Capítulo 13

Dataset e *corpus*

Cláudia Freitas

Publicado em: 26/09/2023

13.1 Introdução

A preparação de bons *datasets* (ou *corpora* anotados) para o PLN é um empreendimento que costuma envolver conhecimentos variados – de computação e linguística, no mínimo. Neste capítulo, fazemos uma apresentação de conceitos básicos e metodologias relacionados à criação de *datasets* (ou *corpora* anotados)[1]. Afinal, se queremos avançar na área, mesmo levando em conta os grandes modelos de linguagem (LLM), precisaremos de *datasets* de alta qualidade, feitos para a nossa língua e cultura.

Um *dataset*, literalmente, é um conjunto (*set*) de dados (*data*). Dados são elementos que, organizados (ou distribuídos) de uma(s) certa(s) maneira(s), isto é, tratados, produzem informação. Praticamente qualquer coisa pode ser um dado. No PLN, os dados que usamos são dados linguísticos; nossa matéria prima é a linguagem humana, e cada língua individualmente.

Os dados podem ser

- palavras, que podem ser classificadas como substantivos, advérbios, verbos etc;
- postagens em rede social, que podem ser classificadas como ofensivas ou não;
- palavras ou unidades maiores, que podem ser classificadas como pessoa, lugar etc;
- pronomes, que podem ser classificados como masculinos, femininos ou neutros;
- documentos ou frases, que podem ser classificados como simples ou complexos;
- pares de frases, que podem ser classificadas como sinônimas (mais precisamente, classificadas quanto ao seu grau de similaridade) ou não;
- segmentos de textos, que podem ser classificados e relacionados como uma pergunta e a resposta a ela associada;
- sequência de caracteres que aparecem entre espaços em branco ou espaços em branco e pontuação, que podem ser classificados como palavras;
- posição que cada palavra ocupa em uma frase ou em um texto inteiro;
- frequência de cada palavra ao longo de um texto;
- sons, que podem ser classificados como fala humana, uma tosse ou uma risada;

[1] Este capítulo contém adaptações de (Freitas, 2022), onde apresento os bastidores do processo de anotação -- planejamento e esquema de anotação, documentação e concordância entre anotadores - com mais algum detalhe, e de um ponto de vista linguístico. E, apesar de ter sido publicado primeiro, o capítulo (Caseli; Freitas; Viola, 2022) aproveitou muito do que já havia sido escrito para este capítulo.

- sons de fala, que podem ser classificados como uma palavra ("agente") ou mais de uma ("a" "gente").

Partindo dos exemplos acima, o elemento "palavra" pode virar um dado quando atribuímos a ele algum valor, como a sua classe gramatical (substantivo, verbo etc), classe semântica (pessoa, lugar etc), sua posição no texto ou a sua frequência.

No PLN, estes valores podem ser atribuídos aos dados de duas maneiras. A primeira delas é de maneira explícita – por exemplo, com cada palavra associada a uma informação do tipo PoS (classe de palavra, do inglês, *part-of-speech*), sendo essa informação do tipo *substantivo*, *verbo*, *pronome*, *advérbio* etc. Ou cada frase (ou palavra) associada a uma informação do tipo *polaridade de opinião*, sendo essa informação do tipo *positiva*, *negativa*, *neutra*. No primeiro caso podemos dizer que organizamos (ou distribuímos, ou classificamos, ou rotulamos) as palavras do texto conforme sua classe morfossintática, no segundo, podemos dizer que organizamos (ou distribuímos, ou classificamos, ou rotulamos) as frases (ou palavras) do texto conforme sua polaridade. O que há em comum em ambos os casos é a organização (ou classificação) dos dados conforme classes pré-estabelecidas que nos parecem relevantes para explorar o conteúdo linguístico, e a partir delas produzimos informação: por exemplo, se há mais opiniões positivas ou negativas (ou neutras) com relação a um determinado objeto.

Mas nem todo *dataset* com conteúdo linguístico precisa ter seus dados organizados de acordo com atributos "externos" ao texto. Grandes modelos de língua – modelos de previsão de palavras – têm como entrada imensos volumes de texto, sem informação linguística explícita associada (Capítulo 15). A informação capturada é a posição da palavra no texto e a frequência com que é usada, e isso já permite saber muito sobre as palavras, desde que tenhamos muitas delas. Mas não é deste tipo de *dataset* (que contém o que chamamos de *textos crus*) que nos ocuparemos aqui, e nem dos procedimentos que transformam posição e frequência em informação linguística, tema do Capítulo 10. Nosso foco está nos dados linguísticos que possuem alguma organização explícita humana, feita conforme classes pré-definidas por nós – *dados anotados*, que são *elementos* linguísticos que possuem classificações, anotações ou rótulos linguísticos que codificam alguma dimensão do nosso entendimento sobre as palavras, frases ou textos. Porque contêm classificações (ou análises) atribuídas aos elementos linguísticos, estes conjuntos de dados também podem ser considerados *corpora anotados*.

13.1.1 *Dataset* ou *corpus* anotado?

Um *corpus* é um conjunto de dados linguísticos. A utilização de *corpus* – palavra latina que significa *corpo* e que tem como plural a palavra *corpora* – vem de longa data nos estudos linguísticos e lexicográficos[2], e ganha força nos anos 1980 com a popularização dos computadores. A partir daí, e cada vez mais, *corpus* diz respeito a uma coleção de textos que pode ser processada por computadores. O material que compõe um *corpus* (os textos) é coletado com algum propósito (investigar ou explorar algum aspecto da linguagem, teórico ou aplicado) e foi produzido "naturalmente", isto é, não estamos diante de frases artificialmente inventadas com o objetivo de construir um *corpus*.

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

Como inicialmente o uso de *corpus* estava vinculado aos estudos linguísticos, o propósito mais comum é (ou era) o de estudar a/uma língua e suas variedades. Mas, à medida que a utilização de *corpus* vai sendo ampliada para áreas além da linguística, como no PLN, os limites (e critérios) do que é um *corpus* vão também se ampliando. A característica da *naturalidade* - enunciados naturalmente construídos em situações reais de trocas linguísticas, aspecto fundamental se o interesse é estudar uma língua – é deixada de lado quando pensamos em *corpora* criados artificialmente (automaticamente) a partir de dados "naturais". Este é o caso de materiais criados por meio de tradução automática, ou por meio do aumento artificial de dados, caso do WinoBias[3] (Zhao et al., 2018), criado para estudar viés de gênero e no qual, por meio de regras, foi criado um material balanceado quanto ao gênero com o mesmo número de entidades (pessoas) do gênero masculino e do gênero feminino. Do mesmo modo, nada impede a utilização de textos criados por grandes modelos de linguagem na elaboração de um *corpus*, ou de um *corpus* composto pelas interações linguísticas entre pessoas e máquinas.

Corpus ou *dataset*? Podemos usar três critérios para diferenciar *corpus* e *dataset*: utilidade, tipo de anotação, e tamanho.

Um *corpus*, para ser considerado um *dataset* linguístico no PLN, precisa ter algum **tamanho**. No mínimo, tamanho suficiente para permitir avaliar de maneira confiável o resultado de uma análise automática e, idealmente, tamanho suficiente para treinar um modelo de aprendizado de máquina. Nos estudos linguísticos, conforme o fenômeno pesquisado, quantidade pode não ser uma exigência. Em ambos os casos (nos estudos linguísticos e no PLN), podemos dispor de material já classificado – ou anotado, ou rotulado, ou etiquetado. Esta anotação ou classificação pode ser de diferentes naturezas e pode estar associada a diferentes segmentos de texto (palavras, expressões, frases, parágrafos ou o texto inteiro). Por exemplo:

- Indicar a classe gramatical de uma palavra;
- Indicar se duas frases são sinônimas ou não;
- Indicar se um tweet expressa discurso de ódio;
- Indicar se uma palavra se refere a uma PESSOA;
- Relacionar uma palavra a pronomes e demais outras maneiras pelas quais esta palavra é referida ao longo de um texto;
- Indicar se uma palavra, expressão, ou tweet é favorável ou desfavorável com relação a alguém ou algo;
- Indicar se uma frase é uma boa legenda para uma imagem;
- Indicar, para um trecho de áudio de um enunciado linguístico, a sua transcrição textual;
- Indicar qual sentimento uma palavra ou segmento de texto evoca.

Todos esses rótulos ou anotações têm em comum o fato de codificarem a interpretação humana, assim outro critério relevante para a distinção entre *corpus* e *dataset* é o **tipo de anotação**. A anotação é um procedimento de análise de textos – um procedimento interpretativo – ainda que esta dimensão nem sempre seja lembrada. Mesmo quando a anotação é incluída em textos de maneira totalmente automática, ela está (ou deveria estar) tentando reproduzir a análise humana. E aqui, além do tamanho, temos outra

[3] https://github.com/uclanlp/corefBias#readme

diferença entre os *corpora* anotados usados no PLN e nos estudos linguísticos: nos estudos linguísticos é possível (e comum) a utilização de *corpora* que foram anotados de maneira completamente automática, desde que a anotação automática seja de qualidade. Já quando pensamos em *datasets* para o PLN, como sua utilidade está na avaliação e treinamento de ferramentas ou modelos, as anotações são feitas por pessoas, ou feitas por máquinas e revistas por pessoas. Neste caso, *datasets* são equivalentes a *corpora padrão ouro* (*gold standard*).

É possível que a diferença quanto à nomeação – *corpus* anotado ou *dataset* – também se deva à **utilidade** em foco: quando aquilo que a Linguística chama de *corpus* anotado passa a ser usado como material para treinar modelos de linguagem, ele também pode ser visto como um *dataset*. Vejamos a apresentação do já referido WinoBias, criado para a verificação de viés de gênero na resolução de correferência (Capítulo 12). Retomaremos o WinoBias mais à frente, mas interessam aqui as duas primeiras frases do resumo do artigo que apresenta o material (grifo meu)[4]:

> Apresentamos um novo benchmark, WinoBias, para resolução de correferência com foco no viés de gênero. Nosso ***corpus*** contém sentenças no estilo do esquema de Winograd com entidades correspondentes às pessoas às quais se faz referência por meio de sua profissão .

Já no repositório do projeto onde está o WinoBias, encontramos a seguinte apresentação (grifo meu)[5]:

> Analisamos diferentes sistemas de resolução de anáfora para entender as questões de viés de gênero presentes em tais sistemas. Dando uma mesma frase como entrada para o sistema, mas apenas mudando o gênero do pronome na frase, há variação no desempenho dos sistemas. Para demonstrar a questão do viés de gênero, criamos o ***dataset*** WinoBias.

Nem todo *dataset* tem dados linguísticos, e nem todo conjunto de dados linguístico é um *corpus* anotado. Nem todo *corpus* anotado é considerado, no PLN, um *dataset* linguístico. Neste capítulo, trataremos de datatsets – *corpora* padrão ouro – que são conjuntos de dados com classificações linguísticas atribuídas por pessoas (Figura 13.1).

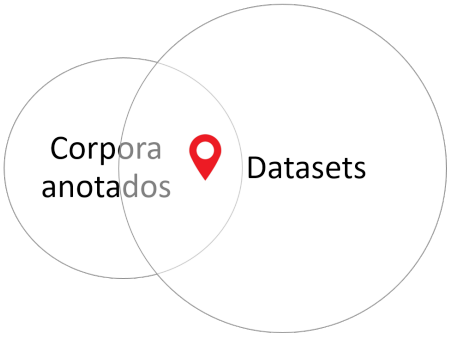
13.1.2 Anotação linguística

A anotação é uma **atividade de classificação**: temos um conjunto de classes (as etiquetas) – também chamado de *tagset* – previamente definidas e critérios que guiam a classificação. O que torna a anotação interessante – ou desafiadora – é que a classificação é feita levando em conta o contexto do enunciado linguístico. Nos exemplos do Quadro 13.1, temos anotações de classes de palavras (anotação de PoS) e anotação de polaridade, utilizada na tarefa de

[4] "Our corpus contains Winograd-schema style sentences with entities corresponding to people referred by their occupation (e.g. the nurse, the doctor, the carpenter)." https://aclanthology.org/N18-2003.pdf

[5] "We analyze different resolution systems to understand the gender bias issues lying in such systems. Providing the same sentence to the system but only changing the gender of the pronoun in the sentence, the performance of the systems varies. To demonstrate the gender bias issue, we created a WinoBias dataset." https://github.com/uclanlp/corefBias/tree/master

Figura 13.1: Nosso ponto de partida



análise de sentimentos. A anotação de polaridade, nos exemplos, está presente em dois níveis: no nível da palavra propriamente – cada palavra recebe uma etiqueta indicando se a polaridade é positiva [+], negativa [-] ou neutra [0] – e no nível da frase – cada frase recebe uma etiqueta indicando se a polaridade é positiva, negativa ou neutra com relação ao objeto sendo analisado. A anotação pode ser codificada (ou formalizada) de diferentes maneiras, e a codificação do Quadro 13.1 é apenas ilustrativa. Na Seção 13.4.4 retomaremos aspectos de codificação e formalização.

Quadro 13.1: Exemplos de frases anotadas com PoS e polaridade de sentimento

1. TRISTE$_{_SUBST_[0]}$ é$_{_V_[0]}$ uma$_{_ART_[0]}$ palavra$_{_SUBST_[0]}$ de$_{_PREP_[0]}$ 6$_{_NUM_[0]}$ letras$_{_SUBST_[0]}$ [frase neutra]
2. Um$_{_NUM_[0]}$ dos$_{_PREP+ART[0]}$ livros$_{_SUBST_[0]}$ mais$_{_ADV_[0]}$ tristes$_{_ADJ_[-]}$ que$_{_PRON-Rel_[0]}$ já$_{_ADV_[0]}$ li$_{_V_[0]}$ [frase neutra?]
3. Sofri$_{_V_[-]}$ com$_{_PREP_[0]}$ a$_{_ART_[0]}$ protagonista$_{_SUBST_[0]}$ a$_{_PREP_[0]}$ cada$_{_PRON_[0]}$ nova$_{_ADJ_[0]}$ página$_{_SUBST_[0]}$; sofri$_{_V_[-]}$ quando$_{_ADV_[0]}$ o$_{_ART_[0]}$ livro$_{_SUBST_[0]}$ acabou$_{_V_[0]}$ [frase positiva]
4. Nunca$_{_ADV_[0]}$ sofri$_{_V_[-]}$ tanto$_{_ADV_[0]}$ para$_{_PREP_[0]}$ ler$_{_V_[0]}$ um$_{_ART_[0]}$ livro$_{_SUBST_[0]}$ [frase negativa]

Como já indicado, é fundamental ter em mente que a anotação é uma **atividade interpretativa**. Como em boa parte das vezes o que está sendo anotado é o resultado de um amplo consenso, ou do "senso comum", ficamos com a impressão (errada) de que estamos diante de uma classificação objetiva. Decidir, por exemplo, se a palavra "TRISTE", na frase 1, deve ser classificada como *substantivo* ou *adjetivo* é uma decisão que irá depender da teoria adotada na anotação. Na frase 2, analisar a palavra "um" como *artigo* ou *numeral* também é fruto de uma escolha (teórica), e não a codificação de um dado objetivo. E, embora aqui o aspecto interpretativo da anotação pareça um detalhe teórico que só interessa a linguistas, veremos na Seção 13.4.5 consequências nem um pouco irrelevantes desta crença na objetividade da classificação, que aparece disfarçada de senso comum.

Quanto à **importância do contexto**, vemos pelos exemplos que ainda que as frases 1 a 3 contenham palavras de polaridade de sentimento considerada negativa ("triste" e "sofri"), no exemplo (1) a frase não tem polaridade (ou tem polaridade neutra), no exemplo (2)

é difícil decidir se estamos diante de um julgamento de valor (negativo) sobre o livro ou se diante de uma constatação, por isso a polaridade sugerida é seguida com um "?", e no exemplo (3) temos uma frase que indica uma opinião positiva sobre um livro, apesar da menção ao sofrimento.

13.2 *Datasets* pra quê?

A existência de *datasets* linguísticos, ou *corpora* padrão ouro, é fundamental para uma série de tarefas e aplicações de PLN. E por que fundamental? São três os motivos, todos igualmente importantes.

O primeiro deles é consequência da popularização, no PLN e na IA, de métodos baseados em aprendizado de máquina (AM): precisamos de exemplos do que se precisa aprender. E mesmo com os avanços da área, **bons exemplos**[6] **continuam necessários**, com a vantagem de agora os algoritmos necessitarem de uma quantidade menor deles, graças ao ajuste fino (Capítulo 15). Já se sabe que quanto mais cuidado na preparação dos dados, melhor o desempenho dos modelos – melhor a qualidade das predições, além da possibilidade de se usar menos dados[7]. Um bom *dataset*, fruto de uma anotação cuidadosa, pode ser visto como um atalho para o aprendizado, como um empurrãozinho que damos nos modelos para que atinjam logo o melhor resultado possível.

Então um motivo para investirmos na criação de *datasets* linguísticos é fornecer exemplos para que certos tipos de aprendizado possam acontecer de maneira eficiente.

O segundo motivo que torna *datasets* fundamentais em diversas tarefas de PLN é que eles **facilitam o processo de avaliação** de um sistema, ferramenta ou modelo, e de **comparação** entre eles. Isto porque se a anotação codifica, no *corpus*, a compreensão humana sobre algo, e o que queremos das máquinas em certas tarefas é que elas reproduzam esta compreensão humana, a melhor maneira de saber em que medida um resultado é bom é comparando-o com o entendimento humano. Nesse contexto, poder dispor de um *corpus* padrão ouro facilita muito as coisas. Sem ele, a alternativa para a avaliação é selecionar uma amostra do material analisado automaticamente e avaliar. Embora esta seja a única opção disponível em certos casos, não é ideal porque dificulta comparações com o desempenho de outros modelos/sistemas/ferramentas. Isto é, se cada ferramenta for avaliada de maneira "independente" a partir de uma amostra do seu resultado, será difícil uma comparação com os resultados de outras ferramentas. Outra desvantagem da avaliação por meio da análise de uma amostra, ainda que mais facilmente contornável, é que quando separamos uma amostra para fazer uma análise de erros, é mais fácil perceber aquilo que foi analisado de maneira errada (falsos positivos) do que aquilo que não foi analisado, mas deveria ter sido (falsos negativos)[8].

[6] Bons exemplos são exemplos representativos, variados e analisados de maneira consistente, como veremos na Seção 13.3.

[7] Os trabalhos (Souza; Nogueira; Lotufo, 2020) e (Pires et al., 2023) trazem dados da língua portuguesa a respeito disso, e a palestra de Rodrigo Nogueira sobre a adaptação e modelos de linguagem para o Português também: Adaptando Modelos de Linguagem para o Português: Passado, Presente e Futuro - com Rodrigo Nogueira

[8] Quando estamos tratando de *datasets* para avaliação no contexto de aprendizado de máquina, é importante que o material usado na avaliação não tenha sido utilizado nem nas etapas de treinamento e nem de validação do modelo. Esta separação do *dataset* em partes diferentes é importante para que a avaliação seja de fato um teste, e não uma avaliação em que o modelo está "roubando", uma vez que seu

O último motivo é um desdobramento dos anteriores: a partir do momento em que temos condições de treinar, avaliar e comparar resultados, temos condições de **avançar no PLN**, uma vez que o avanço na área pode ser medido pelo desempenho em tarefas. Ou seja, no PLN, um *dataset* é criado para ajudar a resolver algum problema ou tarefa[9]. Assim, para que o *dataset* exista, houve uma pergunta/problema/tarefa anterior que motivou a sua existência – alguma dimensão do PLN foi percebida como sensível e foi considerada digna de uma medição: o quanto o PLN é bom em responder perguntas, encontrar palavras de um certo tipo semântico, relacionar palavras que se referem à mesma entidade ao longo de um texto, traduzir, resumir ou simplificar um texto, dentre tantas outras.

E aqui aproximamos *datasets* e **avaliações conjuntas** (*shared tasks*). Uma avaliação conjunta (ou uma *shared task*) tem como principal objetivo incentivar a pesquisa e desenvolvimento de uma área, uma vez que fornece uma estrutura experimental comum (os mesmos conjuntos de dados e as mesmas medidas de avaliação) (Santos, 2007). Além disso, avaliações conjuntas também são maneiras de divulgar uma tarefa. Se achamos que um determinado aspecto do PLN precisa ser avaliado, ou precisa de atenção, a criação de uma avaliação conjunta que a tematize é um caminho. E, como já sinalizado, avaliações conjuntas só existem se existem *datasets* associados (e quanto melhor o *dataset*, mais bem-sucedida a avaliação).

Voltando um pouco no tempo para exemplificar, foi a criação de um *corpus* padrão ouro (chamado "Coleção Dourada") no âmbito da avaliação conjunta HAREM[10], em 2007-2008, que permitiu o avanço na tarefa de identificação e classificação de entidades mencionadas em português. Foi a criação do *corpus* ASSIN que permitiu a realização da avaliação conjunta ASSIN (Avaliação de Similaridade Semântica e Inferência Textual)[11], levando ao avanço em tarefas de similaridade semântica e inferência em português.

No entanto, em ambos os casos estamos diante de tarefas (identificação de entidades mencionadas, de similaridade semântica) que já existiam para outras línguas e que careciam de recursos – *datasets* padrão ouro são *recursos* (Capítulo 1) – para que pudessem ser abordadas também em língua portuguesa. Mas que outras tarefas poderíamos ter? Que desafios ou tarefas o PLN tem e ainda não foram abordados por falta de recursos? Por isso, além de avaliar e treinar, um *dataset* permite, ainda que indiretamente, **pautar os rumos do PLN**, o que não é pouca coisa.

É definindo tarefas que vamos abrindo caminhos no PLN – tanto caminhos previamente explorados para outras línguas, mas ainda não pavimentados para o português, quanto caminhos realmente novos, ainda não explorados em nenhuma língua. Um problema, ou tarefa, é enfrentado quando temos os meios para fazê-lo – e a construção de conjuntos de dados é um dos meios de que precisamos, considerando as abordagens atuais. Para quais tarefas – quais práticas de linguagem – queremos ajuda das máquinas?

Com a utilização cada vez maior de grandes modelos de linguagem que se alimentam de imensos volumes de dados, estratégias e abordagens para mitigar a presença de viés indesejado – como as manifestações, na linguagem, de comportamentos racistas, sexistas, xenofóbicos, dentre outros – têm preocupado pesquisadoras e pesquisadores de PLN

desempenho está sendo testado nos mesmos dados em que foi treinado.

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

[10] https://www.linguateca.pt/HAREM/.

[11] http://propor2016.di.fc.ul.pt/?page_id=381 e https://sites.google.com/view/assin2/.

(Capítulo 29) e têm sido um caminho explorado no que se refere à criação de *datasets*. O que seriam tarefas voltadas para a "monitoria de diversidade"? O já mencionado WinoBias, por exemplo, foi criado para avaliar a presença de viés de gênero. Em 2023 foi lançada uma avaliação conjunta para a detecção de homofobia/transfobia com *datasets* em inglês, espanhol, hindi, tâmil e malaiala – mas não para português, porque naquele momento não havia *datasets*[12]. Levando em conta a proliferação de conteúdo produzido automaticamente por grandes modelos de linguagem, Ignat et al. (2023) sugerem, por exemplo, o desenvolvimento de modelos capazes de identificar as partes interessadas no conteúdo gerado e seus tipos de interesse, como lucros comerciais ou interesses políticos. Novamente, para que tais modelos existam, precisaremos de *datasets*. E por este exemplo, vemos a imensa responsabilidade atribuída aos *datasets* – especificamente, aos dados, como veremos na Seção 13.2.1, e às pessoas responsáveis pela classificação dos dados, como veremos na Seção 13.4.5.

13.2.1 Sobre a importância dos dados

Quando indicamos que, para o AM, *datasets* são relevantes porque fornecem exemplos e permitem criar modelos, nem sempre nos damos conta da importância dos dados de um ponto de vista qualitativo. Isto é, pensamos nos dados como recursos de que precisamos dispor em abundância para produzir bons modelos de linguagem (Capítulo 15), mas não como fonte de "visões de mundo". Afinal, se uma língua constrói e representa as visões de mundo e crenças de seus falantes, um modelo de língua (ou de linguagem) irá, mesmo que indiretamente, codificar visões de mundo e crenças subjacentes aos dados linguísticos com que foi treinado. De forma mais direta: se, no paradigma do AM com representações distribuídas, todo o conhecimento vem dos dados, precisamos entender bem que dados são esses.

Um exemplo famoso da importância dos dados vem da pesquisadora Robyn Speer, que criou um algoritmo de Análise de Sentimento baseado em representações distribuídas (Capítulo 10), usando como dados textos publicados da internet. Ela percebeu que o algoritmo estava classificando restaurantes mexicanos de maneira negativa, o que não encontrava respaldo na pontuação dada pelos usuários para avaliar os restaurantes e nem nos próprios textos das resenhas dos restaurantes. O motivo da avaliação ruim dos restaurantes mexicanos era que a palavra *mexicano* sempre aparecia associada a *ilegal*: *imigrantes mexicanos* estavam associados a *imigrantes ilegais,* levando à classificação de restaurantes mexicanos como algo negativo. A experiência e suas lições estão relatadas no post "*How to make a racist AI without really trying*" ("Como fazer uma IA racista sem fazer muito esforço").

E, apesar das tentativas de mitigar viés indesejado utilizando pessoas comuns para classificar dados (Seção 13.4.5), ainda há muito por fazer. Um estudo de 2021, por exemplo, mostrou que textos ficcionais criados pelo modelo de linguagem GPT-3 reproduzem estereótipos de gênero: se há um personagem feminino, ele tem muitas chances de estar associado a palavras que remetem a ambiente doméstico/familiar e à aparência/corpo; se o personagem é masculino, há muitas chances de estar associado a palavras que remetem à política, guerra e tecnologia (Lucy; Bamman, 2021). E por que isso? Justamente porque é o padrão (e viés) encontrado nos dados. Em uma exploração da caracterização de

[12]Informações e *datasets* disponíveis em https://codalab.lisn.upsaclay.fr/competitions/11077.

personagens literários em obras de língua portuguesa que já estão em domínio público – obras brasileiras e portuguesas –, encontramos um padrão bastante parecido quando analisamos as caracterizações atribuídas preferencialmente a personagens masculinos e personagens femininos: personagens femininos são caracterizados sobretudo quanto à aparência, com destaque para a beleza (ou ausência dela); personagens masculinos caracterizados sobretudo quanto a traços de caráter como excelência, coragem, liberdade e sabedoria (Freitas; Santos, 2023).

Ou seja, os estereótipos produzidos pelo modelo de linguagem nada mais são do que a reprodução dos padrões que já estão nos dados. E os padrões que estão nos dados nada mais são do que comportamentos linguísticos que estão na nossa sociedade. Os dados são sempre um registro do que foi dito em algum tempo/espaço, são sempre dados do passado, e por isso, invariavelmente, modelos de previsão estão fadados a refletir o passado. Isso nos leva a uma reflexão: podemos, por meio de um passado alterado, prever um futuro que queremos? Que futuro queremos? Em outras palavras, será que *datasets* construídos artificialmente com o objetivo de eliminar vieses indesejados, ao estilo do WinoBias, que artificialmente (e indiretamente) constroem relações igualitárias de gênero, raça ou etnia, seriam capazes de produzir linguagem e, consequentemente, de criar mundos, mais igualitários? Quais as implicações éticas disso? "Quem controla o passado, controla o futuro.", dizia o slogan do estado totalitário do romance *1984*, de George Orwell.

13.3 Características de um bom *dataset* linguístico

Quando falamos de um bom *dataset* linguístico, ou de um bom *corpus* anotado, são cinco as características desejáveis: consistência, variedade, representatividade, documentação detalhada e tamanho.

- **Consistência** – por consistência, entenda-se que fenômenos semelhantes devem ser anotados (isto é, analisados) da mesma maneira. A consistência permite que algoritmos de aprendizado de máquina generalizem a partir dos dados e que as avaliações sejam confiáveis. Em outras palavras: se para aprender são necessários exemplos, mas os exemplos são inconsistentes – às vezes *Brasil* em construções do tipo *morar no Brasil* está anotado como LOCAL, às vezes está anotado como ORGANIZAÇÃO –, não será possível uma boa generalização, pois os dados ruidosos trarão dificuldade para esse processo.
- **Variedade** – na medida do possível, um bom *dataset* deve ser variado (ou balanceado) com relação aos fenômenos para os quais foi construído. Por exemplo, se desejamos um *dataset* para a tarefa de análise de sentimento/opinião com resenhas de produtos, um compilado de avaliações de páginas do tipo "Reclame Aqui" não é recomendado, pois conterá, invariavelmente, muito mais avaliações negativas do que positivas. Se desejamos um *dataset* para reconhecimento de fala (Capítulo 3), é importante que ele contemple a diversidade dos sotaques brasileiros.
- **Representatividade** – esperamos que os textos que compõem o *corpus*/*dataset* sejam representativos do tipo de texto que será alvo da aplicação, isto é, ao qual o modelo baseado nesse *dataset* será aplicado. Se a ideia é criar um modelo/ferramenta que irá procurar informações em relatórios técnicos, não é indicado que o modelo seja gerado a partir de um *dataset* que contém apenas textos jornalísticos, por exemplo.

Por outro lado, considerando o que foi dito sobre a presença de viés indesejado nos dados, a representatividade pode ser uma armadilha. Dados representativos poderão conter também estereótipos, então talvez seja mais prudente pensar na representatividade como uma característica que não é absoluta. Podemos desejar um material que seja representativo da "vida real" em vários aspectos, mas talvez não em todos[13].

- **Documentação** – um bom *dataset* é um *dataset* bem documentado, que informa a origem do conteúdo textual, as classes de anotação e as diretrizes usadas para anotar, as características e/ou formação dos anotadores. É por meio da documentação, por exemplo, que poderemos saber que o conteúdo de um *dataset* de análise de sentimento foi extraído do site "Reclame Aqui", e que talvez seja pouco indicado para aprendizado equilibrado de avaliações positivas e negativas. Como já mencionado, outro ponto cada vez mais importante diz respeito aos cuidados relativos à presença de viés indesejado nos dados. Uma vez que não é possível obter dados sem viés nenhum– línguas não são neutras, e uma língua constrói e dissemina visões de mundo e valores de uma comunidade linguística –, o material com que o PLN trabalha (enunciados reais proferidos por pessoas reais, e anotados ou revisados por pessoas reais) pode acabar incluindo e reproduzindo preconceitos. Com o objetivo de minimizar limitações científicas e éticas decorrentes de conjuntos de dados enviesados, tem sido proposto que se inclua na documentação de *datasets* informação detalhada a respeito dos falantes que produziram tais dados ou dos anotadores que analisaram e classificaram os dados em termos de *gênero*, *classe social*, *idade*, *etnia* e o que mais for possível obter (sempre respeitando questões de privacidade). Esta preocupação não é nova (Couillault et al., 2014), mas tem importância crescente no PLN (veja-se, mais recentemente, (Bender; Friedman, 2018)).
- **Tamanho** – por fim, o tamanho é um aspecto que precisa ser levado em conta. E por tamanho não me refiro apenas ao tamanho do *corpus* em *tokens* ou *palavras*, mas também à quantidade de fenômenos rotulados. Na anotação de PoS e de sintaxe, por exemplo, todas as palavras recebem alguma etiqueta; na anotação de entidades, polaridade ou correferência, por outro lado, apenas algumas palavras são classificadas, isto é, recebem uma etiqueta. Consequentemente, um *dataset* de anotação morfossintática poderá ser menor, em número de palavras, do que um de correferência ou de entidades. Ou seja, uma coisa é o número de palavras (ou de *tokens*), e outra o número de palavras (ou *tokens*) que recebem alguma etiqueta. Uma das características dos LLMs (grandes modelos de linguagem) é terem sido treinados com *datasets* gigantescos. Para conseguirem produzir material com o volume desejado e em quantidade de tempo razoável, grandes empresas fazem uso da anotação colaborativa (ou anotação *crowdsourcing*), que cada vez mais é alvo de críticas e preocupações, como veremos na Seção 13.4.5.

[13] De um ponto de vista linguístico, a representatividade é uma característica bastante discutível. Se pensamos em *corpora* "gerais", de que representatividade estamos falando? Ser representativo de algo significa ser uma parte (uma amostra) que contém as principais propriedades e características do todo que ela representa. Quando nosso objeto é uma língua, sabemos qual é o todo? Até onde vai uma língua? No mundo do PLN, a representatividade está associada sobretudo às tarefas – que por sua vez estarão associadas a um ou mais tipos de texto – , o que torna a busca pela representatividade mais factível.

Por fim, como nem sempre teremos à disposição *datasets* com todas as características desejáveis, sobretudo no que se refere à variedade e tamanho, podem ser utilizadas diferentes técnicas para aumentar artificialmente o conjunto de dados linguísticos (método chamado de *data augmentation).*

13.4 Por onde começar?

Para criar um *dataset*, precisaremos de

- um problema ou tarefa: por exemplo, classificar enunciados como ofensivos ou não; classificar palavras ou grupos de palavras como remetendo a PESSOAS etc.
- etiquetas de anotação (*tagset*), com as quais iremos classificar os dados;
- instruções sobre como classificar os dados, conhecidas como esquema ou diretrizes (*guidelines*) de anotação;
- compilação ou criação de um *corpus* adequado para esta tarefa;
- uma maneira de codificar a anotação;
- pessoas responsáveis pela anotação;
- uma ferramenta de anotação (se for o caso);
- estratégias para otimizar o processo de anotação (se for o caso);
- maneiras de avaliar o resultado da anotação, isto é de avaliar a qualidade do *dataset* produzido (embora não tenha a ver com "por onde começar?", avaliar a qualidade do material produzido é parte importante da preparação de *datasets*).

13.4.1 Definição do problema ou tarefa

O primeiro passo é ter clara a motivação para a criação do *dataset* – que problema ele se propõe a ajudar a resolver? Assim, é fundamental ter clareza sobre a motivação para a classificação dos dados: ao organizar os livros em uma estante, diferentes motivações (encontrar rapidamente os livros de que preciso ou tirar fotos de uma estante de livros para decoração) selecionam diferentes critérios para classificação (assunto, por um lado, e tamanho do dos livros, por outro), e levam a diferentes resultados – diferentes maneiras de dispor os livros na estante. Do mesmo modo, uma palavra como "Sofri" em "Sofri para terminar o livro" pode ser classificada como "negativa" se a motivação para a classificação é encontrar opiniões, e também pode ser classificada como um "verbo" se a motivação para a classificação é o comportamento morfológico. A sequência "Abri mão" em "Abri mão do prêmio" pode ser classificada como uma unidade se a motivação para a classificação é encontrar unidades de sentido (equivalente a "abdicar", por exemplo), mas considerada duas unidades se a motivação para a classificação é construir um corretor gramatical, já que será necessário verificar, na correção, se as várias formas do verbo "abrir" estão corretas ("eles abriram mão do prêmio", "abrirei mão do prêmio").

13.4.2 Conjunto de etiquetas e instruções: o esquema de anotação

A tarefa/problema bem definidos também são cruciais para o desenvolvimento das classes de anotação (as etiquetas, e também chamamos de *tagset* o conjunto de etiquetas), que podem ser poucas como "positivo" "negativo" ou "neutro", no caso de anotação de polaridades, ou mais de dez, como na anotação morfossintática.

Definir um conjunto de etiquetas e sua utilização refletem uma maneira de ver a tarefa. Por isso, quanto mais bem definido o problema (a tarefa), mais chances de sucesso. Caso seja necessário criar um esquema de anotação, devemos logo responder às seguintes perguntas: Qual o objetivo da anotação? A que ela serve?

Além disso, um esquema de anotação deve favorecer a generalização, mas sem perder informatividade. Esta generalização é o que fazemos quando, na frase (a), anotamos *nunca* e *tanto* como advérbios; *sofri* e *ler* como verbos; quando anotamos as frases (a) e (b) como frases positivas com relação a um alvo; ou quando anotamos, na frase (c) *sarampo*, *caxumba* e *rubéola* como DOENÇAS e *tríplice viral* como VACINA. Em todos os casos, embora cada uma das palavras ou frases comentadas (e anotadas) seja uma palavra/frase diferente, dizemos que elas são de um mesmo tipo – algumas são do tipo *advérbio*, outras do tipo *verbo*; algumas do tipo *positivo*; outras do tipo *negativo*; algumas do tipo DOENÇA, outras do tipo VACINA. E, apesar de as igualarmos, mantemos informação relevante (para as tarefas). Poderíamos classificar, por exemplo, *tanto* e *livro* como palavras de um mesmo tipo, porque ambas terminam em *o*, mas esta classificação não carrega informação relevante; ou podíamos classificar *sarampo*, *caxumba*, *rubéola* e *tríplice viral* com palavras do campo da SAÚDE ou MEDICINA. Em ambos os casos, embora até estejamos favorecendo alguma generalização, propondo classes mais amplas, estamos perdendo informação.

a. Nunca$_{_ADV_[0]}$ sofri$_{_V_[-]}$ tanto$_{_ADV_[0]}$ para$_{_PREP_[0]}$ ler$_{_V_[0]}$ um$_{_ART_[0]}$ livro$_{_SUBST_[0]}$ [frase negativa]
b. Sério, sofri pra terminar o livro, o livro ainda consegue ser pior que os filmes. [frase negativa]
c. Sarampo[DOENÇA], caxumba[DOENÇA] e rubéola[DOENÇA] são combatidas pela tríplice=viral[VACINA].

Na definição de um esquema de anotação é fundamental uma rodada inicial de anotação para as primeiras observações e ajustes. Porque uma coisa é a teoria e, outra, o contato com os dados. Nessa primeira rodada de anotação, pode ser que classes que julgávamos claras não estejam tão claras assim quando as palavras estão em contexto. Ou pode ser que as classes sugeridas sejam insuficientes para dar conta do que o *corpus* apresenta, e então é necessário criar novas classes. Assim, ao longo de um processo de anotação, as etiquetas iniciais (provisórias) podem ser confirmadas, ou os dados podem levar à reformulação das categorias iniciais, e o processo de anotação recomeça.

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

- Levantamento bibliográfico sobre o que já existe relacionado à questão, em termos teóricos e aplicados: *Existem anotações do mesmo tipo, ou diretamente relacionadas? Existem datasets para tarefas similares?*
- Elaboração de um esquema de anotação que contenha as primeiras generalizações acerca do fenômeno observado, isto é, a primeira proposta de etiquetas (classes);
- Aplicação dessas etiquetas a uma amostra mais ampla;
- Refinamento progressivo do esquema de anotação.

As instruções podem ser complexas como um manual ou uma gramática (como a

documentação do *treebank* Bosque[14], e da coleção dourada do HAREM[15], que codificam morfossintaxe, e entidades genéricas, respectivamente), ou podem conter apenas alguns parágrafos de explicação, como ilustrado no Quadro 13.2, que traz (uma versão traduzida) das instruções para a anotação do *corpus* SNLI (*Stanford Natural Language Inference*), que contém 560 mil pares de frases. Nesse caso, o problema (ou tarefa) associado ao *corpus* é a identificação de certos tipos de relação semântica entre duas frases. As relações de interesse são *acarretamento*, *contradição*, *neutra*. Essas, portanto, são as etiquetas atribuídas aos pares de frases. As instruções são interessantes porque não pedem que os anotadores classifiquem pares de frases com as referidas etiquetas, pede "apenas" que produzam frases a partir de certas instruções, e voltaremos a esse *corpus* (e essas instruções) na próxima seção.

Quadro 13.2: Instruções de anotação do *corpus* SNLI (Bowman et al., 2015)

Vamos te mostrar a legenda de uma foto. Não vamos te mostrar a foto. Usando apenas a legenda e o seu conhecimento de mundo: • Escreva uma legenda alternativa que seja definitivamente uma descrição verdadeira da foto. Exemplo: para a legenda "Dois cães correndo em um campo." você pode escrever "Existem animais ao ar livre". • Escreva uma legenda alternativa que possa ser uma descrição verdadeira da foto. Exemplo: para a legenda "Dois cães correndo em um campo." você pode escrever "Alguns cachorros estão correndo para pegar um graveto." • Escreva uma legenda alternativa seja definitivamente uma descrição falsa da foto. Exemplo: para a legenda "Dois cães correndo em um campo." você poderia escrever "Os animais de estimação estão sentados em um sofá". Isso é diferente da categoria talvez correta porque é impossivel para os cães correr e sentar.

13.4.3 Escolha do *corpus*

A escolha do *corpus* também está diretamente associada à tarefa – se o interesse está na detecção de opinião, é pouco indicado um *corpus* da Wikipédia, por exemplo, a não ser que o problema seja justamente a busca por opinião em textos que supostamente não deveriam conter opinião. Além disso, outras preocupações devem estar associadas à seleção do material:

- Tem direitos autorais ou pode ser usado (e disponibilizado) livremente?
- Leva em conta ou viola a privacidade de quem escreve (Capítulo 28)?
- Como lidar com conteúdo duplicado ou repostado, comum na internet?
- Como é a qualidade do texto?
  - Produzido originalmente em formato eletrônico ou resultado de processamento por OCR?
  - Possui muitas imagens, gráficos, tabelas? Neste caso, como lidar com este material?

[14]https://www.linguateca.pt/Floresta/BibliaFlorestal/.
[15]https://www.linguateca.pt/aval_conjunta/HAREM/directivas.html.

– O texto será normalizado (tudo em minúsculas, por exemplo), ou será mantida a grafia original?

Também é possível que o material textual que compõe um *corpus* tenha sido especificamente produzido para o próprio *corpus*, como o exemplo do SNLI mostrou[16]. Outro exemplo de *dataset* em que o material textual é produzido pelas pessoas com a finalidade de criar o *dataset* é o SQuAD (*Stanford Question Answering Dataset*), criado para a tarefa de perguntas e respostas (para a língua inglesa). Na tarefa de criação do *corpus*, as pessoas receberam parágrafos de documentos da Wikipédia, e sobre este material deveriam formular 5 perguntas sobre o conteúdo, e respondê-las. Além disso, não era possível usar o recurso de copiar & colar, o que forçou as pessoas a usarem suas próprias palavras na formulação das perguntas e das respostas. Nestes casos, a etapa de seleção do *corpus* deixa de existir, e é substituída pela etapa "Crie uma maneira engenhosa de produzir certos fenômenos linguísticos em grande escala". Do mesmo modo, não há anotação ou classificação propriamente, uma vez que os enunciados criados já "nascem" organizados conforme certas classes (do tipo *pergunta* e do tipo *resposta*, ou do tipo *contradição*).

13.4.4 Codificação

Na preparação de um *dataset* é preciso decidir o formato dos dados (txt, csv, tsv, JSON, XML ou outros formatos) e o formato da anotação propriamente. Muitas vezes, o formato é determinado pela ferramenta que se usa para anotar; outras vezes, a ferramenta é escolhida em função (também) dos arquivos que suporta.

Por exemplo, um formato de anotação bastante comum (para a anotação de entidades mencionadas, mas não só) é o formato IOB (Inside-Outside-Beginning), ou BIO, em que o B significa o início (*begin*) de uma entidade, o I (*in*) a continuação dela, e o O (*out*), indica que a palavra em questão não pertence à entidade. Por exemplo, em "Ana conheceu a Serra da Mantiqueira ontem", ou em "Sarampo e rubéola são combatidas pela tríplice viral" teríamos o seguinte, usando um *token* por linha:

Já a anotação de dependências sintáticas, quando feita pela abordagem Universal Dependencies (Capítulo 4), segue o formato CoNLL-U, que por sua vez é um formato TSV (*tab-separated values*, isto é, valores separados por tabulação). São características deste formato:

- Cada frase tem um identificador;
- Na linha seguinte temos o texto da frase;
- Na linha seguinte começa a representação das palavras (*tokens*) da frase, com um *token* por linha;
- A separação entre uma frase e outra é feita por uma linha em branco;
- Cada *token* contém anotação em 10 campos separados por uma tabulação. Cada campo codifica diferentes informações morfossintáticas, e informações não preenchidas são marcadas com o caractere especial "_".

[16]Embora, no PLN, não haja qualquer problema em considerar o SNLI um *corpus*, é possível que, de um ponto de vista dos estudos linguísticos, este material seja discutível no que se refere à dimensão *naturalidade*, isto é, "enunciados naturalmente produzidos".

Figura 13.2: Anotação no formato IOB para entidades mencionadas.

```
Ana             B-PESSOA
conheceu        O
a               O
Serra           B-LOCAL
da              I-LOCAL
Mantiqueira     I-LOCAL
ontem           B-TEMPO

Sarampo         B-DOENÇA
e               O
rubéola         B-DOENÇA
são             O
combatidas      O
pela            O
tríplice        B-VACINA
viral           I-VACINA
```

A Figura 13.3 mostra uma frase no formato CoNLLL-U[17].

Figura 13.3: Anotação no formato CoNLLL-U para dependências sintáticas.

```
# sent_id = 1
# text = Ana conheceu a Serra da Mantiqueira ontem.
1       Ana     Ana     PROPN   _       Gender=Fem|Number=Sing  2       nsubj   _       _
2       conheceu        conhecer        VERB    _       Mood=Ind|Number=Sing|Person=3|Tense=Past|VerbForm=Fin   0       root    _
3       a       o       DET     _       Definite=Def|Gender=Fem|Number=Sing|PronType=Art        4       det     _       _
4       Serra   Serra   PROPN   _       Gender=Fem|Number=Sing  2       obj     _       _
5-6     da      _       _       _       _       _       _       _
5       de      de      ADP     _       _       7       case    _       _
6       a       o       DET     _       Definite=Def|Gender=Fem|Number=Sing|PronType=Art        7       det     _       _
7       Mantiqueira     Mantiqueira     PROPN   _       Number=Sing     4       nmod    _       _
8       ontem   ontem   ADV     _       _       2       advmod  _       _
9       .       .       PUNCT   _       _       2       punct   _       _
```

Também é preciso decidir qual segmento de texto será anotado. Na preparação de *datasets* para detecção de desinformação e discurso de ódio, ou para análise de sentimento, por exemplo, é possível tanto uma anotação localizada, que classifica palavras e expressões ou, por outro lado, uma anotação que classifica segmentos maiores como frases ou documentos inteiros. Na Figura 13.4, temos um exemplo adaptado do *dataset* HateXplain[18], utilizado na tarefa de detecção de discurso de ódio (e que foi utilizado para avaliar a capacidade do modelo de linguagem GPT de identificar este tipo de discurso). A apresentação do material explica que cada postagem no *dataset* é anotada de três perspectivas diferentes: (i) a classificação em 3 classes "frequentemente usadas para este tipo de anotação" ("ódio", "ofensivo" ou "normal"), (ii) a comunidade-alvo (ou seja, a comunidade que foi vítima de

[17]Mais informações sobre o formato em https://universaldependencies.org/format.html.

[18]https://huggingface.co/datasets/hatexplain#dataset-structure e https://github.com/hate-alert/HateXplain.

discurso de ódio/discurso ofensivo na postagem) e (iii) as justificativas para a classificação, ou seja, as partes do texto que justificam a decisão de anotar algo como *odioso*, *ofensivo* ou *normal*. A Figura 13.4 traz um exemplo (traduzido e adaptado) do material, que está no formato JSON. Cada campo codifica o seguinte:

- Id: identificador único de cada postagem
- Anotadores: traz as anotações produzidas por cada anotador
- Label: indica a etiqueta (ou classe) atribuída a cada postagem. Os valores possíveis ódio (0), normal (1) ou ofensivo (2)
- [anotador_id]: o identificador exclusivo atribuído a cada anotador
- [target]: O alvo da postagem (no caso, adaptamos para os "extraterrestres", dentre os quais estão os marcianos)
- justificativas: os elementos do texto (os *tokens*) selecionados como justificativa para a etiqueta atribuída pelos anotadores. Cada justificativa representa uma lista com valores 0 ou 1. Um valor de 1 significa que o *token* faz parte da justificativa selecionada pelo anotador. Para obter o *token* específico, podemos usar a mesma posição de índice em "post_tokens"
- post_tokens : A lista de *tokens* representando a postagem que foi anotada.

Como podemos ver pela Figura 13.4, diferentes porções do texto (diferentes *tokens*) foram selecionadas como justificativas para a classificação geral (ódio, ofensivo ou normal). Além disso, a mesma postagem foi considerada "discurso de ódio" para duas pessoas (anotadores 4 e 3), mas apenas "discurso ofensivo" para uma delas. E com isso passamos a uma dimensão fundamental da anotação: as pessoas que anotam.

13.4.5 Anotação: sabedoria de especialistas ou sabedoria da multidão?

A anotação é valiosa porque codifica o entendimento humano sobre alguma coisa, mas esta maneira de apresentar a anotação dá a entender que o "entendimento humano" sobre mundo é homogêneo, o que sabemos não ser verdade (e a divergência das interpretações sobre a classificação da frase dos marcianos é um pequeno exemplo). Além disso, nem todas as pessoas têm o conhecimento necessário para fazer todos os tipos de anotação.

Quando mencionamos anotadores **especialistas**, a referida especialidade pode ser de diferentes naturezas: conhecimento dos conceitos linguísticos (por exemplo, de classes de palavras, anáfora, acarretamento) e/ou do domínio (por exemplo, medicina em um *corpus* de Medicina) e/ou da tarefa em questão (anotar certos elementos usando certas ferramentas e seguindo certas instruções de anotação).

Inicialmente, boa parte das anotações utilizadas no PLN tinha como fonte ou inspiração os níveis de análise linguística ou alguma teoria linguística: morfologia, sintaxe, semântica, pragmática, papéis semânticos. E, por demandarem este tipo de conhecimento especializado, linguistas sempre foram responsáveis pela anotação. Mas isto não quer dizer que linguistas estão aptos a fazer todo e qualquer tipo de anotação, justamente porque anotações (ou boa parte delas) precisam ser feitas por especialistas, e linguistas são especialistas em linguagem. Há anotações que, idealmente, precisarão do apoio ou supervisão de outros profissionais, como em projetos de anotação e/ou de criação de *datasets* de áreas como medicina, direito, genética, geologia, dentre tantas outras.

Figura 13.4: Anotação de discurso de ódio no formato JSON.

```
{
    "id": "24198545_gab",
    "anotadores": [
      {
        "label": 0,  # ódio
        "anotador_id": 4,
        "target": ["extraterrestres"]
      },
      {
        "label": 0,  # ódio
        "anotador_id": 3,
        "target": ["extraterrestres"]
      },
      {
        "label": 2,  # ofensivo
        "anotador_id": 5,
        "target": ["extraterrestres"]
      }
    ],

    "justificativas":[
    [0,0,0,1,1,1],
    [0,0,0,1,0,0],
    [0,0,0,1,0,0]

    ],

    "post_tokens":
    ["marcianos","são","o","câncer","dessa","nação"]
  }
```

Ao lado de tarefas que demandam conhecimento especializado, há tarefas que demandam "apenas" capacidade de leitura e fluência de escrita, como atribuição de orientação semântica (positiva ou negativa) a enunciados, identificação de entidades genéricas como PESSOA, ORGANIZAÇÃO, ou a criação de frases de determinados tipos, como no *corpus* SNLI, apresentado na Seção 13.4.2. E tarefas de PLN que no fim dos anos 2000 seriam vistas com descrédito por serem complexas e precisarem de uma quantidade brutal de dados são realidade 20 anos depois. A criação dos já referidos *datasets* SNLI e SQuaD são exemplos dessa tendência, apoiada por grandes empresas que pagam (mal) para que pessoas do mundo todo produzam dados de treino para as máquinas. Trata-se da *anotação colaborativa*, ou anotação feita por "trabalhadores de multidão" ou "microtrabalhadores" (*crowdworkers*). Anotações colaborativas envolvem um grande número de anotadores e possibilitam produzir *datasets* enormes em uma quantidade de tempo relativamente baixa.

Apesar de possibilitar a geração de dados em larga escala, esta forma de anotar é alvo de muitas críticas, que vão desde a falta de comprometimento dos anotadores com a tarefa, remuneração baixa e exposição de quem trabalha a material com conteúdo repulsivo (por exemplo, classificar postagens em tarefa de detecção de discurso de ódio) (veja também

Capítulo 29) até a crítica de que o que se chama de inteligência artificial é o fruto de milhares de trabalhadores invisíveis e precarizados[19].

A falta de comprometimento com a tarefa pode levar a características indesejadas nos *datasets*, que por sua vez irão impactar tanto o que é aprendido quanto a avaliação da tarefa. O estudo de Gururangan et al. (2018), por exemplo, mostrou que os anotadores do *corpus* SNLI usaram estratégias bastante previsíveis, como o emprego de (i) advérbios de negação para criar frases com *contradição*, ou de (ii) hiperônimos (relação entre *animal* e *gato*, por exemplo) para frases com *acarretamento* (as instruções para a tarefa estão no Quadro 13.2, na Seção 13.4.2). Em consequência disso, a previsibilidade tornou a tarefa artificialmente fácil para as máquinas, levando a números de desempenho enganosamente altos.

A Amazon Mechanical Turk (AMT)[20] é a mais antiga plataforma para este tipo de anotação, e ficou mais conhecida pelo público pelas reportagens sobre quem são e as condições desumanas a que estão submetidos os "anotadores", chamados *turkeys*. Admitindo que muito do que tem sido gerado pelas IAs tem como fonte os dados produzidos por estas pessoas, não chega a ser um grande exagero pensar que elas são as representantes do nosso senso comum.

Assim, outra crítica a este tipo de trabalho é que, se na imensa maioria das vezes o trabalho envolve anotações que codificam um certo senso comum, não sabemos exatamente de quem é este senso comum. A decisão sobre classificar um determinado enunciado como discurso de "ódio", "ofensivo" ou "normal" (tarefa codificada no *dataset* HateXplain e em muitos outros do mesmo tipo), por exemplo, é feita com base em alguns exemplos fornecidos pela empresa responsável pela anotação. Mas será que, em temas como esses, o que desejamos é a codificação do "senso comum"? Como garantir que a anotação não vai justamente reforçar e amplificar, uma vez que provavelmente irá alimentar modelos de linguagem, o comportamento que deveria detectar e suprimir? Vejamos a apresentação do HateXplain, traduzida abaixo:

> Antes de iniciar a anotação, os anotadores são explicitamente avisados de que a tarefa exibe algum conteúdo de ódio ou ofensivo. Preparamos instruções que explicam claramente o objetivo da tarefa de anotação, como anotar os segmentos de texto e também incluímos uma definição para cada categoria. Fornecemos vários exemplos com classificação, comunidade-alvo e segmentação das anotações para ajudar os anotadores a entender a tarefa.

A apresentação faz parecer que a identificação de algo como "discurso de ódio" (em oposição a "discurso ofensivo", ou "normal") é trivial. A manifestação de discurso de ódio, no Brasil, é crime previsto por lei[21], mas os limites entre discurso de ódio e liberdade de

[19] A intensa divulgação e popularização de grandes modelos de linguagem tem chamado a atenção para a maneira pela qual estes *datasets* são construídos, veja-se http://www.uol.com.br/tilt/reportagens-especiais/a-vida-dura-de-quem-treina-inteligencias-artificiais/, https://www.bbc.com/portuguese/geral-49234093 e http://time.com/6247678/openai-chatgpt-kenya-workers/ .

[20] O nome inspirado no "Turco Mecânico", um robô jogador de xadrez do século 18 que os adversários enfrentavam pensando estar competindo contra uma máquina, quando na verdade havia um mestre de xadrez escondido lá dentro https://www.bbc.com/portuguese/geral-49234093.

[21] Veja-se esta página do Ministério Público Federal dedicada ao tema: https://respeiteadiferenca.mpf.mp.br/www/discurso-odio.html.

expressão são alvo de discussão inclusive no meio jurídico. É razoável confiar no senso comum, ou na sabedoria da multidão, para indicar às máquinas aquilo que especialistas estão debatendo?

E voltamos à relevância da participação de especialistas nos processos de anotação. Se queremos codificar conhecimento, precisamos **qualificar** este conhecimento. Se estamos tentando fazer as máquinas nos ajudarem a classificar um determinado conteúdo, e esta classificação tem implicações jurídicas e limites pouco definidos, é importante que especialistas (do direito e dos direitos humanos, por exemplo) tomem parte no processo. E, neste caso específico, é uma maneira de restringir apenas a especialistas o contato com conteúdo sensível. Sem contar o óbvio: quanto melhor a curadoria, isto é, quanto melhor a classificação dos dados, melhor a qualidade das previsões que serão feitas e menos dados são necessários para um bom desempenho.

Por fim, quando menciono especialistas não-linguistas não quero dizer que o conhecimento linguístico formalizado só é útil em projetos de anotação linguística explícita, isto é, em projetos que envolvem conhecimento de teorias linguísticas. Diferentes tipos de anotação podem se beneficiar se o material já contém alguma anotação linguística anterior, e o ideal é que profissionais de diferentes especialidades colaborem. A anotação linguística, mesmo a mais simples como PoS, já é uma primeira organização dos dados textuais brutos, já codifica informação. O que difere a anotação linguística – de PoS, sintaxe, semântica etc. – das demais é que, por serem genéricas e fornecerem um primeiro nível de organização nos dados, facilitam e impulsionam outros tipos de anotação, como veremos na Seção 13.5.

13.4.6 Ferramentas de anotação

É importante contar com ferramentas que auxiliem o processo de anotação e revisão. Devido à expansão da atividade de anotação para além de tarefas estritamente linguísticas, há um grande número de ferramentas, algumas gratuitas, criadas para este fim. Em geral, a escolha da ferramenta é influenciada pela natureza da anotação ou da tarefa, e pontos para se levar em conta são:

- o formato dos arquivos de entrada e de saída;
- se a ferramenta suporta receber textos com outros tipos de anotação, ou apenas o texto cru que será anotado;
- se é possível corrigir (ou editar) anotações usando regras/padrões linguísticos, ou a revisão/edição só pode ser feita caso a caso;
- se há diferentes maneiras de visualizar os textos que serão anotados;
- se há opção de atalhos usando teclado (precisar clicar para anotar, apesar de aparentemente mais fácil, é pouco produtivo para a maioria das pessoas que anotam);
- se é possível trabalhar online ou apenas localmente;
- se a ferramenta suporta diferentes anotadores trabalhando simultaneamente;
- se a ferramenta é customizável;
- o tempo de familiarização necessário para começar a usar.

Em geral, teremos sempre uma tensão entre uma ferramenta com várias funcionalidades, mas cuja utilização é mais difícil, e uma ferramenta mais fácil de usar, mas que oferece menos funcionalidades. Alguns exemplos de ferramentas são

- BRAT[22]: para anotação de entidades mencionadas, correferência, dependências sintáticas, entre outras.
- Label Studio[23]: para anotação de entidades mencionadas, perguntas e respostas, análise de sentimento, entre outras.
- Inception[24]: para diversos tipos de anotação semântica e discursiva.
- WebAnno[25]: para anotação morfológica, sintática e semântica.
- Arborator[26]: para anotar e buscar dependências sintáticas em formato UD.
- ET: dois ambientes integrados para buscar, revisar e avaliar dependências sintáticas em formato UD: Interrogatório[27], para buscar e revisar anotações e Julgamento[28], para avaliar .

13.4.7 Estratégias de anotação

Um *dataset*/*corpus* padrão ouro pode ser construído (i) de maneira totalmente manual por uma única pessoa, por grupo pequeno ou por centenas ou milhares de pessoas, ou (ii) de maneira híbrida, quando é feita uma primeira rodada de anotação automática que depois é revista por pessoas[29]. Este será o tema da Seção 13.5.

13.4.8 Formas de avaliação

Quem garante que um determinado *dataset* é bom, isto é, que os dados que contém são confiáveis, diversificados, representativos, codificados de maneira consistente e adequados à tarefa? Cada tarefa tem suas especificidades, e é avaliada de uma maneira. Este será o tema da Seção 13.6.

13.5 Procedimentos e estratégias de anotação e revisão

Depois de definido o problema ou a tarefa, escolhido o *corpus*, o esquema de anotação, a codificação, a ferramenta e as pessoas responsáveis pela anotação – e nada impede que a anotação seja feita por uma única pessoa, ainda que esta não seja a situação ideal – é hora de anotar propriamente.

Como vimos, diferentes práticas podem ser consideradas anotação:

- Ler (ou ouvir) um enunciado e atribuir uma etiqueta ao enunciado todo ou a partes dele;
- Produzir uma frase (enunciados em geral) a partir de determinadas instruções;
- Produzir perguntas e respostas a partir de um texto.

[22]https://brat.nlplab.org/
[23]https://labelstud.io/
[24]https://inception-project.github.io/
[25]https://webanno.github.io/webanno/
[26]https://arborator.icmc.usp.br
[27]https://github.com/alvelvis/Interrogat-rio
[28]https://github.com/alvelvis/Julgamento
[29]Já quando falamos de um *corpus* anotado, além dessas duas maneiras, existe a possibilidade de uma anotação 100% automática.

O foco desta seção está nas anotações do primeiro tipo, e em enunciados escritos. Quando a anotação consiste em atribuir uma etiqueta a um elemento do texto, é possível começar usando uma lista de palavras vindas de um léxico ou recursos lexicais (Capítulo 4 e (Freitas, 2022)). A língua portuguesa dispõe de alguns, para diferentes tarefas[30].

13.5.1 Palavras, regras e padrões linguísticos na construção de um *corpus* padrão ouro

Quando a anotação parte de uma lista de palavras (um léxico), a cada vez que uma palavra de interesse é encontrada, ela recebe uma etiqueta (ou mais de uma, caso o *esquema de anotação* preveja isso).

Por exemplo, o Emocionário[31] contém uma lista de palavras que, em algum contexto, descrevem algum tipo de emoção e/ou sentimento da língua portuguesa[32]. O léxico tem o formato de uma lista de lemas associados a uma classe de emoção/sentimento.

Mas, como as palavras podem assumir sentidos diferentes conforme o contexto em que estão sendo usadas, listas de palavras, sozinhas, são insuficientes. Será necessária uma **revisão** da classificação inicial das palavras (da anotação) para corrigir os erros. Léxicos são sempre ótimos pontos de partida, mas não são eficientes para uma boa anotação. Por exemplo, quando usamos o léxico do Emocionário para anotar um texto, o verbo *chocar* é anotado com a emoção SURPRESA nas duas frases abaixo, mas apenas a ocorrência 2 remete à SURPRESA, e deveria ter sido anotada.

1. Após a queda, sua moto ***chocou***-se com o muro e pegou fogo.
2. O mundo ***chocou***-se com a notícia da morte do navegador.

Por isso, após a aplicação do léxico, ficamos diante da necessidade de revisão. Tanto a revisão quanto a anotação podem se beneficiar se o *corpus* já contiver uma análise linguística prévia: quanto mais informação (anotação) disponível no *corpus*, mais otimizada pode ser a revisão. Para a língua portuguesa, podemos contar com modelos de anotação morfossintática com desempenho bastante satisfatório, e esta anotação prévia pode ajudar o trabalho de revisão semântica se estamos diante de anotadores humanos especializados em conhecimentos linguísticos/morfossintáticos. Voltando aos exemplos 1 e 2, contar com uma **anotação prévia** e com uma *ferramenta adequada* permite listar os sujeitos do verbo "chocar", facilitando a análise dos casos corretos e dos que precisam ser corrigidos. Ou seja, ao ler uma lista com palavras como "mundo", "moto", "caminhão", "país", "padre", avião", "asteróide" e "Maria" e já sabendo que estas são palavras que estão sendo usadas como *sujeito* do verbo "chocar" ("mundo se chocou", "moto se chocou", "país se chocou" etc), podemos dividir as palavras em três grupos, como no Quadro 13.3.

[30]Ver GitHub do Brasileiras em PLN: https://github.com/brasileiras-pln
[31]https://www.linguateca.pt/acesso/corpos/dicemocoes.iso.txt
[32]Para mais informações, veja https://www.linguateca.pt/Gramateca/Emocionario.html.

Quadro 13.3: Distribuição de palavras que exercem a função de *sujeito* do verbo "chocar"

Grupo 1 = surpresa	**Grupo 2 = colisão**	**Grupo 3 = dúvida**
mundo	moto	Maria
país	caminhão	
	avião	
	asteróide	

Das 7 ocorrências listadas, precisaremos ler o contexto apenas daquelas listadas no grupo 3, já que pessoas podem ficar *surpresas* (grupo1) mas também podem *colidir* com outras pessoas ou objetos (grupo 2). Após a análise do contexto e decisão sobre os casos de dúvida, precisaremos eliminar a etiqueta SURPRESA quando os sujeitos forem aqueles listados nas palavras do Grupo 2. Com uma ferramenta adequada, podemos fazer todas essas alterações de uma única vez. Será preciso escrever uma regra, verificar se ela faz o que era esperado, e aplicá-la ao *corpus*. A regra poderia ser algo do tipo:

```
token.lemma = "palavras_do_grupo2" and token.deprel == "nsubj"
and token.head_token.lemma == "chocar" >> token.emo == "0"
```

que utiliza informação morfossintática (codificada nos atributos lemma e deprel, que indica a função sintática). Ou seja, a regra diz[33]:

Se

> O lema do *token* é alguma das palavras do grupo 2 (informação do atributo *lemma*) E
>
> A função sintática do *token* é sujeito (informação do atributo *deprel*, e *nsubj* é o código para a função de *sujeito*) E
>
> O "pai" do sujeito é o lema "chocar" (informação do atributo *token.head_token.lemma*)

Então

> A anotação de emoção do *token* será zerada (informação do atributo *emo*)

E por que não usar esta forma de trabalhar, baseada em regras, para todo o PLN? Isto é, se o interesse está na anotação de emoção, por que precisamos de um *dataset* para treinar e criar um modelo, por que não podemos fazer tudo por meio de regras? Acontece que esta forma de trabalhar é bem pouco prática/eficaz para textos diferentes daqueles que originaram as regras de revisão, e também para *corpora* realmente grandes. Para o "mundo controlado"[34] do *dataset*, abordagens baseadas em léxicos e regras são uma boa estratégia, mas nem tudo poderá ser (bem) resolvido por regras, e isto se deve a uma propriedade

[33] O exemplo é baseado na utilização da ferramenta Interrogatório https://github.com/alvelvis/Interrogatorio.

[34] "Controlado" porque sabemos onde começa e onde termina, isto é, é um conjunto finito de enunciados linguísticos.

das línguas: independentemente do tipo de texto e do assunto tratado, sempre haverá um número imenso de fenômenos que não se repetem, e para os quais não teremos regras[35].

Quando observamos a distribuição das palavras em um *corpus*, sempre veremos muitas palavras com baixa ocorrência – muitas palavras que ocorrem 5 vezes, ainda mais palavras que ocorrem 4 vezes, ainda mais palavras que ocorrem 3 vezes, ainda mais palavras com ocorrência 2, e ainda mais palavras com apenas uma única ocorrência. Temos uma imensa proporção de palavras do *corpus* que são ocorrências singulares, de casos que ocorrem apenas uma vez. Este fenômeno tem um nome: *hapax legomenon* (*hapax legomena*, no plural), termo que vem do grego e que significa "sendo dito uma vez".

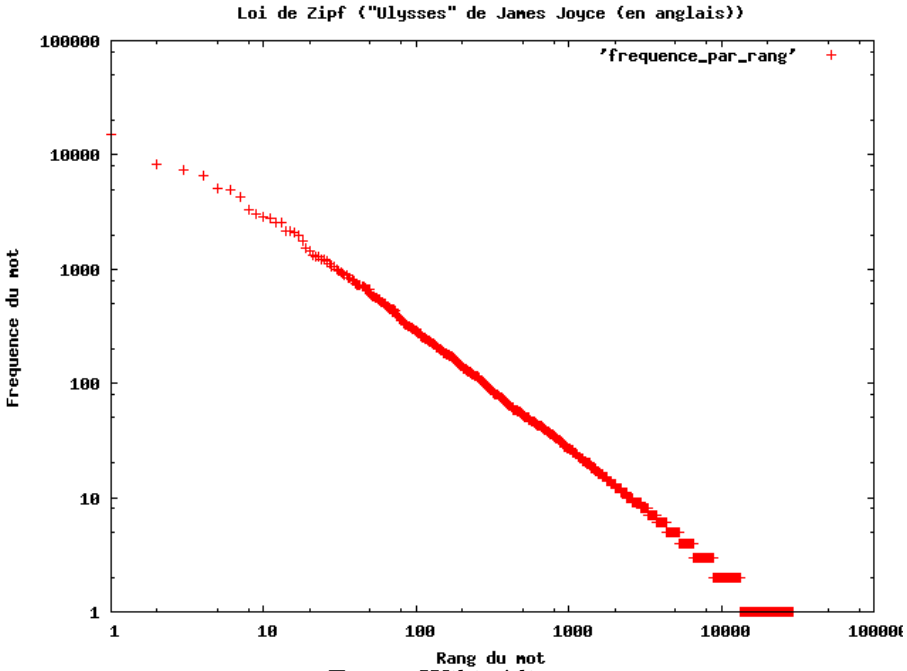
Esta distribuição segue um padrão: poucos casos com muitas ocorrências, um número intermediário de casos com frequência média, e um número enorme de casos de frequência baixa. Além disso, quanto menor a frequência, mais palavras compartilham essa mesma frequência. Por isso há mais palavras de frequência 1 do que palavras de frequência 2, mais palavras de frequência 2 do que palavras de frequência 3 etc. Em *corpora* grandes, 40% a 60% das palavras são *hapax legomena* e outros 10% a 15% são *dis legomena* (ocorrem 2 vezes). E esta distribuição não se aplica apenas a palavras isoladas, mas a qualquer fenômeno, como a distribuição das entidades em um *corpus*, ou a estrutura dos sintagmas nominais. Este tipo de fenômeno é capturado pela lei de Zipf, uma lei empírica formulada no contexto da Linguística, e assim nomeada devido ao trabalho do linguista George Kingsley Zipf (1902–1950).

A Figura 13.5 mostra a distribuição das palavras do livro Ulisses, de James Joyce, e vemos que a frequência de uma palavra é inversamente proporcional à sua posição no *ranking* de frequências (isto é, a palavra na posição 1 tem frequência muito maior que a palavra na posição 10 mil). E vemos que o padrão se repete na Figura 13.6, que mostra a distribuição das 10 milhões de palavras mais frequentes em 30 wikipédias, isto é, em 30 línguas diferentes.

Quando ficamos cientes desta propriedade distribucional, conseguimos entender melhor por que as regras linguísticas são importantes, mas até um certo ponto: as regras capturam regularidades, e regularidades só existem se existe repetição. Por outro lado, se descartamos os casos com frequência 2 ou 1 (que não se repetem, portanto), estaremos olhando para uma língua mutilada, da qual uma imensa parte foi dispensada. Como uma boa parcela da língua não se repete, é difícil, apenas com regras, conseguir generalizar fenômenos. Assim, não importa a quantidade colossal de dados que tenhamos à disposição, sempre teremos uma proporção enorme de casos raros e não previstos. Por isso, abordagens baseadas em regras serão limitadas, e por isso também o aprendizado de máquina apresenta bons resultados: desde que haja dados, algoritmos estão cada vez melhores em prever eventos raros. Apesar da ressalva, o combinado léxico & regras continua sendo uma ótima maneira de construir *datasets* e *corpora* padrão ouro, considerando a alternativa de ler cada frase isoladamente, e anotar cada fenômeno de uma vez. A elaboração do PetroNER, que contém anotação de entidades da área de petróleo, seguiu esta maneira de trabalhar. O *corpus* tem 615 mil *tokens*, quase 20 mil entidades anotadas, e partiu de um léxico inicial de quase 390 mil instâncias de entidades (palavras) fornecidas por especialistas, mas nem todas foram encontradas no *corpus*. Foram criadas quase 2 mil regras para rever e melhorar a aplicação do léxico inicial ao *corpus*, e, mais da metade delas (56,2%) foi aplicada apenas uma vez.

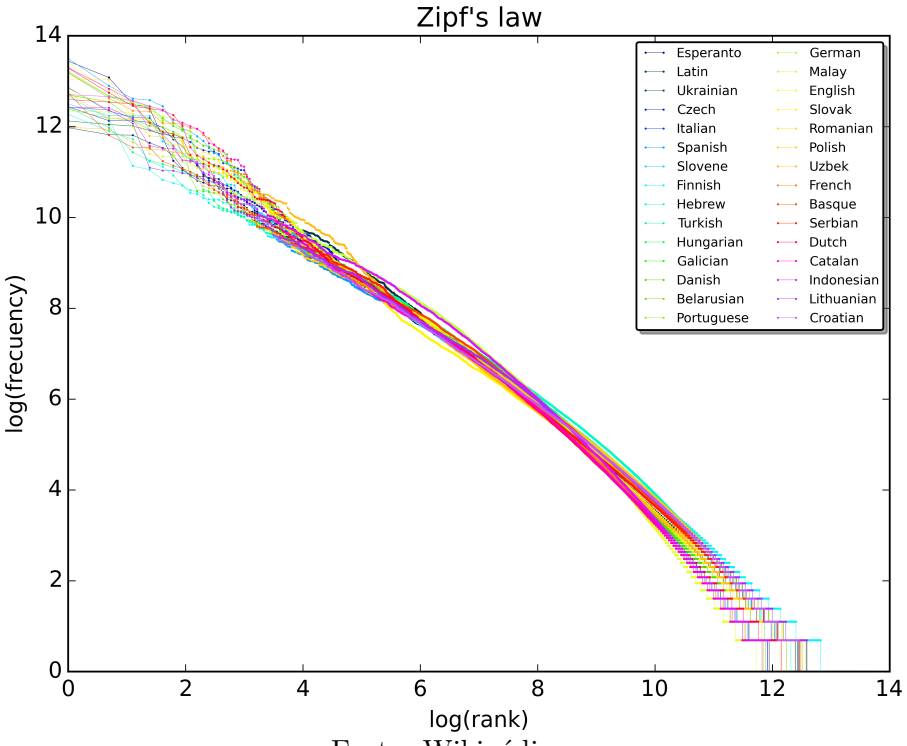
[35] E não custa lembrar que os anos iniciais do PLN foram dominados pela crença de que o processamento linguístico poderia ser inteiramente resolvido por meio de regras linguísticas.

Figura 13.5: Distribuição das palavras do livro Ulisses, de James Joyce.



Fonte: Wikipédia

Figura 13.6: Distribuição das 10 milhões de palavras mais frequentes em 30 wikipédias.



Fonte: Wikipédia

Quando reaplicamos essas regras a um novo *corpus*, com as mesmas características do PetroNER, precisamos criar mais de mil novas regras para que ele ficasse padrão ouro, e quase 55% das regras usadas no PetroNER não foram aplicadas.

13.5.2 Revisão de anotação

A revisão de uma anotação pode ser motivada pelo reconhecimento de que aquilo que se considerava padrão ouro, no fim das contas, não era tão ouro assim, ou porque estamos diante de um *corpus* novo, de um gênero ou domínio novos, o que acaba trazendo novos desafios para a anotação ou, como vimos na seção anterior, ou como parte de uma estratégia de anotação.

O que estas situações têm em comum é já contarem com uma anotação, ainda que de má qualidade. Por isso, o que precisamos neste momento é de **estratégias** para encontrar erros ou inconsistências na anotação. Por que estratégias? Um caminho óbvio é ler cada frase no *corpus*, e esse pode ser o primeiro passo se este é o primeiro contato com o *corpus* e conforme o tipo de anotação. Mas, como única abordagem para a detecção de erros, é uma estratégia limitada, por dois motivos: (i) uma mesma frase pode conter erros de diferentes naturezas, o que tem como consequência dificuldade em manter o foco e a consistência da revisão, deixando o processo de revisão mais suscetível a erros e mais demorado; (ii) conforme o tipo de anotação que está em jogo, há fenômenos que não costumam trazer erros – o reconhecimento de certas formas como artigos, verbos, advérbios, sujeitos, por exemplo – e rever cada unidade da frase envolve rever também estes casos mais fáceis. Ou seja, analisar *token* a *token* (ou palavra por palavra) também pode ser um desperdício de energia, em *certos casos.* Wallis (2003), por exemplo, recomenda trocar uma revisão linear, *token* a *token*, por uma **revisão transversal**, que permitiria ver os fenômenos em questão de forma ampliada e garantiria uma revisão consistente, como ilustramos na seção anterior com o exemplo de *chocar* e SURPRESA. Por isso – porque analisar linearmente, frase a frase, é pouco eficaz – uma etapa importante do processo de revisão consiste em encontrar uma maneira de detectar erros e corrigi-los.

O trabalho de revisão de anotação pode se tornar um imenso labirinto, e para chegar até o (padrão) ouro precisaremos de um mapa e de um guia que nos ajudem a encontrar os erros ou inconsistências: um caminho que seja, de preferência, rápido e seguro, isto é, que nos permita encontrar erros no menor tempo possível sem abrir mão da qualidade. A preocupação é dupla: do ponto de vista do resultado, buscamos análises consistentes e adequadas – uma anotação padrão ouro –; do ponto de vista do processo de revisão, buscamos eficiência – chegar aos melhores resultados consumindo pouca energia, isto é, em pouco tempo.

13.5.3 Até onde precisamos rever?

Quem já trabalhou com revisão de *corpus* sabe que o trabalho parece não ter fim – e por isso também é importante termos estratégias para conduzir e finalizar a anotação.

Do ponto de vista do aprendizado de máquina, a presença de alguns erros aleatórios não chega a prejudicar justamente porque não há o que ser aprendido, já que não será possível generalizar a partir de erros aleatórios. Mas o fato de um *dataset* ruidoso não influenciar a generalização não significa, necessariamente, que não seja necessário torná-lo o melhor

possível. Em primeiro lugar, porque nem todas as ferramentas ou tarefas precisam estar associadas ao paradigma de aprendizado automático. Em segundo lugar, e mais importante, porque se quisermos que o *dataset* produzido também sirva para avaliação, a presença de erros será sempre um problema, uma vez que pode penalizar sistemas injustamente caso a análise automática esteja correta, mas a anotação padrão ouro esteja errada.

De todo modo, no ambiente de aprendizado automático, o que temos visto é que é sim possível aprender com dados um pouco ruidosos, e que uma anotação de alta qualidade não se reflete, necessariamente, em facilidade de generalização. O trabalho de Souza; Freitas (2023) traz alguns dados para informar a discussão no que se refere à língua portuguesa.

Na preparação do *treebank* PetroGold v3[36], cada rodada de revisões foi acompanhada de uma rodada de "avaliação de aprendizagem" (uma *avaliação intrínseca*, como veremos na Seção 13.6) com o objetivo de verificar o impacto das melhorias linguísticas na aprendizagem automática. Os resultados estão na Tabela 13.1[37]:

Tabela 13.1: Evolução de um *corpus* comparando quantidade de revisões e melhoria no aprendizado de máquina.

	***Tokens* revistos**	**F1**
V1	–	88,53
V2	8.802	88,82
V3	9.314	90,22

Pela Tabela 13.1, vemos que apesar do esforço de revisão, o impacto na aprendizagem é limitado, sobretudo entre as versões 2 e 3. No fim das contas, um *corpus* maravilhosamente anotado leva a um desempenho muito semelhante àquele treinado em um *corpus* "apenas" bem anotado. No entanto, levar um melhor desempenho não é garantia absoluta de um *dataset* melhor, como veremos na próxima seção.

13.6 Como avaliar a qualidade do *dataset*?

Na Seção 13.2, vimos o que são e qual a relevância das avaliações conjuntas, que avaliam modelos e ferramentas a partir de um mesmo conjunto de dados/*dataset*. Nesta seção, o foco está em avaliar a qualidade dos *datasets*, já que um *dataset* de baixa qualidade não será capaz de dar suporte a uma avaliação confiável, seja ou não uma avaliação conjunta, e irá gerar um modelo de linguagem (ou previsões) de baixa qualidade[38].

Cada tarefa de PLN tem seus próprios métodos de avaliação. No entanto, quando tratamos de *datasets* anotados, alguns elementos da avaliação são comuns às diferentes tarefas. Uma vez que a anotação pode ser considerada uma tarefa de classificação, a avaliação também é feita nesses moldes.

Para avaliar um modelo ou ferramenta de classificação é comum utilizar as medidas de *precisão* e abrangência. A **precisão** mede (avalia) se a classificação que foi feita está correta (se a palavra analisada como *verbo* é realmente um *verbo*, ou se um comentário classificado como *ofensivo* é realmente *ofensivo*). A **abrangência** mede (avalia) se tudo o

[36]https://petroles.puc-rio.ai/files/Corpora/petrogold-v3.zip
[37]O trabalho está detalhadamente descrito em (Souza, 2023) e (Souza; Freitas, 2023).
[38]Veja mais sobre Avaliação de sistemas de PLN no Capítulo 14.

que deveria ter sido encontrado (e classificado) foi encontrado e classificado corretamente (se todos os verbos ou comentários ofensivos foram encontrados). A precisão mede a *qualidade* das classificações realizadas; a abrangência mede a qualidade da quantidade de elementos classificados, isto é, indica se tudo aquilo que deveria ter sido encontrado foi, de fato, encontrado.

Para calcular precisão, abrangência e a medida F (que é uma média harmônica entre ambas), classificamos os resultados da seguinte maneira:

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

Para calcular a precisão fazemos:

$$Precisão = \frac{VP}{VP + FP}$$

Para calcular a abrangência fazemos:

$$Abrangência = \frac{VP}{VP + FN}$$

Para calcular a medida F fazemos

$$F = 2 * \frac{Precisão * Abrangência}{Precisão + Abrangência}$$

Uma ferramenta pode ser muito precisa – todas as classificações que ela faz são corretas – e pode, igualmente, ter uma baixa abrangência – apesar de acertar bastante, há muitos casos que ficam de fora. Em geral, há uma tensão entre essas duas medidas: se afrouxamos a abrangência para encontrar mais casos, podemos diminuir a precisão, trazendo muitos casos errados. E, tentando melhorar a precisão, corremos o risco de perder em abrangência. Por isso, um bom desempenho se reflete em um equilíbrio entre essas medidas, e esta é a proposta da **medida F**: indicar em um único número uma combinação entre precisão e abrangência que reflita o desempenho geral.

Usamos medidas de F1, precisão e abrangência para avaliar modelos e ferramentas quanto à capacidade de generalizar a partir dos dados a que foram expostos no treinamento. Mas o quanto os dados permitem essa generalização? A capacidade de generalizar a partir dos dados está associada aos algoritmos utilizados, mas *datasets* também têm um papel nessa história – para o bem e para o mal –, pois algoritmos não fazem mágica.

13.6.1 Concordância entre-anotadores

No que se refere a *datasets* com dados anotados, além de bem documentado, de tamanho suficiente, variado e adequado à tarefa, um bom *dataset* é aquele no qual as anotações foram feitas de maneira consistente.

Nos *datasets* criados por meio de anotação *crowdsourcing*, algumas estratégias usadas para garantir a consistência são a revisão das análises por outros anotadores (e apenas aquelas anotações em que não foi necessária correção são consideradas) e o descarte de respostas ou análises desviantes.

Na anotação feita por equipes menores, em geral compostas por especialistas, a consistência de uma anotação é avaliada por meio da concordância entre anotadores (ou **concordância inter-anotadores**), que nada mais é do que a comparação entre duas ou mais anotações (análises) humanas. Quanto maior o índice de concordância, isto é, quanto mais convergência entre as análises, mais confiáveis elas são. Na concordância entre anotadores, a ideia de uma anotação *correta* é substituída pela ideia de uma anotação *consistente* (todos os anotadores analisaram os fenômenos da mesma maneira). O raciocínio subjacente é este: se diferentes pessoas, seguindo as mesmas instruções (esquema e documentação de anotação), analisaram um fenômeno da mesma maneira, esta análise é confiável.

Levando em conta o trabalho envolvido na anotação, é comum que a verificação da concordância seja feita utilizando apenas uma amostra do *corpus*, isto é, temos as mesmas pessoas anotando os mesmos textos apenas em um subconjunto do *corpus*. O resultado da comparação destas anotações é um número que nos diz o grau de confiança que podemos ter nas análises/anotações; que nos diz o quanto as análises/anotações convergiram (ou divergiram), e a partir dele temos uma estimativa de o quão consistente está a anotação no *corpus* todo. Uma alta concordância entre anotadores é indicativa do potencial de reprodutibilidade, isto é, da possibilidade de reprodução das análises por outras pessoas (e espera-se que as máquinas sejam capazes de reproduzir estas mesmas análises).

Quando os resultados da concordância entre anotadores indicam uma baixa consistência entre as análises, é possível explorar algumas alternativas:

- Melhoria das instruções de anotação, como a inclusão de exemplos, tanto positivos ("faça assim nesses casos") quanto negativos ("não faça assim nesses casos");
- Aumento do tempo de familiarização com a tarefa e com o fenômeno sendo analisado, para que as pessoas responsáveis pela anotação estejam seguras do que estão fazendo;
- Reformulação das classes de anotação (do *tagset*). Os resultados da concordância entre anotadores podem ser baixos porque o conjunto de classes não está bem desenhado/modelado, mesmo que ele corresponda às classes de uma teoria.

A prática de medir a concordância entre anotadores é comum não apenas para avaliar a qualidade e consistência da anotação de um *dataset*, mas também em outras tarefas que visam avaliar a qualidade (e confiança) da análise humana. Este é o caso, por exemplo, da **anotação de erros de tradução automática**, uma das maneiras de avaliar a qualidade de uma tradução automática (Seção 21.3.4).

Algumas medidas utilizadas para o cálculo da concordância entre anotadores foram propostas por Cohen (1960), Fleiss (1971) e Krippendorff (1970). O Cohen's Kappa foi originalmente projetado para medir a concordância entre apenas dois anotadores. Se a intenção é calcular a concordância entre múltiplos anotadores, é possível então levar em conta a média de cada par possível de anotadores. Já o Fleiss' Kappa pode ser utilizado com mais de dois anotadores, e o Alfa de Krippendorff é flexível para ser aplicado em cenários com múltiplos anotadores, permitindo considerar diferentes níveis de desacordo.

A interpretação dos valores de Kappa irá depender da tarefa avaliada. Em termos gerais, Cohen sugere que valores 0 indicam nenhuma concordância, 0.01–0.20 indicam concordância fraca a nula, 0.21–0.40 indicam concordância razoável, 0.41–0.60 indicam concordância moderada, 0.61–0.80 indicam concordância substancial e 0.81–1.00 indicam uma concordância quase perfeita.

Por exemplo, na anotação semântica que visa a desambiguação de sentidos (Capítulo 8), as medidas de concordância costumam ser baixas, em torno de 0.7. Já na anotação sintática, que aparentemente seria mais complexa, é comum números superiores a 0.9.

Por fim, existem diferentes maneiras de lidar com as discordâncias, partindo da ideia de que alguns tipos de divergência entre análises humanas são mais graves (ou mais inofensivos) do que outros. O coeficiente Kappa de Cohen, por exemplo, pode ser usado na versão ponderada ou não ponderada. A diferença entre **Kappa ponderado** (*weighted Kappa*) e **Kappa não ponderado** (*non-weighted Kappa*) está na maneira como eles lidam com as discordâncias: o Kappa não ponderado trata todos os pares de categorias/etiquetas da mesma forma, ou seja, todas as discrepâncias são tratadas igualmente, independentemente de sua natureza ou importância. Por outro lado, o Kappa ponderado atribui diferentes pesos às discordâncias entre as categorias/etiquetas, baseando-se em uma escala que reflete a importância atribuída a essas categorias. Por exemplo, em uma escala Likert de 1 a 5, o Kappa não ponderado considera que, no cálculo uma concordância (mais precisamente, de uma discordância), os valores 1, 2, 3, 4 e 5 são equivalentes, isto é, são, todos eles, indicativos de discordância. Já o Kappa ponderado reconhece que apesar dos valores 1, 2, 3, 4 e 5 indicarem a presença de uma discordância, o valor 1 está mais próximo de 2 do que 5, indicando que a discordância entre 1 e 5 é mais grave ou relevante que uma discordância entre 1 e 2.

13.6.2 Avaliação intrínseca

Por meio de uma avaliação da concordância entre anotadores ficamos sabendo que o material está consistente do ponto de vista da análise humana. Porém, não temos muitas pistas sobre o quanto esta análise humana – ou, o quanto o *dataset* (a combinação de análise humana + dados) – permite generalizar. Além disso, consistência na amostra utilizada no cálculo da concordância entre anotadores não significa, necessariamente, que não haja lapsos na anotação ao longo do material.

No PLN, **avaliação intrínseca** é, tradicionalmente, uma avaliação dos modelos e ferramentas (não do *dataset*), e é calculada utilizando medidas de precisão e abrangência apresentadas no início desta seção. Como o nome indica, a avaliação intrínseca é intrínseca à tarefa que está sendo avaliada. Ou seja, se temos uma ferramenta/modelo que faz anotação de sintaxe, ou de entidades, a ferramenta/modelo será avaliada nesta tarefa (parece óbvio, mas pode não ser assim, como veremos na **avaliação extrínseca**). Considerando a geração de um modelo de aprendizado de máquina, para que seja possível realizar uma avaliação intrínseca é importante que o material tenha um *tamanho* que permita todo o ciclo do aprendizado: treino, validação e teste. A partição de *teste* é uma parte do *dataset*, e apenas esta parte será alvo da avaliação (Capítulo 15).

Quando pensamos na avaliação intrínseca de *datasets* (e não dos modelos/ferramentas), estamos fazendo uma *inversão*, ou uma mudança de perspectiva. Se na avaliação intrínseca "original" verificamos a capacidade do modelo ou ferramenta de generalizar a partir dos

dados a que foram expostos, na avaliação intrínseca de *datasets* verificamos (indiretamente) o quanto o *dataset* (mais especificamente, a natureza dos textos + a classificação dos dados) permitiu esta generalização, levando em conta as características do modelo/ferramenta. A partir dessa mudança de perspectiva, quando olhamos para o desempenho de um modelo e vemos números de precisão, abrangência e medida F, não vemos apenas uma avaliação do modelo ou ferramenta, ou o quão bom o modelo/ferramenta é, mas também até onde os dados permitiram ir, considerando os limites do modelo/ferramenta. Podemos entender da seguinte maneira: avaliação intrínseca é uma **análise dos erros** de um modelo/ferramenta que pressupõe que o material que serviu de treino está bem anotado e que o modelo/algoritmo tem um desempenho que não é aleatório. Na Tabela 13.2, por exemplo, vemos o resultado de uma avaliação intrínseca de um modelo/ferramenta que tem um desempenho geral de 89,09%, um desempenho muito bom, portanto. O que vemos na tabela, no entanto, é o desempenho de cada classe individualmente (considerando um *tagset* com 30 classes/etiquetas).

Tabela 13.2: Desempenho de um modelo (desempenho global de 89,09%) para cada classe aprendida

Etiqueta	**Qtde**	**F1**	**Etiqueta**	**Qtde**	**F1**	**Etiqueta**	**Qtde**	**F1**
classe 1	1786	99,27%	classe 11	314	89,81%	classe 21	564	74,47%
classe 2	101	99,01%	classe 12	346	89,02%	classe 22	185	71,89%
classe 3	1920	98,80%	classe 13	1269	88,02%	classe 23	78	62,82%
classe 4	178	96,07%	classe 14	140	86,43%	classe 24	29	62,07%
classe 5	447	94,85%	classe 15	166	86,14%	classe 25	92	61,96%
classe 6	319	93,73%	classe 16	1375	84,73%	classe 26	435	61,84%
classe 7	129	93,02%	classe 17	154	83,77%	classe 27	115	60,00%
classe 8	332	92,77%	classe 18	308	80,19%	classe 28	81	55,56%
classe 9	248	92,34%	classe 19	213	77,93%	classe 29	36	41,67%
classe 10	666	90,09%	classe 20	110	77,27%	classe 30	14	78,57%

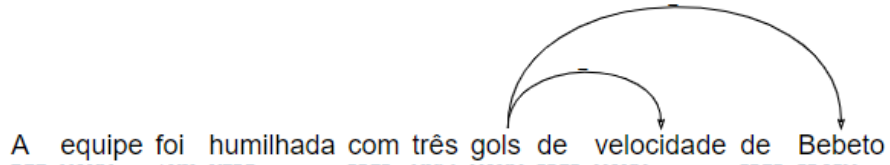
Na tabela, vemos que as classes 14 a 30 foram as mais difíceis, e que as classes/etiquetas 23 a 29 foram especialmente difíceis. Vemos também que, exceto pela classe 26, as classes 23 a 30 são menos frequentes (na partição teste) que as demais. Assim, quando trocamos de perspectiva e olhamos para o desempenho nos dados individualmente, e não para o desempenho global do modelo, pensamos: como fazer para melhorar a classificação prevista nesses casos críticos, sem piorar os demais? O que esses casos têm de especial? Acrescentar mais exemplos para cada classe resolve? Ou o problema está nas instruções de anotação, que não deixam claro como exatamente classificar em certos casos? Ou não há o que fazer, e o problema de generalização é apenas da ferramenta/modelo?

A limitação desta maneira de avaliação é, justamente, a impossibilidade de diferenciar uma dificuldade de generalização oriunda do algoritmo de aprendizado de uma dificuldade de generalização oriunda de análises inconsistentes ou da falta de exemplos suficientes no material padrão ouro. Por isso, é importante garantir que há **consistência** na análise humana, o que vemos na concordância entre anotadores. Outra limitação desta forma de avaliar é que ela se restringe à análise da partição 'teste' de um *dataset*, uma partição pequena. Com tantas limitações, qual a vantagem?

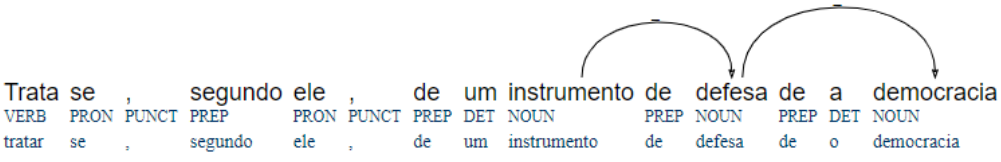
Por meio do resultado de uma avaliação intrínseca das classes, podemos verificar o desempenho relativo à generalização de cada classe utilizada, individualmente (e sempre em contraste com o padrão ouro), e com esse olhar de análise iremos tentar melhorar os

resultados (ou seja, melhorar o modelo) "melhorando" o conjunto de dados, e mexendo apenas nele. Há classificações mais difíceis de serem aprendidas? Por que? Essas perguntas têm como pano de fundo a constatação de que nem sempre o que é fácil para as pessoas é fácil para as máquinas. Para uma pessoa escolarizada, por exemplo, é simples diferenciar a estrutura (1) da estrutura (2). Mas de um ponto de vista formal temos exatamente a mesma sequência: substantivo + preposição + substantivo + preposição + substantivo.

(1)



(2)



Na frase (1), no entanto, o último substantivo da sequência está relacionado ao primeiro elemento (*Bebeto* está associado aos *gols*, e não à *velocidade*); na frase (2), o último substantivo da sequência está relacionado ao segundo elemento (*democracia* está associado à *defesa*, e não a *instrumento*).

O que podemos fazer para que esta distinção fique mais nítida para as máquinas? Mais exemplos ajudam? Ou o material já tem exemplos suficientes de ambas as estruturas (e neste caso não há o que fazer)?

Uma das vantagens deste tipo de avaliação – que joga luz sobre os *datasets*, e não sobre as ferramentas – é a possibilidade de perceber, de forma localizada, os obstáculos para generalização, e então atuar de forma direcionada para construir *datasets* "otimizados" para tarefas. Trata-se de uma forma de avaliar a qualidade da anotação que sinaliza, a partir do "ponto de vista das máquinas", onde há espaço para melhoria no conjunto de dados.

Esta é uma abordagem que nos permite uma visão simultaneamente quantitativa e qualitativa; panorâmica e detalhada dos resultados de uma análise automática, capaz de guiar intervenções e melhorias onde de fato elas são necessárias. Além disso, repetindo o ciclo de aprendizagem após as intervenções é possível verificar se a solução proposta (inclusão de mais exemplos, redefinição das classes, melhoria das instruções de anotação) realmente facilitou a generalização. Ou seja, esta é uma abordagem não só para avaliar, mas que permite também medir o quanto as revisões melhoram um *corpus*, e nos ajuda a decidir um ponto de corte na revisão, respondendo à pergunta "Até onde precisamos rever?", da Seção 13.5.3.

A segunda vantagem desta abordagem já foi esboçada nos parágrafos anteriores: uma avaliação intrínseca de *datasets* permite verificar o resultado de experimentações na classificação dos dados; ela ajuda e informa a tomada de decisões no processo de anotação. Se há duas ou mais maneiras legítimas e adequadas de analisar um mesmo fenômeno, qual escolher? Facilitar a generalização deve ser um critério relevante na escolha, no contexto do PLN. Por exemplo, na frase exemplo (1) do início deste capítulo ("TRISTE é uma palavra de 6 letras"), a palavra TRISTE foi analisada como um *substantivo* (e não

como *adjetivo*), mas essa é uma escolha de anotação: decidir se as classes de palavras são propriedades das palavras (e, portanto, estáticas, e nesse caso TRISTE seria um *adjetivo* em qualquer contexto) ou se as classes são atribuídas em função do papel que exercem na frase (e, portanto, dinâmicas, e nesse caso TRISTE seria um *substantivo* no contexto da frase) deveria ser fruto de uma escolha que leva em conta as tarefas e demandas do PLN (Capítulo 4). Qual maneira de classificar leva à maior generalização sobre a anotação de PoS? E qual maneira de classificar leva ao melhor desempenho em uma outra tarefa – na análise sintática, por exemplo? Esta segunda pergunta se relaciona à *avaliação extrínseca*, tema da próxima seção.

13.6.3 Avaliação extrínseca

A avaliação extrínseca verifica se a informação codificada no *dataset* melhorou o desempenho em tarefas consideradas mais complexas. Por isso, está associada à **adequação**, e não à consistência. Assim como na avaliação intrínseca, esta é uma forma de avaliar aplicada tradicionalmente a modelos ou ferramentas, mas *datasets* também podem ser avaliados dessa maneira (novamente fazendo uma mudança de perspectiva e tomando o ponto de vista dos dados/*dataset*, e do que permitem generalizar). A avaliação é extrínseca porque ela não avalia aquilo que, diretamente, o modelo/ferramenta faz, ou aquilo que o *dataset* codifica. Ela avalia indiretamente, verificando o quanto outras tarefas são beneficiadas quando aquilo que o *dataset* codifica é levado em conta. Ou: uma avaliação extrínseca verifica o quanto a informação codificada em um *dataset* é adequada para as tarefas mais complexas que o *dataset* pretende auxiliar. Por trás dessa ideia, o reconhecimento de que, no PLN, a codificação de certos tipos de informação (principalmente informação linguística especializada, como sintaxe, semântica etc.) é um meio, e não um fim, para resolver as tarefas de PLN. Ou seja, um *dataset* sintático, para tarefas de PLN, interessa não porque codifica informação sintática, mas porque, dispondo de informação sintática, é possível melhorar o desempenho em outras tarefas, como extração de informação[39] (Capítulo 20).

O estudo de Nooralahzadeh; Øvrelid (2018), por exemplo, estava interessado em verificar a contribuição de três tipos de gramáticas diferentes (ou, de três esquemas de anotação) na tarefa de extração de relações semânticas entre entidades, com o objetivo de saber qual o tipo de anotação gramatical (ou, de representação gramatical) mais adequado à tarefa. Ou seja, o estudo queria saber qual gramática ajudaria mais o modelo/ferramenta a chegar aos melhores resultados na extração de relações entre entidades. Para tanto, dispunham de um modelo que utilizava, como entrada, informação de dependências sintáticas[40] (Capítulo 4), e prepararam *datasets* anotados conforme as três abordagens gramaticais: as chamadas dependências sintáticas de Stanford, as dependências sintáticas do projeto Universal Dependencies e dependências sintáticas usadas nas tarefas CONLL[41] (ou seja, exatamente os mesmos textos anotados, a única diferença era o esquema de anotação sintática utilizado em cada um).

Voltando ao exemplo da palavra TRISTE, podemos imaginar dois *datasets* de PoS, que

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

[40] Importante notar que a tarefa não pressupõe a existência de uma anotação sintática anterior.

[41] Considerando o foco deste capítulo, basta sabermos que são três maneiras diferentes de codificar a informação sintática, e para detalhes é possível consultar o artigo original.

contém exatamente os mesmos documentos, mas em um deles as classes são atribuídas independentemente do contexto (e "triste" será sempre um adjetivo); e no outro as classes são atribuídas conforme o papel que exercem em cada frase. Não é difícil prever que a opção pelas classes estáticas ("triste" é sempre um adjetivo) será de mais fácil generalização, e levará a um desempenho melhor em uma avaliação intrínseca. Mas será que essa é a melhor opção se desejamos "aprender" informação sintática?

Em resumo, a avaliação intrínseca verifica a consistência interna da anotação (o que não necessariamente é sinônimo de qualidade, embora, em geral, seja), mas nada nos diz sobre a adequação do *dataset* para uma tarefa específica, e aqui a avaliação extrínseca pode ser útil. E nem sempre um *dataset* que obtêm melhor desempenho na avaliação intrínseca leva aos melhores resultados quando em uma avaliação extrínseca.

No entanto, a avaliação extrínseca só pode ser realizada se a) dispomos do *dataset* para a "tarefa subsequente" – no estudo de Nooralahzadeh; Øvrelid (2018) foi utilizado o *dataset* de uma *avaliação conjunta* de extração de relações entre entidades; b) dispomos de diferentes versões do mesmo *corpus*, que variam apenas quanto ao esquema de anotação (as etiquetas e maneira de utilizá-las): diferentes tagsets de PoS; diferentes representações sintáticas, diferentes representações de papéis semânticos etc.

Para a língua portuguesa, dispomos – ainda que de maneira tímida – de alguns *corpora* com diferentes anotações de um mesmo tipo de atributo, como vemos na Tabela 13.3, que lista *corpora* que são fruto de projetos de pesquisa acadêmica, e estão disponíveis para uso. MacMorpho[42] e Bosque, os mais antigos, são os que contém mais variações, e para o Bosque-UD[43] também há uma versão em que os casos de omissão do sujeito (sujeito oculto, nos termos da Gramática Tradicional) foram explicitados (o material foi criado com o objetivo de verificar o quanto a explicitação de sujeitos nas frases é capaz de facilitar o processamento linguístico ou o quanto a omissão do sujeito – característica da língua portuguesa, mas não da língua inglesa – dificulta o processamento automático do português)[44]. O mais recente deles é o PetroGold[45].

Tabela 13.3: Variações de atributos linguísticos em *corpora* do português

ATRIBUTO	**VARIAÇÃO**	***CORPUS***	**REFERÊNCIA**
PoS	*Tagset* LácioWeb clássico	MacMorpho	(Aluísio et al., 2003)
	Tagset LácioWeb modificado	MacMorpho	(Fonseca; Rosa, 2013)
	UD	MacMorpho	(Freitas et al., 2018)
		Porttinari	(Pardo et al., 2021)
		PetroGold	(Souza; Freitas, 2023)
		Bosque-UD	(Rademaker et al., 2017)
	PALAVRAS/Floresta Sintá(c)tica	Bosque	(Freitas; Rocha; Bick, 2008)

[42]O *corpus* MacMorpho e a documentação estão disponíveis em http://nilc.icmc.usp.br/macmorpho/ e o MacMorpho-UD em https://github.com/own-pt/macmorpho-UD.

[43]O Bosque anotado com a abordagem UD está disponível em https://universaldependencies.org/treebanks/pt_bosque/index.html. Demais versões estão disponíveis em https://www.linguateca.pt/Floresta/levantamento.html

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

[45]O *corpus* está disponível em https://github.com/UniversalDependencies/UD_Portuguese-PetroGold/tree/master

ATRIBUTO	VARIAÇÃO	*CORPUS*	REFERÊNCIA
Sintaxe	UD	Porttinari	(Pardo et al., 2021)
		PetroGold	(Souza; Freitas, 2023)
		Bosque-UD	(Rademaker et al., 2017)
	UD	Bosque-UD com sujeitos ocultos explicitados	(Freitas; Souza, 2021)
	UD - Petrolês	PetroGold	(Souza; Freitas, 2023)
	PALAVRAS/Floresta Sintá(c)tica	Bosque	(Freitas; Rocha; Bick, 2008)

E qual desses é o melhor? Ou, qual dessas maneiras de representar o conhecimento linguístico é melhor? Aliás, melhor para que?

Em termos de avaliação intrínseca, precisamos garantir condições semelhantes de avaliação: mesmo algoritmo/ferramenta e mesmo conjunto de treino e teste. Todos os *corpora*/*datasets* citados estão públicos.

Quanto à adequação à tarefa, a avaliação extrínseca é uma forma interessante de avaliar a qualidade de um *dataset*, mas para que seja possível utilizá-la é fundamental a existência de um segundo *dataset*, que codifica a tarefa mais complexa. E voltamos ao que foi mencionado no início deste capítulo: a relevância de *datasets* para o avanço do PLN.

13.7 Em resumo...

- *Datasets* são importantes para o PLN porque permitem avaliar, treinar e pautar os rumos do PLN. (Seção 13.2)
- *Datasets* bons são consistentes (boa avaliação intrínseca), bem documentados, de tamanho adequado e com dados adequados. (Seção 13.3)
- A anotação de um *dataset* não precisa ser 100% perfeita para funcionar no aprendizado de máquina. (Seção 13.5.3)
- A preparação de um *dataset* se beneficia se a revisão/anotação é feita de maneira sistemática. (Seção 13.5)
- Avaliação intrínseca não significa, necessariamente, melhor adequação à tarefa, que é medida pela avaliação extrínseca. (Seção 13.6)

Um bom projeto cuidadoso de anotação, por sua vez, deverá levar em conta (Seção 13.4):

- Clareza quanto ao fenômeno que será anotado (que se reflete em um bom esquema de anotação);
- Escolha do *corpus* adequado (um *corpus* composto por relatórios de pesquisa é pouco adequado para a anotação de ironia, por exemplo);
- Conhecimento linguístico – para identificação e descrição do fenômeno anotado;
- Conhecimento do problema – para um melhor recorte das classes;
- Uma boa dose de inspiração ou criatividade – para chegar ao equilíbrio em termos de granularidade e generalização;
- Um outro tanto de experimentação – para validação e reformulação das classes, se for o caso;

- Verificação quanto à eficiência da anotação (por exemplo, o tempo levado e o nível de treinamento/conhecimento necessário por parte de quem vai anotar);
- Infraestrutura adequada – diretivas, documentação e ferramenta;
- Avaliação (verificação da concordância entre os anotadores).

A avaliação de um *dataset* pode levar em conta diferentes perspectivas (Seção 13.6):

- A concordância inter-anotadores avalia a consistência das anotações humanas;
- A avaliação intrínseca também avalia a consistência, mas privilegia a capacidade de generalizar a partir dos dados e a possibilidades de experimentações na classificação dos dados, tendo em vista o aprendizado de máquina;
- A avaliação extrínseca leva em conta a adequação de uma anotação para uma determinada tarefa.

Agradecimentos

Agradeço muitíssimo à Helena Caseli e à Graça Nunes pelas sugestões e provocações que deixaram o capítulo bem melhor (espero!).

Referências

ALUÍSIO, S. et al. **An Account of the Challenge of Tagging a Reference Corpus for Brazilian Portuguese**. (N. J. Mamede et al., Eds.)Computational Processing of the Portuguese Language. **Anais**...Berlin, Heidelberg: Springer Berlin Heidelberg, 2003.

BENDER, E. M.; FRIEDMAN, B. Data Statements for Natural Language Processing: Toward Mitigating System Bias and Enabling Better Science. **Transactions of the Association for Computational Linguistics**, v. 6, p. 587–604, 2018.

BOWMAN, S. R. et al. **A large annotated corpus for learning natural language inference**. Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. **Anais**...Lisbon, Portugal: Association for Computational Linguistics, set. 2015. Disponível em: <https://aclanthology.org/D15-1075>

CASELI, H. DE M.; FREITAS, C.; VIOLA, R. Processamento de Linguagem Natural. Em: **Tópicos em Gerenciamento de Dados e Informações: Minicursos do SBBD 2022**. [s.l.] Sociedade Brasileira de Computação, 2022. p. 1–28.

COHEN, J. A Coefficient of Agreement for Nominal Scales. **Educational and Psychological Measurement**, v. 20, n. 1, p. 37–46, 1960.

COUILLAULT, A. et al. **Evaluating corpora documentation with regards to the Ethics and Big Data Charter**. Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14). **Anais**...Reykjavik, Iceland: European Language Resources Association (ELRA), 2014. Disponível em: <http://www.lrec-conf.org/proceedings/lrec2014/pdf/424_Paper.pdf>

FLEISS, J. L. Measuring nominal scale agreement among many raters. **Psychological Bulletin**, v. 76, n. 5, p. 378–382, 1971.

FONSECA, E. R.; ROSA, J. L. G. **Mac-Morpho Revisited: Towards Robust Part-of-Speech Tagging**. Proceedings of the 9th Brazilian Symposium in Information and Human

Language Technology. **Anais**...2013. Disponível em: <https://aclanthology.org/W13-4811>

FREITAS, C. et al. **Tagsets and Datasets: Some Experiments Based on Portuguese Language**. (A. Villavicencio et al., Eds.)Computational Processing of the Portuguese Language. **Anais**...Cham: Springer International Publishing, 2018.

FREITAS, C. **Linguística Computacional**. [s.l.] Parábola Editorial, 2022.

FREITAS, C.; ROCHA, P.; BICK, E. **Floresta sintá (c) tica: bigger, thicker and easier**. International Conference on Computational Processing of the Portuguese Language. **Anais**...Springer, 2008.

FREITAS, C.; SANTOS, D. **Gender Depiction in Portuguese: Distant reading Brazilian and Portuguese literature**. 2nd Annual Conference of Computational Literary Studies. **Anais**...2023. Disponível em: <https://www.linguateca.pt/Diana/download/FreitasSantos2023-2ndCCLS.pdf>

FREITAS, C.; SOUZA, E. Sujeito oculto às claras: uma abordagem descritivo-computacional / Omitted subjects revealed: a quantitative-descriptive approach. **REVISTA DE ESTUDOS DA LINGUAGEM**, v. 29, n. 2, p. 1033–1058, 2021.

GURURANGAN, S. et al. **Annotation Artifacts in Natural Language Inference Data**. Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers). **Anais**...New Orleans, Louisiana: Association for Computational Linguistics, jun. 2018. Disponível em: <https://aclanthology.org/N18-2017>

IGNAT, O. et al. **A PhD Student's Perspective on Research in NLP in the Era of Very Large Language Models**., 2023. Disponível em: <https://arxiv.org/abs/2305.12544>

KRIPPENDORFF, K. Estimating the Reliability, Systematic Error and Random Error of Interval Data. **Educational and Psychological Measurement**, v. 30, n. 1, p. 61–70, 1970.

LUCY, L.; BAMMAN, D. **Gender and Representation Bias in GPT-3 Generated Stories**. Proceedings of the Third Workshop on Narrative Understanding. **Anais**...Virtual: Association for Computational Linguistics, jun. 2021. Disponível em: <https://aclanthology.org/2021.nuse-1.5>

NOORALAHZADEH, F.; ØVRELID, L. **Syntactic Dependency Representations in Neural Relation Classification**. Proceedings of the Workshop on the Relevance of Linguistic Structure in Neural Architectures for NLP. **Anais**...Melbourne, Australia: Association for Computational Linguistics, jul. 2018. Disponível em: <https://aclanthology.org/W18-2907>

PARDO, T. et al. **Porttinari - a Large Multi-genre Treebank for Brazilian Portuguese**. Anais do XIII Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana. **Anais**...Porto Alegre, RS, Brasil: SBC, 2021. Disponível em: <https://sol.sbc.org.br/index.php/stil/article/view/17778>

PIRES, R. et al. **Sabiá: Portuguese Large Language Models**. (M. C. Naldi, R. A. C. Bianchi, Eds.)Intelligent Systems. **Anais**...Cham: Springer Nature Switzerland, 2023.

RADEMAKER, A. et al. **Universal Dependencies for Portuguese**. Proceedings of the Fourth International Conference on Dependency Linguistics (Depling 2017). **Anais**...Pisa,Italy: Linköping University Electronic Press, set. 2017. Disponível em: <https://aclanthology.org/W17-6523>

SANTOS, D. Avaliação conjunta. Em: SANTOS, D. (Ed.). **Avaliação conjunta: um novo paradigma no processamento computacional da língua portuguesa**. Lisboa, Portugal: IST Press, 2007. p. 1–12.

SOUZA, E. DE. **Construção e avaliação de um treebank padrão ouro**. Mestrado—[s.l.] PUC-Rio, 2023.

SOUZA, E. DE; FREITAS, C. **Explorando variações no tagset e na anotação Universal Dependencies (UD) para Português: Possibilidades e resultados com base no treebank PetroGold**. Anais do XIV Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana. **Anais**...Association for Computational Linguistics, 2023.

SOUZA, F.; NOGUEIRA, R.; LOTUFO, R. **BERTimbau: pretrained BERT models for Brazilian Portuguese**. (R. Cerri, R. C. Prati, Eds.)Proceedings of the 2020 Brazilian Conference on Intelligent Systems. **Anais**...Springer International Publishing, 2020.

WALLIS, S. Completing Parsed Corpora. Em: ABEILLÉ, A. (Ed.). **Treebanks: Building and Using Parsed Corpora**. Dordrecht: Springer Netherlands, 2003. p. 61–71.

ZHAO, J. et al. **Gender Bias in Coreference Resolution: Evaluation and Debiasing Methods**. Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers). **Anais**...New Orleans, Louisiana: Association for Computational Linguistics, jun. 2018. Disponível em: <https://aclanthology.org/N18-2003>