão de uma regra linguística que indique que se a palavra “a” antecede um verbo ela é uma preposição.

Para o *empirismo*, o conhecimento humano deriva da experiência; aprendemos vivendo, aprendemos estando no mundo – não há conhecimento inato. De um ponto de vista cognitivo, apesar de também negarem que o aprendizado parta de um estado “tábula rasa”, abordagens empiristas responderão ao argumento da pobreza de estímulos – crucial para o gerativismo – de uma outra maneira: ao invés de postular princípios e procedimentos linguísticos específicos e inatos, esta corrente acredita que o cérebro é dotado de mecanismos probabilísticos gerais, responsáveis por mecanismos de associação, de reconhecimento de padrões, e de generalização. Para o empirismo, somos bons em generalizar, mesmo com dados esparsos (uma outra maneira de ver a pobreza de estímulo), e é isso o que nos possibilita viver em um mundo repleto de incertezas.

Uma breve nota explicativa, para que as ideias do parágrafo anterior não se percam: *argumento da pobreza de estímulos* é o que motiva a postulação da hipótese inatista. Segundo este argumento, os dados de linguagem – o estímulo linguístico – a que as crianças têm acesso é escasso e variável, não sendo suficiente para explicar o domínio de uma habilidade complexa como a linguagem humana em tão pouco tempo. Desse ponto de vista, a explicação possível é dotar biologicamente o ser humano com um aparato cognitivo, inato, especialmente dedicado à linguagem humana. Uma “gramática universal”.

O PLN/IA empirista é representado por algoritmos de aprendizado de máquina estatísticos e redes neurais – algoritmos que possibilitam aos computadores executar uma tarefa automaticamente sem que tenham sido programados explicitamente para isso, com base apenas nos *dados* fornecidos.

A abordagem empirista foi dominante entre 1920 e 1960, e desde os anos 1990 vem se tornando hegemônica na IA e no PLN. A supremacia se deve não apenas à maior disponibilidade de dados, mas também ao desenvolvimento de métodos estatísticos muito mais sofisticados que os modelos dos anos 1940-1950 e ao aumento considerável da capacidade de processamento computacional, capaz de lidar com volumes colossais de dados de forma paralela. Tudo isso tem permitido grandes avanços no tratamento da incompletude e da incerteza (ou dos dados escassos).

Na Linguística, podemos rastrear abordagens empiristas na análise de palavras com base nos seus padrões de co-ocorrência, como vemos na passagem do linguista britânico J.R. Firth: “You shall know a word by the company it keeps” (“Você conhece uma palavra pela companhia que mantém” ou “diga-me com quem anda, e te direi quem és”). A forte crítica de Chomsky contribuiu para o declínio deste tipo de abordagem, mas o desenvolvimento e a popularização dos computadores a partir dos anos 1980s associada à disponibilização de textos em formato eletrônico (corpus) tornou possível armazenar e processar grandes volumes de texto. Consequentemente, viabilizou-se a análise de

padrões da língua de uma maneira confiável - computadores são bons em analisar padrões.

De fato, é possível comparar a importância do corpus para os estudos linguísticos à importância do microscópio ou do telescópio para áreas como Biologia e Astronomia, respectivamente. Com essa nova ferramenta (que já não é tão nova assim), é possível observar a língua de uma perspectiva diferente, até então inacessível. Somos capazes de “ver” fenômenos que não conseguíamos observar antes, e por isso seu enorme potencial para enriquecer os estudos linguísticos.

Um ponto de vista empírico e quantitativo sobre a língua nos informa características nem sempre abordadas na formação linguística, mas relevantes para compreender a limitação de abordagens de PLN baseadas exclusivamente (ou majoritariamente) em regras. A principal delas é familiar para quem já explorou grandes corpora: sempre temos muitos casos com pouca frequência e poucos casos com muita frequência. De um ponto de vista prático: quando observamos a distribuição das palavras em um corpus, sempre teremos 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".

Podemos tomar como exemplo o livro Dom Casmurro, que tem cerca de 65 mil palavras ao todo: as palavras com frequência menor ou igual a 7 correspondem a cerca de 87% de todas as palavras do livro. Por outro lado, as palavras mais comuns (que aparecem 100 ou mais vezes) respondem por apenas 1.4% de todas as palavras do livro. Se, ao invés de todas as palavras, nos limitarmos a apenas uma classe, por exemplo os verbos, vemos que pouquíssimos – apenas 19 verbos, que não chegam a 2% de todos os verbos do livro – aparecem mais de 100 vezes. Já verbos que aparecem pouco, com até 5 ocorrências apenas, respondem por 75% de todos os verbos da obra. E quase 40% de todos os verbos aparecem apenas uma vez (40% de todos os verbos são *hapax legomena*; e 50% de todas as palavras são *hapax legomena*, ou seja, das 65 mil palavras, metade aparece só uma vez).

Diferentemente do que se poderia imaginar, essa imensa variedade não decorre do fato de estarmos analisando uma obra literária. Se formos para textos jornalísticos, o padrão de distribuição se mantém: palavras raras, com frequência de até 10 ocorrências, correspondem a 96% das palavras do corpus CHAVE, composto por notícias de jornal e que contém 99.2 milhões de palavras. Cerca de metade das palavras aparece apenas uma vez e palavras frequentes, que aparecem 100 ou mais vezes, correspondem a apenas 2.6% do total de palavras.

Existe um padrão na distribuição: pouquíssimos 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 estrutura dos sintagmas nominais. Em cerca

da metade dos casos, temos sintagmas compostos por uma estrutura única. Por outro lado, para que o padrão fique visível, é preciso que o conjunto de dados seja grande. Este tipo de fenômeno foi previsto 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).

Quando ficamos cientes desta propriedade distribucional, conseguimos entender melhor por que as regras linguísticas são importantes, mas até um certo ponto: as regras capturam regularidades, e regularidades só existem se existe repetição. Como uma boa parcela da língua não se repete, é difícil capturar regularidades. Por outro lado, se descartamos os casos com frequência 2 ou 1, estaremos olhando para uma língua mutilada, da qual uma imensa parte foi dispensada.

Moral da história: não importa a quantidade colossal de dados que tenhamos à disposição, sempre teremos casos raros e não previstos. Por isso, abordagens baseadas em regras serão limitadas, e por isso também o aprendizado de máquina começa a apresentar bons resultados: desde que haja dados, algoritmos estão cada vez melhores em prever eventos raros.

3.2 A Linguística em perspectiva (ou: provocações linguísticas)

Desde que começou a obter resultados utilizáveis de um ponto de vista aplicado/comercial, o PLN tem tido uma participação modesta da Linguística, o que de forma alguma impediu seu avanço, muito pelo contrário. É conhecida no PLN (mas não na Linguística) a afirmação feita no final dos anos 1980 por um pesquisador da IBM, “Cada vez que demito um linguista, o desempenho do meu sistema melhora”, motivada pelos bons resultados que vinham sendo obtidos utilizando apenas modelos estatísticos.

Para tentar entender melhor a dificuldade da linguística para dialogar com o PLN, voltaremos muito no tempo.

Quando, na tradição ocidental, têm início as primeiras reflexões sistemáticas sobre a linguagem, não era a linguagem que estava no centro das atenções, mas o raciocínio humano, a nossa racionalidade. A linguagem interessava não por ela, linguagem, mas apenas secundariamente, como via de acesso às estruturas (lógicas) do nosso *pensamento*. Por isso Aristóteles dá importância especial à lógica, vista como um reflexo de como funciona o pensamento humano. Com Aristóteles, o estudo sobre a linguagem passa a interessar na medida em que auxiliaria a análise lógica dos silogismos, que seriam, por sua vez, uma manifestação da faculdade racional humana, esta sim, alvo de interesse.

Segundo essa visão, a lógica é um *instrumento para o conhecimento*, e estaria baseada no silogismo – raciocínio formalmente estruturado que supõe certas premissas colocadas previamente para que haja uma conclusão necessária. Daí que, para tentar compreender o que são o conhecimento e a racionalidade humana, seria preciso estudar a estrutura dos silogismos e das proposições (ou: estudar a linguagem que estrutura os silogismos e proposições). Uma proposição, na lógica de Aristóteles, é um tipo especial de frase, composta de um *sujeito*, um *verbo* e um *atributo*, e passível de ser julgada como verdadeira ou falsa. “Sócrates é mortal”, ou “Sócrates canta” são proposições. Em

resumo: a linguagem interessa porque é útil para a análise dos silogismos e proposições, e esta análise é útil porque, indiretamente, permitiria o *acesso à faculdade racional humana*. (E lembremos que, cronologicamente, os *lógicos* vieram antes dos *gramáticos*, e tinham objetivos distintos: para os primeiros, interessava a análise lógica das proposições, e, aliás, foi por isso que postularam as classes dos nomes e dos verbos, que irão fornecer os elementos para o *sujeito* e o *predicado* das proposições; para os segundos, interessava a interpretação de textos literários, não havendo necessidade de analisar frases em termos de suas partes constitutivas.)

Também vem desta mesma tradição a aposta em uma separação clara entre linguagem e mundo. A função primeira da linguagem seria representar, ou relacionar, *realidade e entidades mentais*. E o fato de essa forma de pensar (a linguagem representa o mundo) nos parecer óbvia apenas evidencia o peso da tradição. Assim, diferentes línguas corresponderiam apenas superficialmente a diferentes modos de nomeação da realidade, que afinal é a mesma para todos, visto que vivemos em um mesmo mundo e somos os mesmos (somos humanos).

A força dessa tradição não é pouca. O interesse nas proposições nos ajuda a entender a preferência pelos enunciados declarativos e pelo sentido literal. Essa preferência é necessária para garantir a consistência interna e validade das inferências. Podemos comparar (I) e (II):

(I)	(II)
Todo homem é mortal	Todo açúcar é alimento
Sócrates é homem	Frutose é açúcar
Sócrates é mortal	Frutose é alimento

Por que a conclusão do segundo silogismo é estranha, difícil de julgar quanto a valores de V ou F, apesar de as premissas serem verdadeiras? Temos uma polissemia (vários sentidos) em *açúcar*, foco de nosso interesse agora, que tanto é *alimento* quanto *composto químico*. Mas dificilmente pensaríamos em *açúcar* como um termo polissêmico (e voltaremos à polissemia em 6.4). Para que o raciocínio se estruture da maneira correta, é preciso evitar a multiplicidade de sentidos e a imprecisão.

A noção de estabilidade também é central para esta abordagem: conhecimento *verdadeiro* é aquele que não varia, que independe de ponto de vista. Assim como aquilo que é *universal*. O que oscila e pode variar conforme o ponto de vista é a *opinião*.

Tanto tempo depois, o estudo científico da linguagem continua interessado na linguagem apenas secundariamente, como meio de acesso às faculdades mentais. Podemos reconhecer essas ideias no Curso de Linguística Geral (Saussure, 1917), quando se afirma que o “estudo da linguagem comporta, portanto, duas partes: uma essencial, tem por objeto a língua (...); esse estudo é unicamente psíquico” (1917: 27) e na proposta da linguística gerativa/biolinguística de Noam Chomsky, quando afirma que a “perspectiva biolinguística atualiza abordagens que remontam à tradição filosófica aristotélica, em relação ao que foi posteriormente interpretado como entidades mentais.” (Chomsky 2017:5).

Quando passa a ter como objeto a língua efetivamente usada, vinda das compilações de corpora, o PLN se depara com “problemas” que são fenômenos típicos da língua cotidiana, problemas que envolvem temas talvez considerados básicos demais, específicos ou simplórios demais para linguistas (isto é, problemas cuja solução não irá contribuir em nada para teorias gerais de linguagem ou para iluminar nosso entendimento da relação entre linguagem e mente). Assim, o PLN não encontrava na Linguística as respostas para aquilo que procurava e a Linguística, por sua vez, demonstrava pouco interesse pelos problemas de linguagem surgidos no âmbito das aplicações de PLN, porque linguistas, em sua maioria, estavam preocupados com outras questões.

Dentre várias outras influências, o legado grego nos deixa (a) a preferência pela estabilidade (dos conceitos, das representações do mundo, da verdade; e voltaremos a tratar de estabilidade e instabilidade em 6.4, com representação de sentidos e vetores de palavras contextuais); (b) a preferência pelo declarativo e literal (o que não deixa de ser um desdobramento do ponto anterior) e (c) o privilégio da dimensão cognitiva/mental no estudo da linguagem. Se deixamos o item (c) de lado, porque diretamente associado à linguagem, podemos generalizar (a) e (b) e afirmar que o grande legado é a preferência pela *verdade*.

E esta herança nos chega porque a filosofia de inspiração aristotélica corresponde ao lado dos vencedores. Mas havia um outro lado, o dos sofistas, do qual pouco sabemos porque a história nos é contada pela voz dos vencedores. Mas sabemos que valorizavam a multiplicidade e a dinamicidade em oposição à universalidade e à estabilidade. Para os sofistas, no lugar da verdade está a *eficácia*.

No livro *A Obsessão do Fogo*, Umberto Eco nos lembra que bibliotecas, feitas de madeira, tinham enorme tendência a pegar fogo:

Ainda lemos Eurípides, Sófocles, Ésquilo, que consideramos os três grandes poetas trágicos gregos. Mas quando Aristóteles na *Poética* (...) cita os nomes dos seus mais ilustres representantes, não menciona nenhum dos três. Aquilo que perdemos era melhor, mais representativo do teatro grego, do que o que conservamos? (...) Consolar-nos-emos, sonhando que entre os rolos de papiro desaparecidos no incêndio da biblioteca de Alexandria, e de todas as bibliotecas desfeitas em cinzas, repousavam prováveis ridículas, obras-primas de mau gosto e estupidez?” (2009:12-13)

E se os fragmentos que nos chegaram, que constituem a nossa formação e o nosso senso comum fossem outros? E se o que interessasse da linguagem, em um primeiro momento, não fossem as proposições e silogismos, mas a eficácia retórica? É possível que PropBanks (bancos de *proposições*, e falaremos deles na seção 6.3) não tivessem estado tão cedo na ordem do dia, e que o PLN tivesse começado pela análise de opinião e sentimento, e não pela extração de informação. É possível que a distribuição das palavras em classes morfossintáticas (as partes do discurso) não existisse da forma como a conhecemos, já que as classes de palavras surgem, inicialmente, para dar conta do vocabulário usado nas proposições. Mas essas são apenas especulações, trazidas para lembrar que tudo, sempre, poderia ser diferente (e voltaremos a isso no capítulo final).

4. Recursos lexicais em foco: léxicos computacionais, ontologias

Assim como nos estudos linguísticos, o papel de léxicos no PLN é fornecer informação sobre as palavras. Um léxico computacional tanto pode ser um *recurso* “independente”, isto é, que existe autonomamente, como um componente integrado a um sistema.

Além das palavras propriamente, um léxico (computacional ou não) pode ter associado a cada palavra informações como sentido, propriedades morfológicas (número, gênero, pessoa, tempo ou modo), classe (substantivo, verbo, preposição etc), relações morfológicas (relações flexionais e derivacionais), relações semânticas (hiperonímia; sinonímia, antonímia), polaridade (uma palavra é considerada positiva, como *festa* ou *amar*) ou negativa (como *doença* ou *odiar*), padrões sintáticos com que ocorre (por exemplo, “sofrer de algo” vs “sofrer para fazer algo”), entre outras informações possíveis.

E, assim como teorias linguísticas irão debater a respeito de quais tipos de informação estão associados aos elementos de um léxico, léxicos computacionais podem conter informação de diferentes naturezas, conforme a aplicação (ou motivação) subjacente à sua criação.

A figura abaixo mostra uma parte do OpLexicon, um léxico criado para auxiliar a tarefa de análise de sentimento (falaremos dela na seção 6.9). O OpLexicon é composto por uma lista de palavras classificadas com a sua categoria morfológica, polaridade e a indicação de como a classificação foi atribuída.

boçal	adj	-1	A
bom-samaritano	adj	1	A
devoto	adj	0	A
monótona	adj	-1	M
dar as=costas a	vb n prp	-1	A

Exemplo do OpLexicon

Vejamos como estas informações estão codificadas:

Coluna 1: a palavra propriamente. A última entrada da figura, *dar as costas a*, indica que não apenas unidades simples podem funcionar como uma entrada do léxico (entrada lexical), mas também unidades maiores.

Coluna 2: a classe gramatical da palavra.

Coluna 3: a polaridade (ou orientação semântica) da palavra: neutra (0); positiva (1) e negativa (-1).

Coluna 4: forma atribuição das classificações, isto é, esta coluna indica se a codificação das informações nas colunas anteriores foi feita de forma manual (M) ou automática (A).

As duas primeiras colunas codificam elementos gerais, que costumam integrar qualquer léxico. A terceira coluna codifica informação específica deste tipo de léxico. A última coluna traz um dado adicional para quem pretende usar o léxico, indicando o grau de confiança nas demais colunas.

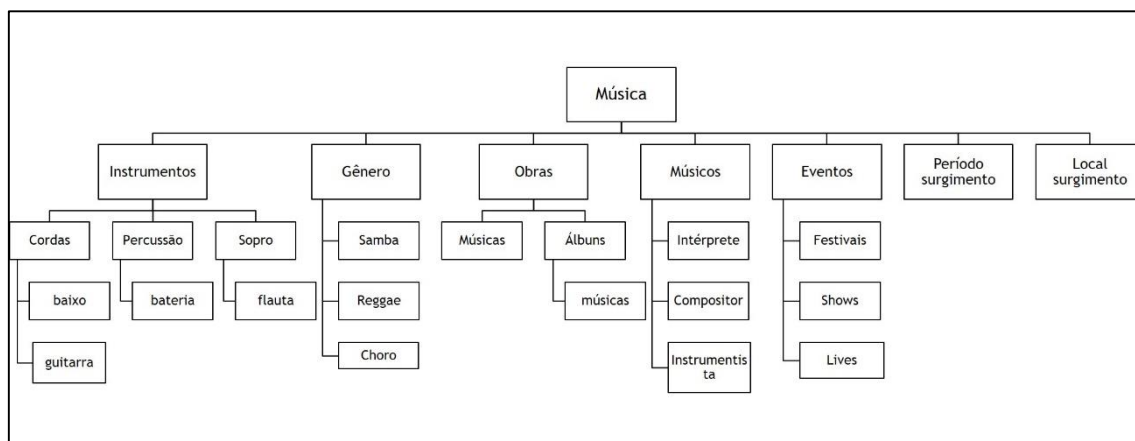
Léxicos podem ser ainda mais simples. Por exemplo, um léxico do corpo humano pode conter uma lista com palavras do corpo humano; um léxico de verbos de elocução apenas uma lista com verbos que sinalizam a presença de um discurso relatado. Comum a todos, a ideia de estruturação como uma lista.

Para fazer referência a objetos que não apenas listam, mas *organizam* e *estruturam* o conhecimento, o PLN utiliza o termo *ontologia*. Ontologias, assim, representam o conhecimento de uma maneira formalmente estruturada, e não como uma lista de termos (como os léxicos). Ser *formalmente estruturada* indica que os termos/conceitos de uma ontologia não são listados livremente, mas obedecem aos princípios de estruturação da ontologia.

Ontologias podem ser entendidas de duas maneiras: em sentido amplo e em sentido estrito. Ontologias em sentido amplo referem-se a qualquer maneira de organizar o conhecimento. Uma taxonomia, por exemplo, que é uma lista de termos organizada como uma hierarquia (podemos nos lembrar das taxonomias da Biologia), pode ser entendida como uma ontologia. Em sentido estrito, ontologias referem-se a um certo tipo de organização do conhecimento, com classes, relações pré-definidas entre as classes e instâncias dessas classes, além das relações hierárquicas.

Quando se trata de formalizar o conhecimento específico de uma área, como por exemplo a Medicina ou uma de suas subáreas, ontologias são criadas com a participação de especialistas (das áreas). São eles que irão determinar as classes, relações e instâncias (instâncias são exemplos das classes, por exemplo *Brasil* é uma instância da classe PAÍSES).

Podemos arriscar a elaboração de uma ontologia simplificada do domínio da Música, ilustrada no quadro abaixo.



O quadro contém as classes da ontologia e relações taxonômicas (relações de inclusão) entre essas classes: guitarra é um instrumento; choro é um gênero; um compositor é um músico; um show é um evento. É possível ainda adicionar relações entre as classes, como as seguintes:

CLASSE	RELAÇÃO	CLASSE
Instrumentista	TOCAR	Instrumento
Instrumento	PARTE DE	Gênero
Compositor	COMPÕE	Obra
Intérprete	EXECUTA	Obra
Músico	PARTE DE	Gênero
Instrumento	PARTE DE	Obra
Gênero	INCLUI	Obra
Gênero	PARTICIPAR DE	Evento
Músico	PARTICIPAR DE	Evento

Ontologias têm uma estrutura fixa. Informações que não se enquadrem em nenhuma das classes são deixadas de fora, a não ser que a ontologia seja reformulada.

Na IA, o interesse na construção de ontologias é típico do trabalho feito nas décadas de 1980 e 1990, mas o trabalho com ontologias continua bastante ativo. Naquele momento inicial, ontologias – que surgem com os gregos há mais de 25 séculos para designar *o estudo do ser* – eram sobretudo recursos que contêm e organizam conhecimento relativo ao nosso senso comum (um conhecimento geral sobre o mundo). Segundo a visão de IA da época, sistemas não eram bons porque não havia maneiras de organizar e representar o mundo, e por isso deveríamos investir na construção de ontologias. Com a hegemonia do aprendizado de máquina e do conhecimento obtido apenas a partir dos dados, ontologias (desde sempre associadas ao PLN simbólico de influência racionalista) cada vez mais são utilizadas como uma maneira de representar o conhecimento específico de um domínio, ou uma maneira de organizar a informação de uma área, e podem integrar sistemas para viabilizar raciocínio lógico ou alimentar, com suas classes e termos, a anotação de corpus, como veremos em 6.5 e 6.6. Na página eletrônica, em *@Ontologias e realidade*, trago mais alguma informação histórica sobre o surgimento das ontologias, além de provocações relativas à relação entre ontologias, linguagem e a possibilidade de descrição do mundo de um ponto de vista externo.

Bases de conhecimento são recursos muito próximos a ontologias – há quem considere ontologias um tipo de base de conhecimento. Diferentemente de ontologias, bases de conhecimento podem ser mais flexíveis quanto às relações que codificam. Além disso, são alteráveis, isto é, os fatos contidos podem ser alterados, diferentemente das ontologias. Ontologias podem funcionar como um esqueleto para bases de conhecimento. Uma maneira de entender a diferença é pensar que as bases de conhecimento contêm declarações como

- 1- Luiz Gonzaga compôs a música Asa Branca
- 2- Asa Branca é do gênero Baião

A ontologia codifica a relação entre as classes (Luiz Gonzaga é músico; Asa Branca é Obra; Asa Branca é Baião), e sistemas de raciocínio poderão concluir que “Luiz Gonzaga compõe Baião”.

Os projetos Wikidata e DBpedia são exemplos de bases de conhecimento. A Wikidata é uma grande base de conhecimento que pode ser editada por qualquer pessoa, assim como a Wikipédia. A Wikidata é dinâmica, e contém bilhões de fatos sobre milhões de tópicos.

A DBPedia (o DB vem de *Data Base*, *base de dados*) é um projeto que visa extrair e organizar em um banco de dados a informação factual contida nas caixas de informação da Wikipédia (apresentadas em 2.2). Por ser editada por pessoas, não há uniformidade nas classes contidas nas caixas, e os mesmos conceitos podem aparecer de forma diferente (por exemplo |local_de_nascimento= e |lugar_de_nascimento), o que faz com que uma mesma informação precise ser procurada de duas formas diferentes. Subjacente à DBPedia existe uma ontologia, descrita como “o coração da DBpedia”, que permite a estruturação da informação.

Bases de conhecimento e ontologias são recursos típicos do PLN (e da IA) baseada em conhecimento, e mesmo com os avanços das redes neurais continuam como recursos relevantes quando se pensa em informação e raciocínio. Técnicas de PLN baseado em regras ou redes neurais podem ser utilizadas na identificação e extração de certos tipos de informações em textos, que por sua vez podem alimentar bases de conhecimento e ontologias, como será exemplificado na seção 6.6. Do mesmo modo, sistemas de pergunta e resposta (baseados em redes neurais ou não) poderão acessar o conhecimento codificado em bases como a Wikidata.

Léxicos e ontologias são exemplos de recursos que lidam com palavras (ou unidades lexicais) e são também extremos no que se refere à organização: léxicos têm a estrutura de uma lista, ontologias em sentido estrito têm relações rigidamente especificadas. Entre um e outro, o PLN tem ainda outras formas de representação do conhecimento linguístico, como a WordNet, que se apresenta como uma “base de conhecimento lexical”, e a FrameNet. É o que veremos a seguir.

4.1 Wordnets

Uma wordnet, como o nome indica, é uma rede (*net*) de palavras (*word*). Porém, a unidade básica de uma wordnet não é a palavra, mas o *synset*, um conjunto (*set*) de sinônimos (*syn*). O substantivo *açúcar*, por exemplo, participa de dois synsets (dentre outros): um contém [*hidratos de carbono*, *carboidrato*, *açúcar*] e outro contém [*açúcar*, *açúcar refinado*]. De forma precisa, as redes de uma wordnet conectam *synsets* segundo diferentes tipos de relações, e seu principal objetivo é dar suporte a tarefas que envolvem a análise automática de textos. Algumas das relações que se estabelecem entre os synsets estão abaixo:

Entre substantivos

hiperonímia: Y é um hiperônimo de X se todo X for um Y (*instrumento* é um hiperônimo de *pandeiro*)

hiponímia: Y é um hipônimo de X se todo Y for um X (*pandeiro* é um hipônimo de *instrumento*)

termos coordenados: Y é coordenado de X se X e Y compartilham um hiperônimo (*pandeiro* é coordenado de *tamborim* e *tamborim* é coordenado de *pandeiro*)

meronímia: Y é um merônimo de X se Y for uma parte de X (*corda* é merônimo de *violão*)

holonímia: Y é um holônimo de X se X for parte de Y (*violão* é um holônimo de *corda*)

Entre verbos

hiperonímia: o verbo Y é um hiperônimo do verbo X se a atividade X for um Y (*falar* é um hiperônimo de *gritar*)

implicação: o verbo X é acarreta Y se ao fazer X você necessariamente está fazendo Y (*roncar* acarreta em *dormir*)

termos coordenados: aqueles verbos que compartilham um hiperônimo comum (*gritar* e *sussurrar* são termos coordenados)

Uma wordnet combina, por um lado, informações tradicionais de dicionário, como definições e exemplos de uso (chamada glosa) e, por outro lado, uma organização adequada para a utilização computacional, o que facilita sua utilização como base de conhecimento léxico-semântico. Embora a WordNet seja frequentemente referida como uma ontologia lexical, esta não é uma posição compartilhada por seus criadores, que a veem como uma *base de dados lexical*.

Além de ser um recurso do PLN, uma wordnet pode ser melhorada fazendo uso de técnicas de PLN. Marti Hearst (1992, 1998), por exemplo, aproveitou o insight de que determinados padrões léxico-sintáticos poderiam sistematicamente expressar determinadas relações semânticas, e com isso propôs a identificação automática de relações de hiponímia/hiperonímia em corpus. Ela trabalhou com os seguintes padrões (originalmente em inglês, mas se aplicam perfeitamente bem ao português):

- (i) SN0 tais como SN1 {, SN2 ... , (e | ou) SNi }
- (ii) tais SN0 como {SN ,}* {(e| ou)} SN1
- (iii) SN1 {, SN}* {,} ou outros SN0
- (iv) SN1 {, SN}* {,} e outros SN0
- (v) SN0 {,} incluindo {SN ,}* {e| ou} SN
- (vi) SN0 {,} especialmente {SN ,}* {e | ou} SN

onde SN0 corresponde ao sintagma nominal (SN) hiperônimo e os demais SNs (SN1, SN2...SNi) a SNs hipônimos:

- a) O estágio adulto é mais específico de [mamíferos] tais como [equinos], [antas] e [capivaras] e, eventualmente, ...
Relações extraídas: mamífero HIPERÔNIMO DE equinos; mamífero HIPERÔNIMO DE antas; mamífero HIPERÔNIMO DE capivaras
- b) Praticar [exercícios], [ioga], [meditação] e outras [atividades físicas] é importante para...
Relações extraídas: atividades física HIPERÔNIMO DE exercícios; atividades física Hipерônimo de ioga; atividades física HIPERÔNIMO DE meditação

Hearst propõe um algoritmo de descoberta de padrões que consiste de 4 etapas:

- decidir qual a relação lexical de interesse (no caso dela, a relação foi de hiperonímia)
- selecionar uma lista de pares de palavra na qual a relação esteja expressa: por exemplo, o par *mamífero-gato*;
- extrair sentenças do corpus em que ambas as palavras (*mamífero* e *gato*) apareçam, registrando o contexto lexical e sintático em que foram encontradas;
- encontrar semelhanças entre esses contextos e tentar generalizar: contextos comuns levam a padrões que indicam a relação de interesse.

Seguindo esta estratégia, podemos chegar a outros padrões:

- c) Existem dois tipos de [cromossomos gigantes]: [cromossomos politênicos] e [cromossomos plumulados]

Relações extraídas: cromossomos gigantes HIPERÔNIMO DE cromossomos politênicos; cromossomos gigantes HIPERÔNIMO DE cromossomos plumulados

- d) O [iogurte] contém [bactérias benéficas]

Relações extraídas: bactérias benéficas MERÔNIMO DE iogurte

Importante notar que este método é altamente dependente de uma boa análise sintática e que hoje em dia a análise sintática é muito mais confiável do que quando o trabalho foi feito, há mais de 25 anos. A principal crítica a esta abordagem é sua pouca abrangência, isto é, provavelmente existem muito mais relações semânticas nos textos do que aquelas explicitamente expressas nos padrões. Por outro lado, a metodologia é bastante precisa e apresenta a vantagem de oferecer grupos de palavras já rotulados com um hiperônimo, e não simplesmente grupos de palavras, o que a faz ser utilizada na criação semi-automática de ontologias e na detecção de relação entre entidades (que veremos em 6.6).

A WordNet original, para a língua inglesa, começou a ser desenvolvida de forma totalmente manual em 1985, e desde então está em constante atualização.

Wordnets de língua portuguesa

A WordNet.BR (Dias-da-Silva, 2010) é uma wordnet para a variante brasileira do português. Em sua elaboração, a equipe do projeto reaproveitou material disponível em outras fontes lexicográficas, como dicionários gerais, dicionários de sinônimos e antônimos, um dicionário analógico e um dicionário de verbos do português. Os resultados da fase inicial do projeto estão disponíveis para consulta e download sob o nome de TeP (Thesaurus eletrônico para o Português do Brasil).

Já a OpenWordnet-PT (de Paiva et al., 2012) se dedica à língua portuguesa em geral e não apenas àquela falada no Brasil. Esta é uma wordnet que foi criada automaticamente para o português a partir do inglês, e desde sua criação ela vem sendo aperfeiçoada de forma manual e semi-automática, usando para isso ferramentas, recursos e conhecimentos linguísticos. As melhorias consistem principalmente em adições (ou correções) linguisticamente motivadas, levando em conta dados de grandes corpora. A filosofia da OpenWN-PT é manter um alinhamento próximo com a wordnet original, mas removendo os maiores erros produzidos pelos métodos automáticos.

A língua portuguesa conta ainda com o PULO (Ontologia Lexical Unificada para o Português) (Simões e Guinovart, 2014), cujo objetivo é, nas palavras dos autores, “ser uma variante da wordnet para a língua portuguesa, enriquecida com outra informação que de algum modo não a torne incompatível com wordnets de outras línguas.”. Temos também a Onto.PT. (Gonçalo Oliveira, 2013), uma ontologia lexical estruturada de forma semelhante à WordNet e construída automaticamente, e o PAPEL (Gonçalo Oliveira et al., 2008), que tem como diferencial ter sido construído automaticamente a partir de um dicionário (a sigla PAPEL significa *Palavras Associadas Porto Editora Linguateca*). Assim como a OpenWordNet.PT, todos estão disponíveis para consulta e para download, mas têm como público principal máquinas, e não pessoas.


As figuras 1a e 1b reproduzem telas da OpenWordnet-PT para a palavra “andar”.

word_pt

andar

Search

[\[Doc | API | Source | Activity | Login \]](#)



30 results found for 'andar'

RDF Type:

- ☐ VerbSynset (27)
- ☐ BaseConcept (8)
- ☐ CoreConcept (8)
- ☐ NounSynset (3)

Lexicographer file:

- ☐ verb.motion (26)
- ☐ noun.act (2)
- ☐ noun.artifact (1)
- ☐ verb.competition (1)

words (pt_BR):

- ☐ 2 (14)
- ☐ 3 (11)
- ☐ 1 (3)
- ☐ 4 (2)

words (en):

- ☐ 1 (20)
- ☐ 2 (6)
- ☐ 3 (2)
- ☐ 4 (2)

Frame:

- ☐ Somebody —s PP (17)
- ☐ Somebody —s (16)
- ☐ Something —s (6)
- ☐ Somebody —s something (4)
- ☐ Something is —ing PP (4)
- ☐ Somebody —s Adjective (1)
- ☐ Somebody —s somebody PP (1)
- ☐ Something —s something (1)

11. [01957529-v](#) ride, sit | **cavalgar, andar**
 - (sit and travel on the back of animal, usually while controlling its motions; "She never sat a horse!"; "Did you ever ride a camel?"; "The girl liked to drive the young mare")
12. [01924023-v](#) tiptoe, tip, tippytoe | **andar na ponta dos pés, andar**
 - (walk on one's toes)
13. [01904930-v](#) walk | **ir, caminhar, andar**
 - (use one's feet to advance; advance by steps; "Walk, don't run!"; "We walked instead of driving"; "She walks with a slight limp"; "The patient cannot walk yet"; "Walk over to the cabinet")
14. [02092476-v](#) retreat | **retirar-se, andar**
 - (move away, as for privacy; "The Pope retreats to Castelgondolfo every summer")
15. [03365991-n](#) floor, storey, story, level | **pavimento, piso, andar**
 - (a structure consisting of a room or set of rooms at a single position along a vertical scale; "what level is the office on?")
16. [01841079-v](#) travel | **ir, viajar, andar**
 - (undergo transportation as in a vehicle; "We travelled North on Rte. 508")
17. [01839538-v](#) ride | **andar, ir, viajar**
 - (move like a floating object; "The moon rode high in the night sky")
18. [01919391-v](#) march | **marchar, andar, caminhar**
 - (walk fast, with regular or measured steps; walk with a stride; "He marched into the classroom and announced the exam"; "The soldiers marched across the border")
19. [02102002-v](#) move_around, travel | **andar, viajar, ir**
 - (travel from place to place, as for the purpose of finding work, preaching, or acting as a judge)
20. [02102398-v](#) ride | **ir, andar, viajar**
 - (sit on and control a vehicle; "He rides his bicycle to work every day"; "She loves to ride her new motorcycle through town")

[« Previous](#) [1](#) [2](#) [3](#) [Next »](#)

Figura 1a: Synsets associados à palavra *andar*, conforme a OpenWordnet-PT

Como podemos ver pela parte superior da tela, o recurso contém 30 resultados – ou seja, 30 *synsets* – com a palavra *andar*. A imagem refere-se aos *synsets* que estão na 2ª tela, os *synsets* de 11 a 20. A parte central da tela lista os *synsets* e a parte esquerda lista informações adicionais. Começaremos por ela.

Dos 30 *synsets* que, em português, contém a palavra “andar”, 27 são *synsets* verbais (*VerbSynset*) e 3 são *synsets* nominais (*NounSynset*). Vemos também que 8 são considerados básicos e 8 são considerados nucleares, mas como esta informação tem a ver com opções técnicas da OpenWordnet-PT, podemos deixá-la de lado aqui. Na parte chamada *Lexicographer file* observamos que, dentre os *synsets* com a palavra *andar*, a grande maioria se refere a verbos de movimento (*verb motion*). A seção “Frame” indica os padrões de subcategorização associados a *andar*. No primeiro caso, por exemplo, temos que “Alguém (*Somebody*) anda com/para/de ...”, isto é, temos que o *andar*, em 17 casos tem um sujeito humano, e recebe complemento introduzido por preposição. Na 2ª linha, temos a indicação de que, em 16 casos, o *andar* tem um sujeito humano, mas o verbo é intransitivo; em na 3ª linha, casos em que o sujeito não é humano (*Something*) e o verbo é intransitivo. Todas essas informações do lado esquerdo da tela funcionam como filtros: ao selecionar a este 3º frame (*Something –s*), veremos apenas os *synsets* com essa característica. Ao selecionar *Noun Synset*, apenas os *synsets* nominais.

Na parte central da tela, vemos que cada *synset* tem um identificador numérico (01957529-v no primeiro *synset* da tela, e o v no final indica que se trata de um verbo), responsável por indicar o alinhamento com a WordNet original. Se clicamos nele, temos informação específica do *synset*, conjugando informações da OpenWordNet-PT e da WordNet (Figura 1b).

word_pt

andar

Search

[Doc](#) | [API](#) | [Source](#) | [Activity](#) | [Login](#)

01924023-v

English

(walk on one's toes)

tiptoe ([drt](#)) • tip • tippytoe

Portuguese

Gloss: empty gloss

Ex.: empty example

andar na ponta dos pés • andar

Relations

- Lexicographer file: (verb.motion)
- Frame: (Somebody ---s PP Somebody ---s)
- RDF Type: (VerbSynset CoreConcept)
- Hyponym of: [[walk](#)]

External resources

- [OMW](#)
- [SUMO](#)
- [Princeton](#)
- [Linked Data](#)

Figura 1b: Detalhe de um dos *synsets* associados à palavra andar, conforme a OpenWordnet-PT.

A figura 1b é uma boa maneira de ver como se estruturam as redes de uma wordnet (ainda que existam algumas diferenças entre os modelos das diferentes línguas, são suficientemente parecidos para se chamarem wordnets). Ela indica que o *synset* 01924023-v está associado ao sentido de.... *andar nas pontas dos pés*, e este exemplo foi escolhido de propósito, para ilustrar um caso em que um *synset* do inglês não se alinha perfeitamente a um do português, visto que, diferentemente do inglês, não temos um verbo específico para *andar na ponta dos pés*, da forma que o fazemos, por exemplo, quando não queremos fazer barulho ao caminhar.

A tela também mostra o trabalho em andamento para a língua portuguesa: a glosa, que é a explicação, está vazia, assim como o exemplo (mas isso não prejudica o entendimento); e a tradução como *andar*, foi eliminada do *synset*, mantendo-se apenas *andar na ponta dos pés*. O contrário também existe – elementos do português que não têm um *synset* equivalente em inglês. Na figura abaixo vemos que *tamborim* e *pandeiro* pertencem ao mesmo *synset*, o que sabemos ser erro. Por outro lado, a OpenWordnet-PT é um projeto colaborativo, em atualização constante, e mão de obra disposta a ajudar é sempre bem-vinda.

1. [04387400-n](#) tambourine | **pandeiro, pandeireta, tamborim**

- (a shallow drum with a single drumhead and with metallic disks in the sides)

A seção de Relações (*Relations*) é importada de outros recursos e nos dá informações específicas (já comentadas) desse *synset*. A novidade é a presença de informação hierárquica: *andar na ponta dos pés* é um *synset* hipônimo (isto é, é uma especificação) do *synset* mais geral *walk* (*caminhar*). E é só clicar em *walk* para ver os caminhos semânticos desse *synset*.

4.2 FrameNets

A FrameNet é um recurso lexical e, ao mesmo tempo, a exemplificação de uma teoria: a teoria da Semântica de Frames, desenvolvida pelo linguista Charles J. Fillmore na década de 80. Podemos pensar em um *frame* como um quadro “conceitual”, como o

enquadramento de uma determinada experiência, no qual esta experiência é descrita de uma maneira altamente especializada. Um frame é decomposto em suas unidades constitutivas, os *elementos do frame*. Na Semântica de Frames, a constituição interna de um *frame* está intimamente relacionada à cultura, o que faz com que, ao menos teoricamente, nem sempre os *frames* sejam exatamente os mesmos entre diferentes línguas ou culturas.

O frame relativo a *cozinhar*, em nossa cultura, normalmente envolve (i) alguém que cozinha; (ii) o alimento que deve ser cozido; (iii) algo para reter a comida enquanto esta é cozida; e (iv) uma fonte de calor. Esses 4 elementos irão compor o frame associado a *cozinhar*.

No projeto FrameNet, o verbo *cozinhar* está associado aos frames APLICAR_CALOR, CRIAÇÃO_CULINÁRIA e ABSORÇÃO_DE_CALOR. Considerando apenas o frame APLICAR_CALOR, os elementos ALIMENTO, INSTRUMENTO DE AQUECIMENTO e RECIPIENTE são chamados *elementos de frame* (FEs) nucleares. Palavras que evocam esse frame, como *dourar*, são chamadas de *unidades lexicais* (LUs) do frame APLICAR_CALOR. Já o *frame* VINGANÇA tem os seguintes elementos de frame: PARTE QUE OFENDE, LESÃO, PARTE OFENDIDA, VINGADOR e PUNIÇÃO.

Como podemos observar, um frame também é uma maneira de organização do conhecimento. A palavra que evoca o frame, em geral, é o verbo, e os elementos que giram em torno do frame são os argumentos e modificadores deste verbo. Mas um nome (uma nominalização, como *retaliação*) ou um adjetivo também podem ser elementos evocadores de frame.

Frames são estruturas detalhadamente descritas em termos de propriedades sintáticas e semânticas, e a validação de um frame acontece por meio da consulta a um corpus. As figuras 2ª, 2b e 2c, retiradas do projeto FrameNet Brasil, apresentam os elementos associados ao frame APLICAR_CALOR, a figura 4 apresenta o elemento COZINHAR e, a seguir, um exemplo da aplicação (ou anotação) dos frames ABSORÇÃO_DE_CALOR e VINGANÇA:

- ... [Cook OS rapazes] ... *assaram* [Food seus peixes] [Heating_instrument na fogueira].
- [Avenger Eu] vou me *vingar* [Offender de vc] [Injury por isso]!

Em geral, as anotações são feitas em frases “avulsas”, ou seja, frases escolhidas para representar uma única unidade lexical, mas existem coleções de textos nos quais todas as palavras que evocam frames foram manualmente anotadas (Baker, 2017).

Segundo a página do projeto, a equipe da FrameNet definiu e detalhou manualmente mais de 1.000 frames, todos vinculados por um sistema de relações que associa frames mais gerais a outros mais específicos. Todo o material está disponível para consulta online.

Assim como a WordNet, a FrameNet foi originalmente criada para a língua inglesa, e hoje existem FrameNets para várias línguas, como espanhol, alemão, chinês e japonês e português, todas elas paralelas e alinhadas ao projeto original, em inglês.

Para a língua portuguesa, temos o projeto FrameNet Brasil (Salomão, 2009), realizado pela UFJF, também em andamento e alvo de constantes melhorias, e disponível para consulta online.

FrameNet Brasil Webtool 3.0 [FNBr-docker]

Reports Grapher PT Language Sign In

Frames

Search Frame Search LU

Aplicar_calor

Definição

Um **Cozinheiro** aplica calor ao **Alimento**, onde a **Configuração_de_temperatura** do calor e a **Duração** da aplicação podem ser especificadas. Um **Instrumento_de_aquecimento**, geralmente indicado por uma frase locativa, também pode ser expresso. Alguns métodos de cozedura envolvem a

Exemplo(s)

Elementos de Frame Nucleares

FE Core:

Alimento [Food]	Entidade na qual é aplicado o calor pelo Cozinheiro .
Configuração_de_temperatura [Temperature_setting] semantic_type: @temperature	A Configuração_de_temperatura do Instrumento_de_aquecimento para o Alimento .
Cozinheiro [Cook] semantic_type: @sentient	O Cozinheiro aplica calor ao Alimento .
Instrumento_de_aquecimento [Heating_instrument] semantic_type: @physical_entity	A entidade que fornece diretamente calor para o Alimento .
Recipiente [Container] semantic_type: @container	O Recipiente contém o Alimento ao qual o calor é aplicado.

FE Core set(s):

{Instrumento_de_aquecimento, Recipiente}

Figura 2a: Frame APLICAR_CALOR (FrameNet Brasil)

FrameNet Brasil Webtool 3.0 [FNBr-docker]

Reports Grapher PT Language Sign In

Frames

Search Frame Search LU

{Instrumento_de_aquecimento, Recipiente}

Elementos de Frame Não-Nucleares

Beneficiário [Beneficiary]	A pessoa para a qual o Alimento é feito
Co-participante [Co-participant] requires: Alimento	Co-participante é o gramaticalmente menos proeminente de dois alimentos aos quais o calor é aplicado pelo Cozinheiro .
Duração [Duration] semantic_type: @duration	Quantidade de Tempo que o calor é aplicado ao Alimento .
Finalidade [Purpose] semantic_type: @state_of_affairs	A Finalidade para a qual se realiza um ato intencional.
Grau [Degree] semantic_type: @degree	O Grau a que ocorre a aplicação de calor.
Lugar [Place] semantic_type: @locative_relation	O Lugar onde a aplicação de calor ocorre.
Maneira [Manner] semantic_type: @manner	Qualquer descrição do evento de culinária que não seja coberta por EFs mais específicos, incluindo efeitos secundários (silenciosamente, em voz alta) e descrições gerais comparando eventos (da mesma forma). Além disso, pode indicar características salientes de um Cozinheiro que também afetam a ação (presunçosamente, friamente, deliberadamente, avidamente, cuidadosamente).
Meio [Means] semantic_type: @state_of_affairs	O Meio pelo qual o calor é aplicado ao Alimento .
Mídia [Medium]	A substância através da qual o calor é aplicado ao Alimento .
Tempo [Time] semantic_type: @time	O Tempo em que ocorre a aplicação de calor.

Figura 2b: Frame APLICAR_CALOR (FrameNet Brasil)

Relações

É causativo de [Absorção_de_calor](#)

Herda de [Afetar_intencionalmente](#) [Atividade](#)

Veja também [Criação_culinária](#)

É usado por [Criação_culinária](#)

Unidades Lexicais

[dourar.v](#)

Figura 2c: Frame APLICAR_CALOR (FrameNet Brasil)

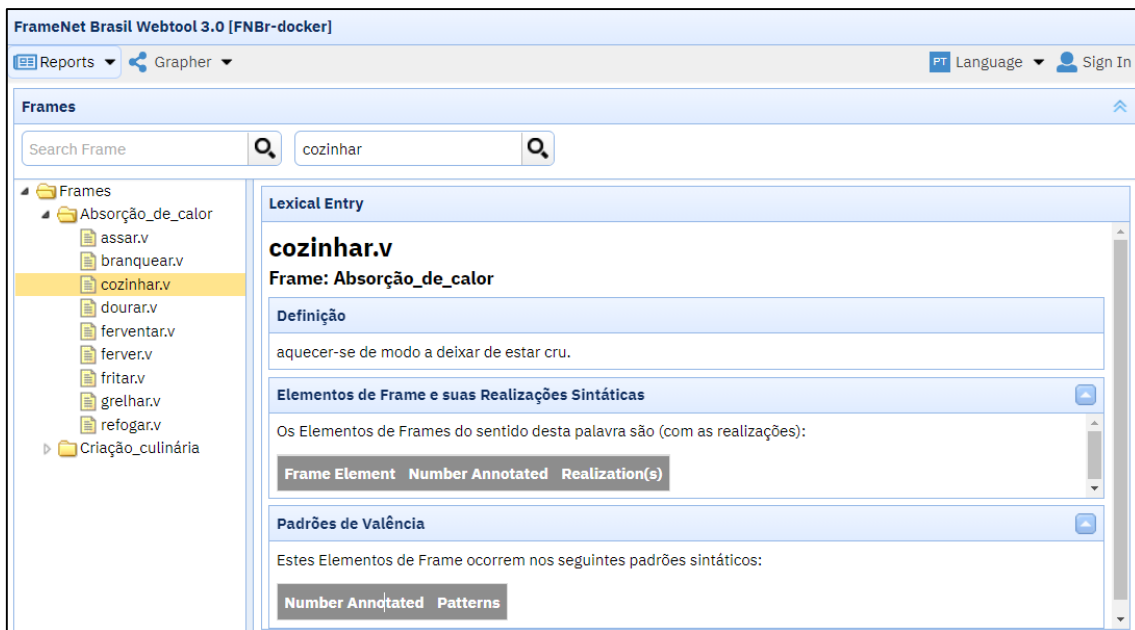


Figura 3: Unidade lexical COZINHAR (FrameNet Brasil)

4.3 VerbNets

A VerbNet (Kipper et al. 2006) apresenta-se como o maior léxico de verbos disponível online para a língua inglesa. Trata-se de um recurso que contém verbos agrupados de acordo com as suas propriedades sintáticas e semânticas, em uma ampliação da proposta original da linguista Beth Levin, de 1993. Levin apresenta classes que são “sintaticamente relevantes” e “semanticamente coerentes” (1993:22), o que traz como resultado uma proposta de 49 classes principais, que por sua vez são compostas por subcategorias. O objetivo principal é distribuir os verbos em classes semânticas segundo o seu comportamento sintático – e a opção de privilegiar a sintaxe sobre a semântica se deve, segundo a autora, à dificuldade de identificação de significados com base apenas na intuição ou na definição de dicionários (e veremos exemplos dessa dificuldade em 6.4).

Uma característica da VerbNet é estar alinhada a outros recursos lexicais, como as já mencionadas WordNet e FrameNet. Para facilitar o acesso a todo o material, foi criado um sistema que junta a informação de diferentes recursos, o *Unified Verb Index*. A figura 4 apresenta uma tela desse sistema para o verbo *cook* (*cozinhar*), e podemos reconhecer os frames da seção anterior.

SEARCH REQUEST: **[COOK]**

SEARCH:

VERBNET MEMBERS
 COOK: **BUILD-26.1**
 COOK: **COOKING-45.3**
 COOK: **PREPARING-26.3-1**
 COOK_UP: **PREPARING-26.3-1**

VERBNET CLASSES
 COOKING-45.3: **COOKING-45.3**

ONTOLOGIES SENSE GROUPINGS
 COOK: **COOK.V**
 COOK.N: **COOK.N.V**

PROPBANK
 COOK: **COOK.V**

FRAMENET
 COOK: **APPLY_HEAT**
 COOK: **COOKING_CREATION**
 COOK: **ABSORB_HEAT**
 COOK_UP: **INVENTION**
 COOK_UP: **COOKING_CREATION**

Figura 4: Tela do índice unificado de Verbos para o verbo *cozinhar*

Para a língua portuguesa temos a VerbNet-BR (Scarton e Aluisio, 2012), construída semi-automaticamente a partir de mapeamentos entre a VerbNet e a WordNet original, por um lado, e os alinhamentos entre a WordNet e a já mencionada WordNet.Br.

Para a construção da VerbNet.Br e seu alinhamento com a VerbNet original, as autoras partiram da hipótese de que as classes de verbos criadas por Levin têm um potencial multilíngue, isto é, podem ser aplicadas a outras línguas, o que as permitiu explorar as possíveis compatibilidades entre o inglês e o português. As autoras do projeto indicam, porém, que decidiram apenas traduzir as alterações sintáticas que se encaixassem “perfeitamente” no português, uma vez que o objetivo é começar pelos alinhamentos, novamente, “perfeitos”, para em seguida passar aos demais casos. A figura 5 traz uma tela da VerbNet.Br para o verbo andar.

[Home](#) [Busca](#) [Publicações](#) [Contato](#)

Membros da VerbNet.Br:
 andar: [continue-55.3](#) | [exist-47.1](#) | [lodge-46](#) | [other_cos-45.4](#) | [run-51.3.2](#) | [sustain-55.6](#)

Figura 5: Verbo andar (VerbNet.Br)

Os endereços dos recursos mencionados aqui estão na página eletrônica do livro, em @Ponteiros, na seção *Recursos Lexicais*.

5. Preparativos para um processamento computacional da língua (ou: pré-processamento também é processamento)

A fase de preparativos para o processamento automático de textos é chamada de pré-processamento (falamos em “pré-processamento do corpus”), mas este nome é enganoso, pois sugere que o verdadeiro “processamento” é o que vai acontecer depois. Embora nem sempre explicitada, e às vezes considerada trivial ou pouco importante, esta fase envolve a tomada de decisões linguísticas tão básicas quanto cruciais, já que irão determinar as unidades primeiras com que as demais ferramentas irão trabalhar: *palavra* e *frase*.

Entretanto, anterior à identificação de palavras e frases é a identificação de tudo o que não é texto para o PLN: tabelas, gráficos, figuras, fórmulas, cabeçalhos e rodapés etc. Quando mantidos, esses elementos são transformados em caracteres estranhos que são confundidos com palavras, atrapalhando o processamento automático.

No mundo do PLN, os procedimentos relativos à segmentação de um texto em palavras e frases são chamados, respectivamente, de *tokenização* e de *sentenciação*.

Tokenização: a etapa de tokenização consiste em segmentar o texto em unidades, chamadas *tokens*. Um *token* pode ser uma palavra ou algum outro sinal que compõe o texto, como os sinais de pontuação ou símbolos como \$. Por isso, em um texto, o número de tokens é superior ao número de palavras. Mas o que é uma palavra?

Para começar, devemos lembrar que nem *palavra*, nem *frase*, são objetos “naturais” (como nenhum outro objeto linguístico). Além disso, não há uma definição única do que seja *palavra*, e a Linguística opera com diferentes conceitos, relativos a diferentes perspectivas: a palavra de um ponto de vista fonológico, de um ponto de vista morfosintático e de um ponto de vista semântico. Como nem sempre haverá convergência entre esses diferentes pontos de vista, uma noção prática e de amplo uso é a de palavra gráfica. Ou seja, é *palavra* tudo aquilo que em um texto escrito aparece entre espaços em branco ou sinais de pontuação. Pelo critério gráfico, a frase anterior tem 20 palavras, e esta frase tem 14. A frase anterior com 20 palavras tem 22 tokens, e a frase seguinte com 14 palavras tem 17 tokens. Entretanto, apesar da aparente simplicidade, há espaço para discussão e decisão linguística em ambas as frases.

Deixar espaços em branco entre as palavras corresponde à possibilidade de pausa entre as unidades na fala. Trata-se de uma prática consolidada na Idade Média, e que é cheia de distorções e imprecisões, uma vez que os dados fônicos da oralidade nem sempre correspondem às unidades lexicais da língua. Em geral, quanto mais antiga a tradição escrita de uma língua, mais teremos segmentações arbitrárias e incoerentes. Este fato traz questões relevantes para a fase de tokenização, e voltamos às nossas frases, repetidas abaixo:

- (1) Ou seja, é palavra tudo aquilo que em um texto escrito aparece entre espaços em branco ou sinais de pontuação.
- (2) Pelo critério gráfico, a frase anterior tem 20 palavras, e esta frase tem 14.

Do ponto de vista gráfico, *ou seja* (frase 1) conta como 2 palavras, mas essa identificação interessa? Se tivermos 2 palavras em “*ou seja*”, qual a classe de cada uma? Sem titubear,

podemos dizer que “ou” é uma conjunção, mas o que dizer de “seja”? Sabemos que verbos flexionam, e sabemos igualmente que “ou sejam”; “ou foi” ou “ou será” não existem no contexto da locução (e quem está familiarizado com a gramaticalização já sabe que “palavras” do tipo *ou seja* estão longe de ser raras). A frase (2) contém a forma “pelo”, que tanto pode ser desmembrada em “por” e “o”, como pode ser trada como uma unidade. Diferentemente do “ou seja”, o caso das contrações (presentes em *do; da; no; na; neste; naquilo* etc) não envolve discussão, apenas a tomada de uma decisão (descontrair ou não), e há sistemas que trabalham com as contrações desfeitas e sistemas que separam as contrações em duas palavras.

Voltando ao exemplo (1), questões de delimitação associadas ao que chamamos genericamente de *locução* têm se tornado cada vez mais frequentes, e podem ser agrupadas sob o rótulo guarda-chuva das expressões multipalavra (ou mwe, do inglês *multi-word expression*). O que fazer com “levar em conta” ou “abrir mão”? Levando em conta o critério gráfico, temos 3 palavras no primeiro caso, e 2 palavras no segundo. Isso quer dizer que “em conta” ou “mão” são objetos/argumentos dos verbos? Apesar de tantas dúvidas, aqui temos uma certeza: não queremos uma análise que indique que “mão” ou “conta” são argumentos de “levar” e “abrir”, respectivamente. Além de não fazer sentido, estamos criando um problema para a sintaxe, pois se “mão” for argumento de “abrir”, como analisar “prêmio” em “abrir mão do prêmio”? A linguista Maria Teresa Biderman afirma que só a dimensão semântica oferece a chave decisiva para identificar a unidade do léxico expressa no discurso (Biderman, 2001). Porém, nossas intuições semânticas são muito menos precisas que as intuições formais. Formas são visíveis, têm materialidade, e o sentido não tem. Por isso, a delimitação de unidades lexicais complexas continua sendo uma questão nos estudos linguísticos, e uma questão que tem consequências para o PLN.

Especificidades de gêneros textuais trazem desafios adicionais: textos técnicos podem conter “palavras” como *poli(bisfenol A-co-epicloridrina)*, que contém hífen, espaço, parêntese e travessão, mas que é considerado um único item. A comunicação digital traz *hashtags*, que com frequência constituem um sintagma ou frase completa (*#soquenao*), emojis e elementos como *rsrsrs; ;-) , :-}}, :-P*.

Na criação dos modelos de língua (2.1.1), o primeiro passo consiste em traduzir as palavras de um texto em representações numéricas, já que lidar com números facilita o trabalho das máquinas. Mas quais unidades virarão números? Qual o conceito de palavra subjacente? São três as estratégias de tokenização: baseada em palavra, baseada em caractere e baseada em subpalavra.

A tokenização baseada em palavra usa o critério gráfico na delimitação de tokens. O ponto positivo desta abordagem é que as representações criadas a partir das “palavras” serão bastante informativas do ponto de vista do sentido e do contexto, já que cada representação corresponde a uma palavra. Porém, usando apenas o critério gráfico, a abordagem lida com unidades que apesar de separadas por espaços em branco correspondem a uma mesma palavra, como “camisa de força”, “de repente” e “engolir sapo”. O critério gráfico também não captura relações importantes entre as palavras, como a relação entre as formas “palavra” e “palavras”. Isto é, cada uma das formas “palavra” e “palavras” é tratada como uma unidade distinta, e receberá um identificador diferente. Outro ponto negativo diz respeito a uma característica da língua que vimos no

capítulo 3: a imensa quantidade de palavras diferentes. Se a intenção é criar um modelo que consiga lidar da forma correta com todas frases do corpus (todas as frases contidas no material de treino), precisaríamos de um identificador para cada palavra, e logo a quantidade de palavras do modelo ficaria gigantesca, tornando o processamento mais pesado. Uma solução utilizada para este problema é ignorar as palavras menos frequentes. Neste caso, cada palavra com frequência abaixo de um determinado limite recebe uma etiqueta [desconhecida], o que significa que palavras pouco usadas como *cotidianidade* e *dançomania* serão representadas exatamente da mesma maneira. E como também já vimos no capítulo 3, essa pode não ser uma boa ideia, pois muito da língua ficaria de fora.

A tokenização baseada em caracteres (letras e símbolos), apesar de parecer estranha, tem a vantagem de ser mais econômica: ao invés de criar uma representação numérica para cada palavra da língua, é criada uma representação para cada letra. Com isso, mesmo palavras com erros de ortografia ou com ortografia pouco comum (*ameeeeei*) são tokenizadas. No entanto, diferentemente das palavras, os caracteres praticamente não carregam informação. Ou seja, será necessário um esforço adicional para atribuir sentido a um conjunto de tokens que corresponda a uma palavra. Outra característica dessa estratégia é imensa quantidade de tokens que precisará ser processada. Se na abordagem anterior contamos 2 tokens em “palavra” e “palavras”, agora passamos a contar 7 e 8 tokens, respectivamente, o que também tem impacto no processamento: para conseguir dar conta de tantos tokens, será necessário limitar a quantidade de texto a ser processada.

A tokenização baseada em subpalavra é uma abordagem intermediária, e permite lidar melhor com as palavras não vistas (diferentemente da abordagem de palavras) sem aumentar a quantidade de tokens (diferentemente da abordagem baseada em caracteres). Na tokenização baseada em subpalavra os algoritmos aprendem, por meio de estatística, aquilo que chamamos de morfema – os algoritmos aprendem a separar as palavras em pedaços que têm algum significado a partir da recorrência desses pedaços. Palavras frequentes são consideradas na íntegra, isto é, são indecomponíveis, e palavras pouco frequentes são quebradas em partes significativas, as tais “subpalavras” que dão nome à estratégia. Assim, palavras raras como *cotidianidade* e *dançomania*, que na abordagem baseada em palavras são igualmente tratadas como [desconhecida], ganham alguma representação devido à recorrência de *_idade* e *_mania*, e a quantidade de palavras desconhecidas que consegue ser tratada aumenta. Pelo mesmo princípio, “palavras” será quebrada em “palavra” e “s”, evitando com isso duas representações dissociadas para o caso de singular e de plural. Os tokenizadores baseados em subpalavras conseguem identificar prefixos e sufixos, dando conta, por exemplo, de “desmerecimento”, que será transformado em três tokens. O fato de levar em conta a frequência das palavras é uma maneira interessante de evitar segmentações equivocadas, como em “desenvolvimento”, em que não queremos *des_* e *_envolvimento* ou *des_ confiar*, mas queremos *des_ proteger*. Atualmente, os modelos de língua com os melhores resultados utilizam essa estratégia de tokenização e, linguisticamente, entendemos o motivo de ela levar aos melhores resultados.

Abordagens baseadas em regras linguísticas também utilizam pistas morfológicas no reconhecimento de palavras nunca vistas. A diferença é que prefixos e sufixos serão manualmente inseridos pelo programador.

Sentencição: A sentencição diz respeito à identificação de frases, e seu primeiro desafio está em reconhecer quando um ponto final não está sendo usado para delimitar frases. É frequente o uso de ponto final para abreviar nomes próprios (Philip B. Morris), ou comuns (“em conformidade com o art. 4º), e não queremos, em nenhum desses casos, que o ponto seja considerado um separador de frases. Também há a situação oposta, quando o fim da frase não é indicado por qualquer sinal de pontuação, mas por uma mudança de linha, caso de manchetes de jornais, títulos ou nomes de seção em textos técnicos ou acadêmicos. Uma consequência da ausência do sinal de pontuação é que diferentes frases ficam concatenadas como se fossem uma única frase, dificultando as etapas posteriores de análise linguística. A figura a seguir ilustra o ponto, com um trecho do Estatuto da Criança e do Adolescente.

<p>Parágrafo único. São também princípios que regem a aplicação das medidas: (Incluído pela Lei nº 12.010, de 2009)</p> <p>I - condição da criança e do adolescente como sujeitos de direitos: crianças e adolescentes são os titulares dos direitos previstos nesta e em outras Leis, bem como na Constituição Federal; (Incluído pela Lei nº 12.010, de 2009)</p> <p>II - proteção integral e prioritária: a interpretação e aplicação de toda e qualquer norma contida nesta Lei deve ser voltada à proteção integral e prioritária dos direitos de que crianças e adolescentes são titulares; (Incluído pela Lei nº 12.010, de 2009)</p> <p>III - responsabilidade primária e solidária do poder público: a plena efetivação dos direitos assegurados a crianças e a adolescentes por esta Lei e pela Constituição Federal, salvo nos casos por esta expressamente</p>
<p>Parágrafo único. São também princípios que regem a aplicação das medidas: (Incluído pela Lei nº 12.010, de 2009) I - condição da criança e do adolescente como sujeitos de direitos: crianças e adolescentes são os titulares dos direitos previstos nesta e em outras Leis, bem como na Constituição Federal; (Incluído pela Lei nº 12.010, de 2009) II - proteção integral e prioritária: a interpretação e aplicação de toda e qualquer norma contida nesta Lei deve ser voltada à proteção integral e prioritária dos direitos de que crianças e adolescentes são titulares; (Incluído pela Lei nº 12.010, de 2009) III - responsabilidade primária e solidária do poder público: a plena efetivação dos direitos assegurados a crianças e a adolescentes por esta Lei e pela Constituição Federal, salvo nos casos por esta expressamente</p>
<pre># text = Parágrafo único. # text = São também princípios que regem a aplicação das medidas: (Incluído pela Lei nº 12.010, de 2009) I - condição da criança e do adolescente como sujeitos de direitos: crianças e adolescentes são os titulares dos direitos previstos nesta e em outras Leis, bem como na Constituição Federal; (Incluído pela Lei nº 12.010, de 2009) II - proteção integral e prioritária: a interpretação e aplicação de toda e qualquer norma contida nesta # text = Lei deve ser voltada à proteção integral e prioritária dos direitos de que crianças e adolescentes são titulares; (Incluído pela Lei nº 12.010, de 2009) III - responsabilidade primária e solidária do poder público: a plena efetivação dos direitos assegurados a crianças e a adolescentes por esta Lei e pela Constituição Federal, salvo nos casos por esta expressamente</pre>

Exemplo de um pré-processamento mal feito no que se refere à sentencição, em 3 momentos.

A figura traz o mesmo trecho em 3 momentos: o documento original (em pdf) com as marcas de parágrafo; a versão convertida em texto simples (*texto simples* é um arquivo de texto sem informações de estilo ou de formatação. Arquivos criados no “Bloco de

Notas”, o editor de texto encontrado no sistema operacional Windows, são exemplos de arquivos de texto simples, que têm a extensão “.txt”); e a saída de um programa de análise linguística. Exceto pela primeira frase, “Parágrafo único.”, todo o trecho que vem a seguir não tem ponto final e é a formatação que nos informa que estamos diante de unidades de informação distintas. Porém, a consequência disso é que tudo o que vem após “Parágrafo único.” pode ser considerado uma única frase ou, no exemplo do terceiro momento da figura, o trecho foi considerado duas frases, com uma segmentação completamente aleatória.

Onde gostaríamos de identificar frases no trecho selecionado? Qual a segmentação certa? Cada parágrafo do documento original deve ser considerado uma frase, mesmo que esteja separado do parágrafo seguinte por ponto e vírgula? Já em textos literários, os dois pontos podem sinalizar o início da fala de personagens no discurso direto, frequentemente introduzida com um novo parágrafo. Do mesmo modo que na tokenização, diferentes gêneros textuais trarão diferentes desafios e precisarão de soluções diferenciadas.

5.1 StopWords: por que retirá-las?

É comum, em trabalhos com processamento automático de textos, a menção à retirada de *stopwords* do texto como uma das etapas de pré-processamento. Independentemente da decisão de retirá-las ou não, o importante é que esta seja uma decisão informada.

Stopwords são palavras consideradas irrelevantes (... e o que é irrelevante para uma pessoa/tarefa pode não ser para outra...) e que, por isso, podem ser descartadas. As palavras mais frequentes de um corpus – artigos, preposições, conjunções, verbos auxiliares, por exemplo – costumam ser consideradas irrelevantes, porque pouco significativas com relação ao conteúdo de um documento.

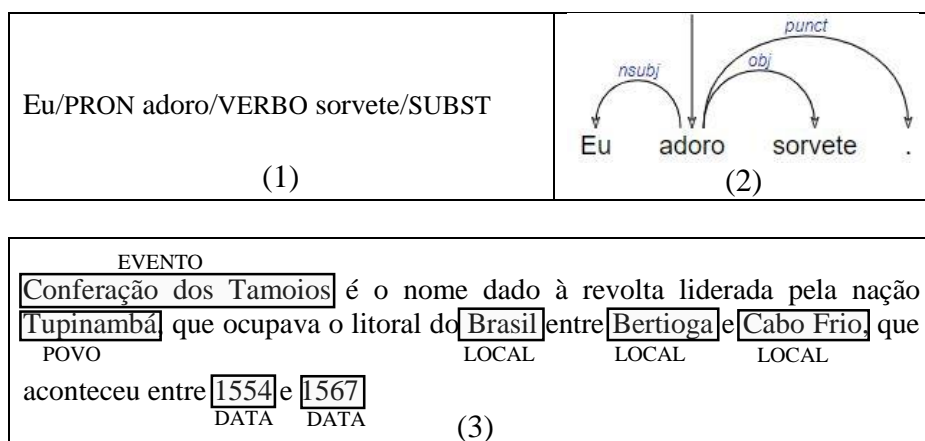
A ideia de trabalhar com uma lista de palavras que podem ser eliminadas do texto surge no contexto da indexação automática de documentos: utilizar estas palavras frequentes como elementos de um índice, além de não ser informativo, exige um espaço maior de armazenamento. A ideia subjacente às stopwords é que ao se deparar com as palavras de uma lista, deve-se interromper (*stop*) a indexação e o processamento dos documentos.

Porque as palavras mais frequentes irão corresponder ao que chamamos de *palavras gramaticais* (em oposição a *palavras lexicais*), é comum também uma confusão entre esses termos. Qualquer palavra que se deseje eliminar de um documento por ser pouco informativa pode ser considerada uma *stopword*, independentemente de sua classe gramatical, e poderá integrar uma lista de palavras que serão canceladas (uma *stoplist*).

Uma crítica que se pode fazer à eliminação irrefletida de palavras gramaticais tem a ver com a noção de palavra. Quando se justifica a remoção de palavras gramaticais com o argumento de que palavras lexicais e palavras gramaticais têm valores semânticos diferentes – comparemos o sentido de *lagoa* e o sentido de *de* – fica explícito que o critério semântico é um critério relevante. Entretanto, se eliminamos a palavra “de”, *camisa de força* vira duas palavras, *camisa* e *força*, e isso poderá ter impacto em tarefas que envolvem sentido ou conteúdo dos textos. Se levamos em conta que terminologias, ou palavras específicas de um domínio, são com frequência unidades de mais de uma palavra é importante ter ciência do que se ganha e do que se perde com a decisão de eliminar ou manter as palavras gramaticais.

6. Anotação: Principais tipos

Anotação linguística é a adição de informação (linguística) a um corpus. Tecnicamente, anotar consiste em delimitar um segmento de texto (palavra, sintagma, frase, parágrafo...) e atribuir-lhe uma etiqueta, definida conforme o objetivo da anotação. No exemplo (1) abaixo, temos uma anotação atribuída a cada palavra, no exemplo (2) temos uma anotação que relaciona palavras e no exemplo (3) uma anotação atribuída a palavras e segmentos de texto.



A lista dos tipos de anotação é potencialmente infinita, e praticamente não há limites quanto à informação linguística que se pode adicionar a um corpus. O conjunto das etiquetas utilizadas na anotação de um corpus é chamado de *tagset* (porque uma etiqueta, em inglês, é uma *tag*). De um ponto de vista formal, existem algumas maneiras de se representar um texto anotado. No capítulo 2 usei uma representação bastante comum, reproduzida na tabela a seguir:

- cada palavra fica em uma linha
- cada palavra recebe um identificador (id) (coluna 1)
- cada tipo de informação linguística fica em uma coluna distinta (colunas 3, 4 e 5, que representam etiquetas relativas a classe de palavras, tipo de relação sintática, e dependência sintática, respectivamente, e as colunas 6 e 7 codificam informações relacionadas ao sentido)

Id	palavra	POS	Tipo de relação sintática	Dependência entre as relações sintáticas	Classe semântica	Polaridade
1	Eu	PRON	Sujeito	2	Pessoa	--
2	adoro	VERBO	Núcleo da Oração	0	Verbo de opinião	Positiva
3	sorvete	SUBST	Objeto direto	2	Alimento	--
4	de	PREP	?	5		--
5	laranja	SUBST	Adj. adnominal	3	Alimento	--
6	com	PREP	?	7	--	--
7	gengibre	SUBST	Adj. adnominal	5	alimento	--
8	da	PREP	?	9	--	--
9	sorveteria	SUBST	?	3 ou 10 (?)	Organização	--
10	Sabores	PROP ou SUBST	Adj. adnominal	9 ou 10 (?)	Organização	--

11	do	PREP (ou PROP?)	?	12	Organização	--
12	Campo	PROP ou SUBST	?	10	Organização	--

Ainda que codifiquem informações que podem se relacionar, as colunas são independentes, e um corpus anotado pode ter apenas uma (qualquer uma) ou várias colunas (camadas) de anotação. A coluna 5 deve ser lida da seguinte maneira:

- a 1ª linha indica que a palavra (ou token) 1, “eu” é dependente do token 2 (adoro), e essa relação é de “sujeito” (informação que está na coluna 3).

- a 2ª linha indica que a palavra (ou token) 2, “adoro”, não depende de ninguém, é o núcleo da oração principal (ou raiz), e por isso ela *depende* do token 0. Por convenção, 0 sempre é usado para indicar o núcleo da frase. Assim, em *Cuidado! Cão bravo*, a palavra cuidado será 0, porque é o único elemento dessa frase, e a frase *cão bravo* terá o 0 atribuído a *cão*.

- a 3ª linha indica que a palavra (ou token) 3, “sorvete” é dependente do token 2 (adoro), e essa relação é de “objeto direto” (informação que está na coluna 3).

- a 4ª linha indica que o token 4 (*de*) é *dependente* do token 5 (*laranja*), e essa relação é de....? Não estamos habituados, na chamada Gramática Tradicional (GT), a atribuir uma função sintática específica às preposições (ou conjunções). Justamente para mostrar diferentes maneiras de lidar com a gramática, quis trazer um exemplo de uma outra abordagem – e formalismo –, chamada *gramática de dependências*, comum no PLN. Nesta abordagem, as relações não se estabelecem entre sintagmas ou grupos de constituintes, mas entre cada palavra individualmente (retomaremos esse ponto mais à frente). Existem algumas formas de fazer análise sintática, e no exemplo do quadro o “de” é dependente de “laranja”, e não de “sorvete”. Essa é uma escolha decorrente de uma abordagem gramatical, e há abordagens em que o “de” é dependente de “sorvete”.

- a 5ª linha indica que o token 5, “laranja” é dependente do token 4 (sorvete), e essa relação é de “adjunto adnominal”.

- a 6ª linha indica que o token 6, “com” é dependente do token 7 (gengibre), e essa relação é de ... Toda a explicação sobre o token 4 (“de”) se aplica aqui.

- a 7ª linha indica que o token 7, “gengibre” é dependente do token 4 (laranja), e essa relação é de “adjunto adnominal”.

É a coluna 5, portanto, das dependências, que indica que “de gengibre” se relaciona a (lê-se como “é dependente de”) *laranja*, e não a *sorvete*. Isto é, o sabor do sorvete é “laranja com gengibre”.

Antes de avançar, é possível que você esteja se perguntando: o que um programa faz com essa tabela?

A tabela ilustra uma porção de um *corpus anotado*. E um corpus anotado por pessoas – um corpus padrão ouro – é um elemento crucial para o desenvolvimento da área (como já vimos na seção 2.1.2). Retomamos aqui dois usos muito comuns: um corpus anotado serve para fornecer exemplos do que um sistema precisa aprender (para treino) e para fornecer exemplos para avaliar a qualidade da análise automática (para avaliação). No

primeiro caso, e considerando apenas a coluna 3, por exemplo, ele “verá” que *adoro* é um verbo e, ao lado de muitos outros exemplos, irá aprender – se tudo correr bem – como classificar algo como um verbo, inclusive palavras que nunca viu. Mas nem sempre o procedimento será bem-sucedido, e já vi a palavra *dióxido* ser classificada como a forma participial do verbo (inexistente) *dioxa*. Na página eletrônica associada ao livro, na seção @Exercícios, você verá como utilizar uma ferramenta de anotação gramatical criada com a abordagem de aprendizado de máquina, ou seja, sem qualquer informação linguística explicitamente codificada relativa à gramática da língua portuguesa.

A figura 13 traz mais algumas provocações gramaticais:

- no trecho “sorveteria Sabores do Campo”, quem é o núcleo e quem é o modificador? Algumas gramáticas trazem a classe do aposto especificativo restritivo, mas este é um ponto controverso. “Sabores do Campo” especifica “sorveteria” (especifica sobre qual sorveteria estamos falando) ou “sorveteria” especifica “Sabores do Campo” (especifica que “Sabores do Campo” é uma sorveteria, e não um restaurante, por exemplo)? Estamos habituados a pensar que a noção de dependência relacionada à importância, à hierarquia: se um elemento A depende de um outro, B, este outro, B, é um elemento mais importante do que A. No nosso exemplo, quem é mais importante, “sorveteria” ou “Sabores do Campo”? Eu, sinceramente, não tenho uma opinião definida, e aceito de bom grado qualquer uma das leituras. Na anotação, o importante é que sejamos *consistentes*, isto é, sempre que passarmos por fenômenos como esse, estejam analisados (anotados) da mesma maneira. E, se decidimos que “Sabores do Campo” depende de “sorveteria”, então “sorveteria” será adjunto adnominal de “sorvete”.

- Por fim, o que fazer com “Sabores do Campo”? São uma única “palavra”, “Sabores_do_Campo”? Ou 3 palavras (Sabores; do; Campo)? Ou 4? (Sabores; de; o; Campo?) No caso de serem 3 (ou 4), qual é a classe de cada uma delas? *Sabores* será substantivo próprio ou substantivo comum, cujo lema é “sabor”? E “do”, será nome próprio? E “Campo”?

A anotação nos obriga a pensar (e a formalizar) questões linguísticas em seus mínimos detalhes, e por isso acaba sendo uma fonte inesgotável de estudo para os linguistas (computacionais ou não). Por isso, também, a anotação é o espaço mais evidente de relação entre a linguística e o PLN.

Mais do que sugerir “vejam como a anotação é cheia de problemas difíceis!”, interessa aqui enfatizar a anotação como uma atividade de *interpretação*: não está dado de antemão, e precisaremos decidir, dentre outras coisas, o que é uma palavra (Sabores do Campo é 1 palavra ou 4?) e como atribuir as classes gramaticais (o “do” de “Sabores do Brasil” é parte de um nome próprio e, portanto, deve ser anotado como tal? E “Sabores”?) e funções sintáticas (aposto ou adjunto adnominal?). São problemas talvez pouco interessantes para teorias linguísticas, mas é esta análise que será reproduzida por sistemas, e que conseqüentemente nos será “devolvida” por estes mesmos sistemas quando quisermos analisar um texto qualquer.

Ainda é pouco comum no PLN a discussão sobre as análises linguísticas subjacentes à anotação. Mas vemos como ela é necessária quando mesmo distinções consideradas óbvias, como entre o numeral “um” e o artigo “um”, podem não ser tão óbvias assim. No trecho abaixo estamos diante de um artigo ou de um numeral?

A Media Capital, que gere a TVI, apresenta assim um argumento para a assembleia de credores que se realiza no próximo dia 14.

E para demonstrar que a diferença entre artigos e numerais não é tão óbvia assim, trago as palavras da *Gramática de Usos do Português*:

Do ponto de vista da quantidade, isso significa que, no caso do artigo indefinido, fala-se de “pelo menos um”, enquanto, no caso do numeral, fala-se de “exatamente um”. (...). Apesar disso, *em muitos enunciados tal diferença é neutralizada, pois fica difícil concluir-se se o que está no primeiro plano é um ou outro valor.* (Neves, 2000:518, grifo meu)

Por isso, não é exatamente correto afirmar que um corpus anotado é uma fonte objetiva de informações linguísticas que poderão ser usadas para o aprendizado ou para a avaliação de sistemas, mas sim que um corpus anotado reflete a *interpretação* dos anotadores, isto é, reflete a interpretação de seres humanos, com relação a um dado fenômeno, sendo fruto de um processo de análise linguística. O que acontece é que alguns fenômenos, como as classes de palavras e as funções sintáticas, já estão tão internalizados devido aos anos de escolarização que ficamos com a falsa impressão de que são classificações mais “objetivas”, enquanto outras são consideradas mais subjetivas.

O fato de não termos hábito de discutir as análises gramaticais – desde a escola são quase sempre ensinadas em termos de *acerto* e *erro*, e nunca em termos de classes que são construtos humanos, situados historicamente e vinculados a uma certa teoria ou visão de mundo ou interesse – contribui para a ideia de que estamos diante de classificações objetivas. Some-se a isso o fato de que, dentre todas as disciplinas científicas, a gramática é aquela que possui o vocabulário teórico próprio mais estável e mais antigo de todas: as classes de palavras.

A seguir estão listados os tipos mais frequentes de anotação. Para cada tipo, é apresentada uma descrição geral da anotação e da tarefa a ela associada, desafios linguísticos e corpora padrão ouro em português. A lista dos corpora não é exaustiva, e pode facilmente ficar desatualizada, mas serve como um incentivo para explorações futuras nos materiais.

6.1 POS – as classes de palavras

A sigla POS vem do inglês *part of speech* e é o equivalente ao nosso “partes do discurso” ou “classes de palavras”. As classes de palavras são a classificação linguística mais antiga, aliás, a classificação *científica* mais antiga de que se tem notícia, e surgem da necessidade de fornecer um vocabulário teórico que pudesse ser usado na análise lógica das proposições (falamos delas e de sua relação estreita com os primeiros estudos da linguagem em 3.2). É esse vocabulário cristalizado pela tradição que comparece até hoje nos estudos linguísticos e na anotação de POS, ainda que com pequenas variações. E assim como não há, nos estudos gramaticais, consenso absoluto sobre quais sejam as classes de palavras, pode haver diferentes conjuntos de POS (chamados *tagsets* de POS) disponíveis para uma mesma língua. Para a língua portuguesa, temos atualmente os conjuntos abaixo, e a página eletrônica deste livro (@Ponteiros) remete a cada um dos projetos, que por sua vez contém a explicação das etiquetas:

- Tagset do projeto Lácio-Web/Mac-Morpho (Aluisio et al., 2003) – 23 etiquetas na primeira versão
- Tagset do sistema PALAVRAS (Bick, 2000) – 14 etiquetas (são as mesmas utilizadas pelo projeto AC/DC)
- Tagset do projeto/ corpus Tycho Brahe (Galves et al., 2017) – cerca de 28 etiquetas principais

Além disso, existem os projetos multilíngues, que propõem tagsets independentes de língua, e, portanto, aplicáveis também ao português. O projeto Eagles, de 1996, e a proposta mais recente do projeto *Universal Dependencies* são exemplos de projetos multilíngues.

Projetos de tagsets compartilháveis e independentes de língua pretendem oferecer uma metalinguagem e uma análise linguística comuns para todas as línguas. Se tais projetos são/serão bem-sucedidos, no sentido de permitirem uma análise igualmente satisfatória das várias línguas, é uma questão em aberto.

Desafios linguísticos

Em uma tarefa de anotação de POS é preciso atribuir, para cada palavra, a classe gramatical adequada – e o que deve contar como palavra, como já vimos, está longe de ser óbvio. Do mesmo modo, o que conta como “classe adequada” vai depender da teoria subjacente, ou das decisões linguísticas subjacentes à anotação. Por exemplo, como deve ser classificada a palavra *proibidos* na frase abaixo: como verbo ou como adjetivo?

1. Empresas usam liminares para manter painéis proibidos
 - 1a. Empresas usam liminares para manter painéis (*que foram*) **proibidos** (verbo)
 - 1b. Empresas usam liminares para manter painéis **proibidos** (=ilegais) (adjetivo)

E como lidar com o “o” na frase abaixo?

2. Isto poderá significar, simplesmente, que o controle voluntário da mão direita é mais fácil do que o da mão esquerda.

Em (2), o “o” pode ser analisado como um pronome substantivo demonstrativo (2a) ou como um artigo (2b), e neste caso se considera uma elipse do substantivo “controle”. As leituras têm como resultado final o mesmo sentido, estão igualmente corretas, e não adianta ampliar o contexto para decidir por uma outra análise:

- 2a. Isto poderá significar, simplesmente, que o controle voluntário da mão direita é mais fácil do que **o** (= aquele) da mão esquerda.
- 2b. Isto poderá significar, simplesmente, que o controle voluntário da mão direita é mais fácil do que **o** (= controle) da mão esquerda.

A classificação automática de palavras em classes morfossintáticas é uma tarefa antiga da Linguística Computacional /PLN (existe desde 1958), e inicialmente era encarada como uma tarefa de desambiguação, cujo objetivo era eliminar as muitas possibilidades de classificação de uma mesma palavra (pensemos nas várias classes que uma palavra como “que” pode ter. Segundo o dicionário Houaiss, são 17 classes possíveis e há um trabalho (Martins, 1985) que menciona 27 classes para o “que”!)

Atualmente, o desempenho de ferramentas de anotação de POS para o português está em torno de 97%, indicando que a análise automática é bastante boa. No entanto, isto não significa que ao anotar um texto qualquer iremos obter uma qualidade de 98%. Sabemos que textos podem variar de muitas maneiras: cada gênero textual e cada domínio têm estruturas sintáticas e vocabulário característicos, e esta variação influencia a qualidade análise linguística. Como boa parte das ferramentas – baseadas em AM ou em regras – é criada a partir de corpora de textos jornalísticos, podemos esperar um bom desempenho quando estivermos diante deste tipo de material. Para textos técnicos ou literários, veremos alguma queda na qualidade da anotação. Voltamos aqui à necessidade de corpora padrão ouro, neste caso, relativos a textos não jornalísticos.

Corpora em português

Para cada um dos tagsets de POS mencionados, temos pelo menos um corpus padrão ouro associado. Para o português dispomos dos seguintes corpora anotados com POS: corpus MacMorpho, corpus Bosque (ambos disponíveis em várias versões e formatos) e o corpus Tycho Brahe.

6.2 Sintaxe

Corpora anotados sintaticamente são chamados de *florestas sintáticas* (em inglês, *treebanks*, literalmente *bancos de árvores*), pois cada frase analisada sintaticamente produz uma *árvore sintática*. A motivação para a criação de florestas sintáticas vem tanto do lado teórico quanto do lado aplicado da Linguística Computacional/PLN: do lado teórico, trata-se de um recurso para a investigação de teorias linguísticas, que podem ser postas à prova em um ambiente real, isto é, com frases naturalmente construídas pelos falantes. E, de maneira complementar, o resultado da anotação – o *treebank* – associado a uma ferramenta de busca, pode ser um auxiliar valioso para pesquisas linguísticas. Do lado aplicado, a anotação sintática fornece pistas sobre a frequência de certas estruturas linguísticas, facilitando a desambiguação sintática: a seleção da análise adequada pode ser feita levando em conta as chances de a estrutura sintática em questão ser possível na língua. A frase abaixo, embora não seja difícil para nós (sabemos que a palavra *Sérvia* está relacionada a *sanções*, e não a ao verbo *suspendeu*) tem uma estrutura ambígua, como vemos no contraste com a frase (2), na qual a palavra em negrito se refere ao verbo).

- (1) O país suspendeu as sanções contra a **Sérvia**
- (2) O inimigo jogou irmãos contra **irmãos**

Além disso, e como qualquer outro tipo de anotação em corpus, florestas sintáticas são de grande utilidade para a avaliação e treino de sistemas.

A maioria dos *treebanks* é pouco especificada em termos teóricos, e esta decisão tem a vantagem de permitir sua utilização por pessoas de diferentes filiações teóricas. Mas, assim como no caso das classes de palavras, “leveza” teórica não é sinal de neutralidade: o que se tem é uma forte inspiração (ou influência) de análises derivadas de gramáticas tradicionais, que acabam funcionando como metalinguagem supostamente neutra.

Existem dois grandes tipos de formalismos sintáticos, que levam a dois grupos de *treebanks*: aqueles que levam em conta uma estrutura sintagmática (isto é, que leva em conta, na anotação, o conceito de sintagma), e aqueles que levam em conta uma estrutura de dependências. Cada formalismo pode ser visualizado de diferentes maneiras, e pode

seguir diferentes abordagens gramaticais. Ou seja, uma coisa é um formalismo gramatical, isto é, uma maneira formal de representar algum aspecto da gramática (no nosso caso, a sintaxe), e outra coisa é uma abordagem de gramática, uma maneira de entender como se estrutura uma língua (em nosso caso, como se estrutura sintaticamente uma língua). Uma abordagem de gramática, portanto, pode ser representada de diferentes maneiras.

Uma abordagem (ou gramática) de dependências comporta teorias gramaticais baseadas na relação de *dependência*, e uma abordagem (ou gramática) sintagmática comporta teorias baseadas na relação entre *constituíntes*.

Nas gramáticas de dependência, os elementos que se conectam para compor a estrutura da frase são as *palavras*, e o verbo é considerado o núcleo (ou raiz) da oração. Todas as outras unidades são direta ou indiretamente conectadas a ele, em relações que são chamadas de relações de *dependência*. Já nas gramáticas de constituíntes (ou de estrutura sintagmática), os elementos que se conectam para compor a estrutura da frase são *constituíntes*. Constituíntes são grupos de palavras que possuem alguma hierarquia (por exemplo, “vaca amarela” forma o sintagma nominal – e um sintagma nominal é um constituinte – “vaca amarela”, que por sua vez pode ser decomposto nos constituíntes menores “vaca” e “amarela”).

Considerando a frase “A vaca amarela fugiu”, na abordagem de dependências é apenas a palavra *vaca* que se relaciona – iato é, que “depende de” – *fugiu*. As palavras *a* e *amarela*, por sua vez, dependem de *vaca*. Na abordagem sintagmática, o que se relaciona com *fugiu* é todo o sintagma *a vaca amarela*, que por sua vez tem como núcleo a palavra *vaca*.

Outra diferença importante diz respeito à estrutura básica da oração: na gramática de constituíntes (de inspiração gerativa, sobretudo), a estrutura é uma divisão binária entre sujeito (sintagma nominal - SN) e predicado (sintagma verbal - SV), o que não acontece na gramática de dependências. Podemos usar uma representação de colchetes para mostrar essa diferença, e a divisão binária está à esquerda:

[Eu] [vi a vaca amarela]

[Eu] [vi] [a vaca amarela]

Nos últimos anos, a Linguística Computacional impulsionou o desenvolvimento de teorias baseadas na dependência. No entanto, a noção de dependências entre unidades gramaticais existe desde as primeiras gramáticas registradas, e se encontra, por exemplo, na obra de Paṇini, gramático indiano nascido em 520 a.C.. As gramáticas de dependência modernas, por sua vez, começam principalmente com o trabalho do linguista francês Lucien Tesnière (Percival, 1990).

A distinção entre gramáticas de dependência e de estrutura sintagmática deriva em grande parte da divisão inicial da oração. Tesnière, no entanto, argumentou contra a divisão binária, e posicionou o verbo como a raiz de todas as estruturas oracionais: seu entendimento era de que a divisão sujeito-predicado deriva da lógica aristotélica e não faz sentido na Linguística. Seu principal trabalho, *Éléments de syntaxe structurale*, publicado postumamente em 1959, nunca foi traduzido para o português, e só recentemente foi traduzido para o inglês.

De volta à diferença entre formalismos e abordagens, podemos ter uma árvore (um formalismo) de constituintes, assim como uma árvore de dependências. E podemos ainda ter uma árvore de constituintes que siga ou não a exigência de divisões binárias (exigência de modelos associados à abordagem gerativa).

Como exemplo de treebanks sintagmáticos, temos o Penn Trebank (de língua inglesa) e o corpus TychoBrahe (corpus diacrônico de língua portuguesa); como exemplo de treebanks de dependência, temos todos os corpora do já mencionado projeto *Universal Dependencies*, que atualmente conta com mais de 150 *treebanks* para as mais variadas línguas. O projeto Floresta Sintá(c)tica, pioneiro na construção de *treebanks* para a língua portuguesa, disponibiliza seus *treebanks* em ambos os formatos: árvores sintagmáticas e de dependência.

O quadro abaixo traz exemplos de análises (simplificadas) sintagmáticas (esquerda) e dependenciais (direita).

<pre>=SUBJ:SN ==>N:art('o' <artd> M S) O ==H:n('país' M S) país =P:SV ==MV:v-fin('suspendeu' PS 3S IND) suspe =ACC:SV ==>N:art('o' <artd> F P) as ==H:n('sanção' <np-idf> F P) sanções ==N<:SPrep ===H:prp('contra') contra ===P<:SN ====>N:art('o' <artd> F S) a ====H:prop('Sérvia' F S) Sérvia =.</pre>	<table><tr><td>1</td><td>O</td><td>o</td><td>DET</td><td>2</td><td>det</td></tr><tr><td>2</td><td>país</td><td>país</td><td>NOUN</td><td>3</td><td>nsubj</td></tr><tr><td>3</td><td>suspendeu</td><td>suspendeu</td><td>VERB</td><td>0</td><td>root</td></tr><tr><td>4</td><td>as</td><td>o</td><td>DET</td><td>5</td><td>det</td></tr><tr><td>5</td><td>sanções</td><td>sanção</td><td>NOUN</td><td>3</td><td>obj</td></tr><tr><td>6</td><td>contra</td><td>contra</td><td>ADP</td><td>8</td><td>case</td></tr><tr><td>7</td><td>a</td><td>o</td><td>DET</td><td>8</td><td>det</td></tr><tr><td>8</td><td>Sérvia</td><td>Sérvia</td><td>PROPN</td><td>5</td><td>nmod</td></tr><tr><td>9</td><td>.</td><td>.</td><td>PUNCT</td><td>3</td><td>punct</td></tr></table>	1	O	o	DET	2	det	2	país	país	NOUN	3	nsubj	3	suspendeu	suspendeu	VERB	0	root	4	as	o	DET	5	det	5	sanções	sanção	NOUN	3	obj	6	contra	contra	ADP	8	case	7	a	o	DET	8	det	8	Sérvia	Sérvia	PROPN	5	nmod	9	.	.	PUNCT	3	punct
1	O	o	DET	2	det																																																		
2	país	país	NOUN	3	nsubj																																																		
3	suspendeu	suspendeu	VERB	0	root																																																		
4	as	o	DET	5	det																																																		
5	sanções	sanção	NOUN	3	obj																																																		
6	contra	contra	ADP	8	case																																																		
7	a	o	DET	8	det																																																		
8	Sérvia	Sérvia	PROPN	5	nmod																																																		
9	.	.	PUNCT	3	punct																																																		

Desafios linguísticos

Do mesmo modo que a anotação de POS, a anotação sintática envolve a tomada de uma série de decisões linguísticas nem sempre abordadas em gramáticas ou teorias linguísticas com o detalhe necessário para a anotação. A frase abaixo ilustra desafios simples, e bastante frequente em textos de jornal:

No realismo moderno de John Cassavetes (1929-89), o ator construía o personagem em tempo real diante da câmera, baseado no improviso.

O trecho sublinhado traz desafios de várias naturezas: quantas palavras temos em 1929-89? Se a escolha for por duas palavras (dois numerais, o que faz sentido, por um lado, mas contraria o conceito de palavra gráfica), como se relacionam sintaticamente? Estamos diante de uma coordenação? E qual a relação entre a data e *John Cassavetes*: estamos diante de um aposto, de um modificador de um nome? Assumir a elipse de uma oração (1929-89 = *cujo período de vida foi de 1929 a 1989* ou *que nasceu em 1929 e morreu em 1989*) é uma opção delicada para o processamento automático, porque envolve analisar elementos que não estão na frase.

Em frases com locuções conjuntivas do tipo *quer dizer, por exemplo, isto é, ou seja...* estamos diante de coordenação ou subordinação? A gramática de Mario Vilela e Ingedore Koch (2001) os considera conectores discursivos, o que faz muito sentido, mas precisamos decidir como exatamente formalizar isto na análise sintática – ou já estamos na análise discursiva?

Durante muito tempo, linguistas computacionais acharam que era desnecessário ser muito explícito sobre os alvos dos sistemas de análise automática, uma vez que nossa herança cultural (linguística) compartilhada já está estabelecida há muito tempo. Mas o equívoco desta ideia, nos conta Sampson (2000), foi verificado experimentalmente. Em 1991, em um workshop da já referida ACL, pesquisadores de nove grupos de pesquisa diferentes receberam um mesmo conjunto de frases (em inglês) para analisar (conforme as análises que seus grupos de pesquisa considerariam corretas). Não eram frases especialmente complicadas ou confusas, mas eram frases extraídas de corpus. Para simplificar, a comparação entre as análises não levaria em conta os rótulos dos constituintes, apenas a estrutura (a forma) das árvores, ou seja, a ideia era verificar apenas se os pesquisadores identificariam as mesmas subsequências de palavras como constituintes gramaticais, sem levar em conta a classificação desses constituintes. Para uma grande surpresa na época, o nível de concordância entre as análises foi extremamente baixo. Especificamente, apenas as duas subsequências marcadas por colchetes (abaixo) foram identificadas como constituintes por todos os nove participantes (e os resultados para outros casos foram na mesma direção):

One of those capital-gains ventures, in fact, has saddled him [with [Gore Court]].
Um desses empreendimentos de ganhos de capital, na verdade, o sobrecarregou
[com [Gore Court]].

Se especialistas concordaram tão pouco entre si, e, conseqüentemente, sobre o que *os sistemas deveriam fazer na análise automática*, isto sinaliza a necessidade de alguma discussão linguística – explícita e pública – sobre estas questões, e não apenas discussões

técnicas ou puramente computacionais. Na seção @Sobre avaliação sintática abordo a avaliação automática da anotação sintática.

Corpora em português

Diversos grupos no Brasil estão trabalhando para a criação de *treebanks* de qualidade, e em breve devemos ter um cenário favorável para a anotação sintática. Atualmente, temos o corpus Bosque, que originalmente integra o projeto Floresta Sintá(c)tica, hoje está disponível em diferentes formatos e conforme diferentes modelos gramaticais, e o também já referido corpus do projeto TychoBrahe, de português diacrônico.

6.3 Papéis semânticos

A anotação de papéis semânticos corresponde à formalização de um fenômeno linguístico não tão familiar como classes gramaticais ou funções sintáticas: os papéis semânticos. Papéis semânticos atribuem sentido aos constituintes sintáticos, e podem ser entendidos como a contraparte semântica dos argumentos sintáticos, como “cara & coroa”. É a análise de papéis semânticos que indica que embora

- (1) O lobo mordeu Pedro
- (2) Pedro foi mordido pelo lobo

tenham estruturas sintáticas diferentes, veiculam a mesma ideia. Isto porque em ambas as frases *lobo* recebe o papel semântico de *agente*, apesar de ser sujeito em (1) e agente da passiva em (2). Do mesmo modo, considerando

- (3) Ela abriu a porta com esta chave
- (4) Esta chave abriu a porta

chave é o *instrumento* (papel semântico) com que se abre a porta em ambas as frases, mas em (3) exerce a função sintática de um adjunto adverbial e, em (4), de sujeito.

No mundo PLN, a identificação de papéis semânticos aparece como facilitadora de tarefas como a resposta automática a perguntas e extração de informação. Na Linguística, a primeira proposta de papéis semânticos vem de Charles Fillmore e seu projeto de uma Gramática de Casos. Inicialmente, Fillmore buscou capturar as relações entre *padrões formais* que, apesar de diferentes, indicavam os mesmos significados, como a relação entre orações na voz ativa e na voz passiva. Desde então, surgiram algumas propostas de papéis semânticos – Fillmore (1968), por exemplo, propõe 9 papéis (*agente, experimentador, instrumento, objeto, fonte, objetivo, localização, tempo e caminho*).

A anotação de um corpus com papéis semânticos apresenta dois principais desafios: o primeiro é a alta dependência de uma análise sintática de qualidade, levada a cabo em uma etapa anterior; o segundo desafio é a escolha de quais papéis usar na anotação, isto é, a escolha do tagset. Diferentemente das anotações anteriores, não há uma tradição milenar, ou ensino escolar, subjacente aos papéis semânticos capaz de disfarçá-los de senso comum ou de “levemente teóricos”.

PropBank é o nome de um corpus que contém a anotação de *proposições* (novamente elas) e seus argumentos. No PropBank, o conceito de *proposição* é tomado da semântica de Frames proposta por Fillmore (1968), na qual uma proposição é um conjunto de

relações entre nomes e verbos, sem informação relativa a tempo, modo, aspecto, negação ou modificadores modais.

O primeiro PropBank foi criado para o inglês, em 2005. Ele é um pedaço do corpus *PennTreebank* – ou seja, uma porção de corpus que já foi analisado sintaticamente –, com uma camada adicional de papéis semânticos. Para escolher o tagset usado na anotação do PropBank, o critério de seleção foi o de convergência, isto é, foram usados os papéis semânticos comuns às diferentes propostas, o que levou a 5 etiquetas diferentes para os argumentos (Arg0-Arg5) e 18 etiquetas para modificadores, padronizadas “na medida do possível”, segundo os autores (a página eletrônica deste livro, em @Ponteiros, remete ao projeto de anotação do PropBank, que contém a explicação das etiquetas). São chamados *argumentos de verbos* os complementos considerados “essenciais”, como os objetos, e chamados *modificadores* os complementos considerados “acessórios”, como os adjuntos adverbiais.

No PropBank, os argumentos são

- Arg0 = agente
- Arg1 = paciente/tema/experienciador
- Arg2 = instrumento/beneficiário/atributo
- Arg3 = ponto de partida/ atributo
- Arg4 = ponto de chegada
- ArgM = modificador

Os 18 modificadores (ArgM), por sua vez, indicam desde valores tradicionalmente associados aos advérbios, como *causa, tempo, lugar, modo, negação* a elementos como *discurso direto*. Na frase

- *Mês passado, Ana comprou um apartamento à vista.*

Temos os seguintes papéis:

- Arg0: Ana
- Arg1: um apartamento
- ArgM-MNR (maneira): à vista
- ArgM-TMP(tempo): mês passado.

Associando agora este tipo de informação às anteriores (pos e sintaxe), temos o seguinte:

Id	palavra	lema	pos	Dependência sintática	Tipo de relação sintática	Anotação PropBank
1	Mês	mês	SUBST	5	adj adv	ArgM-TMP
2	passado	passado	ADJ	1	adj adn	
3	,	,	PUNCT	1	punct	
4	Ana	Ama	PROP	5	sujeito	Arg0
5	comprou	comprar	VERBO	0	núcleo da or.	V*
6	um	um	ART	7	adj adn	Arg1
7	apartamento	apartamento	SUBST	5	objeto	
8	a	a	PREP	10	caso	ArgM-MNR
9	a	o	ART	10	adj adn	
10	vista	vista	SUBST	5	adj adv	

A última coluna da figura do quadro acima não contém os papéis semânticos propriamente, apenas informações genéricas como Arg0 ou Arg1. Esta opção pela generalidade é uma especificação do modelo PropBank e não da anotação de papéis semânticos em geral. Ou seja, o que a anotação do PropBank faz é atrelar, à análise sintática – mais especificamente, aos constituintes identificados na análise sintática – categorias genéricas que vão de Arg0 a Arg4, além dos ArgM. Para saber que valor terá o Arg1 (*paciente, experienciador* ou *tema*) é preciso consultar o *frame* (ver seção 4.2) do verbo ao qual os argumentos se relacionam, ou seja, é preciso consultar o frame do verbo *comprar*.

Consultar onde? Em um outro recurso, a VerbNet. Como já mencionado no capítulo 4, a VerbNet contém grupos de verbos associados a suas propriedades sintáticas e semânticas. O verbo *comprar* tem as seguintes informações na VerbNet:

Arg0-Agente: comprador
 Arg1-tema: coisa comprada
 Arg2-fonte: vendedor
 Arg3-valor: preço pago
 Arg4-beneficiário: beneficiário

E um verbo como *abrir* teria

Arg0-Agente: “abridor”
 Arg1-tema ou paciente: coisa aberta
 Arg2-maneira: instrumento
 Arg3- beneficiário: destinatário/alvo

O final do processo de alinhamento entre PropBank e VerbNet produz algo como

(5) Mês passado, ela[agente] comprou um apartamento[tema] à vista.

Mais alguns exemplos. Seguindo a anotação apenas do Propbank, teríamos

- (6) Ela[Arg0] comprou um apartamento[Arg1] para os seus netos[Arg4], por um milhão de reais[Arg3].
- (7) A Costa Rica[Arg0] abriu o mercado de telecomunicações[Arg1] para o capital privado[Arg3].
- (8) Ela[Arg0] abriu a janela[Arg1] com a alavanca[Arg2].
- (9) A alavanca[Arg2] abriu a janela[Arg1].
- (10) A janela[Arg1] abriu.

A atribuição de papéis semânticos sinaliza que as frases (8) (9) e (10) têm sentidos relacionados, mesmo que os argumentos do verbo exerçam funções sintáticas distintas. No entanto, sem o alinhamento com a VerbNet não temos informação sobre a “semântica” dos papéis argumentais. A seguir temos as mesmas frases, com os papéis explicitados.

- (6) Ela[Arg0:agente] comprou um apartamento[Arg1:tema] para os seus netos[Arg4:beneficiário], por um milhão de reais[Arg3:valor].
- (7) A Costa Rica[Arg0:agente] abriu o mercado de telecomunicações[Arg1:tema] para o capital privado[Arg3:beneficiário]
- (8) Ela [Arg0:agente] abriu a janela[Arg1:tema] com a alavanca[Arg2:instrumento]
- (9) A alavanca[Arg2:instrumento] abriu a janela[Arg1:tema]
- (10) A janela[Arg1:tema] abriu

Desafios linguísticos

Assim como nos demais casos de anotação, a escolha por um determinado tipo de papel semântico durante a anotação pode não ser óbvia:

- (11) Moreira Alves também votou a favor de Collor em outros dois mandatos de segurança, mas **foi vencido** nos dois julgamentos.

Qual modificador deve ser usado no trecho sublinhado: *tempo* (foi vencido quando?) ou *local* (foi vencido onde?)? Seria ótimo poder escolher ambas as análises, mas a classificação múltipla não é uma opção de anotação do PropBank (Não ser possível no PropBank não significa, obviamente, que não seja possível na anotação de papéis semânticos.). E, como veremos nas próximas páginas, quanto mais dependente do sentido é a anotação, mais chances de interpretações variadas.

Já deve ter ficado claro por que a anotação de papéis semânticos é dependente de uma análise sintática de qualidade. E uma boa análise sintática, por sua vez, depende da identificação adequada das unidades linguísticas básicas, foco do pré-processamento. Nas frases abaixo, é importante que *por culpa* e *abrir brechas* sejam compostas por duas unidades cada, já que os verbos *por* e *abrir* têm suas propriedades flexionais individuais preservadas (*ele põe a culpa*; *eles põem a culpa*). Por outro lado, para a análise do sentido, à qual se associa também a atribuição de papéis semânticos, é crucial que *por a culpa* e *abrir brechas* sejam considerados uma unidade.

- (12) A medida provisória[Arg0] abre brechas para contestações[Arg1].
(13) Schumacher[Arg0] põe culpa[Arg1] em grau[Arg2]

Vale lembrar que não estamos diante de um fato novo para os estudos linguísticos: é justamente esse limite frequentemente impreciso entre léxico e gramática que motiva a abordagem da lexicogramática. No entanto, a formalização que a tarefa de anotação demanda nos obriga a tomar uma decisão acerca de juntar ou separar as unidades.

O PropBank não é o único recurso a tratar de papéis semânticos para o português, ainda que seja bastante popular. O analisador PALAVRAS, mencionado no capítulo 2, também realiza anotação de papéis semânticos. Para tanto, usa 51 categorias (a informação sobre as etiquetas (ou tagset) está disponível na página eletrônica deste livro, na seção @Ponteiros), e oferece um índice de acerto de 88% mas, por ter o seu próprio sistema de anotação e tagsets, é difícil comparar esse resultado com os demais sistemas de anotação.

Corpora em português

O único corpus padrão outro para o português com anotação de papéis semânticos é o PropBank.Br, cujo esquema de anotação é bastante similar ao do projeto original a fim de facilitar o alinhamento multilíngue.

6.4 Anotações semânticas – e algumas considerações sobre o sentido

Sob o rótulo *anotação semântica* podemos agrupar diversos tipos de anotação, que têm em comum a atribuição de uma informação de natureza semântica a uma palavra ou expressão, em contexto.

Na anotação semântica em sentido estrito, as classes de anotação podem vir de classes genéricas, representativas de um campo semântico (por exemplo, *cor*, ou *emoção*, ou *doenças*), e nesse caso o resultado da anotação é a anotação de um campo semântico; podem vir de um inventário de sentidos como as acepções de um dicionário ou os *synsets* (conjunto de sinônimos) de uma wordnet, vistos na seção 4.1.

A anotação semântica é a que menos precisa de um conhecimento linguístico teórico especializado quando comparada às anotações anteriores, em que é necessário um conhecimento morfosintático profundo (anotação de pos e de sintaxe) ou de uma teoria (anotação de papéis semânticos). Ou seja, identificar, em contexto, se uma dada palavra se refere a uma cor, ou a uma parte do corpo, a relações de parentesco, a uma pessoa ou a uma organização é algo que qualquer falante nativo com boa capacidade de interpretação de texto é capaz de fazer. Mesmo assim, e mais uma vez, a anotação nos confronta com casos simples mas inesperados, que a tornam uma tarefa pouco óbvia.

Os exemplos (1) e (2) ilustram a anotação do campo semântico do corpo humano. Diferentemente de (1), no exemplo (2) a presença da palavra *orelha* não faz referência a uma parte do corpo, mas a um tipo de sorriso. É uma opção do esquema de anotação decidir se a palavra receberá uma etiqueta relativa a partes do corpo e, em caso positivo, se haverá alguma especificação para indicar que não se trata de um uso convencional de *orelha*.

- (1) Lampião era cruel o bastante para, pessoalmente, arrancar olhos ou cortar línguas e orelhas.
- (2) Milena era um sorriso só, de orelha a orelha.

Existem campos semânticos especialmente complicados, como o das emoções e sentimentos. Neste caso, a dificuldade está menos na anotação propriamente, e mais na própria definição do que seja uma emoção. Em um trabalho as pessoas deveriam selecionar trechos de obras literárias que contivessem emoção e anotar as emoções presentes nesses trechos, os autores, que inicialmente previam trabalhar com espectro amplo de emoções, precisaram reduzir as classes de anotação a apenas duas, *medo* e *felicidade*, porque a divergência nas anotações propostas foi tanta que inviabilizou o esquema de anotação original (Heuser et al., 2016).

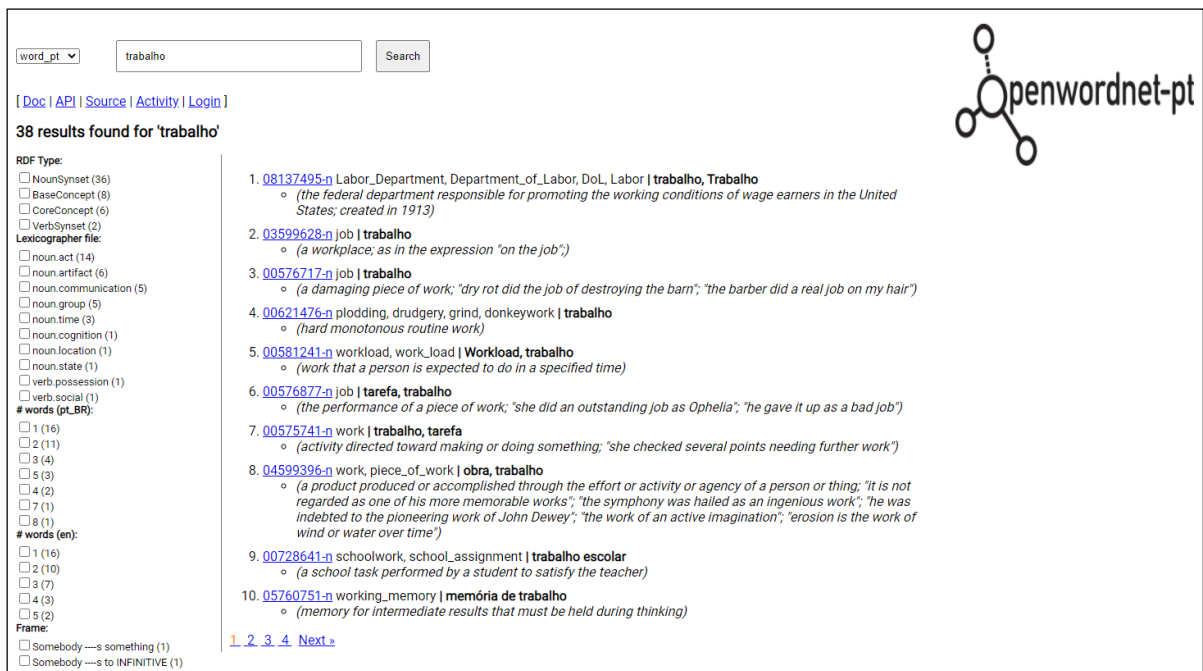
O segundo tipo de anotação semântica – anotação de sentidos (*word senses*) – busca codificar no corpus o sentido de uma palavra, no contexto em que está inserida. Por isso, às vezes essa anotação é descrita como um trabalho de desambiguação, visto que a tarefa propriamente consistiria em selecionar, dentre os vários sentidos possíveis de uma palavra, aquele invocado no contexto da frase. Tomando a palavra *trabalho* sublinhada na frase anterior, a tarefa consiste em escolher um sentido dentre os vários listados em recursos externos, como uma wordnet. Como wordnets do português ainda não são 100% confiáveis, e para que não se pense que a dificuldade está no recurso, e não na tarefa, podemos substituir por enquanto uma wordnet por um dicionário. A tarefa de anotação, portanto, consiste em escolher, dentre opções listadas no quadro abaixo, retiradas do

dicionário Caldas-Aulete online, aquela adequada ao contexto. Me parece que as acepções 1, 2, 3 e 14 são aceitáveis no contexto da frase, o que já é um problema se precisamos escolher apenas um sentido.

1. Emprego da força física ou intelectual para realizar alguma coisa
2. Aplicação dessas forças como ocupação profissional: *Seu trabalho é de gari.*
3. Local onde isso se realiza: *Mora longe do trabalho.*
4. Esmero, cuidado que se emprega na confecção ou elaboração de uma obra
5. A confecção, elaboração ou composição de uma obra
6. Obra realizada: *Essa cômoda é um belo trabalho de marcenaria.*
7. Grande esforço; TRABALHÃO; TRABALHEIRA
8. Exercício para treino: *A professora passou muito trabalho para casa.*
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.*

Figura 1: Acepções da palavra *trabalho* conforme dicionário.

As figuras 2 e 3 a seguir referem-se aos *synsets* (ver 3.1) que contém a palavra *trabalho* conforme a OpenWordNet-PT.



The screenshot shows the OpenWordNet-PT web interface. At the top, there is a search bar with 'trabalho' entered and a 'Search' button. Below the search bar, there are links for 'Doc', 'API', 'Source', 'Activity', and 'Login'. The main content area displays '38 results found for 'trabalho''. On the left side, there are filters for 'RDF Type' (NounSynset, BaseConcept, CoreConcept, VerbSynset, Lexicographer file) and 'Frame' (Somebody —s something, Somebody —s to INFINITIVE). The main list of results shows 10 synsets, each with a unique identifier, a list of related terms, and a brief description. The synsets are numbered 1 through 10, and the list is paginated with links for '1', '2', '3', '4', and 'Next'.

Figura 2: Acepções da palavra *trabalho* conforme a OpenWordNet-PT



Figura 3: Acepções da palavra *trabalho* conforme a OpenWordNet-PT

Desafios linguísticos

Para ilustrar os desafios da anotação semântica, retomo brevemente a experiência de construção do corpus *SemCor*. No fim dos anos 1980 e início dos 1990, quando a WordNet se tornou um recurso popular no PLN, uma equipe de pesquisadores decidiu criar o corpus *SemCor*, um corpus anotado semanticamente segundo os *synsets* da WordNet. Nele, cada palavra do corpus (apenas substantivos, verbos e adjetivos) estaria alinhada ao *synset* adequado. Ao olhar retrospectivamente para o processo de criação do *SemCor*, os autores admitem que o trabalho de anotação foi feito tendo como pano de fundo a hipótese – reconhecida agora como claramente ingênua – de que a anotação semântica seria um trabalho simples de alinhamento entre uma palavra e um *synset* (Baker et al., 2017). No entanto, continuam os autores, a experiência mostrou que o trabalho não tinha o menor traço de simplicidade: em vários casos, nenhum dos sentidos disponíveis para o alinhamento parecia se encaixar, ou o sentido da palavra em questão era capturado por mais de um *synset* da WordNet. A conclusão geral a que os autores chegam é que frequentemente os sentidos das palavras resistem a uma representação discreta e estável como a que os dicionários supõem.

Ainda durante a criação do *SemCor*, foi feito um outro estudo em que foram comparadas as anotações (alinhamentos) feitas por alunos e aquelas feitas por lexicógrafos e linguistas treinados (Fellbaum et al., 1997). Os resultados indicaram

- concordância geral em torno de 70% na atribuição dos sentidos;
- concordância mais alta quando as palavras eram substantivos (em oposição a adjetivos e verbos);
- diminuição da concordância relacionada ao aumento da polissemia;
- preferência por escolher os primeiros *synsets* apresentados, isto é, como se, considerando as figuras 2 e 3 com a palavra *trabalho*, houvesse uma preferência a escolher os *synsets* apresentados logo na primeira tela (fig. 2), embora haja 4 páginas/telas de *synsets*.

Anos depois, a mesma equipe fez um outro estudo. Os autores escolheram uma amostra com dez palavras razoavelmente frequentes, mas moderadamente polissêmicas, para serem anotadas/alinhadas com a WordNet. O quadro 1 apresenta as palavras utilizadas,

bem como o número de sentidos que recebem na WordNet (em inglês) e a frequência no corpus. Como as palavras estão descontextualizadas, a tradução é apenas uma aproximação. O número de sentidos associados a cada uma sinaliza também a dificuldade de tradução sem o contexto.

Palavra	POS	N. sentidos	N. ocorrências
fair (~justo)	Adj	10	463
long (~longo)	Adj	9	2706
quiet (~quieto)	Adj	6	244
land (~terra)	Subst	11	1288
time (~tempo)	Subst	10	21790
work (~trabalho)	Subst	7	5780
know (~saber)	Verbo	11	10334
say (~dizer)	Verbo	11	20372
show (~mostrar)	Verbo	12	11877
tell (~dizer)	Verbo	8	4799

Quadro 1: Lista de palavras usadas no estudo de Passoneu et al., 2009.

Novamente, os resultados evidenciaram a variação nas interpretações sobre o sentido e, novamente, verbos levaram às maiores divergências. Para cada classe, as palavras que mais variaram foram *long*, *work* e *tell*.

Os autores concluíram que os seguintes fatores contribuem para a baixa concordância:

- (i) quanto maior especificidade do contexto em que a palavra está inserida, maior concordância;
- (ii) sentidos mais concretos levam a uma maior concordância que sentidos abstratos;
- (iii) um inventário de sentidos (um conjunto de *synsets*) com elementos parecidos entre si leva a uma menor concordância.

Considerando o ponto (i), sabendo que a palavra *longo* possui os seguintes sentidos (dentre outros):

- 1. *extensão espacial*;
- 2. *extensão temporal*;
- 3. *maior que o normal ou que o necessário*

a frase (a) leva a uma maior concordância que a frase (b)²:

- a) Durante 18 longos meses, Michael não conseguiu encontrar um emprego.
- b) Depois de enviar o manuscrito, meu editor sugeriu uma série de cortes para agilizar o que já era um longo e complicado capítulo sobre as ideias de Brian.

Para ilustrar os itens (ii) e (iii) vamos voltar à palavra *trabalho*, considerando as acepções de dicionário (figura 1). Com relação a (ii), podemos imaginar que é mais difícil uma concordância a respeito dos sentidos de *trabalho* em (c) que em (d)

- c) Nosso *trabalho* foi relevante para a população da cidade
- d) O *trabalho* mais famoso de Van Gogh é O Comedor de Batatas

Com relação ao ponto (iii), em uma frase como “Eu trabalho como professora”, é compreensível uma divergência entre as acepções 2 e 12 .

Ainda tentando entender por que este tipo de análise é tão difícil (difícil quanto à concordância nas interpretações) mais um estudo foi feito. Dessa vez, muitos anotadores

² As frases são traduções das frases usadas no estudo original.

treinados deveriam anotar as *mesmas palavras*, nas *mesmas frases*, para ver se pelo menos nessas condições seria possível obter um consenso em torno dos sentidos escolhidos. Os resultados sinalizaram que ainda assim a variação entre os anotadores foi bem grande, bem maior do que o previsto (Baker et al., 2017).

Para uma boa parte da comunidade de PLN e para o senso comum, 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 (o que se manifesta nas acepções separadas por números) faz parecer que os significados se organizam naturalmente dessa maneira. Nos dicionários, porém, as palavras têm os significados que têm não porque esses sentidos lhes sejam intrínsecos, mas porque alguém (um conjunto de lexicógrafos) assim o decidiu. E assim o decidiu porque *precisavam* delimitar os significados - lexicógrafos são obrigados a descrever palavras como se todas elas tivessem um conjunto de significados discretos, que não se sobrepõem, porque *é assim* que dicionários funcionam, e dicionários são objetos de mercado. Profissionais da Lexicografia estão conscientes de que decidir se agrupam ou separam os significados de uma palavra é um trabalho inevitavelmente subjetivo (interpretativo), e frequentemente a decisão alternativa a respeito de agrupar ou separar – a decisão que não foi tomada - seria igualmente válida. Uma comparação entre dicionários nos mostra exatamente isso, e as divergências na anotação de sentidos também.

Com a anotação, a dificuldade na identificação dos sentidos das palavras como unidades discretas não é apenas impressão, ela pode ser medida: todos os trabalhos com anotações semânticas que buscam a desambiguação de sentidos costumam ter as mais baixas medidas de concordância entre anotadores, em torno de 70%. Na análise sintática, tipo de anotação aparentemente mais complexa, os números são superiores a 90%.

Levando em consideração essas limitações, podemos tentar um outro ângulo: o sentido não é uma propriedade intrínseca das palavras, mas uma abstração que só irá se concretizar no uso – traduzido no PLN como ocorrências da palavra em um corpus – e enquanto *decorrência de algum objetivo* ou *tarefa* (no caso da Lexicografia, o objetivo é o de fazer dicionários). No PLN, corpora diferentes e usos diferentes irão levar a representações diferentes de significados, e isso é o que vemos acontecer nos vetores de palavras contextuais, apresentados na próxima seção.

Corpora em português

Para a língua portuguesa, não há nenhum corpus no estilo do SemCor. Em 2015, participei de estudo piloto com o objetivo de medir tanto o grau de dificuldade que seria enfrentado na anotação de um corpus com synsets da OpenWordNet-PT quanto à correção de seus synsets (Freitas et al., 2015). Selecionamos 30 frases do corpus Bosque, anotamos (isto é alinhamos) todos os substantivos das frases segundo os *synsets* da OpenWordNet.PT. Apenas substantivos foram considerados justamente para facilitar a tarefa, já que verbos e adjetivos são mais polissêmicos. Diferente da proposta do SemCor, permitimos uma classificação múltipla, e mais de um *synset* poderia ser escolhido. O exercício contou com a colaboração de alunos de graduação em Letras, e relato algumas de nossas impressões:

Em 20% dos casos não foi possível escolher um *synset* adequado, mas nem sempre a impossibilidade era decorrência da inexistência do *synset*, mas de problemas nas etapas

anteriores de pré-processamento do corpus, como tokenização e lematização. Por exemplo, há palavras cujos sentidos variam ligeiramente quando estão no singular ou no plural: *recursos* pode ser o plural de *recurso*, mas com o sentido de *bens*, *riquezas*, *recursos financeiros*, será usado sempre no plural. Além disso, quando a tokenização é feita palavra por palavra, é difícil apontar para o *synset* adequado se ele for composto por uma unidade multipalavra. Alguns desses erros estão no quadro abaixo.

Palavra	Contexto
troca	em troca
realidade	na realidade
cadeia	transmissão em cadeia nacional
carteira	carteira de títulos
ferro	a ferro e fogo

Exemplos de palavras que não puderam ser alinhadas com a OpenWordNet

Com relação à anotação de campos semânticos, a língua portuguesa conta com os corpora do projeto AC/DC (e a sigla significa Acesso a Corpos/Disponibilização de Corpos), criado e mantido pela Linguateca. O material contém anotação relativa aos verbos do dizer/discurso relatado, campo semântico das cores, do corpo, doenças, relações familiares, locais, sentimentos e predicacões humanas. Nem todos os corpora contém todos os tipos de anotação. O material foi anotado utilizando regras linguísticas e léxicos, e está parcialmente revisto. Como se trata de um acervo dinâmico, novas anotações (novos campos semânticos) podem ser incorporadas (e lembro que o endereço eletrônico do AC/DC, bem como dos demais recursos mencionados, estão na página eletrônica, em @Ponteiros).

6.4.1. Palavras viram números – vetores de palavras

A dificuldade de lidar com os significados como unidades discretas e estáveis deixa aberto o caminho para que se invista em formas alternativas de lidar com o sentido, como é o caso da abordagem dos vetores de palavras (*word embeddings*). A ideia subjacente é que palavras que participam de contextos linguísticos similares tendem a ser similares. Por exemplo, *carro* e *bicicleta* estão próximos porque aparecem em contextos como “fui para o trabalho de carro” e “vou para o trabalho de bicicleta”.

Esta maneira de olhar para as palavras não é nova nos estudos da linguagem, pelo contrário (Wittgenstein, 1953; Harris, 1954; Firth, 1957). Na transposição para o PLN, uma palavra é representada como um vetor – e imagine um vetor como uma lista de números, como na figura abaixo. Cada posição nessa lista de números, que no exemplo tem 5 posições (“locomção”, “lazer”, “sustentável”, “comestível” e “quente”), corresponde a uma dimensão.

Locomoção	Lazer	Sustentável	Comestível	Quente
1	0	0	0	0
1	1	1	0	0
0	1	0	1	0

carro

bicicleta

sorvete

Cada palavra é representada por uma lista que tem lugar para 5 números diferentes, isto é, valores para 5 dimensões, e cada palavra tem uma combinação única desses 5 números, levando em conta as dimensões “locomotoção”, “lazer”, “sustentável”, “comestível” e “quente”:

carro [1,0,0,0,0]

bicicleta [1,1,1,0,0]

sorvete [0,1,0,1,0]

Continuando com a figura, vemos que *bicicleta* está próxima de *carro* na dimensão “locomotoção”, e próxima de *sorvete* na dimensão “lazer”. *Sorvete* e *carro*, por sua vez, estão mais distantes – distantes conforme as 5 características elencadas, é bom lembrar.

Assim, quando tratamos de vetores de palavras, o primeiro passo é entender que cada palavra é representada como um conjunto de números atribuídos em função das palavras que co-ocorrem com ela e das dimensões, e não como uma entrada de dicionário ou como um conceito. A transformação em números facilita o processamento pelos computadores.

Algumas perguntas devem ter surgido aqui:

De onde vêm esses números? O que significam? E de onde vêm as dimensões? Quem determina quais e quantas são as dimensões? Por que, no exemplo, foram usadas as dimensões “locomotoção”, “lazer”, “sustentável”, “comestível” e “quente”? Afinal, sabemos que as palavras têm muitas dimensões e podem ser parecidas em várias delas: forma (*engenheiro*; *motoqueiro*; *cinzeiro* – *esfarelar*; *cantar*; *dançar*); sentido (*cerveja*; *breja*; *gelada*); polaridade (*odiar*, *doença*; *pesadelo* - *comemorar*; *saúde*; *festa*), dentre muitas outras.

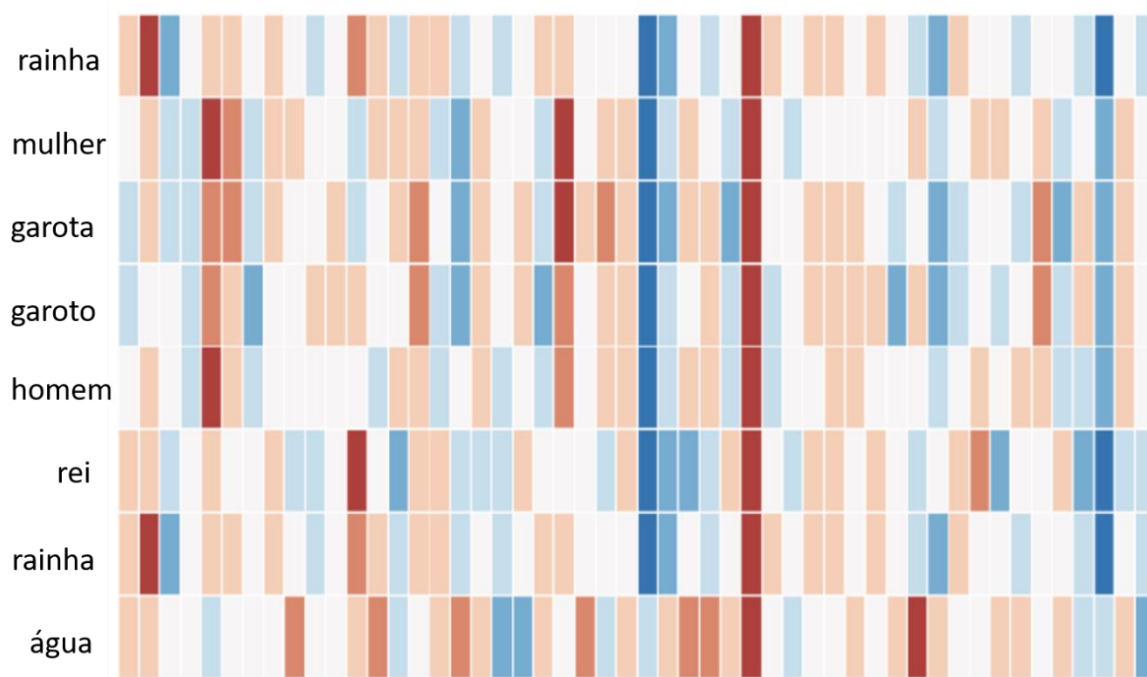
O que cada dimensão codifica (ainda) não sabemos. Trata-se de uma representação que foi aprendida pelas máquinas a partir dos dados – e esta é também uma das críticas que se faz este tipo de abordagem, como vimos na seção 2.1.

O que sabemos com certeza é que quanto mais dimensões, maior o trabalho no processamento automático. Há representações que usam 50 dimensões, representações que usam 100, representações que usam 300 dimensões...

Abaixo está um exemplo da representação vetorial da palavra *rei*, considerando um vetor de 50 dimensões, e por isso temos 50 números (em breve falaremos sobre a origem desses números):

[0.50451 , 0.68607 , -0.59517 , -0.022801, 0.60046 , -0.13498 , -0.08813 , 0.47377 , -0.61798 , -0.31012 , -0.076666, 1.493 , -0.034189, -0.98173 , 0.68229 , 0.81722 , -0.51874 , -0.31503 , -0.55809 , 0.66421 , 0.1961 , -0.13495 , -0.11476 , -0.30344 , 0.41177 , -2.223 , -1.0756 , -1.0783 , -0.34354 , 0.33505 , 1.9927 , -0.04234 , -0.64319 , 0.71125 , 0.49159 , 0.16754 , 0.34344 , -0.25663 , -0.8523 , 0.1661 , 0.40102 , 1.1685 , -1.0137 , -0.21585 , -0.15155 , 0.78321 , -0.91241 , -1.6106 , -0.64426 , -0.51042]

A lista acima nada nos informa relativamente à palavra *rei*. Para facilitar o entendimento, se transformamos cada número em uma cor, e comparamos os vetores das palavras *rei*, *rainha*, *homem*, *mulher*, *garota*, *garoto* e *água*, teríamos o seguinte (a figura é originalmente colorida e foi retirada de Alammr, J., (2019). Na página eletrônica do livro, em @Palavras viram números, é possível ver a imagem colorida, o que facilita a análise):



- *mulher* e *garota* são semelhantes em muitas posições; assim como *homem* e *garoto*;
- *garoto* e *garota* também têm semelhanças entre si, mas que não são compartilhadas com *mulher* ou *homem*. Podemos especular que a dimensão das semelhanças codifica algum tipo de *juventude*...
- *rei* e *rainha* também são semelhantes entre si em algumas posições, e distintos de todos os outros. Podemos especular que a semelhança codifica uma concepção vaga de *realeza*...
- Há uma coluna escura (em vermelho, no original) idêntica em todas as palavras, indicando que ao longo de uma dimensão (ao longo desta coluna escura), as palavras são semelhantes. Não sabemos para que serve cada dimensão, e podemos especular que a coluna escura talvez indique algo relativo ao caráter nominal das palavras representadas...
- Exceto pela palavra *água*, todas as outras representam *pessoas*. Há uma outra coluna escura (em azul escuro, no original) idêntica em todas as palavras, exceto em *água*. É possível que esta coluna tenha a ver com isso.

Quanto mais dimensões, mais pulverizada (ou distribuída) é a representação da palavra. Voltando à figura 1, podemos imaginar que ao invés de 5, teríamos 8 dimensões, com "locomção" se distribuindo entre "locomção terrestre"; "locomção aérea" e "locomção aquática".

Como já foi dito, o que cada dimensão contém será aprendido *automaticamente*, em função do número de dimensões que for estabelecido. E quem estabelece o número de dimensões?

Até o momento, não há pesquisas que demonstrem uma regra clara para a escolha da quantidade de dimensões; a decisão tem sido baseada na experiência, a partir do teste com diferentes números de dimensões. Daí que outra crítica a este tipo de abordagem é que a análise voltada para a quantidade de dimensões dos vetores de palavras não tem recebido atenção suficiente.

E de onde vêm os números que estão em cada uma das dimensões? Voltando aos 50 números que caracterizam a palavra *rei*: de onde surgiram?

Cada um dos números, de cada uma das dimensões, é o resultado de cálculos matemáticos que levam em conta as palavras que co-ocorrem com a palavra alvo (que co-ocorrem com *rei*) em um corpus. Este cálculo, por sua vez, é fruto de uma série de procedimentos – e é quanto aos tipos de procedimentos que as diferentes maneiras de construir os vetores irão variar: tamanho do contexto; quantas palavras à direita, ou à esquerda, ou em ambas as direções... Em comum, a necessidade de um corpus grande, sobre o qual serão feitos os cálculos.

Por outro lado, sabemos que uma mesma palavra pode aparecer em contextos diferentes – desde contextos completamente distintos, nos casos de homonímia como *banco* e *manga*, até contextos ligeiramente diferentes, como os exemplos de *trabalho*, que já vimos.

Aliás, é importante notar que, diferentemente do que supõe o senso comum, os casos de *banco* ou *manga*, apesar de fartamente citados como exemplos de ambiguidade, estão longe de ser prototípicos. Pelo contrário, são raros os casos em que dois sentidos se apresentam tão claramente distintos. O mais comum são casos como a palavra *trabalho*, em que os sentidos se cruzam e sobrepõem, como vimos na seção sobre anotação semântica. Voltando ao nosso exemplo: será que *rei do gado*, *rei da cocada preta* e *rei da Espanha* podem ser representados pelo mesmo vetor *rei*? Além disso, *rei da cocada preta* nos lembra da noção de palavra e do processo de tokenização, discutidos no capítulo 5: faz sentido pensar (ou calcular) 3 ou 4 vetores para *rei (da) cocada preta*? Ou estamos diante de uma unidade de sentido, e, portanto, de um vetor?

Independentemente da noção de palavra e considerando os algoritmos mais complexos, cada *uso* das palavras terá um vetor diferente. Por isso, nesses casos falamos de *vetores contextuais* (*contextual word embeddings*). Vetores produzidos a partir de conjuntos de dados diferentes levarão a representações diferentes.

Relembro aqui a importância dos dados no PLN, já mencionada nas páginas iniciais. A pesquisadora Robyn Speer comenta que, quando foi aplicar um algoritmo construído por ela para análise de sentimento baseado em vetores de palavras, percebeu que o algoritmo estava classificando restaurantes mexicanos como ruins. No entanto, esta avaliação negativa não encontrava respaldo nem na quantidade de estrelas usadas pelos usuários para avaliar os restaurantes, nem no texto das resenhas. Ela então descobriu que o motivo da avaliação ruim era aquilo que o sistema havia aprendido a partir dos dados: o sistema aprendeu a palavra “mexicano” a partir dos textos da internet, e na internet (aliás, nos

textos da internet nos Estados Unidos) a palavra *mexicano* sempre aparece associada à palavra *ilegal*, especialmente para associar *imigrantes mexicanos* a *imigrantes ilegais*. Com isso, o sistema acabou aprendendo que *mexicano* significa algo semelhante a *ilegal* e, portanto, deve significar algo ruim.

Mas se a completa dependência de dados é um ponto fraco deste tipo de abordagem, é também o seu ponto forte, e é incontestável que a incorporação de informação de vetores tem levado a resultados melhores em uma série de tarefas de PLN.

Por ser dependente dos dados (das palavras de um corpus), a representação do significado será, sempre, provisória, posto que outros dados poderiam levar a outras representações. Um exemplo esclarecedor a esse respeito, e desconcertante, diz respeito aos trechos abaixo: duas traduções *consagradas* dos *mesmos versos* (canto XI) da *Odisseia*, comentadas por Jorge Luis Borges no ensaio “As versões homéricas”:

Tradução 1 (Pope, 1725)	Tradução 2 (Butler, 1900)
<i>Quando os deuses coroaram de conquista as armas, quando os soberbos muros de Troia fumegaram por terra, a Grécia, para recompensar as galhardas fadigas de seu soldado, cumulou sua armada de incontáveis despojos. Assim, grande em glória, voltou seguro do estrondo marcial, sem uma cicatriz hostil, e embora as lanças se fechassem à sua volta em tormentas de ferro, seu jogo inútil foi inocente de ferimentos.</i>	<i>Uma vez ocupada a cidade, ele pôde apanhar e embarcar sua parte de benefícios havidos, que era uma forte soma. Saiu sem um arranhão de toda essa perigosa campanha. Já se sabe: tudo está em ter sorte.</i>

Quando pensamos no processo de tradução como transporte de significados entre língua A e língua B e, de maneira análoga, na representação computacional dos significados como o transporte (ou codificação) de significados entre *conceitos* e algum tipo de *representação semântica computacional* (como wordnets, framenets, ou outros tipos de representação semântica), acreditamos ser o significado do texto original, no primeiro caso, e dos conceitos, no segundo caso, um objeto estável, transportável, de contornos claros. No entanto, as traduções acima e os percalços na anotação de sentidos como aqueles relatados no projeto SemCor trazem dificuldades para esta visão do significado.

As traduções 1 e 2 contêm discrepâncias incontestáveis, que não podem ser descritas como paráfrases ou variações estilísticas. Tais discrepâncias, como já dito, são um desafio para uma visão de significado como a exposta anteriormente: se não estamos diante de paráfrases ou sinônimas e se cada tradução traz significados diferentes, como explicar que sejam, apesar de tão distintas, traduções não apenas aceitas, mas igualmente consagradas, de um mesmo original?

Para sair desse impasse, uma alternativa é compreender significados e conceitos como objetos heterogêneos e instáveis, e não como objetos diferentes apenas na forma (*amor* e *love*) e equivalentes no sentido ou em termos de mesmas *entidades mentais*. Compreender línguas como sistemas dinâmicos indissociáveis de práticas e valores históricos e sociais. Deste outro ângulo, deixamos de ver as palavras (e os textos) como repositórios de

significados que se deslocam inalterados pelo tempo e pelo espaço, mas como marcas de um significado que, ao se deslocar no tempo e no espaço, é compreendido/interpretado segundo contextos sociohistóricos. Esta alternativa explica porque interpretações de um mesmo texto podem variar, como as diferentes versões da Bíblia e os versos homéricos, e ajuda a entender por que podemos prescindir de uma representação estável e bem delimitada do significado e ainda assim (ou por isso mesmo) ter bons resultados em uma série de tarefas, inclusive a tradução automática. Os métodos de representação de palavras por meio de vetores contextuais, com a sua dependência dos dados vindos de grandes corpora, dão corpo a este caráter dinâmico próprio dos sentidos (codificam a “estabilidade provisória” dos sentidos).

Mas esta posição não deve ser vista como defendendo a irrelevância de maneiras convencionais de representar o sentido. Seria ingenuidade negar a utilidade dos dicionários, do mesmo modo que de recursos ao estilo wordnet. São maneiras diferentes de lidar com o significado: de um lado, abordagens baseadas em dados; de outro, em conhecimento.

Podemos agora retomar alguns pontos dos capítulos 2 e 5, e fechar o ciclo da geração de modelos de língua. A etapa de tokenização transforma o texto em “unidades de trabalho” (palavras, caracteres, subpalavras). Estas unidades são transformadas em vetores, e esses vetores serão processados pelas redes neurais.

6.5 REM (ou NER): anotação de entidades

A anotação de entidades mencionadas (em inglês *named entities*) é bem popular no PLN. A sigla NER (em inglês) ou REM (em português) refere-se à tarefa, e o R vem de *reconhecimento (recognition)*: Reconhecimento de Entidades Mencionadas (*Named Entities Recognition*). Entidades são elementos nominais relevantes para uma área de conhecimento ou tarefa.

A tarefa de REM consiste em duas tarefas relacionadas: *identificar* algo como em entidade e *classificar* conforme o contexto, isto é, anotar esta entidade de acordo com categorias pré-definidas como PESSOA, LUGAR, ORGANIZAÇÃO, TEMPO, QUANTIDADE etc. Por isso, o reconhecimento de entidades é considerado a primeira etapa do processamento semântico de textos:

- (1) A partida, originalmente marcada para esta noite em [Quito]_{LOCAL}, no [Equador]_{LOCAL}, será realizada na próxima sexta, em [Assunção]_{LOCAL}, no [Paraguai]_{LOCAL}.
- (2) Confira os convocados da [Seleção Brasileira]_{ORG} para jogos com [Equador]_{ORG} e [Paraguai]_{ORG}.

Originalmente, entidades referiam-se aos nomes próprios, mas a pista da maiúscula tem sido deixada de lado e qualquer substantivo incluído no conjunto de classes (o *tagset*) pré-determinado pode ser anotado como entidade, como vemos no quadro a seguir, que apresenta um texto anotado com entidades do domínio da Música.

O [*choro*]_{GENERO}, popularmente chamado de [*chorinho*]_{GENERO}, é um gênero de música popular e instrumental brasileira, que surgiu no [*Rio de Janeiro*]_{LOCAL} em meados do [*século XIX*]_{TEMPO}.

Os primeiros conjuntos de [*choro*]_{GENERO} surgiram por volta da [*década de 1870*]_{TEMPO}, nascidos nas biroschas do bairro [*Cidade Nova*]_{LOCAL} e nos quintais dos [*subúrbios cariocas*]_{LOCAL}. O flautista e compositor [*Joaquim Antônio da Silva Calado*]_{PESSOA}, os pianistas [*Ernesto Nazareth*]_{PESSOA} e [*Chiquinha Gonzaga*]_{PESSOA}, e o maestro [*Anacleto de Medeiros*]_{PESSOA} compuseram [*quadrilhas*]_{GENERO}, [*polcas*]_{GENERO}, [*tangos*]_{GENERO}, [*maxixes*]_{GENERO}, [*xotes*]_{GENERO} e [*marchas*]_{GENERO}, estabelecendo os pilares do [*choro*]_{GENERO} e da música popular carioca da virada do [*século XIX*]_{TEMPO} para o [*século XX*]_{TEMPO}. Herdeiro de toda essa tradição musical, [*Pixinguinha*]_{PESSOA} consolidou o [*choro*]_{GENERO} como gênero musical

, levando o virtuosismo na [*flauta*]_{INSTRUMENTO} e aperfeiçoando a linguagem do contraponto com seu [*saxofone*]_{INSTRUMENTO} e organizou inúmeros grupos musicais, tornando-se o maior compositor de [*choro*]_{GENERO}.

Quadro 1: Exemplo de anotação de entidades do domínio da Música

Quanto mais “novidade” há no domínio que será anotado com entidades, mais trabalho será necessário para escolher as categorias de anotação (o *tagset*) relevantes para o domínio. Na Biologia, genes e processos genéticos podem ser entidades; no Direito, leis, atores e argumentos podem ser entidades. No quadro 1, que exemplifica uma anotação no domínio da Música, além de classes gerais como PESSOA, LUGAR e TEMPO, há etiquetas específicas como GÊNERO e INSTRUMENTO, e poderia haver outras.

Desafios linguísticos

A anotação de entidades traz os seguintes desafios linguísticos:

- Decidir o que deve ser anotado (identificar uma entidade)

No exemplo do quadro 1, é possível questionar o motivo de a palavra *biroschas* não ter sido anotada, já que também é um lugar e pode fornecer informações interessantes – músicas nascidas em espaços de elite e nascidas em espaços populares, por exemplo. A pergunta ou tarefa que motiva a anotação é determinante na definição do que vai ser considerado entidade.

Também é em função da motivação que palavras como *flautista*, *compositor*, *pianista*, e *maestro* serão anotadas. Se forem anotadas, é preciso decidir se serão do tipo PESSOA, que já existe, ou formarão uma nova classe PROFISSÃO, que pode ser um tipo de PESSOA. O mesmo se aplica aos estilos de música: *música popular*, *instrumental*. Neste caso, será preciso decidir se integram a classe GÊNERO ao lado de *choro*, *xote* etc. Ou seja, conforme as escolhas de anotação o texto do quadro 1 poderá ser anotado de algumas maneiras, todas corretas.

- Classificar a entidade

Como vimos na anotação de sentido, a polissemia está presente boa parte das palavras, e na anotação de entidades este traço é evidente. Os exemplos (1) e (2) no início da seção mostram que *Equador* e *Paraguai* podem ser LOCAL ou ORGANIZACAO conforme o contexto. Podemos (e devemos) discutir se *Seleção Brasileira*, *Equador* e *Paraguai*, quando utilizados para fazer menção aos times, são ORGANIZACAO ou PESSOA (os jogadores). Pode ainda ser decidido que é aceitável usar mais de uma classe simultaneamente, e todos no exemplo ficarão anotados como PESSOA|ORGANIZACAO. Na frase abaixo, podemos interpretar (e classificar) *Brasil* como LOCAL, PESSOA ou ambos, já que uma leitura não exclui a outra – e se não fosse a necessidade imposta pela anotação de classificar considerando essas etiquetas, a dúvida sobre ser LOCAL ou PESSOA nem existiria, uma vez que não impede ou diminui nossa compreensão da frase. Faz parte das decisões de anotação aceitar mais de uma classificação simultaneamente.

Mais de 600 mil pessoas morreram na pandemia que atingiu o **Brasil**.

- Segmentar a entidade

Uma entidade também pode ser decomposta de maneiras diferentes. Decisões de segmentação dizem respeito a considerar entidades maiores ou entidades mais granulares, que correspondem a uma leitura mais composicional. No quadro 1 temos [década de 1870] e [século XX]. Mas diferentes alternativas são aceitáveis, como vemos:

[década de 1870]	[século XX]
década de [1870]	século [XX]
[década] [de] [1870]	[século] [XX]

No trecho anotado abaixo, é discutível a classificação de *Paraguais*, mas se não fosse o

Um livro de geografia usado nas escolas públicas do [Estado de São Paulo]_{ORG} traz dois [Paraguais]_{LOCAL} e exclui o [Equador]_{LOCAL} de um mapa da [América do Sul]_{LOCAL}. A [Secretaria de Educação de São Paulo]_{ORG} informou que a [Fundação Vanzolini]_{ORG}, responsável pela impressão dos [Cadernos do Aluno]_{OBRA}, substituirá os [500 mil]_{QUANTD} livros que têm o mapa errado.

plural dificilmente hesitaríamos. Quanto à segmentação, as possibilidades abaixo são igualmente aceitáveis e às vezes a cada segmentação corresponde uma classificação diferente.

[Estado] _{ORG} de [São Paulo] _{ORG} ou LOCAL
[Secretaria de Educação] _{ORG} de [São Paulo] _{ORG} ou LOCAL
[Secretaria] _{ORG} de [Educação] _{SABER} de [São Paulo] _{ORG} ou LOCAL

Outro elemento que contribui para hesitações relativas à segmentação é o uso pouco sistemático que costumamos fazer das maiúsculas, como podemos nas frases a seguir

Mil dólares foram apreendidos anteontem no **porto** de Santos (SP).

O **Porto de Santos** é um porto estuarino, localizado nos municípios de Santos, Guarujá e Cubatão, no estado de São Paulo.

Estes são exemplos simples, mas mostram como é preciso combinar (e documentar) o que será anotado, mesmo em casos fáceis.

Se dispomos de uma ontologia e de uma lista de palavras relativas à área que estamos anotando – e assumindo que ontologia e anotação servem a interesses parecidos – podemos pular toda a parte relativa à identificação, porque já sabemos o que estamos procurando; já sabemos o que devemos identificar. Como vimos no capítulo 4, uma ontologia define as classes de um domínio, e podemos usá-las como etiquetas na anotação. Mas partir de uma ontologia ou léxico não significa que não existe trabalho a ser feito. Uma vez que a classificação de uma entidade pode variar conforme o contexto, a estratégia de lidar com entidades a partir de léxicos (por exemplo, uma lista de entidades do tipo LOCAL ou do tipo PESSOA) é insuficiente. Na preparação de material padrão ouro, a utilização de regras pode ser uma boa estratégia para a desambiguação.

Podemos explorar um pouco a formalização simplificada usada em 2.1.1. e tomar como problema a desambiguação das duas frases exemplo do início desta seção (repetidas abaixo). Sabemos que *Quito* e *Equador* são ambíguos quanto às classes LOCAL e PESSOA:

- (1) A partida, originalmente marcada para esta noite em [Quito], no [Equador], será realizada na próxima sexta, em [Assunção], no [Paraguai].
- (2) Confira os convocados da [Seleção Brasileira] para jogos com [Equador] e [Paraguai].

Analisando os contextos das frases 1 e 2, podemos criar uma regra

SELECT (LOCAL) IF (-1 “no”)

que diz *selecione a etiqueta LOCAL se à sua esquerda está a palavra “no”*. Mas essa regra é muito específica, que só irá selecionar a classe LOCAL nesse caso exato. Analisando as demais ocorrências de LOCAL no corpus, podemos fazer

SELECT (LOCAL) IF (-1 “em|no|na”)

Do mesmo modo, podemos fazer

SELECT (ORGANIZACAO) IF (-2 “jogo|jogar”) (-1 “com|contra”)

que diz *selecione a etiqueta ORGANIZACAO se 2 palavras à esquerda houver a palavra jogo ou a palavra jogar, e se 1 palavra à esquerda está a palavra “com” ou a palavra “contra”*. Se a decisão for anotar times como entidades do tipo PESSOA, é preciso alterar a regra, trocando ORGANIZACAO por PESSOA.

Podemos ainda fazer uso de outras camadas de anotação. Na regra abaixo, deixamos de lado a informação de posição (quantidade de palavras à direita ou à esquerda) para usar informação sintática (ser argumento de certos verbos, não importa a posição ou a quantidade de palavras entre o verbo e o argumento). A regra indica que *se a entidade é o argumento de verbos como regressar, chegar, voltar ela será do tipo LOCAL*. O que fazemos aqui é usar a anotação de outras camadas para melhorar a anotação das entidades, e já vislumbramos uma função as anotações gramaticais. E para que a regra faça o que queremos, já definimos que a classe ARG não contempla os sujeitos dos verbos.

SELECT (LOCAL) IF (ARG: “regressar|chegar|voltar”)

Corpora em português

A língua portuguesa conta com o material produzido para a avaliação conjunta HAREM, chamado Coleção Dourada do HAREM (falamos dela em 2.1.2). Os textos da coleção são de gêneros variados, e o conjunto de classes e subclasses foi escolhido levando em conta os interesses da comunidade de PLN na época, contando com 10 categorias principais: ABSTRACAO, ACONTECIMENTO, COISA, LOCAL, OBRA, ORGANIZACAO, PESSOA, TEMPO, VALOR e OUTRO. Outras opções de anotação do HAREM foram anotar apenas entidades grafadas com nome próprio, aceitar a classificação múltipla e segmentações alternativas para uma mesma entidade. Todo o material está disponível na página do HAREM.

Outro material é o corpus Summit++, um corpus multicamadas que contém anotação de entidades classificadas conforme as orientações do HAREM.

6.6 Relações entre entidades e extração de informação

Após a detecção e classificação das entidades, o passo seguinte na estruturação das informações dos textos é *relacionar* essas entidades. Por exemplo, uma ORGANIZACAO se localiza em um LOCAL; uma PESSOA está vinculada a uma ORGANIZACAO, uma PESSOA nasce em um LOCAL etc. Neste tipo de anotação o que se faz pré-definir tipos de relações que interessam e procurar essas relações no texto. Por exemplo, interessa detectar relações de *causalidade* entre entidades do tipo DOENÇA e SINTOMA, *vínculos profissionais* entre PESSOA e ORGANIZACAO, relações de localização entre ACIDENTE GEOGRÁFICO e LOCAL. Assim como na anotação de entidades, o inventário das relações que serão anotadas pode vir da tarefa, dos interesses do domínio ou de recursos externos como ontologias. Poder contar com anotação sintática no processo de anotação (saber quem é sujeito, quem é objeto, quais elementos estão coordenados com quais elementos etc), sempre será uma grande vantagem.

Vamos continuar com nosso pequeno corpus sobre o *choro* usado para exemplificar a anotação de entidades (quadro 1), e vamos aproveitar as relações já esboçadas na ontologia da música (capítulo 4). O resultado da anotação seria a criação das seguintes relações:

ARG1: choro ARG2: chorinho RELAÇÃO: <u>identidade</u>	ARG1: Joaquim Antônio da Silva Calado; Ernesto Nazareth; Chiquinha Gonzaga; Anacleto de Medeiros; Pixinguinha ARG2: choro RELAÇÃO: <u>parte de</u>
ARG1: Rio de Janeiro; Cidade Nova; subúrbios cariocas ARG2: choro RELAÇÃO: <u>local de nascimento</u>	ARG1: flauta; saxofone ARG2: choro RELAÇÃO: <u>parte de</u>
ARG1: década de 1870; século XIX ARG2: choro RELAÇÃO: <u>data de nascimento</u>	

O resultado da extração de relações pode ser usado para a construção de uma base de fatos do domínio, uma base de conhecimento, ou ainda para alimentar uma ontologia com instâncias, isto é, exemplos das classes.

Do nosso corpus podemos extrair os seguintes fatos

Choro é o mesmo que chorinho
 Choro nasceu no Rio de Janeiro
 Choro nasceu na década de 1870
 Joaquim Antônio da Silva Calado é parte de choro
 Flauta é parte de choro

Desafios linguísticos

Se o conjunto de relações de interesse (o tagset de relações) já está definido, o desafio será encontrar as relações no corpus e anotá-las. Uma maneira de conduzir esta anotação é buscar padrões linguísticos, utilizando uma estratégia bem parecida com aquela apresentada em 4.1, já que relações entre entidades também são relações semânticas. Primeiro precisamos analisar o material, e começamos com a frase inicial do nosso pequeno corpus, que é rica em informações (e isto não é sorte: frases e parágrafos introdutórios de textos informativos/técnicos são carregadas de informação):

O [*choro*], popularmente chamado de [*chorinho*], é um gênero de música popular e instrumental brasileira, que surgiu no [*Rio de Janeiro*] em meados do [*século XIX*].

O trecho

O *choro*, popularmente chamado de *chorinho*,

Nos fornece o padrão abaixo (parênteses indicam opcionalidade), que sinaliza uma relação de identidade:

N “,” (ADV) chamado de N “,”

Após mais alguma exploração do corpus, percebemos que o “conhecido como” (*choro*, *também conhecido como chorinho*,) leva ao mesmo tipo de relação. Teremos então dois padrões que detectam a presença de uma relação de identidade:

N “,” (ADV) chamado de N “,”

N“,” (ADV) conhecido como N “,”

Relação: identidade

Na anotação de relações, sabemos a relação que estamos procurando (no caso, *identidade*) e vamos procurar exemplos dessa relação no corpus. A busca por padrões típicos de cada relação auxilia e acelera o processo de anotação, mas não é uma exigência.

O segmento seguinte da frase também é informativo, e nos fornece um hiperônimo de *choro*, “música popular” (também podemos discutir se a entidade é “música popular” ou apenas “música”, isto é, se modificadores do núcleo do sintagma são considerados parte da entidade. Esta é uma decisão vinculada à anotação de entidades, e não de relações entre entidades).

O *choro*, popularmente chamado de *chorinho*, é um gênero de *música popular*...

Porém esta relação é mais facilmente capturada se contamos com análise sintática, uma vez que sujeito e predicado estão distantes na frase:

N/SUJ é um gênero de SN/PREDICADO

Relação: hiperonímia

Se pudermos contar com a anotação de entidades, o processo é mais rápido, pois teremos apenas as estruturas que envolvem termos do domínio:

N/SUJ/GENERO é um gênero de SN/PREDICADO

Como já foi indicado, é importante saber que nem sempre a relação será materializada por um padrão linguístico, mas encontrar padrões ajuda o processo de anotação. A frase abaixo deve ser anotada com uma relação entre *choro* e *gênero musical*, mesmo que não seja expressa por um padrão.

Até o surgimento do *choro* não existia ainda no país um *gênero musical* que pudesse ser considerado brasileiro.

Porque as relações não necessariamente precisam estar expressas por certos padrões, este é um tipo de anotação que precisa ser feito com a colaboração de especialistas da área em questão, pois em muitos casos apenas especialistas saberão reconhecer a existência de uma relação entre dois termos em uma frase.

O segmento seguinte de nossa frase inicial também é informativo, mas é mais complexo de ser tratado. O motivo da complexidade é o pronome relativo, *que*, que é correferente de *choro*.

O *choro* (...) é um gênero de música popular (...), que surgiu no *Rio de Janeiro*

Se quisermos identificar que o sujeito de *surgir no Rio de Janeiro* é o *choro*, precisaremos resolver a cadeia de correferência entre o *que* e *choro* (e anotação de correferência é a próxima seção). Isto é, podemos anotar uma relação entre *que* e *Rio de Janeiro*, mas esta relação só será útil se soubermos o valor do *que*. Vamos fingir que isso já foi resolvido, e descrever mais um padrão para detectar relações. A anotação de entidades ajuda ainda mais aqui, uma vez que *surgir* pode ser completado com informações de tempo e de lugar. Por isso também, já podemos criar padrões para relação de lugar e de tempo:

N/SUJ/GENERO surgir ADJUNTO_ADVERBIAL/LOCAL

Relação: local
de nascimento

N/SUJ/GENERO surgir ADJUNTO_ADVERBIAL/TEMPO

Relação: data
de nascimento

Em ambos os casos também é possível generalizar e incluir o verbo *nascer* ao lado de *surgir* (dentre alguns outros) como verbos que indicam relação de temporalidade e de localização entre as entidades.

Cada padrão pode levar à criação de uma regra de anotação, e quanto melhores os padrões, menos casos errados retornarão. Por outro lado, quanto mais camadas de anotação envolvemos nas regras, mais complexa é também sua formulação. Para os dois últimos padrões, podemos ter regras como as abaixo (não estamos nos importando com a formalização, a regra é apenas uma maneira de ver o que é preciso codificar):

a:[pos=N & func=SUJ & ner=GENERO] b:[lema=surgir & func=PRED:a] c:[func=AADV:b & ner=LOCAL] >> b:[lema=surgir & func=PRED:a & relner=localsurgimento:c_a]

Lê-se da seguinte maneira: a regra conta com três elementos: “a”; “b”; “c”. A detecção do padrão corresponde ao lado esquerdo, antes do sinal “>” e o que será preciso fazer está do lado direito. Tudo no texto que satisfizer às condições indicadas do lado esquerdo será alvo de uma transformação, indicada do lado direito. As condições do lado esquerdo são:

Elemento a: deve ser um N; deve ter a função sintática de sujeito; deve ser uma entidade da classe GENERO.

Elemento b: deve ter o lema “surgir”; deve ter a função sintática de predador do elemento “a”. Ou seja, é preciso garantir que o predador esteja relacionado ao elemento “a”, e não a qualquer outro sujeito da frase.

Elemento c: deve ter a função sintática de adjunto adverbial; deve ser uma entidade da classe LOCAL.

Se essas condições forem satisfeitas, a regra faz o seguinte – neste caso, a regra só altera o elemento “b”:

Elemento b: continua com o lema “surgir”; continua com a função sintática de predador do elemento “a”; recebe a etiqueta de relações entre entidades “relner” com o valor “localsurgimento:c_a”, que indica que se trata de uma relação de “local de surgimento” entre os elementos “c” e “a”. Esta maneira de formalizar indica também que a relação entre entidades estará anotada no verbo que faz a relação. Poderia não ser assim, poderíamos estabelecer que a anotação das relações estará em cada entidade envolvida. Existem algumas maneiras de fazer as mesmas coisas, mas não iremos nos ocupar disso aqui.

Apesar de muito úteis, não devemos nos esquecer que padrões muito precisos frequentemente têm o inconveniente de deixar muita coisa de fora (podem ser pouco abrangentes), e esta é uma característica deste tipo de abordagem de anotação. Sabemos que os padrões capturam muitas coisas, mas não resolvem tudo, e sempre será necessário rever o resultado da aplicação das regras no corpus, pois sempre há casos não previstos (como indica a lei de Zipf). Se tudo pudesse ser resolvido com regras, o aprendizado estatístico não faria tanto sucesso.

A anotação de relações torna explícita uma série de relações entre entidades, mas ela também é mais complexa do ponto de vista da execução. Para poder usar uma estratégia de regras ao invés de ler o corpus completo, frase a frase, buscando relações exclusivamente a partir da nossa leitura estamos pressupondo algumas coisas:

1. Precisamos do texto lematizado (para saber que *surgiram*, *surgiu* etc) são formas do verbo *surgir*;
2. Precisamos de análise de POS para encontrar advérbios, substantivos etc;
3. Precisamos da análise sintática para encontrar sujeitos e adjuntos adverbiais, independentemente da posição que estejam na frase;
4. Precisamos de cadeias de correferência para poder construir relações informativas (apenas a palavra “que” como sujeito não é informativa);
5. Precisamos de anotação de entidades para facilitar a seleção e a revisão.

Nem sempre teremos tudo isso à disposição. Conforme o tipo de texto, podemos confiar em ferramentas para fazer automaticamente 1, 2 e 3. O tipo de texto é um aspecto importante porque a maioria das ferramentas para o português é treinada em corpus jornalístico, então tudo o que se afasta disso tende a ter um desempenho pior.

Além disso, é fundamental uma ferramenta que faça a interface entre o corpus e as regras que irão procurar os padrões no corpus, assim como uma forma de procurar os padrões no corpus (uma sintaxe) e uma ferramenta para incluir as relações em cada caso.

Justamente porque demanda uma série de processamentos linguísticos prévios, além da necessidade de especialistas de linguística (porque estão mais aptos a perceber e descrever os padrões linguísticos usando informações como adjunto adverbial, objeto etc) e em alguns casos especialistas também do domínio, métodos estatísticos irão tentar capturar as relações entre entidades de uma outra maneira, a partir exclusivamente dos dados. Observando os contextos em que cada palavra aparece (como vimos em na representação do sentido com os vetores de palavras), teríamos como resultado o agrupamento de palavras como choro-chorinho; choro-flauta, choro-música, mas sem rótulos que explicitem a relação entre elas. Sabemos “apenas” que são palavras relacionadas (não que isto seja pouco). Para sistemas de raciocínio, que irão produzir conhecimento novo – tudo o que for dito sobre *choro* se aplica igualmente a *chorinho*, já que há uma relação de identidade entre eles, por exemplo –, rótulos são importantes; para a estruturação explícita da informação em documentos rótulos são importantes, mas isto não se aplica a todas as tarefas do PLN.

Corpora em português

Para a língua portuguesa, temos três corpora padrão ouro com relações entre entidades.

O primeiro é dataset relativo ao ReReLEM (Reconhecimento de Relações entre Entidades Mencionadas), uma tarefa piloto da edição de 2008 do HAREM cujo objetivo relacionar semanticamente as entidades anotadas no HAREM. Originalmente, o ReReLEM contou com 4 classes – identidade; inclusão; ocorrência; outra. Posteriormente, a classe *outra* foi especificada, levando a um total de 24 relações. A página do evento exemplifica cada uma das relações, que estão reproduzidas abaixo.

autor_de/obra_de; causador_de; data_morte; data_nascimento; datado_de/data_de; ident; inclui/incluido; local_morte; localizado_em/localizacao_de; natural_de/local_nascimento_de; nome_de/nomeado_por; ocorre_em/sede_de; outra_edicao; outrarel; participante_em/ter_participacao_de; periodo_vida; personagem_de; praticado_em/pratica_se/praticante_de/praticado_por; produtor_de/produzido_por; proprietario_de/propriedade_de; relacao_familiar; relacao_profissional; residente_em/residencia_de; vinculo_inst
--

O corpus Summ-it++ também contém anotações padrão ouro com relações entre as entidades. Especificamente, estão anotadas todas as relações possíveis entre entidades do tipo ORGANIZACAO, PESSOA e LOCAL.

O corpus de Garcia e Gamallo (2014) é um corpus multilíngue que contém material em Português, Galego e Espanhol. O corpus é anotado com relações de IDENTIDADE entre

entidades do tipo PESSOA, e por isso é mais apropriado descrevê-lo como um corpus padrão ouro anotado quanto à correferência, tema de 6.7.

6.6.1. Extração de informação aberta

Uma maneira menos restrita de criar relações entre elementos de um texto é por meio da extração de informação aberta (*open information extraction*). Neste tipo de extração de informação não há predefinição de relações nem exigência de que as relações aconteçam entre entidades. As relações são estabelecidas entre os argumentos dos verbos (sujeito e objeto) e a partir disso são criadas relações com a estrutura ARGUMENTO1_verbo_ARGUMENTO2, chamadas triplas.

Este tipo de extração de informação é chamado de “aberta”, justamente por não predefinir que tipos de relação devem ser encontradas. Na extração de informação aberta não está em jogo, ao menos inicialmente, a etiquetagem das relações. Como as relações são identificadas a partir dos verbos, a quantidade de verbos diferentes corresponderá à quantidade de relações diferentes. Porque depende da informação linguística que está no corpus, a anotação sintática anterior é imprescindível para a extração das triplas.

6.7 Correferência

A anotação de correferência é mais uma etapa na estruturação da informação dos textos, e se aproxima da anotação de relação entre entidades quando consideramos apenas a relação de identidade. Por outro lado, na anotação de correferência, nem sempre a cadeia de correferência precisa se estabelecer entre as entidades de um texto.

Apesar das aproximações, a anotação de correferência costuma ser apresentada como uma tarefa independente, na qual são estabelecidas cadeias de correferência entre os elementos de um texto, que podem ser substantivos (comuns ou próprios), pronomes (possessivos, demonstrativos, relativos, clíticos) e elementos elípticos, dentre outros.

A anotação de correferência não se restringe aos limites de uma frase, podendo se estabelecer entre parágrafos, ou no âmbito do documento inteiro. Trata-se de uma tarefa com uma longa tradição no PLN, visto sua ampla relevância na extração de informação e na sumarização automática. No entanto, trata-se igualmente de uma tarefa bastante complexa do ponto de vista das máquinas, demandando uma boa dose de ‘interpretação’ de texto. Cada um dos trechos abaixo indica, em negrito, elementos correferentes:

1. **Trump**[1] pede recontagem em **Wisconsin**[2] e tenta suspender apuração na Pensilvânia, Geórgia e Michigan.
Campanha de reeleição do **republicano**[1] tenta revisão dos votos no **estado em que ficou atrás**[2] e a interrupção da contagem em dois estados em que está na frente e um em que foi superado.
2. O próximo ano ficará marcado como o ano em que o Governo **decidiu** construir um novo aeroporto, uma **decisão** polêmica.

Do ponto de vista humano, o grau de dificuldade pode ser medido de uma outra maneira: em avaliações (escolares) que têm como objetivo verificar a capacidade de “interpretação de texto” dos alunos, costuma haver pelo menos uma questão que envolva a resolução de correferência (por exemplo, saber a quem um determinado pronome se refere). Ou seja, se achamos que seres humanos, após mais de uma década de contato com a língua e dos anos de escolarização, podem não ser capazes de identificar com sucesso as cadeias de correferência em um texto, não deve espantar que, para máquinas, esta também seja uma atividade complexa. O desempenho da resolução automática da correferência, para o

inglês, está em 80%, e a melhoria no desempenho das máquinas se deve à incorporação de informação linguística oriunda dos vetores de palavras, comentados na seção 6.4.1.

Desafios linguísticos

A anotação de correferência tem como desafio o desenvolvimento de estratégias para guiar e otimizar a anotação. O estabelecimento correto de cadeias de correferência depende de informação semântica e contextual, o que explica a melhoria do desempenho das máquinas quando leva em conta informação dos vetores de palavras. Por outro lado, como vimos em 6.4, não temos material com anotação de sentidos em português, o que facilitaria o desenvolvimento de estratégias de anotação – por exemplo, candidatos a correferência compartilham o mesmo sentido ou têm um mesmo hiperônimo, como *Wisconsin* e *estado*, no exemplo (1) acima.

Corpora disponíveis

A língua portuguesa conta com pelo menos três corpora anotados com correferência: o corpus Corref-PT, o corpus Summ-it e o corpus de Garcia e Gammalo (2014), que contém anotação de correferência entre entidades do tipo pessoa.

6.8 Inferências e Similaridade semântica

No mundo do PLN, Similaridade Semântica refere-se à determinação do grau de semelhança entre duas porções de um texto. A detecção de similaridade é utilizada em programas de detecção de plágio e em tarefas de NLI (*Natural Language Inference* – algo como *Inferências de Linguagem Natural*), que lidam com raciocínio lógico. A anotação é uma tarefa de classificação a respeito do tipo de relação lógica entre duas frases (ou proposições): a frase “hipótese” e a frase “premissa”. A classificação é feita conforme três etiquetas: acarretamento, contradição e neutro.

acarretamento - o sentido de uma frase está incluído no sentido de outra. Ou, a frase hipótese é verdadeira dada a frase premissa;

contradição - o sentido de uma frase contradiz o sentido de outra. Ou, a frase hipótese não é verdadeira dada a frase premissa;

neutro - a frase hipótese pode ser verdadeira dada a frase premissa.

O corpus SICK (*Sentences Involving Compositional Knowledge*) e o corpus SNLI (*Stanford Natural Language Inference*) exemplificam este tipo de anotação. Em ambos os casos, as frases vêm de legendas de imagens. Esta é uma maneira de garantir frases declarativas e no tempo presente (inferências precisam de estabilidade, como vimos em 3.2).

No corpus SICK, ambas as frases são fornecidas para os anotadores, que devem também avaliar a adequação da relação semântica atribuída. Cada relação de inferência é classificada em uma escala de 5 pontos, sendo 5 uma relação que está perfeitamente exemplificada pelo par de frases, e 1 uma relação fracamente exemplificada pelo par de frases. Abaixo estão alguns exemplos de frases anotadas retiradas do corpus SICK-BR.

Frase premissa: Uma criança risonha está segurando uma pistola de água e sendo espirrada com água.

Frase hipótese: Uma criança está segurando uma pistola de água.
Etiqueta (anotação): acarretamento. Pontuação da similaridade: 4.5

Frase premissa: Não há nenhum homem de jaqueta preta fazendo truques em uma moto.
Frase hipótese: Uma pessoa de blusa preta está fazendo truques em uma moto.
Etiqueta (anotação): contradição. Pontuação da similaridade: 3.6

Frase premissa: Ninguém está dirigindo uma bicicleta com uma roda.
Frase hipótese: Uma pessoa de blusa preta está fazendo truques em uma moto.
Etiqueta (anotação): neutro. Pontuação da similaridade: 2.8

Diferentemente do SICK, no corpus SNLI não há pontuação de similaridade e as frases hipótese são construídas pelos anotadores a partir da frase premissa, conforme algumas instruções. Como as instruções são fundamentais para entender a qualidade do que é criado, estão reproduzidas abaixo:

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.

Desafios linguísticos

É difícil abordar desafios linguísticos desta tarefa porque tanto o SICK como o SNLI foram criados utilizando um processo de anotação colaborativa em grande escala (*crowdsourcing annotation*), quando se contrata pessoas de qualquer formação para a tarefa de classificação. Este tipo de estratégia é comum quando se precisa de um dataset grande (o SICK tem 10 mil pares de frases e o SNLI tem quase 560 mil pares) e a tarefa de anotação não precisa de conhecimento linguístico específico, como vimos nas instruções do SNLI.

Por outro lado, o olhar linguístico para este tipo de material é pode perceber características linguísticas capazes de impactar o desempenho dos sistemas. Após analisar as frases de datasets como o SNLI, uma equipe de pesquisadores descobriu que as instruções para a formulação das frases hipóteses davam margem a que anotadores usassem sistematicamente certas estratégias, que os autores chamam de “artefatos linguísticos”. Devido a esses “artefatos”, que foram aprendidos pelas máquinas, era possível acertar a relação entre premissa e hipótese olhando apenas para a hipótese e ignorando a premissa. Especificamente, Gururangan e sua equipe (2018) descobriram que (i) frases de acarretamento tendiam a remover informação de gênero e número e usar termos genéricos

como *pessoa*, *animal*, *instrumento*; (ii) frases neutras eram frequentemente construídas pela adição de uma *oração adverbial final* à oração principal ou pela adição de *adjetivos modificadores*; e (iii) frases com contradição eram criadas por meio da adição de negações como *não*, *nunca*, *ninguém*.

O uso recorrente destas estratégias é explicado por serem exatamente as estratégias presentes nas instruções da tarefa (que estão no quadro 1). Os autores então criaram um dataset fácil, apenas com as frases construídas por meio dos artefatos linguísticos, um dataset difícil, com frases que não faziam uso dos artefatos, e compararam com o dataset completo, com ambas as frases. O que encontraram foi uma piora de 10% quando compararam os resultados do dataset completo com o dataset “difícil”. Além disso, criaram um sistema muito simples, que deveria apenas prestar atenção nos artefatos linguísticos e ignorar a frase premissa, e este sistema conseguiu acertar mais da metade dos casos no dataset completo.

No SICK, a atribuição de pontuação é certamente um ponto delicado. Na construção do corpus SICK-BR, os autores comentam sobre discordâncias relativas à pontuação atribuída pelo corpus original. Uma vez que no corpus SICK não há instruções claras indicando como pontuar a relação de similaridade entre as frases, nem sempre há consistência na atribuição das etiquetas. No entanto, como o objetivo dos autores do SICK-BR não era reanotar o corpus SICK, mas produzir material para a língua portuguesa que também pudesse estar alinhado com o material em inglês, os eventuais deslizos de anotação foram mantidos na versão brasileira do material.

Corpora disponíveis

Para avançar com a área de raciocínio semântico em português, foi criado o corpus SICK-BR, que consistiu na tradução e adaptação do material do corpus SICK. Segundo seus autores, a tradução foi especialmente cuidadosa, a fim de garantir que (i) as traduções mantivessem o mesmo valor de verdade dos pares originais; (ii) os mesmos fenômenos discutidos no SICK pudessem ser discutidos no SICK-BR; (iii) houvesse um alinhamento entre as frases do SICK e do SICK-BR; (iv) as frases soassem naturais em português (Real et al., 2018). Subjacente à tradução, a hipótese de que os fenômenos lógicos seriam similares em ambas as línguas, e que as relações de acarretamento e de contradição funcionariam da mesma maneira em inglês e em português, o que se confirmou. Na avaliação de 800 pares de frases, em apenas 20 deles os autores discordaram da relação atribuída, sendo que em 14 deles a relação já estava errada no material original em inglês, e nos restantes 6 pares a atribuição de relações foi considerada discutível também no original.

O corpus SICK-BR serviu como base para o corpus ASSIN, que por sua vez foi o corpus criado para a avaliação conjunta ASSIN-2, que propôs as tarefas Similaridade Semântica Textual e Reconhecimento de Inferência para a língua portuguesa (ver também seção 2.1.2).

6.9 Opinião e sentimento

As áreas de mineração de opinião e análise de sentimento lidam com a identificação de opiniões, avaliações e atitudes direcionadas a entidades como pessoas, produtos e organizações. Embora muitos trabalhos da área façam uso de léxicos de sentimentos e

polaridades (o capítulo 4 traz um exemplo de um léxico de sentimentos, o OpLexicon), a anotação de corpus também tem um papel relevante porque, assim como no caso da anotação de entidades, léxicos podem ser insuficientes, por um lado, e levar a ambiguidades, por outro. Nosso pequeno corpus de música usado para ilustrar a anotação de entidades exemplifica o cuidado que precisamos ter na utilização ingênua de léxicos: o texto é sobre o *choro*, um gênero musical, e *chorões* é o nome dado a quem toca choro. Qualquer léxico de polaridade atribui valor negativo a *choro* ou *chorões*, o que levaria o texto a ser classificado como negativo.

A anotação de opiniões e/ou sentimentos pode assumir diferentes formatos: a anotação de polaridades a qualquer palavra, sintagma ou frase (com a atribuição de valores como positivo, negativo e neutro às palavras); a inclusão de escalas para cada uma dessas categorias (muito positivo, pouco positivo etc); a atribuição de polaridade apenas a palavras indicativas de algum tipo de emoção ou sentimento (*amor* como positivo; *ódio* como negativo; *surpresa* conforme o contexto); ou ainda a detecção de palavras indicativas de emoção ou sentimento conforme uma dada teoria.

Desafios linguísticos

Quando a anotação não parte de um léxico prévio que já contém as palavras de interesse, um desafio é a detecção do elemento que carrega a polaridade e a opinião. Abaixo, em destaque, alguns trechos selecionados para ilustrar a questão. Quais elementos devem ser anotados como índices de opinião – a frase inteira ou apenas algumas palavras?

1. Mas em um contexto todo, *o livro conseguiu suprir minhas expectativas*.
2. *impossível abandonar o livro pela metade*
3. *a tradução deveria ser CRAPúsculo*.

No exemplo (1), devemos anotar *suprir minhas expectativas*, apenas *suprir*, ou todo o trecho em itálico? Em (2), é difícil apontar um trecho específico que carregue a opinião, que é positiva sobre o livro, apesar de, isoladamente, as palavras *impossível* e *abandonar* serem consideradas negativas (e ambas estão codificadas como negativas no OpLexicon e no Senti-Lex, outro léxico de polaridades para análise de sentimento). Por fim, em (3), é apenas o reconhecimento de que *crap* é uma palavra da língua inglesa cujo significado é *lixo* que permite compreender o trocadilho *crapúsculo/crepúsculo*, que contém uma opinião negativa. Mas será que apenas a identificação de *crapúsculo* é suficiente, ou a opinião está no segmento completo?

Outro ponto importante é a necessidade de levar em conta o contexto na atribuição de polaridades, e por isso léxicos podem ser insuficientes. Palavras e expressões convencionalmente consideradas negativas podem integrar um comentário positivo, e vice-versa, como vimos com *impossível* e *abandonar*. Abaixo, mais alguns exemplos de frases com palavras de polaridade convencionalmente negativa, mas que são usadas para uma avaliação positiva (o objeto avaliado era um livro):

- a. *Lamentável* tê-lo lido somente agora.
- b. Fui à nocaute, sem direito a (re)contagem. *Chorei*. *Fiquei meio deprê*.
- c. *Tenso*. É o único livro que me deixou *nervoso* e *apreensivo* enquanto lia.
- d. *Dolorido*, *pavoroso*, *nojento*, *repugnante* e *nauseante*. Por todos esses adjetivos que o livro nos causa, ele consegue ser bom.

Corpora disponíveis

Para o português, temos atualmente uma série de corpora anotados tendo em vista a tarefa de análise de sentimento, sendo talvez pioneiros o Senti-Corpus e o ReLi. O Senti-Corpus foi construído a partir da anotação de comentários (posts) em matérias sobre eleições em Portugal, no domínio da política e em uma escrita típica da internet. Já o ReLi contém resenhas de livros, também publicadas na internet.

6.10 Relações discursivas e retóricas

Também é possível relacionar estruturas retóricas e discursivas em textos por meio da anotação. Neste caso, as relações podem se estabelecer entre sintagmas, frases e parágrafos. Na anotação de relações discursivas, elementos como *mas*, *e*, *portanto*, *por isso*, *por outro lado*, etc os responsáveis pelas conexões.

Uma maneira de indicar relações discursivas em um texto é utilizando a teoria RST (Rhetorical Structure Theory), proposta por Mann e Thompson (1988), que distribui e classifica relações discursivas de acordo com um inventário próprio de relações. Próxima à teoria RST está a teoria CST (*Cross-document Structure Theory*), que envolve o estabelecimento de relações entre documentos, o que é bastante relevante na área de sumarização automática.

Corpora disponíveis

Para a língua portuguesa, temos pelo menos dois corpora padrão-ouro com relações discursivas: o já referido corpus Summ-it, que contém relações RST, e o corpus CST-News, que contém relações CST.

7. Anotação: Bastidores

Podemos abordar o *como anotar* de duas maneiras diferentes, mas que não se excluem. De um ponto de vista da arquitetura, precisamos planejar de onde vêm as categorias de anotação e como são organizadas, isto é, planejar e descrever o *esquema de anotação*. De um ponto de execução, precisaremos definir como a anotação será feita.

7.1 Planejamento e esquema de anotação

O conjunto de etiquetas de anotação é um *tagset*, e subjacente a um *tagset* há um *esquema de anotação*, isto é, uma forma de descrever e organizar as classes (etiquetas) que o compõem.

Enquanto tarefa do PLN, a anotação está atrelada a alguma aplicação ou tarefa. Mas, porque o PLN trata de problemas de linguagem, a anotação no PLN frequentemente busca reproduzir os níveis tradicionais de análise de linguística: morfologia, sintaxe, semântica, pragmática. Nesses casos, o conjunto de etiquetas pode vir diretamente de uma teoria ou pode apenas se inspirar em algum modelo teórico; pode ser a simplificação de uma teoria ou a convergência de diferentes teorias que se debruçam sobre o mesmo fenômeno, como vimos na anotação de papéis semânticos. Por outro lado, as etiquetas podem ser criadas tendo em vista uma determinada tarefa/aplicação em mente, como a anotação de polaridades e a anotação de entidades.

Excetuando-se o primeiro cenário, quando todas as categorias já estão dadas de antemão pela própria teoria, e as situações em que se deseja replicar (para uma outra língua ou para um outro corpus) uma determinada anotação que já existe, nos demais contextos será necessário criar um *esquema de anotação*.

Definir um conjunto de etiquetas (um conjunto de classes) e sua utilização reflete 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?*

Um aspecto importante de um esquema de anotação é que ele favoreça a generalização. A terminologia de uma área, embora contenha termos específicos daquele domínio, não é exatamente um conjunto de entidades desse mesmo domínio: é importante que esses diferentes termos sejam agrupados em classes mais amplas, ou seja, os termos podem ser instâncias de categorias de anotação. Por exemplo, “conglomerado”, “lamito” e “arenito” são termos da Geologia. Na anotação da frase (a) precisaremos de uma classe ampla que os agrupe e generalize – por exemplo, uma classe “rocha”. Do mesmo modo, se precisarmos anotar textos com entidades do domínio da música, iremos anotar, na frase (b), “choro” e “violão” como “gênero musical” e “instrumento”, respectivamente.

- a. A Formação Abaeté é constituída por conglomerados, arenitos conglomeráticos, arenitos e lamitos.
- b. No estilo choro, o violão caracteriza-se por frases de contraponto geralmente em escala descendente, utilizando-se somente as cordas graves.

Mudando um pouco a perspectiva, esta generalização é o que fazemos quando, nas mesmas frases (a) e (b), anotamos “conglomerados”, “arenitos”, “lamitos”, “estilo”, “choro”, “frases”, “contraponto” etc como “substantivos”.

Outro aspecto associado à generalização é a *consistência* da anotação. A anotação é uma atividade de interpretação e *consistência*, nesse contexto, significa garantir que fenômenos do mesmo tipo sejam interpretados (anotados) da mesma maneira ao longo de

um projeto de anotação. A consistência permite que algoritmos de AM generalizem corretamente a partir dos dados e que as avaliações sejam confiáveis. Em outras palavras: se para aprender são necessários exemplos, mas os exemplos estão inconsistentes – às vezes Brasil em “morar no Brasil” está anotado como LOCAL, às vezes está anotado como ORGANIZAÇÃO – não será possível uma boa generalização, pois os dados não permitirão. Do ponto de vista dos estudos linguísticos, não é diferente: seja para estudar certos fenômenos linguísticos, seja simplesmente para encontrá-los no corpus (e criar exercícios, por exemplo), desejamos que fenômenos semelhantes sejam analisados da mesma maneira.

Certamente o grau de complexidade na definição de um esquema de anotação é desigual. Etiquetas referentes às classes de palavras são relativamente consensuais. Na anotação de entidades mencionadas, frequentemente as classes são definidas em função dos interesses do domínio.

No desenvolvimento de um esquema de anotação, uma alternativa é usar recursos pré-existentes: uma ontologia de uma determinada área pode se transformar em um esquema de anotação para a extração de informação dessa área. Uma ontologia de Geologia pode informar as classes relevantes do domínio (por exemplo, “rocha”), e uma ontologia da Música pode informar as classes relevantes do domínio (por exemplo, “gênero” e “instrumento”), e isto guiará o esquema de anotação.

Independente da sua origem (se a própria tarefa, ou se recursos externos como uma ontologia), 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 seja 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. E para que este não seja um processo infinito de reformulações, é importante que o problema que motiva a anotação esteja bem definido.

O processo de anotação com refinamento do esquema de anotação segue as seguintes etapas:

- a. Levantamento bibliográfico sobre o que já existe relacionado à questão, em termos teóricos (descrição linguística, por exemplo) e aplicados (existem anotações do mesmo tipo, ou diretamente relacionadas? Quais os problemas enfrentados?)
- b. Elaboração de um esquema ou modelo de anotação, que contém as primeiras generalizações acerca do fenômeno observado, isto é, a primeira proposta de etiquetas (categorias);
- c. Aplicação dessas etiquetas a uma amostra mais ampla;
- d. Refinamento progressivo do esquema de anotação;
- e. Observação dos casos em que as generalizações não se aplicam (com a ressalva de que as irregularidades devem estar igualmente marcadas no corpus).

Com relação às etiquetas:

- f. Criar uma etiqueta do tipo MISCELÂNEA, ou OUTROS, é útil para os casos não previstos. Esta etiqueta pode incluir casos que serão especificados no futuro;

- g. Admitir a indeterminação, isto é, que duas ou mais etiquetas estejam igualmente adequadas, no mesmo contexto. A possibilidade de anotação múltipla permitiria que as etiquetas AVISO, AMEAÇA e CONSELHO sejam atribuídas à frase *Se eu fosse você, deixaria a cidade imediatamente*, considerando uma anotação de Atos de Fala, por exemplo.

7.2 Execução

Do ponto de vista da execução, a anotação pode ser feita de diferentes maneiras, que variam conforme o volume de trabalho humano envolvido: totalmente manual, semiautomática ou totalmente automática (e, nesse caso, temos as *ferramentas* de anotação).

A anotação é totalmente manual quando estamos desenvolvendo um esquema ou projeto de anotação, ou quando não há anotação automática razoavelmente confiável. Por envolver um trabalho mais moroso, costuma ser usada em corpora de dimensões modestas, que pode servir como material de avaliação de sistemas de anotação automática.

A anotação semiautomática é a mais frequente quando se pensa na constituição de um corpus *padrão ouro*. Nesse cenário, o que se costuma fazer é a anotação automática como primeiro passo, seguida da revisão humana.

Cada procedimento (manual ou semiautomático) tem vantagens e desvantagens. Quando a anotação é feita de maneira semiautomática, o tempo gasto na preparação do material é consideravelmente menor, mas ainda assim trata-se de um processo custoso. Por outro lado, a anotação manual de um corpus a partir do zero é capaz de levantar questões que poderiam não aparecer na revisão, já que se a análise fornecida pela máquina estiver correta a tendência será acatá-la sem hesitar. Consequentemente, casos de ambiguidade ou de vagueza, em que análises alternativas são igualmente possíveis, dificilmente serão percebidas/anotadas.

Por fim, a anotação pode ser feita de maneira completamente automática. No entanto, para estimar o quão boa é uma ferramenta de anotação, a maneira mais comum é comparar o resultado da máquina com o desempenho humano na mesma tarefa, e voltamos à necessidade de anotação humana.

Vejamos agora um exemplo de um projeto de anotação:

Anotação relativa ao léxico do corpo humano

O objetivo da anotação: (a) Descobrir quais sentidos (além do corpo humano) estão associados às palavras do corpo humano; (b) Investigar como descrevemos a aparência física em português.

A que tarefas ela serve: Este projeto de anotação interessa mais diretamente aos estudos linguísticos, e indiretamente à tarefa de análise de opinião e sentimento.

Classes de anotação: após a observação de algumas ocorrências em corpus, foi definida a seguinte estratégia:

- Criar a primeira grande classificação relevante para as motivações do projeto: palavras do corpo humano que se referem ao corpo humano vs. palavras e expressões do corpo humano que se distribuem por outros campos semânticos (uma divisão fácil de fazer);
- Criar subclasses que organizem as palavras e expressões do corpo humano por outros campos semânticos a partir da análise das ocorrências.

O processo de anotação foi iniciado com cinco classes, criadas após a análise preliminar do corpus: corpo, opinião, sentimento, lugar e outros.

Corpo: machuquei o *pé*
 Opinião: tem sempre um *orelhudo* na conversa
 Sentimento: ficar com o *coração* apertado
 Lugar: A casa fica ao *pé* da montanha
 Outros: a banda tem uma *veia* pop

À medida que a anotação e revisão avançaram, foram sendo criadas novas classes. Uma nova classe só era criada após uma discussão – e consenso – entre todos os envolvidos na anotação. A cada alteração havia revisões retroativas, a fim de garantir uniformidade na anotação. Para decidir se deveríamos criar uma classe nova, usamos dois critérios:

- a classe deveria englobar diferentes palavras do corpo que compartilhassem o mesmo tipo de sentido; ou
- a classe poderia conter poucas palavras do corpo, mas seriam palavras com um uso muito frequente e sistemático.

Além dos critérios, havia duas preocupações:

- evitar uma classificação muito granular, com classes pouco ocorrências;
- não inchar a classe “outros” com usos sistemáticos.

Como resultado, chegamos às seguintes classes:

CLASSES	EXEMPLOS
CORPO	torceu o <i>pé</i> na corrida; ter <i>olhos</i> azuis
CORPO:ANIMAL	<i>orelha</i> de porco;
CORPO:CENTRALIDADE	Seu departamento é o <i>cérebro</i> da operação; Sem revelar o <i>coração</i> do plano, Itamar rebatizou o conjunto de medidas(...)
CORPO:DOENÇA	não tenho medo do <i>pé</i> de atleta; esta medonha epidemia de <i>bexiga</i>
CORPO:FACULDADE	uma provocação plástica para <i>olhos</i> e <i>ouvidos</i> livres
CORPO:GRUPO	<i>corpo</i> docente; <i>coluna</i> do exército
CORPO:LUGAR	no <i>coração</i> da floresta amazônica
CORPO:MEDIDA	dois <i>dedos</i> de pinga; onda de 3 <i>pés</i>
CORPO:MOVIMENTO	ir a <i>pé</i> ; assim que pôs os <i>pés</i> na cidade
CORPO:OPINIAO	ele é um <i>bundão</i> ; tem sempre um <i>orelhudo</i> na conversa..
CORPO:PARTE	<i>boca</i> do fogão; <i>braço</i> da máfia
CORPO:POSICAO	suplicou de <i>joelhos</i> ; dormiu em <i>pé</i>
CORPO:SENTIMENTO	com o <i>coração</i> apertado; o meu <i>sangue</i> ferve por vocês
CORPO:VEGETAL	<i>dente</i> de alho; <i>pé</i> de laranja
CORPO:OUTROS	<i>boca</i> da noite; uma <i>veia</i> pop

7.3 Anotação e neutralidade

Qualquer que seja a motivação para a utilização de um determinado tagset, é importante lembrar que as classes que ele contém não “emergem” naturalmente, mas buscam responder a alguma pergunta inicial. Por isso, quanto mais clara está a pergunta, mais sabemos se as classes estão ou não adequadas. Por isso, também, não existe anotação neutra ou atórica. No caso específico de modelos de anotação que lidam com níveis de análise mais populares, como morfologia e sintaxe, o senso comum (ou atórico, ou levemente teórico) corresponde a classes e divisões cristalizadas na gramática tradicional, como se esta fosse um bloco homogêneo, e como se prescindisse de uma visão de língua.

Do mesmo modo, ferramentas não serão neutras. Uma ferramenta que mede a complexidade textual, e usa a quantidade de verbos por frase como uma das medidas de complexidade, está subordinada à maneira contar verbos aprendida pela máquina, que está codificada no corpus. Como lidar com as formas participais, como vimos em 6.1? Como lidar com locuções verbais? A gramática de Cunha & Cintra, por exemplo, considera que em *Paulo quer vender sua moto*, não há locução verbal, e cada verbo pertenceria a uma oração distinta; já as gramáticas de Rocha Lima e Evanildo Bechara consideram *quer vender* um caso de locução verbal, e *querer* é um verbo auxiliar. Quantas orações, afinal, temos em *Paulo quer vender sua moto*?

Nem *tagsets*, nem classes linguísticas “convencionais”, são neutras, mas refletem o peso da tradição linguística. Os rótulos das classes de palavras são uma maneira, dentre muitas outras, de distribuir as palavras de uma língua em grupos e, como em qualquer classificação, têm seus critérios de definição escolhidos de acordo com os interesses subjacentes à classificação. Ao longo do tempo, diferentes escolas de pensamento se interessaram pelas palavras e sua classificação, enfatizando diferentes aspectos. Classes de palavras, ou qualquer outra classificação linguística, representam tentativas artificiais e arbitrárias de categorizar e sistematizar fenômenos linguísticos, elaborados por seres humanos em um contexto histórico e social particular. Olhar para as classificações dessa perspectiva abre as portas para pesquisas sobre tagsets (de diferentes naturezas) no âmbito dos estudos linguísticos e do PLN.

No PLN, é a anotação que guia o desempenho em uma tarefa e dá pistas sobre o que se pode aprender. Por isso, não é exagero dizer que, no contexto da anotação, os linguistas (computacionais) são os designers.

7.4 Documentação

Associada a qualquer esquema de anotação deve existir uma documentação linguística do processo. Quando se trata de interpretação, não existem soluções únicas, e por isso é fundamental o registro de qual foi a solução adotada em um dado momento para o tratamento de um fenômeno. É esta *documentação* que garante a consistência na anotação, sobretudo quando temos várias pessoas anotando. De um ponto de vista estritamente linguístico, a documentação é uma *descrição* do fenômeno anotado: captura as regularidades, as exceções e os casos difíceis, pouco claros, para os quais talvez seja necessário propor uma análise arbitrária.

Sintetizo abaixo o processo de documentação do projeto Floresta Sintá(c)tica (Afonso et al., 2002), mas que são aplicáveis a qualquer projeto de anotação:

Fase 1: Identificação de casos problemáticos por cada anotador. O caso problema pode incluir tópicos não descritos nas gramáticas tradicionais, por exemplo.

Fase 2: Discussão em comum com a equipe linguística, com o objetivo de encontrar boas análises para cada um dos casos problemáticos. Esta decisão deve ser sistematizada/generalizada, ou seja, deve poder ser aplicada a vários casos do mesmo tipo.

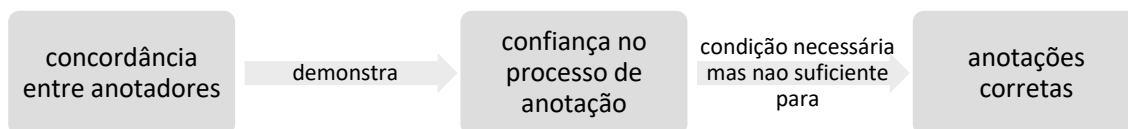
Fase 3: Confirmação de que a solução é generalizável.

Fase 4: Escrita da documentação propriamente, isto é, documentar o caso problema, a solução adotada e os exemplos em que se aplica.

7.5 A concordância entre anotadores (*quem julga o juiz?*)

A anotação automática sempre será avaliada por contraste ao desempenho humano na tarefa. E quem avalia o desempenho humano? Além disso, lembrando que a anotação é um processo interpretativo, como avaliar *interpretações*?

No PLN, temos duas maneiras de aferir a qualidade das interpretações: comparar entre si as anotações produzidas pelos anotadores (humanos), ou comparar essas anotações e um gabarito (e este gabarito foi produzido por alguém...). Como nem sempre existe um gabarito à disposição – ou porque aquilo que queremos é testar a adequação de um esquema de anotação ou porque o que pretendemos é justamente a construção desse gabarito – a comparação entre análises de diferentes anotadores acaba sendo a solução adotada. Assim, a ideia de uma anotação *correta* (supostamente fornecida pelo gabarito) é substituída pela ideia de uma anotação *consistente* (todos os anotadores analisaram os fenômenos da mesma maneira). O raciocínio subjacente é o seguinte: se diferentes pessoas, seguindo as mesmas instruções (esquema e documentação de anotação), analisaram algo da mesma maneira, esta maneira de análise é confiável. O nome que se dá a esse procedimento de avaliação é *concordância entre anotadores* (em inglês IAA – *inter annotator agreement*).



Levando em conta o trabalho envolvido na anotação, é comum que a verificação da concordância seja feita utilizando uma amostra do corpus. O resultado da comparação de todas as anotações, que costuma ser feita utilizando programas específicos, é um número que nos diz o grau de confiança, isto é, que nos diz o quanto as análises 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 valida um esquema de anotação, isto é, valida uma maneira de analisar e classificar os dados linguísticos, e 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). Quanto maior a consistência, maior o grau de confiança, porque a análise é passível de reprodução.

7.5.1 Concordância baixa, e agora?

Uma resposta que gostaria de evitar, ao menos inicialmente, é que “a concordância está baixa porque a tarefa é complexa”. Complexa para humanos? Em que sentido? Complexa porque cada anotador entende o que é para fazer de uma maneira diferente? Complexa porque, mesmo entendendo da mesma maneira o que deve ser feito, as interpretações divergem? Complexa porque demanda um conhecimento muito especializado, então é importante que anotadores estejam bastante familiarizados?

O argumento da “complexidade da tarefa” deve ser acionado depois que as seguintes alternativas tenham sido exploradas:

- Melhoria das instruções de anotação com a inclusão de exemplos, tanto positivos (“faça assim nesses casos”) quanto negativos (“não faça assim nesses casos”);
- Aumento do tempo de familiarização, para que os anotadores estejam seguros 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.

Em resumo, um projeto cuidadoso de anotação deverá levar em conta:

- 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).

7.6 Por que anotar?

São, pelo menos, quatro os motivos que justificam o investimento na anotação humana, três deles diretamente vinculados ao PLN.

Avaliação: Como já indicado, a principal motivação para a anotação humana é fornecer o resultado da análise humana para um problema que deverá ser resolvido pela máquina. Um corpus padrão ouro com erros fornecerá dados imprecisos sobre o seu desempenho. E, independente da tecnologia do momento, o desempenho humano na tarefa sempre será valioso, pois é ele que desejamos reproduzir. O corpus MacMorpho, cuja primeira versão é de 2003, a coleção dourada do HAREM, tornada pública em 2008, e o corpus Bosque, cuja primeira versão data de 2002, são exemplos de recursos muito usados até hoje. E um bom corpus anotado (consistente e bem documentado), se hoje é pouco usado porque leva a um baixo desempenho da máquina – porque codifica uma tarefa complexa – poderá ser utilizado amanhã, com o desenvolvimento de novas técnicas.

Treino: a segunda grande motivação é oferecer exemplos consistentes e variados para sistemas baseados em aprendizado de máquina. Neste caso, quantidade de material anotado precisa ser maior que a aquela preparada para a avaliação.

Estatísticas sobre a língua: sistemas baseados em regras também podem incorporar módulos estatísticos, que levarão em conta, por exemplo, a frequência que um determinado verbo tem de aparecer em um tipo de construção, resolvendo certas ambiguidades.

Pesquisa e descrição linguística: arrisco afirmar que a melhor maneira de compreender um fenômeno linguístico (sintático, semântico, pragmático) é tentando anotá-lo em um corpus, por menor que seja o corpus, mas sem deixar de lado casos difíceis, pois eles enriquecerão a pesquisa. É igualmente interessante, embora nem sempre seja possível, pedir para que outros façam o mesmo, isto é, anotem as mesmas frases que nós, para depois comparar e discutir as análises.

7.7 Anotação e Categorização

A anotação (humana ou automática) é uma tarefa de classificação e categorização: distribuir elementos em classes, de acordo com certos critérios ou instruções, incluindo-os em categorias.

Uma língua, qualquer uma, é também um sistema categorização, que organiza nossa experiência no mundo por meio de palavras, para que possamos falar dele e experimentá-lo. Dificilmente nos damos conta dessa condição, e tratar dela - da classificação - ajuda a entender os limites e o potencial da anotação. (E, ao sinalizar que uma língua, qualquer língua, organiza nossa experiência no mundo por meio de palavras, a distinção entre língua e mundo deixa de ser tão nítida.)

E como se dá, exatamente, o processo de categorização? Quais os procedimentos envolvidos? Afinal, é isso o que nós e as máquinas fazemos quando anotamos um texto: classificamos os segmentos de interesse (os alvos da anotação) como sendo do tipo x, y, ou z (as classes da anotação). Nas pessoas, o processo de categorização envolve basicamente dois procedimentos: *simplificação* e *homogeneização*.

A simplificação consiste em reduzir a diversidade e a complexidade da experiência, ou dos objetos com que estamos lidando por meio da seleção de determinados atributos, consequentemente deixando outros de lado. Quando classificamos A, B e C como “quente” (ou “bom”, ou “ruim”, ou “substantivo”, ou “folha”, ou “mulher”, ou “linguista”, ou “violência”), estamos abrindo mão das especificidades de A, B e C, e igualando-os segundo algum critério. Uma cerveja *quente*, uma sopa *quente* e um dia *quente* não têm exatamente a mesma temperatura, mas segundo certos critérios (o que é considerado uma temperatura esperada para cervejas, sopas e dias, por exemplo), as temperaturas são igualadas, e as chamamos “quente”. E mesmo duas cervejas quentes não precisam estar exatamente na mesma temperatura para serem consideradas *quentes*. Pensemos agora na diversidade daquilo que chamamos *folhas* (estou me referindo apenas ao mundo da botânica): tamanhos, formatos, cores, texturas... Apesar da diversidade, chamamos todas de *folha*. Abrimos mão de todas as diferenças e categorizamos todas como *folhas*. O mesmo vale para *mulheres*, *linguistas*, *violência*, *maçã*, *cadeira*, *amor*, *verbos*... Todas essas palavras relacionam, igualando, elementos, vivências e situações diferentes entre si, mas que tornamos iguais ao organizá-las como membros de uma mesma classe – a classe das *mulheres*, das *linguistas*, da *violência*, das *maçãs*, das *cadeiras*, do *amor*, dos *verbos*, respectivamente. Esta igualdade é construída justamente a partir da simplificação: abrimos mão de uma série de especificidades e atributos de cada um dos elementos que compõem um grupo, e os tornamos iguais por meio das palavras-rótulos *quente*, *mulher*, *linguista*, *violência*, *maçã*, *cadeira*, *amor*, *verbo*.

Quando classificamos *aluna*, *tempestade* e *tênis* como *substantivo*, estamos abrindo mão de uma série de características – *aluna* tem uma palavra semanticamente relacionada, *aluno*, mas *tênis* e *tempestade* não têm; *aluna* se refere a pessoa, *tênis* se refere a artefato, *tempestade* se refere a evento; *tênis* não forma plural com acréscimo de -s, *aluna* e *tempestade* formam... – e *construindo* uma homogeneidade.

Desta perspectiva, a categorização pode ser compreendida como um processo *ativo* de construção de homogeneidade, uma construção *interessada*, que acontece por meio do descarte de *certas* diferenças – apenas aquelas que nos interessam descartar. Esse processo não é natural: o que será descartado, por um lado, e a homogeneidade construída, por outro, são sempre decorrências do interesse de quem classifica, seja um indivíduo ou um grupo. Vimos na seção 3.2, por exemplo, a relação entre a análise linguística (gramatical) e o interesse nas proposições enquanto manifestação do conhecimento e racionalidade humanas.

Ter em mente que categorizar é igualar o que não é igual, por meio de simplificações e de uma homogeneidade construída, nos ajuda a entender quando (e por que) a anotação não acontece como planejamos. Desse ponto de vista, o que as diretivas de anotação fazem é ajudar os anotadores com a construção dessa homogeneidade. E o que fazemos, enquanto anotamos, é uma atividade completamente artificial de classificar segmentos de texto segundo *nossos interesses* (os interesses da anotação que, por sua vez, refletirão os interesses da tarefa). Nem sempre nossos interesses estão claros, e nem sempre os interesses de classificação dos mesmos objetos são convergentes – podemos comparar o interesse dos gramáticos gregos ao classificar as palavras de uma língua em partes do discurso, e o nosso interesse, no contexto do PLN. Assumir que as classificações de um servirão tal & qual para o outro é sinal de um grande otimismo.

Por isso, também, frequentemente a consistência que se deseja (na anotação e nos dados) só será construída artificialmente, por meio de diretivas muito precisas. Por outro lado, ter claro e bem delimitado o motivo da anotação, isto é, a que tarefa do PLN ela responde, é um passo importante para uma anotação bem-sucedida e, conseqüentemente, para um problema bem resolvido.

8. Análise de erros

Na seção 2.3 vimos que a avaliação de ferramentas, sistemas ou modelos de PLN costuma ser feita com a utilização de medidas como precisão e abrangência, calculadas a partir de uma comparação entre um gabarito e os resultados da análise automática. Porém, nem sempre há um corpus padrão ouro à disposição, e neste caso a única possibilidade de avaliação é a análise de uma amostra do material (ou do material completo, conforme o tamanho).

Medidas como precisão e abrangência nos dão números, não fornecem pistas qualitativas sobre o desempenho das máquinas. Por outro lado, a análise manual de uma amostra dos resultados permite ver além dos números e pode dar pistas sobre como obter resultados melhores.

A análise de erros é um trabalho que demanda conhecimento linguístico especializado: é preciso olhar os erros e *entender* o que dizem; agrupá-los de forma sistemática de maneira a contribuir para a descrição e posterior solução do problema. Voltando à seção 6.8, foi a análise linguística dos resultados da identificação de similaridade que levou à descoberta dos “artefatos linguísticos”, uma estratégia indesejável capaz de mascarar positivamente os resultados dos sistemas.

Qualquer que seja a abordagem da avaliação (contraste com corpus padrão ouro ou análise manual de uma amostra), uma análise qualitativa dos resultados, embora seja trabalhosa, é uma atividade que nos permite ver (e talvez entender) por que as coisas não estão acontecendo como o esperado.

A análise de erros segue os seguintes passos:

1. Seleção da amostra que será analisada
2. Análise de cada caso da amostra
3. Organização dos erros por tipo
4. Análise global dos resultados

A partir daí é possível detectar os pontos fracos de uma ferramenta, o que por sua vez pode guiar (i) a criação de regras para corrigi-los; (ii) a incorporação no material de treino de mais exemplos dos fenômenos que levaram a erros e (iii) a melhoria das instruções de anotação.

Vamos tomar como exemplo a identificação do apostro, uma relação bastante informativa do ponto de vista da extração de informação. Desejamos verificar o quão confiável é a anotação automática de apostos em textos literários, e para isso precisaremos analisar uma amostra dos resultados, pois não temos um gabarito. Abaixo está uma pequena amostra da saída de uma ferramenta de anotação, e o trecho analisado como apostro está em **negrito**. A partir dos resultados, o quanto podemos confiar na análise automática?

1. Uma tarde lhe foram dizer que a sua filha caçula, a bela **Marianinha**, fora vista fazendo namoros com o supradito sapateiro.
2. Manuel Pedro da Silva, mais conhecido por Manuel Pescada, era um português de uns cinqüenta anos, **forte**, vermelho e trabalhador.

3. E depois, fazendo um grande esforço para se controlar, pediu a um dos africanos presentes, o **Lima da Alfândega**, que explicasse a Nicolau o que se pretendia dele.
4. Com a morte do monarca, e em conformidade com as leis locais, sucedeu-lhe no trono um seu sobrinho, D. Pedro VI, devidamente coroado mediante a aprovação de sua majestade el-rei de Portugal, **D. Luís I**.
5. Nada lhes escapava, nem mesmo as escadas dos pedreiros, os cavalos de pau, o **banco** ou a ferramenta dos marceneiros.
6. Mas que milagre o trouxe a estas horas cá por casa, seu **compadre**?
7. Para não ficar só com a filha “que se fazia uma mulher” convidou a sogra, **D. Maria Bárbara**, a abandonar o sítio em que vivia e ir morar com ele e mais a neta.
8. Tinha-lhe birra; não podia sofrer aquele cabelo à escovinha, aquele cavanhaque sem bigode, aqueles **dentes** sujos, aquela economia torpe e aqueles movimentos de homem sem vontade própria.
9. -- Isto é um sonso, minha **afilhada**! olhe em que estado ele traz as orelhas!
10. Dentro em pouco, no agasalho carinhoso daquelas asas de mãe, **Raimundo**, de feio que era, tornou-se uma criança forte, sã e bonita.

O primeiro passo é separar as frases certas e as frases erradas. O segundo passo é organizar os erros, para buscar alguma sistematicidade que possa ser corrigida. Após a análise, temos os seguintes dados:

Frases corretamente analisadas: 1, 3, 4, 7

Frases erradas: 2, 5, 6, 8, 9, 10

Distribuição dos erros:

Coordenação: 2, 5, 8

Vocativo: 6, 9

Outros: 10

Pela análise, vemos que a análise automática erra mais do que acerta (40% de acertos e 60% de erros). Quando nos detemos nos tipos de erros, percebemos que não são aleatórios: coordenação e vocativo, elementos nominais que costumam aparecer entre vírgulas, levam a vários erros. Podemos apresentar os resultados da análise assim:

TIPO DE ERRO	FREQUÊNCIA
COORDENAÇÃO	50%
VOCATIVO	33%
OUTROS	16%
TOTAL DE ERROS DA AMOSTRA	6

Resultado da análise de erros do aposto

O que fazemos com esses resultados? Se nos interessa apenas usar a análise automática, os resultados mostram que talvez esta não seja uma boa ideia, pois há mais erros que acertos. Se desejamos avaliar a ferramenta para melhorá-la, temos duas possibilidades. Se estamos diante de uma ferramenta baseada em regras devemos ser capazes de pensar regras ou estratégias capazes de eliminar ou minimizar os erros. Já sabendo que os principais problemas são coordenação e vocativo, devemos pensar soluções direcionadas para cada um deles. Se estamos diante de uma ferramenta de aprendizado de máquina podemos enriquecer o material de treino com mais exemplos corretos de vocativo e de coordenação, e torcer para que com mais exemplos seja possível generalizar melhor.

Mesmo quando dispomos de corpus padrão outro, a análise de uma amostra dos resultados pode ser enriquecedora. Em um artigo já antigo (de 2011), com o sugestivo título de *Part-of-speech tagging from 97% to 100% - Is it time for some linguistics?* (algo como *Anotação de classes de palavras de 97% a 100% - Está na hora da linguística?*), Christopher Manning sugere a adoção de um ângulo linguístico como maneira de melhorar o desempenho de ferramentas em tarefas de PLN (especificamente, a anotação de *pos*, objeto do artigo). Ao realizar uma análise de erros da saída de um sistema, Manning detecta que mais da metade dos erros cometidos pela máquina têm origem linguística, e se distribuem da seguinte maneira:

- (a) erros que decorrem de lacunas nas diretivas de anotação, que não explicitavam o tratamento de determinados fenômenos;
- (b) erros que correspondem a uma construção linguística complexa para as máquinas, sendo necessário recorrer a um contexto de análise maior do que aquele levado em conta pela ferramenta;
- (c) erros em que não havia clareza sobre qual classe atribuir, como as formas participiais, que frequentemente poderiam ser adjetivo ou verbo;
- (d) erros que decorrem da anotação do corpus padrão ouro (erros da anotação humana, portanto).

Se excluirmos o item (b), vemos que soluções exclusivamente linguísticas dariam conta de mais da metade do total de erros. Nesse sentido, a exploração de um caminho linguístico como maneira de melhorar resultados envolveria um investimento em classes de análise (*tagsets*) bem desenhadas e definidas, e investimento na qualidade dos anotadores (humanos).

A matriz de confusão

Uma matriz de confusão (MC) é uma tabela que indica convergências e divergências entre análises. Normalmente, é usada para visualização de erros e acertos por meio de uma comparação entre a previsão feita pelo modelo e o resultado esperado, que está no gabarito.

A figura abaixo é uma MC relativa o desempenho de um anotador de POS (cujo índice de acerto foi 96%, e portanto há muito mais acertos do que erros):

	ADJ	ADP	ADV	AUX	CCONJ	DET	INTJ	NOUN	NUM	PRON	PROPN	PUNCT	SCONJ	SYM	VERB
ADJ	8869	18	61	0	0	22	0	463	12	3	86	0	2	0	287
ADP	19	32040	129	6	13	101	0	88	4	45	294	0	136	0	40
ADV	43	42	5204	1	64	86	4	39	0	11	31	0	50	0	13
AUX	3	0	0	2737	1	1	0	12	0	0	13	0	0	0	298
CCONJ	1	1	19	0	4742	0	4	3	0	1	90	0	6	0	0
DET	21	284	72	1	4	30099	1	25	109	249	154	0	13	1	10
INTJ	0	0	0	0	1	1	7	0	1	0	1	0	0	0	1
NOUN	510	98	86	6	0	7	3	39759	41	17	654	8	14	0	261
NUM	10	2	0	0	0	8	0	86	4671	3	50	10	0	1	0
PRON	5	2	11	0	2	68	0	15	1	3050	28	0	148	0	4
PROPN	86	54	16	3	9	7	1	706	34	8	20058	1	2	1	30
PUNCT	26	4	0	0	2	7	0	27	3	6	45	29513	2	1	4
SCONJ	2	29	39	0	17	16	0	4	2	81	18	0	4588	0	7
SYM	0	2	0	0	0	0	0	1	0	0	1	1	0	813	0
VERB	295	6	10	615	7	10	0	271	6	1	80	0	3	0	19228

Na MC, os números na horizontal referem-se à análise automática (considerando uma abordagem de aprendizado de máquina, os números indicam o que a máquina aprendeu, isto é, o que ela previu a partir dos dados a que foi exposta) e os números na vertical referem-se à análise humana, isto é, à anotação padrão ouro.

Começando da esquerda para a direita, e de cima para baixo, temos que em 8.869 casos, humanos analisaram uma palavra como ADJ (adjetivo) e o sistema também – temos 8.869 casos de convergência, ou seja, 8.869 acertos. Ainda na primeira coluna, descendo para a linha NOUN, temos que em 510 casos humanos analisaram a palavra como NOUN (substantivo), mas o sistema analisou como ADJ - temos 510 casos de divergência, ou de confusão, entre ADJ e NOUN. Por isso o nome matriz de *confusão*. Os valores altos, que estão na diagonal da esquerda para a direita, indicam os casos em que houve convergência. Os números que fogem dessa diagonal são as análises divergentes, as confusões. A forma usual de se analisar MC é simplesmente considerar aquilo que não está na diagonal, e que chamamos divergência, como *erro*. Afinal, são os casos em que a máquinas e pessoas divergiram, e pessoas devem saber o que fazem. No entanto, o olhar linguístico atento para estes *erros* pode perceber coisas interessantes.

Vejamos a coluna dos adjetivos. Além da confusão entre adjetivos e substantivos (já velha conhecida da Linguística) temos uma confusão entre adjetivos e verbos, que acontece em 295 casos. A frase abaixo corresponde a um desses casos, e reconhecemos imediatamente outra questão companheira de longa data dos estudos linguísticos: as formas participiais.

A corretora Antonio Delapieve e seus dirigentes foram multados por negociar títulos de renda fixa a preços **superavaliados** (ADJ – VERB)

A classe dos verbos, por sua vez, também está envolvida em várias outras confusões: verbos e adjetivos (287), verbos e verbos auxiliares (AUX) e verbos e substantivos (NOUN). O olhar linguístico treinado percebe que as confusões tendem a reproduzir temas da discussão linguística/gramatical. Alguns exemplos estão no quadro abaixo:

Confusão entre...	Exemplo de confusão
adjetivos (ADJ) e substantivos (NOUN)	O ranking de as gigantes <u>multinacionais</u> que dividem o bolo mudou.
Adjetivos (ADJ) e verbos (VERB)	O fotógrafo expôs 29 fotos <u>coloridas</u> que retratam o cotidiano da guerra civil em Angola.
Verbos (VERB) e verbos auxiliares (AUX)	Os deputados <u>começam</u> a ter os elementos comprobatórios.
Conjunções (SCONJ) e preposições (ADP)	A realização da conferência da ONU tem <u>como</u> objetivo a divulgação de políticas para o controle de natalidade.
Artigos (DET) e numerais (NUM)	E, depois de <u>uma</u> semana com gripe, o holandês recupera a sua forma.

Alguns outros casos, se não correspondem exatamente a problemas linguísticos, são de confusão compreensível:

- As formas “o” e “a” podem ser pronome ou artigo (*A liberdade sexual privada, como a concebemos, era impensável na Grécia*);
- Certas formas podem ser substantivos (*Luta no Afeganistão deixa dez mortos*) ou verbos (*Ela luta com todas as suas forças*);
- Qualquer palavra pode se transformar em um nome próprio (PROPN), o que explica a grande quantidade de confusões entre estes e classes variadas (*O esgotamento do modelo baseado em a intervenção do Estado*; *Não temos novidades sobre o estado de saúde do jogador*.)

O que podemos fazer com essas análises? Pelo ângulo linguístico-computacional, sabemos que essas confusões nada têm de aleatórias em boa parte dos casos. Elas refletem dificuldades de generalização por parte das máquinas, e é possível que essa dificuldade

tenha sido causada por inconsistência na análise humana. Deste ângulo, precisamos criar estratégias para eliminar, ou ao menos minimizar, essa dificuldade de generalização.

Embora estejamos cientes das discussões linguísticas a respeito de certos fenômenos, nem sempre estamos atentos a eles como possíveis elementos disparadores de inconsistência. Vamos tomar como exemplo a confusão entre pronomes e artigos já usada em 6.1, e que também pode ser encarada como um dos “casos problemáticos” mencionados nas etapas da documentação em 7.4.

O controle voluntário da mão direita é mais fácil do que o da mão esquerda.

Suponhamos que na construção do corpus padrão ouro a pessoa 1 tenha anotado como (1) e a pessoa 2 tenha anotado como (2), e essa inconsistência tenha contribuído para que o algoritmo se confundisse quanto a artigos e pronomes.

1) O controle voluntário da mão direita é mais fácil do que o _ART (CONTROLE) da mão esquerda

2) O controle voluntário da mão direita é mais fácil do que o _PRON (AQUELE) da mão esquerda

Para o entendimento do texto, não há diferença entre as análises. Mas a necessidade de concordância transforma essa divergência em defeito, em algo a ser eliminado. Se o esquema de anotação prevê a codificação de apenas uma única análise, será preciso decidir, no âmbito do projeto de anotação, qual será a solução adotada e, mais importante, será necessário documentar essa decisão. Temos aqui um exemplo de *homogeneidade construída*.

Para além de pistas para melhoria das instruções de anotação, uma grande vantagem da análise da MC é direcionar nosso olhar linguístico para os casos que realmente levam a erros ao invés de selecionarmos uma amostra aleatória para análise. Uma MC de relações sintáticas mostra claramente a confusão entre casos de aposto e de coordenação, que no exemplo anterior descobrimos “manualmente” após a análise da amostra.

9. Mais algumas palavras sobre Linguística e Linguística Computacional

9.1 Com o aprendizado profundo, perdemos o bonde?

Com o desenvolvimento de tecnologias cada vez mais sofisticadas, e maior poder de processamento, as máquinas hoje conseguem – ou estão em vias de conseguir, depende do tipo de desafio proposto – chegar aos resultados esperados dispensando a nossa supervisão e criando seus próprios caminhos, ainda que estes caminhos sejam, até o momento, incompreensíveis para nós, seres humanos, como vimos em 2.1.1. Assistimos a um deslocamento de esforços: o *custo humano* envolvido na anotação dá lugar ao *custo computacional*, derivado do processamento paralelo, e em camadas, de muitos dados.

Apresento quatro argumentos que me fazem acreditar que tão cedo o conhecimento linguístico dedicado ao PLN não ficará obsoleto.

O primeiro argumento é que nem todo PLN é voltado para aplicações da indústria, e há aplicações linguísticas que continuarão existindo, como descrições linguísticas para ensinar uma língua e lexicografia. Trabalhar com corpus anotado com informação linguística convencional como classes de palavras e funções sintáticas é de muita utilidade para o desenvolvimento de material didático, para a confecção de dicionários e de corretores gramaticais.

O segundo argumento já apareceu diversas vezes ao longo do livro: quando se trata de aprender algum tipo de classificação (de anotação), uma amostra do que seja o conhecimento humano esperado na tarefa será sempre necessária, pois avaliar o desempenho da máquina é uma etapa que não pode ser ignorada. No que se refere à construção de gabaritos de tarefas linguísticas, é difícil que conhecimentos linguísticos especializados não sejam acionados. A criação de datasets, benchmarks, ou corpora padrão ouro é facilitada quando lançamos mão de conhecimento linguístico “clássico”: regras. Regras para corrigir resultados da análise automática e/ou regras para acelerar o processo de anotação, regras associadas a um léxico. Algo que percebemos no PLN estatístico é um reposicionamento do papel das regras linguísticas: deixam de ser vistas como responsáveis pela estruturação da língua, mas aparecem como necessárias na preparação dos conjuntos de dados. Regras que refinam e melhoram os resultados de análise automática, como vimos no capítulo 2 e na seção 6.6.

O terceiro argumento é uma resposta ao ponto levantado por Strubell, Ganesh, e McCallum (2019), e diz respeito aos custos computacionais e financeiros de se resolver tarefas complexas como o processamento automático da linguagem humana. Os autores comparam a emissão de CO₂ em quatro atividades humanas e no treino de modelos de linguagem. É assustador saber que o treino de um único modelo de língua é tão poluente quanto a utilização de 4 carros durante uma vida inteira. Ou que a quantidade de energia utilizada pelo Google para treinar o AlphaGo, a inteligência artificial que venceu humanos no Go (ver seção 2.1) teria sido “suficiente para ferver o oceano”. Do ponto de vista financeiro, o treino de uma inteligência artificial como a GPT-3, alimentada com o

equivalente a 70 mil anos de leitura e que responde a perguntas, traduz e escreve artigos, dentre muitas outras peripécias com a linguagem, custou cerca de 4.6 milhões de dólares. Ou seja, se com o desenvolvimento das redes neurais profundas conseguimos diminuir o trabalho humano e aperfeiçoar os resultados em tarefas de PLN, não podemos esquecer que os melhores modelos custam caro para treinar e se desenvolver, e esse custo é elevado tanto de um ponto de vista financeiro quanto ambiental.

O que a anotação faz é oferecer um atalho para a generalização. A anotação organiza os dados, e com isso os estrutura, conforme certos critérios e interesses (os interesses da tarefa). Cada camada de anotação distribui (e organiza) as palavras de uma determinada maneira, sinalizando que *essas palavras são de um tipo, essas são de outro* segundo critérios definidos por nós. Nada impede que essa mesma organização seja construída/aprendida pelas máquinas sem intervenção humana, mas quando partem dos dados brutos (do texto “cru”), sistemas têm apenas a *forma* como ponto de partida – caracteres. Por isso o esforço computacional para se chegar à generalização é alto, e daí a necessidade de muitos dados e capacidade de processamento. Vimos na seção 3.1 a imensa quantidade de formas que ocorrem pouco. Com poucas ou apenas uma única ocorrência, é mais trabalhoso perceber padrões.

Voltando à pergunta que abre esta seção: Perdemos o bonde? Não, e pelo contrário. Diria que ganhamos um parque de diversões para fazer linguística e, de quebra, contribuir com o PLN.

O quarto argumento é filosófico. A Linguística – estudo científico da linguagem – e as ciências nascem e existem no interior do paradigma da *verdade*, como vimos em 3.2. Mas a IA baseada em dados tem nos mostrado um alinhamento com o que podemos chamar de paradigma da *eficácia*. É possível obter bons resultados sem saber exatamente como foram obtidos, sem entender exatamente como as máquinas funcionam e o que “compreendem” (sem saber a “verdade”): bons resultados – eficácia – são suficientes. Para a Ciência, eficácia sem compreensão, sem explicação, isto é, sem verdade, não tem valor, não é ciência.

Se deixamos a verdade de lado, ficamos sem chão, sem fundamento. Vamos para onde o vento sopra – e o vento, no nosso caso, são os dados. Por isso é tão fundamental a atenção com a curadoria e diversidade dos dados. Mas a anotação também pode ajudar a lidar com a falta de chão. A anotação codifica o conhecimento humano, é uma âncora simbólica no imenso mar de dados. A anotação não ingênua, que não se disfarça de neutra e que sabe que está a serviço de uma determinada pergunta ou tarefa é a garantia de que somos nós no comando.

9.2 Diálogos possíveis

Grande parte do que vem sendo feito hoje sob o rótulo Linguística Computacional/PLN não precisa de formação linguística como a conhecemos, mas de formação computacional e/ou matemática. Apesar disso, Linguística e PLN têm assunto para muitas conversas (e na seção *@História do PLN*, na página eletrônica, trago um panorama histórico do PLN que privilegia as diferentes maneiras de interação com a Linguística ao longo do tempo).

O nome Linguística sugere unidade, mas a Linguística pode ser várias – de base empirista ou racionalista, só para começar. Um PLN baseado em dados se aproxima mais harmoniosamente de uma Linguística que vê as línguas como objetos dinâmicos (lembramos da diversidade e dos casos não previstos, na seção 3.1); que não vê o sentido das palavras como uma propriedade estável e fixa que as acompanha, e que considera limitada a proposta de uma representação única e com limites claramente definidos (vimos os desafios de alinhar um corpus a uma wordnet, por um lado, e o sucesso de vetores de palavras contextuais, por outro). Uma Linguística que reconhece que suas classes de análise, suas taxonomias, não são conhecimentos a-históricos, mas antes produtos humanos, situados no tempo e motivados por problemas específicos – que, no caso da Linguística, nunca foram o processamento automático das línguas.

Podemos então reinterpretar a alfinetada “cada vez que demito um linguista, meus resultados melhoram”. O problema não estava (e nem poderia estar) na incorporação de conhecimento linguístico, mas no tipo de conhecimento linguístico que se considerava relevante para o PLN.

Olhando para o que vem sendo feito, acredito que linguistas e ideias linguísticas têm um espaço de grande relevância no PLN. E este espaço não é apenas o da criação de recursos lexicais e datasets (não que isto seja pouco). Explorar o papel das estruturas linguísticas como estratégia capaz de facilitar o aprendizado no PLN parece um caminho promissor, mas não temos garantias disso. E nem teremos, se nós, linguistas, não ocuparmos esse espaço, que é o mais linguístico de todos. Um exemplo muito simples: verificamos experimentalmente que quando explicitamos (desocultamos) os sujeitos ocultos nas frases, facilitamos o aprendizado de análise sintática (Freitas e de Souza, 2021). Vemos aqui uma maneira linguisticamente motivada de melhorar resultados sem a necessidade de mais dados ou capacidade de processamento.

Temos uma avenida para explorar!