REVISTA DA **ABRALIN**

RELATO DE PESQUISA

Avaliação da anotação automática de dependências sintáticas

Elvis DE SOUZA
Pontifícia Universidade Católica do Rio de Janeiro (PUC-Rio)

Cláudia FREITAS
Pontifícia Universidade Católica do Rio de Janeiro (PUC-Rio)

OPEN ACCESS

EDITADO POR
- Jorge Baptista (UAlg)
- Adriana Silvina Pagano (UFMG)
- Marta Deysiane Alves Faria Sousa (UFS)

AVALIADO POR
- Roana Rodrigues (UFS)
- Alisson Hudson Veras Lima (IFS)

SOBRE OS AUTORES
- Elvis de Souza
  Curadoria dos dados, Análise formal, Escrita – rascunho original – e Escrita – análise e edição.
- Cláudia Freitas
  Conceptualização, Metodologia, Escrita – rascunho original – e Escrita – análise e edição.

DATAS
- Recebido: 19/11/2022
- Aceito: 20/12/2022
- Publicado: 09/09/2024

COMO CITAR
De Souza, E.; Freitas, C. (2024). Avaliação da anotação automática de dependências sintáticas. *Revista da Abralin,* v. 22, n. 2, p. 308-331, 2024.

RESUMO

Considerando a importância que dependências sintáticas vêm assumindo em tarefas de Processamento de Linguagem Natural (PLN) e, consequentemente, nos estudos linguísticos voltados para o processamento automático das línguas, apresentamos aqui uma avaliação qualitativa de um *treebank* padrão ouro recém lançado para a língua portuguesa, com o objetivo de identificar (a) os padrões linguísticos que apresentam maior dificuldade para anotadores automáticos, (b) os motivos que podem levá-los a errar essas análises, e (c) ampliar as possibilidades de diálogo entre os estudos linguísticos e a linguística computacional. A anotação sintática foi realizada conforme as diretrizes do projeto *Universal Dependencies* (UD), e a avaliação da anotação foi realizada utilizando ferramentas de código aberto, em três etapas: em primeiro lugar, fizemos uma avaliação intrínseca de um modelo de dependências sintáticas, assumindo que este tipo de avaliação reflete indiretamente a consistência da anotação do *corpus* com o qual o modelo foi treinado; em seguida, detalhamos os resultados desta avaliação, apresentando o índice de acertos de cada classe linguística individualmente, o que nos deu um panorama das dificuldades linguísticas para o aprendizado automático e, também, informação quanto à confiança na análise automática de cada classificação linguística. Por fim, selecionamos as classes com maior número de erros e analisamos todos os casos errados. Os resultados sugerem que, do lado linguístico, já podemos contar com análises consistentes e em quantidade, ao menos aparentemente, suficiente. No que se refere à qualidade dos *parsers* automáticos, o espaço para melhorias linguísticas é cada vez menor.

ABSTRACT

Considering the importance that syntactic dependencies have been assuming in Natural Language Processing (NLP) tasks and, consequently, in linguistic studies focused on the automatic processing of languages, we present here a qualitative evaluation of a recently released gold-standard treebank for the Portuguese language, with the objective of identifying (a) the linguistic patterns that are more difficult for automatic annotators, (b) the reasons that can lead them to make mistakes in these analyses, and (c) to expand the possibilities of dialogue between the linguistic studies and computational linguistics. The syntactic annotation was performed according to the guidelines of the *Universal Dependencies* (UD) project, and the evaluation of the annotation was performed using open source tools, in three steps: firstly, we performed an intrinsic evaluation of a syntactic dependencies annotation model, assuming that this type of evaluation reflects the consistency of the *corpus* annotation with which the model was trained; then, we detail the results of this evaluation, presenting the number of correct answers for each individual linguistic class, which gave us an overview of the linguistic difficulties for automatic learning and, also, information regarding the confidence in the automatic analysis of each linguistic classification. Finally, we selected the classes with the highest number of errors and analyzed all the wrong cases. The results suggest that, on the linguistic side, we can already count on consistent analyzes and in quantity, at least apparently, enough. As far as the quality of automatic parsers is concerned, the room for linguistic improvements is getting smaller and smaller.

PALAVRAS-CHAVE

Anotação linguística. Dependências sintáticas. Treebank de linguística portuguesa. Linguística Computacional. Avaliação de dependências sintáticas.

KEYWORDS

Linguistc annotation. Dependency parsing. Brazilian Portuguese Treebank. Computational Linguistics. Dependency parsing evaluation.

Introdução

Considerando a importância que dependências sintáticas vêm assumindo em tarefas de Processamento de Linguagem Natural (PLN) e, consequentemente, nos estudos linguísticos voltados para o processamento automático das línguas, apresentamos aqui uma avaliação qualitativa de um *treebank* padrão ouro recém-lançado para a língua portuguesa, o *corpus* PetroGold v2[1] (DE SOUZA; FREITAS, 2022).

Um *treebank* (literalmente banco de árvores, ou floresta sintática) é um *corpus* anotado com informação morfossintática, chamado dessa forma pois cada frase é anotada estruturalmente e pode ser representada como uma árvore sintática (BICK *et al.*, 2007). Esse tipo de recurso pode ser utilizado tanto na linguística descritiva quanto no processamento automático do português:

> enquanto a linguística descritiva vê uma floresta como favorecendo sobretudo o ensino e a extracção de dados quantitativos, ou estatísticas, sobre a realidade da língua, um linguista computacional utiliza uma floresta como uma ferramenta para treinar e medir o seu analisador sintáctico (BICK et al., 2007, p. 291).

Uma estratégia que pode ser usada para medir a consistência da anotação de um *corpus* é a avaliação intrínseca. De amplo uso na avaliação de sistemas e modelos de PLN desenvolvidos com abordagens de aprendizado de máquina, a avaliação intrínseca é interna a uma tarefa específica: se a tarefa é anotação morfológica, os dados de treino deverão conter anotação desse mesmo tipo e a avaliação será feita contrastando as anotações previstas pela máquina com o gabarito (as anotações fornecidas por pessoas) em um mesmo subconjunto do *corpus*. Contudo, além de avaliar a qualidade de uma anotação automática, este tipo de avaliação também nos fornece, indiretamente, evidências acerca da consistência desse material, uma vez que, em condições ideais, boas medidas de precisão e abrangência são índices de um bom aprendizado, o que se obtém com dados de treinamento consistentes – isto é, contendo análises semelhantes para fenômenos linguísticos semelhantes –, uma propriedade fundamental em projetos de anotação[2].

No entanto, as medidas de precisão e abrangência da avaliação intrínseca nos dão uma visão geral, quantitativa, do desempenho das máquinas, mas nada nos informam acerca da dificuldade de construções linguísticas específicas. De um ponto de vista linguístico – ou, mais precisamente, linguístico-computacional – pode ser instrutivo analisar fenômenos linguísticos individualmente, medindo o grau de dificuldade que impõem tanto às máquinas, na generalização automática, quanto aos seres humanos, na consistência das análises codificadas na anotação humana.

Avaliação é um termo que pode ser utilizado em diferentes contextos no PLN (AFONSO, 2004; SAMPSON; BABARCZY, 2008; MANNING, 2011; BAIA; PRATES; CLARO, 2020). Podemos avaliar, de um lado, a) se um esquema de anotação está bem definido e se ele é adequado aos objetivos de uma

---

[1] Disponível para download em https://petroles.puc-rio.ai

[2] Deve-se notar, porém, que consistência nem sempre significa correção, pois como observa Freitas (2022), uma anotação errada mas consistentemente aplicada no padrão ouro pode ser aprendida com facilidade de gerar bons números na avaliação intrínseca, a despeito de ser incorreta.

tarefa. De outro, podemos avaliar b) o alinhamento dos anotadores humanos entre si e com relação às diretivas do projeto de anotação, isto é, o quanto uma anotação humana pode ser considerada correta. E podemos ainda avaliar c) o desempenho de ferramentas automáticas de anotação, baseadas em conhecimento linguístico ou em aprendizado de máquina.

Neste trabalho, avaliação se relaciona aos três contextos, ainda que a avaliação de um modelo de linguagem (contexto c) interesse apenas indiretamente. Utilizamos a avaliação de um modelo – uma avaliação intrínseca – para identificar os fenômenos linguísticos mais difíceis para a análise automática. Depois, tentamos identificar motivos para os erros de análise, que podem ter origem em uma deficiência do algoritmo de aprendizado de máquina, sobre o qual não temos controle do ponto de vista linguístico, ou podem ter origem na anotação (isto é, na análise linguística) do recurso que treinou o modelo. O objetivo final é a avaliação da anotação de um *treebank*, o que nos permite visualizar um panorama das dificuldades linguísticas para o aprendizado automático, bem como avaliar a qualidade do trabalho humano – a anotação propriamente, o esquema de anotação e a documentação que guiou a revisão.

Especificamente, avaliamos a anotação de um *treebank* analisado linguisticamente conforme as diretrizes do projeto *Universal Dependencies* (UD) (DE MARNEFFE et al., 2021). Inspirados pelo trabalho de Manning (2011), que faz uma classificação dos erros da saída de uma ferramenta de anotação de classes de palavras, chegamos a uma conclusão contrária à dele, passados mais de dez anos: do lado linguístico, já podemos contar com análises consistentes e em quantidade, ao menos aparentemente, suficiente. No que se refere à qualidade dos *parsers* automáticos, o espaço para melhorias linguísticas é cada vez menor.

1. O *corpus* e a anotação

O *corpus* utilizado neste estudo é o PetroGold v2 (DE SOUZA; FREITAS, 2022), um *treebank* padrão ouro do domínio do petróleo. O *corpus* é composto por 19 teses e dissertações – documentos do gênero acadêmico, e/ou técnico-científico –, resultando em 8.949 frases e 250.595 *tokens*. Os documentos foram processados na integralidade: foram removidos apenas elementos como tabelas, figuras, fórmulas, e seções como índice e referências bibliográficas. Os textos foram obtidos do Petrolês (CORDEIRO, 2020), um corpus de documentos públicos do domínio do óleo e gás.

O *corpus* foi anotado conforme o formato e a gramática do projeto *Universal Dependencies* (UD), uma iniciativa que visa promover um ambiente de anotação gramatical (morfologia e sintaxe) que seja consistente entre as mais de 100 línguas que compõem o projeto. O projeto conta com um conjunto de etiquetas de 17 classes gramaticais e 37 relações sintáticas e o modelo gramatical da anotação é o de dependências sintáticas. Embora inicialmente apresentado como uma abordagem gramatical para anotação, recentemente UD se declarou explicitamente como uma teoria linguística, fruto de um esforço coletivo que tem como objetivo fornecer um arcabouço teórico para a anotação consistente de diferentes línguas (DE MARNEFFE *et al.*, 2021). Enquanto arcabouço teórico, UD tem

ambições variadas, voltadas tanto para aplicações de PLN – facilitar a criação de sistemas de PLN multilíngues, permitir a transferência de aprendizado (automático) multilíngue e possibilitar comparar as dificuldades de anotação entre as diferentes línguas do projeto, por exemplo – como para estudos linguísticos comparativos.

Alguns princípios linguísticos guiam a gramática UD, e listamos dois deles apenas com o intuito de sinalizar que, apesar do que o nome *universal* sugere, estamos diante de uma teoria fruto de escolhas e interesses, como qualquer outra[3]. O primeiro deles é a independência entre os níveis de classe de palavra (POS) e de sintaxe. Ou seja, no nível morfológico, as classes são atribuídas às palavras não em função do contexto sintático em que ocorrem, mas conforme sua classe sintática "padrão". Assim, em "Os ausentes não puderam ser contemplados" (exemplo fictício), a palavra *ausentes* deve ser classificada como um adjetivo. Uma consequência bastante discutível desta decisão se refere a certos nomes de entidades, que passam a ser anotados literalmente. Em uma frase como "Adorei Corra!" (encontrada na internet), a palavra *Corra* – nome de um filme, neste contexto – deve ser anotada como um verbo, e não como um nome próprio. Outra decisão que vale a pena mencionar porque se aplica a fenômenos comuns na língua portuguesa refere-se à classe do aposto, concebida de maneira bastante restrita: aposto deve ser um elemento nominal, o que exclui a análise de orações apositivas, como no exemplo (fictício): "Fez o que deveria ser feito: *resolveu* a questão assim que pôde".

A anotação do PetroGold foi obtida a partir de análises automáticas que foram intensamente revistas por quatro anotadores experientes: quando submetidos a um teste de concordância interanotadores, obtiveram no mínimo 91,9% de concordância utilizando a métrica κ (ARTSTEIN, 2017), que calcula o quanto os anotadores concordaram nas suas análises e desconsidera no cálculo a probabilidade de haver concordância por acaso. A revisão da anotação foi realizada utilizando três métodos de revisão, apresentados em Freitas e de Souza (2023) e implementados ao longo das versões publicadas do *treebank* (DE SOUZA; SILVEIRA *et al.*, 2021b; DE SOUZA; FREITAS, 2022; DE SOUZA, 2023). O *corpus* é particionado em subconjuntos de treinamento e de teste seguindo a proporção de 95% e 5% das frases, de maneira que a partição de treino tem 8.671 frases.

2. Trabalhos relacionados

Em um trabalho de 2011, o cientista da computação Christopher Manning discute possibilidades para melhorar a anotação automática de classes gramaticais (POS) que, à época, conseguia resultados próximos a 97,3% de acertos por *token*, embora apenas 56% das frases estivessem inteiramente corretas. O autor sugere que há pouco espaço para melhorar o sistema de anotação automática (Stanford Part-of-Speech Tagger) do ponto de vista computacional, cabendo à linguística a tarefa de alavancar os

3 Importante também notar que as diretrizes linguísticas estão constantemente sendo discutidas pela comunidade de colaboradores.

resultados. O autor realiza uma análise de erros de uma amostra com 100 casos em que o sistema errou para identificar que pontos poderiam ser melhorados, e distribui os erros em 7 classes. Dessas, 3 são de origem linguística, derivadas de falhas na concepção do esquema de anotação ou na sua aplicação pelos anotadores humanos, correspondendo a 55,5% dos erros analisados pelo autor[4].

A primeira das três classes, que compreende 12% dos erros, diz respeito à pouca especificação ou clareza das categorias do esquema de anotação[5]. O autor utiliza como exemplo o fato já conhecido de que em muitos contextos é difícil decidir entre uma etiqueta de verbo ou de adjetivo para as formas participiais, como para a palavra "discontinued" na frase "it will take a $10 million fourth-quarter charge against *discontinued* operations". Manning (2011) observa que, para casos limítrofes, testes linguísticos podem ser decisivos na escolha entre uma classe ou outra, mas que não funcionam em todos os contextos, como o apresentado.

Para esses casos, uma boa documentação de um projeto de anotação talvez fosse capaz de estabelecer critérios que definem – mesmo que artificialmente – quando o termo deve ser anotado como verbo ou como adjetivo. Sampson e Babarczy (2008) questionam esse tipo de prática pois, para eles, enquanto o uso linguístico é inerentemente ambíguo, certas distinções puramente lógicas com o objetivo de desambiguar casos nebulosos não correspondem a nenhum significado linguístico real. Para o PLN, contudo, critérios artificiais podem ser úteis quando facilitam a obtenção de consistência, facilitando a generalização, desde que o custo para isso não seja a perda de informatividade.

Um exemplo de consistência buscada artificialmente está no trabalho de Freitas, *Trugo et al.* (2018), em que os autores decidiram adicionar uma etiqueta específica para as formas participiais com o único objetivo de eliminar a já referida pouca especificação entre verbo e adjetivo em certos contextos. Essa pequena modificação levou à melhoria na identificação das classes – as confusões[6] entre adjetivos e verbos caíram de 295 para 53, e as confusões entre verbos e adjetivos caíram de 287 para 32 – dando suporte à tese de Manning (2011) de que uma seara a ser explorada para melhorar o desempenho dos anotadores está na modelagem linguística.

A segunda classe de erros de anotação identificada por Manning (2011) se relaciona a erros derivados da falta de consistência no padrão ouro ou da falta de diretivas claras[7]. Diferentemente da primeira classe, nesses casos uma resposta correta é plausível, mas ou o esquema de anotação não previu o fenômeno ou as diretivas de anotação falharam em apontar como casos específicos deveriam ser anotados. Em decorrência da falha na documentação, os anotadores humanos foram

---

4 O número de 55,5% de erros de origem linguística é resultado de uma soma feita por nós, uma vez que o autor não diferencia tão claramente o que são erros de origem linguística e erros de origem dos sistemas. Para chegar a este número, consideramos os erros do tipo "Padrão ouro errado" (Gold standard wrong), "Inconsistente/sem padrão" (Inconsistent/no standard) e "Subespecificado/pouco claro" (Underspecified/unclear).

5 No original: "Underspecified/unclear".

6 Chamamos de *confusões* as divergências entre as análises automáticas e humanas.

7 No original: "Gold standard inconsistent or lacks guidance".

inconsistentes. Por exemplo, na frase “Orson Welles‘s Mercury Theater in the ‘30s”, não havia indicação na documentação se o número deveria ser anotado como numeral cardinal ou como substantivo, uma questão que não é derivada de fronteiras linguísticas pouco claras como no caso das formas participiais, mas de falta de um direcionamento que deveria vir da documentação. Esse tipo de erro é o mais frequente de todos, correspondendo a 28% de todos os erros do sistema.

A terceira e última classe de erros decorre da aplicação errada de um esquema de anotação[8] – nesse caso, o esquema é claro e a documentação direciona os anotadores corretamente, mas ainda assim a anotação gabarito está incorreta por outros motivos, como falta de compreensão do esquema por parte do anotador ou lapso na sua aplicação. Esse tipo de erro corresponde a 15,5% dos erros encontrados na amostra analisada pelo autor.

Para concluir que uma anotação humana (gabarito) estava incorreta, Manning (2011) precisou realizar a sua própria análise, que considerou ser a correta. O autor avaliou se um erro era considerado erro do sistema ou subespecificação das diretivas ou do esquema, o que nem sempre os anotadores podem fazer durante um projeto de anotação que já está em andamento. Além disso, um gabarito humano é ainda uma interpretação humana e, como lembram Sampson e Babarczy (2008), “uma vez que não existe uma única ‘anotação correta’ universalmente aceita de qualquer forma linguística, é difícil ter uma ideia de quão consistente e refinado qualquer conjunto utilizável de convenções de anotação pode ser”[9] (p. 472).

Por fim, e embora o trabalho de Manning (2011) tenha sido importante para nos motivar a realizar uma análise de erros, lembramos que a ideia de analisar erros de anotação linguisticamente – e qualitativamente – com o objetivo de tentar melhorar o desempenho de anotadores automáticos já foi explorada em outros trabalhos. Para a língua portuguesa, Oliveira et al. (2006) analisam a saída de um identificador de sintagmas nominais, propondo uma tipologia de 9 classes de erros com o objetivo de melhorar templates de regras do Aprendizado Baseado em Transformações (*Transformation-Based Learning*, ou TBL) algoritmo de aprendizado de máquina utilizado na época. Afonso (2004) e Sampson e Babarczy (2008) discutem erros da concordância interanotadores e classificam os erros com base (a) no tipo de informação linguística divergente, no primeiro trabalho, e com base (b) nos motivos que levaram os anotadores a cometê-los, no segundo trabalho.

3. Metodologia

8 No original: “Gold standard wrong”.

9 Tradução livre. Original: "since there is no single universally-agreed 'correct annotation' of any linguistic form, it is hard to get a feeling for how consistent and refined any usable set of annotation conventions can be"

A avaliação da anotação foi feita em três etapas: em primeiro lugar, fizemos uma avaliação intrínseca de um modelo gerado utilizando como conjunto de treino o *treebank* PetroGold (versão 2[10]). Em seguida, detalhamos os resultados desta avaliação, apresentando o índice de acertos do modelo para cada classe linguística individualmente. Por fim, selecionamos as classes com o maior número de erros e analisamos todos os casos errados.

O modelo alvo da avaliação intrínseca foi gerado utilizando o programa UDPipe em sua versão 1.2.0 (STRAKA; HAJIC; STRAKOVÁ, 2016) e utilizando o *script* de avaliação conjunta do CoNLL de 2018 (ZEMAN *et al.*, 2018), que fornece medidas de avaliação para dependências sintáticas.

3.1 Medidas de avaliação intrínseca

Das métricas de avaliação discutidas em Zeman *et al.* (2018), relatamos aqui, especificamente, as métricas de UPOS (*universal part-of-speech score*), responsável por avaliar a atribuição de etiqueta de classe gramatical, UAS (*unlabeled attachment score*), responsável pela avaliação do encaixe de dependência sintática sem avaliar a etiqueta atribuída para a relação, LAS (*labeled attachment score*), quando se conta, além do encaixe, também a etiqueta designada, e CLAS (*content-label attachment score*), que diz respeito ao encaixe e à anotação apenas para palavras de conteúdo lexical, excluindo palavras funcionais e pontuações. A métrica CLAS foi desenvolvida sob a justificativa de que, sendo o UD um projeto multilíngue, seria necessário utilizar uma medida de avaliação que descartasse as imensas diferenças relativas à frequência com que cada língua emprega palavras funcionais, focando a avaliação apenas nas palavras que, teoricamente, seriam utilizadas em frequência comparável entre todas as línguas, garantindo uma forma de avaliar que não sofre de viés aritmético (NIVRE; FANG, 2017).

3.2 Avaliação qualitativa e tipologia de erros

Analisamos linguisticamente os erros[11] encontrados na avaliação intrínseca do modelo de duas formas. Primeiro, listamos os erros por categoria morfossintática utilizando a ferramenta Julgamento[12] (DE SOUZA; FREITAS, 2021), que nos permite analisar o desempenho do modelo por classe e, consequentemente, conseguimos saber as classes linguísticas que o analisador automático mais erra. Em seguida, selecionamos as classes com pior desempenho e classificamos os erros encontrados em cada uma segundo uma tipologia inspirada em Manning (2011).

---

[10] Para uma apresentação do *corpus* e das diferentes versões do recurso, ver Souza (2023).

[11] De forma mais precisa, o que chamamos de erros são divergências entre a saída do modelo e a análise humana codificada no *corpus* padrão ouro.

[12] Disponível para download em https://github.com/alvelvis/Julgamento

Classificamos os erros em dois grandes grupos: aqueles que são de fato erros do analisador automático (erros do tipo 1) e aqueles que são erros da análise humana (erro do tipo 2), como indica a tabela 1. Quando o erro tem origem humana, corrigimos o *corpus* – tanto a partição de teste como todo o *corpus*, caso o erro tenha sido consistentemente introduzido e seja possível encontrar outros casos parecidos. Além dessa primeira classificação, procuramos possíveis causas para o erro. As classes de erros utilizadas são explicadas a seguir.

1	**Erro do analisador automático**
1.1	Fenômeno pouco frequente no *corpus*
1.2	Ambiguidade estrutural
1.3	Sem explicação aparente
2	**Erro do padrão ouro**
2.1	Faltam diretivas claras
2.2	Diretivas não seriam o suficiente
2.3	Sem explicação aparente

TABELA 1 – Tipologia de erros
Fonte: Elaborado pelos autores

Tipo 1.1 – Fenômeno pouco frequente: Essa categoria se destina aos erros do analisador automático possivelmente derivados do fato de que o fenômeno em questão é pouco frequente no *corpus*, resultando em uma difícil generalização no processo de aprendizado automático.

1. Arenitos laminados podem ser mais sensíveis à invasão de sólidos do que os arenitos homogêneos por os seguintes fatos: • Os finos gerados pela penetração nessas lâminas tendem a ser menores e mais difíceis de serem **removidos**, sendo mais propícios a invadir e aprisionar se em a formação. •[13]

2. Observou-se que a Argila A4 quando hidratada **desfolha**.

Na frase 1, a locução verbal “serem removidos”, cujo governante é o verbo no particípio “removidos”, modifica o adjetivo “difíceis”. Em uma versão anterior das diretivas de anotação do projeto UD, esse fenômeno receberia a etiqueta de relação “ccomp”, etiqueta atribuída às orações substantivas objetivas, fenômeno que ocorre 667 vezes no *corpus*, contra apenas 18 ocorrências em que a etiqueta é utilizada com o verbo modificando um adjetivo, como na frase em destaque. Essa baixa ocorrência do fenômeno (complementos oracionais de adjetivos) poderia ter causado o erro do

13 Este exemplo e os demais foram retirados do *corpus* PetroGold.

analisador automático que, embora tenha acertado o governante da relação, a classificou como de tipo “advcl” (etiqueta atribuída a orações adverbiais)[14].

No exemplo 2, "desfolha" foi anotado pelo analisador automático como um substantivo, objeto do verbo "hidratada", quando na verdade funciona como núcleo da oração subordinada substantiva objetiva direta da oração principal, cujo núcleo é "observou-se". Uma explicação possível diz respeito à estrutura sintática: na frase, há uma oração do tipo objeto direto (cujo núcleo é "desfolha") à qual se encaixa, à sua esquerda, uma oração do tipo adverbial (cujo núcleo é "hidratada"). Esse ordenamento das orações (subordinada adverbial anteposta à oração principal) é menos comum (700 ocorrências) do que o ordenamento padrão (2373 ocorrências), e apenas 22,8% das frases com oração subordinada adverbial seguem esse ordenamento. Além disso, dessas 700 ocorrências, apenas 219 (31,2%) não têm sinal de pontuação dependente da oração adverbial, o que é o padrão quando deslocamos uma oração subordinada para o início da frase (como nas frases 3, com sinal de pontuação, e 4, sem sinal de pontuação). A associação desses fatores faz com que a frase do exemplo figure entre um grupo restrito de 7% de frases com oração adverbial que, apesar de a terem à esquerda da principal, não têm sinal de pontuação marcando o seu deslocamento.

3. ***com vírgula*** – Quando se **observa** este mapa isoladamente , é difícil distinguir a presença de estes quatro corpos , que formam um alinhamento curvo aproximadamente N-S , muito bem delineado em o mapa de amplitude de o sinal analítico.

4. ***sem vírgula*** – Para **vencer** estes desafios é necessário identificar oportunidades para a indústria reduzir seus custos de produção com menor impacto em a natureza.

Tipo 1.2 – Ambiguidade estrutural: Essa categoria corresponde aos erros do analisador automático ao se deparar com estruturas que admitem duas ou mais análises sintáticas distintas, e a ambiguidade só pode ser resolvida acionando conhecimentos que estão além da sintaxe.

5. Primeiramente , deve-se calcular a mobilidade de o gás carbônico (CO2 mobility) dentro de o reservatório , que **corresponde** a a permeabilidade absoluta de o reservatório dividido por a viscosidade de o CO2 sob as condições de pressão e temperatura de o reservatório .

14 Uma possível solução já está em curso por parte do projeto UD, pois em versões mais recentes das diretivas do já é recomendado igualar os modificadores oracionais de adjetivos aos modificadores oracionais de substantivos e advérbios, isto é, como “acl” (oração adjetiva).

6. Em razão de a grande extensão de o trecho de o carboduto entre as fontes de captura e as plataformas de injeção de CO2 , será necessário construir duas estações de recompressão (**booster** station) .

7. Uma de as principais características a ser observada durante a escolha de o tipo de processamento a ser adotado é a razão entre os volumes produzidos de gás associado e óleo (**RGO**) .

Na frase 5, há uma oração relativa, cujo núcleo é "corresponde", e uma série de substantivos dos quais a oração poderia ser dependente, como "mobilidade", "gás" e "reservatório". Associar a oração relativa à "mobilidade" é resultado de um conhecimento que extrapola os limites de uma análise sintática, relacionado a como os enunciados matemáticos são construídos.

Em 6, "estação de recompressão" é um sintagma com dois nominais, de tal maneira que ambos poderiam ser o referente para o que está inserido nos parênteses, "booster station". Uma tradução direta nos auxiliaria a afirmar que *booster station* se refere à *estação de recompressão*, e não apenas à *recompressão*, motivo pelo qual sabemos qual deve ser a anotação correta, a despeito da ambiguidade sintática.

Já na frase 7, sabemos que "RGO" é uma sigla para "razão gás-óleo", portanto, um aposto do *token* "razão". Sem o conhecimento da sigla, um anotador desavisado poderia associar o *token* "RGO" a qualquer um dos nominais: razão, volume, gás ou óleo, motivo pelo qual o erro é do tipo ambiguidade estrutural.

Tipo 1.3 – Sem explicação aparente: A terceira categoria de erros do analisador automático foi atribuída àqueles para os quais não foi possível encontrar uma explicação que justificasse o erro – se esperaria que, por serem casos frequentes e sem ambiguidade, teriam sido analisados corretamente pelo anotador.

8. As dimensões de as folhas tetraédricas e octaédricas são tais que podem se reajustar ou se encaixar entre si para formar camadas compostas por duas ou mais folhas , em uma variedade de maneiras , as quais **dão** origem a a maioria de as estruturas fundamentais de os argilominerais conhecidos .

Na frase 8, por exemplo, há uma oração relativa cujo núcleo é "dão" (ou "dão origem", em uma leitura multipalavras) que deve modificar "maneiras". Contudo, o analisador automático anotou a oração relativa como dependente de "encaixar", outra oração. Não parece haver nenhum motivo para esse erro do ponto de vista linguístico – o fenômeno das orações relativas é muito comum e está em uma estrutura padrão, além de que o corpus está consistentemente anotado, pois não há nenhuma ocorrência de orações relativas dependentes de outras orações em todo o material, não havendo, portanto, ocorrências de onde o modelo possa ter aprendido a realizar a anotação errada.

Tipo 2.1 – Faltam diretivas claras: Os erros da categoria 2.1 têm sua origem na falta de documentação sobre um fenômeno linguístico, levando os anotadores humanos a serem inconsistentes na sua análise, portanto, gerando dados pouco confiáveis para o aprendizado automático. São questões que poderiam ser facilmente resolvidas caso tivessem sido previstas e devidamente documentadas. Nesses casos, é difícil dizer se a anotação do padrão ouro ou a anotação do analisador automático é a correta, pois não há um padrão a ser seguido.

9. Gráfico 13 : Água e óleo - Duto 3D 500 A – **500 Hz**

10. Reprodutibilidade : O catalisador 5% MoC/CBV-740 **apresentou** a mesma atividade de o que catalisador de mesma composição e suporte preparado por ROCHA [9] .

Na frase 9, o analisador automático coordenou "Hz" à palavra "Duto", ao passo que o padrão ouro coordena a palavra "Água". Ambas as anotações podem ser corretas se houvesse documentação específica para casos como esse – coordenações múltiplas, sem conectivos ou, mais especificamente, legendas de gráficos. Na ausência de diretivas sobre o fenômeno, não se pode afirmar qual anotador errou.

Em 10, o anotador automático analisou "apresentou" como núcleo da oração principal, enquanto o padrão ouro analisou "reprodutibilidade" como núcleo da oração principal, sendo "apresentou" *parataxis* subordinado à oração principal. Há uma tendência de se anotar o primeiro *token* como o núcleo da oração em casos como esse, mas a tendência está deficientemente documentada, com poucos exemplos, tornando a anotação do padrão ouro inconsistente, o que pode ter gerado o erro.

Utilizamos três documentos para guiar a revisão da anotação do PetroGold: as diretrizes oficiais do projeto[15], um documento com as diretrizes de UD específicas para a língua portuguesa[16] (DE SOUZA; CAVALCANTI et al., 2020), e um outro específico para as questões do PetroGold, construído durante o processo de revisão do *corpus*[17] (DE SOUZA; SILVEIRA *et al*., 2021a).

Tipo 2.2 – Diretrizes não seriam o suficiente: A categoria de erros 2.2 diz respeito àqueles que têm sua origem em fenômenos linguísticos conhecidamente complexos, cuja anotação depende de escolher entre uma ou outra interpretação possível, sendo que por vezes nem documentar os casos em que uma outra decisão foi tomada é suficiente, pois cada contexto suscita uma interpretação distinta. Acrescentam-se a essa classe as frases de difícil interpretação, seja porque estão mal

[15] Disponível em https://universaldependencies.org/guidelines.html

[16] Disponível em https://comcorhd.letras.puc-rio.br/Documenta-o-UD-PT

[17] Disponível em https://www.researchgate.net/publication/365597977_Documentacao_da_anotacao_morfossintatica_do_PetroGold

escritas, seja porque a interpretação depende de conhecimento de domínio muito específico, tornando difícil decidir qual a anotação correta.

11. Apesar de os estudos de estes autores terem sido restritos à região de Canabrava , **nordeste** de a área estudada em este trabalho , notam se semelhanças entre os resultados de aqueles autores e os dados aqui apresentados .

12. Pode se observar em seção a anomalia associada a o alinhamento , para a qual a estimativa de a Deconvolução de Euler sugere uma associação com dique (**índice** estrutural igual a 0,7 , próximo a 1) .

13. A Sub-Bacia Abaeté passou a ser preenchida por sedimentos mais pelíticos , mas ainda com **contribuição** arenosa , possivelmente em decorrência de aumento em a umidade

14. Tabela IV.18 – Força máxima (F) para desprender o êmbolo de a **argila** A4

Na frase 11, é difícil dizer se "nordeste" deve se relacionar à região ou a Canabrava, pois ambas podem se situar "a nordeste", sendo a diferença entre ambas as anotações sutil e de difícil distinção para quem não é especialista do domínio. Nota-se que, neste caso, mesmo especialistas podem ter dificuldade em dizer qual o referente de "nordeste" se de fato ambas as palavras puderem ocupar este lugar, o que parece ocorrer de fato.

De modo semelhante, na frase 12, a estrutura do sintagma entre parênteses é a mesma utilizada tanto para a estrutura de aposição quanto de parataxis. Um não especialista do domínio (como os autores deste trabalho) teria dificuldade em dizer se "índice" é um aposto de “estimativa”, aposto de “Deconvolução de Euler” ou ainda um aposto de “associação”. A hipótese de “índice” ser parataxis de “sugere” também não está descartada, pois trata-se de um sintagma sem conectivo que pode se relacionar ao verbo, adicionando informação cujo conteúdo, infelizmente, não conseguimos distinguir de que tipo é.

Na frase 13, há muitas possibilidades de encaixe para “contribuição” – a oração pode funcionar como coordenada à oração cujo núcleo é “preenchida”; tendo ocorrido a elipse do verbo na oração coordenada, pode funcionar como coordenada a “pelíticos”, ou ainda como coordenada a “sedimentos”. É necessário, para anotar a frase, entender ao que se está contrastando o que se chamou “contribuição arenosa”, sendo necessário conhecimento específico do domínio suficiente para se estabelecer a relação – de oposição, de coordenação ou de concessão – entre os elementos da frase.

Por fim, na frase 14, há ambiguidade possível entre uma estrutura adnominal – “êmbolo da argila” como um objeto do verbo “desprender” (sendo “argila” *nmod* de “êmbolo”) – e entre uma estrutura argumental do verbo – “da argila” como lugar de onde se deve desprender o êmbolo (“argila” como

obl:arg do verbo “desprender”). Ambas as leituras são aceitáveis, sobretudo com pouco contexto como ocorre na frase.

Tipo 2.3 – Sem explicação aparente: Os erros do tipo 2.3 são aqueles que não são derivados nem da falta de diretrizes claras nem de fenômenos linguísticos inerentemente complexos, de difícil categorização. São, portanto, erros sem explicação aparente, inconsistências introduzidas ou negligenciadas pelos anotadores humanos durante a revisão do *corpus* mas que podem ser facilmente corrigidas uma vez que foram identificadas.

15. O Escudo Sul-Rio-Grandense localiza se em a porção meridional de a Província Mantiqueira (Almeida et al. 1981, **Hasui** et al. 1985) , englobando o Orógeno Dom Feliciano , corresponde a a área do Estado de o Rio Grande de o Sul que é marcada por a ocorrência de rochas ígneas , metamórficas e sedimentares pré-paleozóicas , cuja origem é relacionada a os ciclos Transamazônicos (Paleoproterozóico) e Brasiliano/Pan-Africano (Neoproterozóico) .

16. II.3 – **ENCERAMENTO** DE BROCA

Na frase 15, “Hasui” estava anotado como aposto de “Almeida”, quando na verdade há uma coordenação (conj) entre os termos. Na frase 16, embora tenhamos definido que caracteres indicadores de seção, como no caso de “II.3”, sejam casos de adjunto adnominal (nummod) do primeiro nome (“ENCERAMENTO”), a frase tinha permanecido anotada erroneamente como sendo o numeral a raiz da frase e o substantivo o seu modificador.

4. Resultados e análise: avaliações linguísticas

A tabela 2 apresenta os números relativos à avaliação intrínseca do *corpus*, e vemos que, apesar de altos, ainda há espaço para melhorias, sobretudo quando considerando as medidas de LAS e CLAS.

Corpus	**LEMMA (%)**	**UPOS (%)**	**UAS (%)**	**LAS (%)**	**CLAS (%)**
PetroGold v2	98,54	98,40	90,92	89,09	**84,07**

Tabela 2 – Resultados da avaliação intrínseca do PetroGold
Fonte: Elaborado pelos autores

A seguir, podemos conferir os números de avaliação para cada uma das categorias morfossintáticas e uma análise qualitativa dos erros.

Classes gramaticais: Na tabela 3, a primeira coluna corresponde à etiqueta da classe, a segunda ao número de *tokens* com a classe no padrão ouro, e a terceira ao número de acertos da classe - uma média harmônica entre precisão e abrangência (F1). Os dados da tabela dizem respeito somente à ocorrência das classes na partição de teste, uma vez que é ela que está sendo analisada em uma avaliação intrínseca.

POS	#	F1 (%)	POS	#	F1 (%)	POS	#	F1 (%)
CCONJ	319	100	NUM	349	98,85	PROPN	521	95,59
PUNCT	1431	100	VERB	976	98,46	ADJ	791	94,56
DET	1809	99,67	NOUN	2866	98,19	**SCONJ**	**83**	**86,75**
ADP	2099	99,52	PRON	271	97,05	**X**	**8**	**75**
AUX	320	99,06	ADV	335	96,12	**SYM**	**36**	**69,44**

Tabela 3 – Distribuição dos acertos de POS no *corpus*
Fonte: Elaborado pelos autores

As únicas classes com menos de 90% de acertos (SCONJ, X, SYM) são as que têm menos de 100 ocorrências na partição teste. Um dos motivos para o baixo índice de acerto, portanto, poderia ser a pouca ocorrência no *corpus*, dificultando a generalização. Por outro lado, como são poucas as ocorrências na partição teste, um único erro impacta muito negativamente o número relativo de acertos.

O índice de acertos das classes NOUN, PRON, ADV, PROPN e ADJ ficou abaixo da média das outras classes. Ao longo do processo de revisão do *corpus*, percebemos que muitos erros de anotação (erros oriundos de uma primeira anotação automática do *corpus*) vinham de confusões entre artigos (DET) e pronomes (PRON), devido sobretudo à ambiguidade da forma "o" em português; entre substantivos comuns (NOUN) e próprios (PROP); e entre adjetivos (ADJ) e verbos (VERB), sobretudo devido às formas participiais. No decorrer da revisão, demos um tratamento sistemático para estas questões, mas o número de divergências entre anotadores nos indica que são classes cuja identificação continua trazendo algum grau de dificuldade para as máquinas.

Tiveram desempenho acima da média, por sua vez, as classes CCONJ, PUNCT, DET, ADP, AUX, NUM e VERB - classes fechadas em sua maioria ou com muitas categorias de flexão, como é o caso dos verbos, tornando mais fácil a sua identificação. Importante também lembrar que, em UD, locuções prepositivas e conjuntivas, por exemplo, devem receber, no nível de POS, uma análise literal de seus elementos. Assim, as locuções *a partir de* e *isto é* devem ser anotadas como *preposição verbo preposição* e *pronome verbo*, respectivamente, o que, sem dúvida, é estranho, mas contribui para acertos automáticos.

Relações sintáticas: Na tabela 4, a primeira coluna é a etiqueta da classe sintática, a segunda é o número de *tokens* com a classe no padrão ouro, a terceira é o número de acertos da relação de dependências, independentemente do encaixe da dependência, a quarta é o número de acertos da relação e do encaixe ao mesmo tempo (LAS). A tabela está organizada por LAS de forma decrescente. Classes com LAS menor que o F1 da avaliação intrínseca (89,10%), em ordem decrescente: nsubj,

nmod, nsubj:pass, fixed, punct, mark, advmod, aux, acl, xcomp, obl, advcl, csubj, obl:arg, ccomp, acl:relcl, conj, appos, parataxis[18].

Dessas classes, algumas apresentaram mais erros relacionados ao encaixe da dependência do que à atribuição da etiqueta relativa à natureza da relação. Elas são, em ordem decrescente de dificuldade: acl:relcl, conj, appos, acl, punct, advmod e advcl. São as classes em

REL	#	HIT (%)	LAS (%)	REL	#	HIT (%)	LAS (%)
flat	3	100	100	punct	1375	99.56	84.73
flat:foreign	1	100	100	mark	154	87.66	83.77
obl:agent	52	100	100	advmod	308	94.81	80.19
det	1786	99.66	99.27	aux	14	78.57	78.57
expl	101	100	99.01	acl	213	92.96	77.93
case	1920	99.43	98.80	xcomp	110	78.18	77.27
aux:pass	178	96.07	96.07	obl	564	79.96	74.47
root	447	94.85	94.85	advcl	185	83.24	71.89
obj	319	94.67	93.73	csubj	7	71.43	71.43
cop	129	96.90	93.02	obl:arg	78	62.82	62.82
cc	332	97.29	92.77	ccomp	29	65.52	62.07
nummod	248	95.16	92.34	acl:relcl	92	95.65	61.96
amod	666	94.14	90.09	conj	435	88.28	61.84
flat:name	314	91.08	89.81	appos	115	81.74	60
nsubj	346	90.75	89.02	parataxis	81	64.20	55.56
nmod	1269	92.75	88.02				
nsubj:pass	140	86.43	86.43				
fixed	166	86.14	86.14				

Tabela 4 – Distribuição sintáticos no *corpus*
Fonte: Elaborado pelos autores

que a identificação da relação correta - uma relação de adjetivação, aposto, condição (ou qualquer outro dos sentidos adverbiais) - não é tão difícil de ser feita por seres humanos quanto o é para as máquinas devido a ambiguidades estruturais.

No exemplo 17, o trecho "concentração da amostra de clorofórmio" contém dois adjuntos adnominais facilmente identificáveis (pois são substantivos modificando substantivos[19]) - "amostra" e "clorofórmio". A dificuldade - para as máquinas - porém, reside na distinção entre a estrutura [concentração [da amostra [de clorofórmio]]] e [concentração [da amostra] [de clorofórmio]].

[18] O inventário das relações sintáticas está em https://universaldependencies.org/u/dep/index.html, e deve-se notar que a convenção do projeto (também utilizada neste trabalho) é empregar as etiquetas em caixa alta quando são etiquetas de POS, e em caixa baixa quando são etiquetas de relação sintática.

19 Além de serem substantivos modificando substantivos, deve-se notar também que UD não distingue adjuntos adnominais de complementos nominais, tornando ainda mais fácil atribuir a etiqueta de relação correta.

17. De esse modo , a concentração de a **amostra** de **clorofórmio** corresponde a a concentração de óleo em a água.

Do outro lado, temos as classes cujo tipo de relação sintática foi atribuído erroneamente, não adiantando analisar o governante da relação, já que um erro na identificação da classe frequentemente resulta em erro de encaixe. Em ordem decrescente de dificuldade, as etiquetas são: obl:arg, parataxis, ccomp, csubj, xcomp, aux e obl.

4.1 Análise de erros

Para uma tentativa de compreensão linguística dos erros produzidos pela análise automática, analisamos todos os erros encontrados na avaliação intrínseca (isto é, na partição teste) das seis classes com o pior desempenho (parataxis, appos, acl:relcl, conj, obl:arg e ccomp) para o analisador automático segundo as métricas utilizadas, totalizando 250 erros de relação sintática e/ou encaixe de dependência. Classificamos os erros segundo a tipologia descrita na seção 4.3 e os resultados se encontram no gráfico da figura 1 e na tabela 5.

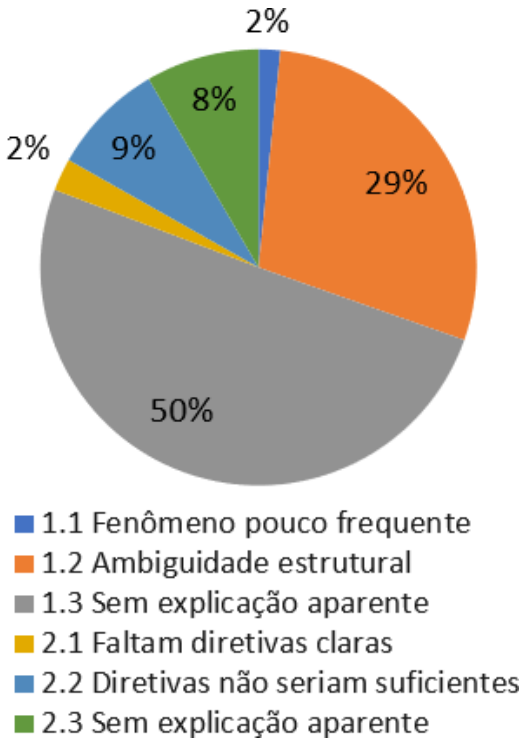

Figura 1 – Distribuição dos tipos de erros encontrado
Fonte: Elaborado pelos autores

REL	1.1	1.2	1.3	2.1	2.2	2.3	Total
acl:relcl	1	2	23	0	1	3	30
appos	0	16	4	0	5	0	25
ccomp	3	0	5	0	0	2	10
conj	0	28	78	3	11	14	134
obl:arg	0	24	2	0	1	1	28
parataxis	0	2	14	3	3	1	23
Total	4	72	126	6	21	21	250
	202			48			

Figura 1 – Classificação dos erros com pior desempenho
Fonte: Elaborado pelos autores

A grande maioria dos erros encontrados são erros do analisador automático (80,8%). Desses, poucos são derivados de fenômenos linguísticos de baixa frequência no *corpus* (1,9%, tipo 1.1), portanto não se pode dizer, a partir dessa análise, que providenciar mais exemplos dos fenômenos linguísticos difíceis seria uma boa estratégia para diminuir a quantidade de erros. Se, por um lado, resolver ambiguidades estruturais que necessitem conjugar diferentes conhecimentos tem sido um desafio, generalizar conhecimento a partir de estruturas que ocorrem pouco já é uma realidade.

Um grande número de erros do analisador automático (35,6%) são erros de ambiguidade estrutural (1.2), quando um anotador realiza o encaixe de dependências sem de fato compreender a frase, pois ligou duas palavras que, sintaticamente, podem ser unidas, mas na prática a união entre as duas não é a correta. Nesses casos, um humano saberia identificar qual a árvore sintática correta fazendo uso de diversos conhecimentos, de que, parece, que o algoritmo ainda não dispõe. Para esses erros, soluções que passem por simplificar a anotação dessas construções no *corpus* - como, por exemplo, definir que, em casos difíceis, o núcleo será sempre o *token* mais próximo - embora possam alavancar os dados de avaliação intrínseca, tornariam a análise linguística artificial, pouco informativa e, muitas vezes, simplesmente errada. O desenvolvimento de tecnologias de anotação que sejam informadas semanticamente sobre o conteúdo das palavras, seja essa informação obtida de conhecimento linguístico ou de métodos estatísticos, pode funcionar para alavancar o número de acertos.

A grande quantidade de erros do analisador automático injustificados (1.3), que corresponde a 62,3% do número de erros do analisador automático, pode estar diretamente relacionada à forma como foi desenvolvido o algoritmo utilizado pelo UDPipe, não se reproduzindo em outros sistemas. Contrastar esses erros com os de outros analisadores treinados no mesmo material pode ser elucidativo sobre aspectos do *corpus* para os quais um ou outro sistema esteja dando maior ênfase, durante o seu processo de treinamento, pois não sendo erros derivados nem da composição do *corpus* nem da estrutura linguística em si, podem ter sua origem na forma como o algoritmo foi construído.

Os erros que não são do analisador automático somam 19,2% do total de erros. 8,4% são erros injustificados do padrão ouro (erros do tipo 2.3), os quais não foram detectados pelos anotadores humanos ou pelos métodos de revisão utilizados, mas que já foram corrigidos para a nova versão do PetroGold.

Outros 8,4% dos erros são derivados de fenômenos linguísticos complexos ou frases cuja interpretação não é trivial (2.2). Não estamos incluindo essa classe de erros no número de erros do padrão ouro, pois não se pode afirmar que nem o anotador automático nem o anotador humano acertou a análise devido à dificuldade de julgar qual a interpretação correta[20].

Por fim, 2,4% dos erros são derivados da falta de diretivas claras sobre algum fenômeno linguístico ou estrutura do *corpus* (2.1). Nesses casos, também não se pode afirmar quem acertou ou errou, mas sabe-se que houve uma falha humana uma vez que se apresentaram lacunas na documentação da anotação linguística do *corpus*, motivo pelo qual incluímos o número no somatório de erros do padrão ouro.

Como conclusão, vimos que 10,8% dos erros da avaliação intrínseca são erros humanos. Como comparação, Manning (2011) encontrou uma taxa de 55,5% de erros no padrão ouro, analisando uma amostra de 100 erros apenas de classes gramaticais, na seção do *Wall Street Journal do Penn Treebank* (TAYLOR; MARCUS; SANTORINI, 2003). Manning (2011) conclui que mais da metade dos erros da anotação de POS, portanto, deve ser solucionada melhorando a qualidade da anotação do *corpus*, e não dos modelos de aprendizagem apenas. Nosso estudo e o de Manning (2011) não são facilmente comparáveis - o autor pretendia investigar o que seria necessário para melhorar a anotação automática de POS, que já alcançava um alto índice de acertos (97%), enquanto que o nosso objetivo é entender melhor os números de avaliação sintática de um treebank, onde ainda há muito espaço para melhora.

Dadas as circunstâncias, porém, diferentemente do que concluiu o autor para POS, a melhor saída para melhorar os números de avaliação sintática para um *treebank* padrão ouro como o Petro-Gold passaria pela melhoria dos algoritmos, e não do *corpus* em si, o que pode se justificar pelo fato de que o *corpus* já foi alvo de intensas etapas de revisão da sua anotação linguística e, como vimos, poucos dos erros de avaliação intrínseca parecem ser fruto da anotação do padrão ouro ou da falta de exemplos sobre fenômenos linguísticos específicos no *corpus*.

Considerações finais

Relatamos aqui uma avaliação qualitativa da anotação automática de um *treebank* dependencial, visando uma análise linguística sobre o resultado desta anotação. Nossa intenção é não apenas lançar alguma luz sobre o que acontece, linguisticamente, quando temos uma anotação automática, mas também explorar possibilidades de contribuições linguísticas para a melhoria dos resultados dos algoritmos.

Os resultados sugerem que, do ponto de vista linguístico, temos pouco a fazer. A minoria dos erros cometidos pelo analisador automático analisado tem origem na baixa ocorrência de exemplos dos fenômenos linguísticos no *corpus* ou na falta de diretivas de anotação claras, e a maioria dos erros linguísticos (apenas 9% do total de erros encontrados) diz respeito a casos cuja ambiguidade inerente à linguagem impede a definição de apenas uma anotação correta. Do ponto de vista

[20] Buscamos sempre minimizar a quantidade de casos como esses durante a anotação de um *corpus*, mas a existência de ocorrências de difícil interpretação é inerente à natureza da tarefa de categorização (SAMPSON; BABARCZY, 2008).

computacional, a imensa maioria dos erros (50% do total) não tem explicação aparente, não sendo tampouco explicável por ambiguidade estrutural.

Embora fuja ao escopo estritamente linguístico, uma pergunta que nos fizemos durante a análise de dados diz respeito à quantidade de informação linguística necessária para a discriminação correta das classes. Em outras palavras, gostaríamos de algum parâmetro relativo a quantos exemplos precisamos para que haja um aprendizado efetivo, o que pode ser investigado em trabalhos futuros.

Outras lacunas a serem exploradas no futuro se relacionam ao fato de que, neste trabalho, utilizamos um modelo gerado por apenas um sistema, não sendo possível comparar a qualidade e os resultados linguísticos gerados por diferentes arquiteturas de aprendizado de máquina. Assim, uma seara a ser explorada é a diversificação desses sistemas – observando os *outputs* linguísticos que eles geram – e a tentativa de descrever melhor os erros do analisador automático que, por ora, classificamos como "sem explicação aparente".

Agradecimentos

Elvis de Souza agradece ao Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) e à Fundação de Amparo à Pesquisa do Estado do Rio de Janeiro (FAPERJ) pelas bolsas de mestrado de nº, respectivamente, 130495/2021-2 e 202.433/2022.

Informações complementares

Avaliação e resposta dos autores

Avaliação: https://doi.org/10.25189/rabralin.v22i2.2114.R

Editores

Jorge Manuel Evangelista Baptista
Afiliação: Universidade do Algarve
ORCID: https://orcid.org/0000-0003-4603-4364

Adriana Silvina Pagano
Afiliação: Universidade Federal de Minas Gerais
ORCID: https://orcid.org/0000-0002-3150-3503

Marta Deysiane Alves Faria Sousa
Afiliação: Universidade Federal de Sergipe
ORCID: https://orcid.org/0000-0002-0480-0422

RODADAS DE AVALIAÇÃO

Avaliador 1: Roana Rodrigues
Afiliação: Universidade Federal de Sergipe
ORCID: https://orcid.org/0000-0002-7748-8716

Avaliador 2: Alisson Hudson Veras Lima
Afiliação: Instituto Federal de Alagoas
ORCID: https://orcid.org/0000-0001-6597-6547

AVALIADOR 1

No artigo intitulado "Avaliação da anotação automática de dependências sintáticas", os autores apresentam a avaliação qualitativa de um treebank para a língua portuguesa anotado sintaticamente de acordo com as diretrizes da UD. Para a discussão, são analisadas a anotação humana e a anotação automática, com o intuito de verificar pontos em que a anotação pode melhorar, a partir da identificação de construções linguísticas específicas que causam dificuldade no processo de anotação (humano e automático).

Os resultados demonstram que, dentre as classes mais difíceis, a maior parte dos erros decorrem da anotação automática sem explicação aparente. Segundo os autores, isso, somado a outros aspectos, aponta para a conclusão de que a anotação humana analisada é consistente e a anotação automática apresenta desvios que não condizem com o input linguístico, o que sugere menor espaço para melhorias linguísticas.

O artigo é conciso e os procedimentos metodológicos são precisos e bem descritos. Salienta-se, no entanto, a necessidade de uma revisão textual do artigo, pois há alguns lapsos de concordância verbal e nominal e de ortografia ao longo do trabalho. Além disso, é preciso revisar a formatação do texto, sobretudo o local de inserção das tabelas e figura/gráfico. Como trabalhos futuros, os autores poderiam se dedicar, além da análise da anotação de pontuação (já em andamento), na verificação da saída de outros analisadores automáticos treinados com o mesmo material, conforme mencionado no próprio artigo. Como sugestão final, os autores podem considerar:

- Em "Introdução", especificar/exemplificar a maneira como as dependências sintáticas impactam nas tarefas de PLN.
- Em "2 O corpus e a anotação", poderia haver um exemplo "concreto" sobre as limitações de anotação de *aposto* segundo as diretrizes da UD – da mesma maneira em que houve

um exemplo para o caso da anotação de *corra* como verbo, mesmo se referindo a um *nome*.

Reitera-se a relevância do trabalho para os estudos linguístico-computacionais e para os avanços de trabalhos em UD no Brasil.

AVALIADOR 2

Parabenizo os autores pelo excelente relato de pesquisa e encaminho, em anexo, algumas considerações feitas por mim e que, se forem aceitas, podem ajudar o trabalho a ficar ainda mais enriquecido. As sugestões foram feitas para cada parte do texto, indicando o que ajudaria na leitura da versão final. Algumas questões tangem a questões de referência aos trabalhos que fazem com quem nasça a pesquisa e outras tratam de questões de cunho de escrita, tanto de escrita acadêmica quanto de disposição de elementos textuais. Sugiro que as alterações sejam aceitas e que o texto seja modificado para publicação, uma vez que se trata de uma contribuição relevante para a área de Processamento de Linguagem Natural.

Conflito de Interesse

O autor e a autora não têm conflitos de interesse a declarar.

Protocolo e Pré-Registro de Pesquisa

Avaliando os roteiros propostos pela Equator Network, consideramos que nenhum deles se mostra relevante para a pesquisa em tela. Também informamos que a pesquisa desenvolvida não foi pré-registrada em repositório institucional independente.

Declaração de Disponibilidade de Dados

O compartilhamento de dados não é aplicável a este artigo, pois nenhum dado novo foi criado ou analisado neste estudo.

REFERÊNCIAS

AFONSO, Susana. **Avaliação do grau de concordância entre anotadores:** análise e discussão dos resultados do processo de re-revisão, 2004.

ARTSTEIN, Ron. Inter-annotator agreement. In: **HANDBOOK of linguistic annotation**. [S.l.]: Springer, 2017. P. 297–313.

BAIA, Jardel; PRATES, Arley; CLARO, Daniela. CoNLL Dependency Parser: Extrinsic Evaluation through the Open Information Extraction task. In: **SBC. ANAIS do VIII Symposium on Knowledge Discovery, Mining and Learning.** [S.l.: s.n.], 2020. P. 193–200.

BICK, Eckhard et al. Floresta Sintá (c) tica: Ficção ou realidade. In: **AVALIAÇÃO Conjunta, Um novo paradigma no processamento computacional da língua portuguesa.** [S.l.]: IST Press, 2007. P. 291–300.

CAVALCANTI, Tatiana *et al.* Os limites da palavra e da sentença no processamento automático de textos. **Revista Brasileira de Iniciação Científica**, v. 8, e021033–e021033, 2021.

CORDEIRO, Fábio Corrêa. Petrolês-como construir um corpus especializado em óleo e gás em português. **PUC-Rio, Rio de Janeiro, RJ-Brasil: PUC-Rio**, 2020.

DE MARNEFFE, Marie-Catherine *et al.* Universal dependencies. **Computational linguistics**, MIT Press One Rogers Street, Cambridge, MA 02142-1209, USA journals-info . . ., v. 47, n. 2, p. 255–308, 2021.

DE SOUZA, Elvis. **Construção e avaliação de um treebank padrão ouro.** 2023. Mestrado – PUC-Rio.

DE SOUZA, Elvis; CAVALCANTI, Tatiana *et al.* Diretivas e documentação de anotação UD em português (e para língua portuguesa), 2020.

DE SOUZA, Elvis; FREITAS, Cláudia. ET: A Workstation for Querying, Editing and Evaluating Annotated Corpora. In: **PROCEEDINGS of the 2021 Conference on Empirical Methods in Natural Language Processing:** System Demonstrations. Online e Punta Cana, Dominican Republic: Association for Computational Linguistics, nov.2021. P. 35–41. DOI: 10.18653/v1/2021.emnlp-demo.5. Disponível em: https://aclanthology.org/2021.emnlp-demo.5.

DE SOUZA, Elvis; FREITAS, Cláudia. Polishing the gold–how much revision do we need in treebanks? In: **PROCEEDINGS of the Universal Dependencies Brazilian Festival.** [S.l.: s.n.], 2022. P. 1–11.

DE SOUZA, Elvis; SILVEIRA, Aline *et al.* **Documentação da anotação morfossintática do PetroGold**, 2021.

DE SOUZA, Elvis; SILVEIRA, Aline *et al.* PetroGold–Corpus padrão ouro para o domınio do petróleo. In: **SBC. ANAIS do XIII Simpósio Brasileiro de Tecnologia da Informaçãoo e da Linguagem Humana**. [S.l.: s.n.], 2021. P. 29–38.

FREITAS, Cláudia. **Linguística Computacional.** [S.l.]: Editora Parábola, 2022.

FREITAS, Cláudia; DE SOUZA, Elvis. A study on methods for revising dependency treebanks: in search of gold. **Language Resources and Evaluation**, Springer, p. 1–21, 2023.

FREITAS, Cláudia; TRUGO, Luiza F. *et al.* Tagsets and datasets: some experiments based on Portuguese language. In: **Computational Processing of the Portuguese Language: 13th International Conference, PROPOR 2018, Canela, Brazil, September 24–26, 2018, Proceedings 13**. Springer International Publishing, 2018. p. 459-469.

MANNING, Christopher D. Part-of-speech tagging from 97% to 100%: is it time for some linguistics?. In: **International conference on intelligent text processing and computational linguistics**. Berlin, Heidelberg: Springer Berlin Heidelberg, 2011. p. 171-189.

NIVRE, Joakim; FANG, Chiao-Ting. Universal dependency evaluation. In: **Proceedings of the NoDaLiDa 2017 Workshop on Universal Dependencies (UDW 2017)**. 2017. p. 86-95.

OLIVEIRA, Claudia et al. A set of np-extraction rules for portuguese: Defining, learning and pruning. In: **Computational Processing of the Portuguese Language: 7th International Workshop, PROPOR 2006, Itatiaia, Brazil, May 13-17, 2006. Proceedings 7**. Springer Berlin Heidelberg, 2006. p. 150-159.

RADEMAKER, Alexandre et al. Universal dependencies for Portuguese. In: **PROCEEDINGS of the Fourth International Conference on Dependency Linguistics (Depling 2017).** [S.l.: s.n.], 2017. P. 197–206.

SAMPSON, Geoffrey; BABARCZY, Anna. Definitional and human constraints on structural annotation of English. **Natural Language Engineering**, Cambridge University Press, v. 14, n. 4, p. 471–494, 2008.

STRAKA, Milan; HAJIC, Jan; STRAKOVÁ, Jana. UDPipe: trainable pipeline for processing CoNLL-U files performing tokenization, morphological analysis, pos tagging and parsing. In: **Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)**. 2016. p. 4290-4297.

TAYLOR, Ann; MARCUS, Mitchell; SANTORINI, Beatrice. The Penn treebank: an overview. **Treebanks:** Building and using parsed corpora, p. 5-22, 2003.

ZEMAN, Daniel et al. CoNLL 2018 shared task: Multilingual parsing from raw text to universal dependencies. In: **Proceedings of the CoNLL 2018 Shared Task: Multilingual parsing from raw text to universal dependencies**. 2018. p. 1-21.