Pontifícia Universidade Católica
do Rio de Janeiro

Elvis Alves de Souza

Construção e avaliação de um treebank padrão ouro

Dissertação de Mestrado

Dissertação apresentada como requisito parcial para obtenção do grau de Mestre em Letras/Estudos da Linguagem pelo Programa de Pós-graduação em Estudos da Linguagem da PUC-Rio.

Orientadora: Maria Cláudia de Freitas

Rio de Janeiro
Abril 2023

Pontifícia Universidade Católica
do Rio de Janeiro

Elvis Alves de Souza

Construção e avaliação de um treebank padrão ouro

Dissertação apresentada como requisito parcial para obtenção do grau de Mestre pelo Programa de Pós-graduação em Estudos da Linguagem da PUC-Rio. Aprovada pela Comissão Examinadora abaixo:

Maria Cláudia de Freitas
Orientadora
Departamento de Letras – PUC-Rio

Adriana Silvina Pagano
UFMG

Thiago Alexandre Salgueiro Pardo
USP

Diana Maria de Sousa Marques Pinto dos Santos
University of Oslo

Rio de Janeiro, 17 de abril de 2023.

Ficha Catalográfica

de Souza, Elvis Alves

Construção e avaliação de um treebank padrão ouro / Elvis Alves de Souza ; orientadora: Maria Cláudia de Freitas. – 2023.
193 f. : il. color. ; 30 cm

Dissertação (mestrado)–Pontifícia Universidade Católica do Rio de Janeiro, Departamento de Letras, 2023.
Inclui bibliografia

1. Letras – Teses. 2. Processamento de linguagem natural. 3. Linguística computacional. 4. Treebanks. 5. Anotação de corpus. 6. Descrição do português. I. Freitas, Maria Cláudia de. II. Pontifícia Universidade Católica do Rio de Janeiro. Departamento de Letras. III. Título.

CDD: 400

Para todas as professoras que me inspiraram,
de todas as formas possíveis.

Agradecimentos

Aos meus pais, por me permitirem seguir estudando e, assim, chegar mais longe do que poderiam sonhar.

À minha irmã, pelo apoio incondicional nesses dois anos turbulentos.

À Aline Silveira e à Ana Carolina Albuquerque, amigas que a PUC me deu para a vida inteira, pela companhia e pelas ricas discussões sobre tudo.

Às parceiras de projeto, Tatiana Cavalcanti, Maria Clara Castro e Wograine Evelyn, pelas contribuições e pelos momentos de descontração.

Aos colegas de laboratório do ICA, pelo suporte computacional em diversos momentos.

A todos os professores e colegas do PPGEL e da graduação em Letras da PUC-Rio, parte fundamental da minha formação.

À minha orientadora, professora Cláudia Freitas, que desde antes de iniciar minha graduação na PUC-Rio, pelo destino ou acaso, era mais uma que me incentivava a entrar para o mundo das Letras (e, depois, dos computadores também).

Ao CNPq, à FAPERJ, à ANP e à PUC-Rio pelos auxílios concedidos, sem os quais este trabalho não poderia ter sido realizado.

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Código de Financiamento 001

Resumo

de Souza, Elvis Alves; Freitas, Cláudia (Orientadora). **Construção e avaliação de um treebank padrão ouro**. Rio de Janeiro, 2023. 193p. Dissertação de Mestrado – Departamento de Letras, Pontifícia Universidade Católica do Rio de Janeiro.

Esta dissertação apresenta o processo de desenvolvimento do PetroGold, um *corpus* anotado com informação morfossintática – um *treebank* – padrão ouro para o domínio do petróleo. O desenvolvimento do recurso é abordado sob duas lentes: do lado linguístico, estudamos a literatura gramatical e tomamos decisões linguisticamente motivadas para garantir a qualidade da anotação do *corpus*; do lado computacional, avaliamos o recurso considerando a sua utilidade para o processamento de linguagem natural (PLN). Recursos como o PetroGold recebem relevância especial no contexto atual, em que o PLN estatístico tem se beneficiado de recursos padrão ouro de domínios específicos para alimentar o aprendizado automático. No entanto, o *treebank* é útil também para tarefas como a avaliação de sistemas de anotação baseados em regras e para os estudos linguísticos. O PetroGold foi anotado segundo as diretivas do projeto *Universal Dependencies*, tendo como pressupostos a ideia de que a anotação de um *corpus* é um processo interpretativo, por um lado, e utilizando o paradigma da linguística empírica, por outro. Além de descrever a anotação propriamente, aplicamos alguns métodos para encontrar erros na anotação de *treebanks* e apresentamos uma ferramenta criada especificamente para busca, edição e avaliação de *corpora* anotados. Por fim, avaliamos o impacto da revisão de cada uma das categorias linguísticas do *treebank* no aprendizado automático de um modelo alimentado pelo PetroGold e disponibilizamos publicamente a terceira versão do *corpus*, a qual, quando submetida à avaliação intrínseca de um modelo, alcança métricas até 2,55% melhores que a versão anterior.

Palavras-chave

Processamento de Linguagem Natural; Linguística Computacional; treebanks; anotação de corpus; descrição do português.

Abstract

de Souza, Elvis Alves; Freitas, Cláudia (Advisor). **Building and evaluating a gold-standard treebank**. Rio de Janeiro, 2023. 193p. Dissertação de Mestrado – Departamento de Letras, Pontifícia Universidade Católica do Rio de Janeiro.

This thesis reports on the development process of PetroGold, a gold-standard annotated *corpus* with morphosyntactic information – a treebank – for the oil and gas domain. The development of the resource is seen from two perspectives: on the linguistic side, we study the grammatical literature and make linguistically motivated decisions to ensure the quality of *corpus* annotation; on the computational side, we evaluate the resource considering its usefulness for natural language processing (NLP). Resources like PetroGold receive special importance in the current context, where statistical NLP has benefited from domain-specific gold-standard resources to train machine learning models. However, the treebank is also useful for tasks such as evaluating rule-based annotation systems and for linguistic studies. PetroGold was annotated according to the guidelines of the Universal Dependencies project, having as theoretical assumptions the idea that the annotation of a *corpus* is an interpretative process, on the one hand, and using the empirical linguistics paradigm, on the other. In addition to describing the annotation itself, we apply some methods to find errors in the annotation of treebanks and present a tool created specifically for searching, editing and evaluating annotated *corpora*. Finally, we evaluate the impact of revising each of the treebank linguistic categories on the automatic learning of a model powered by PetroGold and make the third version of the *corpus* publicly available, which, when performing an intrinsic evaluation for a model using the *corpus*, achieves metrics up to 2.55% better than the previous version.

Keywords

Natural Language Processing; Computational Linguistics; treebanks; corpus annotation; Portuguese description.

Sumário

Lista de figuras

Lista de tabelas

1
Apresentação

1.1
Motivação e objetivos

Esta dissertação apresenta o processo de desenvolvimento de um *corpus* anotado com informação morfossintática – um *treebank* – padrão ouro, discute teoricamente as decisões linguísticas tomadas durante a sua construção, e o avalia considerando sua utilidade para o Processamento de Linguagem Natural (PLN).

O treebank PetroGold (DE SOUZA et al., 2021b; DE SOUZA; FREITAS, 2022a) integra o projeto Petrolês, uma iniciativa de colaboração interinstitucional liderada pelo Centro de Pesquisas e Desenvolvimento da Petrobras (CENPES), em parceria com PUC-Rio, UFRGS e PUC-RS[1]. Além de um projeto, o Petrolês é ainda um repositório de recursos para o PLN em língua portuguesa do domínio do petróleo, o que inclui modelos de anotação sintática, anotação semântica, modelos de palavras vetorizadas (GOMES et al., 2021), bem como ontologias e outros.

Um dos objetivos do projeto é desenvolver recursos que subsidiem a construção de um sistema de buscas semanticamente orientado para o domínio do petróleo. Para a construção de sistemas como esse, Cleverley e Burnett (2015) concluem que métodos linguísticos e métodos estatísticos são complementares, e quando utilizados em conjunto são capazes de aumentar a taxa de acerto das buscas e facilitar a descoberta de novas informações por meio das interfaces de busca. No entanto, a aplicação de métodos linguísticos necessita de subsídios (também linguísticos) que os viabilizem. Para subsidiar a construção desse sistema, o presente trabalho tem como objetivo principal, em termos específicos, a construção de um treebank padrão ouro.

O PetroGold é um treebank que deve se caracterizar pela sua anotação de qualidade. O recurso será empregado tanto no treinamento de sistemas de aprendizado de máquina quanto na avaliação de sistemas de anotação automática, contribuindo para a pesquisa no âmbito do Petrolês por um lado, e, por outro, para o PLN em língua portuguesa como um todo.

A construção do treebank se justifica pelo fato de que embora para línguas como o inglês, com abundância de dados textuais anotados para o PLN, e para alguns domínios com maior tradição nos estudos linguísticos, como o

[1]Mais informações sobre o projeto estão disponíveis no endereço: <https://petroles.puc-rio.ai>. Acesso em 10 jan. 2023.

domínio jornalístico, os resultados de tarefas básicas de PLN sejam em grande parte satisfatórios, quando há diferenças entre o tipo de texto que alimenta o treinamento dos algoritmos e o tipo de texto onde eles são aplicados, os resultados do processamento automático pioram consideravelmente. Enquanto um anotador de classes gramaticais obtém desempenho de até 97,96% (BOHNET et al., 2018) na anotação de um corpus de textos jornalísticos (a seção de artigos do Wall Street Journal do Penn Treebank), a mesma tarefa vê seu desempenho cair para 83,5% em um corpus de resumos de artigos científicos da área da biomedicina (THOMPSON; ANANIADOU; TSUJII, 2017). Nesse contexto, uma solução para o processamento de textos em um domínio específico envolve preparar um corpus desse mesmo domínio para que os sistemas sejam capazes de generalizar informação linguística tendo como fonte textos com características semelhantes.[2]

Os últimos anos viram um aumento na quantidade de corpora anotados em língua portuguesa, desde a criação da Floresta Sintá(c)tica (AFONSO et al., 2002; FREITAS; ROCHA; BICK, 2008), a Coleção Dourada do HAREM (FREITAS et al., 2010), o corpus Summit++ (FONSECA et al., 2016) e o PropBank-Br (DURAN; ALUÍSIO, 2012). Contudo, ainda são poucos os treebanks padrão ouro, sobretudo para domínios (ou áreas) específicos e, no limite, cada domínio pode necessitar de recursos próprios.

Construir um corpus com as características do PetroGold envolve discussões linguísticas e decisões metodológicas importantes, como a seleção e o pré-processamento dos documentos que irão compor o corpus e que são relevantes para o projeto, além das decisões relativas à anotação linguística do material, que deve ser projetada para permitir bons resultados na análise morfossintática automática em português, especialmente em textos do domínio do petróleo, sem descuidar da correção das análises do ponto de vista linguístico.

Para isso, temos como objetivos secundários aplicar e avaliar metodologias para a revisão da anotação de corpora, tendo como horizonte a adequação dessa anotação ao PLN. Ao mesmo tempo, desenvolvemos e disponibilizamos algumas ferramentas que foram sendo construídas durante o processo de revisão do corpus e uma documentação linguística detalhada acerca de fenômenos que encontramos nos textos que compõem o corpus, indicando como a sua anotação foi realizada no corpus com as justificativas correspondentes.

Nesse contexto, este trabalho busca responder a algumas perguntas:

[2]Recentemente, porém, ao realizar experimentos com três corpora em português Brasileiro anotados conforme diretivas do mesmo projeto de anotação e de gêneros textuais diferentes – entre os quais o gênero acadêmico representado pelo PetroGold –, Silva (2023) sugere que o gênero textual e o domínio específico não são assim tão determinantes na qualidade do aprendizado de anotação morfológica, desde que o formalismo e a gramática da anotação sejam os mesmos.

1. Sendo uma tarefa de interpretação, quais os limites da anotação linguística, ainda que condicionada à tradição gramatical e às diretivas de um projeto de anotação?

2. Quais os fenômenos linguísticos mais complexos para o aprendizado automático?

3. Qual o impacto da *modelagem* da anotação linguística nos resultados do treinamento de anotadores automáticos?

4. Qual o impacto da *revisão* da anotação linguística nos resultados do treinamento de anotadores automáticos?

5. Quais os métodos de revisão de anotação mais eficazes na construção de um treebank padrão ouro?

Como o Petrolês é um projeto cujo horizonte é a construção de um sistema de buscas semanticamente orientado para o domínio do petróleo, modelamos as etapas do projeto de maneira que tenhamos como resultado o desenvolvimento de recursos que facilitem: (1) o pré-processamento de boletins, relatórios técnicos e documentos acadêmicos – os gêneros textuais presentes no Petrolês –, o que envolve a segmentação desses textos em frases e palavras; (2) a análise automática de morfossintaxe nesses textos, e (3) a anotação semântica de entidades relevantes para o projeto. O foco desta dissertação está na etapa (2) – pesquisar as melhores práticas para a construção de um treebank para o PLN, o PetroGold, composto apenas por teses e dissertações do domínio do petróleo.

1.2 Um recurso para o PLN

O campo de que este estudo faz parte recebeu muitos nomes ao longo do tempo: Processamento ("estatístico" ou "automático") de Linguagem Natural (PLN), Linguística Computacional (LC), Engenharia da Linguagem entre outros (FREITAS, 2022). Há implicações práticas em dizer que o PetroGold é um recurso desenvolvido *para o PLN* – e não para a linguística, por exemplo.

O cientista da computação Kay (2005 apud FREITAS, 2022) atribui o surgimento do nome "Linguística Computacional" para a área a uma necessidade de se estabelecer um campo que pudesse ser tomado como científico, capaz de receber recursos financeiros para pesquisa. A este objetivo se sucedeu, à época (na década de 1950), uma valorização do caráter acadêmico da área em detrimento do lado aplicado. Havia quem defendesse, como o próprio Martin Kay, que modelos de aprendizado de máquina sem conhecimento

linguístico codificado fossem uma "aberração". O cientista se referia ao paradigma do PLN estatístico, em que as tarefas de linguagem são realizadas com conhecimentos linguísticos derivados dos dados estatisticamente, sem que seja necessário explicitá-los *a priori*.

Adeptos da ideia de que o PLN (ou LC) deveria ser mais acadêmico/teórico e menos técnico/aplicado julgavam que aplicações de cunho estatístico não compreendiam de fato uma língua, apenas reproduziam padrões por meio de "truques" textuais. Do outro lado, entusiastas do PLN estatístico julgavam que o PLN baseado em conhecimento linguístico codificado era frágil uma vez que as regras linguísticas não poderiam jamais ter granularidade o suficiente para contemplar todos os usos linguísticos do mundo real, sendo útil apenas para testar teorias linguísticas, mas de pouca eficácia para as aplicações do PLN mais comuns, como tradução automática, sumarização de textos e extração de informação (FREITAS, 2022).

Como nota Freitas (2022), embora houvesse inicialmente uma diferenciação entre o que se fazia no PLN e na LC, hoje já não é mais possível delimitar tão rigidamente a prática linguística e a prática computacional no PLN. Nesse contexto, dizer que o PetroGold é um recurso desenvolvido para o PLN significa apenas que, ao tomar as decisões cabíveis para anotá-lo da melhor forma possível, o principal critério que utilizamos foi a sua adequação ao processamento automático do português. Não abrimos mão de interpretações coerentes durante a anotação linguística do corpus, mas quando confrontados com impasses, o que se sobressai é o objetivo de categorizar os fenômenos linguísticos de maneira que facilite a generalização. Em resumo, embora o treebank seja útil também aos estudos linguísticos, deixamos claro desde já que o seu objetivo é subsidiar tarefas subsequentes à anotação morfossintática automática, no âmbito do projeto Petrolês, e isso se reflete em algumas escolhas de anotação e na forma como o avaliamos.

Notadas as possíveis diferenças entre as denominações utilizadas para se referir a esta "área que mistura Linguística e Computação" (FREITAS, 2022), observo que durante esta dissertação o nome PLN será utilizado indistintamente para se referir ao campo do qual o estudo faz parte, seja para se referir às questões linguísticas, seja para se referir às questões computacionais, ambas presentes e de igual relevância neste trabalho.

1.3 Estrutura do trabalho

O trabalho se divide da seguinte forma: no segundo capítulo, serão apresentados os pressupostos teóricos da pesquisa. Primeiro, contextualizo o

trabalho em um breve histórico do PLN. Em seguida, coloco em discussão a prática da anotação linguística. Então, discuto a visão de linguagem (ou de trabalho com a linguagem) que guia esta dissertação, chamada linguística empírica.

No terceiro capítulo, realizo uma revisão da literatura (1) sobre os treebanks já disponíveis, (2) sobre as diferentes formas de se revisar a anotação linguística nesses recursos, e (3) sobre as formas de avaliar a qualidade de treebanks. No quarto capítulo, então, apresento de fato o percurso utilizado para desenvolver o PetroGold, isto é, a metodologia da pesquisa – (i) que ferramentas e procedimentos utilizamos para anotar o treebank, (ii) quais métodos de revisão aplicamos para corrigir a anotação automática e para adequá-la aos nossos objetivos, e (iii) que métodos empregamos para avaliar a qualidade do recurso.

No quinto capítulo, apresento os detalhes sobre a construção do PetroGold, o que envolve uma descrição do recurso e das suas diferentes versões de desenvolvimento. Na seção de questões linguísticas, realizo estudos sobre três fenômenos muito presentes no corpus – argumentos verbais introduzidos por preposição, expressões multipalavras e o pronome "se" – cuja anotação não seria possível sem antes realizar estudos descritivos e consultar recursos como léxicos, gramáticas e documentações de outros projetos. Por fim, coloco em perspectiva todos os métodos de revisão aplicados na construção do corpus, indicando quais foram os mais eficientes.

No sexto capítulo, realizo a avaliação do corpus do ponto de vista computacional, isto é, em que medida a anotação (e revisão) das categorias linguísticas propicia um aprendizado automático de qualidade. Em um primeiro momento, realizo a avaliação intrínseca de uma versão do PetroGold anterior a esta dissertação (a versão 2), para encontrar as categorias mais difíceis para o processamento automático e discutir possíveis motivos que possam ter resultado na dificuldade. Em seguida, avalio também por meio de avaliação intrínseca o impacto no aprendizado automático das correções das três questões linguísticas discutidas. Por fim, discuto as características e a avaliação do PetroGold v3, resultado de todas as discussões presentes nesta dissertação.

Em relação às questões linguísticas, o sexto capítulo funciona como uma contraparte do quinto capítulo. No capítulo quinto, as decisões são tomadas, as mudanças são implementadas e os resultados são discutidos do ponto de vista linguístico. No capítulo sexto, são avaliados os resultados computacionais das modificações realizadas e apresento e discuto o que um modelo de anotação automática aprendeu, bem como os fenômenos linguísticos em que ainda não obtém bom desempenho.

No sétimo capítulo, reservado às considerações finais, faço uma última avaliação de quais as contribuições deste trabalho. Nesse momento, considero também algumas das lacunas no estudo que não puderam ser preenchidas por restrições de tempo e indico alguns caminhos para o futuro a partir dos resultados obtidos e dos recursos desenvolvidos.

2
Pressupostos teóricos

2.1
Contextualização

Na revisão histórica que a cientista da computação Karen Sparck Jones realiza, há uma divisão entre quatro fases do PLN (JONES, 1994). Essa divisão se dá menos por questões filosóficas – relativas às crenças de como as línguas funcionam[1] –, e mais por questões práticas e temporais – como a disponibilidade de computadores potentes o suficiente para processar frases em um tempo razoável (há relatos de computadores dos anos 1960 que demoravam 7 minutos para processar uma única frase) e mesmo recursos adequados para o processamento de uma língua, como dicionários que pudessem ser lidos por computadores e teorias que fossem adequadas aos dados da língua em uso.

Entre o final dos anos 1940 e o final dos anos 1960, no que a cientista chamou de primeira fase do PLN, os esforços estavam principalmente voltados para a tradução automática. No início, as aplicações eram feitas por meio de buscas em dicionários palavra a palavra, de maneira que logo os cientistas esbarraram nos problemas da ambiguidade sintática, semântica e, enfim, em todos os obstáculos derivados do fato de que a língua é flexível. O início do PLN tinha como objetivo uma das tarefas mais complexas – a tradução automática –, que logo foi desencorajada pelo relatório de 1966 do ALPAC (*Automatic Language Processing Advisory Committee*), que minou o financiamento para as pesquisas em tradução automática tendo em vista que, segundo o relatório, os resultados ainda estavam longe de serem bons. O relatório, no entanto, encorajava os estudos em linguística computacional – uma das conclusões é que, embora as aplicações de tradução automática ainda não fossem boas, uma das formas de contornar o problema era investir em pesquisa básica em estudos linguísticos para posteriormente realizar o processamento computacional de uma língua.

Na segunda fase do PLN, entre o final da década de 1960 e o final da década de 1970, os cientistas voltaram os esforços para as questões relativas à representação do significado. Dado o fracasso dos sistemas de tradução automática que tinham como base o processamento de dicionários, isto é, os significados das palavras descontextualizadas, tentou-se construir bases de conhecimento de mundo, à moda do que vinha sendo feito na inteligência

[1] O embate filosófico entre as duas visões de linguagem presentes nos estudos linguísticos (e que se refletem no PLN) será apresentado na seção 2.3.

artificial. Embora muito avanço tenha sido constatado, logo se percebeu que, no entanto, esse conhecimento representado nas bases de dados não era facilmente convertido em aplicações de PLN, nem mesmo para domínios muito restritos de linguagem.

Dadas as dificuldades de continuar na empreitada semanticamente orientada, os cientistas voltaram os esforços para a construção de gramáticas computacionais do final da década de 1970 ao final da década de 1980. Jones (1994) chama essa terceira fase do PLN de lógico-gramatical. Era também uma resposta à difusão de gramáticas na teoria linguística, de maneira que se tornava possível testar as hipóteses linguísticas com métodos computacionais e representar as estruturas gramaticais de forma lógica. Apesar da ênfase no desenvolvimento da Linguística (computacional), com o desenvolvimento de gramática computacionais por exemplo, também foram publicados recursos úteis para o PLN, como analisadores sintáticos.

Por fim, na década de 1990, entra em voga, mais forte do que nunca, o que ficou conhecido por processamento estatístico de linguagem (*statistical language processing*). Essa quarta fase do PLN é possibilitada pela análise automática de dados, uma "ideia antiga" mas que só nesse momento poderia ser adequadamente aplicada, dada a difusão de dados legíveis por máquina e o poder computacional disponível para processá-los (JONES, 1994).

É nesse contexto, de computadores capazes de processar grandes quantidades de dados e de algoritmos de aprendizado de máquina, que este trabalho se insere. O PetroGold se soma, portanto, à prateleira de recursos legíveis computacionalmente, com informação linguística codificada – corpora anotados – que passam a alimentar esses algoritmos.

2.2 Anotação linguística

Anotação pode ser entendida de duas formas. Por um lado, é a prática de adicionar informação linguística interpretativa a um corpus eletrônico. Sendo interpretativa, essa informação linguística é produto da mente humana, de uma interpretação específica dos textos que compõem o corpus, não havendo forma objetiva de dizer qual é a etiqueta correta para um dado fenômeno linguístico (LEECH, 1997).

Por outro lado, anotação é também o produto dessa prática (LEECH, 1997). São resultados da anotação linguística os corpora anotados, como os corpora com anotação morfológica, sintática (os *treebanks*), anotação de proposições (os *propbanks*), os corpora com anotação discursiva, de entidades mencionadas e afins (FREITAS, 2022).

Os textos de um corpus podem vir de diferentes fontes e gêneros e podem ser segmentados de diferentes maneiras, como em função do tempo, caso os dados sejam em formato de áudio, por exemplo. Tanto a segmentação dos dados em partes – parágrafo, frase, palavra, tempo ou qualquer outra unidade relevante – quanto a atribuição de etiqueta são tarefas interpretativas, que refletem um conjunto de escolhas que foram feitas no momento da concepção do projeto de anotação (FREITAS, 2022).

Quando há anotação padrão ouro, garante-se que as análises contidas em um corpus são as consideradas corretas segundo os critérios definidos para o projeto de anotação. Se as análises estão corretas, a anotação do corpus está apta a funcionar como modelo para sistemas de aprendizado automático, os quais devem generalizar os dados codificados no corpus para aprender a realizar as análises em textos inéditos (FREITAS, 2022). Além disso, anotações padrão ouro podem ser utilizadas para estudar a língua ou para avaliar os resultados de um sistema automático de anotação (SANTOS, 2008), seja ele constituído de regras linguísticas ou baseado em aprendizado de máquina.

A anotação é feita utilizando um conjunto de etiquetas (um *tagset*), como, por exemplo, as de substantivo ou de oração subordinada, considerando *tagsets* de classes de palavras e funções sintáticas, respectivamente. Como tratamos de recursos linguísticos, as etiquetas tendem a reproduzir os níveis linguísticos clássicos – morfologia, sintaxe, semântica, pragmática etc. – mas podem também codificar outras informações linguísticas, como polaridade, emoções, entidades e afins. As etiquetas podem ser importadas diretamente de alguma teoria, ser inspiradas por ela, por várias, ou ainda direcionadas para alguma aplicação do PLN. Contudo, as etiquetas não bastam, pois é ainda necessário descrever a que fenômenos cada uma delas se aplica e organizá-las dentro de um conjunto seguindo determinados objetivos. Em outras palavras, é necessário ter desenvolvido um esquema de anotação (FREITAS, 2022).

A anotação pode ter diferentes motivações, e o esquema de anotação deve se adequar a elas. Se a tarefa para a qual o recurso anotado será utilizado estiver bem definida, há maiores chances de o esquema de anotação também ser bem definido. Assim, se o objetivo da anotação é construir um recurso que seja útil para estudos linguísticos, um bom esquema de anotação será aquele que seja mais consensualmente aceito pela comunidade linguística. Se, por outro lado, o objetivo do recurso for treinar um modelo de extração de informação específico para um domínio técnico como o biomédico, o esquema de anotação deverá ser pensado para melhor retornar as categorias mais úteis a quem interessa essa extração de informação (FREITAS, 2022).

Em ambos os casos, estudos linguísticos e tarefas de PLN, a anotação

tem como objetivo agrupar itens. Freitas (2022) utiliza exemplos do domínio da geologia (frase 6), onde os termos em negrito, dentro de um esquema de anotação hipotético, poderiam ser todos agrupados sobre uma categoria ampla "rocha", e do domínio da música (frase 7), onde os termos poderiam ser agrupados em categorias distintas: "choro" na categoria ampla "gênero musical", e "violão" na categoria "instrumento". Como pode ser visto nestes mesmos exemplos, algumas categorizações necessitam de um conhecimento mais específico do domínio em questão, enquanto outras são mais genéricas.

6. A Formação Abaeté é constituída por **conglomerados**, **arenitos conglomeráticos**, **arenitos** e **lamitos**.

7. No estilo **choro**, o **violão** caracteriza-se por frases de contraponto geralmente em escala descendente, utilizando-se somente as cordas graves.

Além do desenvolvimento do esquema de anotação, é necessária uma metodologia para implementá-lo. Muitas são as variações possíveis neste campo: relativas à aplicação apenas de ferramentas automáticas ou apenas de análises humanas, ou ainda uma execução híbrida, em que se utiliza a análise humana como um refinamento da análise automática; relativas à ferramenta que será utilizada pelos anotadores humanos para conduzir a anotação, o que pode facilitar ou dificultar a tarefa, e relativas a como garantir a qualidade da anotação.

Uma aliada na construção de consistência na anotação – fenômenos semelhantes devem receber sempre a mesma classificação – é a documentação (FREITAS, 2022; ARCHER, 2012; SAMPSON, 1995). Ela funciona como o material de apoio para os anotadores humanos pois descreve os fenômenos sendo anotados, direcionando-os para as regularidades encontradas, para as exceções, além de exemplificar ocorrências complexas ao lado da solução encontrada para cada caso. Funcionando como um manual, a documentação permite que pessoas sejam consistentes durante o processo de anotação, sendo necessário modificá-la quando novos casos difíceis são encontrados no corpus, possivelmente tendo que se reformular alguma categoria que não tenha se mostrado adequada aos textos do corpus. Freitas (2022) nota, porém, que o processo de reformulação das categorias não é infinito uma vez que se sabe a motivação para anotação, pois é essa motivação que delimita aquilo que deve ser ainda mais aprimorado e aquilo que pode ser relevado no esquema de anotação.

Como já observado, anotar é enquadrar um segmento de texto dentro de uma categoria, o que significa interpretá-lo de uma certa maneira. Ainda

assim, Archer (2012) considera que a anotação linguística é uma informação nova mas bem-vinda ao corpus uma vez que possibilita uma série de atalhos na busca por fenômenos linguísticos (a autora fala em um contexto de estudos linguísticos). Para isso, a autora nota que as análises devem vir acompanhadas de documentação, pois é ela que garante que uma interpretação será justificada e poderá ser compartilhada por diferentes pessoas, que então poderão fazer seus próprios julgamentos relativos às fraquezas e pontos fortes da anotação de um corpus.

Do ponto de vista do PLN por aprendizado automático, entendemos que a anotação é importante pois um dos objetivos das tarefas é possibilitar que a máquina reproduza as anotações consideradas relevantes por um grupo de especialistas em uma determinada área, o que pode ser feito com maior velocidade e exigindo menos recursos humanos uma vez já tendo sido treinada. Para isso, é necessário providenciar exemplos a partir dos quais a máquina aprenderá a anotação correta. Esses exemplos deverão ser muito ou relativamente volumosos em quantidade a depender do algoritmo sendo utilizado, motivo pelo qual agrupar diferentes fenômenos sob um mesmo rótulo é crucial, pois garante que haverá uma repetição substancial das etiquetas, facilitando a generalização (FREITAS, 2022).

2.3 Linguística empírica

Línguas humanas são tradicionalmente consideradas sistemas simbólicos, isto é, sistemas que relacionam um código – a palavra e os demais níveis de organização linguística – a um significado (MANNING, 2003). A natureza do significado, porém, é alvo de disputa teórica. O estudo desse sistema pode ser feito por meio de duas abordagens filosóficas distintas, como notam Manning e Schutze (1999): a abordagem racionalista e a empirista.

A abordagem racionalista entende que uma parte considerável do conhecimento linguístico é dada por herança genética, antes mesmo de o bebê entrar em contato com o mundo e com os falantes de uma língua. Essa abordagem foi amplamente difundida na linguística pelos argumentos de Noam Chomsky, sobretudo o argumento da pobreza de estímulo – o linguista considera que o ser humano detém um conhecimento linguístico muito complexo a despeito de ainda ter muito pouco contato com uma língua quando ainda pequeno, o que seria justificado pelo conhecimento internalizado de origem genética. Na inteligência artificial, essa abordagem se caracteriza pela tentativa de codificar o máximo de conhecimento linguístico específico nos sistemas inteligentes.

Do outro lado, a abordagem empirista considera que o conhecimento ini-

cial humano é muito básico, se restringindo às estruturas gerais que permitem ao bebê *aprender*. Conhecimento linguístico específico, como os princípios de uma língua, são aprendidos pelas crianças apenas no contato com o mundo. A abordagem tem visto um ressurgimento na inteligência artificial recentemente, onde os sistemas voltam a ser programados com apenas algoritmos básicos de aprendizado geral, de maneira que os parâmetros das línguas são definidos apenas no contato com os dados da língua em uso – um corpus, por exemplo (MANNING; SCHUTZE, 1999).

Racionalistas realizam julgamentos categóricos, por exemplo, das frases que são bem ou mal formadas em uma determinada língua para um falante nativo; empiristas como Geoffrey Sampson, por sua vez, na tarefa de anotação de um corpus como o SUSANNE (SAMPSON, 1995), precisam classificar as estruturas de um treebank, isto é, declarar se são sintagmas nominais, verbais, preposicionais etc. Ambas as formas de se trabalhar com a língua podem ser consideradas categóricas se não utilizam nas análises categorias contínuas, como propõe Manning (2003).

Manning defende, nesse trabalho, a utilidade de uma abordagem empirista não-categórica, o que se refletiria, na linguística, em estudos gramaticais probabilísticos baseados em dados de corpus. O autor considera que a influência de Chomsky no século XX acabou desencorajando os estudos da linguagem de cunho estatístico, mantendo a tradição grega de se pensar a gramática em termos inteiramente categóricos, mas defende que nem mesmo os estudos linguísticos deveriam ser assim.

Abordagens racionalistas como a de Chomsky fazem uso de uma descrição linguística que, segundo Chambers (2003), utiliza unidades discretas e qualitativas em prejuízo de unidades de análise contínuas e quantitativas. Chambers nota que essa tradição de categoricidade já se fazia presente entre os estruturalistas americanos, cujo ápice pode ser encontrado em Joos:

> "Ordinary mathematical techniques fall mostly into two classes, the continuous (e.g., the infinitesimal calculus) and the discrete or discontinuous (e.g., finite group theory). Now it will turn out that the mathematics called 'linguistics' belongs to the second class. It does not even make any compromise with continuity as statistics does, or infinite-group theory. Linguistics is a quantum mechanics in the most extreme sense. All continuities, all possibilities of infinitesimal gradation, are shoved outside of linguistics in one direction or the other." (JOOS, 1950 apud CHAMBERS, 2003, pp. 701-702)

A ideia de classes discretas já se fazia presente também nos primeiros estudos de organização do discurso e de categorização do pensamento, na Grécia Antiga, realizados por Aristóteles em sua obra Categorias. Nela, se apresentavam as dez unidades discretas básicas do pensamento que se refletiriam na linguagem[2], sendo a origem de grande parte das classificações gramaticais que se seguiriam, seja na morfossintaxe ou mesmo na categorização de sentidos.

Ao contrário de Chomsky e de parte da tradição linguística racionalista e mesmo empirista, Manning (2003) argumenta em favor de estudos linguísticos probabilísticos na sintaxe, com unidades de análise contínuas e quantitativas. A aplicação de uma abordagem sintática probabilística encontraria utilidade, por exemplo, no processamento automático de linguagem natural, pois, uma vez que coleções de textos eletrônicos maiores são disponibilizadas e há capacidade computacional para processá-las, é possível encontrar, estatisticamente – em substituição aos métodos simbólicos –, os padrões mais comuns que auxiliem, entre outros, na resolução de ambiguidade, que é característica das línguas humanas e com que lidamos a todo momento.

Para os estudos linguísticos e para o ensino de línguas, obter informações de distribuição das palavras em seus contextos de uso pode ser mais adequado do que "intuir" quais palavras ou estruturas linguísticas são ou não são utilizadas em uma língua. Um aprendiz de uma língua precisa aprender a utilizar as palavras adequadas nos contextos certos, informação essa que se pode depreender de uma análise distributiva da língua orientada por dados produzidos por falantes, de forma que o ajude a decidir quais palavras e estruturas são as preferíveis nas diferentes variedades e situações de comunicação. Confiar apenas nas intuições que um ou outro falante individualmente possam ter sobre as estruturas mais relevantes de sua língua e como elas são aplicadas pode nos guiar ao erro. Como declara Sampson (2002), "the data of 'intuitions' may be abundant, but they are hopelessly unreliable" (p. 2).

Manning (2003) vai além, e defende que, do ponto de vista cognitivo e de aquisição de linguagem, o processamento humano também ocorre de forma probabilística. Assim, a criança em período de aquisição de linguagem, em face de *inputs* pouco informativos, realiza cálculos incertos sobre o que ouve nos dados. Podemos errar a interpretação que damos aos dados e reajustamos nosso raciocínio iterativamente. Quando alguém diz "está frio aqui", há um cálculo probabilístico de como a declaração deve ser interpretada de acordo com os diferentes contextos que conhecemos em contraste com o contexto em que a frase foi de fato enunciada – e há a chance de o cálculo falhar

[2]A visão simbólica de Aristóteles sobre a linguagem pode ser sintetizada no seguinte trecho: "Spoken words are the symbols of mental experience and written words are the symbols of spoken words." (On Interpretation I, traduzido por E. M. Edghill).

(podemos entender mal uma intenção), ou acertar, ambas as situações podendo ser previstas estatisticamente. Nesses casos, uma abordagem probabilística de alto nível seguiria a distribuição $P(sentido|enunciado, contexto)$, sendo um mapeamento da forma enunciada para um sentido específico condicionado pelo contexto.

Em relação ao argumento de Chomsky de que estudos probabilísticos não explicam a língua internalizada do falante, autores como Manning e Sampson defendem que qualquer teoria de linguagem deve poder ser testada empiricamente. Manning e Schutze (1999) notam que, embora o grande sucesso que abordagens estatísticas têm obtido no processamento automático de linguagem natural não seja decisivo para validar a hipótese de que a aquisição de linguagem ocorra de forma probabilística, deve servir de subsídio para apoiá-la. Sampson (2002) abre concessão para alguns objetos de estudo, como princípios morais, que não podem ter sua validade testada empiricamente – "there is no way that observation can possibly 'refute' or 'confirm' a moral principle such as 'one should obey the law even if one disagrees with it'" (p. 1) –, mas afirma que a língua é um objeto que não tem nenhum impedimento para que seja estudado empiricamente, testando-se hipóteses sobre os padrões nos dados e, à luz de novos dados que confrontem uma hipótese, propor reformulações que expliquem os novos dados e os antigos.

Teorias linguísticas categóricas, de acordo com Manning (2003), se dizem mais poderosas do que são. Por um lado, se apresentam como capazes de distinguir uma categoria da outra estabelecendo limites rígidos sobre estruturas que, na realidade, quando confrontadas com as questões da convencionalidade e da criatividade humana, vemos que não têm fronteiras claras. Além disso, segundo o autor, elas explicam pouco. Não se tenta explicar, com essas teorias, por que certas estruturas ou significados são preferidos em certos contextos, ou, do ponto de vista do receptor, qual interpretação é a preferida.

Com o auxílio dos cálculos estatísticos, de acordo com Manning (2003), temos resultados mais concretos e informativos sobre qualquer hipótese linguística. Uma análise distributiva de sintagmas nominais anotados em um corpus, por exemplo, propõe analisar uma abstração teórica – o sintagma –, e esta pode ser analisada empiricamente, utilizando as ferramentas estatísticas e recursos – corpora – adequados. Ou seja, diferentemente do que alega Chomsky, análises de corpora não se restringem a análises superficiais, podendo ser realizadas também em estruturas complexas, que visem representar conhecimento linguístico abstrato.

Ao longo desta dissertação, nem sempre seremos capazes de apresentar os dados probabilísticos para cada uma das análises linguísticas realizadas. E, no

fundo, nossa tarefa é categórica – apesar das discussões que levantaremos para cada uma das classificações que faremos, o produto da anotação linguística é uma categorização das palavras presentes no treebank em apenas uma classe. Contudo, em diversos momentos ficará claro para quem lê que encaramos a anotação linguística como um processo que lida com fenômenos que não são discretos e, portanto, não podem ser facilmente categorizados sem que sofram alguma deformação. A ideia é que, conforme discutimos, precisamos reduzir os fenômenos linguísticos a algumas características em comum para que possam caber nas caixinhas às quais queremos submetê-los. E essa tarefa ingrata tem um objetivo prático: facilitar o aprendizado automático das interpretações que adicionamos aos dados linguísticos.

3
Revisão da literatura

3.1
Sobre *treebanks*

Treebank (literalmente um banco de árvores, ou floresta sintática) é o nome dado ao recurso linguístico que contém um conjunto de árvores sintáticas (BICK et al., 2007). Acredita-se que o nome *treebank* tenha sido adotado por Geoffrey Leech, fazendo alusão ao fato de que a representação mais comum para análises sintáticas é em formato de árvores (SAMPSON, 2003).

Treebanks são úteis tanto para a linguística descritiva quanto para o PLN: "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).

Treebanks são construídos, frequentemente, com base em corpora já previamente existentes, isto é, conjuntos de texto não anotados ou com anotação morfológica, acrescentando-se anotação sintática ao material (NIVRE, 2008). Uma consequência dessa prática é que os treebanks herdam as características dos corpora de que são originários, como é o caso do SUSANNE Corpus (SAMPSON, 1995), baseado em uma seção do Brown Corpus of American English (KUCERA; FRANCIS, 1967), e a Floresta Sintá(c)tica (AFONSO et al., 2001), baseada inicialmente em uma fração do corpus CETEMPúblico (ROCHA; SANTOS, 2000).

As diferenças entre treebanks podem ser de duas naturezas: relativas às características dos textos que compõem o treebank e relativas à anotação sintática propriamente.

Em relação às características dos textos, Afonso (2004b) as separa em características "externas" e "internas". As características externas compreendem o tipo de texto presente no material em relação ao gênero textual (corpus jornalístico, literário, etc.), ao registro (escrito, falado, dialetal, etc.) e ao tempo (corpus contemporâneo ou histórico, por exemplo). Já as características internas dizem respeito à forma como os textos estão dispostos no corpus, como as decisões relativas a que seções dos textos devem ou não compor o material, o que deve ser considerado palavra, quais os critérios de divisão de frases, a organização por assuntos e a identificação por metadados.

De acordo com Nivre (2008), a maioria dos treebanks de língua escrita são baseados em corpora de textos de jornal contemporâneos em apenas uma língua, pois são textos relativamente fáceis de se obter. Dois exemplos de treebanks nesse modelo são a seção Wall Street Journal do Penn Treebank (MARCUS; SANTORINI; MARCINKIEWICZ, 1993), em inglês, e a já mencionada Floresta Sintá(c)tica, em português, que inicialmente continha apenas extratos de jornal.

Há, contudo, exceções à supremacia dos treebanks baseados em linguagem contemporânea, como é o caso dos treebanks históricos como o Penn-Helsinki Parsed Corpus of Middle English (TAYLOR; KROCH, 1994), o Partially Parsed Corpus of Medieval Portuguese (ROCIO et al., 2003), e o corpus Tycho Brahe de português histórico (GALVES, 2018). Há também os treebanks paralelos, compreendendo duas ou mais línguas alinhadas, como o Czech-English Penn Treeebank (ČMEJREK; HAJIČ; KUBOŇ, 2004).

Em relação às diferenças na anotação sintática, treebanks podem se diferenciar na codificação e no esquema de anotação. A codificação diz respeito à forma como a anotação ficará disponível no treebank, isto é, se será utilizada alguma linguagem de marcação específica, como HTML ou XML, se será armazenada junto ao texto ou em arquivos separados, em colunas ou em texto corrido e afins (NIVRE, 2008).

O esquema de anotação diz respeito à forma como serão analisados os fenômenos linguísticos e pode variar a depender dos objetivos da anotação. Alguns treebanks se alinham a teorias sintáticas específicas, como a *Head-Driven Phrase Structure Grammar* (HPSG), que tem treebanks em inglês (OEPEN et al., 2002) e búlgaro (SIMOV et al., 2002), a *Functional Generative Description* (SGALL et al., 1986), à qual se alinha o Prague Dependency Treebank, e a *Combinatory Categorial Grammar* (HOCKENMAIER; STEEDMAN, 2002), à qual se alinha o CCG-bank, uma versão do Penn Treebank.

Bick et al. (2007) notam que, embora seja comum o desejo de se criar treebanks para representar teorias linguísticas específicas, a maior parte dos recursos que tenham tamanho considerável se julgam corpora de referência para "análise sintáctica generalizada, para uma dada língua" (p. 292). Os autores acrescentam que, mesmo que se considerem gerais, e não específicos de uma teoria, "é impossível evitar totalmente a orientação ou influência de uma teoria linguística, por mais neutro que se deseje ser" (p. 292).

A anotação de constituintes foi a principal forma de anotação nos primeiros projetos de anotação em larga escala (NIVRE, 2008), como o Lancaster Parsed Corpus (GARSIDE; LEECH; VÁRADI, 1995). Nesse tipo de anotação, após anotar a classe gramatical de cada uma das palavras, anotam-

```
[N Vous_PPSA5MS N]
[V accedez_VINIP5
   [P a_PREPA
      [N cette_DDEMFS session_NCOFS N]
   P]
   [Pv a_PREP31 partir_PREP32 de_PREP33
      [N la_DARDFS fenetre_NCOFS
         [A Gestionnaire_AJQFS
            [P de_PREPD
               [N taches_NCOFP
               N]
            P]
         A]
      N]
   Pv]
V]
```

Figura 3.1: Exemplo de anotação de constituintes no IBM Paris Treebank. Fonte: Nivre (2008).

se os sintagmas, chamados constituintes, com categorias do tipo NP para sintagmas nominais, VP para verbais, etc. A identificação de constituintes é comumente chamada de *bracketing* (anotação em colchetes).

A anotação funcional, por sua vez, estabelece a função sintática dos termos da frase. O modelo tem sido cada vez mais utilizado nos últimos anos, como na anotação de dependências do Prague Dependency Treebank of Czech (HAJIČ, 1998), onde a anotação de estrutura de dependências é adicionada sobre a camada de anotação morfológica sem utilizar qualquer tipo de anotação de constituintes (figura 3.2). Mais recentemente, o projeto Universal Dependencies tem disponibilizado também uma série de treebanks para diferentes línguas utilizando esse modelo de anotação funcional (MARNEFFE et al., 2021).

A tarefa de anotação humana de treebanks é muito laboriosa. Uma consequência disso é que, embora corpora de 1 milhão de palavras já sejam considerados pequenos hoje, treebanks do mesmo tamanho são escassos (NIVRE, 2008). Uma estratégia comum para contornar o trabalho de anotação humana é a pós-correção de apenas seções do treebank, seja manualmente ou utilizando analisadores não-determinísticos que apontam possibilidades entre as quais o anotador deverá desambiguar[1].

A Floresta Sintá(c)tica foi um projeto pioneiro para língua portuguesa. Como mencionado, o treebank era inicialmente composto por textos de jornal em língua portuguesa contemporâneos: o CETEMPúblico para a variedade

[1] A estratégia foi utilizada no TreeBanker (CARTER, 1997), por exemplo.

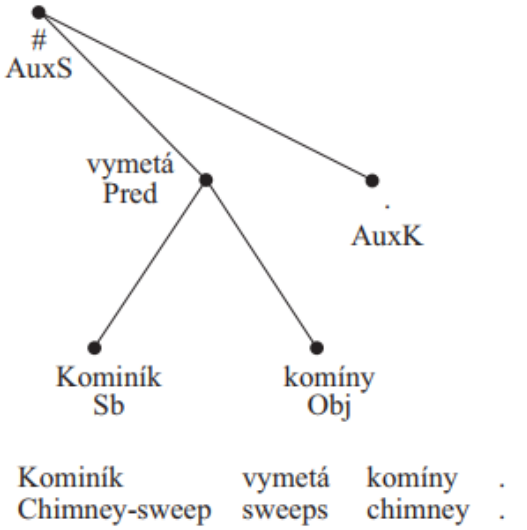


Figura 3.2: Exemplo de anotação funcional no Prague Dependency Treebank. Fonte: Nivre (2008).

europeia e o CETENFolha para a brasileira. Dadas as dificuldades de se atingir um grande número de frases anotadas sintaticamente com qualidade, utilizou-se a estratégia de revisar a anotação automática de apenas uma seção da Floresta Sintá(c)tica, a que se chamou de Bosque. Assim, Afonso et al. (2001) relatam que, à época, a porção revista do treebank continha 1.427 árvores revisadas (1.405 frases, 36.408 tokens), enquanto que a porção não revista, chamada Floresta Virgem, continha 41.406 árvores (41.406 frases, 1.072.857 tokens). A diferença no número de árvores e de frases no Bosque se deve ao fato de que, em caso de ambiguidade, as diretivas de anotação do projeto abriam a possibilidade de mais de uma análise da árvore, o que só foi aplicado na revisão humana.

A Floresta Sintá(c)tica foi anotada pelo PALAVRAS (BICK, 2014), que é um analisador morfossintático e de dependências baseado em léxico e em regras linguísticas utilizando a metodologia de Constraint Grammar (KARLSSON et al., 2011) para desambiguar a morfologia e mapear as funções sintáticas de maneira dependente de contexto. Portanto, trata-se primeiramente de uma anotação funcional, na classificação de Nivre (2008). Na revisão do Bosque, essa primeira anotação funcional foi revista por humanos antes de se seguir para a anotação de constituintes, utilizando um outro programa baseado em Phrase Structure Grammar (PSG).

Na imagem 3.3, vemos a anotação no formato Constraint Grammar, em que cada um dos tokens possui informação de lema, atributos morfológicos, classe gramatical, função sintática e direção do governante da relação de dependências, visível pelos símbolos de maior (>) e menor (<). Essa frase

```
Queremos     [querer] <fmc> V PR 1P IND VFIN @FMV
que          [que] KS @SUB @#FS-<ACC
especialistas [especialista] N M/F P @<ACC
internacionais [internacional] ADJ M/F P @N<
e            [e] <co-postnom> KC @CO
nacionais    [nacional] ADJ M/F P @N<
pensem       [pensar]  V PR 3P SUBJ VFIN @FMV
em           [em] PRP @<PIV
as           [o] <artd> DET F P @>N
possibilidades [possibilidade] N F P @P<
que          [que] KS @SUB @#FS-<ACC
existem      [existir] V PR 3P IND VFIN @FMV
de           [de] PRP @<ADVL
abordagem    [abordagem] N F S @P<
de           [de] PRP @N<
o            [o] <artd> DET M S @>N
tema         [tema] N M S @P<
em           [em] PRP @N<
o            [o] <artd> DET M S @>N
contexto     [contexto] N M S @P<
de           [de] <sam-> PRP @N<
a            [o] <-sam> <artd> DET F S @>N
sociedade    [sociedade] N F S @P<
de           [de] <sam-> PRP @N<
a            [o] <-sam> <artd> DET F S @>N
informação   [informação] N F S @P<
:
```

Figura 3.3: Exemplo de anotação no formato Constraint Grammar no HAREM. Fonte: Afonso et al. (2002).

teria sido corrigida pelos anotadores do Bosque nos tokens "que" (a segunda ocorrência), que não é conjunção integrante, mas pronome relativo, exercendo função de sujeito (possibilidades *que existem*), e "de", que não introduz função adverbial, mas de adjunto adnominal (possibilidades *de abordagem*).

Uma vez tendo sido corrigida essa representação funcional, as frases foram direcionadas ao analisador de constituintes (indicados pelo sinal de igual na figura 3.4), que utiliza as informações prévias para realizar sua análise. Assim, a análise humana da anotação funcional na etapa anterior auxilia na identificação automática dos constituintes, e a revisão humana é iterativa, sendo também realizada sobre o resultado da segunda etapa, garantindo cada vez menos erros na anotação das frases.

Um dado interessante é que, embora a seção Floresta Virgem da Floresta Sintá(c)tica não tenha sido revista por humanos, a melhoria da anotação do Bosque resultou na melhoria também do analisador automático PALAVRAS que, a cada iteração, precisou ter suas regras melhoradas para se adequar à qualidade das análises humanas. Assim, a Floresta Virgem não é um treebank padrão ouro, mas se beneficiou de um sistema melhorado de anotação (AFONSO et al., 2001).

```
A1
STA:fcl
(...)
ADVL:pp
=H:prp( de '<sam->)                de
=P<:np
==>N:pron-det( esse '<-sam> <dem> F P)        essas
==H:n( revista 'F P)                revistas
==N<:pp
===H:prp( de )  de
===P<:np
====H:n('viagem' F P)          viagens
====N<:cu
=====CJT:fcl
======SUBJ:pron-indp('que' <rel> F P)      que
======ADVL:adv('agora' <kc>)               agora
======P:v-fin('proliferar' PR 3P IND)      proliferam
=====CO:conj-c('e' <co-vfin> <co-fmc>)     e
=====CJT:fcl
======SUBJ:pron-indp('que' <rel> F P)      que
======P:v-fin('perpetuar' PR 3P IND)       perpetuam
======ACC:np
=======>N:art('o' F P)         as
=======H:n('fantasia' F P)     fantasias
=======N<:pp
========H:prp('sobre')         sobre
========P<:np
========H:n('ilha' F P)        ilhas
========N<:adj('exótico' F P)              exóticas
```

Figura 3.4: Exemplo de anotação de constituintes no formato Constraint Grammar. Fonte: Afonso et al. (2002).

Outra curiosidade é que, enquanto projeto, a Floresta Sintá(c)tica foi encabeçada por duas organizações com objetivos diferentes. O VISL (Visual Interactive Syntax Learning), sendo um projeto de pesquisa e ensino da Southern Denmark University, tinha maior interesse linguístico do que computacional na Floresta, como a criação e propagação de conhecimento linguístico, enquanto a (atualmente chamada) Linguateca tinha sobretudo interesse computacional, na produção de um recurso para avaliação de analisadores sintáticos e outras ferramentas baseadas em recursos públicos e validados linguisticamente (AFONSO et al., 2001).

Ambos os objetivos para o desenvolvimento de um recurso não são conflitantes, mas podem influenciar escolhas, por exemplo, no esquema de anotação. Afonso (2004b), assim como Bick et al. (2007), apontam duas utilidades principais para a Floresta Sintá(c)tica: a pesquisa linguística e a avaliação de sistemas. Enquanto pesquisa, a autora realiza um estudo descritivo das orações do Bosque, caracterizando-as em relação à finitude ou infinitude do verbo principal e em relação à transitividade verbal assim como a estrutura do

objeto direto, seja ele um sintagma nominal ou uma oração finita, um sintagma adjetival etc. Tal análise só pode ser possível com anotação sintática, e é mais acurada quando a anotação é revista por humanos.

Enquanto recurso para avaliação, a autora nota que o Bosque, sendo rico em informação linguística validada, pode avaliar sistemas de análise de morfologia e sintaxe. Além disso, indica que o formato e o esquema de anotação do treebank são flexíveis, de tal maneira que é possível adaptar o formato em que se codificaram as anotações para compará-las com outros treebanks, além de ser possível simplificar as etiquetas de modo a comparar esquemas de anotação, como é o caso relatado das palavras "segundo" e "como" que, em contextos específicos, podem ter sua anotação facilmente simplificada para a de preposições (no lugar da anotação primária "advérbio").

3.2 Sobre revisão de *treebanks*

Por melhor que seja a qualidade da anotação de um treebank, erros sempre podem existir e irão impactar negativamente, em maior ou menor grau (dependendo da quantidade e padronização dos erros), o treinamento e a avaliação de modelos de PLN. Erros são mais difíceis ainda de serem evitados quando o recurso anotado é de tamanho considerável, de difícil inspeção manual.

Pode-se prevenir a criação de novos erros na anotação de um treebank utilizando algumas estratégias, como documentar as decisões de anotação e verificar se há um alto grau de concordância interanotadores, assuntos discutidos nas seções 2.2 e 3.3, respectivamente. Mas há também uma vasta bibliografia sobre estratégias para detecção e correção de erros em corpora anotados. Tanto a detecção quanto a correção dos erros podem ocorrer de forma manual, semiautomática ou mesmo automática, dependendo das necessidades do projeto, que pode exigir maior quantidade de dados ou maior qualidade na anotação.[2]

Quando o recurso é volumoso, uma possibilidade é revisar apenas uma pequena amostra do todo (JÄRVINEN, 2003). A detecção de erros na amostra pode servir como indicador de quais são as questões mais complicadas que merecem revisão no corpus todo, seja porque a decisão de anotação não foi tomada corretamente, seja porque não houve aplicação consistente da decisão tomada no corpus (DICKINSON, 2005). De modo semelhante, Kilgarriff (1998) sugeriu olhar o corpus sob a lente da distribuição das palavras, para que a partir

[2]O levantamento bibliográfico desta seção expande o conteúdo já apresentado em Freitas e de Souza (2023).

delas se encontrem os erros de anotação semântica, tornando desnecessário revisar o corpus todo.

Outro procedimento que se pode realizar a partir de uma pequena amostra revisada é utilizá-la como material de treinamento para um anotador automático e, com este novo modelo, anotar o restante do material, que deve incorporar as correções realizadas previamente (SKUT et al., 1997). É uma ideia próxima à de "aprendizado ativo", quando se utiliza um material anotado para selecionar o que anotar a seguir (DICKINSON, 2005) – por exemplo, sabendo-se que uma certa frase está anotada corretamente, pode-se procurar por frases parecidas, que devem ser anotadas da mesma forma. De maneira semiautomática, Hinrichs et al. (2000) propõem que, ao selecionar frases cuja anotação está correta, sejam retornadas frases semelhantes com suas respectivas anotações, de modo que seja possível distinguir a anotação mais comum daquelas que destoam, sendo uma forma de verificar o nível de consistência na anotação do fenômeno.

Há também as formas de revisar corpus orientadas por conhecimento linguístico. Oliva (2001), por exemplo, desenvolve regras manualmente para identificar erros em corpus. Dickinson (2005) nota que, nos casos em que é necessário que alguém especifique quais são os padrões de erro, não há nenhuma garantia de que todos os erros ou tipos de erros serão identificados, pois dependem primeiro da identificação humana.

Mas ainda é possível realizar o contrário: no lugar de especificar quais são os padrões de erros, pode-se especificar quais são os padrões corretos. Ou seja, constrói-se uma gramática paralelamente ao treebank, de maneira que um recurso alimenta o outro, conforme adotado, por exemplo, por Oepen, Flickinger e Bond (2004). Dickinson (2005) considera essa abordagem útil e teoricamente muito atrativa, mas de difícil aplicação pois requer muito esforço para desenvolver uma gramática que seja robusta o suficiente para considerar todos os casos do corpus.

De um ponto de vista estatístico, pode-se partir do princípio de que erros são anomalias, isto é, eventos que destoam de um padrão. Usando modelos probabilísticos como o de Eskin (2000), é possível detectar anomalias (*outliers*) – padrões locais de etiquetas raros – em um corpus. Quando aplicado no Penn Treebank (TAYLOR; MARCUS; SANTORINI, 2003), a estratégia retornou 7.055 anomalias, no entanto, apenas 44% delas eram de fato erros, segundo uma inspeção manual.

Para explicar a baixa precisão do método, devemos nos lembrar de que fenômenos pouco frequentes ou raros nem sempre são erros, pois a língua é repleta de fenômenos de baixa frequência e que nem por isso estão

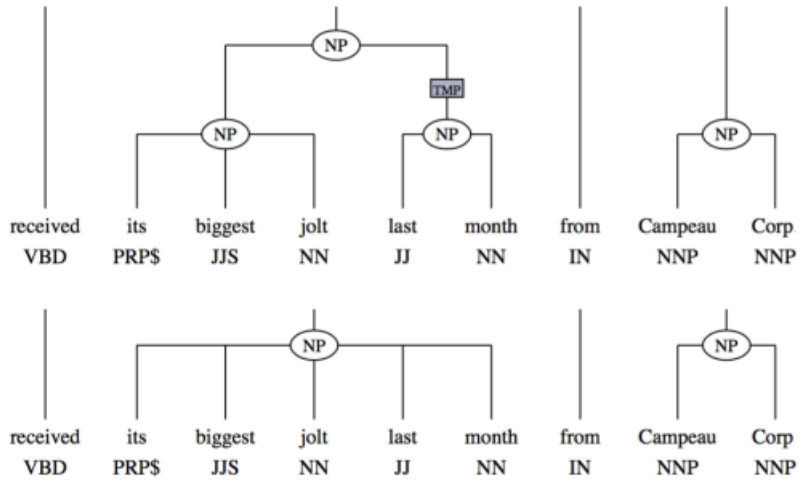


Figura 3.5: Duas frases com as mesmas palavras mas com variação na anotação de constituintes. Fonte: Boyd, Dickinson e Meurers (2008).

errados, o que Manning e Schutze (1999) detectaram usando a lei de Zipf. Do mesmo modo, erros de anotação nem sempre são raros, pois podem ter sido introduzidos de forma padronizada, seja por um dos anotadores humanos ou como efeito colateral de mudanças nas diretivas que não foram corretamente revistas no corpus.

Outra forma de detectar erros na anotação de corpora que foi muito desenvolvida ao longo do tempo em diferentes contextos é a que busca por variação na anotação de trechos semelhantes. Dickinson (2005) mostra que esse tipo de aplicação pode ser realizado para anotação sequencial de POS, para anotação de sintaxe de constituintes ou mesmo para sintaxe de dependências. A ideia é a de que em corpora muito volumosos é possível encontrar pedaços de texto idênticos cuja anotação seja diferente, cabendo ao revisor (automático ou manual) decidir se a diferença está correta ou se uma das duas anotações é a correta e deve ser replicada nas demais. No exemplo 3.5, dois trechos com as mesmas palavras tiveram anotações distintas para o sintagma “last month”.

Boyd, Dickinson e Meurers (2008) desenvolvem algumas restrições para melhorar a eficiência desse método de detecção de erros. A primeira heurística, chamada “non-fringe”, entende que, para os erros identificados pelo método serem de fato erros, é importante que as palavras que estão ao redor do núcleo com variação sejam as mesmas também. Na figura 3.5, como vimos, além de o núcleo “last month” ter anotação distinta nas duas frases, as palavras que cercam o núcleo também são as mesmas em ambas as frases.

Outra restrição proposta em Boyd, Dickinson e Meurers (2008) é a chamada “dependency context”, específica para anotação de dependências,

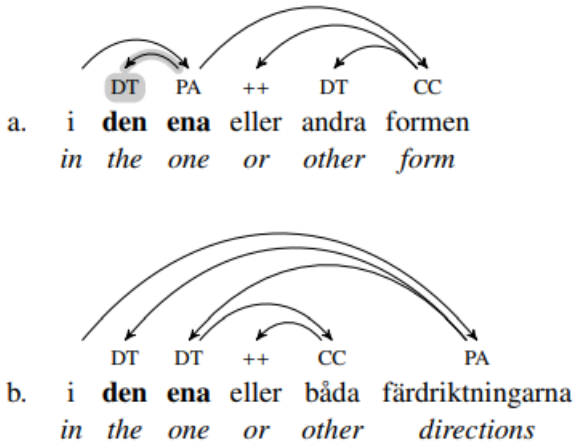


Figura 3.6: Duas frases com as mesmas palavras sendo anotadas de formas diferentes, mas sem relação de dependência entre si. Fonte: Boyd, Dickinson e Meurers (2008).

exigindo que, para que a variação na anotação seja considerada erro, a relação de dependência entre as duas palavras que compõem o núcleo da variação seja a mesma. Na figura 3.6, as palavras do núcleo "den ena", apesar de estarem rodeadas pelas mesmas palavras, não estão relacionados entre si nas duas frases, indicando que se trata de uma ambiguidade legítima, e não um erro, segundo os autores.

Trabalhando com corpora em inglês, francês e finlandês, de Marneffe et al. (2017) propuseram duas adaptações ao trabalho de Boyd, Dickinson e Meurers (2008). Primeiro, no lugar de considerar trechos idênticos aqueles que têm as mesmas palavras, os autores foram mais abrangentes e consideraram trechos com os mesmos lemas. Além disso, os autores removeram a restrição original de que, para serem considerados idênticos, os trechos precisavam ter as mesmas anotações de POS.

Analisando 100 pares de palavras retornados pelas variações do método, os autores concluem que adicionar a restrição *dependency context* melhora consideravelmente a precisão do método, que chega a até 76% em inglês (ou seja, 76% das frases retornadas estão de fato anotadas erroneamente), sendo que o custo dessa restrição é identificar apenas 66% dos erros identificados pelo método mais abrangente, *non-fringe*. No entanto, o resultado não se reproduz para todas as línguas: o finlandês, língua morfologicamente mais rica que o inglês e com menos palavras funcionais, tem apenas 21% de precisão utilizando a heurística *dependency context*.

Por fim, outra forma de detectar erros em anotação de corpus é utilizando o princípio por trás dos anotadores automáticos. Se o objetivo de um modelo que aprende uma tarefa é conseguir replicar um padrão, quando contrastamos

o resultado de um anotador automático treinado em um dado corpus com a anotação padrão ouro desse mesmo corpus, conseguimos verificar quais setores do padrão ouro não estão seguindo o padrão que o modelo aprendeu. Enquanto o método retorna um bom número de discordâncias entre as duas anotações no trabalho de Halteren (2000), apenas 20,49% das discordâncias são erros de fato do padrão ouro, confirmando a hipótese de que é mais frequente o modelo generalizar um erro do que o padrão ouro estar mal anotado.

Para aprimorar a busca por erros utilizando anotadores automáticos, Halteren (2000) propõe, embora não chegue a testar de fato, a utilização de anotadores diferentes para detectar erros na anotação do corpus. A ideia é a de que construções com anotação errada no corpus serão de difícil anotação para a maioria dos anotadores automáticos, resultando em desacordo entre as previsões. De certo modo, essa proposta, não testada pelo autor, é semelhante a uma das estratégia que descrevemos em Freitas e de Souza (2023) e que utilizamos para revisar o PetroGold, assunto que abordaremos na seção 4.2.

3.3 Sobre avaliação de *treebanks*

Avaliação é um termo que pode ser utilizado em diferentes contextos no PLN. Podemos avaliar, de um lado, se um esquema de anotação está bem definido e se ele é adequado aos objetivos de uma tarefa. De outro, podemos avaliar o alinhamento dos anotadores humanos às diretivas do projeto de anotação, isto é, o quanto uma anotação humana pode ser considerada correta. E podemos ainda avaliar o desempenho de sistemas automáticos de anotação baseados em conhecimento linguístico ou em aprendizado de máquina. Todas essas avaliações podem ser utilizadas para avaliar, no fim, a qualidade de um recurso como um treebank.

Em um trabalho de 2011, 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 conclui que há pouco espaço para melhorar o sistema de anotação automática (Stanford Part-of-Speech Tagger) do ponto de vista computacional, cabendo à modelagem linguística a tarefa de alavancar os resultados. O autor realiza uma análise de erros de uma amostra com 100 casos que o sistema errou para identificar que pontos poderiam ser melhorados, e distribui os casos em 7 classes. Dessas, 3 são 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.

A primeira das três classes, que compreende 12% dos erros de anotação automática, diz respeito à pouca especificação ou clareza das categorias do esquema de anotação. 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 na frase "it will take a $ 10 million fourth-quarter charge against discontinued operations" (MANNING, 2011, p. 7). O autor 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 – artificialmente – quando o termo deve ser anotado como verbo ou como adjetivo. Sampson e Babarczy (2008) questiona esse tipo de prática pois, 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 uma vez que facilitam a obtenção de consistência, consequentemente facilitando a generalização.

Em um contexto parecido, Freitas et al. (2018) decidiram adicionar uma etiqueta, PCP, no corpus Mac-Morpho (ALUÍSIO et al., 2003) convertido para a gramática do projeto UD, com o objetivo de eliminar a mesma pouca especificação entre verbo e adjetivo nas formas participiais em certos contextos. Essa pequena modificação levou a um aumento na acurácia de um dado sistema que, sem a etiqueta PCP, conseguia acurácia de 0,9607 na atribuição de POS, e com a etiqueta alcançou 0,9624. Embora seja uma melhora sutil, dá suporte à tese de Manning de que uma seara a ser explorada para melhorar o desempenho dos anotadores está na modelagem linguística. Além disso, Freitas et al. (2018) concluem também que, embora seja razoável afirmar que esquemas de anotação granulares sejam mais difíceis de serem aprendidos – como vimos, sistemas precisam generalizar a partir de exemplos categorizados em classes amplas, de modo que novas etiquetas tendem a dificultar a criação de agrupamentos robustos –, é também verdade que fenômenos pouco especificados, quando são de difícil distinção, podem se beneficiar de uma classe específica para agrupá-los, como é o caso dos PCP.

A segunda classe de erros de anotação identificada por Manning é a "gold standard inconsistent or lacks guidance". 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 – por isso, retomamos a importância

de conceber o esquema de anotação e documentá-lo tendo em vista o corpus e o objetivo da tarefa. Em decorrência da falha na documentação, os anotadores foram inconsistentes, resultando nos erros do sistema automático. Por exemplo, na frase "Orson Welles's Mercury Theater in the '30s", não havia indicação na documentação se o número deveria ser anotado como número cardinal ou como substantivo, uma questão que não é derivada de fronteiras linguísticas pouco delimitadas como na primeira classe, 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 – nesse caso, o esquema é claro e a documentação direciona os anotadores corretamente, mas ainda assim a anotação padrão ouro está incorreta por quaisquer 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 padrão ouro estava incorreta, Manning precisou realizar a sua própria análise, que julgou ser a correta de fato. Nesse caso em especial, o autor pôde julgar se um erro era considerado de fato 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. No entanto, em outro contexto, um gabarito humano é ainda uma interpretação humana e, como lembra Sampson e Babarczy (2008), "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" (p. 472). Contudo, embora não exista anotação correta, existem algumas estratégias para averiguar o quão alinhados ao esquema de anotação estão os anotadores, tornando a anotação consistente, e o quão adequado é um esquema de anotação a uma tarefa.

Uma estratégia é realizar um tipo de avaliação chamada concordância entre anotadores. Se, por um lado, um gabarito também é uma interpretação humana, e às vezes o próprio objetivo da anotação é a construção de um gabarito que ainda não existe, a única forma de medir consistência no processo de anotação, garantindo que todos os anotadores estão compreendendo o esquema de anotação do mesmo jeito e estão conseguindo aplicá-lo adequadamente, é comparando as anotações fornecidas pelos anotadores sobre os mesmos fenômenos. Freitas (2022) indica que é comum que esse tipo de estratégia seja utilizado em uma pequena amostra do corpus por conta da quantidade de recursos humanos necessários para o trabalho, e existem algumas métricas que podem ser utilizadas para medir convergência e divergência entre as anotações,

como as da família *kappa/alpha*, que são as mais utilizadas pois desconsideram do cálculo as concordâncias que podem ter ocorrido por acaso, e não por consistência de fato (ARTSTEIN, 2017).

Outro modo de avaliar a concordância entre anotadores foi descrito em Afonso (2004a) durante o desenvolvimento do recurso Floresta Sintá(c)tica. A avaliação tinha o objetivo de não apenas medir a qualidade da anotação, mas também indicar caminhos para melhoria do recurso e também da ferramenta de anotação automática. Assim, partindo de um baseline – uma anotação básica realizada por um anotador automático –, um anotador fez a revisão do material indicando as categorias que precisariam ser alteradas ou adicionadas a cada token. Posteriormente, um segundo anotador realizou um processo que foi chamado de "re-revisão": tendo em mãos tanto a análise original do parser quanto as modificações realizadas pelo revisor 1, julgou os tokens que ou não foram modificados pelo revisor 1 e precisariam de modificação, ou foram modificados pelo revisor 1 mas ainda estavam incorretos.

Como resultado, o revisor 1 realizou 2.110 modificações à análise original – o maior número de modificações – enquanto que o revisor 2 realizou apenas 1488 modificações à análise original – concluindo que quase 60% das correções foram realizadas pelo revisor 1. Uma pequena parcela de modificações do revisor 2 foi revisão de erros do revisor 1 – 363 foram os tokens modificados pelo segundo revisor que haviam sido erroneamente modificados pelo primeiro. Além de possibilitar aferir consistência a partir da concordância entre os anotadores, esse estudo permitiu uma descrição detalhada de quais tipos de modificações cada um dos revisores introduziu ao corpus, direcionando os esforços para as áreas mais difíceis da anotação tanto para anotadores quanto para o sistema.

A concordância entre anotadores é uma forma de medir o quanto as anotações são confiáveis. Utilizando a mesma ideia, Sampson e Babarczy (2008), anotadores do treebank SUSANNE, que contém um dos esquemas de anotação considerado dos mais bem definidos e consistentes (p. 472), se propõem a testar o quanto das divergências entre dois anotadores é derivado de limitações na definição do esquema de anotação ou da aplicação desse esquema mesmo que por profissionais já bastante familiarizados com as diretivas de anotação. Os autores entendem que, embora seja tentador pensar que anotadores divergem na anotação porque têm convenções gramaticais diferentes, bastando a um gabarito indicar a gramática "correta" para o projeto, esse mesmo gabarito foi feito por pessoas suscetíveis aos mesmos questionamentos e incertezas sobre questões gramaticais mal definidas, portanto não havendo garantia de que o esquema de anotação ideal, perfeitamente bem definido, existe.

Os autores concluem que embora exista uma inerente vagueza (*fuzziness*) nas questões linguísticas mais complexas, limitando o grau de precisão humana na interpretação dos fenômenos linguísticos, mesmo dentro dos limites mais razoáveis os anotadores humanos carecem de habilidade para distinguir onde a nuvem que torna os fenômenos nebulosos começa e onde termina, na analogia proposta em Babarczy, Carroll e Sampson (2006).

Após realizar uma categorização das discrepâncias entre ambos os anotadores do estudo, os resultados são os seguintes:

(A) 58,8% das discrepâncias é fruto de violação de um ponto explícito no esquema de anotação

   a. 38,4% são erros por negligência de detalhes do esquema
   b. 11,9% são erros por negligência do sentido do texto

(B) 19,7% das discrepâncias é fruto de falta de definição no esquema de anotação

   a. 19,4% porque o esquema é vago sobre os limites entre as classes
   b. 0,3% porque o esquema é contraditório

(C) 21,4% das discrepâncias é fruto de ambiguidade estrutural do texto

   a. em 7,8% dos casos, anotações diferentes correspondem a significados linguísticos diferentes
   b. em 13,6% dos casos, anotações diferentes são apenas diferenças lógicas sem significado real diferente

Desse modo, concluímos que a categoria de discrepâncias mais comum é a (A).(a), quando os anotadores negligenciam detalhes do esquema, mesmo quando os anotadores são muito bem familiarizados com ele, como é o caso do estudo. Nesse cenário de erros por distanciamento do esquema de anotação por parte dos anotadores, realizar um teste de concordância entre anotadores antes da tarefa de anotação pode ajudar a indicar o grau de confiança das anotações, mas deve-se ter em mente que há limites humanos para a aplicação de esquemas de anotação mesmo que bem definidos e os anotadores sejam bem treinados. Além disso, nos casos nebulosos (C), na maioria das vezes (b) é irrelevante, do ponto de vista linguístico, distinguir uma classe de outra pois não têm significado linguístico real. Do ponto de vista do PLN, porém, pode ser interessante uma vez que melhora o desempenho dos modelos de aprendizado de máquina.

Por fim, em relação à adequação do esquema de anotação aos objetivos da tarefa, existem três avaliações que podem ser feitas. Às vezes o objetivo

da tarefa é simplesmente realizar a anotação de alguma informação linguística com qualidade similar à humana. Nesse caso, avaliar o desempenho de um sistema baseado em regras ou em aprendizado de máquina, ao mesmo tempo que avalia a qualidade do sistema do ponto de vista computacional, avalia também a adequação do esquema de anotação à tarefa uma vez que um esquema inadequado pode não ser generalizado. É importante salientar, porém, que nesse cenário é necessário distinguir os erros de anotação automática que são fruto de deficiências do sistema daqueles que são deficiências do esquema de anotação, como Manning (2011) fez no contexto da anotação de POS.

Esse tipo de avaliação é chamado de intrínseca pois mede o desempenho interno de uma tarefa: utiliza-se, por exemplo, um *parser* (um sistema de anotação sintática) treinado em um treebank (um recurso anotado sintaticamente) para avaliar o desempenho dos recursos na mesma tarefa, a anotação sintática. Para realizar a avaliação intrínseca, é necessário possuir um *benchmark* – um recurso com anotação padrão ouro que servirá de gabarito. O recurso pode servir tanto como treinamento para sistemas de aprendizado de máquina como para avaliação, comparando-se os resultados do anotador às anotações do *benchmark*. No caso do treinamento de sistemas de IA, é necessário que o conteúdo do treinamento não seja o mesmo da avaliação, pois espera-se que o sistema seja capaz de generalizar o aprendizado de modo a acertar as análises em textos com que não havia tido contato durante o treinamento. As métricas utilizadas em uma avaliação intrínseca são a precisão, a abrangência e a medida F – uma média harmônica entre precisão e abrangência.

Ainda na esteira da avaliação de sistemas, Freitas (2022) observa que tornou-se prática no PLN a realização de avaliações conjuntas para medir o desempenho de diversos sistemas para uma mesma tarefa. A ideia é a de que os desenvolvedores de um sistema podem não ser as pessoas mais confiáveis para testá-lo e nem dispõem de meios para fazê-lo de uma forma neutra e comparável com a avaliação de outros sistemas. Comparar os resultados com os de outros desenvolvedores pode incentivar a melhoria dos índices com o compartilhamento dos conhecimentos adquiridos e dos recursos desenvolvidos. Santos (2007) indica, como alguns dos objetivos da avaliação conjunta, a investigação de uma tarefa como um todo – o que envolve definir o que deve ser avaliado, isto é, a motivação, bem como a metodologia da avaliação que será aplicada nos competidores. O sistema vencedor da competição é aquele que alcança os melhores resultados, que passarão a se chamar estado da arte. Outro objetivo da avaliação conjunta que beneficia a comunidade para além da divulgação de resultados de desempenho dos sistemas diz respeito à criação de recursos reutilizáveis – tanto os sistemas quanto os dados linguísticos que os

alimentam (*datasets*) e os avaliam (*benchmarks*) devem idealmente se tornar disponíveis à comunidade de modo a tornar os resultados reprodutíveis em um paradigma científico. Quando há anotação nos dados, como já vimos, a avaliação conjunta ajuda a pensar a melhor forma de adequá-la aos objetivos da tarefa, pois enquanto não há apenas uma anotação correta, existe aquela que dá origem aos melhores resultados.

Por fim, mais uma forma de verificar o desempenho dos sistemas e a adequação de um esquema de anotação à tarefa é a realização de uma avaliação extrínseca, isto é, utilizar os recursos anotados como etapa intermediária para a realização de alguma tarefa que não se espelha diretamente nos dados anotados em si, mas que se alimenta deles. Quando um projeto de anotação tem como objetivo a resolução de tarefas como a extração de informação. de opinião, de humor ou qualquer outro contexto de uso, seria adequado que o esquema de anotação – em qualquer um dos níveis linguísticos, como na morfologia ou na sintaxe – fosse adaptado para obter os melhores resultados na tarefa final, a despeito de alinhamento com qualquer teoria linguística ou de resultados de avaliação intrínseca.

Um exemplo de tarefa do PLN que pode ser utilizada como avaliação extrínseca para outras tarefas é a Extração de Informação Aberta (EIA). A EIA tem como objetivo a extração de dados estruturados utilizando dados não estruturados (um texto corrido, por exemplo). Essas informações vêm no formato de triplas, como no exemplo “Fulano é aluno de mestrado da Universidade”, do qual se pode extrair as informações (“Fulano”, “é aluno”, “da Universidade”) e (“Fulano”, “é aluno”, “de mestrado”). Contudo, da frase “Se ele tirar 10 na prova, Fulano será aprovado na disciplina” não se pode extrair a informação (“Fulano”, “será aprovado”, “na disciplina”), pois esta é condicionada a uma outra informação por subordinação.

Os exemplos são os empregados por Baia, Prates e Claro (2020) e ilustram a importância da análise sintática – uma análise correta – para que se extraiam informações válidas na EIA. O objetivo dos autores é avaliar anotadores de dependências sintáticas baseando-se na sua utilidade para a EIA. Foram avaliados 4 sistemas de anotação automática de dependências – os sistemas que melhor pontuaram na avaliação conjunta do CoNLL 2018, utilizando o mesmo material de treinamento –, os quais deveriam fornecer análises sintáticas de um conjunto de 100 frases para o melhor sistema de EIA. Dois anotadores foram responsáveis por verificar as informações extraídas pelo sistema de EIA utilizando como base as anotações sintáticas de cada um dos quatro anotadores automáticos para verificar as informações que são válidas e as que são inválidas, classificando os sistemas de acordo com a utilidade na EIA por meio de diversas

métricas relevantes para a tarefa.

Os autores notam que, embora existam diferentes modelos de anotação sintática, o modelo de anotação de dependências tem obtido precisão e abrangência melhores, sendo o escolhido para a avaliação extrínseca de todos os anotadores automáticos sendo avaliados. Mais uma vez, vemos a importância da escolha de um esquema de anotação adequado para satisfazer as necessidades de uma tarefa, sob pena de ter o seu desempenho prejudicado.

Os resultados de Baia, Prates e Claro (2020) são indicativos também de que nem sempre os sistemas que obtêm melhor desempenho na avaliação intrínseca ou em avaliações conjuntas são os melhores para realizar tarefas específicas, como a EIA. Nesse sentido, o melhor anotador automático na avaliação conjunta mencionada alcançou a terceira posição de quatro na avaliação extrínseca, ao passo que os sistemas nas posições 2 e 3 na avaliação intrínseca subiram para as posições 1 e 2 na extrínseca.

4
Metodologia

4.1
Como anotamos o *treebank*

A anotação morfossintática do PetroGold teve origem automática, tendo passado por extensas revisões humanas, as quais serão explicadas na seção 4.2. O modelo utilizado para anotar o corpus, no entanto, já havia sido customizado para realizar uma boa anotação do tipo de texto que compõe o treebank, e este processo inicial será descrito a seguir.

A primeira etapa na construção do PetroGold foi a familiarização dos anotadores humanos – 4 estudantes de Letras (incluindo o autor desta dissertação) – com as teses e dissertações que iriam compor o treebank. O formato escolhido para anotar o corpus foi o Universal Dependencies, um projeto que já conta com treebanks para mais de 100 línguas, devido à facilidade de processar textos neste formato bem como devido à abundância de ferramentas disponíveis, como concordanciadores e anotadores automáticos, além da já familiaridade dos anotadores com o formato, tendo inclusive desenvolvido um documento com diretivas para anotação do projeto em língua portuguesa (DE SOUZA et al., 2020).

No entanto, alguns elementos do PetroGold ainda eram novidade e precisavam ser discutidos entre a equipe para decidir como anotá-los, como as referências bibliográficas, as notas de rodapé, as listas em tópicos, as fórmulas químicas e matemáticas e afins. Esse processo de familiarização dos anotadores gerou alguns resultados: (1) uma documentação de como a segmentação de frases e tokens deveria ocorrer no PetroGold (CAVALCANTI et al., 2021), (2) uma documentação sobre a anotação morfossintática do PetroGold, com os fenômenos linguísticos típicos de teses e dissertações (DE SOUZA et al., 2021a), (3) a disponibilização de um corpus inteiramente revisto no que diz respeito à segmentação de frases e tokens em documentos acadêmicos do domínio do petróleo, com 1.139 frases selecionadas por apresentarem algum desafio à segmentação automática[1], e (4) a disponibilização de um corpus padrão ouro, inteiramente revisto, de anotação morfossintática UD composto por uma pequena seleção de textos do Petrolês, totalizando 818 frases[2].

Este último recurso, mais especificamente o Petro1, por ser um padrão ouro de anotação morfossintática, alimentou um novo modelo de anotação

[1]PetroTok, disponível em <https://petroles.puc-rio.ai>. Acesso em 10 jan. 2023.
[2]Petro1 e Petro2, disponíveis em <https://petroles.puc-rio.ai>. Acesso em 10 jan. 2023.

automática que desenvolvemos para anotar o PetroGold. Assim, o treebank foi anotado utilizando um modelo customizado do anotador Stanza (QI et al., 2020), treinado utilizando um conjunto de frases que continha as frases do Bosque-UD (RADEMAKER et al., 2017), treebank padrão ouro composto por textos jornalísticos em português brasileiro e europeu, e as frases do Petro1, composto por textos do mesmo tipo do PetroGold. Neste material de treino, as frases do Petro1 responderam por 7% do total de frases do conjunto.

O desenvolvimento do Petro1 possibilitou um processo iterativo de discussão, solução de dúvidas, tomada de decisão e documentação sobre a anotação de questões específicas do tipo de texto do Petrolês. Ao final, calculamos a concordância interanotadores, uma forma de avaliar o quanto os anotadores estão sintonizados e anotando os mesmos fenômenos da mesma forma, garantindo consistência e, paralelamente, garantindo também que o esquema de anotação está adequado.

Para isso, utilizamos a métrica *Cohen's Kappa* (κ), uma forma de medição da confiança da anotação que é robusta pois, além de considerar a quantidade de anotações iguais entre os pares de anotadores, não descarta a possibilidade de que anotações aleatórias também sejam iguais (ARTSTEIN, 2017). O melhor par na tarefa de anotação de relações sintáticas (a tarefa mais complexa) obteve um resultado de 95,1% de concordância, enquanto o pior par (a dupla de anotadores com mais divergências) obteve um resultado de 91,9%.

4.1.1 A abordagem *Universal Dependencies*

Tanto o formato quanto a gramática escolhidos para a anotação morfossintática do PetroGold são os disponibilizados pelo projeto Universal Dependencies, um *framework* atualizado semestralmente e que, na sua versão 2.11 (de novembro de 2022), já contava com mais de 200 treebanks para mais de 100 línguas.

A escolha por UD se deu por motivos práticos: por um lado, a difusão de estudos linguístico-computacionais utilizando a gramática e os recursos do projeto, o que resultou na consolidação de uma comunidade ativa, engajada no debate das diretivas gramaticais e mesmo dos rumos do projeto[3], além de uma profusão de ferramentas computacionais para lidar com recursos UD. Por outro lado, minha experiência anterior, trabalhando no desenvolvimento das versões 2.4, 2.5 e 2.6 do Bosque-UD (RADEMAKER et al., 2017), facilitaria

[3]As discussões da comunidade de UD em língua portuguesa, da qual faço parte, foram muito relevantes no desenvolvimento e no alinhamento das soluções para análise das questões linguísticas presentes nesta dissertação.

o processo de adaptação do conhecimento já adquirido para um novo projeto – a anotação do PetroGold em UD.

Segundo Marneffe et al. (2021), UD é, ao mesmo tempo, um *framework* para anotação morfossintática consistente entre diferentes línguas, um esforço coletivo, de comunidade aberta, para criar corpora anotados, um repositório para armazená-los. Como um projeto aberto à comunidade, as direções do projeto e mesmo as diretivas de anotação – a gramática UD – estão abertas para discussão entre os mais de 300 contribuidores, grupo do qual faço parte.

Enquanto um *framework* para anotação morfossintática, os autores esclarecem que a gramática UD não é eclética, mas uma teoria coerente e desenvolvida pela comunidade ao longo do tempo, cujo objetivo é ser aplicável a línguas tipologicamente distintas. Essa anotação deve ser útil, ao mesmo tempo, para *Natural Language Understanding* (NLU, compreensão de linguagem natural) e para estudos linguísticos mais amplos.

A sintaxe do UD utiliza um modelo chamado de gramática de dependências, em que cada sintagma tem um núcleo e todas as outras palavras são seus dependentes. A gramática de dependências adota uma relação binária e assimétrica entre as palavras de uma frase, sendo que o projeto representa as relações de dependências com setas partindo das palavras que governam as relações em direção aos seus dependentes, como na figura 4.1, obtida de Marneffe et al. (2021)[4].

De maneira informal, pode-se dizer que os núcleos das relações de dependência são as palavras mais importantes dos sintagmas. Sintagmas nominais têm núcleos nominais, orações têm núcleos verbais (exceto quando o predicado é nominal) etc. E a escolha do projeto UD foi por colocar esses núcleos como os governantes das relações de dependência, de tal maneira que todas as demais palavras funcionais são dependentes das palavras de conteúdo lexical.

Uma consequência (e também uma das causas) dessa escolha dentro do paradigma da gramática de dependências é que, dessa forma, as representações linguísticas privilegiam as estruturas de predicado-argumento, que estão no centro das dependências, em detrimento da observação da estrutura interna de cada sintagma individualmente. Segundo os autores, essa escolha é importante para a comparação entre diferentes línguas – na figura 4.1, duas línguas tipologicamente muito distintas, inglês e finlandês, têm uma representação de dependências similar, uma vez que a estrutura predicado-argumento é a mesma, a despeito de as palavras funcionais utilizadas em cada língua serem

[4]Neste trabalho, podemos utilizar os termos "núcleo", "head" ou "governante" para designar os núcleos das relações de dependência sem nenhuma diferenciação no significado.

muito distintas. O mesmo ocorre com uma frase em português, recortada do PetroGold (figura 4.2).

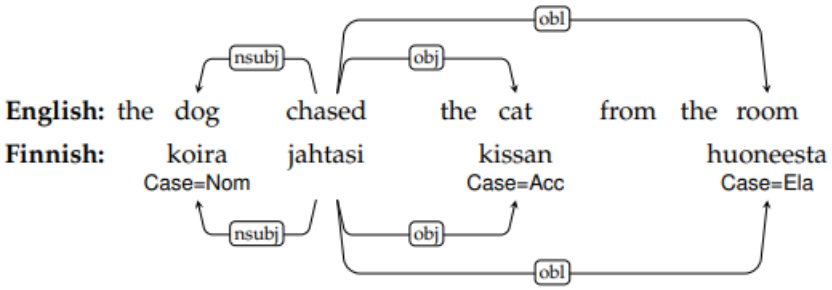


Figura 4.1: Frases de línguas diferentes anotadas usando a gramática UD. Fonte: Marneffe et al. (2021).

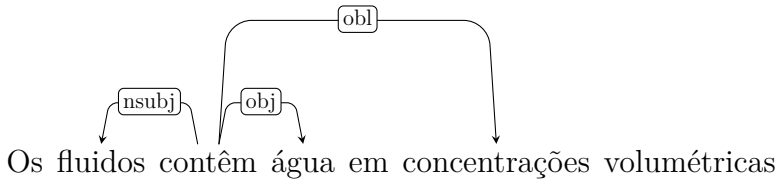


Figura 4.2: Anotação de dependências na gramática UD de parte de uma frase do PetroGold

O formato escolhido pelo projeto para codificar a anotação dos treebanks é o CoNLL-U, uma versão revisada do CoNLL-X (BUCHHOLZ; MARSI, 2006). Na figura 4.3, a frase "Os fluidos contêm água em concentrações volumétricas", a mesma da figura 4.2, foi anotada nesse formato, onde se vê uma palavra por linha e uma informação linguística distinta por coluna.

A primeira coluna de cada palavra (ou token, pois já passaram pelo processo de *tokenização*, isto é, transformação do texto em unidades mínimas de processamento) tem um número, o qual representa o identificador do token na frase. A segunda coluna apresenta a forma da palavra tal qual ela aparece no texto, e a terceira coluna mostra o lema do token, isto é, a forma de dicionário (não flexionada e em masculino singular). A quarta coluna é reservada à classe gramatical (*Universal Part-Of-Speech tag*, ou UPOS) da palavra e a quinta coluna à classe gramatical da palavra em outro formato de anotação, caso o treebank tenha sido importado de outro projeto (a coluna não é obrigatória – e não é utilizada no PetroGold –, podendo ser preenchida por um *underscore*). A sexta coluna contém informações flexionais da palavra, como gênero, número e tempo verbal, separadas por uma barra vertical. A sétima coluna indica o token do qual ele depende, representado pelo seu número de identificação, e a oitava coluna indica de que tipo é essa relação de dependência entre o token e o seu governante (*dependency relation*, ou DEPREL). A nona coluna

sent_id = 1

text = Os fluidos contêm água em concentrações volumétricas

1	Os	o	DET	_	Definite=Def|Gender=Masc|Number=Plur|PronType=Art	2	det	_	TokenRange=0:2
2	fluidos	fluido	NOUN	_	Gender=Masc|Number=Plur	3	nsubj	_	TokenRange=3:10
3	contêm	conter	VERB	_	Mood=Ind|Number=Plur|Person=3|Tense=Pres|VerbForm=Fin	0	root	_	TokenRange=11:17
4	água	água	NOUN	_	Gender=Fem|Number=Sing	3	obj	_	TokenRange=18:22
5	em	em	ADP	_	_	6	case	_	TokenRange=23:25
6	concentrações	concentração	NOUN	_	Gender=Fem|Number=Plur	3	obl	_	TokenRange=26:39
7	volumétricas	volumétrico	ADJ	_	Gender=Fem|Number=Plur	6	amod	_	SpaceAfter=No|TokenRange=40:52

Figura 4.3: Anotação de parte de frase do PetroGold no formato CoNLL-U

Open class words	Closed class words	Other
ADJ	ADP	PUNCT
ADV	AUX	SYM
INTJ	CCONJ	X
NOUN	DET	
PROPN	NUM	
VERB	PART	
	PRON	
	SCONJ	

Figura 4.4: Lista de classes gramaticais (UPOS) do projeto UD. Fonte: Captura de tela.

não é obrigatória (e não é utilizada no PetroGold), estando reservada a informações sintáticas mais detalhadas, como quando o token pode ter mais de um governante, por exemplo nos casos de coordenação ou elipse. A décima coluna, por fim, é destinada a informações extras acerca dos tokens, como informações para treinamento do tokenizador automático.

A gramática UD na versão 2 conta com 17 classes gramaticais e 37 relações de dependências, como mostram as figuras 4.4 e 4.5[5]. Durante este trabalho, usaremos a nomenclatura da nossa gramática tradicional (GT) para explicar os fenômenos linguísticos e, quando estivermos nos referindo às anotações realizadas no corpus, usaremos as etiquetas correspondentes de modo a facilitar a leitura, apontando as diferenças entre GT e UD quando necessário.

A única diferença entre a anotação do PetroGold (utilizada nesta dissertação e disponível no endereço do projeto Petrolês) e a gramática UD (publicada na versão 2.11 do projeto UD) é a criação da etiqueta de relação sintática "nmod:appos", que originalmente não existe no projeto UD, tendo sido criada por nós tendo em vista a tarefa de anotação de entidades mencionadas, uma

[5]Figuras obtidas, respectivamente, dos endereços <https://universaldependencies.org/u/pos/index.html> e <https://universaldependencies.org/u/dep/index.html>. Acesso em 10 jan. 2023.

	Nominals	Clauses	Modifier words	Function Words
Core arguments	nsubj obj iobj	csubj ccomp xcomp		
Non-core dependents	obl vocative expl dislocated	advcl	advmod* discourse	aux cop mark
Nominal dependents	nmod appos nummod	acl	amod	det clf case
Coordination	**MWE**	**Loose**	**Special**	**Other**
conj cc	fixed flat compound	list parataxis	orphan goeswith reparandum	punct root dep

Figura 4.5: Lista de relações de dependência (DEPREL) do projeto UD. Fonte: Captura de tela.

etapa posterior do projeto Petrolês. A etiqueta é utilizada para os termos cuja relação com o substantivo governante é clara ao leitor, mas não está explícita na oração.[6]

Por exemplo, na frase 8, sabemos que "MG" é o estado em que se situa a cidade de Presidente Olegário, a despeito de não haver nenhuma sinalização na frase de que o primeiro termo é uma cidade, e o segundo a sigla de um estado. Sabemos disso pelo nosso conhecimento de geografia ou do padrão de escrita "cidade (UF)". O fenômeno é diferente do aposto, pois este tem uma relação de identificação com o substantivo; é diferente do *nmod*, pois este modifica um substantivo e o tipo de modificação é expressa pela preposição, por exemplo; e não é *parataxis*, pois nestes casos o tipo de relação que o termo mantém com o governante não é claro, nem mesmo para o leitor.

8. Próximo a Presidente Olegário (**MG**) foram descritos em este estudo pacotes siliciclásticos relativamente espessos (até 60 m) pertencentes a esta formação .

9. Para fluidos Newtonianos , como a **água** e o ar , a viscosidade é independente de a taxa de cisalhamento .

[6]Mantemos, portanto, duas versões do PetroGold – uma para o projeto UD e outra para as questões específicas do projeto Petrolês, ambas disponíveis nos respectivos sites. Embora atualmente com apenas esta diferença, a versão que consideramos mais completa é a do Petrolês.

10. Esses últimos autores postulam a hipótese de ambiente transicional (laguna) , com periódicas ingressões marinhas para o paleoambiente de o Membro Romualdo e , ainda que não totalmente comprovada tal hipótese , esta encontra subsídios em as evidências paleontológicas , pois a presença concomitante de fauna marinha , como **dinoflagelados** e foraminíferos , atesta a influência marinha em a área .

11. A capacidade de absorção de o solvente é proporcional a a pressão parcial em a unidade de absorção (Gupta , **2003**) .

12. Santos (**2003**) estudou experimentalmente o comportamento de molhabilidade de superfícies de aço carbono , aço galvanizado , aço inoxidável e vidro borassilicato em meio aquoso e em solução de 1 % (p/p) metassilicato de sódio .

Além disso, utilizamos a etiqueta também para estruturas de hiperonímia, quando um termo faz parte de um campo semântico expresso por um substantivo, como na frase 9, em que "água" e "ar" são termos que participam do campo semântico de "fluidos Newtonianos", e a frase 10, onde "dinoflagelados" e "foraminíferos" participam do campo semântico de "fauna marinha". Essa etiqueta é importante sobretudo pois auxilia na anotação de entidades nomeadas no PetroGold, uma vez que a própria estrutura linguística pode indicar quais termos fazem parte de uma classe de entidades.

Finalmente, a etiqueta também é utilizada para a data de publicação de referências bibliográficas no corpo do texto, como nos exemplos 11 e 12. Todas as referências bibliográficas com ano de publicação, esteja o ano entre parênteses ou não, têm o ano anotado como "nmod:appos" dependente da primeira palavra da referência bibliográfica.

Para o lançamento do corpus no projeto Universal Dependencies, todas as etiquetas "nmod:appos" são convertidas em "nmod". A conversão não é adequada, pois os sintagmas não são adjuntos adnominais propriamente, mas essa é a única leitura possível dentro da abordagem UD.

Além disso, a divisão de frases nas partições de treinamento, teste e desenvolvimento do PetroGold é diferente no projeto Petrolês e no projeto UD: no projeto Petrolês, selecionamos frases aleatórias para cada uma das partições e mantemos a mesma divisão para a primeira, a segunda e a terceira versão do corpus, de maneira a permitir a comparação; já no projeto UD, há a preferência por manter, na medida do possível, documentos inteiros em cada

partição[7]. Mais informações sobre como realizamos a avaliação do PetroGold nesta dissertação serão apresentadas na seção 4.3.

4.2 Como revisamos o *treebank*

Utilizamos cinco métodos distintos para a revisão da anotação do PetroGold. Alguns dos métodos foram totalmente manuais, necessitando de exploração do corpus para identificar os erros e para corrigi-los; outros foram semiautomáticos na detecção, utilizando os dados do próprio corpus, submetidos a determinados algoritmos, para identificar potenciais erros. Em nenhum dos casos a correção foi automática, sendo sempre necessária a interferência humana para indicar a anotação correta, embora em alguns dos métodos já seja sugerido, automaticamente, como corrigir a anotação.

Os três métodos semiautomáticos que aplicamos – regras linguísticas, busca por *n-grams* inconsistentes e IAD (Inter-Annotator Disagreement) – foram descritos e testados em relação à precisão, abrangência e eficiência em um trabalho anterior utilizando uma fração totalmente revista do Petrolês (FREITAS; DE SOUZA, 2023). Na seção 5.3 avaliaremos os métodos utilizando as mesmas métricas, já tendo sido finalizada a revisão do PetroGold v3, para indicar qual foi a contribuição de cada um deles na construção do recurso.

4.2.1 Consulta a gramáticas

O primeiro método utilizado foi a consulta a gramáticas para auxiliar na anotação de questões gramaticais pouco discutidas nas diretivas do projeto UD para língua portuguesa, sendo necessário um estudo sistemático do português e dos dados no corpus. Trata-se de um tipo de estudo comumente realizado na área denominada Linguística de Corpus, por meio da descrição do material e da língua, com o auxílio de dicionários e gramáticas tradicionais (GT). Uma vez identificadas questões linguísticas mal anotadas no corpus por meio da exploração do material, procuramos soluções linguísticas que melhorassem o aprendizado automático mas que também fossem linguisticamente corretas e, na medida do possível, que não se distanciassem do conhecimento compartilhado representado pela GT.

Não é o objetivo das gramáticas tradicionais providenciar instruções sobre a anotação de um corpus – muito menos no formalismo UD e de textos do domínio do petróleo, como é o nosso caso. Contudo, muitas gramáticas

[7]Uma discussão sobre os motivos que levam o projeto UD a preferir esse modelo de particionamento dos treebanks pode ser encontrada no endereço: <https://github.com/UniversalDependencies/UD_Portuguese-PetroGold/issues/3>. Acesso em 10 de mar. 2023.

baseiam suas análises na descrição dos fenômenos linguísticos, e embora essas análises não possam ser facilmente transpostas para a anotação de corpus, elas fornecem um ponto de partida para pensar os fenômenos, muitas vezes com bons exemplos, que nos auxiliam na tomada de decisão quando confrontamos as análises com os exemplos do corpus.

De modo geral, em primeiro lugar nós consultamos as diretivas do projeto UD, pois são o principal guia para a anotação morfossintática do treebank. Uma vez cientes de todas as decisões e lacunas das diretivas para língua portuguesa, investigamos as gramáticas do português de modo a encontrar o que já se havia discutido e estabelecido como padrão de análise para as questões em foco. Paralelamente, consultas ao PetroGold iam nos mostrando em que pontos as análises tradicionais poderiam ser aplicadas satisfatoriamente e, quando não fosse possível alinhar as GTs com as construções do corpus, procuramos soluções alternativas que fossem consistentes e capazes de melhorar o aprendizado automático da anotação. Por fim, aplicamos as modificações ao corpus, utilizando metodologia específica para cada uma das questões, que serão aprofundadas na seção 5.2.

4.2.2 Consulta a léxico

O segundo método utilizado para revisar a anotação do corpus foi a consulta a um léxico computacional. O léxico foi utilizado nas questões relativas à lematização, à anotação de POS e à atribuição de características morfológicas. O léxico foi disponibilizado pelo projeto POeTiSA e se chama PortiLexicon-UD (LOPES et al., 2022). O recurso é gratuito e inclui 1.221.218 entradas (palavras ambíguas foram contabilizadas como entradas diferentes quando tinham anotações distintas) com informações morfológicas de acordo com a gramática UD.

Comparamos todas as entradas do léxico com todos os tokens no PetroGold considerando quatro das informações presentes na anotação no corpus: forma da palavra, lema, POS e atributos morfológicos. Para as formas das palavras que encontramos no léxico[8], verificamos se alguma das anotações do PetroGold não era prevista pelo léxico e, quando era o caso, corrigimos o corpus. Para comparar o léxico e corpus, desconsideramos as palavras cuja etiqueta de POS fosse PROPN, NUM ou X (nomes próprios, numerais e palavras estrangeiras, respectivamente), pois poucas dessas palavras existiam no léxico.

[8]Nem todas as palavras do PetroGold foram encontradas no léxico, principalmente as que dizem respeito ao domínio do petróleo.

4.2.3 Regras linguísticas

O terceiro método utilizado para revisão do treebank foi a criação de regras linguísticas para detecção de padrões que podem indicar erros na anotação. Como discutido na seção 3.2, esse é um tipo de método orientado por conhecimento linguístico, pois foi necessário o desenvolvimento de um conjunto de regras que representassem padrões de erro na gramática utilizada para anotar o treebank. As regras foram desenvolvidas tendo como base (i) o conhecimento da gramática UD, (ii) o conhecimento da gramática do português, (iii) a exploração dos erros mais comuns do anotador automático (assunto que será discutido na seção 6.1), e (iv) a exploração de erros detectados por outros métodos e que puderam se transformar em regras de detecção. Não há garantia de que a lista de regras desenvolvidas seja abrangente e, além disso, padrões estranhos nem sempre são erros na anotação, sendo necessária verificação humana para corrigir os erros identificados pelo método.

As regras se dividem em quatro grandes grupos[9]. O primeiro grupo de regras diz respeito ao formato de anotação utilizado no projeto UD, e tem como objetivo corrigir sobretudo os erros introduzidos pelos anotadores durante a revisão do corpus. Embora tenham sido devidamente treinados e a concordância interanotadores, conforme mostramos na seção 4.1, seja alta, falhas humanas, como erros de digitação, podem acontecer, principalmente quando não estão utilizando ferramentas adequadas.

Por exemplo, a regra (a) da tabela 4.1 procura pelos tokens cujo identificador do token que o governa (*head_token.id*) seja o próprio identificador do token. É uma regra que busca por ciclos, quando um token é dependente de si mesmo. É um erro grave, que inutiliza a árvore sintática da frase, mas que é comumente introduzido sem que o anotador humano perceba. Além disso, a regra indica que o erro é na coluna *dephead* e apresenta uma mensagem indicando qual o erro encontrado.

O segundo grupo de regras engloba aquelas cuja revisão pode ser semi-automática, pois a própria natureza do erro já indica qual deve ser a correção aplicada. No exemplo (b) da tabela, busca-se por um token cuja etiqueta de POS seja AUX ou VERB (verbos auxiliares ou principais) e cuja anotação de atributos morfológicos não inclua o valor *VerbForm*, destinado à informação de forma verbal (forma finita, infinitivo, particípio ou gerúndio).

O terceiro grupo de regras destina-se àquelas cuja solução precisa de uma

[9] As regras de detecção de erros são escritas na linguagem Python, atualmente somam 64 e são constantemente atualizadas no endereço: <https://github.com/alvelvis/ACDC-UD/blob/master/validar_UD.txt>. Acesso em 10 jan. 2023.

Tabela 4.1: Quadro com os 4 grupos de regras de detecção de erros

(a) FORMAT VALIDATION
dephead|erro: Token dependent on itself
token.head_token.id == token.id

(b) SEMIAUTOMATIC REVISION
feats|erro: VERB should have the "VerbForm" feature
upos = "(AUX|VERB)" and feats != ".*VerbForm=.*"

(c) LINGUISTIC REVISION
deprel|erro: Adjective has wrong number
deprel = "amod" and head_token.number != number

(d) NEED THOROUGH ANALYSIS OF THE TREE
upos|erro: Conj should be dependent on same POS
@deprel = "conj" and head_token.upos != upos

análise linguística do contexto. Por exemplo, na regra (c) da tabela, o padrão procurado é o de adjuntos adnominais do tipo adjetivo (*amod*) cujo token governante tem a informação morfológica de número diferente da informação de número do token dependente. A regra baseia-se na ideia de que adjetivos devem concordar em número com os substantivos que modificam, no entanto, para solucionar o erro da frase, o anotador deve optar entre três escolhas: (i) modificar a anotação de número do substantivo, que pode estar errada; (ii) modificar a anotação de número do adjetivo, que pode estar errada, ou ainda (iii) manter a anotação como está, pois embora adjetivo não concorde com substantivo, esse foi um erro de escrita introduzido pelo autor do texto e que não pode ser solucionado na anotação.

O quarto grupo de regras é mais complexo, e exige que o anotador faça uma análise mais abrangente de toda a frase para entender por que houve a detecção de um possível erro. Por exemplo, na regra (d) da tabela, a busca é por tokens com anotação sintática de *conj* (um sintagma nominal ou verbal coordenado) cuja etiqueta de POS seja diferente da etiqueta de POS do seu governante. Ainda que haja um consenso linguístico de que elementos coordenados tendem a ter a mesma classe gramatical, essa não é uma exigência nem da gramática UD, nem da língua em uso – com frequência, fazemos coordenações de palavras de classes diferentes. Cabe ao revisor, portanto, identificar se o que foi detectado pela regra não é um erro, se a classe gramatical de algum dos elementos coordenados está errada, ou ainda se a relação entre os elementos não é de coordenação.

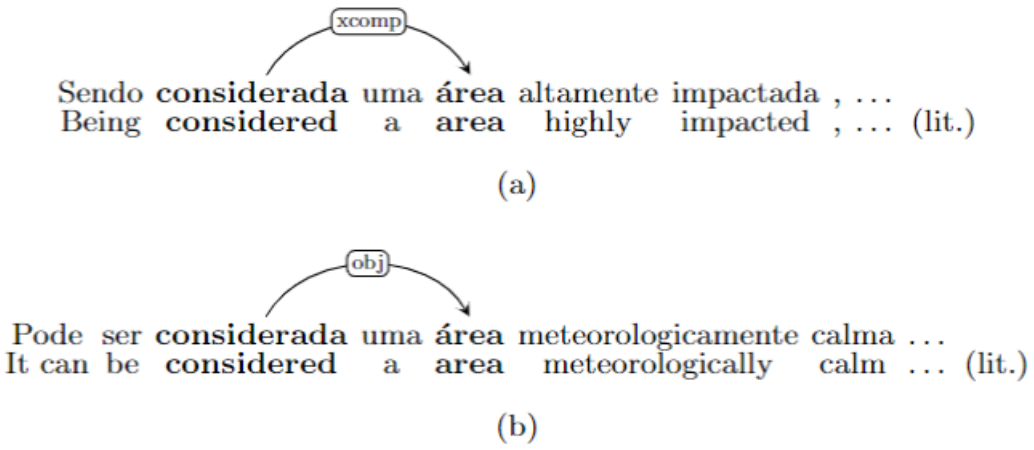


Figura 4.6: Duas frases com os mesmos lemas relacionados mas com anotações sintáticas distintas – uma delas está errada

4.2.4 Busca por *n-grams* inconsistentes

O quarto método para revisar as frases do PetroGold utiliza o princípio discutido na seção 3.2 de busca por variação na anotação de trechos de texto semelhantes – ou seja, buscamos por "*n-grams* inconsistentes", forma como chamamos o método. Assim como de Marneffe et al. (2017), buscamos pelos lemas das palavras para encontrar *n-grams* iguais e não exigimos que tenham a mesma etiqueta de POS. Por exemplo, na figura 4.6, obtida de Freitas e de Souza (2023), encontramos duas frases com o mesmo par de lemas (considerar, área), no entanto, a relação de dependência entre ambos é diferente – na frase (a), "área" é *xcomp* de "considerada", enquanto na frase (b) "área" é *obj* de "considerada". A análise correta para ambas é *xcomp*, por tratar-se de um predicado verbo-nominal, de maneira que a frase (b) precisou ser corrigida para se tornar consistente com a análise de (a).

Para se adequar melhor ao PetroGold (e conforme indicamos em Freitas e de Souza (2023)), fizemos algumas adaptações do método discutido em de Marneffe et al. (2017). Diferentemente do que os autores fizeram, não exigimos que as palavras ao redor do par em análise fossem as mesmas, tampouco que a relação do par de palavras com o restante da frase fosse a mesma em todas as frases em análise (as restrições foram chamadas, respectivamente, de *non-fringe* e *dependency context heuristics* por Boyd, Dickinson e Meurers (2008)), pois vimos que as restrições diminuem drasticamente o número de resultados[10].

Para compensar a queda na precisão que resultaria do relaxamento das

[10]Cabe lembrar que o método de busca por *n-grams* inconsistentes é um método que naturalmente é pouco abrangente, pois só é capaz de detectar inconsistência em *n-grams* que sejam semelhantes, o que pode ser difícil de encontrar em corpora não tão volumosos.

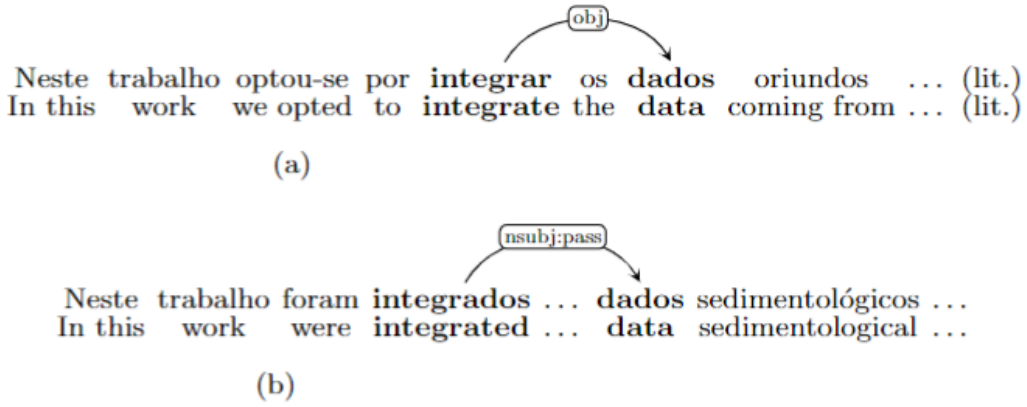


Figura 4.7: Duas frases com os mesmos lemas relacionados mas com anotações sintáticas distintas – nenhuma delas está errada

restrições, adicionamos uma outra que evitaria muitos dos falsos positivos que encontramos na aplicação no PetroGold: só consideramos os pares de palavras como iguais quando, além de terem o mesmo lema, também compartilhavam a mesma informação de voz (passiva ou ativa). Dessa forma, em frases como as da figura 4.7, em que se encontra o par de lemas (integrar, dado), a diferença na anotação entre as palavras nas frases não é retornada como um possível erro pelo algoritmo, pois na primeira frase o verbo está na voz ativa e, na segunda, na voz passiva. De fato, as anotações, apesar de diferentes, estão corretas: na frase (a), "dados" é objeto de "integrar", enquanto na frase (b) "dados" é sujeito paciente de "integrados".

4.2.5 *Inter-Annotator Disagreement* (IAD)

O quinto método utilizado, chamado IAD (*Inter-Annotator Disagreement*), se baseia no contraste entre a anotação de dois anotadores automáticos distintos, treinados utilizando o mesmo corpus (FREITAS; DE SOUZA, 2023). Os anotadores utilizados foram o Stanza (QI et al., 2020) e o UDPipe 1.2.0 (STRAKA; HAJIC; STRAKOVÁ, 2016), ambos treinados no material apresentado na seção 4.1, composto pelo Bosque-UD (RADEMAKER et al., 2017) e por uma fração do Petrolês inteiramente revista. O primeiro anotador automático, Stanza, nós chamamos de "anotação guia", pois fornece resultados melhores que o segundo, UDPipe, a anotação "sistema". A ideia subjacente ao método é a de que dois modelos irão discordar da anotação nos casos mais complexos, que não foram vistos no material de treino ou cuja anotação fornecida no treinamento estava inconsistente.

A anotação guia que utilizamos no método IAD também era a anotação automática original do PetroGold. Assim, quando houve divergência entre a

anotação guia e a anotação sistema, e a anotação guia estava correta segundo o nosso julgamento, nada precisou ser feito. Quando a anotação sistema estava correta, foi necessário corrigir o alvo da divergência. Houve ainda casos em que nenhuma das duas anotações estava correta, sendo necessária a análise humana para decidir qual outra anotação seria correta.

Uma lacuna desse método é que há a possibilidade de os dois anotadores automáticos errarem da mesma forma em um mesmo caso. Esses são os erros que chamamos de "invisíveis", pois os anotadores concordaram na análise e, portanto, apesar de errada, não retornaram a frase para inspeção humana.

4.3 Como avaliamos o *treebank*

Uma estratégia para medir a consistência da anotação de um corpus para o PLN é a partir da avaliação intrínseca de um modelo treinado utilizando o corpus como dataset. Segundo Freitas (2022), 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 do sistema será feita contrastando as anotações do modelo gerado com o gabarito de morfologia. No entanto, essa avaliação é útil não só para avaliar o modelo, mas também, indiretamente, a qualidade de um treebank, pois a avaliação de um modelo que utilizou um padrão ouro como material de treino nos fornece evidências acerca da consistência desse material, uma vez que facilitou ou dificultou o aprendizado automático.[11]

Todas as avaliações intrínsecas são realizadas utilizando o mesmo programa para treinamento do modelo de dependências sintáticas – o programa utilizado é o UDPipe na versão 1.2.0 (STRAKA; HAJIC; STRAKOVÁ, 2016) – e utilizando o mesmo código de avaliação, disponibilizado durante a avaliação conjunta do CoNLL de 2018 (ZEMAN et al., 2018)[12]. Daremos maior atenção às métricas de LEMMA (*lemmatization*), responsável por avaliar a lematização dos tokens, UPOS (*universal part-of-speech score*), responsável pela atribuição de etiqueta de classe gramatical, UAS (*unlabeled attachment score*), responsável pela avaliação do encaixe de dependência sintática sem avaliar a

[11] 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 e gerar bons números na avaliação intrínseca, a despeito de ser incorreta.

[12] Embora nosso objetivo seja apenas comparar o impacto de diferentes anotações do mesmo corpus na avaliação intrínseca, não desconsideramos a importância de utilizar os programas e os códigos mais utilizados pela comunidade (no nosso caso, a comunidade UD), garantindo a reprodutibilidade técnica e a comparação entre os resultados.

etiqueta atribuída para a relação, e LAS (*labeled attachment score*), quando se conta, além do encaixe, também a etiqueta designada.

Além dessas métricas, também utilizaremos CLAS (*content-label attachment score*), que diz respeito ao encaixe e à relação sintática apenas para palavras de conteúdo lexical, excluindo palavras funcionais e pontuações. A métrica foi desenvolvida sob a justificativa de que, sendo o UD um projeto multilíngue, seria necessário utilizar uma métrica que descartasse as 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).

Para nós, é interessante utilizar a métrica CLAS sobretudo quando estamos lidando com fenômenos que não envolvem palavras funcionais e pontuações, uma vez que, sendo mais frequentes no corpus e comumente mais fáceis de serem acertadas, sempre elevam as métricas de avaliação, tornando mais difícil de visualizar o impacto das mudanças realizadas na anotação do corpus. A relação de pontuação (*punct*) não participa do rol de palavras de conteúdo nem funcionais, portanto, embora tenha um grande número de ocorrências nos corpora, não é considerada no cálculo da avaliação intrínseca nem de LAS nem de CLAS.

A divisão de frases entre as partições de treinamento, teste e desenvolvimento segue sempre a proporção de 90%, 5% e 5%, respectivamente. Optamos por essa proporção para alinhar com a versão 2.8 do Bosque-UD (RADEMAKER et al., 2017), de modo a permitir comparações com um dos mais relevantes *datasets* do projeto UD para português. As frases de cada partição foram escolhidas aleatoriamente na versão 1 do PetroGold e seguiram sendo as mesmas para as versões 2 e 3, facilitando a comparação entre as versões, e em todas as avaliações intrínsecas juntamos as partições de treinamento e de desenvolvimento para alimentar o anotador automático.

4.4 Integrando busca, edição e avaliação: a ET

Uma série de ferramentas foi desenvolvida para auxiliar na construção do PetroGold. Essas ferramentas foram sendo incorporadas, ao longo do tempo, em um ambiente integrado que chamamos de ET – uma Estação de Trabalho para busca, edição e avaliação de corpora anotados (DE SOUZA; FREITAS, 2021; DE SOUZA; FREITAS, 2019).

A ferramenta está disponível gratuitamente em português, em inglês, e pode ser utilizada em diversos outros projetos que envolvam corpora anotados.

Contudo, tendo em vista que a ET foi sendo desenvolvida durante a construção do PetroGold, ela não é otimizada para tarefas que não foram necessárias nesse processo. Por exemplo, há ferramentas mais adequadas para a anotação manual de corpora, como Arborator (GERDES, 2013), ConlluEditor (HEINECKE, 2019) e UD Annotatrix (TYERS; SHEYANOVA; WASHINGTON, 2017), assim como há ferramentas também específicas para auxiliar nos estudos linguísticos com grandes corpora, como AntConc (ANTHONY, 2005), CQPWeb (HARDIE, 2012) e AC/DC (SANTOS; BICK, 2000), para nomear algumas[13]. Além disso, deve-se notar que a ET funciona apenas com corpora no formato CoNLL-U, o escolhido para o PetroGold (conforme apresentado na seção 4.1.1).

Um dos módulos mais centrais da ET, porque auxilia em todas as etapas da construção e avaliação de um corpus, é o Interrogatório, o ambiente de busca e edição[14]. O ambiente de busca é otimizado para lidar com a anotação de dependências sintáticas, de maneira que é possível encontrar as mais complexas estruturas linguísticas utilizando as diferentes sintaxes de busca do programa.

Por exemplo, na figura 4.8, onde vemos a página inicial do Interrogatório, há uma expressão de busca designada para encontrar objetos de verbos[15]. A figura 4.9 apresenta o resultado da busca, que retornou 9.346 ocorrências de objetos em 5.718 frases em uma versão inicial do PetroGold.

A busca do Interrogatório conta com uma série de utilidades para auxiliar no estudo de fenômenos linguísticos e para corrigi-los no corpus. Como veremos na seção 5.2.1, um dos estudos realizados nesta dissertação tem como objetivo padronizar e corrigir a anotação de objetos indiretos. Para isso, foi necessário consultar gramáticas, consultar as diretivas do projeto UD e, não menos importante, consultar o corpus, pois sabemos que são os casos reais, encontrados no material que estamos anotando, que devem validar as nossas decisões de anotação.

Para consultar o corpus, além de visualizar as frases retornadas pela expressão de busca, podemos visualizar o seu contexto (isto é, as frases que aparecem antes ou depois delas nos documentos) e visualizar a anotação

[13]Conforme é apresentado em de Souza e Freitas (2021), o *design* da ET foi fortemente influenciado pelos serviços da Linguateca, em particular o AC/DC. No entanto, a despeito da influência, são notórias as diferenças com relação ao enfoque, o que resulta na diferença de qualidade das ferramentas para as funções a que se propõem.

[14]Disponível para instalação no endereço: <https://github.com/alvelvis/Interrogat-rio>. Acesso em 3 de mar. 2023.

[15]A expressão de busca é *@token.deprel = "obj" and token.head_token.upos = "VERB"*, onde o foco, representado pelo arroba, está no token cuja anotação de deprel (relação de dependência) é "obj" e a anotação de upos (part-of-speech) do token que está acima dele na árvore de dependências (ou seja, o seu *head_token*) é de "VERB". Ao longo da dissertação, todas as expressões de busca utilizadas para ilustrar alguma busca estarão nesse formato.

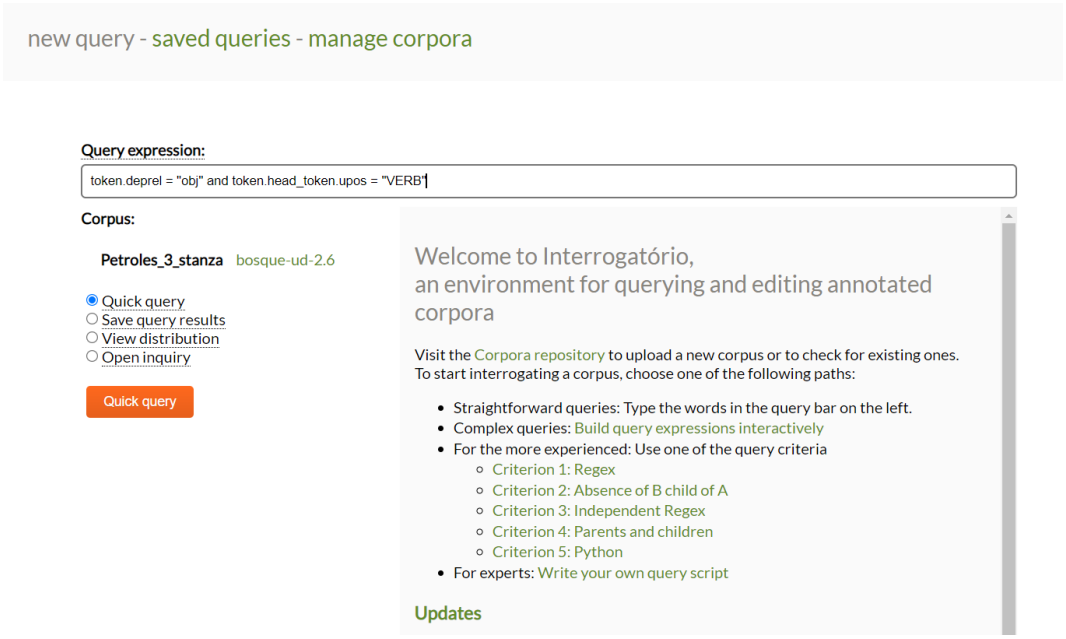


Figura 4.8: Página inicial do Interrogatório

Figura 4.9: Página com resultados de uma busca no Interrogatório

lemma	frequency	in files
valor	119	16
quantidade	83	17
processo	80	18
concentração	76	12
área	71	12
%	67	17
comportamento	63	14
região	61	11
resultado	61	18
estrutura	59	14
óleo	59	11
anomalia	58	3
característica	57	16
aumento	55	18
efeito	54	14
formação	53	16
presença	53	8
fluido	52	8
dado	50	9
interpretação	49	7
parte	49	17
propriedade	49	12
tipo	48	16
tempo	47	11
viscosidade	47	12
forma	46	15
custo	45	11
água	45	14
superfície	44	18
condição	43	13

Figura 4.10: Distribuição de lemas para uma busca no Interrogatório

linguística delas para identificar a forma como os fenômenos estão atualmente anotados no corpus. Podemos ainda realizar refinamentos na busca (chamados de *filtros*), manualmente ou por meio de outras expressões de busca, para especificar ainda mais o fenômeno que estamos procurando (por exemplo, após a busca por objetos, podemos selecionar apenas aqueles objetos que são substantivos).

Após verificar os resultados da busca e dos seus refinamentos, podemos visualizar a distribuição das palavras em foco, como na figura 4.10, em que vemos uma lista dos objetos mais frequentes no corpus distribuídos pelo lema das palavras[16], ou podemos ainda modificar a anotação das frases, seja manualmente (na figura 4.11, o usuário pode modificar qualquer célula da tabela), seja via regras (na figura 4.12, o usuário escreveu uma regra para transformar todos os tokens anotados como *obj* de uma busca em *iobj*).

[16]O lema é a forma canônica, não flexionada da palavra. Distribuir os resultados de uma busca por lema tem como objetivo agrupar as diferentes flexões de uma palavra nos resultados.

Edit any cells:

text = Este histórico é determinante para classificar o campo de petróleo como um campo maduro, que com o passar dos anos em exploração e produção de petróleo, vão se tornando maduros..

sent_id = 398-20160721-MONOGRAFIA_0-49

1	Este	este	DET	_	Gender=Masc|Number=Sing|PronType=Dem	2	det	O	_
2	histórico	histórico	NOUN	_	Gender=Masc|Number=Sing	4	nsubj	O	_
3	é	ser	AUX	_	Mood=Ind|Number=Sing|Person=3|Tense=Pres|VerbForm=Fin	4	cop	O	_
4	determinante	determinante	ADJ	_	Gender=Masc|Number=Sing	0	root	O	_
5	para	para	SCONJ	_	_	6	mark	O	_
6	classificar	classificar	VERB	_	VerbForm=Inf	4	advcl	O	_
7	o	o	DET	_	Definite=Def|Gender=Masc|Number=Sing|PronType=Art	8	det	O	_
8	campo	campo	NOUN	_	Gender=Masc|Number=Sing	6	obj	O	_
9	de	de	ADP	_	_	10	case	O	_
10	petróleo	petróleo	NOUN	_	Gender=Masc|Number=Sing	8	nmod	O	_
11	como	como	ADP	_	_	13	case	O	_
12	um	um	DET	_	Definite=Ind|Gender=Masc|Number=Sing|PronType=Art	13	det	O	_
13	campo	campo	NOUN	_	Gender=Masc|Number=Sing	8	nmod:appos	O	_
14	maduro	maduro	ADJ	_	Gender=Masc|Number=Sing	13	amod	O	_
15	,	,	PUNCT	_	_	6	punct	O	_
16	que	que	PRON	_	Gender=Masc|Number=Sing|PronType=Rel	32	nsubj	O	_

Figura 4.11: Página para edição de uma frase no Interrogatório

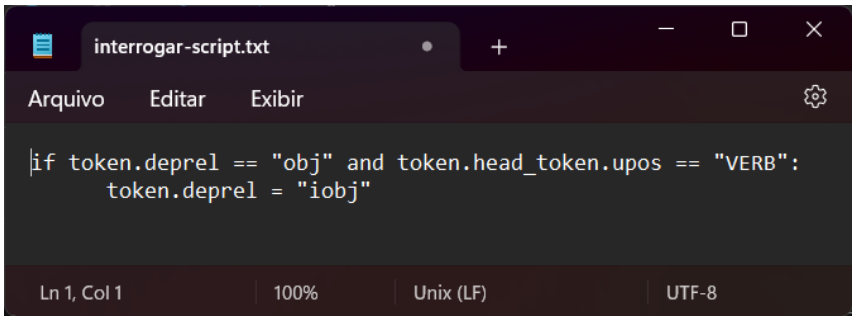


Figura 4.12: Exemplo de regra para edição de frases no Interrogatório

Tanto a edição manual quanto a edição via regras contam com mecanismos de validação para prevenir a introdução de erros pelos anotadores. Após realizar modificações manuais na anotação de uma frase, todas as regras de detecção de erros desenvolvidas para o método de revisão discutido na seção 4.2 são aplicadas somente à frase sendo editada, de maneira que erros introduzidos pelo usuário são apresentados assim que ele realiza as modificações. Já na edição via regras, a tela de aplicação é precedida por uma simulação das edições, de maneira que o usuário pode verificar se as correções estão corretas antes de aplicá-las.

Além da consulta ao corpus para a tomada de decisões linguísticas e da possibilidade de edição da anotação manualmente ou via regras, a ET dispõe ainda do Julgamento[17], o ambiente que reúne alguns dos métodos específicos para revisão de treebanks que discutimos na seção 4.2. Nele, os possíveis erros detectados pelos métodos são apontados para o usuário e também é possível

[17]Disponível para instalação no endereço: <https://github.com/alvelvis/Julgamento>. Acesso em 3 de mar. 2023.

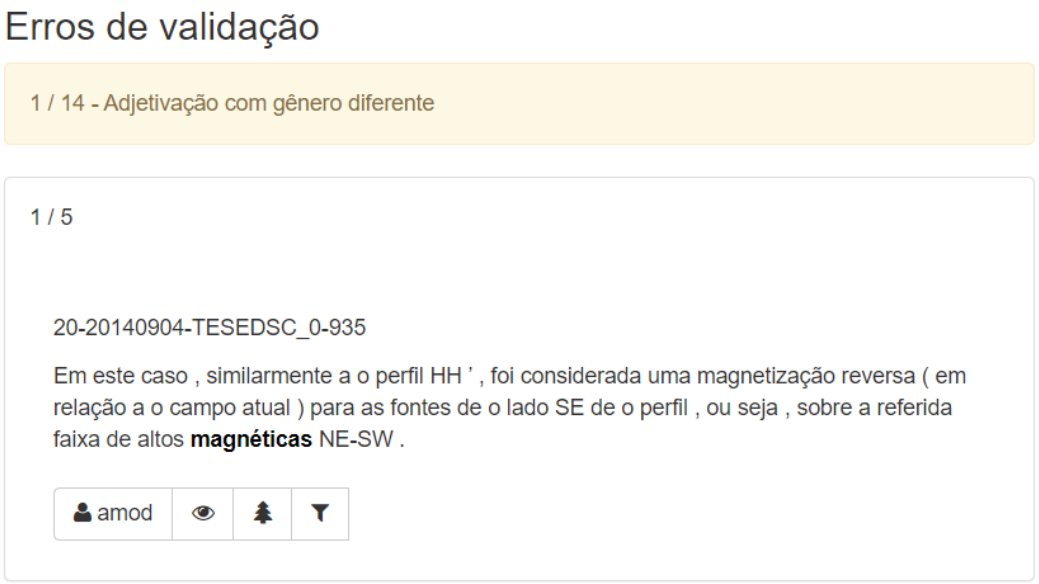


Figura 4.13: Página com os erros identificados pelo método das regras linguísticas no Julgamento

corrigi-los manualmente pela interface do programa. Como os ambientes da ET são integrados, o corpus utilizado em qualquer uma das ferramentas é compartilhado entre as demais, de maneira que o usuário conta sempre com a versão mais atualizada da anotação do recurso.

O primeiro método de revisão de treebanks disponível pelo Julgamento é o de detecção de erros via regras linguísticas. A interface do programa (figura 4.13) agrupa as frases por tipo de erro detectado (cada regra linguística designa um erro) e apresenta ao usuário a explicação do erro e o token onde o erro ocorre negritado, para facilitar a identificação e correção do problema na frase.

Na figura, vemos no quadro em amarelo que foram encontrados 14 tipos de erros, sendo que o primeiro erro tem a explicação "Adjetivação com gênero diferente"[18]. Esse erro está presente em 5 frases, sendo que apenas a primeira foi ilustrada na figura, com a palavra "magnéticas" em negrito. Ao clicar em "amod", o usuário verá a anotação da frase e poderá editá-la – neste caso, não haverá correção na anotação a ser feita, pois o erro de concordância foi cometido pelo autor do documento, e não pelo anotador da frase.

O segundo método de revisão disponível pelo Julgamento também foi discutido na seção 4.2. Ele se chama *N-grams* inconsistentes, e busca por pares de tokens relacionados sintaticamente que tenham anotação diferente nas frases em que aparecem. Na figura 4.14, o quadro em amarelo indica que há 31 pares

[18]Esse erro foi encontrado por meio da regra *deprel = "amod" and head_token.gender != gender*, que busca por tokens cuja relação de dependência seja "amod" (adjunto adnominal do tipo adjetivo) cujo governante da relação tenha gênero diferente do gênero do adjetivo, o que é um erro, pois adjetivos devem concordar em gênero com o token que modificam.

Figura 4.14: Página com os erros identificados pelo método de busca por n-grams inconsistentes no Julgamento

de palavras com anotação inconsistente, e o 9º par é o das palavras “estufa” e “efeito”, sendo “efeito” o governante da relação.

Esse par de palavras aparece em três frases, das quais em duas a anotação de “estufa” é *nmod* (adjunto adnominal), e em uma a anotação é de “compound” (termo composto), motivo da inconsistência encontrada. Conforme veremos adiante, a classe dos *compound* foi eliminada do PetroGold, de maneira que a anotação correta é *nmod*. Nesse momento, o usuário pode abrir a anotação da primeira frase e corrigi-la.

O terceiro método de revisão disponível pelo Julgamento é o IAD (Inter-Annotator Disagreement), baseado no contraste entre duas análises diferentes para as mesmas frases. O método é acessível por meio de matrizes de confusão de POS e de DEPREL na interface do programa, e para acessá-lo é necessário, além de enviar o corpus que se quer analisar para o sistema,

UPOS confusion matrix

sistema	ADJ	ADP	ADV	AUX	CCONJ	DET	NOUN	NUM	PRON	PROPN	PUNCT	SCONJ	SYM	VERB	X	All
golden																
ADJ	762	0	0	0	1	0	12	0	0	2	0	0	0	12	0	789
ADP	0	2085	0	0	0	1	0	0	0	0	1	1	0	0	0	2088
ADV	1	5	336	0	0	0	1	0	2	0	0	3	0	0	0	348
AUX	0	0	0	315	0	0	0	0	0	1	0	0	0	1	0	317
CCONJ	0	1	1	0	317	0	0	0	0	0	0	1	0	0	0	320
DET	0	1	1	1	0	1794	0	0	1	0	0	0	0	0	0	1798
NOUN	13	0	0	0	0	0	2819	3	2	16	1	0	4	2	0	2860
NUM	1	0	0	0	0	0	2	349	0	1	0	0	0	0	0	353
PRON	1	0	4	0	0	2	1	0	245	0	0	5	0	0	0	258
PROPN	1	2	0	0	0	0	12	4	0	500	3	0	0	0	0	522
PUNCT	0	0	0	0	0	0	0	0	0	0	1421	0	0	0	0	1421
SCONJ	0	0	1	0	0	0	0	0	5	0	0	71	0	1	0	78
SYM	0	0	0	0	0	0	0	0	0	0	12	0	24	0	0	36
VERB	8	0	0	1	0	2	7	0	0	0	0	0	0	963	0	981
X	0	0	0	0	0	0	1	0	0	0	0	0	0	0	6	7
All	787	2094	343	317	318	1799	2855	356	255	520	1438	81	28	979	6	12176

Figura 4.15: Página com matriz de confusão que dá acesso ao método IAD no Julgamento

enviar uma outra versão do corpus com análises diferentes para as mesmas frases. Os dois arquivos podem ter sido anotados por pessoas diferentes, por analisadores automáticos diferentes, ou podem ser uma anotação padrão ouro e a sua contraparte analisada automaticamente, como na figura 4.15, onde contrastamos a partição de teste do padrão ouro do PetroGold e a sua contraparte obtida durante a avaliação intrínseca de um modelo treinado utilizando o corpus. Na figura, pedimos a matriz de confusão de classes gramaticais (UPOS).

Pela matriz de confusão o usuário pode observar os núcleos de divergência, isto é, as classes com maior discordância entre as duas análises. Já a diagonal indica a concordância, quando as análises são iguais. Por exemplo, há 762 tokens no corpus que foram anotados como adjetivo tanto no padrão ouro quanto na análise automática, e há 13 casos de divergência nos quais a análise do padrão ouro é a de substantivo, e na análise automática a etiqueta é a de adjetivo. Clicando no número "13", o usuário visualiza as 13 ocorrências de divergência entre NOUN e ADJ e pode corrigir o padrão ouro quando julgar necessário.

Por fim, o Julgamento auxilia na avaliação de corpora de mais duas formas. Quando o usuário envia para o programa uma partição de teste de um corpus com anotação padrão ouro e a sua contraparte analisada por um

Metrics from conll18_ud_eval.py

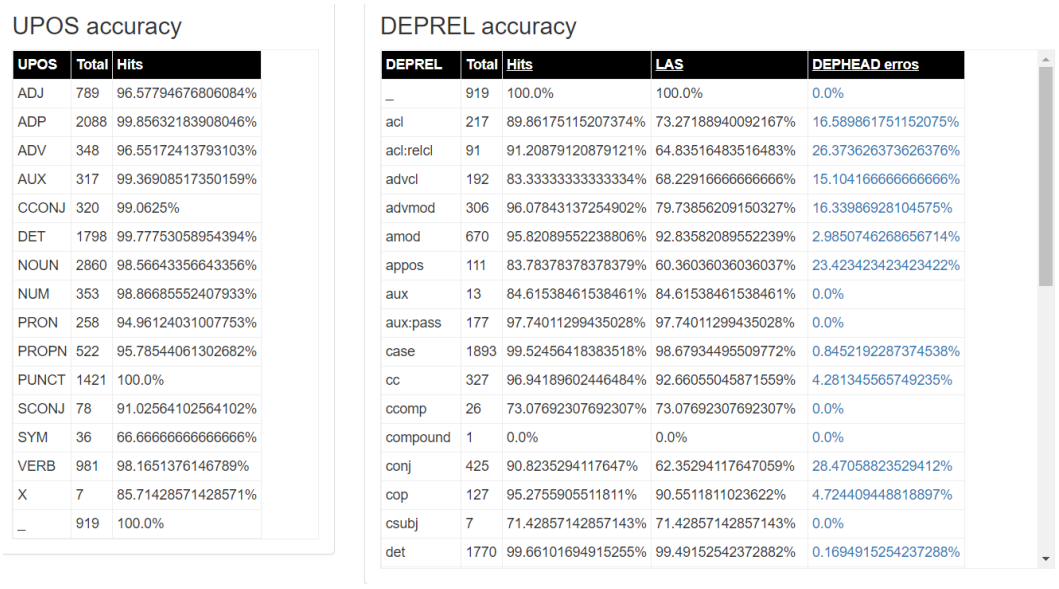
Metric	Precision	Recall	F1 Score	AligndAcc
Tokens	100.00	100.00	100.00	
Sentences	100.00	100.00	100.00	
Words	100.00	100.00	100.00	
UPOS	98.61	98.61	98.61	98.61
XPOS	100.00	100.00	100.00	100.00
UFeats	98.26	98.26	98.26	98.26
AllTags	97.66	97.66	97.66	97.66
Lemmas	98.56	98.56	98.56	98.56
UAS	91.15	91.15	91.15	91.15
LAS	89.42	89.42	89.42	89.42
CLAS	84.56	84.13	84.34	84.13
MLAS	81.87	81.46	81.67	81.46
BLEX	82.82	82.40	82.61	82.40

Figura 4.16: Página do Julgamento com resultados da avaliação intrínseca

anotador automático, a ferramenta disponibiliza a consulta às ferramentas de avaliação intrínseca discutidas na seção 4.3, como na figura 4.16. Por meio dessa função é possível ver a avaliação da anotação do modelo utilizando métricas globais, como a medida F1 de UAS (91,15%), LAS (89,42%) etc.

Além disso, e mais interessante do ponto de vista linguístico, é observar o índice de acertos para cada categoria morfossintática. A figura 4.17 mostra o panorama de acertos de um modelo para as classes de POS e de DEPREL. A coluna "UPOS" na primeira tabela e "DEPREL" na segunda mostra qual a etiqueta sendo avaliada; a coluna "Total" diz o número de tokens com essa etiqueta no corpus (ou na partição de teste, caso contrário não seria possível realizar a avaliação intrínseca); a coluna "Hits", na primeira tabela, indica o número de acertos da atribuição de POS, e a coluna "LAS", na segunda tabela, o acerto de relação e encaixe de dependência; e a coluna "DEPHEAD errors" indica a quantidade de erros de encaixe de dependência, sendo que o usuário pode clicar em cada um dos números para verificar quais são as frases com erro no encaixe e corrigir o padrão ouro quando necessário[19].

[19]Para visualizar não o erro de encaixe, mas de relação, o usuário pode usar a matriz de confusão de DEPREL apresentada anteriormente.



UPOS accuracy

UPOS	Total	Hits
ADJ	789	96.57794676806084%
ADP	2088	99.85632183908046%
ADV	348	96.55172413793103%
AUX	317	99.36908517350159%
CCONJ	320	99.0625%
DET	1798	99.77753058954394%
NOUN	2860	98.56643356643356%
NUM	353	98.86685552407933%
PRON	258	94.96124031007753%
PROPN	522	95.78544061302682%
PUNCT	1421	100.0%
SCONJ	78	91.02564102564102%
SYM	36	66.66666666666666%
VERB	981	98.1651376146789%
X	7	85.71428571428571%
_	919	100.0%

DEPREL accuracy

DEPREL	Total	Hits	LAS	DEPHEAD erros
_	919	100.0%	100.0%	0.0%
acl	217	89.86175115207374%	73.27188940092167%	16.589861751152075%
acl:relcl	91	91.20879120879121%	64.83516483516483%	26.373626373626376%
advcl	192	83.33333333333334%	68.22916666666666%	15.104166666666666%
advmod	306	96.07843137254902%	79.73856209150327%	16.33986928104575%
amod	670	95.82089552238806%	92.83582089552239%	2.9850746268656714%
appos	111	83.78378378378379%	60.36036036036037%	23.423423423423422%
aux	13	84.61538461538461%	84.61538461538461%	0.0%
aux:pass	177	97.74011299435028%	97.74011299435028%	0.0%
case	1893	99.52456418383518%	98.67934495509772%	0.8452192287374538%
cc	327	96.94189602446484%	92.66055045871559%	4.281345565749235%
ccomp	26	73.07692307692307%	73.07692307692307%	0.0%
compound	1	0.0%	0.0%	0.0%
conj	425	90.82352941176647%	62.35294117647059%	28.47058823529412%
cop	127	95.27559055118111%	90.5511811023622%	4.724409448818897%
csubj	7	71.42857142857143%	71.42857142857143%	0.0%
det	1770	99.66101694915255%	99.49152542372882%	0.1694915254237288%

Figura 4.17: Página do Julgamento com avaliação da anotação automática de cada categoria morfossintática

5
Construção de um *treebank* padrão ouro

5.1
Sobre o PetroGold

PetroGold é um treebank padrão ouro composto por teses e dissertações – textos acadêmicos – do domínio do petróleo em português brasileiro processados integralmente: apenas elementos como sumário, resumo, apêndices e seções de referências bibliográficas foram descartadas, assim como figuras, gráficos, fórmulas e tabelas. O corpus já conta com três versões, cada uma representando um avanço importante na qualidade do material e utilizando metodologias distintas para revisão. A terceira versão do PetroGold coincide com o lançamento do corpus no projeto Universal Dependencies, uma iniciativa para uniformização da anotação de treebanks em diferentes línguas, e é também um produto das modificações descritas nesta dissertação, contando com 19 documentos, 8.946 frases e 250.605 tokens.

A anotação morfossintática segue as diretrizes do projeto UD, que prevê uma gramática própria no formato de dependências, e teve origem automática mas foi manualmente revista por quatro anotadores. Os 19 documentos – teses e dissertações do domínio do petróleo – que compõem o PetroGold foram selecionados aleatoriamente do Petrolês (CORDEIRO, 2020), um corpus com 4.302 documentos de interesse para a área do petróleo que incluem, além dos textos acadêmicos, notas, boletins e outros textos técnicos.

A revisão do corpus foi realizada utilizando alguns métodos para revisão de erros em treebanks, conforme já explicado em 4.2. Além disso, algumas questões linguísticas específicas precisaram ser revistas tendo em vista as especificidades dos textos acadêmicos e tendo como objetivo a facilitação do aprendizado automático para a futura tarefa de reconhecimento de entidades mencionadas, momento em que precisamos realizar estudos linguísticos com base no corpus, o que veremos na seção 5.2.

5.1.1
A primeira e a segunda versão

A primeira versão do PetroGold foi publicada em 2021 (DE SOUZA et al., 2021b). Nessa versão, o corpus já havia passado pela etapa do pré-processamento, o que envolveu tanto a conversão dos arquivos, originalmente no formato PDF, para TXT (SILVEIRA et al., 2019), quanto a segmentação

	v2	**v1**
Tokens	250.595	253.640
Correções	8.802	12.832
Palavras	221.208	223.707
Frases	8.949	9.127
Documentos	19	19

Tabela 5.1: Características do PetroGold v2 em comparação com a v1

dos textos em frases e tokens (CAVALCANTI et al., 2021). Posteriormente, revisamos o corpus para incorporar uma série de padronizações linguísticas relativas ao gênero acadêmico e ao domínio do petróleo, e aplicamos um dos métodos de revisão que foram apresentados na seção 4.2, o IAD (Inter-Annotator Disagreement). Já a segunda versão do treebank, publicada em 2022 (DE SOUZA; FREITAS, 2022a), reuniu uma série de melhorias na anotação de alguns fenômenos linguísticos e expandiu a lista de métodos de revisão utilizados para torná-lo padrão ouro. Desse modo, a segunda versão do corpus serviu também para avaliar o impacto da revisão da anotação linguística de um corpus que já é padrão ouro nos números de avaliação intrínseca.

A tabela 5.1 ilustra as diferenças de características entre as duas versões do PetroGold[1]. O decréscimo de 178 frases da primeira para a segunda versão, resultando em decréscimo também do número de tokens, ocorreu por conta da eliminação de frases que posteriormente julgamos serem incompatíveis com o corpus que queríamos desenvolver, seja porque continham erros de pré-processamento, quando o texto foi convertido do texto original para o formato de texto sem formatação, seja porque eram textos em outra língua ou inteiramente fórmulas matemáticas.

O número de revisões feitas no corpus a partir da anotação automática, resultando na primeira versão do PetroGold, foi de 12.832 tokens que tiveram alguma das suas informações linguísticas revistas, o que respondia por 5% dos tokens do corpus. Da primeira para a segunda versão do corpus houve um acréscimo de 8.802 erros corrigidos, chegando a um total de 21.634 tokens (8,63%) do corpus tendo passado por alguma correção na segunda versão do PetroGold.

Neste trabalho, estamos apresentando uma terceira versão do PetroGold,

[1]Token é a unidade mínima para o processamento computacional, sendo ele que recebe, tradicionalmente, a anotação linguística em um corpus anotado. Já palavra é um conceito de difícil definição (CAVALCANTI et al., 2021), podendo ser delimitada do ponto de vista lexical, sintático, fonético, semântico etc. Na tabela, a diferença entre token e palavra se dá apenas pela não inclusão dos tokens da classe de pontuação (PUNCT) na contagem das palavras, sendo que em ambas as contagens, as contrações (como da preposição com o artigo em "da" – *de+a*), foram contadas como palavras/tokens individuais.

em que se incorporaram revisões relativas a questões linguísticas que serão discutidas na seção 5.2, além de todos os métodos de revisão apresentados na seção 4.2.

5.2 Questões linguísticas

A seguir, veremos no detalhe algumas das questões linguísticas mais relevantes na anotação do PetroGold. Neste capítulo, veremos as discussões e resultados linguísticos relativos à anotação dos fenômenos em questão. Já no capítulo 6, avaliaremos do ponto de vista computacional as modificações realizadas no corpus. Exceto pela anotação dos argumentos verbais introduzidos por preposição (5.2.1), que já havia sido introduzida na versão 2 do corpus (a qual será avaliada na seção 6.1), as demais questões foram introduzidas apenas na versão 3, que será avaliada na seção 6.2.

5.2.1 Argumento verbal introduzido por preposição

Na tradição gramatical, de modo geral, a oração se divide em termos essenciais, integrantes e acessórios. Os termos essenciais são o sujeito e o predicado; os integrantes, os complementos nominais e verbais (objetos direto e indireto), e os termos acessórios, por fim, são os adjuntos adnominal, adverbial e o aposto. De um ponto de vista categórico, absolutamente todas as palavras, sintagmas ou termos de uma frase devem ser classificadas em uma (e apenas uma) dessas categorias. Entretanto, em muitos momentos a distinção entre uma classe e outra não é trivial.[2]

Uma das questões gramaticais que abordaremos é a distinção entre objeto indireto (um dos argumentos verbais) e adjunto adverbial. A dificuldade na distinção entre um objeto indireto e um adjunto adverbial decorre do fato de que ambos os sintagmas são preposicionados (no caso do adjunto adverbial, a dificuldade só ocorre quando é preposicionado) e ambos são dependentes do núcleo do predicado. A tendência das gramáticas é simplificar o assunto e apresentar apenas frases prototípicas em que a distinção seja mais facilmente realizada. Nossa preocupação, porém, é com frases reais de corpus, como as frases 13-16, encontradas no Bosque (FREITAS; ROCHA; BICK, 2008), cuja distinção entre complemento verbal e adjunto adverbial não é tão simples de ser feita para os termos em negrito.

[2]O conteúdo desta seção é uma ampliação da discussão realizada em de Souza e Freitas (2022b).

13. Os jogadores se dividem **pelos dez quartos** do alojamento, equipados com frigobar, ar condicionado, televisão e telefone.

14. Outros profissionais brasileiros, que atuam **nos EUA**, também participam.

15. Papa indica mulher **para secretaria**

16. O PDT pretende reduzir os impostos federais **a quatro**.

Para distinguir um objeto indireto de um adjunto adverbial precisamos ter clara a diferença entre um argumento e um adjunto. Bechara (2012) distingue o termo argumental do não argumental (também o chamaremos de adjunto) pelo significado lexical referido pelo verbo. Se o termo aparece "solicitado" ou "regido" por ele, trata-se de um argumento, caso contrário, um adjunto. Ser "solicitado" pelo verbo, nas palavras do autor, significa dizer que "está mais estreitamente ligado ao conteúdo do pensamento designado pelo verbo" (p. 437), como ocorre com o termo "experiências amargas", na frase 17, mas não com o termo "durante sua vida", na mesma frase, segundo sua argumentação. É uma definição semântica, mas o autor fala também em dependência sintática, já que, para ele, a frase 18 parece razoável, enquanto 19, sem o termo que considera argumento, tem sua "estruturação sintático-semântica" prejudicada. Haveria, portanto, uma subordinação entre "experiências amargas" e o verbo "conheceu".

17. Graciliano conheceu experiências amargas durante sua vida.

18. Graciliano conheceu experiências amargas.

19. * Graciliano conheceu durante sua vida.

Vilela e Koch (2001) reconhecem que a distinção argumento/adjunto "tem merecido algumas reflexões e ainda não se chegou a uma conclusão definitiva" (p. 347). Essa conclusão definitiva, de um ponto de vista categórico, deveria ser alcançada através do refinamento dos critérios qualitativos de identificação do argumento e do adjunto, entendendo-se que a distinção é adequada, os critérios de distinção é que falham.

Os autores indicam alguns critérios na distinção argumento/adjunto. Para se distinguir quando o termo não é preposicionado, tenta-se transformar a frase em voz passiva, procedimento que só será possível caso o termo seja objeto direto. Assim, a frase 20 não pode se tornar 21, sendo o termo um adjunto adverbial, e não objeto.

20. A aula durou três horas

21. * Três horas foram duradas pela aula

Outro critério é a conversão do complemento em oração adverbial ou coordenada, por meio dos denominados pró-verbos "fazer", para os verbos de ação/atividade, e "acontecer/passar-se com", para os verbos de processo. Se a conversão for bem sucedida, como ocorre da frase 22 para as 23 e 24, o sintagma é um adjunto adverbial; caso contrário, é um argumento do predicado, como ocorre na conversão de 25 para 26.

22. Ele estudou três horas para preparar a aula

23. Ele estudou para preparar as aulas e fê-lo durante três horas

24. Trabalhando arduamente durante três horas ele estudou para preparar a aula.

25. A aula durou três horas

26. * A aula durou e isto aconteceu durante três horas.

Veja que a mesma frase, "A aula durou três horas", ora teve o termo "três horas" classificado como adjunto, ora como argumento. Os autores não resolvem a confusão, se restringindo a indicar que esses critérios não são os mais adequados para a frase. Para eles, o critério essencial é a regra de que, assim como postulou Bechara, ao se suprimir o termo, a frase se torna incompleta (frase 27), indicando ser o termo um argumento.

27. * A aula durou

Outro elemento essencial para os autores na distinção argumento/adjunto é a interrogação ao verbo para identificar aqueles termos que "estão instalados no próprio significado do predicado" (p. 347). Caso o termo responda às perguntas "quem, qual, que, onde, quanto, como" feitas ao verbo, trata-se de um argumento; se, do contrário, o sintagma responder às perguntas "onde, por que, como, quando", trata-se de um adjunto adverbial. Vemos, porém, que há interrogações que se repetem nas duas classificações (onde, como e quando), sendo pouco úteis para a distinção das classes. Como nas frases abaixo, onde temos que, tanto em 28 como em 29, O Francisco responde à pergunta "quem colocou/descobriu", sendo classificado, pois, como argumento (do tipo sujeito), mas há dificuldade na classificação do sintagma "na prateleira", pois ambos respondem às perguntas onde colocou/descobriu, forma que cabe tanto aos argumentos quanto aos adjuntos, segundo seu critério.

28. O Francisco colocou a enciclopédia na prateleira.

29. O Francisco descobriu a enciclopédia na prateleira.

Nesse caso, as intuições dos autores lhes diriam que, para o verbo "colocar", "na prateleira" é argumento e, para "descobrir", adjunto. A palavra intuição é mesmo a utilizada pelos autores, indicando acreditarem na existência de um conhecimento linguístico anterior que deve ser respeitado:

> "Ora, a nossa intuição diz-nos que os 'locativos' não têm, relativamente aos predicados (colocar e descobrir), mesmo estatuto: em relação a 'colocar' constitui um complemento nuclear ou actancial, em relação a 'descobrir', é um circunstante." (VILELA; KOCH, 2001, p. 348)

O sentido de intuição entendido pelos autores da gramática é parecido com o analisado por Neto (2011) na sua resenha sobre uma outra gramática, momento em que o autor provoca: "Talvez analfabetos possam ter 'intuições' sobre a língua, linguistas relembram análises com que tiveram contato" (p. 69). O autor sugere que se trata apenas de um processo de reafirmação das mesmas categorias por repetição de análises já realizadas anteriormente por algumas pessoas e que se travestem de "intuição", algo supostamente natural.

Vilela e Koch (2001) buscam então "critérios suplementares" para justificar sua intuição. Voltando ao critério de supressão do termo, a ideia é a de que ao se suprimir um adjunto, a frase continuaria completa – segundo eles, podemos dizer "O Francisco descobriu a enciclopédia" e a frase continua completa, mas não seria aceitável terminar a outra frase em "O Francisco colocou a enciclopédia" sem o complemento de lugar.

Realizamos uma breve exploração para verificar a afirmação de que o verbo colocar exige um complemento de lugar. Uma busca no corpus "todos juntos", no serviço AC/DC da Linguateca, retorna 313.047 ocorrências do verbo "colocar"[3]. No início da lista, encontramos pelo menos quatro frases com o verbo "colocar" sem o complemento de lugar prototípico (frases 30-33).

30. Eles têm o monopólio do mercado e **colocam** o preço que quiserem.

31. Para aproveitar o contra-ataque, Ramirez vai **colocar** os volantes Ney e Cristóvão exercendo uma forte marcação no meio-campo.

[3]O corpus "Todos juntos", conforme a página do AC/DC, "foi criado pela Linguateca de forma a permitir que todos os corpos da Linguateca pudessem ser interrogados duma só vez, evitando ao mesmo tempo que sobreposições de material produzissem repetição escusada e enganadora de concordâncias ou frequências" (Disponível em: <https://linguateca.pt/acesso/corpus.php?corpus=TODOS>. Acesso em 5 de mar. 2023). A busca foi realizada utilizando a expressão *[lema="colocar"]*.

32. A situação da democracia no Peru, após uma eleição cercada por suspeitas de fraude, pode **colocar** Fujimori no alvo de reprimendas.

33. Para situar nosso questionamento no modelo lógico da Política Nacional de Monitoramento e Avaliação da Atenção Básica8, é necessário **colocar** a aquisição de novos conhecimentos e a melhoria do desempenho do Sistema Único de Saúde (SUS) como suas principais finalidades.

Para além do confronto com a intuição, a busca em corpus já sinaliza a necessidade de um primeiro esclarecimento: a que verbo "colocar" os autores se referem? A polissemia verbal é um fato linguístico conhecido, e a cada sentido podem corresponder diferentes formas de complementação verbal. Assim, quando nos referimos ao verbo "colocar" e à suposta exigência de complemento de lugar, estamos tratando de uma definição específica do verbo:

> "1. Pôr (algo ou alguém, inclusive si mesmo) (num lugar, em certa posição etc.), botar, dispor [tda. : Coloque os documentos aqui, o processo naquela prateleira.] [td. : Antes de subir na escada, colocou um reforço: Encontrou-a na fila e colocou -se a seu lado.: Coloque essas peças em ordem.]"[4]

A definição não se aplicaria, em sentido estrito, a nenhuma das frases que encontramos, justificando a ausência de um complemento de lugar. A tarefa passaria, portanto, a ser a desambiguação dos sentidos do verbo, o que já é sabidamente difícil de ser feito, e não mais a discriminação entre argumentos e adjuntos verbais.

A "intuição" dos gramáticos de que uma frase com o verbo "colocar" pede um complemento de lugar deve se sustentar na maioria das ocorrências, mas na breve exploração encontramos ocorrências que confrontam a hipótese. Em face dos dados encontrados, quando trabalhamos com corpus, a hipótese deveria ser reformulada (como discutido em Sampson (2002)).

Vilela e Koch (2001) parecem ter ciência também de que por vezes o argumento não aparece na frase por variados motivos, a despeito de serem "exigidos" pelo verbo. Por isso, escrevem sobre argumentos facultativos e argumentos obrigatórios. Para os gramáticos, um mesmo verbo, mesmo que tenha um complemento inscrito no seu significado, pode abrir mão dele sem que isso provoque inaceitabilidade da frase. É o caso da frase 34, com todos os espaços do verbo "escrever" preenchidos, embora seja possível abandoná-los, como em 35, 36 e 37.

[4]Definição do dicionário Aulete Digital. Disponível em: <https://aulete.com.br/colocar>. Acesso em 10 de jan. 2022.

34. A Joana escreveu uma carta ao Antão

35. A Joana escreve (muitas vezes)

36. A Joana escreveu uma carta

37. A Joana escreve (frequentemente) ao Antão

Acrescentamos algumas frases mais complexas: que função seria atribuída aos sintagmas "em inglês" e "em Python" (frases 38 e 39)?

38. A Joana escreveu uma carta em inglês ao Antão.

39. A Joana escreve seus códigos em Python.

Diferentemente do verbo "escrever", o verbo "oferecer" não permitiria esse tipo de elisão enquanto "propriedade sistemática" do verbo – isto é, embora as frases 40 e 41 sejam aceitáveis, 42 não o é, pois o verbo permite omissão de objeto indireto, mas não de objeto direto.

40. A Joana ofereceu um livro ao Antão

41. A Joana ofereceu um livro

42. * A Joana ofereceu ao Antão

E além da elisão que é propriedade de cada verbo, há ainda a elisão contextual – quando o argumento não está na frase mas é recuperável pelo contexto, portanto independendo das propriedades do verbo para se tornar aceitável.

Importante notar, contudo, que a postulação de um argumento facultativo (ou, nos termos dos autores, "actante facultativo") parece alargar o limite do razoável. Se, na definição, o argumento é obrigatório, e um dos principais testes para identificá-lo é tentar suprimi-lo, a existência de um argumento que não é obrigatório descaracteriza qualquer esforço de definir o fenômeno. Estamos diante de uma tentativa de manter a tradição a qualquer custo, a despeito do fato de que os dados nem sempre confirmam as classificações tradicionais e as definições das categorias já não fazem mais sentido. Para Neto (2011), esse tipo de manipulação conceitual para atingir a tradição impede o avanço da teoria: "como se as teorias modernas não pudessem se desvencilhar da carga da tradição – e esses laços tolhem significativamente suas possibilidades de desenvolvimento" (p. 55). Os critérios de distinção entre classes podem se tornar

consensuais pela força da tradição e da academia, mas, no limite, os critérios suplementares tendem a ser infinitos e, para cada nova frase que não se enquadre aos critérios estabelecidos, um novo critério poderá ser criado.

Sobre os argumentos introduzidos por preposição que se confundem com adjuntos adverbiais, Bechara (2012) tem o cuidado de não chamá-los nem objeto preposicionado (caso de Amar a Deus sobre todas as coisas), tampouco objetos indiretos (O diretor escreveu cartas aos pais). Nas frases 43-46, o autor chama o termo em negrito de "complemento relativo", sendo semelhante ao objeto direto em termos semântico-sintáticos, exceto pela presença de preposição.

43. Todos nós gostamos **de cinema**.

44. O marido não concordou **com a mulher**.

45. Poucos assistiram **ao concerto**.

46. O comerciante não confiou **no empregado**.

O gramático indica que cada verbo vem acompanhando de sua própria preposição pelo que chama de "servidão gramatical". Assim, "depender de", "concorrer com" e "agregar a" são previsíveis, embora haja exceções: primeiro, o caso em que a norma permite o emprego de mais de uma preposição (Ela se parece ao/com o pai), e segundo, o caso de variação linguística (diatópica, diastrática e diafásica[5]), como com os verbos "socorrer", "contentar" e outros, que em diferentes variedades do português podem ser usados com ou sem preposição. Essa posição é atualizada por Bagno (2012), que apresenta exemplos de mudança histórica, e não apenas de variação do português brasileiro, como nos casos desagradar (a) alguém, desobedecer (a) algo, aspirar (a) algo etc.

Bagno (2012) igualmente nota que, na gramática tradicional, os objetos indiretos seriam todos os complementos introduzidos por preposição, mas modernamente já se consideram objetos indiretos apenas aqueles com o traço semântico [beneficiário], isto é, aquele que se beneficia com a ação indicada pelo verbo, como nas frases 47 e 48. Assim, cita o complemento relativo de Bechara e o complemento oblíquo de Castilho (2010) como exemplos dos outros complementos preposicionados.

47. Comprei um perfume delicioso **para você.**

48. Esses computadores pertencem **à escola**.

[5]Respectivamente, variação de lugar, de grupo social e de contexto de comunicação.

Por fim, Bechara insere uma importante observação sobre o fenômeno que denominou complemento relativo:

> "Não há unanimidade entre os estudiosos em considerar tais argumentos do predicado complexo como complementos relativos. Levando em conta exclusivamente o aspecto semântico, muitos preferem considerar tais termos como adjuntos circunstanciais ou adverbiais (...)" (BECHARA, 2012, p. 446)

No paradigma empírico não-categórico, Manning (2003) entende que existem termos que são claramente argumentos (sujeitos e objetos diretos) e termos que são claramente adjuntos (de tempo ou de local externo), mas há igualmente uma gama de outros sintagmas no meio do caminho entre as duas classificações. Assim, não seria possível afirmar categoricamente que um verbo exige ou não complemento. O que podemos fazer é, cientes de que a classificação é uma interpretação, verificar quais foram as análises codificadas em um corpus, segundo critérios específicos, e modelar estatisticamente a distribuição de argumentos para um determinado verbo nos seus diferentes contextos.

McEnery e Hardie (2011) argumentam que adicionar anotação linguística a um corpus, como a de argumentos e adjuntos, é tornar explícita uma análise que estava implícita nos dados (p. 31). Podemos discutir se, de fato, estamos adicionando informação nova ou se ela já estava presente nos dados – ao contrário dos autores, entendo que estamos adicionando informação nova ao corpus uma vez que toda classificação é fruto de interpretação humana e que aos mesmos dados poderiam ter sido atribuídas análises distintas –, mas fato é que, quando analisamos distribucionalmente anotações codificadas em corpus, estamos lidando com análises explícitas. Sendo explícitas, nossas hipóteses podem ser corroboradas ou contestadas por meio dos dados, que são reanalisáveis, procedimento que não poderá ser feito quando nosso objeto contém análises de base intuitiva, que não são explicitadas, e portanto não são reprodutíveis.

Para realizar estudos derivados da anotação de adjuntos e argumentos, precisaremos antes esclarecer os critérios que utilizamos na interpretação de ambas as classes. A fim de anotar o corpus PetroGold, tomamos uma posição que julgamos levar a uma maior concordância interanotadores: consideramos exclusivamente o aspecto semântico, como Bechara notou que alguns estudiosos fazem, e consideramos os termos como adjuntos adverbiais sempre que o sentido sendo analisado é um dos tradicionalmente adverbiais (quando o termo refere-se a tempo, lugar, modo, finalidade, etc.). Nossa prática de anotação

dessas categorias no PetroGold, assim como alguns experimentos estatísticos, são objeto das próximas seções.

5.2.1.1 Metodologia

Como identificar adjuntos e argumentos

Como já observamos, a anotação do corpus PetroGold segue a abordagem morfossintática de dependências do projeto multilíngue Universal Dependencies (MARNEFFE et al., 2021), um esforço coletivo de tornar comparáveis as análises de treebanks de diferentes línguas para a realização de tarefas de processamento de linguagem natural. Nas diretivas gramaticais do projeto, a questão argumento/adjunto segue o mesmo direcionamento desde a primeira versão do projeto, reforçado em Zeman (2017) e outros. As dificuldades relatadas nas gramáticas do português também se fazem presentes em grande parte das línguas que compõem o projeto, motivo pelo qual UD decidiu eliminar a distinção entre argumento e adjunto[6].

Przepiórkowski e Patejuk (2018) endossam e refinam a escolha por essa distinção na gramática UD, tendo já discutido em muitos de seus trabalhos anteriores as dificuldades de se distinguir adjunto de argumento, uma dicotomia que chamam "má-definida" (p. 3838). Os autores encontram solução, na gramática léxico-funcional, para substituir a dicotomia adjunto-argumento pela nuclear-oblíquo (na terminologia UD, *core-oblique*), sendo uma dicotomia de natureza distinta, motivo pelo qual defendem o mesmo procedimento para UD. Assim, desloca-se a tensão de um lugar para outro. Esse deslocamento, porém, veremos que é apenas parcial segundo nossos objetivos.

A ideia da distinção nuclear-oblíquo é assegurar comparabilidade entre línguas, tendo em vista que cada uma codifica seus dependentes de formas particulares. Como explicam Marneffe et al. (2021), "the core-oblique distinction has to do with the morphosyntactic encoding of dependents, not with their status as obligatory or selected by the predicate" (p. 268).

Partindo da ideia de que algumas relações de dependências são mais universais do que outras, os termos nucleares são aqueles que seriam menos

[6] "[...] the argument/adjunct distinction is subtle, unclear, and frequently argued over. For instance, syntacticians at certain times have argued for various obliques to be arguments, while at other times arguing that they are adjuncts, particularly for certain semantic roles such as oblique instruments or sources. We take the distinction to be sufficiently subtle (and its existence as a categorical distinction sufficiently questionable) that the best practical solution is to eliminate it." (Disponível em <https://universaldependencies.org/u/overview/syntax.html>. Acesso em 10 de dez. 2021.)

variáveis entre línguas e ocorrem da mesma forma na superfície, sendo eles o sujeito e o objeto, quando ocorrem de forma "não marcada"[7]. Os critérios de marcação ou não marcação das formas do sujeito e objeto, como notam Marneffe et al. (2021), são específicos para cada língua, contudo, alguns critérios são recorrentes, entre os quais destaco:

(i) Verbos geralmente concordam apenas com termos nucleares (o sujeito, em português)

(ii) Termos nucleares comumente aparecem sem preposição enquanto *t*ermos oblíquos são marcados por preposição e outros marcadores gramaticais

(iii) Operações como passivização comumente só permitem promover e demover termos nucleares

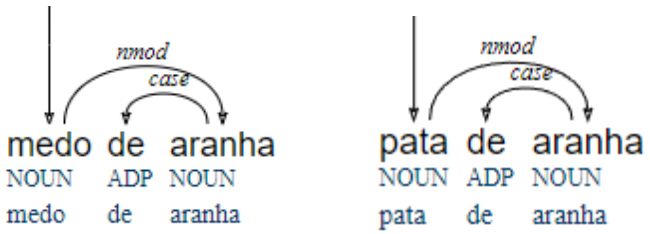
Considerando os critérios (ii) e (iii) em conjunto, concluímos que sintagmas precedidos por preposição não podem ser termos nucleares. Zeman (2017) nota que um critério simples de distinção entre termos nucleares e oblíquos no treebank em inglês é a presença ou não de preposição, postura que também pode ser adotada para a língua portuguesa.

Dessa forma, um argumento verbal é *obj* (objeto direto) quando não é precedido por preposição, é *iobj* (objeto indireto) apenas quando já há um objeto direto na frase e este objeto indireto deve necessariamente ser um pronome oblíquo, pois ocorre no caso dativo sem ser preposicionado, e *obl* para todos os demais casos, tanto de argumentos preposicionados quanto de adjuntos adverbiais.

Trata-se de uma solução simples mas pouco informativa do ponto de vista linguístico. Se quisermos encontrar os argumentos de um verbo no corpus, e sabendo que muitos deles preposicionados (objetos indiretos, complementos relativos ou objetos diretos preposicionados), torna-se impossível distinguí-los de adjuntos adverbiais, pois todos esses fenômenos estariam anotados da mesma forma. Ou ainda no caso de verbos que podem ter seus complementos precedidos ou não por preposição devido a mudança ou variação linguística, como no caso do verbo "assistir (a) [algo]", teríamos duas análises distintas para enunciados com o mesmo sentido.

A ausência de distinção entre argumento e adjunto quando o argumento é preposicionado, porém, não é uma escolha inédita em UD. Como notam Marneffe et al. (2021) em outro momento do texto, a distinção nuclear-oblíquo ocorre apenas no nível da frase, de maneira que todos os dependentes dos

[7] "All or nearly all languages have a standard way of encoding the one or two arguments of most verbs, and this unmarked form of argument expression defines core arguments for that language." (MARNEFFE et al., 2021, p. 267)



(a) complemento nominal (b) adjunto adnominal

Figura 5.1: Anotação de complementos nominais e adjuntos adnominais em UD

nominais são oblíquos, não havendo distinção, portanto, entre o complemento (frase da figura 5.1-a) ou adjunto (frase da figura 5.1-b) adnominal, na terminologia da gramática tradicional.

Para solucionar o impasse da falta de distinção entre objeto precedido por preposição e adjuntos adverbiais, pois, além da distinção primária nuclear (nsubj, obj e iobj) e oblíquo (obl), Zeman (2017) propõe uma subespecificação do oblíquo, o *obl:arg*, para quando, além de preposicionado, o sintagma também é considerado argumento do verbo. Assim, as etiquetas trocam de rótulo mas a dificuldade de se distinguir o argumento do adjunto continua – a confusão, na GT, entre objeto indireto e adjunto adverbial aparece no projeto UD para língua portuguesa entre as etiquetas *obl* (sintagma preposicionado, dependente do verbo) e *obl:arg* (sintagma preposicionado, também dependente do verbo), sendo a primeira um adjunto adverbial e a segunda, um argumento verbal.

Concordamos com Przepiórkowski e Patejuk (2018) e, como já demonstramos, os testes linguísticos em grande medida são incompatíveis entre si ou insuficientes para distinguir adjuntos de argumentos, dando margem a que cada treebank escolha livremente os critérios que deseja adotar na anotação da subespecificação *obl:arg*. Nosso objetivo é garantir uma boa consistência interna, traduzida na concordância interanotadores, e tornar as análises informativas, uma vez que não ignora que haja diferença entre argumentos e adjuntos. Queremos distinguir 49 de 50, que serão *obl:arg* (argumento) e *obl* (adjunto), respectivamente, e igualar 51 e 52, que serão *obj* (argumento) e *obl:arg* (argumento), respectivamente.

49. Gostar de **sorvete**.

50. Viajou de **carro**.

51. Assistiu o **filme**.

52. Assistiu ao **filme**.

Para isso, não analisamos o fenômeno considerando espaços argumentais dos verbos, isto é, se ele necessitaria ou não de um complemento e se o sintagma preposicionado estaria exercendo a função de complementá-lo; no lugar, olhamos para o sentido do sintagma preposicionado – se o seu sentido é o sentido tradicionalmente associado a um advérbio (tempo, lugar, modo, finalidade, causalidade, conformidade), anotamos como *obl* e, na ausência de uma semântica adverbial, trata-se de um *obl:arg*, conforme indicamos na seção 5.2.1. Assim, deslocamos o foco da sintaxe – no espaço argumental do verbo – para características semântico-contextuais do sintagma nominal a ele associado. Nas frases do corpus Bosque já apresentadas, temos, portanto, as anotações das frases 53-56.

53. obl – adj. adv. de lugar: Os jogadores se dividem pelos dez **quartos** do alojamento, equipados com frigobar, ar condicionado, televisão e telefone.

54. obl – adj. adv. de lugar: Outros profissionais brasileiros, que atuam nos **EUA**, também participam.

55. obl – adj. adv. de lugar: Papa indica mulher para **secretaria**[8]

56. obl:arg – não tem sentido comumente adverbial: O PDT pretende reduzir os impostos federais a **quatro**.

Estratégia de revisão

A anotação linguística do PetroGold foi herdada do anotador Stanza (QI et al., 2020), que teve seu modelo para língua portuguesa treinado a partir do corpus Bosque-UD (RADEMAKER et al., 2017) v.2.5. Nesta versão do corpus, os argumentos e adjuntos foram anotados tentando manter um alinhamento entre corpus e informações lexicográficas obtidas pela consulta a dicionários, de tal maneira que se respeitasse a subcategorização de cada verbo de acordo com as informações dicionarizadas. O processo de transformação dessa anotação para a que propomos, utilizando o critério estabelecido, teve o auxílio de um método de agrupamento, já que o corpus é composto por 20.210 ocorrências de verbo (1.080 lemas diferentes), sendo muito custoso analisá-las caso a caso. Guiamos as revisões pelas preposições que se associam a cada

[8]Neste caso, o lugar não é físico – refere-se a um cargo, sendo a mulher indicada para se tornar secretária. Nosso direcionamento, como dissemos, é primeiro olhar para a função do nome – podendo ocupar o lugar de um advérbio (onde, quando, com que finalidade, etc.), anotamos como obl e, apenas na ausência dessas leituras, anotamos como obl:arg.

verbo (na árvore de dependências, as preposições são dependentes dos nomes, que são dependentes dos verbos) e agrupamos os verbos por lema.

Utilizando uma planilha, agrupamos os verbos por preposição e os anotadores do corpus indicaram se haveria ou não a possibilidade de o verbo, seguido por aquela dada preposição, ter o sintagma preposicionado como seu argumento utilizando o critério semântico. Os anotadores tiveram à disposição o corpus para consulta à procura de ocorrências em que o sintagma fosse argumento – e a presença de uma única frase com sintagma preposicionado sendo seu argumento já é o suficiente para indicar que o verbo pode ter argumento, pois, enquanto qualquer predicado pode ser modificado por um adjunto adverbial, nem todos os predicados podem ter argumento. Neste momento, optamos por agrupar os verbos na planilha por preposição considerando que, muitas vezes, o lema da preposição tende a introduzir um adjunto ou um argumento – por exemplo, na maioria das vezes a preposição "em" introduz um adjunto adverbial, enquanto que a preposição "de" tende a introduzir argumento.

Tendo completamente preenchido a lista de verbos e preposições associadas, realizamos as modificações no corpus em lote – por exemplo, indicamos na planilha que o verbo "acarretar" relacionado à preposição "em" pode ter no sintagma preposicionado seu argumento; portanto, todas as ocorrências de "acarretar em" tiveram o substantivo seguinte anotado como argumento do verbo (obl:arg). Consequentemente, casos como o da frase 57, onde há um sintagma preposicionado encabeçado por "em" que não é argumento do verbo "acarretar", foram indicados para os anotadores revisarem – o analisador automático havia identificado "em geral" como adjunto adverbial, mas a planilha indicou que "acarretar em" era sempre objeto, uma divergência que precisou de nossa solução. Assim, contrastando análise automática e planilha, os anotadores foram capazes de realizar a anotação dos argumentos e adjuntos utilizando o critério semântico.

57. Segundo Souza (2009), a estabilidade conferida às emulsões devido à presença dos agentes emulsionantes naturais acarreta, em **geral**, em um incremento significativo na sua viscosidade

Em seguida, verificamos a distribuição dos verbos no PetroGold em relação à sua subcategorização. Dividimos a subcategorização verbal em quatro tipos: os verbos plenamente transitivos são sempre acompanhados por argumento, os intransitivos nunca são acompanhados por argumento, e há ainda os verbos que estão entre os dois polos, ora são transitivos, ora intransitivos, considerando o resultado do processo de revisão da anotação.

Neste estudo, não estamos considerando os predicados nominais, pois na anotação em UD os verbos de ligação não são anotados como verbos, mas como auxiliares, e estamos considerando como argumentos verbais apenas os objetos – diretos ou indiretos, isto é, obj, iobj ou obl:arg –, já que o sujeito não nos interessa neste estudo da subcategorização verbal por não apresentar a dificuldade de anotação sendo estudada. No entanto, contabilizamos os sujeitos pacientes (nsubj:pass), de frases na voz passiva (frases 58-59), como argumento verbal, pois ocupariam a posição de objeto na voz ativa, implicando que o verbo permite um complemento no contexto da frase. E classificamos também como argumentos as orações subordinadas substantivas objetivas, anotadas como xcomp e ccomp em UD, postura também defendida por Przepiórkowski e Patejuk (2018).

58. Em todas as seções **foi feita** a descrição das fácies sedimentares, a documentação fotográfica e a medição de paleocorrentes, quando possível.

59. Geralmente **são utilizados** como métodos de estudo a análise química, a determinação da capacidade de troca de cátions, a análise térmica diferencial, a microscopia eletrônica, a difração de raio X e a espectroscopia no infravermelho.

E removemos da análise também todos os verbos no particípio (como nas frases 60-61) e os verbos com pronome "se" expletivos (frases 62-63) dependentes deles.

60. Isso pode ter **ocorrido** devido o clorofórmio extrair também o tensoativo.

61. Viscosidade vs. taxa de cisalhamento de poliacrilamida **hidrolisada**.

62. Estas fontes **se sobressaem** no mapa de amplitude do sinal analítico referido acima.

63. As rochas de a Bacia Sanfranciscana **assentam-se**, em discordância erosiva e angular, sobre rochas paleoproterozóicas do embasamento, rochas neoproterozóicas do Grupo Bambuí, rochas neoproterozóicas do Grupo Natividade e sobre rochas sedimentares da Bacia do Parnaíba (Sgarbi, 2011).

Nas frases com particípio, há dificuldade em distinguir, de forma automática, quais verbos não aceitam complemento ("[isso] ocorre Ø") e quais poderiam aceitá-lo ("[alguma reação] hidrolisou [a poliacrilamida]"). Nas frases

# verbos	preposição	# verbos	preposição
718	em	15	dentre
371	com	13	apesar_de
307	a	11	de_o_que
305	para	7	com_relação_a
250	de	6	diante_de
220	por	5	em_relação
121	através_de	5	mediante
103	dever_a	5	com_vista_a
94	a_partir_de	5	contra
83	durante	4	assim_como
80	após	3	de_que
78	entre	3	em_torno_a
78	até	2	como_exemplo_de
71	sobre	2	pra
63	segundo	2	exceto
60	de_acordo_com	2	a_de
41	sob	2	por_exemplo
39	desde	2	além_de
35	em_relação_a	2	a_exemplo_de
33	conforme	1	uma_vez_em
32	como	1	via
31	em_torno_de	1	de_ainda_acordo_com
23	quanto_a	1	perante
22	sem	1	a_favor_de

Tabela 5.2: Quadro com lista de preposições associadas a verbos no corpus

com "se", também há dúvida em relação à presença ou não de um objeto: ora o verbo funciona de fato como intransitivo ("[algo] se sobressai"), ora o verbo poderia ser interpretado como aceitando um complemento ("[algum fenômeno natural] assentou [as rochas]"). Um estudo sobre o fenômeno será realizado na seção 5.2.3.

5.2.1.2 Resultados

Como parte do método de agrupamento para facilitar a análise dos verbos, o quadro 5.2 traz a quantidade de verbos que se associam a cada uma das preposições no PetroGold.

O estudo contou com 9.653 ocorrências verbais que se distribuem em 719 lemas, 66% do total de lemas verbais no corpus. O gráfico abaixo ilustra quantos lemas verbais nunca são acompanhados por objeto, quantos são acompanhados por objetos menos de 30% das vezes, entre 30% e 70%, mais de 70% das vezes, e quantos sempre são acompanhados por objetos (obj, iobj,

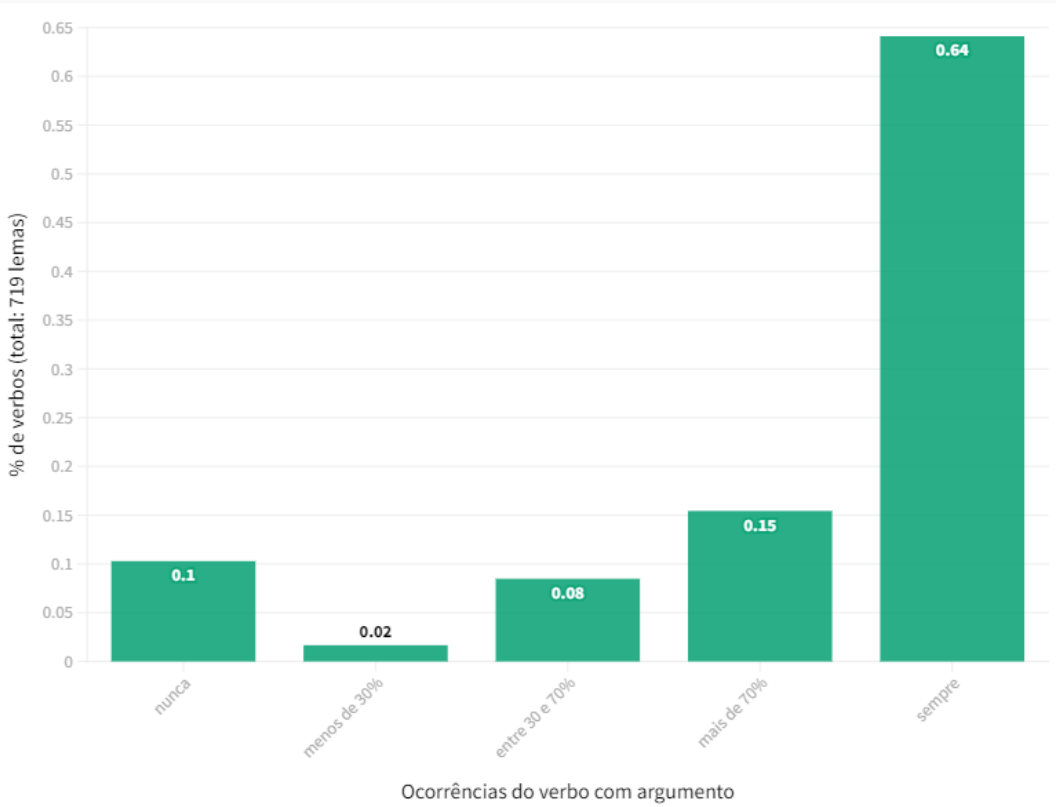


Figura 5.2: Número de verbos distribuídos pela frequência que ocorrem em cada subcategorização verbal

obl:arg, xcomp e ccomp).

Dos 719 lemas, 10% nunca ocorrem com argumento associado a eles, seja preposicionado (obl:arg) ou não (obj, iobj, xcomp, ccomp). Esses são os casos em que, empiricamente, baseados neste corpus, podemos confirmar que o verbo sempre se comportou como intransitivo. 64% dos lemas, por sua vez, sempre são acompanhados por algum argumento, sendo os chamados verbos transitivos da GT.

Todos os outros lemas, porém, podem ou não vir acompanhados de objeto – 2% dos lemas têm argumento em menos de 30% das suas ocorrências, portanto pode-se dizer que tendem à intransitividade; 15% dos lemas têm argumento em mais de 70% das suas ocorrências, tendendo à transitividade, e outros 8% de lemas apresentam comportamento mais indefinido, ocorrendo com argumentos entre 30 e 70% das vezes em que aparecem. Essa fatia de lemas que estão no meio, entre o "nunca" e o "sempre", corresponde a 25,8% dos lemas no corpus. Ou seja, um quarto dos lemas verbais estão exatamente no meio do caminho entre a intransitividade e a transitividade. Para todos esses casos, não se pode dizer, por um lado, que quando carecem de um complemento a frase está incompleta, e, por outro, não se pode dizer que o verbo não permite um complemento sem que se erre um tanto considerável de vezes com essa

afirmação. Esse tipo de informação estatística que obtivemos escapa a uma descrição categórica dos verbos, onde a irregularidade – um quarto dos lemas neste corpus que ora são empregados com argumento e ora sem – ocuparia espaço periférico.

Por exemplo, analisando apenas um dos verbos mais frequentes que têm entre 30% e 70% de chance de ocorrerem com argumento, vejamos o verbo "diminuir". Ele ocorre 43 vezes, sendo 27 das vezes com argumento (frases 64 e 65) e 16 das vezes sem argumento (frases 66 e 67). Como vimos, classificá-lo como transitivo seria correto em 63% das vezes, mas é também correto classificá-lo como intransitivo em outros 37%. Descrevê-lo como um verbo que se comporta, na maioria das vezes (63%), como transitivo, porém, é uma informação mais realista – uma descrição contínua, quantitativa e não-categórica, como defendeu Manning (2003).

64. A peça interna que prende a pastilha de a amostra é de aço , e possui furos para **diminuir** seu peso.

65. Teoricamente , conforme a carga aumenta , a oxidação de o combustível aumenta formando mais CO , CO2 e NOx , e **diminuindo** as emissões de O2.

66. Conforme as reservas de óleo leve **diminuem** , a produção de óleo pesado pode expandir se para atender a demanda.

67. Isto , pois quando o campo envelhece , a pressão natural de o reservatório **diminui** , e o poço pode ter reforço em a forma de compressão , a fim de produzir mais e levar esta produção até a plataforma.

No apêndice A, listamos os verbos presentes em cada uma das distribuições exploradas no gráfico. Alguns dos lemas podem parecer incomuns: "afundar", por exemplo, na coluna dos verbos que nunca vêm acompanhados de objeto, é um verbo que conseguimos encontrar seguido por um objeto direto. Sendo um corpus relativamente pequeno (cerca de 200 mil tokens) e com apenas duas ocorrências do verbo "afundar" (frases 68 e 69), podemos até concluir que não se trata de um material adequado para estudar a língua portuguesa – a tese de que "afundar" seria intransitivo é facilmente contestada em outros corpora, como no material da Floresta Sintá(c)tica (FREITAS; ROCHA; BICK, 2008) (frases 70 e 71)[9], onde encontramos ocorrências do verbo com objetos direto.

[9]A busca foi realizada utilizando a expressão [lema="afundar"] na ferramenta AC/DC da Linguateca, disponível em <https://linguateca.pt/ACDC>. Acesso em 10 de dez. 2021.

68. Durante boa parte do Cenozóico (58-20 Ma) a crosta continental fendeu-se e **afundou** em diversas áreas lineares formando corredores de grábens (riftes) paralelos à costa.

69. Elas podem **afundar** e reflutuar, dependendo do balanço das ondas no momento do derrame.

70. Temporal **afunda** duas embarcações

71. Falar Que O Show "Opinião" **Afundou** A Bossa Nova E Levantou A Mpb ?

De fato, não propomos afirmar que o verbo é intransitivo, pois estaríamos retornando a uma forma de classificação categórica. Antes, o que estamos realizando é uma descrição de como o verbo se comporta no corpus PetroGold, dadas as particularidades do material e como foi anotado. Como resultado, podemos obter uma descrição granular e fiel à realidade das formas de utilização dos verbos e dos seus argumentos.

Listamos também no apêndice B, a título de curiosidade, os verbos cujos únicos complementos são do tipo oracional (xcomp ou ccomp), os verbos que só se realizam no particípio, e os verbos que só se realizam acompanhados pelo pronome "se" expletivo. Lembramos que os dois últimos foram excluídos da análise realizada, enquanto que os primeiros foram adicionados ao grupo dos verbos transitivos.

5.2.2 Expressões multipalavras

Expressões multipalavras (MWEs) são construções que podem assumir muitas formas em uma língua, como nomes compostos (guarda-chuva e óleo diesel), frases institucionalizadas (comes e bebes) ou locuções funcionais (apesar de, de acordo com). Ramisch (2012) mostra que não há uma única definição para as MWEs na literatura, e que elas estão na zona cinzenta entre o léxico e a sintaxe, apresentando um problema relevante para o PLN uma vez que são difíceis de tratar e, ao mesmo tempo, muito frequentes, seja na comunicação do dia a dia, seja nas formas de comunicação mais especializadas.

Embora não haja uma definição consensual, o autor nota algumas características das expressões: (1) elas são arbitrárias, uma vez que expressões perfeitamente gramaticais podem não ser aceitas em determinados contextos; (2) elas são institucionalizadas, isto é, fazem parte de toda comunicação e são aceitas e entendidas pelos falantes como uma forma convencional de se dizer algo; (3) elas têm variação semântica limitada, pois não passam pelo processo

de composicionalidade semântica das demais construções da língua. Assim, não se pode substituir certas partes de uma MWE por quaisquer outras palavras ou construções, pois a expressão não é fruto de uma composição de palavras (tampouco se pode traduzir a MWE palavra a palavra); (4) elas têm variação sintática limitada, já que as regras gramaticais convencionais não se aplicam a essas expressões, sendo difícil dizer quais delas fariam parte do léxico ou da gramática de um falante (e muitas vezes são também extragramaticais, isto é, imprevisíveis e difíceis de entender para um falante aprendiz da língua que aprendeu apenas a gramática da língua), e (5) elas são heterogêneas, cobrindo um número enorme de fenômenos da língua, cada um com características linguísticas específicas, de maneira que aplicações de PLN não deveriam utilizar uma metodologia unificada para processá-las.

Ao tratar do fenômeno, McEnery e Hardie (2011) realizam um contraste entre o que chamam de locuções e colocações – de um lado, as locuções representam unidades linguísticas funcionais, isto é, que funcionam como uma única expressão do ponto de vista linguístico, ao passo que as colocações representam empiricamente, com dados estatísticos oriundos apenas do corpus, as combinações mais frequentes da língua (ou dos textos estudados)[10]. Classificar uma expressão como locução é uma tarefa que necessita de interpretação linguística, enquanto as colocações não necessariamente. As locuções são identificadas tendo como base hipóteses linguísticas pré-concebidas para identificar as expressões que, segundo uma análise, apresentam um comportamento linguístico específico, enquanto as colocações podem ser coletadas inteiramente de um corpus, utilizando métodos estatísticos.

Câmara (1978) considera que locuções são casos de gramaticalização, de modo que mais de uma palavra, mantendo suas formas fonéticas e morfêmicas, veem seu significado lexical desaparecer, funcionando como uma unidade funcional (prepositiva, adverbial, conjuntiva). Já a colocação, de acordo com Manning e Schutze (1999), representa as formas convencionais de se dizer algo com mais de uma palavra.

Delimitar e anotar corretamente expressões multipalavras é uma tarefa importante na construção de um treebank padrão ouro. Do ponto de vista do aprendizado automático, é importante tornar consistente a anotação de MWEs para não fornecer pistas dúbias sobre quais palavras, quando juntas, devem ser tratadas como uma unidade em determinados contextos. Sem indicação de quais são as MWEs que serão anotadas como locução em um corpus,

[10]Por exemplo, uma gramática pode considerar "a despeito de" uma locução, pois funciona como uma unidade (do tipo prepositiva), e os dados estatísticos de um corpus podem indicar que a expressão "comes e bebes" é uma colocação, dada a frequência com que as palavras co-ocorrem.

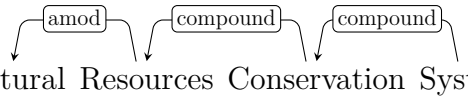


Figura 5.3: Possível anotação para a estrutura de “dois nominais”

a anotação morfossintática desses fenômenos pode se tornar inconsistente, sendo diferente nas diversas ocorrências ou, no limite, pode ser impossível realizar qualquer anotação que faça sentido do ponto de vista morfossintático para certas expressões sem que sejam encaradas como uma multi-palavra – a expressão “isto é”, por exemplo, quando é utilizada como locução conjuntiva, não pode ser anotada como composta por um sujeito (o pronome “isto”) e um verbo de ligação, sob perigo de inviabilizar a anotação sintática de todo o restante da oração.

O projeto UD dispõe de três classes para anotar expressões multipalavras, a saber: (1) fixed, para as expressões fixas que correspondem a expressões gramaticalizadas, que se comportam como palavras funcionais ou adverbiais curtas[11]; (2) flat, para as expressões “semi-fixas”, embora não haja definição para o que são exceto por uma lista de exemplos (nomes de pessoas, datas, números compostos e sintagmas estrangeiros)[12], e (3) compound, para as expressões que, diferentemente das demais, têm uma palavra funcionando como governante sintático, como no exemplo “apple pie” (ou, no nosso caso, óleo diesel)[13].

UD tem uma abordagem bastante econômica com relação ao que considera expressões multipalavras. Por exemplo, uma estrutura muito comum para o inglês é a de dois nominais – expressões como “phone book” e “Natural Resources Conservation Service”, por exemplo, ou “planta piloto” e “fase rifte”, no PetroGold. As diretivas indicam que, quando há critérios claros na documentação de uma língua para distinguir expressões compostas, a expressão pode ser anotada como uma MWE do tipo *compound*, onde todas as palavras da expressão são anotadas como dependentes da palavra principal com a relação *compound* (figura 5.3).

Nas versões 1 e 2 do PetroGold, ensaiamos anotar como *compound* algumas expressões que pudessem ser úteis para o domínio do petróleo, como “óleo diesel” e “meio ambiente”. No entanto, na ausência de um estudo mais abrangente sobre o assunto e por conta das limitações de tempo do projeto, optamos por, nesta terceira versão do corpus, abrir mão da etiqueta *compound*, como sugerem as diretivas de UD, dando lugar ao adjunto adnominal regular

[11] <https://universaldependencies.org/u/dep/fixed.html>
[12] <https://universaldependencies.org/u/dep/flat.html>
[13] <https://universaldependencies.org/u/dep/compound.html>

(nmod), isto é, uma anotação transparente. Na frase 72, "desvio padrão", outrora considerado MWE, com "padrão" como dependente de "desvio" com a relação *compound*, passou a ter a segunda palavra anotada com a relação de *nmod* (adjunto adnominal). A relação de "desvio" para "valores", por sua vez, não alterou, permanecendo como do tipo *conj* (coordenação).

72. Os cálculos de os valores médios , **desvio padrão** e erros padrões estão em a Tabela IV.23 .

Outro tipo de construção que pode ser considerada MWE em certos contextos é a que contém verbos leves (ou verbos suporte), como nas expressões "dar um grito", "ter em mente", "tirar um cochilo", "tomar uma decisão" ou "fazer vista grossa". Nesses exemplos, os verbos dar/ter/tirar/tomar/fazer (além de outros) são verbos "(...) de significado bastante esvaziado que formam, com seu complemento (objeto direto), um significado global, geralmente correspondente ao que tem um outro verbo da língua" (NEVES, 2000). Bagno (2012) nota que nem todas as expressões com verbo suporte podem ser substituídas por outros verbos da língua, o que justifica o uso do advérbio "geralmente" por Neves, pois enquanto "dar um grito" e "tomar a decisão" correspondem a "gritar" e "decidir", respectivamente, "fazer questão" e "soltar balão" não correspondem a nenhum verbo (e tampouco são composicionais[14]), de maneira que as expressões estariam suprindo uma lacuna do léxico do português por meio de uma expressão (um sintagma, portanto)[15]

De qualquer modo, em UD definiu-se que tais expressões, com verbos suporte, devem ter uma anotação transparente, sendo o substantivo um objeto do verbo[16]. Assim, na frase 73, "origem" é objeto do verbo "dar" (obj), e "pacote" é um objeto preposicionado do mesmo verbo (obl:arg). Já na frase 74, "parte" é objeto do verbo "fazer", e "Projeto" é objeto preposicionado do mesmo verbo, a despeito de "fazer parte" poder ser considerada uma expressão multipalavras na língua portuguesa.

[14]Ambas as expressões, "fazer questão" e "soltar balão", podem ser lidas como MWEs ou não em diferentes contextos, como no caso de um professor que formula (ou faz) questões de provas (ou que "faz questão" de que alguém compareça a uma festa) ou uma criança que deixa cair (ou solta) uma bexiga (balão de festa) no chão após um pedido da mãe (ou um adulto que, antes de se tornar proibido, "soltava balão" de festa junina).

[15]Para o autor, o fato de que algumas dessas expressões não encontram no léxico um correspondente é um argumento para o fim da diferenciação entre léxico e gramática, pois léxico e gramática interagem a todo instante, sendo difícil delimitar o espaço de cada uma.

[16]Embora digam que cada língua deve definir quais os critérios para anotar *compound*, as diretivas UD indicam que, para o inglês, a anotação "transparente" é a mais adequada para construções "leves" (tomar uma decisão) e combinações adjetivo + substantivo (*hot-dog*). Fonte: <https://universaldependencies.org/u/dep/compound.html>. Acesso em 5 de mar. 2023.

73. A linha de charneira Cretácea , formada por as falhas normais que romperam o embasamento raso de as áreas proximais e o lançaram a grande profundidade **dando origem** a o espesso pacote sedimentar de as referidas bacias (Macedo 1989 , Almeida e Carneiro 1998) , está incluída parcialmente em a área de pesquisa (figura 1.1) .

74. Este levantamento foi realizado em o ano de 1978 , **fazendo parte** de o Projeto Aerogeofisico São Paulo – Rio de Janeiro de a CPRM .

Assim, resta para UD em português apenas um tipo de expressão multipalavras – as expressões fixas. Embora sigamos a definição de expressões multipalavras presente no projeto UD, não há nenhuma lista, definitiva ou em progresso, de expressões multipalavras para língua portuguesa seguindo as diretivas do projeto.

Por isso, buscamos três fontes distintas de expressões multipalavras fixas: (1) as expressões previamente identificadas pelo anotador PALAVRAS (BICK, 2014) e presentes no corpus Bosque-UD; (2) as colocações do tipo [PREP (DET)? N PREP], como "de acordo com", encontradas por meio de métodos estatísticos inspirados na abordagem de Oliveira, Nogueira e Garrao (2004) no Petrolês, e (3) uma lista de palavras que co-ocorrem sem ambiguidade, isto é, que estão sempre anotadas da mesma forma quando estão juntas, compilada no âmbito do POeTiSA (LOPES; DURAN; PARDO, 2021).

As três listas de expressões foram filtradas visando adequá-las à definição de expressões multipalavras do projeto UD utilizando um critério de distinção para MWEs e, como resultado, geramos duas listas – uma junção de todas as MWEs encontradas utilizando os três métodos (150), e um subconjunto desta, com todas as MWEs encontradas no PetroGold (112) e a sua anotação correspondente. No final, 2.467 MWEs tiveram a anotação revista no PetroGold. O critério e as listas serão apresentados a seguir.

5.2.2.1 Metodologia

Como identificar MWEs do tipo *fixed*

A base para a distinção de uma MWE, conforme discutido em McEnery e Hardie (2011), é a não-composicionalidade do termo, isto é, o fato de que o sentido da expressão não é obtido pela "soma" dos sentidos de cada uma das palavras que a compõem. É um critério semântico, indicativo de que houve gramaticalização ou lexicalização, e pode resultar em anotações divergentes

no corpus uma vez que identificar se uma expressão está sendo usada de forma transparente ou opaca depende também da interpretação da expressão no contexto.

De fato, a expectativa não é a de que as MWEs que anotamos no PetroGold sejam unanimidade em todas as análises para língua portuguesa; no entanto, propomos um critério para garantir maior consistência às análises. Assim, anotamos uma expressão como MWE do tipo *fixed* quando a anotação só é possível dessa forma, pois qualquer outra análise é incorreta, de tão gramaticalizada que a expressão está. O critério não descarta a possibilidade de que as palavras que compõem a MWE possam ter anotação alternativa em outros contextos, pois o que de fato interessa é o contexto em que a expressão está inserida.

Em expressões como "visto que" e "a partir de", uma análise sintática transparente resultaria em árvores sintáticas pouco convencionais e possivelmente erradas do ponto de vista do formato da gramática. Na frase 75, com a expressão "visto que", uma leitura composicional julgaria que "visto" é um verbo no particípio, núcleo de uma oração adverbial dependente da oração principal, e que a oração seguinte, cujo núcleo é "dependente", seria uma oração subordinada substantiva subjetiva, funcionando como sujeito de "visto". A análise não é formalmente errada, pois pode ser codificada na anotação, mas é errada para um padrão-ouro pois não corresponde à leitura humana, onde "visto" não pode ser uma oração, mas parte de uma locução conjuntiva, "visto que".

De modo semelhante, na frase 76, uma leitura não-fixa julgaria que "partir" é um verbo no infinitivo que funcionaria, talvez, como oração adverbial da oração adverbial. Nessa hipótese, "processos" só poderia ser, formalmente, objeto de "partir", o que não encontra nenhuma correspondência na realidade quando a frase é interpretada. Assim, a leitura adequada é a de que "a partir de" é uma locução prepositiva, servindo de ligação entre o adjunto adverbial "processos" e o núcleo da oração anterior, "realizada".

75. O que já era esperado , **visto que** o fluxo de o campo magnético induzido é dependente de a corrente fornecida .

76. Como visto , a recuperação de petróleo pode ser realizada **a partir de** processos convencionais e avançados .

Há casos também onde a leitura transparente para as palavras que compõem a MWE é possível, mas em contexto diferente. Nas duas frases a seguir, por exemplo, o mesmo conjunto de palavras, "isto é", pode ser considerado MWE ou não dependendo do contexto. Na frase 77, "isto é"

funciona como locução conjuntiva, o que acarreta em uma anotação de MWE fixa, e o que garante que essa análise é a adequada é o fato de que, caso contrário, se a anotação fosse transparente, "isto" seria sujeito de um verbo auxiliar, "é", dando origem a uma outra interpretação da frase – a interpretação da frase 78. Para garantir a diferenciação entre ambas as leituras, anotamos a primeira frase como MWE e a segunda, não.

77. ***MWE fixed:*** A viscosidade , em o entanto , também é influenciada por o estado físico de os asfaltenos , **isto é** , o tamanho e estrutura de as micelas formadas por a interação com resinas e aromáticos .

78. ***nsubj:*** O aquecimento indutivo para este teste não foi satisfatório , **isto é** demonstrado em o Gráfico 1 .

Obtenção dos candidatos a MWE

Tendo em mente o critério explicitado, utilizamos três métodos para obter candidatos a expressões multipalavras do tipo fixa. A primeira fonte de expressões multipalavras fixas foi obtida a partir da herança de anotação do PetroGold. Como já apresentado na seção 4.1, o corpus foi originalmente anotado por um modelo treinado no corpus Bosque-UD, cuja anotação, por sua vez, é fruto de uma conversão com correções manuais da anotação do sistema PALAVRAS. Assim, a primeira lista de expressões que analisamos é a de expressões anotadas como unidades no PetroGold como herança do PALAVRAS e fruto de revisões pontuais no corpus.

Na preparação da primeira e da segunda versão do PetroGold, algumas das revisões realizadas no corpus diziam respeito à anotação de expressões multipalavras. No entanto, nenhuma das revisões foi tão abrangente, detalhada, e nem foi avaliada isoladamente, como será feito para essa terceira versão do corpus. Dessa vez, todas as expressões anotadas no corpus foram revistas, sendo avaliado se estão adequadas ao critério estabelecido e adicionando informação relativa à classe gramatical da expressão multi-palavra.

A segunda fonte de expressões multipalavras foi obtida a partir da aplicação de métodos estatísticos ao Petrolês, o grande conjunto de textos do projeto, contendo 330 documentos acadêmicos do domínio do petróleo, sendo que o PetroGold corresponde a apenas 4,3% do tamanho desse corpus. Os resultados foram posteriormente filtrados manualmente utilizando o critério de identificação de MWEs já apresentado.

Os métodos estatísticos aplicados para encontrar colocações foram inspirados por Oliveira, Nogueira e Garrao (2004), que lidam com as noções de

colocação e locução com o objetivo de verificar em que casos as locuções prepositivas são tanto locuções – unidades linguísticas – quanto colocações – palavras que co-ocorrem frequentemente. As locuções analisadas são as do tipo [PREP (DET)? N PREP], como "de acordo com", em que não há o determinante, e "em o caso de" (no caso de), com o determinante.

O estudo de Oliveira, Nogueira e Garrao (2004) tomou por base duas amostras: (i) uma lista de 169 locuções prepositivas pré-compilada pelo NILC (Núcleo Interinstitucional de Lingüística Computacional) com adições oriundas de Oliveira et al. (2009), e (ii) uma lista de colocados gerada utilizando como base o corpus CETENFolha[17] (Corpus de Extractos de Textos Eletrônicos – NILC/Folha de São Paulo), contendo cerca de 24 milhões de palavras em Português brasileiro e compilado pela Linguateca.

Os autores concluem, a partir da lista de locuções e colocados do CETENFolha, que o melhor método de identificação de colocados retorna 86% das 169 locuções prepositivas nos primeiros 10.000 colocados do corpus. O número de colocados retornados para um corpus é consideravelmente maior que o de locuções reconhecidas como tal por linguistas (nesse caso, a proporção é de 10.000 colocados para menos de 169 locuções), sendo uma tarefa humanamente custosa analisar uma grande quantidade de colocados para filtrar aquelas que podem ser consideradas locuções.

Realizamos algo semelhante ao trabalho de Oliveira, Nogueira e Garrao (2004) no Petrolês: calculamos estatisticamente, por meio da biblioteca NLTK (Natural Language Toolkit) (BIRD, 2006) para a linguagem de programação Python, quais os colocados presentes no corpus e verificamos quais deles se enquadrariam como locuções prepositivas (e não apenas colocações) em uma amostra dos resultados (as 40 entradas melhor avaliadas pelo algoritmo[18]).

O método que utilizamos para a identificação dos colocados foi o *Likelihood-ratio*, um dos utilizados por Oliveira, que mede a probabilidade de eventos que aconteceram ao mesmo tempo não serem fruto de coincidência. Assim, calcula-se duas hipóteses: (i) a de que as palavras têm a mesma probabilidade de aparecerem juntas ou separadas, e (ii) a de que as palavras têm mais probabilidade de aparecerem juntas do que separadas. A métrica nos diz o quanto a hipótese (ii) é mais provável que (i) e, se for o caso, a sequência de palavras é considerada uma colocação (MANNING; SCHUTZE, 1999).

A terceira e última fonte de expressões multipalavras foi compilada no âmbito do projeto POeTiSA, como parte de uma série de recursos linguísticos

[17]Disponível a partir do endereço: <https://www.linguateca.pt/CETENFolha/>. Acesso em 1 jul. 2021.

[18]O limite de 40 foi estabelecido pois, a partir desse número, tornou-se muito difícil encontrar locuções prepositivas por meio de análise manual

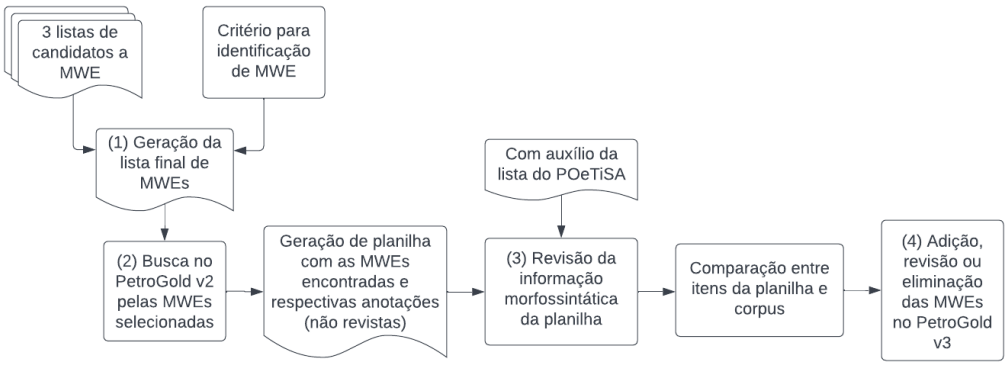


Figura 5.4: Etapas da revisão de MWEs

para melhorar a qualidade da anotação de POS de um corpus. Um desses recursos é uma lista de palavras que, quando co-ocorrem, não são ambíguas e devem ter sempre a mesma anotação de POS, compilada por um linguista durante a anotação do corpus[19]. Os autores notam que nem todas as entradas da lista são necessariamente MWEs, portanto realizamos uma análise para filtrar aqueles casos que, segundo o nosso critério, deveriam ser retirados da lista.

Estratégia de revisão

Para realizar a revisão da anotação das MWEs no PetroGold, utilizamos metodologia e ferramentas específicas. O fluxograma da figura 5.4 ilustra as quatro etapas da revisão das MWEs no corpus, começando pela geração da lista final de MWEs, que foi compilada a partir das três listas de candidatos a MWE obtidas pelos métodos discutidos na seção 5.2.2.1 utilizando o critério para identificação de MWE discutido na seção 5.2.2.1.

Uma vez tendo sido gerada a lista de MWEs, procuramos por cada uma delas no PetroGold (etapa 2) e precisamos revisar a anotação de POS, relação e dependente sintáticos (etapa 3).

As expressões fixas são, segundo UD, um tipo de MWE que é empregado quando a expressão não tem estrutura interna (exceto em uma perspectiva histórica, segundo as diretivas[20]), de modo que seria arbitrário escolher qualquer anotação para a expressão. Assim, optou-se por anotar todas as palavras como dependentes da primeira, sendo a relação do tipo "fixed".

[19]Agradecemos à Magali Duran e à equipe do POeTiSA pela disponibilização da lista e de todos os demais recursos do projeto.

[20]Disponível em: <https://universaldependencies.org/u/dep/fixed.html>. Acesso em 26 de ago. 2022.

Desse modo, na frase 79, as duas expressões em negrito, “de acordo com” e “em função de”, foram anotadas como MWE, ou seja, as palavras “acordo” e “com” são dependentes de “de” com a relação “fixed”, assim como “função” e “de” também são dependentes de “em” com a mesma relação de dependência. Em ambas as expressões, a primeira palavra, núcleo da expressão (“de” e “em”, respectivamente) é dependente do substantivo seguinte (“taxa” e “quantidade”, respectivamente) com a relação de dependência “case”, designada às preposições – assim, encara-se que são ambas locuções prepositivas.

79. A taxa de produção extra de óleo varia **de acordo com** a taxa de injeção de CO2 de os poços e a taxa unitária de produção de óleo **em função de** a quantidade de CO2 que é injetado .

Contudo, as palavras, isoladamente, seguem tendo POS e características morfológicas anotadas tal como se fossem palavras isoladas – em ambos os casos, “ADP”, “NOUN” e “ADP”, respectivamente. Para essas expressões, isso não é um problema, pois a palavra é introduzida por uma preposição e a locução é mesmo do tipo prepositiva. Já na frase 80, há uma incompatibilidade entre a classe gramatical das palavras e a sua função na oração: composta por preposição mais adjetivo, a expressão “em geral” funciona como um advérbio, de maneira que a palavra “em”, núcleo da MWE, foi anotada como dependente de “reativas”, núcleo da oração[21], com a relação de dependência “obl”, utilizada quando quem exerce a função de advérbio não é da classe dos advérbios.

80. Pode se observar que as argilas hidratadas que apresentam afinidade por o metal são , **em geral** reativas .

Para contornar a disparidade entre função sintática e classe gramatical, utilizamos uma estratégia já utilizada no Bosque-UD e aplicamos a todas as expressões fixas do PetroGold – na coluna de informações do tipo miscelânea da primeira palavra da expressão, adicionamos a anotação “MWEPOS”. Assim, na frase acima, a palavra “em” tem a informação “MWEPOS=ADV” na coluna “misc”, indicando que a expressão, que funciona como uma locução adverbial, tem POS de advérbio, a despeito de nenhuma das palavras ser um advérbio isoladamente. Essa anotação tem dois objetivos: (1) fornecer uma informação adicional, que não é necessária nem será utilizada pelo projeto UD, e (2) permitir avaliar o impacto no aprendizado caso, no lugar da POS das palavras isoladamente, anotássemos as palavras com a POS da locução (prepositiva,

[21]Em UD, verbos de ligação não são núcleo de oração, de maneira que “reativas” ocupa esse papel.

adverbial ou conjuntiva), como ocorre em outros modelos de anotação, como o do PALAVRAS, o que testaremos no capítulo de avaliação do treebank.

Foram utilizadas duas estratégias para viabilizar a correção das MWEs no corpus. Primeiro, em uma planilha, listamos as informações de POS, MWE-POS (quando disponível), relação sintática da MWE com o restante da oração e características morfológicas das palavras que compõem a MWE conforme já estavam anotadas no corpus, com o único objetivo de economizar tempo na revisão humana das informações. Para facilitar ainda mais a revisão da planilha, contamos com as informações disponíveis para as MWEs encontradas na lista do projeto POeTiSA, uma vez que elas já vinham acompanhadas de informação revista relativa à POS de cada uma das palavras que compõem a expressão. Uma vez tendo sido revista a planilha – disponível no apêndice D –, prosseguiu-se uma revisão semiautomática do corpus utilizando a ferramenta CoSMO, que permite contrastar duas análises distintas para uma frase: neste caso, contrastamos a anotação do corpus ao resultado de regras de correção automática para as MWEs seguindo as informações da planilha (no fluxograma da figura 5.4, corresponde à etapa 4)[22].

O uso de uma ferramenta como o CoSMO foi importante pois a anotação das MWEs no corpus estava em grande parte inconsistente, sendo impossível prever sem erros como todas as MWEs estavam anotadas para sugerir regras que sempre fossem corrigir corretamente o corpus. Assim, uma vez tendo realizado o contraste com as novas regras de correção, era possível realizar quaisquer outras correções na frase dentro da interface do mesmo programa, como a mudança de núcleo das relações de dependência.

Por exemplo, viu-se que expressões como “antes de”, “depois de” e “diante de” estavam anotadas como expressões fixas no PetroGold. Contudo, decidimos retirar tais expressões da lista, uma vez que os advérbios podem ter o papel de núcleo para um complemento nominal ou verbal, como na frase a seguir: “Este tipo de gás, antes de ser distribuído, precisa ser separado do óleo.”

Nela, já com a anotação revista, a oração cujo núcleo é “distribuído” modifica o advérbio “antes” – o governante da relação –, o que é uma análise correta. Anteriormente, expressões como essa eram consideradas MWEs do tipo prepositiva: “antes de” era uma expressão fixa, servindo de conectivo para a oração “distribuído” e a oração principal, mas a anotação foi desfeita pela possibilidade da leitura transparente. Foi preciso realizar uma mudança de núcleo do sintagma – o núcleo, que era “distribuído”, passou a ser “antes”

[22] A ferramenta CoSMO está disponível gratuitamente para Windows e Linux no endereço: <https://github.com/alvelvis/conllu-merge-resolver>. Acesso em 5 de mar. 2023.

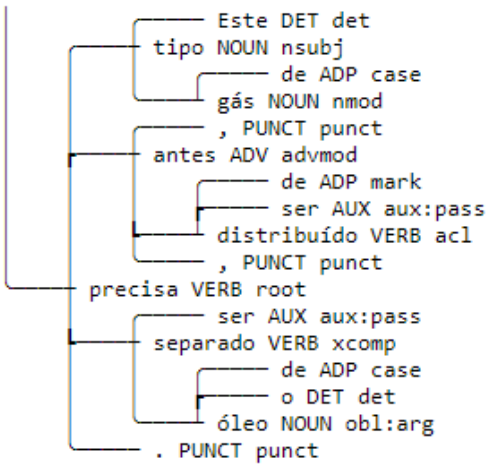


Figura 5.5: Anotação de "antes de" após as revisões

–, o que envolve mudar a dependência também das vírgulas, por exemplo, além da preposição "de", que passou a servir de conectivo para a oração seguinte. Embora o processo tenha sido auxiliado por regras de correção, precisou ser semiautomático por conta da imprevisibilidade da anotação de todas as orações.

A correção semiautomática, por sua vez, foi realizada em duas etapas – primeiro, removemos a anotação de *fixed* de todas as expressões que foram removidas da planilha após ter passado por revisão, como é o caso da frase exemplificada acima. Depois, adicionamos a anotação de *fixed* às expressões da planilha que já não estivessem anotadas dessa forma no corpus, realizando também a correção de POS e características morfológicas, além do MWEPOS, informação disponível na coluna "misc" de todas as expressões do PetroGold v3.

5.2.2.2 Resultados

O primeiro método para obtenção de uma lista de MWEs foi a revisão do PetroGold, que havia sido anotado por um modelo treinado no Bosque-UD inicialmente. A lista de MWEs fixas no PetroGold continha, inicialmente, 148 itens. Desses, apenas 101 restaram após a análise do critério utilizado para anotar expressões. A tabela 5.3 lista as expressões consideradas corretas; contudo, ela ainda não é a lista definitiva de expressões anotadas no PetroGold, pois a ela acrescentamos as expressões do Petrolês e do projeto POeTiSA, assunto que discutiremos em seguida.

O segundo método utilizado foi o estatístico. A tabela 5.4 lista os 21 trigramas (sem o determinante) e 7 quadrigramas (com o determinante)

a a medida em que	além de o mais	de tal forma que	por causa de
a a medida que	além de o que	desde que	por conseguinte
a as vezes	além de o quê	devido a	por exemplo
a exemplo de	apesar de	em a faixa de	por exemplos
a favor de	assim como	em a verdade	por fim
a fim de	assim por diante	em função de	por meio de
a longo de	assim sendo	em geral	por o menos
a medida que	até que	em o entanto	por sua vez
a não ser que	bem como	em razão de	por vezes
a o contrário de	cada vez mais	em relação a	quanto a
a o invés de	caso contrário	em relação as	quanto mais
a o longo de	cerca de	em relação á	sem que
a o menos	com isso	em torno de	sempre que
a o menos que	com relação a	em vez de	sendo assim
a o passo que	com vistas a	enquanto que	sendo que
a o ponto de que	como relação a	isto é	tais como
a o que	como também	junto a	tal como
a partir de	de acordo com	já que	tanto quanto
a principio	de acordos com	mesmo assim	tanto que
a princípio	de aí	mesmo que	toda vez que
a priori	de forma a	não que	um pouco
a seguir	de maneira a	não só	uma vez em
ainda que	de modo a	ou seja	uma vez que
além de	de modo que	para que	visto que
além de isso	de o que	para tal	é que
além de isto			

Tabela 5.3: Quadro com lista de MWEs obtidas pelo primeiro método

encontrados que foram considerados locução prepositiva e, portanto, devendo ser anotados como MWE fixas se encontradas no PetroGold. A precisão do método, considerando que foram analisadas as 40 primeiras entradas dos resultados retornados pelo algoritmo, foi de 70%, sendo que o número cairia drasticamente caso analisássemos expressões retornadas com menor pontuação atribuída pelo algoritmo. Encontra-se, no apêndice E, uma lista com as 100 primeiras entradas encontradas pelo algoritmo, de trigramas e quadrigramas.

O terceiro método, compilado pelo POeTiSA, tinha 110 expressões com palavras que co-ocorrem sem ambiguidade. Após nossa análise, sobraram 55 MWEs. O resultado pode ser observado na tabela 5.5.

As três listas apresentadas – expressões encontradas na anotação original do PetroGold, extraídas por meio de métodos estatísticos do Petrolês e compiladas no projeto POeTiSA – têm entradas em comum e outras que são únicas, ressaltando a eficiência da estratégia utilizada para obter cada uma das listas. O diagrama 5.6 mostra a proporção de MWEs encontradas por método.

Como complemento, a tabela 5.6 apresenta as entradas que foram iden-

a cargo de	a o longo de	em a faixa de	em termos de
a despeito de	a partir de	em direção a	em torno de
a favor de	a respeito de	em face de	em vez de
a fim de	a título de	em função de	por causa de
a o contrário de	com base em	em o caso de	por meio de
a o invés de	com relação a	em o tocante a	por parte de
a o largo de	de acordo com	em relação a	por volta de

Tabela 5.4: Quadro com lista de MWEs obtidas pelo segundo método

à toa	ainda que	em relação a	por muito que
a despeito de	apesar de que	em seguida	pelo menos
a fim de que	assim que	em separado	por pouco que
a menos que	caso contrário	em vão	por sua vez
a não ser que	de agora em diante	já que	pouco a pouco
ao certo	de forma que	logo que	se bem que
ao passo que	de maneira que	mais do que nunca	sem mais nem menos
ao todo	de modo que	nada que	sem que
ao vivo	de sorte que	não obstante	sempre que
aos poucos	de tal forma que	nem ao menos	tudo quanto
a partir de	desde que	nem mesmo	um a um
a ponto de	devido a	nem sequer	um por um
a seguir	em geral	no entanto	volta e meia
ainda mais que	em razão de	por mais que	

Tabela 5.5: Quadro com lista de MWEs obtidas pelo terceiro método

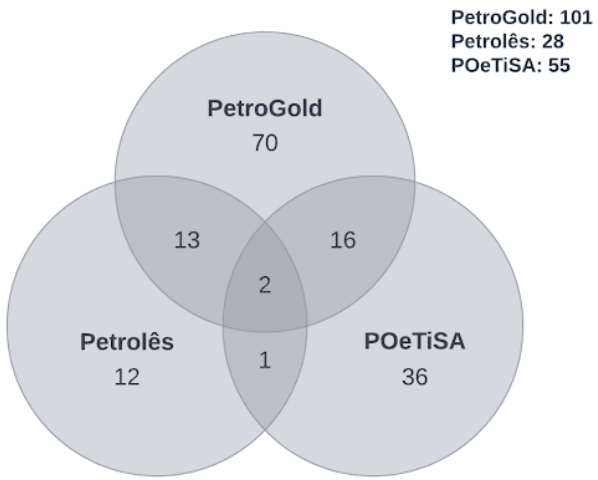


Figura 5.6: Quantidade de MWEs obtidas por cada método

tificadas por apenas um dos métodos – 70 expressões na anotação automática do PetroGold, 12 com os métodos estatísticos aplicados ao Petrolês, e 36 na lista compilada do projeto POeTiSA. A análise das expressões encontradas por cada um dos três métodos individualmente parece indicar que são expressões que poderiam ser encontradas em qualquer texto em língua portuguesa contemporâneo, independentemente de gênero. Os métodos são complementares, portanto, apenas na quantidade de expressões diferentes que retornam.

Já nas interseções entre os resultados de cada um dos métodos, duas MWEs foram encontradas por todos os métodos – a partir de, em relação a – e uma foi encontrada tanto pelo método estatístico no Petrolês, quanto pela lista compilada manualmente no projeto POeTiSA – a despeito de. 13 expressões foram encontradas no Petrolês e na anotação do PetroGold – em torno de, em vez de, em função de, de acordo com, a favor de, em a faixa de, por causa de, a o invés de, com relação a, por meio de, a o contrário de, a fim de, a o longo de –, e 16 foram as expressões encontradas na anotação automática do PetroGold e na lista do POeTiSA – em o entanto, ainda que, caso contrário, devido a, por sua vez, desde que, a seguir, já que, de tal forma que, em geral, sempre que, a o passo que, a não ser que, em razão de, de modo que, sem que.

Proporcionalmente, a lista do PetroGold é a que mais trouxe MWEs que nenhum outro método encontrou – 69,3% das entradas são únicas –, seguida de perto pela lista do POeTiSA (65,45%), e, distante dos demais métodos, a lista do Petrolês (42,85%). Embora o método estatístico tenha sido o menos eficiente na busca pelas MWEs, pode-se investigar futuramente se a busca realizada – por expressões do tipo [PREP DET? N PREP] – e os algoritmos utilizados são o motivo da baixa abrangência de MWEs retornadas pelo método.

Finalmente, a tabela 5.7 compila as três listas retornadas, sem as repetições, contendo 150 entradas. As 150 entradas foram aplicadas ao PetroGold para que as MWEs fossem corrigidas, conforme explicado.

Tendo como base os critérios de anotação de MWE discutidos e utilizando os métodos de obtenção de candidatos a MWE e a estratégia apresentada, o resultado é um total de 2.467 ocorrências de expressões multipalavras do tipo fixa na terceira versão do PetroGold. A lista com todas as 112 MWEs encontradas no PetroGold e a anotação morfossintática correspondente encontra-se no apêndice D.

O número de MWEs, 2.467, não é tão diferente da quantidade de expressões na versão anterior, que não continha revisão consistente das MWEs – na versão 2 do PetroGold havia 2.356 ocorrências de MWE anotadas fruto da anotação automática e de alguma revisão pontual. No entanto, embora a diferença quantitativa seja pouca (111), esse número é a diferença entre

PetroGold		**Petrolês**	**POeTiSA**
quanto a	de acordos com	em o caso de	a o vivo
a longo de	em relação á	a respeito de	em separado
a o que	assim por diante	com base em	sem mais nem menos
em relação as	como também	a título de	nem sequer
sendo que	ou seja	em face de	tudo quanto
a princípio	a exemplo de	em o tocante a	de agora em diante
por vezes	bem como	em termos de	se bem que
além de o quê	uma vez em	em direção a	um por um
de o que	enquanto que	por parte de	não obstante
para tal	além de o que	a cargo de	a o todo
a a medida que	mesmo assim	a o largo de	nem a o menos
tal como	além de o mais	por volta de	a fim de que
tais como	de aí		ainda mais que
um pouco	além de		nem mesmo
mesmo que	tanto quanto		a o certo
para que	quanto mais		pouco a pouco
visto que	assim como		por pouco que
além de isto	a priori		aos poucos
não só	a a medida em que		a menos que
por exemplos	uma vez que		a ponto de
tanto que	é que		assim que
de forma a	a principio		em seguida
além de isso	de modo a		nada que
por conseguinte	cada vez mais		em vão
isto é	por exemplo		por mais que
toda vez que	cerca de		por muito que
a o ponto de que	até que		apesar de que
a o menos	sendo assim		logo que
em a verdade	apesar de		pelo menos
a o menos que	de maneira a		volta e meia
como relação a	assim sendo		de maneira que
por fim	com vistas a		de forma que
por o menos	a medida que		a a toa
a as vezes	não que		de sorte que
com isso			mais de o que nunca
junto a			um a um

Tabela 5.6: Quadro com lista de MWEs identificadas por apenas um dos métodos

a a medida em que	além de isto	em a verdade	por conseguinte
a a medida que	além de o mais	em direção a	por exemplo
a a toa	além de o que	em face de	por exemplos
a as vezes	além de o quê	em função de	por fim
a cargo de	aos poucos	em geral	por mais que
a despeito de	apesar de	em o caso de	por meio de
a exemplo de	apesar de que	em o entanto	por muito que
a favor de	assim como	em o tocante a	por o menos
a fim de	assim por diante	em razão de	por parte de
a fim de que	assim que	em relação a	por pouco que
a longo de	assim sendo	em relação as	por sua vez
a medida que	até que	em relação á	por vezes
a menos que	bem como	em seguida	por volta de
a não ser que	cada vez mais	em separado	pouco a pouco
a o certo	caso contrário	em termos de	quanto a
a o contrário de	cerca de	em torno de	quanto mais
a o invés de	com base em	em vez de	se bem que
a o largo de	com isso	em vão	sem mais nem menos
a o longo de	com relação a	enquanto que	sem que
a o menos	com vistas a	isto é	sempre que
a o menos que	como relação a	junto a	sendo assim
a o passo que	como também	já que	sendo que
a o ponto de que	de acordo com	logo que	tais como
a o que	de acordos com	mais de o que nunca	tal como
a o todo	de agora em diante	mesmo assim	tanto quanto
a o vivo	de aí	mesmo que	tanto que
a partir de	de forma a	nada que	toda vez que
a ponto de	de forma que	nem a o menos	tudo quanto
a principio	de maneira a	nem mesmo	um a um
a princípio	de maneira que	nem sequer	um por um
a priori	de modo a	não obstante	um pouco
a respeito de	de modo que	não que	uma vez em
a seguir	de o que	não só	uma vez que
a título de	de sorte que	ou seja	visto que
ainda mais que	de tal forma que	para que	volta e meia
ainda que	desde que	para tal	é que
além de	devido a	pelo menos	
além de isso	em a faixa de	por causa de	

Tabela 5.7: Quadro com lista final de MWEs obtidas por todos os métodos

	MWE	**Lemas de MWEs**	**Novas MWEs**	**MWEs removidas**	**Tokens modificados**
v2	2.356	148	549	438	5.632
v3	2.467	112			

Tabela 5.8: Número de MWEs revistas no corpus

o número de novas expressões anotadas como MWE (549) e o número de expressões cuja anotação de MWE foi desfeita (438), totalizando, na verdade, uma diferença na anotação de 987 construções do corpus.

O número de lemas de expressões multipalavras encontradas no PetroGold nesta nova versão é menor que na versão anterior – como vimos, nesta versão encontramos 112 MWEs, enquanto que na versão anterior esse número era de 148. Outro dado importante é que, como já havíamos notado, o número de tokens modificados no corpus para realizar a revisão de MWEs é muito maior que o número de MWEs corrigidas em si – no total, foram 5.632 tokens modificados, o que corresponde a uma média de 5,7 tokens modificados por MWE que teve sua anotação modificada nessa versão, o que inclui não apenas todas as palavras da MWE, mas também as vírgulas, núcleos de sintagmas, verbos auxiliares etc. envolvidos nas orações. A tabela 5.8 resume os números apresentados.

No apêndice C encontra-se uma lista com todas as expressões que ora foram anotadas como MWE, ora de forma transparente, pois apesar de terem as mesmas palavras, se comportam de forma diferente nas frases.

5.2.3 O pronome "se"

"Se" é uma palavra que pode assumir diferentes papéis no português. A primeira grande distinção é a da classe gramatical – o termo "se" pode ser conjunção subordinativa, funcionando como conectivo para orações subordinadas substantivas objetivas diretas ou orações subordinadas adverbiais condicionais, assim como pode ser classificado também como pronome, servindo a uma série de funções distintas, cada uma carregando informações semânticas e sintáticas diferentes.

A distinção entre pronome e conjunção subordinativa para a palavra "se" já havia sido satisfatoriamente realizada no corpus Bosque-UD, que alimentou o modelo utilizado para anotar o PetroGold. Contudo, o Bosque-UD carece de subespecificação para os diferentes tipos de "se" pronominais, motivo pelo qual nos debruçamos sobre a palavra durante a construção do PetroGold.

A revisão dos pronomes "se" é motivada, por um lado, pela inserção

recente das classes “expl:impers”, “expl:pass” e “expl:pv” nas diretivas do projeto UD, não fazendo parte da anotação inicial do corpus PetroGold nas suas versões 1 e 2. Por outro lado, a sua revisão é importante pois a distinção entre o “se” que funciona como índice de indeterminação do sujeito, partícula integrante do verbo e pronome apassivador auxilia no processamento automático do português, o que será avaliado no capítulo de avaliação do PetroGold mas que já era levantado por Duran et al. (2013), onde os autores destacam benefícios para a análise sintática e a atribuição de papeis semânticos (SRL), além da tradução automática.

Além disso, como mostramos em Freitas e de Souza (2021), até 46% das orações em português, considerando corpora de diferentes gêneros textuais, não têm sujeito explícito, e a reconstituição de sujeitos ocultos pode melhorar o resultado do aprendizado automático de dependências em até 2%. Contudo, um dos entraves para essa reconstituição do sujeito na época era o fato de que os diferentes usos do pronome “se” não eram indicados na anotação – faltava especificar quando o pronome “se” é utilizado para indeterminar o sujeito e quando é utilizado acompanhando um verbo que tem sujeito, estando ele explícito ou implícito na oração. Quando se sabe que esse pronome indica voz passiva ou verbo pronominal, sabe-se também que há um sujeito que pode ser reconstituído quando está implícito na oração; quando o pronome marca indeterminação do sujeito, pelo contrário, não será possível reconstruí-lo.

A classe “expl” em UD é destinada às palavras expletivas ou pleonásticas que ocupam um lugar de argumento na oração mas não satisfazem nenhum papel semântico do predicado[23]. Nessa classe estão incluídos os pronomes reflexivos, que podem ser subcategorizados em expl:pv, para os reflexivos que se ligam a verbos inerentemente reflexivos, em expl:pass, para os reflexivos que se ligam a verbos transitivos e que funcionam como marcadores de voz, e em expl:impers, para o uso impessoal[24].

Além das definições para os expletivos em UD, utilizaremos gramáticas da língua portuguesa para nos ajudar a definir os critérios que diferenciam os fenômenos linguísticos em que aparece o pronome “se”. Nosso horizonte é (1) a obtenção de consenso, na medida do possível, relativo à anotação do pronome em língua portuguesa; (2) a obtenção de consistência interna, garantindo que

[23] “This relation captures expletive or pleonastic nominals. These are nominals that appear in an argument position of a predicate but which do not themselves satisfy any of the semantic roles of the predicate.” – <https://universaldependencies.org/u/dep/expl.html>. Acesso em 31 de out. 2022

[24] No original, “expl:pv for reflexive clitics attached to inherently reflexive verbs (also called pronominal verbs in some grammars); expl:pass for reflexive clitics attached to transitive verbs and acting as a voice marker (passive or mediopassive); expl:impers for impersonal usage (works also with intransitive verbs)” – <https://universaldependencies.org/u/dep/expl.html>. Acesso em 31 de out. 2022

frases parecidas terão a mesma anotação segundo os critérios estabelecidos; e (3) a facilitação da reconstituição de sujeitos ocultos, permitindo distinguir frases em que a ausência de um sujeito ocorre por causa de elipse ou porque o sujeito é de fato indeterminado.

Na seção de metodologia, explicamos os critérios que utilizamos para classificar o "se" nas orações do PetroGold e apresentamos o passo a passo que resultou na anotação de todas as frases do corpus em que o pronome aparece. Na seção de resultados e análise, mostramos um quantitativo para cada uma das novas classes, listamos os verbos que aparecem associados a um ou mais de um dos tipos de "se", e analisamos os casos difíceis e de ambiguidade.

5.2.3.1 Metodologia

Como classificar o "se"

Utilizaremos as classes do projeto UD tal como já estão designadas nas diretivas, embora não tenham sido extensamente documentadas – por exemplo, não há indicação de como diferenciar o "se" da voz passiva do "se" usado para indeterminação do sujeito nas diretivas do projeto –, motivo pelo qual usaremos como parâmetro também uma adaptação da GT, compilada em Bechara (2012) e em Bagno (2012).

Neste trabalho, não questionaremos a escolha da classe dos expletivos (expl) para os diferentes tipos de "se" pronominais. Uma possível crítica, por exemplo, segundo Bouma et al. (2018), é a dificuldade de afirmar que alguns dos usos dessas palavras consideradas expletivas de fato seguem o requisito de não conterem nenhum valor semântico[25], nos limitando a aplicar as classes que já estão previstas no projeto.

Além disso, embora a palavra "impers" da classe "expl:impers" diga respeito à impessoalização (impersonal, em inglês), entendemos, junto com Bagno (2012), que tanto a indeterminação do sujeito quanto a voz passiva sintética são utilizadas como estratégias de impessoalização da oração. No caso da indeterminação do sujeito (expl:impers), a impessoalização ocorre pela supressão do sujeito, que não pode ser recuperado na oração, enquanto no caso da voz passiva sintética (expl:pass) a impessoalização ocorre pelo deslocamento do objeto direto para a posição de sujeito paciente, sendo que o agente também não é recuperável dentro da oração. A diferenciação entre ambas as classes,

[25] A gramática do anotador automático PALAVRAS, por exemplo, anota o "se" em orações de sujeito indeterminado como sujeito, atribuindo um valor a essas palavras que, portanto, não seriam expletivas (BICK, 2000).

portanto, não se dá pela noção de "impessoalização", mas por outros critérios que serão explicados a seguir.

1. expl:impers – índice de indeterminação do sujeito

A primeira das novas classes do "se" expletivo é a "expl:impers", quando a palavra funciona como um índice de indeterminação do sujeito. Em todos os casos de "se" expletivos, a sua anotação será feita com a classe gramatical de pronome – PRON – e sem informações flexionais, isto é, a anotação de características morfológicas será preenchida por underline, conforme o modelo UD.

Um dos requisitos para a anotação de indeterminação do sujeito é a ausência de um sujeito sintático para o verbo a que o "se" está associado. Em frases com indeterminação do sujeito, não se pode pensar que o sujeito esteja implícito seja por elipse ou qualquer outra estratégia, pois um sujeito simplesmente não cabe na oração. Quando se puder inserir um sujeito para o verbo, provavelmente o fenômeno será de oração passiva sintética ou verbo pronominal, fenômenos que veremos a seguir.

Comumente, o verbo cujo sujeito está indeterminado é intransitivo ou transitivo indireto, pois caso tivesse algum objeto direto, este provavelmente se tornaria um sujeito paciente, como veremos na anotação da voz passiva sintética.

Na frase 81, o verbo, "chega", não tem um sujeito sintático, e tampouco seu objeto, "expressão", poderia ocupar esse lugar. Nesse caso, embora o autor da frase não tenha feito a contração do determinante "a" com a preposição "a" (crase) que antecede "expressão", por conta da transitividade do verbo sabemos que o "a" funciona como preposição, sendo o objeto de tipo indireto, preenchendo, portanto, os requisitos para uma oração de sujeito indeterminado.

81. Considerando se que o ciclo de o motor é realizado a cada duas rotações completas de o eixo de manivelas , **chega se** a expressão de a massa de combustível (em kg) utilizada em cada ciclo : (3.1)

2. expl:pass – pronome apassivador

A segunda das classes para o "se" expletivo é a que indica que ele funciona como um pronome apassivador, marcando voz passiva sintética, quando o agente da passiva é suprimido da oração, dando lugar a um sujeito que é paciente da ação do verbo. Porque o agente da passiva foi suprimido, tanto

aqui como na oração com indeterminação do sujeito há impessoalização – não se sabe quem é o agente da ação do verbo ou não se quer revelá-lo, utilizando como estratégia a indeterminação do sujeito ou a voz passiva. Contudo, diferentemente da oração onde há indeterminação do sujeito, nesta há um sujeito sintático, anotado como "nsubj:pass" – sujeito paciente. Esse sujeito, quando está implícito, deve poder ser recuperado na frase ou no texto, diferentemente da oração com indeterminação do sujeito.

É condição para a ocorrência de voz passiva sintética um verbo com transitividade direta ou transitividade direta e indireta segundo a GT. Em outras palavras, o requisito é um objeto direto, o qual, na transformação para voz passiva sintética, será anotado como sujeito do ponto de vista sintático, embora seja paciente do ponto de vista semântico.

Na frase 82, ambos os verbos em negrito são núcleo de um "se" que é "expl:pass", pois tanto "procedimento" quanto "argilas" podem ocupar o lugar de sujeito sintático das orações embora sejam semanticamente pacientes do conteúdo verbal – entende-se que a/as pessoa/as que "seguiram o procedimento" e "secaram as argilas" foram propositalmente suprimidas da frase, como estratégia de impessoalização, utilizando a voz passiva sintética, portanto colocando os objetos na posição de sujeito paciente.

82. Para a análise de as argilas estudadas , **seguiu se** o seguinte procedimento ; **secou se** as argilas em 39 uma estufa a 80ºC durante 18 horas , e após terem sido retiradas de a estufa , foram moídas em um moinho de bolas durante 18 horas.

O fato de que "argilas", sendo sujeito da oração, não concorda em número com "secou", não é decisivo para definir se houve indeterminação do sujeito ou voz passiva sintética, pois entendemos, assim como Bagno (2012), que em ambos os casos o objetivo é impessoalizar a oração, de tal maneira que já se tornou usual conjugar o verbo na terceira pessoa do singular, independentemente de o fenômeno empregado ser a voz passiva sintética. Assim, embora segundo a GT tenha ocorrido um erro de concordância verbal, ele é explicado pela intenção do autor, que não distingue entre um sujeito paciente e uma oração de sujeito indeterminado na hora de impessoalizá-la.

Embora os critérios sintáticos sejam importantes na distinção entre as classes do "se" – ausência ou não de um sujeito sintático, utilização de um verbo VTD, VTDI, VI ou VTI –, não se pode abrir mão de interpretar as orações individualmente, pois um mesmo termo – como "procedimento", no exemplo anterior – poderia ser considerado sujeito paciente ([alguém]sujeito seguiu o procedimento), a leitura que de fato fazemos, ou não ([o procedimento]sujeito

seguiu/ocorreu), se em um contexto diferente, dando origem a diferentes anotações para o "se" e para o sujeito em si – no primeiro caso, "expl:pass" e "nsubj:pass", e no segundo caso, "expl:pv" e "nsubj", fenômeno que veremos a seguir.

Orações com o verbo no gerúndio também são anotadas como casos de voz passiva quando o verbo é transitivo direto. No exemplo da frase 83, "afinidade" é sujeito paciente – "nsubj:pass" – da oração cujo núcleo é "analisou" (o "se" recebe a anotação de "expl:pass"), assim como "solução" também é sujeito paciente de "utilizando", um VTD.

83. **Analisou se** a afinidade de a argila A3 por o aço , **utilizando se** solução de NaOH pH= 10 (S1) as variáveis de entrada estão descritas em a Tabela III.2 e III.3.

No exemplo da frase 84, o verbo "considerando", que é transitivo direto, não tem um sujeito paciente, pois o seu sujeito seria uma outra oração, cujo núcleo é "realizado", mas não há previsão para sujeito paciente oracional no modelo UD. Nesse caso, a anotação de voz passiva sintética se mantém, com a palavra "se" recebendo a etiqueta "expl:pass", mas o sujeito paciente, cujo núcleo é "realizado", está sendo anotado como oração subordinada substantiva objetiva direta – "ccomp".

84. **Considerando se** que o ciclo de o motor é realizado a cada duas rotações completas de o eixo de manivelas , chega se a expressão de a massa de combustível (em kg) utilizada em cada ciclo : (3.1)

Por fim, no exemplo 85, o sujeito paciente está implícito. No modelo de dependências, "fator" é a raiz da sentença e "considerar" é núcleo de uma oração que adjetiva "fator", portanto sendo anotada como "acl" dependente do substantivo. Como o verbo está no infinitivo, não tem sujeito sintático, contudo, a leitura da frase nos permite identificar que "fator" é, semanticamente, sujeito paciente do verbo – o agente da ação de "considerar" foi suprimido propositalmente como uma estratégia de impessoalização e, no lugar, o "fator", seu objeto, assumiria a posição de sujeito paciente da oração, que está na voz passiva sintética. Assim, por uma restrição do modelo de dependências, o sujeito de "considerar" não está explícito na anotação, mas a anotação do "se" ainda é a de "expl:pass".

85. Consequentemente, a composição é o principal fator a se **considerar** para o controle das suas propriedades.

3. expl:pv – verbo pronominal

O terceiro "se" expletivo indica o emprego de um verbo pronominal e recebe o nome, na GT, de partícula integrante do verbo. Essa é uma análise que, para nós, assim como as demais categorias, depende de interpretação, não sendo característica intrínseca de nenhum verbo em especial exigir ou não o pronome "se". O verbo está sendo usado de forma pronominal quando há um sujeito sintático e ele não é paciente da ação verbal[26]

No exemplo 86, "refere" é núcleo de uma oração subordinada adjetiva explicativa cujo sujeito está representado pelo pronome relativo "que", o qual retoma "viscosidade". Semanticamente, "viscosidade" não é sujeito paciente da ação do verbo "referir" e, na verdade, no contexto apresentado a construção "refere-se" é equivalente a um verbo de ligação (a viscosidade [é] a resistência...). Encaramos que o verbo está sendo considerado pronominal, sendo a palavra "se" anotada com a relação sintática "expl:pv" e o seu sujeito, a palavra "que", simplesmente "nsubj".

86. Dentre os parâmetros reológicos mais usuais , destaca se a viscosidade , que **se refere** a a resistência que uma substância apresenta a o fluxo , e , em o campo , citam se como principais propriedades reológicas de interesse o índice de comportamento , o índice de consistência , a viscosidade aparente , a viscosidade plástica , o limite de escoamento e a força gel (THOMAS et al , 2001).

No exemplo 87, a oração cujo núcleo é o verbo "acumularam" tem como sujeito sintático o substantivo "sedimentos". Não se pode dizer que a frase está na voz passiva sintética pois a sua leitura não nos permite inferir que haja um agente sendo propositalmente omitido, como estratégia de impessoalização – os sedimentos (sujeito) acumularam-se (verbo intransitivo). Diferentemente da segunda frase de exemplo (frase 88), hipotética, onde o mesmo verbo, em negrito, não está sendo usado pronominalmente, mas como voz passiva sintética, pois omitiu-se o agente da oração estrategicamente e as "dívidas" são sujeito paciente da ação de acumular.

[26]Note que, em alguns casos, o sujeito não é nem paciente e nem agente, quando um verbo causativo tornou-se incoativo pelo uso do pronome -se, como na frase "O esporte popularizou-se", levantada em Duran et al. (2013). Nela, o verbo, originalmente transitivo direto, está sendo empregado no aspecto incoativo – "o esporte ficou popular" – indicando uma mudança de estado e, por isso, "esporte" não é nem agente nem paciente de um verbo de ação, mas "sede" da mudança de estado indicada pelo verbo (de estado), termo empregado por Cunha e Cintra (2016). Frases do tipo foram anotadas como pronominais e, embora seja difícil assegurar que duas pessoas lerão as mesmas frases como tendo verbos no aspecto incoativo, no corpus essa anotação do "-se" foi realizada apenas por mim, garantindo consistência interna. Mais discussão sobre esse tipo de construção pode ser encontrada em Cançado e Amaral (2010).

87. ***expl:pv:*** A fase rifte em a região de Santos ocorreu entre o Neocomiano e o Eoaptiano , quando **acumularam se** os sedimentos lacustres de a Formação Guaratiba , escassamente amostrados em a bacia.

88. ***expl:pass:*** **Acumularam-se** muitas dívidas com os fornecedores.

Em Duran et al. (2013), os autores têm a preocupação de construir uma lista de verbos pronominais para auxiliar no processamento automático do português, utilizando a terminologia "verbos inerentemente reflexivos" para se referir a eles. Nesse contexto, faria sentido enquadrar o lema do verbo seguido da partícula "se", como uma única palavra. Aqui, contudo, não pretendemos lidar com a noção de inerência reflexiva do verbo, pois entendemos que potencialmente qualquer verbo, conforme o contexto, pode estar sendo empregado pronominalmente – nos termos da GT, são portanto verbos pronominais acidentais –, não havendo, assim, nenhuma forma de diferenciar aqueles verbos que só se conjugam pronominalmente daqueles que foram empregados pronominalmente para satisfazer um contexto específico.

4. obj / iobj – objeto direto ou indireto

A palavra "se" pode ainda ser anotada como objeto direto ou indireto quando a ação do verbo se estende à terceira pessoa do singular ou plural, na forma do pronome "se". Nesse caso, o "se" recebe as informações morfológicas de pronome pessoal oblíquo, na terceira pessoa do singular ou plural.

Nesses casos, o sujeito sintático também é agente e paciente da ação ao mesmo tempo, de modo que o verbo precisa ter um objeto, direto ou indireto. O fenômeno recebe o nome de pronome reflexivo na GT.

Bechara (2012) indica que o pronome "se" como objeto exige um sujeito animado para o verbo, pois somente dessa forma o sujeito será agente e paciente da ação ao mesmo tempo. Assim, o autor exemplifica o pronome com as frases a seguir.

89. João se banha.

90. João e Maria se amam.

91. João e Maria se escrevem.

92. João e Maria se gostam.

93. Ele se barbeou.

94. Eles se cumprimentaram.

95. Ela se arroga essa liberdade.

Em todas elas, o sujeito é um ser animado e funciona como agente e paciente da ação do verbo. Já na frase "O banco só se abre às 10 horas", para o autor, o sujeito inanimado impede a ocorrência de pronome reflexivo, sendo um caso de voz passiva. Já no caso de "Ele se chama João", sabe-se que, embora animado, o sujeito não é agente da ação, restando a anotação de verbo pronominal.

5. mark – conjunção subordinativa

Por fim, a palavra "se" pode ainda funcionar como uma conjunção subordinativa. Embora não seja pronome, deixamos claros os critérios para anotação da conjunção subordinativa "se" de modo a evitar quaisquer dúvidas.

Nesses casos, o "se" receberá a etiqueta de classe gramatical SCONJ, não terá informações flexionais, e a etiqueta de relação sintática será "mark". A palavra é dependente do núcleo da oração subordinada, seja ele um verbo ou não, no caso de orações com verbo de ligação, que não são considerados núcleo em UD. A diferenciação entre o uso do "se" como conjunção subordinativa de uma oração subordinada adverbial ou substantiva objetiva direta se dá pela relação sintática do verbo, que será "advcl" para a oração adverbial ou "ccomp" para a oração substantiva.

No primeiro exemplo, o "se" funciona como conjunção para uma oração subordinada adverbial condicional, cujo núcleo é "submetermos" – corretamente conjugado no subjuntivo, inclusive – o qual, por sua vez, se conecta à oração principal, cuja raiz é "decrescerá", como uma "advcl" (oração adverbial).

Na segunda frase, o "se" conecta uma oração subordinada substantiva objetiva direta, cujo núcleo não é um verbo – "viável" – a uma oração subordinada adverbial, cujo núcleo é "verificar". A relação de um verbo para o outro é de "ccomp", mas em ambos os casos, na primeira e na segunda frase, o "se" recebeu a mesma anotação sintática para conjunção subordinativa, "mark", associando-se ao núcleo da oração adverbial/objetiva.

96. **Se** , a seguir , **submetermos** o sistema a cisalhamento , a uma velocidade de agitação constante , a viscosidade aparente decrescerá com o tempo até atingirmos o equilíbrio entre a quebra e a reconstrução de a estrutura organizada de o fluido.

97. Para verificar **se** o projeto é **viável** economicamente , foi elaborado dois gráficos com as receitas e as despesas de o projeto.

Para realizar a anotação dos tipos de "se", precisamos reunir todas as frases em que o pronome ocorre e as pré-agrupamos seguindo alguns critérios. No final, contudo, a revisão foi realizada interpretando frase a frase, por meio de uma estratégia que será explicada a seguir.

Estratégia de revisão

A revisão do "se" foi realizada em 6 etapas, descritas a seguir:

(1) Primeiro, foram corrigidos todos os casos de "se" que não tinham relação nem de "expl" nem de "mark". Por exemplo, havia no corpus alguns casos de "sudeste" abreviados como SE e erroneamente lematizados como "se". Ora foi necessário corrigir o lema, ora a relação de dependência desses tokens, como na frase 98, onde se observa um "SE" com relação de "obl", o que é correto, mas cujo lema deve ser também em letras maiúsculas, para não confundi-lo com os demais "se".

98. ***obl:*** Pode se notar que **a SE** ocorre o maior espessamento sedimentar (depocentro controlado por a falha de a borda sul) que , também , de a mesma forma que em o perfil sísmico , se acunha para o norte em direção a o gráben interno definido por os referidos autores.

(2) Em segundo lugar, foram revisados todos os casos de "se" com relação "mark", garantindo que eram mesmo casos de conjunção subordinativa e estavam anotados com POS SCONJ e sem informação flexional. Verifiquei também se, nos casos de subordinação, os verbos estavam corretamente anotados como oração subordinada adverbial condicional (advcl) ou oração subordinada objetiva (ccomp).

(3) Classifiquei todos os lemas verbais associados a um pronome "se" com frequência maior ou igual a 3 nas categorias impers, pass e pv. Para facilitar a geração dessa lista de lemas verbais, adicionei à coluna misc dos verbos de que dependem os pronomes "se" a etiqueta "Se=Yes". Assim, pude pedir a distribuição desses lemas verbais que se associam a "se" somente utilizando essa informação do misc[27]. O objetivo dessa classificação dos verbos distribuídos por lema é realizar um filtro bruto dos verbos que, com o pronome, provavelmente deverão ser colocados em uma das três classes, visando guiar as correções caso a caso que serão feitas a seguir.

(4) Realizei quatro buscas: uma para todas as frases com verbos que classifiquei como impers (245 frases), outra para as frases que classifiquei como pass (995 frases), outra pv (392 frases), e outra para todos os verbos que não

[27]No Interrogatório, a expressão de busca por esses lemas verbais é: *misc = ".*Se=Yes.*"*

classifiquei porque tinham frequência menor que 3 (259 frases). Li todas as frases, de todas as buscas, para assegurar que a classificação inicial estava correta, selecionando os casos divergentes para posterior correção manual – essa classificação inicial, a partir dos lemas, é descontextualizada, sendo necessário uma segunda revisão tendo interpretado todas as frases pois os verbos, como vimos, são empregados de formas diferentes nas frases. No exemplo 99, "se dá" é verbo pronominal, pois tem sujeito (água) e ele não é paciente da ação verbal, enquanto que na frase 100, "dá-se" é voz passiva sintética, uma vez que "nome" é sujeito paciente da ação do verbo, executada por um agente que foi omitido da frase.

99. ***pv:*** Veil (2004) explica que , a origem de a água produzida **se dá** em as formações subterrâneas produtoras , que geralmente são permeadas por diferentes fluidos , tais como óleo , gás e água com alto teor de salinidade.

100. ***pass:*** À esta propriedade **dá se** o nome de tixotropia , a qual é definida como um decréscimo contínuo de a viscosidade com o tempo quando um escoamento é aplicado a uma amostra que tenha estado previamente em repouso e a subsequente recuperação de a viscosidade em o tempo quando o escoamento é descontinuado.

(5) Então, por meio de um script na ET, atribui aos pronomes "se" a anotação "expl:impers,pv,pass" de acordo com o resultado dessa análise das frases. Na coluna misc dos verbos, substitui "Se=Yes" por "Se=impers,pass,pv" para facilitar explorações futuras. Além disso, removi qualquer informação morfológica do "se" (coluna feats).

(6) Por fim, confirmei que: (a) verbos com misc "Se=impers" não têm sujeito; (b) verbos com misc "Se=pv" ou têm sujeito explícito ou implícito; e (c) verbos com misc "Se=pass" tiveram o objeto direto convertido em sujeito paciente (nsubj:pass), exceto nos casos de oração ("Observa-se que [VERB]"), onde a relação do verbo subordinado à oração na voz passiva permanece como ccomp, como discutido. Essa conversão dos sujeitos pacientes foi importante pois a anotação original do corpus, proveniente do Stanza (QI et al., 2020), entendia esses verbos como em voz ativa seguidos de objeto direto, e não como voz passiva sintética com sujeitos pacientes. Assim, nos exemplos 101 e 102, o termo em negrito passou de objeto direto a sujeito paciente (nsubj:pass).

101. Dentre suas propriedades, *destacam-se* a **formação** de bolhas e espumas na superfície de um líquido e a adsorção nas superfícies ou interfaces líquido-líquido, líquido-gás e sólido-líquido, promovendo redução significativa da tensão superficial ou interfacial.

deprel	**mark**	**expl:impers**	**expl:pass**	**expl:pv**
freq.	75	278	807	800

Tabela 5.9: Frequência dos tipos de "se"

102. Somente ao misturar as duas fases é que *se adiciona* o **agente** modificador.

5.2.3.2 Resultados

O corpus PetroGold conta com 1960 ocorrências de "se", sendo 75 conjunções subordinativas e 1.885 pronomes expletivos, conforme tabela 5.9. Todos os 1.885 usos expletivos passaram por revisão, uma vez que precisaram ter sua etiqueta modificada para se acrescentar a informação relativa à indeterminação do sujeito, passivização ou uso pronominal do verbo.

Distribuímos os pronomes "se" expletivos pelos verbos aos quais se associam, e alguns dos verbos podem aparecer associados a mais de um tipo de pronome "se" expletivo, como já discutido. A relação de lemas verbais para cada um dos tipos de pronome "se" pode ser encontrada na tabela 5.10, onde percebemos uma proporção muito menor de verbos possíveis de serem empregados de forma a indeterminar o sujeito.

deprel do "se"	**nº de verbos**	**lemas verbais**
expl:impers	21	acreditar, aliviar, atender, ater, chamar, chegar, dever, entender, exigir, falar, necessitar, optar, partir, passar, poder, proceder, prosseguir, ter, trabalhar, voltar

expl:pass	154	abrir, acionar, acrescentar, adelgaçar, adicionar, admitir, adotar, agitar, aguardar, ajustar, alcançar, alterar, alçar, amplificar, analisar, aplicar, apresentar, apurar, armazenar, assumir, atingir, atribuir, aumentar, buscar, calcular, capturar, caracterizar, citar, classificar, coletar, colocar, combinar, comparar, comprovar, concluir, conduzir, conectar, confundir, conhecer, conseguir, considerar, constatar, construir, correlacionar, criar, dar, definir, deixar, delimitar, delinear, depreender, descrever, desejar, desenvolver, destacar, determinar, diluir, diminuir, dividir, dizer, efetivar, efetuar, elaborar, empregar, encaixar, encontrar, escolher, esperar, esquematizar, estabelecer, estimar, evacuar, evitar, expandir, explorar, extrair, fazer, fechar, filtrar, fixar, imaginar, impor, incluir, incorporar, inferir, inibir, iniciar, injetar, instalar, intensificar, interpretar, introduzir, inverter, julgar, levar, ligar, limitar, manter, medir, melhorar, minimizar, misturar, monitorar, mostrar, multiplicar, notar, objetivar, observar, obter, otimizar, pensar, perceber, perfurar, permitir, pesar, plotar, ponderar, preparar, presumir, pretender, procurar, produzir, projetar, propor, reajustar, realizar, recolher, recomendar, reconhecer, reduzir, remover, reportar, ressaltar, retirar, reutilizar, saber, secar, seguir, somar, substituir, subtrair, sugerir, supor, tentar, tirar, titular, tomar, usar, utilizar, variar, varrer, ver, verificar, zerar

expl:pv	142	acentuar, acumular, acunhar, adaptar, adentrar, adsorver, afastar, aglomerar, aglutinar, ajustar, alargar, aliar, alicerçar, alimentar, alinhar, alojar, alongar, anexar, apoiar, apresentar, aprisionar, aprofundar, aproveitar, aproximar, assemelhar, assentar, associar, atenuar, basear, beneficiar, caracterizar, chamar, completar, comportar, concentrar, condensar, consolidar, constituir, correlacionar, curvar, dar, decompor, dedicar, demonstrar, denominar, depositar, descrever, desenvolver, deslocar, desprender, destinar, deteriorar, dever, diferenciar, diferir, difundir, direcionar, dispersar, dissipar, dissociar, dissolver, distinguir, distribuir, dividir, elevar, encerrar, encontrar, enquadrar, equivaler, espalhar, espessar, estabelecer, estabilizar, estender, estreitar, expressar, fazer, fender, formar, fundamentar, fundir, gastar, hidratar, horizontalizar, individualizar, informar, iniciar, inserir, instabilizar, intemperar, interdigitar, interditar, introduzir, ionizar, juntar, justificar, ligar, limitar, liquefazer, localizar, manifestar, manter, misturar, mostrar, mover, movimentar, ocorrer, organizar, orientar, originar, paralelizar, pautar, perder, posicionar, prender, prestar, projetar, prolongar, propagar, propor, reduzir, referir, relacionar, repetir, reproduzir, restringir, resumir, reunir, revelar, romper, separar, situar, sobrepor, sobressair, somar, subdividir, tornar, traduzir, transformar, tratar, unir, vincular
duas categorias	25	ajustar, apresentar, caracterizar, chamar, correlacionar, dar, descrever, desenvolver, dever, dividir, encontrar, estabelecer, fazer, iniciar, introduzir, ligar, limitar, manter, misturar, mostrar, projetar, propor, reduzir, somar, tratar

três categorias	0	–

Tabela 5.10: Quadro com lista de verbos que se associam aos tipos de pronome "se"

Os resultados da anotação do "se", além de poderem ser encontrados no PetroGold v3, serão avaliados computacionalmente no próximo capítulo. Além disso, no apêndice se encontra uma lista com todos os verbos aos quais se associam a palavra "se", categorizados por tipo de construção (indeterminação do sujeito, voz passiva sintética ou verbo pronominal), frequência e exemplo de frase do corpus.

No PetroGold não foram encontradas ocorrências de "se" como pronome reflexivo (objeto direto ou indireto). Para confirmar que a inexistência do pronome reflexivo no PetroGold está correta, verificamos todos os sujeitos de verbos a que se associam o pronome "se"[28]. A análise dos 281 lemas não retornou nenhum sujeito animado, o que justificaria a ausência do pronome reflexivo no corpus, sugerindo ser uma característica dos textos do domínio a frequência baixa ou nula de frases em que o sujeito é animado, um dos requisitos elencados por Bechara (2012).

Como se vê na tabela, 22 verbos podem se associar a pronomes "se" de dois tipos diferentes – 20 (91%) ora são usados pronominalmente e ora na voz passiva sintética, enquanto somente 2 (9%) são usados ora como verbo pronominal, ora como oração de sujeito indeterminado. Nenhum dos verbos se associa a três categorias ao mesmo tempo, e não encontramos no corpus nenhum verbo que se associe ora a um pronome "se" que indique voz passiva sintética, ora índice de indeterminação do sujeito.

O maior número de verbos que podem ser utilizados tanto na voz passiva sintética como na forma pronominal já era esperado – Azeredo (2000), por exemplo, comenta sobre o fenômeno da cristalização do "se" em verbos de voz passiva, os quais, pela frequência de uso, vão se tornando pronominais. Sintaticamente, o fato de que muitos verbos podem ser empregados das duas formas explica-se pela semelhança estrutural – ambas requerem um sujeito sintático na frase, sendo que a diferenciação é realizada semanticamente ao interpretar se o sujeito é paciente do conteúdo verbal ou não. Por exemplo, nas frases 103 e 104, o verbo em destaque é "ajustar". O primeiro, porém, tem como sujeito o substantivo "modelos", sendo que a minha interpretação da frase não permite inferir que haveria um agente, propositalmente omitido da oração, responsável por ter ajustado os modelos no contexto em que o verbo foi

[28]A busca pode ser realizada na ET utilizando a expressão: deprel = "nsubj" and head_token.misc = ".*Se=.*"

utilizado, diferentemente da segunda oração, onde um agente não identificado ajustou a "frequência", que é sujeito paciente da oração.

103. ***expl:pv:*** A partir do coeficiente de correlação, percebe-se que todos os modelos **se ajustaram**.

104. ***expl:pass:*** A bomba de água foi acionada com uma freqüência de 30 Hz e então **ajustou-se** a freqüência baseando-se na vazão de água desejada.

Já a diferenciação entre orações com sujeito indeterminado e uso pronominal do verbo pode ser explicada em termos puramente sintáticos. Nos exemplos 105 e 106, o verbo em destaque é "chamar". Na primeira frase, o pronome relativo "que" retoma "pasta oleosa", sujeito da oração relativa cujo núcleo é "se chama" e cujo complemento verbal é "petróleo". Por tratar-se de uma oração com sujeito sintático, a anotação é a de verbo pronominal, diferentemente da segunda frase, em que "bentos" é objeto direto (sem a possibilidade da leitura como sujeito paciente) e "organismos" objeto indireto, faltando um sujeito sintático para a oração (que tampouco está elíptico), marcando, portanto, indeterminação do sujeito.

105. ***expl:pv:*** Admite-se que o petróleo foi formado há milhões de anos pelo acumulo de diferentes seres vivos como a decomposição de plânctons - seres que são geralmente encontrados na zona costeira, mares, oceanos e estuários - esses seres teriam se acumulados no fundo dos mares, rios e lagos e soterrados pela ação do movimento da crosta terrestre e posteriormente com o passar dos anos transformando-se em uma pasta oleosa que hoje **se chama** petróleo (VAZ, 2011).

106. ***expl:impers:*** *Em biologia marinha e limnologia, **chama-se** bentos aos organismos que vivem no substrato, fixos ou não, em contraposição com os pelágicos, que vivem livremente na coluna de água.

A ausência de verbos compartilhando o -se na voz passiva sintética e o -se como indeterminação do sujeito pode ser explicada pelo fato de que os dois fenômenos são muito distintos sintaticamente, sendo que no primeiro há um sujeito sintático na oração, que é marcada por um VTD ou VTDI, e no segundo fenômeno não há sujeito, sendo utilizado um verbo VI ou VTI.

Cabe notar, por fim, que o verbo "tratar-se" recebeu tratamento especial no corpus. É um verbo ao mesmo tempo utilizado para impessoalizar a oração e é sempre conjugado pronominalmente. Quando o verbo é empregado corretamente (segundo a GT), a oração tem sujeito indeterminado (como no

exemplo 107), contudo, encontramos exemplos no corpus em que o verbo é utilizado com sujeito (na frase 108, o sujeito tem "reservatório" como núcleo). Optamos, especificamente para esse verbo, por atribuir a etiqueta expl:pv ao -se, relativa a verbo pronominal, para ambos os casos – mesmo que o verbo tenha sujeito indeterminado na maioria das vezes. Nosso objetivo é não destoar do consenso gramatical já estabelecido de que o verbo é do tipo pronominal e possibilitar o alinhamento entre o nosso e outros treebanks além dos léxicos e gramáticas do português, a despeito dos critérios que desenvolvemos para a anotação do -se.

107. ***expl:pv:*** Como **se trata** de uma faixa estreita , se encontra em águas rasas de a plataforma continental.

108. ***expl:pv:*** O reservatório de Marlim **se trata** de um reservatório turbidítico da formação Carapebus, formado no período Paleógeno, de idade oligocênica, a uma profundidade de 2.631 metros.

Os apêndices F e G trazem exemplos de todos os lemas verbais associados a "se" que foram utilizados em dois contextos diferentes e que, portanto, receberam anotações distintas no corpus.

5.3 Métodos de revisão em perspectiva

Ao final do processo de revisão do PetroGold, somamos 30.948 tokens corrigidos, o que corresponde a 12,3% do total de tokens do corpus. Esses tokens são considerados como "revistos" quando, ao compará-los à anotação original (automática) do PetroGold, nota-se que qualquer informação morfossintática foi alterada. Esses tokens podem ter sido modificados manualmente ou via regras, pelo Interrogatório ou pelo Julgamento, e tendo como motivação qualquer um dos métodos de revisão anunciados ou mesmo nenhum deles, pois o erro pode ter sido encontrado ao acaso.

A figura 5.7 mostra a parcela de contribuição dos três métodos semiautomáticos – regras linguísticas, n-grams inconsistentes e IAD (Inter-Annotator Disagreement) – na finalização do PetroGold com o número de erros que foram detectados pelos métodos (e que foram devidamente corrigidos). Além dos três números, há também o campo "outros", para os erros que não foram detectadas por esses métodos, seja porque foram corrigidos via consulta às gramáticas, léxico, ou por meio de qualquer outra forma de exploração manual do corpus, menos sistemática.

O número de tokens da figura soma 33.577, quantidade superior ao número de tokens revistos no PetroGold apresentado anteriormente (30.948).

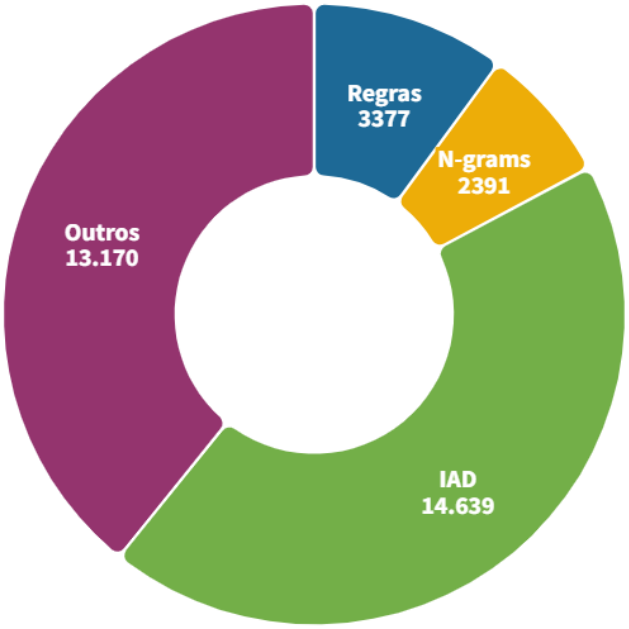


Figura 5.7: Número de erros detectados por cada método semiautomático

Isso se deve ao fato de que um ou mais métodos podem ter detectado erro nos mesmos tokens, de maneira que os métodos podem se sobrepor.

O método de revisão via consulta ao léxico do PortiLexicon-UD resultou em 2.991 tokens corrigidos[29]. Subtraindo-se este número do número de tokens corrigidos por "outros" métodos na figura 5.7 (o que não é metodologicamente preciso, pois não há garantia de que não há sobreposição entre as correções do léxico e as dos três métodos semiautomáticos), restam ainda 10.179 erros (32,9% do total) que foram identificados e corrigidos de outras formas pelos anotadores – como resultado dos estudos gramaticais ou de outras formas exploratórias e menos sistemáticas.

Entre os três métodos semiautomáticos (e mesmo incluindo os "outros" métodos), o método que mais detectou erros foi o IAD (43,6%), seguido de "outros" (39,2%), regras linguísticas (10%) e n-grams inconsistentes (7,1%). A eficiência de um método, porém, não se dá apenas pelo número de erros que detecta (verdadeiros positivos), pois também deve-se considerar a quantidade de falsos positivos, ou seja, tokens que o método identificou como erro mas cuja anotação não está errada. Quanto maior for esse número, maior será o esforço em vão demandado dos anotadores, portanto menor a eficiência do método.

A figura 5.8 traz a proporção de verdadeiros positivos (VP) e falsos positivos (FP) para os métodos de revisão sendo avaliados. Infelizmente, o número de falsos negativos (FN) não é um dado que conseguiremos representar,

[29]Os tokens foram corrigidos por meio da criação e aplicação de 198 regras automáticas.

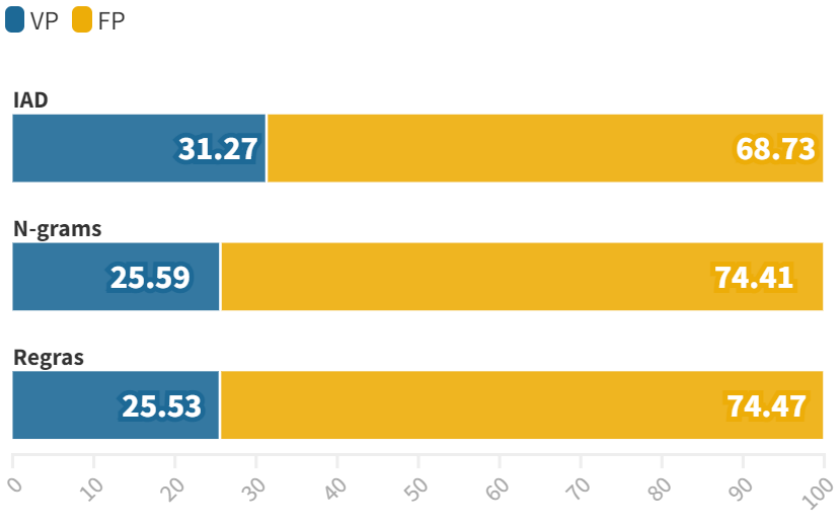


Figura 5.8: Proporção de VP e FP para os métodos de revisão avaliados

pois para consegui-lo precisaríamos garantir que todos os erros do PetroGold foram corrigidos, o que só poderia ser feito com inspeção cuidadosa de *todos* os tokens do corpus, tarefa inviável dadas as características do projeto Petrolês.[30]

Os dados da figura nos mostram que, para os três métodos, mais de 2/3 dos tokens que retornam não são de fato erros a serem corrigidos, mas falsos positivos. Em ordem decrescente, o método com mais falsos positivos são as regras linguísticas (74,47%, ou 9.849 tokens), seguido de n-grams inconsistentes (74,41%, ou 6.952 tokens), e IAD (68,73%, ou 32.171 tokens). Conclui-se portanto que, para o PetroGold, o método IAD foi, além de aquele que mais corrigiu erros, também aquele que menos retornou falsos positivos proporcionalmente, tendo sido o mais eficiente na construção do recurso.

[30]Para uma avaliação mais completa dos métodos de revisão, em Freitas e de Souza (2023) apresentamos inclusive o número de falsos negativos – erros que os métodos não detectam – e calculamos a abrangência de cada um dos três métodos, o que foi possível pois os métodos estavam sendo aplicados em uma fração pequena do Petrolês, inteiramente revista manualmente.

6
Avaliação de um *treebank* padrão ouro

Um treebank padrão ouro é aquele cuja anotação codifica corretamente a interpretação humana dos fenômenos linguísticos. No entanto, além de codificar a nossa leitura sobre os textos do corpus, um treebank deve também ser avaliado de acordo com a sua adequação aos objetivos a que se propõe. O PetroGold, um treebank desenvolvido para o PLN, deve ser um bom material de treino e de avaliação para as tarefas de anotação automática de POS e de dependências sintáticas.

Utilizamos a avaliação intrínseca ao longo de todo este capítulo como uma estratégia para medir a consistência interna do corpus[1]. Considerando que a avaliação intrínseca mede o desempenho de um modelo na realização de uma tarefa para a qual foi treinado, e que os dados de treinamento e de teste desse modelo são um mesmo corpus, esse tipo de avaliação funciona para medir a qualidade do corpus sobretudo quando este tem como objetivo possibilitar um bom aprendizado automático para o PLN.

6.1
Avaliação intrínseca do PetroGold v2

Para comparar as diferentes avaliações intrínsecas que serão realizadas neste capítulo, utilizamos sempre as mesmas ferramentas e métricas, variando apenas a anotação presente no *dataset*, que é o objeto de estudo. Assim, poderemos medir o impacto de cada uma das revisões individualmente na geração de modelos de anotação de dependências. As primeiras avaliações desse capítulo contrastam as versões 1 e 2 do PetroGold, que foram desenvolvidas por meio de métodos de revisão de anotação distintos, e então seguiremos para a avaliação da revisão das questões linguísticas discutidas na seção 5.2.

Os números de avaliação intrínseca serão ainda detalhados a fim de descrever os erros cometidos pelo anotador automático, uma análise qualitativa. Analisaremos a avaliação de cada uma das categorias morfossintáticas da anotação UD, o que nos permitirá visualizar um panorama das dificuldades linguísticas para o aprendizado automático e, por fim, classificaremos todos os erros encontrados na análise da avaliação intrínseca segundo uma taxonomia de erros. A ideia é pensar hipóteses linguísticas que expliquem os erros quando possível – assim, poderemos desenvolver maneiras de melhorar os resultados,

[1]Informações técnicas sobre a forma como efetuamos a avaliação intrínseca foram apresentadas no capítulo de metodologia (seção 4.3).

seja corrigindo ou modificando a anotação do corpus, seja aumentando a quantidade de exemplos de fenômenos difíceis para o analisador automático.

6.1.1 Metodologia

Além da análise quantitativa, procuramos entender linguisticamente os erros encontrados na avaliação intrínseca do PetroGold v2 de duas formas. Primeiro, pedimos a distribuição dos erros por categoria morfossintática, utilizando a ferramenta Julgamento, por onde conseguimos entender quais são as classes que o analisador automático mais erra. Depois, classificamos cada um dos erros encontrados para as categorias linguísticas mais difíceis segundo uma taxonomia inspirada em Manning (2011), que buscou entender os motivos pelos quais havia uma dificuldade, à época, na anotação automática de classes gramaticais – nenhum sistema conseguia alcançar mais de 97% de acertos por token (56% de acertos por frases inteiras)[2].

A taxonomia que utilizamos para classificar os erros do analisador automático os divide entre aqueles que são de fato erros do analisador automático e aqueles que são erros ou deficiências do padrão ouro[3]. A divisão entre erros do anotador automático e erros do padrão ouro pode parecer à primeira vista contrária à proposta da avaliação intrínseca, pois uma vez que o padrão ouro representa as análises corretas, todos os erros do analisador automático deveriam ser, portanto, erros do analisador. No entanto, isso não é totalmente verdadeiro, pois não há *dataset* linguístico à prova de erros e, como estamos analisando cada uma das divergências entre o analisador e o corpus, poderemos distinguir e quantificar os casos nos quais o erro não está na análise automática, mas no padrão ouro.

Quando o erro é do padrão ouro, corrigimos o corpus, seja só a partição de teste onde o erro foi encontrado ou mesmo todo o corpus, quando conseguimos encontrar outros casos semelhantes. Além disso, classificamos os erros de acordo com possíveis causas que possam tê-lo originado. A taxonomia (Tabela 6.1) será explicada a seguir.

[2]Assim como Manning (2011), a classificação dos erros do analisador automático foi realizada apenas por mim, não havendo como avaliar se diferentes anotadores concordariam sobre as mesmas classificações.

[3]Estou chamando de "deficiência do padrão ouro" os casos que não soube julgar se era o padrão ouro ou o anotador automático quem estava errado, seja porque se trata de um fenômeno subespecificado, de difícil interpretação e de difícil documentação, seja porque a frase estava mal construída e/ou difícil de ser interpretada por não especialistas do domínio.

1	**Erro do analisador automático**
1.1	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 6.1: Quadro com taxonomia para classificação de erros

1.1 Pouco frequente no corpus

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.

109. Arenitos laminados podem ser mais sensíveis a a invasão de sólidos de o que os arenitos homogêneos por os seguintes fatos : • Os finos gerados por a penetração em essas lâminas tendem a ser menores e mais difíceis de serem **removidos** , sendo mais propícios a invadir e aprisionar se em a formação . •

110. Observou se que a Argila A4 quando hidratada **desfolha** .

Na frase 109, a oração “serem removidos”, cujo núcleo é o verbo no particípio “removidos”, modifica o adjetivo “difíceis”. Em uma versão anterior da documentação do projeto UD, esse fenômeno receberia a etiqueta de relação “ccomp”, etiqueta comumente atribuída às orações substantivas objetivas (667 vezes), 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 poderia ter causado o erro do analisador automático que, embora tenha acertado o governante da relação, a classificou como de tipo “advcl”. Se, de fato, a baixa ocorrência de “ccomp” nesse contexto for a causa do erro, uma possível solução já está em curso, pois em versões mais recentes das diretivas do projeto UD já se propõe igualar os modificadores de adjetivos em forma de oração aos modificadores de substantivos e advérbios, isto é, como “acl”.

No exemplo 110, “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”. Há duas explicações plausíveis para

o erro: primeiro, a baixa frequência com que o verbo "desfolhar" ocorre no corpus – apenas esta única vez, na partição de teste, sem que tenha servido ao treinamento do anotador automático, além de outras duas ocorrências do substantivo "desfolhamento", que não compartilha nenhuma anotação linguística com o verbo "desfolhar" senão por alguns dos caracteres do lema.

A segunda hipótese 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). Em números relativos, 22,8% das frases com oração subordinada adverbial seguem esse ordenamento[4]. Soma-se a essa explicação, porém, o fato de que, 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 111, com sinal de pontuação, e 112, sem sinal de pontuação)[5]. A soma 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.

111. ***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 .

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

[4]Na ET, encontramos o número de sentenças com oração subordinada adverbial anteposta à oração principal por meio da expressão de busca: deprel = "advcl" and id < head_token.id (lê-se: buscar um token anotado como advcl – oração adverbial – cujo id seja menor – portanto, esteja à esquerda – que o id do seu governante – o núcleo da oração principal). O ordenamento canônico, por sua vez, foi encontrado invertendo o sinal de "menor" para "maior": deprel = "advcl" and id > head_token.id

[5]Na ET, as frases com pontuação dependente da oração adverbial foram encontradas por meio da seguinte expressão de busca: deprel = "punct" and head_token.deprel = "advcl" and head_token.id < head_token.head_token.id, sendo que as demais frases são as que não têm pontuação.

1.2 Ambiguidade estrutural

Essa categoria corresponde aos erros do analisador automático cuja origem tem uma explicação plausível no fato de que a estrutura admite duas ou mais anotações sintáticas distintas, sendo que a solução para a ambiguidade necessita de uma interpretação contextual e o acionamento de outros conhecimentos, o que pode fugir ao escopo do analisador automático.

113. 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 .

114. 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) .

115. 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 113, 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 114, "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 a apenas *recompressão*, motivo pelo qual sabemos qual deve ser a anotação correta, a despeito da ambiguidade sintática.

Já na frase 115, 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.

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 acertados pelo anotador.

116. 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 .

117. Uma unidade consiste de duas folhas de um aglomerado de oxigênios ou hidroxilas junto com alumínio , ferro , ou magnésio formando uma estrutura octaédrica (**Figura** III.2a) .

118. Os zeólitos atuam de a mesma forma que as **resinas** de troca iônica empregados em a adsorção de compostos orgânicos dissolvidos em as águas produzidas .

Na frase 116, 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", uma 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.

Na frase 117, o erro do analisador se deu ao anotar "Figura" como aposto de "estrutura", o que também se poderia encarar como um problema de ambiguidade estrutural, já que as classes de palavras envolvidas permitiriam essa anotação. Contudo, trata-se de uma estrutura extremamente frequente no texto acadêmico, tanto é que, como vimos, a ela foi dado um tratamento sistemático e específico relativo às referências bibliográficas e figuras – a anotação correta é a de parataxis para a raiz da frase. Assim, dada a frequência do fenômeno e o grau de cuidado por que o corpus passou para sistematizar a anotação do fenômeno, o modelo não deveria errar em construções semelhantes.

Em 118, não há motivos para justificar o fato de que o anotador automático analisou "resinas" como sujeito de "atuam", sendo *resinas* núcleo

de uma oração encaixada à oração principal e sendo que "atuam" já tem um sujeito, "zeólitos". A análise correta entende que "resinas" é um complemento de "forma" (nmod), assim como em "(...) atuam *igual* às **resinas**" a palavra em negrito seria complemento da palavra em itálico (obl). Não é um fenômeno pouco comum, não há ambiguidade estrutural e a análise proposta pelo anotador, além de errada, esbarra em restrições do modelo gramatical, pois propõe dois sujeitos para um mesmo verbo.

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.

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

120. 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 119, o analisador automático coordenou "Hz" à palavra "Duto", ao passo que o padrão ouro coordena à 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 120, 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.

2.2 Diretivas 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. Acrescenta-se a essa classe as frases de difícil interpretação, seja porque mal escritas, seja porque a interpretação depende de conhecimento de domínio muito específico, tornando difícil decidir qual a anotação correta.

121. Apesar de os estudos de estes autores terem sido restritos a a 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 .

122. 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) .

123. 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

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

Na frase 121, é difícil dizer se "nordeste" deve se relacionar a região ou 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 122, 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 eu) 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 123, 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 124, há ambiguidade possível entre uma estrutura adnominal – "êmbolo da argila" como um objeto do verbo "desprender" (nmod) – e entre uma estrutura argumental do verbo – "da argila" como lugar de onde se deve desprender o êmbolo (obl:arg). Ambas as leituras são aceitáveis, sobretudo com pouco contexto como ocorre na frase.

2.3 Sem explicação aparente

Os erros do tipo 2.3 são aqueles que não são derivados nem da falta de diretivas 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.

125. A desconformidade é evidenciada por o substrato erosivo irregular de os arenitos (fácies St) e conglomerados (fácies Gt) de a base de o Membro Romualdo sobre os folhelhos (fácies **Fl**) e gipsita (fácies GIP) de as Camadas Ipubi ; tais evidências são verificadas em a seção Pedra Branca (Figura 27) .

126. 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 de o 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) .

127. II.3 – **ENCERAMENTO** DE BROCA

Na frase 125, "Fl" é um tipo de fácies, um aposto especificativo que, segundo as diretivas UD e segundo já documentado para o PetroGold, deve

Corpus	LEMMA	UPOS	UAS	LAS	CLAS
PetroGold v2	98,54%	98,40%	90,92%	89,09%	**84,07**%
PetroGold v1	98,48%	98,19%	90,65%	88,53%	**82,96**%
Bosque-UD v2.8	96,95%	96,52%	85,83%	81,59%	**73,80**%

Tabela 6.2: Comparação entre números de avaliação intrínseca

ser anotado como adjunto adnominal, não havendo motivo para que, no padrão ouro, "Fl" estivesse anotado como *parataxis* de fácies. Na frase 126, "Hasui" estava anotado como aposto de "Almeida", quando na verdade há uma coordenação (conj) entre as referências bibliográficas. Esses são casos que já foram consistentemente documentados e revistos no corpus, não havendo explicação para o erro. Já na frase 127, 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 raíz e o nome o seu modificador.

6.1.2 Resultados e análise

Avaliação quantitativa

A tabela 6.2 apresenta os números relativos à avaliação intrínseca para três *datasets*: o PetroGold v1, o PetroGold v2 e o Bosque-UD v2.8 (RADEMAKER et al., 2017). A comparação entre a primeira versão do corpus e o Bosque-UD v2.8 foi realizada em de Souza et al. (2021b), e tinha como objetivo contrastar um dos principais treebanks para língua portuguesa com o recém-desenvolvido PetroGold. Já a comparação entre ambas as versões do PetroGold foi realizada em de Souza e Freitas (2022a) e tinha como objetivo avaliar o impacto da revisão de diversos fenômenos linguísticos em um *dataset* já considerado padrão ouro.

Visando permitir a comparação entre as avaliações, garantimos a mesma proporção de frases para as partições de treinamento e teste – respectivamente, 95% e 5% das frases, como sugere o projeto UD –, resultando em um número semelhante de frases de treinamento para o PetroGold e o Bosque-UD – respectivamente, 8.671 e 8.328 –, já que são corpora de tamanhos parecidos. Além disso, as frases que constam das partições de treino, teste e desenvolvimento são as mesmas para ambas as versões do PetroGold, excetuando-se as frases que foram removidas da segunda versão.

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 6.3: Avaliação por POS

Os dados da tabela nos mostram que a primeira versão do PetroGold atingiu um desempenho até 12,4% melhor que o Bosque-UD, no que diz respeito à anotação sintática de palavras de conteúdo (CLAS). A diferença de CLAS é significativa, sugerindo um grande impacto positivo na qualidade de um corpus quando utilizamos os métodos de revisão para corrigir sistematicamente os erros de uma análise automática[6].

Por outro lado, da primeira para a segunda versão do PetroGold, a melhoria foi de apenas 1,33%, também para a métrica CLAS. Esse aumento, mais tímido, indica que melhorar a anotação de um corpus já considerado padrão ouro, seja corrigindo erros a partir da execução de um leque maior de métodos de revisão, seja anotando questões linguísticas de formas distintas, não apresenta tanto impacto positivo na avaliação quanto o esforço despendido nessa revisão fina sugeriria. Contudo, ressaltamos a importância dessa revisão, pois recursos padrão ouro devem conter a análise humana que julgamos correta, a despeito de serem facilmente generalizáveis ou não.

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

Avaliação qualitativa

Classes gramaticais

Na tabela 6.3, 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. 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.

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

[6]Outro fator para os resultados melhores do PetroGold pode estar relacionado ao gênero textual – o texto acadêmico é muito mais previsível e formulaico que o texto jornalístico.

o baixo índice de acerto, portanto, poderia ser a pouca ocorrência no corpus, dificultando a generalização. Outro motivo é que, sendo poucas ocorrências na partição de teste, um único erro impacta muito negativamente o número relativo de acertos, enquanto que nas classes robustas, com um grande número de ocorrências, os erros tendem a se diluir no todo.

As classes NOUN, PRON, ADV, PROPN e ADJ tiveram menos acertos do que o F1 de UPOS na avaliação intrínseca, isto é, ficaram abaixo da média das outras classes. Todas elas são classes abertas, o que pode justificar a maior dificuldade em acertá-las uma vez que seu número é grande em comparação às demais classes e há muitas palavras dessas classes com baixa ocorrência. 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.

Relações sintáticas

Na tabela 6.4, que diz respeito a relações sintáticas, a primeira coluna é a etiqueta da classe, 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 de LAS: nsubj, nmod, nsubj:pass, fixed, punct, mark, advmod, aux, acl, xcomp, obl, advcl, csubj, obl:arg, ccomp, acl:relcl, conj, appos, parataxis, compound, goeswith.

Dessas classes, algumas apresentaram mais erros relacionados ao encaixe da dependência do que à atribuição da etiqueta em si. Elas são, em ordem decrescente de dificuldade: acl:relcl, conj, appos, acl, punct, advmod e advcl. São as classes em que a identificação da relação – uma relação de adjetivação, aposição, condição (ou qualquer outro dos sentidos adverbiais) – não é tão difícil de ser identificada quanto o é para as máquinas identificar o governante da relação devido à ambiguidade sintática.

No exemplo abaixo, o trecho "concentração da amostra de clorofórmio" contém dois adjuntos adnominais facilmente identificáveis (pois são substantivos modificando substantivos[7]) – "amostra" e "clorofórmio". A dificuldade,

[7]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.

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%	compound	36	47,22%	41,67%
nsubj:pass	140	86,43%	86,43%	goeswith	1	0	0
fixed	166	86,14%	86,14%				

Tabela 6.4: Avaliação por REL

porém, reside na distinção entre a estrutura [concentração [da amostra [de clorofórmio]]] e [concentração [da amostra] [de clorofórmio]], pois muitas vezes só conseguimos realizá-la semanticamente – nesse caso, sabemos que "amostra" é um substantivo comumente adjetivado para indicar de que tipo de amostra se está falando. Assim, a única análise correta é que "clorofórmio" é um adjunto adnominal dependente de "amostra", que por sua vez é adjunto adnominal de "concentração".

128. 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 cuja etiqueta foi atribuída erroneamente, portanto 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: compound, obl:arg, parataxis, ccomp, csubj, xcomp, aux e obl.

Para a classe de "compound", como vimos na seção 5.2.2, realizamos a sua substituição pela classe dos adjuntos adnominais (nmod). Com isso, abrimos mão de distinguir termos compostos – "cachorro quente" e "óleo diesel", por exemplo – tendo em vista (1) alinhar a anotação do corpus à visão sobre expressões multipalavras (MWEs) do projeto Universal Dependencies, e (2) deixar de fornecer exemplos dúbios e com pouca ocorrência de uma classe que

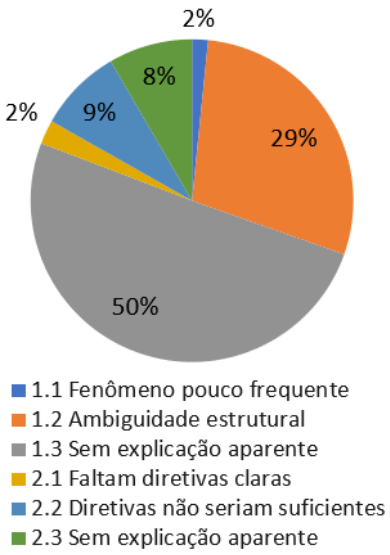


Figura 6.1: Distribuição dos tipos de erros encontrados na avaliação intrínseca do PetroGold v2

não foi tratada nem durante a confecção do PetroGold nem do corpus que alimentou o modelo utilizado para a sua anotação automática, o Bosque-UD.

Tipos de erros

Foram analisados todos os erros encontrados na avaliação intrínseca de seis das classes mais difíceis (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. Para as classes acl:relcl (orações relativas), appos (apostos) e conj (coordenações), verificamos todos os erros de encaixe de dependências, pois eram os mais frequentes; já para as classes ccomp (oração subordinada substantiva objetiva direta), obl:arg (objeto preposicionado) e parataxis (orações inseridas sem conectivo), verificamos todos os erros de identificação da relação sintática. Classificamos os erros segundo a taxonomia descrita em 6.1.1 e os resultados se encontram no gráfico 6.1 e na tabela 6.5.

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.

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			

Tabela 6.5: Classificação dos erros das classes com pior desempenho na avaliação intrínseca do PetroGold v2

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, 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 do ponto de vista computacional.

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[8].

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 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 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 PetroGold 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.

6.2 Avaliação das questões linguísticas

Nesta seção, avaliaremos o impacto no PLN das correções das questões linguísticas aplicadas na seção 5.2 por meio da avaliação intrínseca de um

[8]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.

	LEMMA	UPOS (%)	UAS (%)	LAS (%)	CLAS (%)
Sem *obl:arg*	98.54	98.40	90.66	88.82	83.48
	(0)	(0)	(-0,26)	(-0,27)	(-0,59)

Tabela 6.6: Avaliação do modelo quando *obl:arg* é convertido para *obl*

modelo de anotação em diferentes cenários, para cada revisão individualmente.

6.2.1 Avaliação do argumento verbal introduzido por preposição

As correções relativas ao argumento verbal introduzido por preposição (etiqueta *obl:arg*) já haviam sido incorporadas na segunda versão do PetroGold e discutidas anteriormente (DE SOUZA; FREITAS, 2022a; DE SOUZA; FREITAS, 2022b). Como consequência, a análise da avaliação intrínseca do PetroGold v2 (realizada na seção 6.1) já considerou os erros do analisador automático para a classe *obl:arg*. Nesta seção, iremos avaliar o impacto da correção de *obl:arg* nas métricas de avaliação intrínseca.

Para avaliar esse impacto, contrastamos os resultados da avaliação do PetroGold v2, que contém a etiqueta *obl:arg*, aos resultados para o mesmo *dataset* quando eliminamos a distinção entre *obl:arg* e *obl*, transformando tudo em *obl*. Essa modificação aproxima a anotação do PetroGold à proposta de Zeman (2017) para o UD, já que o projeto não prevê distinção entre os sintagmas preposicionados que são argumentos do verbo e os que não são, tornando a anotação menos granular e, portanto, mais simples.

Realizamos essa modificação de caráter experimental apenas para avaliar o impacto do acréscimo da etiqueta *obl:arg*, e o resultado foi a transformação de 1.488 tokens anotados como *obl:arg* em *obl*. Os tokens se distribuem por 14,8% das frases do corpus de maneira que, embora sejam poucos, se espalham por um número alto de frases.

Os resultados da tabela 6.6 mostram uma queda em todas as métricas relacionadas à anotação de dependências (UAS, LAS e CLAS), principalmente CLAS, com performance pior sem a etiqueta *obl:arg*. As métricas relativas à lematização e à atribuição de POS não sofreram alteração quando a etiqueta *obl:arg* foi alterada para *obl*.

A primeira conclusão que podemos tirar dos números é que diferenciar os sintagmas preposicionados que são argumentos do verbo daqueles que não são argumentos na anotação do PetroGold foi benéfico para o aprendizado de dependências sintáticas como um todo, uma vez que os números de avaliação intrínseca foram melhores com essa diferenciação. Essa conclusão pode parecer contraintuitiva, uma vez que a adição da etiqueta *obl:arg* introduz uma

granularidade à anotação linguística e, ainda por cima, uma de difícil distinção em muitos momentos, como vimos na seção 5.2.1. Nossa expectativa era a de que seria mais fácil aprender a anotar se a única etiqueta existente fosse a de *obl*, o que não se confirmou nos números.

Para entender os motivos da melhora nos resultados, verificamos especificamente os acertos da etiqueta *obl:arg* no PetroGold v2, e concluímos que é uma classe difícil, com apenas 62,8% de acertos. Ao mesmo tempo, verificamos que a classe *obl* é melhor aprendida quando a etiqueta *obl:arg* não existe – sem ela, os acertos de *obl* são 86,4%; com *obl:arg*, os acertos de *obl* são 79,9%. Esses números sozinhos não justificariam, portanto, a melhoria na avaliação do PetroGold v2.

No entanto, verificamos uma melhoria considerável em outras classes quando introduzimos a etiqueta *obl:arg* no corpus – *ccomp* cresce de 58,6% para 65,5% e *acl:relcl* cresce de 93,4% para 95,6%. As duas categorias, que foram mais afetadas positivamente, são destinadas a núcleos de orações (substantivas objetivas, no caso de *ccomp*, e relativas, no caso de *acl:relcl*). Embora não pareça haver nenhuma relação direta entre essas classes e o *obl:arg*, a melhoria nelas explica a melhoria também das métricas de avaliação intrínseca do PetroGold v2 observada.

6.2.2 Avaliação das expressões multipalavras

Nesta seção, avaliamos o impacto da revisão das MWEs no aprendizado automático e discutimos os erros do anotador automático na avaliação intrínseca. Além disso, testaremos o impacto ao mudar a POS das palavras que compõem uma MWE do tipo *fixed* para a POS da locução, e não das palavras isoladamente.

Os números da tabela 6.7 indicam que o corpus com revisão de MWEs alcança resultados até 1,43% piores, na avaliação intrínseca, no que diz respeito à análise sintática de palavras de conteúdo (CLAS). A tendência de piora se encontra também nas métricas de LEMMA, UAS e LAS, sendo que houve melhora apenas na análise de POS, equivalente a 0,13%. A expectativa era que, ao contrário, os números fossem relativamente melhores uma vez que a quantidade de expressões novas anotadas como MWE é maior que o número de expressões que passaram a ser anotadas de forma transparente, o que talvez pudesse tornar o aprendizado facilitado uma vez que a anotação de expressões multipalavras é "plana", sem estrutura interna.

Contudo, uma hipótese para a queda nos números do aprendizado se relaciona ao fato de que um dos critérios utilizados para anotar certas

	LEMMA	**UPOS**	**UAS**	**LAS**	**CLAS**
sem revisão	**98.55**	98.35	**91.04**	**89.10**	**84.06**
com revisão	98.51	**98.48**	90.04	88.30	82.85

Tabela 6.7: Avaliação intrínseca do corpus após a revisão das MWEs

previsto **padrão-ouro**	**advmod**	**case**	**det**	**mark**	**nmod**	**nsubj**	**obl**	**root**
fixed	1	4	1	2	3	1	2	1

Tabela 6.8: Erros do anotador automático envolvendo MWEs no padrão ouro

padrão-ouro **previsto**	**case**	**mark**	**obl**
fixed	1	1	2

Tabela 6.9: Erros envolvendo MWEs na anotação automática

expressões como MWE, como discutido, é o contraste ao interpretar a mesma expressão em contextos de MWE ou de construção transparente. É o caso, por exemplo, de expressões como "isto é", que ora pode ser locução conjuntiva, ora sujeito seguido de verbo auxiliar. Como realizamos uma revisão abrangente desse e de outros casos, introduzimos, assim, uma dimensão interpretativa a fenômenos que, anteriormente, tinham anotação simplificada, mas incorreta.

Outra forma de procurar entender a queda nos números da avaliação intrínseca é a análise da matriz de confusão da partição de teste do PetroGold após as revisões de MWE. No que diz respeito à etiqueta "fixed", reservada às expressões multipalavras que revisamos, o padrão-ouro da partição de teste contém 202 tokens anotados com a etiqueta, dos quais 15 o anotador automático errou (7,4%), segundo a tabela 6.8. Há ainda 4 tokens que o anotador automático identificou como "fixed" que não deveriam ser (Tabela 6.9), em um movimento de hipergeneralização das MWEs para além dos contextos desejáveis.

No exemplo da frase 129, o sistema previu uma sequência de adjuntos adnominais – "bomba" como adjunto adnominal de "meio" que, por sua vez, é adjunto adnominal de "água". Essa análise não se sustenta no contexto da frase – embora em construções como "meio de comunicação" a palavra "meio" tenha um adjunto adnominal, o sentido empregado neste sintagma é diferente do sentido da frase em questão, motivo pelo qual a análise da expressão como locução prepositiva é a correta.

Na frase 130, por sua vez, vemos uma repetição da palavra "já" que, no contexto, acabou recebendo a mesma anotação pelo anotador automático nas duas ocorrências – adjunto adverbial do verbo "esperado". A análise correta,

porém, é a de locução conjuntiva "já que". Em ambos os casos, o anotador automático anotou isoladamente as palavras, "meio" e "já", sem considerar que fazem parte de uma locução e portanto devem receber anotação distinta.

129. O sistema é ainda inovador em relação a os equipamentos tradicionais que transportam , separadamente , o petróleo e a água por meio **de** bomba e o gás natural por meio de compressor , já que este permite a adição de energia a a corrente multifásica sem que seja requerido qualquer pré-condicionamento de essa corrente .

130. Esse resultado já era esperado **já** que o poder calorífico de o diesel é maior que o biodiesel fazendo com que a temperatura em a descarga tenda a ser maior .

Já nos casos que o anotador automático identificou como MWEs construções que não são, encontramos exemplos como a seguir. Na frase 131, o anotador estendeu o alcance da MWE "por o menos" (pelo menos), inserindo ao final dela a preposição "em" ("pelo menos em"), que não havia sido prevista na anotação do padrão-ouro, pois não faz parte da expressão e pode funcionar normalmente como preposição sem prejuízo para a anotação da frase.

No exemplo 132, vemos um erro de digitação, onde foram inseridas aspas antes da preposição "a" que integra a MWE "com relação a". Por conta dessa interrupção da forma da MWE, a anotação do padrão-ouro não incluiu a construção com anotação fixa, pois UD requer que expressões do tipo fixa sejam contíguas, sem tokens intervenientes, como é o caso das aspas. Nesse caso, o anotador automático previu que a MWE é "com relação", sem a preposição, o que não havíamos previsto.

Já na frase 133, a expressão "dessa forma" foi analisada pelo anotador automático como locução adverbial, a despeito do fato de que no padrão-ouro a expressão foi anotada de forma transparente, como adjunto adverbial (formalmente similar a "de manhã", por exemplo), uma decisão tomada pois a anotação transparente não compromete a interpretação da expressão no contexto em que está inserida. Embora seja um erro do anotador automático, há respaldo em muitos compêndios gramaticais para a análise.

131. Estes valores poderiam ser explicados , por o menos **em** a porção superficial de os modelos , por a presença de o pacote sedimentar que se espessa suavemente para SE .

132. A Lei Nº 9.966 , de 28 de abril de 2000 " Lei de o Óleo " atribuiu a o MMA responsabilidades em a identificação , localização e definição de os

	LEMMA	UPOS	UAS	LAS	CLAS
sem MWEPOS	**98.51**	**98.48**	90.04	88.30	82.85
com MWEPOS	98.50	98.28	**90.41**	**88.67**	**83.35**

Tabela 6.10: Avaliação intrínseca com e sem informação de POS da MWE

limites de as áreas ecologicamente sensíveis com **relação** " a a poluição causada por lançamento de óleo e outras substâncias nocivas ou perigosas em águas sob jurisdição nacional (BRASIL , 2005)

133. Coletou se uma aliquota de a fase orgânica , esta foi levada a um espectrofotômetro de absorção molecular UV-visível , a um comprimento de onda de 260 nm-1 , de essa **forma** , a concentração de a amostra de ciclohexano corresponde a concentração de óleo em a água .

Por fim, realizamos também um experimento em que modificamos a classe gramatical das palavras que compõem as MWEs. Conforme apresentamos, o projeto UD requer que as palavras tenham a anotação de POS das palavras isoladas, a despeito do contexto em que estão inseridas ou de fazerem parte de expressões multipalavras. Para contrapor a essa anotação, adicionamos uma anotação extra para todas as MWEs, na última coluna de anotação, reservada às informações do tipo miscelânea, relativa à POS que a MWE exerce na oração. Agora, utilizamos essa informação para anotar todas as palavras que compõem a MWE com essa POS e, assim, podemos comparar os resultados de avaliação intrínseca com e sem a modificação. Em uma expressão como "isto é", por exemplo, em que as palavras haviam sido anotadas com POS pronome e verbo auxiliar, respectivamente, mesmo nos contextos em que funciona como locução conjuntiva, ambas as palavras passaram a ter a anotação de POS para conjunção coordenativa durante o experimento.

A tabela 6.10 mostra a diferença entre o corpus sem MWEPOS, isto é, com POS de cada uma das palavras isoladas (os números já haviam sido apresentadas na tabela anterior, pois corresponde ao corpus "com revisão" de MWEs), e o corpus com MWEPOS, isto é, o mesmo corpus, com revisões, porém com POS das palavras referente à POS da MWE como uma unidade.

Embora tenha havido uma queda no aprendizado na anotação morfológica (lema e POS) de até 0,2%, há o aumento nas métricas de análise sintática, alcançando até 0,6% para as palavras de conteúdo lexical. O decréscimo nas métricas de morfologia pode ser explicado pelo fato de que palavras que, até então, sempre tinham a mesma POS, passaram a ter POS diferentes a depender do contexto em que aparecem. O aumento nas métricas de sintaxe, por sua vez, é explicado pelo fato de que, anteriormente, havia um descolamento entre

LEMMA	UPOS	UAS	LAS	CLAS
98.46 (-0,09)	98.40 (+0,05)	90.59 (-0,45)	88.71 (-0,39)	83.33 (-0,73)

Tabela 6.11: Avaliação intrínseca após as revisões do pronome "se"

POS e relação de dependência – usando como exemplo a expressão "isto é", a anotação anterior era a de um pronome sendo usado como conjunção coordenativa, pouco usual, dando lugar à anotação de uma conjunção coordenativa com função de conjunção coordenativa, o que é o mais natural.

Assim, optar por uma anotação ou outra pode ser apenas uma escolha prática a depender da tarefa a que o corpus deve se prestar. A avaliação intrínseca quando anotamos a POS da MWE como um todo, segundo os resultados, parece compensar as perdas no aprendizado de uma informação linguística (lema e POS) com ganhos semelhantes em outra (sintaxe).

6.2.3 Avaliação do pronome "se"

Realizamos uma avaliação intrínseca do corpus com as novas etiquetas, "expl:impers", "expl:pass" e "expl:pv", para comparar os resultados com a avaliação do corpus PetroGold v2, sem as correções citadas e as novas etiquetas. Assim, conseguimos medir o impacto da subespecificação do "se" na avaliação do corpus.

Os resultados podem ser verificados na tabela 6.11, com a variação desde o PetroGold v2 entre parênteses.

A métrica relativa ao aprendizado de classe gramatical foi a única que obteve uma melhora em relação à avaliação intrínseca do PetroGold v2 (+0.05%). Isso pode ser explicado por alguns motivos: (1) foram realizadas correções sistemáticas relativas a quando o "se" é pronome ou conjunção subordinativa, facilitando o aprendizado automático de POS; (2) foi realizada uma simplificação das informações morfológicas do pronome expletivo "se" – originalmente, eram anotados como tendo atributos morfológicos de um pronome de terceira pessoa, como se fosse um objeto, portanto herdando as características do objeto da oração. Como se trata de um pronome expletivo, que não representa nem um sujeito nem um objeto, removemos completamente suas informações morfológicas, o que pode ter facilitado o aprendizado do etiquetador.

O decréscimo de 0,09% na avaliação de lematização, por sua vez, pode ser explicado pelo fato de que havia 22 palavras "sudeste" ou "Sergipe", abreviadas como "SE", mas que tinham o lema "se", em letras minúsculas, o que facilitava o aprendizado uma vez que, independentemente de a palavra estar em caixa

alta ou não, o lema era sempre o mesmo. Quando desfizemos essa anotação de lema, diferenciando o "se" pronome do "SE" abreviação para "sudeste", introduzimos um pequeno obstáculo que pode ter refletido no decréscimo.

Já as métricas que dizem respeito ao aprendizado de dependências (UAS, LAS e CLAS), seja na anotação da relação ou no encaixe da dependência, tiveram o desempenho piorado em até 0,86% para CLAS. Esse dado pode ser explicado por termos introduzido uma granularidade previamente inexistente no corpus quando adicionamos três novas classes para o pronome "se" – expl:impers, expl:pass e expl:pv. No PetroGold v2, a classe expl obtinha 100% de acertos pois era a única para todos os casos de pronome "se". Nessa nova versão, os resultados de acerto para as três novas classes foram de, respectivamente, 82.3%, 91.3% e 86.8%. Soma-se a isso o fato de que os verbos são polissêmicos e, dependendo do contexto, as orações – e portanto o "se" – podem ser interpretadas de uma forma ou de outra, conforme discutimos.

A etiqueta mais difícil de ser aprendida, expl:impers, tem 17 ocorrências na partição de teste do PetroGold, sendo que em duas delas o sistema previu a anotação como expl:pass e em uma, expl:pv. Uma das confusões com "expl:pass" é a da frase 134, um dos poucos casos onde o verbo que sofreu indeterminação do sujeito é transitivo direto. Nela, um verbo frequentemente empregado com objeto direto (o verbo "ter"), que portanto deveria poder ser empregado na voz passiva, está sendo usado de forma impessoal para indeterminar o sujeito, fenômeno já conhecido pelas gramáticas – segundo a tradição gramatical, o fenômeno ocorre com os verbos ter, haver e existir, caso em que os verbos não flexionam em número pois não têm sujeito.

134. Por um pequeno período , **teve se** a produção somente de o fluido.

Já no caso da frase 135, em que o sistema anotou como "expl:pv", verificamos o mesmo verbo, "ter", tendo o sujeito indeterminado a despeito de o verbo ser transitivo direto. A anotação de uso pronominal do verbo se justificaria uma vez que o anotador automático entendeu que "onde" era sujeito do verbo, embora não seja correto, pois a oração não tem sujeito.

135. Onde **se teve** primeiramente a extração e depois a filtração em o equipamento montado em o Nupeg (Figura 3.3) , esse sistema era composto por um compressor que matinha o reservatório pressurizado em uma pressão de 1 Bar , o reservatório tinha um volume de 500 mL , e logo após tínhamos uma válvula conectada entre o reservatório e o leito poroso.

A classe da voz passiva sintética, "expl:pass", foi erroneamente anotada como "expl:pv" 4 vezes apenas. Na frase 136, a anotação automática como

"expl:pv" se mostra errada quando a frase é interpretada, pois entendemos que as semelhanças às quais a oração se refere são notadas pelos autores da dissertação. Uma distinção semântica que, parece, não foi realizada pelo anotador automático.

136. Apesar de os estudos de estes autores terem sido restritos a a 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.

Já a classe do uso pronominal dos verbos, "expl:pv", foi erroneamente anotada como "expl:pass" 1 vez e 3 vezes como "expl:impers". A confusão com a classe da voz passiva sintética, na frase 137, imprimiria um sentido diferente à frase – caso a interpretação fosse a de que o estado do Ceará tivesse sido encontrado [por alguém], sendo esse alguém propositalmente omitido. Contudo, mais uma vez, o anotador automático falhou ao distinguir a interpretação correta.

137. Parte de essa bacia onde **se localiza** o estado de o Ceará , onde se pode encontrar em o litoral de o mapa a seguir o município de Icapuí , localizado em o litoral de o Estado de o Ceara.

Os erros de "expl:pv" que foram anotados como índice de indeterminação do sujeito, por sua vez, dizem respeito a uma dificuldade imposta pelo modelo UD uma vez que ignora a existência de locuções verbais modais, como a da frase 138, onde a locução "podem formar" tem um "se" interveniente[9]. A anotação UD prevê que, nesses casos, não há locução verbal, mas uma subordinação por "xcomp" (complemento oracional fechado, compartilhando o mesmo sujeito). Nessa estrutura, não há clareza se o pronome deve depender do primeiro ou do segundo verbo, o que pode ter acarretado o erro da anotação automática – no treebank, o "se" depende do verbo "formar" como partícula integrante do verbo, enquanto que na previsão automática o "se" dependeu de "podem" como índice de indeterminação do sujeito, ambas as análises razoáveis, sendo a ambiguidade ocasionada pela lacuna nas diretivas do projeto mencionada.

138. - Pluma de descarte estreita - Tendência a o acúmulo de cascalho (dependendo de a situação **podem se formar** pilhas submarinas)

[9]Em UD, são consideradas locuções verbais e, portanto, anotadas com um verbo auxiliar e outro principal, apenas as locuções verbais de tempo composto e a voz passiva com "ser" auxiliar. O projeto permite que cada língua decida quais verbos podem ser anotados como auxiliares, no entanto, pela dificuldade de estabelecer uma lista de verbos que compõem locuções verbais aspectuais e modais, optou-se por não incluir essas locuções em UD para português. Para uma discussão aprofundada, ver de Souza e Freitas (2019).

Há ainda um erro de anotação do "expl:pv" que diz respeito ao encaixe da dependência para o qual não conseguimos encontrar justificativa possível. Na frase 139, o modelo anotou o "se" em negrito como dependente do verbo "demonstraram", também em negrito, quando na verdade ele deveria ter como dependente o verbo "manteve".

139. Comparados com a referência , os resultados se **demonstraram** satisfatórios pois mesmo não mantendo a tendência de diminuição de o O2 com o aumento de a carga entre os dois primeiros pontos de operação essa tendência **se** manteve posteriormente e a ordem de grandeza de os resultados coincidiu.

6.3 Em busca do ouro: o PetroGold v3

A terceira versão do PetroGold[10] marca também o lançamento do treebank no projeto Universal Dependencies, realizado em 15 de novembro de 2022[11]. Como já comentado na seção 4.1.1, só há duas diferenças entre o PetroGold v3 e o PetroGold UD: (1) neste último, a etiqueta *nmod:appos* dá lugar à etiqueta *nmod*, referente aos adjuntos adnominais, conforme discutido na seção 4.1.1; e (2) a versão UD tem uma divisão de frases por partição (treino, teste e desenvolvimento) diferente do particionamento da versão do Petrolês, sendo que em UD as partições contêm documentos completos e no Petrolês as partições contêm frases aleatórias, mas seguindo a mesma seleção das versões 1 e 2 do corpus.

Nesta seção do trabalho, para facilitar a comparação entre todas as versões disponíveis, convertemos a etiqueta *nmod:appos* em *nmod* de todas as versões. Embora o ideal fosse avaliar a pertinência dessa etiqueta e o seu impacto no aprendizado automático, esse não foi um assunto abordado nesta dissertação e pode ser um dos tópicos para um trabalho futuro.

A tabela 6.12 mostra as características do PetroGold v3 em comparação com a v2, sendo que os números entre parênteses demonstram o acréscimo/-decréscimo em cada uma. O aumento no número de tokens sinaliza apenas a correção da tokenização de 10 frases que haviam sido mal tokenizadas previamente. A tabela 6.13, por sua vez, mostra os números de F1 da avaliação intrínseca do PetroGold v3 do projeto Petrolês em comparação com a mesma versão no projeto UD e o PetroGold v2.

[10]Disponível em <https://petroles.puc-rio.ai>. Acesso em: 10 jan. 2023.

[11]O treebank pode ser encontrado na página do projeto UD, seguindo o endereço: <https://github.com/UniversalDependencies/UD_Portuguese-PetroGold>. Acesso em 10 jan. 2023.

	v3	
Tokens	250.605	(+10)
Correções	30.948	(+9.314)
Frases	8.946	
Documentos	19	

Tabela 6.12: Características do PetroGold v3 em comparação com a v2

	LEMMA	**UPOS**	**UAS**	**LAS**	**CLAS**
v2	98,54	98,40	90,92	89,09	84,07
v3 (petrolês)	98,60	**98,63**	**91,04**	**89,36**	**84,22**
v3 (ud)	**98,77**	98,42	90,50	88,63	83,30

Tabela 6.13: Avaliação intrínseca do PetroGold v3 em comparação com a v2

Comparar a versão UD do PetroGold com as demais pode ser considerada uma comparação de elementos desiguais, uma vez que as frases que constam de cada partição são diferentes, de maneira que os números não significam necessariamente melhora ou piora na qualidade do *dataset*. Pode-se concluir apenas que o particionamento realizado no projeto Petrolês proporciona números quase sempre maiores que os números do particionamento do projeto UD, chegando a até 0,92 ponto percentual de diferença de CLAS entre as duas variações do PetroGold v3. A versão UD ganha da versão Petrolês apenas na anotação de lema, com uma diferença de 0,17 ponto percentual.

A comparação entre o PetroGold v2 e o PetroGold v3 mostra uma melhoria na qualidade do modelo treinado nesta última versão, com diferença de CLAS chegando a 0,15 ponto percentual e superior em todas as demais métricas de avaliação. Vale lembrar, porém, que o PetroGold v3 tem três etiquetas de anotação sintática novas em relação à v2: *expl:pv*, *expl:pass* e *expl:impers*. A comparação entre os *datasets* pode ter sido influenciada de maneiras imprevisíveis pela adição das etiquetas, por isso mostramos a seguir também a diferença de avaliação para cada uma das categorias morfossintáticas do treebank, indicando quais foram as categorias mais afetadas positivamente e negativamente pelas revisões e pelas novas etiquetas introduzidas.

A tabela 6.14 mostra os números de acerto de etiqueta e de acerto tanto da etiqueta quanto do encaixe (LAS) na avaliação intrínseca do PetroGold v3 em comparação com a avaliação da v2, sendo que os números entre parênteses demonstram a variação nos números. As etiquetas *expl:impers*, *expl:pass* e *expl:pv* não têm variação em relação à v2, visto que nesta versão as etiquetas não existiam. A tabela está organizada por ordem decrescente da melhoria no LAS.

Do ponto de vista linguístico, pode ser mais interessante comparar os

REL	**#**	**HIT**		**LAS**	
expl:impers	16	75,00%	(N/A)	75,00%	(N/A)
expl:pass	45	82,22%	(N/A)	82,22%	(N/A)
expl:pv	40	90,00%	(N/A)	87,50%	(N/A)
ccomp	26	73,08%	(+7,56%)	73,08%	(+11,01%)
xcomp	111	85,59%	(+7,41%)	84,68%	(+7,41%)
aux	13	84,62%	(+6,05%)	84,62%	(+6,05%)
obl	576	85,59%	(+5,63%)	76,39%	(+1,92%)
parataxis	82	67,07%	(+2,87%)	58,54%	(+2,98%)
conj	425	90,82%	(+2,54%)	62,35%	(+0,51%)
nummod	234	97,44%	(+2,28%)	96,58%	(+4,24%)
appos	111	83,78%	(+2,04%)	60,36%	(+0,36%)
amod	670	95,82%	(+1,68%)	92,84%	(+2,75%)
aux:pass	177	97,74%	(+1,67%)	97,74%	(+1,67%)
advmod	306	96,08%	(+1,27%)	79,74%	(-0,45%)
flat:name	222	91,89%	(+0,81%)	90,99%	(+1,18%)
fixed	202	86,63%	(+0,49%)	85,64%	(-0,50%)
punct	1420	99,86%	(+0,30%)	86,55%	(+1,82%)
case	1893	99,52%	(+0,09%)	98,68%	(-0,12%)
advcl	192	83,33%	(+0,09%)	68,23%	(-3,66%)
det	1770	99,66%	(+0,00%)	99,49%	(+0,22%)
flat	3	100,00%	(+0,00%)	100,00%	(+0,00%)
flat:foreign	2	100,00%	(+0,00%)	100,00%	(+0,00%)
goeswith	1	0,00%	(+0,00%)	0,00%	(+0,00%)
csubj	7	71,43%	(+0,00%)	71,43%	(+0,00%)
root	445	94,83%	(-0,02%)	94,83%	(-0,02%)
cc	327	96,94%	(-0,35%)	92,66%	(-0,11%)
nsubj:pass	172	86,05%	(-0,38%)	85,47%	(-0,96%)
mark	155	87,10%	(-0,56%)	83,87%	(+0,10%)
obl:arg	79	62,03%	(-0,79%)	62,03%	(-0,79%)
nmod	1331	91,81%	(-0,94%)	86,70%	(-1,32%)
obj	283	93,64%	(-1,03%)	92,93%	(-0,80%)
nsubj	351	89,46%	(-1,29%)	86,32%	(-2,70%)
cop	127	95,28%	(-1,62%)	90,55%	(-2,47%)
acl	217	89,86%	(-3,10%)	73,27%	(-4,66%)
acl:relcl	91	91,21%	(-4,44%)	64,84%	(+2,88%)
obl:agent	52	94,23%	(-5,77%)	94,23%	(-5,77%)

Tabela 6.14: Avaliação de REL do PetroGold v3 em comparação com a v2

	LEMMA	UPOS	UAS	LAS	CLAS
v2 simplif.	98,54	98,40	90,66	88,82	83,48
v3 simplif.	**98,60**	**98,63**	**92,02**	**90,22**	**85,61**
bosque-ud 2.11	96,95	96,52	84,78	80,86	73,04

Tabela 6.15: Avaliação intrínseca de versões simplificadas do PetroGold e Bosque-UD 2.11

corpora quando têm as mesmas etiquetas no seu *tagset*, pois assim garantimos que a variação nos números refletem apenas o grau de revisão dos fenômenos linguísticos. Criamos uma versão simplificada do PetroGold, tanto na v2 quanto na v3, convertendo *obl:arg* em *obl* e as variações do *expl* (pass, pv e impers) em somente *expl*. Essas conversões nos permitem também comparar o treebank com o Bosque-UD, um dos principais treebanks do projeto UD para português e que não contém essas etiquetas[12].

A tabela 6.15 mostra os números da avaliação intrínseca das versões simplificadas do PetroGold e do Bosque-UD 2.11. Em comparação com o Bosque-UD, o PetroGold v3 atinge um resultado de CLAS 17,20% melhor (12,57 pontos percentuais), indicando que, utilizando essa estratégia de avaliação, o PetroGold se consolida como o *dataset* cuja anotação pode estar mais consistente, pois proporciona o melhor aprendizado automático entre os *datasets* sendo avaliados. Em comparação com a v2 do PetroGold, a v3 alcança resultados até 2,55% melhores (2,13 pontos percentuais). Em resumo, e de forma simplificada, todas as revisões linguísticas discutidas nesta dissertação, quando aplicadas em um corpus já considerado padrão ouro, foram capazes de alavancar as métricas de avaliação intrínseca em 2,13 pontos percentuais, garantindo todas as condições de comparabilidade – mesmas frases nas partições de treino, desenvolvimento e teste e mesmo *tagset* –, resultados muito superiores ao 0,15 ponto percentual de diferença observado quando comparamos as versões completas do treebank.

[12]O Bosque-UD utilizado nesta seção está na versão 2.11, correspondendo ao número da versão UD do PetroGold, garantindo que ambos os recursos foram anotados utilizando as mesmas diretivas gramaticais, e o treinamento foi realizado utilizando as mesmas configurações utilizadas para o PetroGold.

7
Considerações finais

Este trabalho teve como objetivo apresentar o processo de desenvolvimento do PetroGold, um treebank padrão ouro do domínio do petróleo. Treebanks específicos de domínio são escassos e recebem relevância especial no contexto recente, em que o PLN estatístico tem se beneficiado de recursos linguísticos de qualidade para alimentar o aprendizado automático.

Tendo em vista que o desenvolvimento do treebank tem como objetivo subsidiar a construção de um sistema de buscas para o projeto Petrolês, esclarecemos desde o início que o recurso estava sendo desenvolvimento especificamente para o PLN. Isso significa que, embora útil para outras áreas, como os estudos linguístico-descritivos, a anotação do recurso considerou sobretudo a capacidade que o treebank tem de treinar bons modelos de aprendizado de máquina para anotação de dependências sintáticas para o domínio do petróleo e para o gênero acadêmico.

Notamos também alguns pressupostos importantes que consideramos durante a tarefa de anotação do treebank. Por um lado, situamos a anotação linguística como um processo interpretativo e que exige consenso para agrupar fenômenos distintos sob um mesmo rótulo segundo critérios específicos (LEECH, 1997; ARCHER, 2012; FREITAS, 2022). Por outro lado, para tomar as decisões de anotação mais adequadas do ponto de vista linguístico, buscamos sempre que possível utilizar as premissas da linguística empírica (SAMPSON, 2002; SAMPSON, 2003) e da linguística probabilística (MANNING; SCHUTZE, 1999; MANNING, 2003).

Optamos por dividir a apresentação do PetroGold em duas etapas: a etapa da construção e a etapa da avaliação do recurso.

O capítulo sobre a construção envolveu todas as discussões relativas à anotação e à revisão do corpus. Apresentamos as características do corpus e nos detivemos sobre três questões linguísticas – os argumentos verbais introduzidos por preposição, as expressões multipalavras e o pronome "se". Consultamos as diretivas do projeto Universal Dependencies, gramáticas do português, estudos linguístico-computacionais e o próprio corpus para investigar quais as melhores anotações para os fenômenos e explicamos a metodologia adotada para executá-las, assim como os resultados dessa anotação do ponto de vista linguístico. Por fim, olhamos para os métodos computacionais utilizados na revisão do corpus – regras linguísticas, *n-grams* inconsistentes e IAD – e discutimos a contribuição de cada um deles na construção do treebank,

momento em que concluímos que o método IAD é ao mesmo tempo aquele que mais detecta erros e aquele cujos erros detectados retornam menos falsos positivos proporcionalmente.

O capítulo sobre a avaliação do PetroGold utilizou a estratégia da avaliação intrínseca para verificar a qualidade do treebank, considerando que um treebank é um bom recurso para o PLN quando proporciona um bom aprendizado automático de dependências sintáticas. Primeiro, verificamos as categorias morfossintáticas mais difíceis para o aprendizado automático e tentamos identificar pontos no corpus que possam ter fomentado a dificuldade, tendo como objetivo corrigi-las. Concluímos que 10,8% dos supostos erros cometidos por um anotador automático não são de fato erros cometidos pelo algoritmo, mas falhas no padrão ouro que precisaram ser corrigidas. Dos erros de fato do anotador automático, a maioria parece não ser explicável linguisticamente, enquanto um pouco menos da outra metade dos erros é decorrente de ambiguidade estrutural e a minoria dos erros é fruto da falta de exemplos dos fenômenos linguísticos no treebank.

Então, observamos o impacto das correções das questões linguísticas tratadas nesta terceira versão do treebank nas métricas da avaliação intrínseca, e identificamos que, a despeito de introduzir uma nova etiqueta, a revisão dos argumentos do verbo introduzidos por preposição melhorou as métricas de avaliação intrínseca em comparação ao mesmo corpus quando sem essas revisões. A revisão das expressões multipalavras, por sua vez, diminuiu os números da avaliação, mas vimos também que anotar a classe gramatical das palavras de uma MWE com a classe gramatical da expressão como um todo melhora os números da anotação sintática. Já em relação ao pronome "se", vimos, como era esperado, um decréscimo nas métricas da avaliação quando introduzimos as três novas etiquetas, especificando o tipo de pronome.

Por fim, comparamos a nova versão do corpus, o PetroGold v3 (já disponível na página do projeto Petrolês e no projeto Universal Dependencies), com a versão anterior utilizando a estratégia da avaliação intrínseca para indicar o grau de consistência interna do treebank. Vimos que o PetroGold v3 atinge números melhores que a v2 em todas as métricas, indicando que a melhoria na qualidade de um corpus que passou por uma nova leva de revisões se reflete também nos resultados da avaliação, e vimos que a versão do corpus disponibilizada no projeto Petrolês leva a números mais altos que a versão do projeto UD exceto por uma das métricas, apontando diferença considerável nos resultados quando mudamos as frases que compõem as partições de treinamento e teste. Do mesmo modo, quando igualamos o *tagset* das versões 2 e 3 do PetroGold, simplificando as etiquetas, vimos uma melhora ainda maior

nos números da última versão em comparação com a anterior.

Algumas etapas do projeto Petrolês que são posteriores à criação do PetroGold já foram finalizadas ou estão em andamento: (1) treinamos um modelo de anotação morfossintática utilizando o PetroGold e as melhores ferramentas disponíveis; (2) anotamos um conjunto maior de textos utilizando esse modelo, e (3) anotamos este novo corpus com entidades mencionadas do domínio do petróleo, fomentando assim a criação do sistema de buscas semanticamente orientado. As três etapas só foram possíveis ou foram consideravelmente facilitadas pela finalização do PetroGold, por um lado, e por conta das ferramentas que foram desenvolvidas durante a sua construção, como a ET, o ambiente de busca, edição e avaliação de corpora anotados. Ambos os recursos estão publicamente disponíveis como contribuição deste trabalho.

Embora durante toda a dissertação tenhamos treinado diversos modelos de anotação morfossintática para realizar a avaliação intrínseca do corpus, temos ciência de que a metodologia empregada para o treinamento e os programas utilizados não são os melhores, mais adequados ou atualizados. Uma das lacunas deste trabalho que pode ser compensada em um trabalho futuro é a utilização de outras ferramentas e algoritmos para a realização dos testes de avaliação intrínseca, seja para obter resultados mais verdadeiros de quantos pontos percentuais o corpus alcança nas métricas de avaliação, seja para comparar a anotação de fenômenos linguísticos específicos realizada por cada ferramenta quando utilizando o mesmo material de treino.

Outra lacuna que podemos notar neste trabalho se refere às tentativas de entender o que os modelos de IA têm aprendido bem e os fenômenos em que têm maior dificuldade. Utilizando as ferramentas como o Julgamento (módulo da ET), conseguimos verificar os resultados de um anotador automático e compará-lo com o padrão ouro, como fizemos em várias partes da dissertação. No entanto, entender por que o modelo anotou de uma forma e não de outra ainda é um ponto com poucas conclusões neste trabalho. A pesquisa e a aplicação de métodos para tornar a IA explicável poderiam ser benéficas a um estudo como este.

Além disso, em vários momentos do trabalho, realizei categorizações dos fenômenos linguísticos ou mesmo dos erros cometidos pelos anotadores automáticos sozinho. O processo foi acompanhado de perto por muitas discussões com minha orientadora, porém sou o responsável final por muitas das classificações realizadas. Embora haja um lado positivo dessa prática (que também foi utilizada por Manning (2011), por exemplo), pois assim garanto consistência nas minhas próprias decisões (as quais estão documentadas neste trabalho), de modo geral a prática não é bem avaliada, pois uma classificação sem discus-

são intensa e sem ser testada quanto à concordância por diferentes anotadores tende a ser inconfiável.

O PetroGold é um recurso público que já está sendo utilizado como subsídio para diversas tarefas de PLN no projeto Petrolês. Futuramente, gostaríamos de avaliar a qualidade do recurso não só para o aprendizado de informação morfossintática, como fizemos, mas também para essas outras tarefas subsequentes à anotação morfossintática (em um *pipeline* tradicional). E embora não tenha sido desenvolvido especificamente com este propósito, seria interessante também observar como o corpus se comporta em estudos linguístico-descritivos, seja sobre a língua portuguesa, de modo geral, seja sobre o gênero acadêmico.

8
Referências bibliográficas

AFONSO, S. Avaliação do grau de concordância entre anotadores: análise e discussão dos resultados do processo de re-revisão. 2004. Citado na página 42.

AFONSO, S. A floresta sintá(c)tica como recurso. 2004. Disponível em: <https://www.linguateca.pt/documentos/Afonso2004Recurso.pdf>. Citado 2 vezes nas páginas 29 e 34.

AFONSO, S.; BICK, E.; HABER, R.; SANTOS, D. Floresta sintá(c)tica: um treebank para o português. In: **Actas do XVII Encontro Nacional da Associação Portuguesa de Linguística (APL 2001)**. Lisboa, Portugal: APL, 2001. p. 533–545. Citado 4 vezes nas páginas 29, 32, 33 e 34.

AFONSO, S.; BICK, E.; HABER, R.; SANTOS, D. Floresta sintá (c) tica: a treebank for Portuguese. In: RODRIGUES, M. G.; ARAUJO, C. P. S. (Ed.). **Proceedings of the Third International Conference on Language Resources and Evaluation (LREC 2002)**. [S.l.], 2002. Citado 4 vezes nas páginas 10, 15, 33 e 34.

ALUÍSIO, S.; PELIZZONI, J.; MARCHI, A. R.; OLIVEIRA, L. de; MANENTI, R.; MARQUIAFÁVEL, V. An account of the challenge of tagging a reference corpus for brazilian portuguese. In: SPRINGER. **Computational Processing of the Portuguese Language: 6th International Workshop, PROPOR 2003 Faro, Portugal, June 26–27, 2003 Proceedings**. [S.l.], 2003. p. 110–117. Citado na página 40.

ANTHONY, L. Antconc: design and development of a freeware corpus analysis toolkit for the technical writing classroom. In: IEEE. **IPCC 2005. Proceedings. International Professional Communication Conference, 2005.** [S.l.], 2005. p. 729–737. Citado na página 62.

ARCHER, D. Corpus annotation: a welcome addition or an interpretation too far. **Outposts of Historical Corpus Linguistics: From the Helsinki Corpus to a Proliferation of Resources**, VARIENG Helsinki, v. 10, 2012. Citado 3 vezes nas páginas 23, 24 e 154.

ARTSTEIN, R. Inter-annotator agreement. In: **Handbook of linguistic annotation**. [S.l.]: Springer, 2017. p. 297–313. Citado 2 vezes nas páginas 42 e 48.

AZEREDO, J. C. d. Fundamentos de gramática do português. **Rio de Janeiro: Jorge Zahar**, 2000. Citado na página 121.

BABARCZY, A.; CARROLL, J.; SAMPSON, G. Definitional, personal, and mechanical constraints on part of speech annotation performance. **Natural Language Engineering**, Cambridge University Press, v. 12, n. 1, p. 77–90, 2006. Citado na página 43.

BAGNO, M. **Gramática pedagógica do português brasileiro**. [S.l.]: Parábola Ed., 2012. Citado 4 vezes nas páginas 79, 93, 109 e 111.

BAIA, J.; PRATES, A.; CLARO, D. 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.], 2020. p. 193–200. Citado 2 vezes nas páginas 45 e 46.

BECHARA, E. **Moderna gramática portuguesa**. [S.l.]: Nova Fronteira, 2012. Citado 6 vezes nas páginas 74, 79, 80, 109, 114 e 121.

BICK, E. **The parsing system palavras: Automatic grammatical analysis of Portuguese in a constraint grammar famework**. [S.l.]: Aarhus Universitetsforlag, 2000. Citado na página 109.

BICK, E. PALAVRAS, a constraint grammar-based parsing system for Portuguese. **Working with Portuguese corpora**, Bloomsbury Academic, p. 279–302, 2014. Citado 2 vezes nas páginas 32 e 94.

BICK, E.; AFONSO, S.; SANTOS, D.; MARCHI, R. Floresta sintá (c) tica: Ficçao ou realidade. In: **Avaliação Conjunta, Um novo paradigma no processamento computacional da língua portuguesa**. [S.l.]: IST Press, 2007. p. 291–300. Citado 3 vezes nas páginas 29, 30 e 34.

BIRD, S. NLTK: the natural language toolkit. In: **Proceedings of the COLING/ACL 2006 Interactive Presentation Sessions**. [S.l.: s.n.], 2006. p. 69–72. Citado na página 97.

BOHNET, B.; MCDONALD, R.; SIMÕES, G.; ANDOR, D.; PITLER, E.; MAYNEZ, J. Morphosyntactic tagging with a meta-BiLSTM model over context sensitive token encodings. In: **Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**. Melbourne, Australia: Association for Computational Linguistics, 2018. p. 2642–2652. Disponível em: <https://aclanthology.org/P18-1246>. Citado na página 15.

BOUMA, G.; HAJIC, J.; HAUG, D.; NIVRE, J.; SOLBERG, P. E.; ØVRELID, L. Expletives in universal dependency treebanks. In: **Proceedings of the Second Workshop on Universal Dependencies (UDW 2018)**. [S.l.: s.n.], 2018. p. 18–26. Citado na página 109.

BOYD, A.; DICKINSON, M.; MEURERS, W. D. On detecting errors in dependency treebanks. **Research on Language and Computation**, Springer, v. 6, n. 2, p. 113–137, 2008. Citado 4 vezes nas páginas 10, 37, 38 e 58.

BUCHHOLZ, S.; MARSI, E. CoNLL-X shared task on multilingual dependency parsing. In: **Proceedings of the tenth conference on computational natural language learning (CoNLL-X)**. [S.l.: s.n.], 2006. p. 149–164. Citado na página 50.

CÂMARA, J. M. **Dicionário de lingüística e gramática, referente à língua portuguesa**. [S.l.]: Editora Vozes, 1978. Citado na página 91.

CANÇADO, M.; AMARAL, L. Representação lexical de verbos incoativos e causativos no português brasileiro. **Revista da ABRALIN**, v. 9, n. 2, p. 123–147, 2010. Citado na página 113.

CARTER, D. The TreeBanker: A tool for supervised training of parsed corpora. 1997. Citado na página 31.

CASTILHO, A. d. Nova gramática da língua portuguesa. **São Paulo: Contexto**, 2010. Citado na página 79.

CAVALCANTI, T.; SILVEIRA, A.; DE SOUZA, E.; FREITAS, C. Os limites da palavra e da sentença no processamento automático de textos. **Revista Brasileira de Iniciação Científica**, v. 8, p. e021033–e021033, 2021. Citado 2 vezes nas páginas 47 e 72.

CHAMBERS, J. **Sociolinguistic Theory: Linguistic Variation and Its Social Significance**. Wiley, 2003. (Language in Society). ISBN 9780631228820. Disponível em: <https://books.google.com.br/books?id=FZ0K9z2PRDwC>. Citado na página 25.

CLEVERLEY, P. H.; BURNETT, S. The best of both worlds: highlighting the synergies of combining manual and automatic knowledge organization methods to improve information search and discovery. **Knowledge organization**, Ergon Verlag, v. 42, n. 6, 2015. Citado na página 14.

ČMEJREK, M.; HAJIČ, J.; KUBOŇ, V. Prague Czech-English dependency treebank: Syntactically annotated resources for machine translation. In: EAMT. **Proceedings of EAMT 10th Annual Conference**. [S.l.], 2004. Citado na página 30.

CORDEIRO, F. C. Petrolês-como construir um corpus especializado em óleo e gás em português. **PUC-Rio, Rio de Janeiro, RJ-Brasil: PUC-Rio**, 2020. Citado na página 71.

CUNHA, C.; CINTRA, L. **Nova gramática do português contemporâneo**. [S.l.]: LEXIKON Editora Digital ltda, 2016. Citado na página 113.

DE MARNEFFE, M.-C.; GRIONI, M.; KANERVA, J.; GINTER, F. Assessing the annotation consistency of the universal dependencies corpora. In: **Proceedings of the Fourth International Conference on Dependency Linguistics (Depling 2017)**. [S.l.: s.n.], 2017. p. 108–115. Citado 2 vezes nas páginas 38 e 58.

DE SOUZA, E.; CAVALCANTI, T.; SILVEIRA, A.; EVELYN, W.; FREITAS, C. Diretivas e documentação de anotação UD em português (e para língua portuguesa). 2020. Disponível em: <http://comcorhd.letras.puc-rio.br/Documenta-o-UD-PT>. Citado na página 47.

DE SOUZA, E.; FREITAS, C. Et: uma estação de trabalho para revisão, edição e avaliação de corpora anotados morfossintaticamente. In: **WORKSHOP DE INICIAÇÃO CIENTÍFICA EM TECNOLOGIA DA INFORMAÇÃO E DA LINGUAGEM HUMANA (TILic 2019), VI**. [S.l.: s.n.], 2019. p. 15–18. Citado na página 61.

DE SOUZA, E.; FREITAS, C. (Re)começando a discutir as locuções verbais. **Anais da VI Jornada de Descrição do Português (JDP), Salvador, Brasil**, 2019. Citado na página 149.

DE SOUZA, E.; FREITAS, C. 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 and Punta Cana, Dominican Republic: Association for Computational Linguistics, 2021. p. 35–41. Disponível em: <https://aclanthology.org/2021.emnlp-demo.5>. Citado 2 vezes nas páginas 61 e 62.

DE SOUZA, E.; FREITAS, C. Polishing the gold–how much revision do we need in treebanks? In: **Procedings of the Universal Dependencies Brazilian Festival**. [S.l.: s.n.], 2022. p. 1–11. Citado 4 vezes nas páginas 14, 72, 135 e 142.

DE SOUZA, E.; FREITAS, C. Still on arguments and adjuncts: the status of the indirect object and the adverbial adjunct relations in Universal Dependencies for Portuguese. In: **Proceedings of the Universal Dependencies Brazilian Festival**. Fortaleza, Brazil: Association for Computational Linguistics, 2022. p. 1–10. Disponível em: <https://aclanthology.org/2022.udfestbr-1.5>. Citado 2 vezes nas páginas 73 e 142.

DE SOUZA, E.; SILVEIRA, A.; CAVALCANTI, T.; CASTRO, M. C.; FREITAS, C. Documentação da anotação morfossintática do PetroGold. 2021. Disponível em: <https://www.researchgate.net/publication/365597977_Documentacao_da_anotacao_morfossintatica_do_PetroGold>. Citado na página 47.

DE SOUZA, E.; SILVEIRA, A.; CAVALCANTI, T.; CASTRO, M. C.; FREITAS, C. PetroGold–Corpus padrão ouro para o domínio do petróleo. In: SBC. **Anais do XIII Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana**. [S.l.], 2021. p. 29–38. Citado 3 vezes nas páginas 14, 71 e 135.

DICKINSON, M. **Error detection and correction in annotated corpora**. [S.l.]: The Ohio State University, 2005. Citado 3 vezes nas páginas 35, 36 e 37.

DURAN, M. S.; ALUÍSIO, S. Propbank-br: a brazilian treebank annotated with semantic role labels. In: **Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)**. [S.l.: s.n.], 2012. p. 1862–1867. Citado na página 15.

DURAN, M. S.; SCARTON, C.; ALUÍSIO, S.; RAMISCH, C. Identifying Pronominal Verbs: Towards Automatic Disambiguation of the Clitic 'se' in Portuguese. In: **Proceedings of the 9th Workshop on Multiword Expressions**. [S.l.: s.n.], 2013. p. 93–100. Citado 3 vezes nas páginas 108, 113 e 114.

ESKIN, E. Automatic corpus correction with anomaly detection. In: **Proceedings of the First Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-00)**. [S.l.: s.n.], 2000. p. 148–153. Citado na página 36.

FONSECA, E. B.; ANTONITSCH, A.; COLLOVINI, S.; AMARAL, D.; VIEIRA, R.; FIGUEIRA, A. Summ-it++: an enriched version of the summ-it corpus. In: **Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)**. [S.l.: s.n.], 2016. p. 2047–2051. Citado na página 15.

FREITAS, C. **Linguística Computacional**. [S.l.]: Editora Parábola, 2022. Citado 10 vezes nas páginas 16, 17, 21, 22, 23, 24, 41, 44, 60 e 154.

FREITAS, C.; CARVALHO, P.; OLIVEIRA, H. G.; MOTA, C.; SANTOS, D. Second HAREM: advancing the state of the art of named entity recognition in Portuguese. In: EUROPEAN LANGUAGE RESOURCES ASSOCIATION. **Proceedings of the International Conference on Language Resources and Evaluation (LREC 2010)**. [S.l.], 2010. Citado na página 15.

FREITAS, C.; DE 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. Citado na página 108.

FREITAS, C.; DE SOUZA, E. A study on methods for revising dependency treebanks: In search of gold. **Language Resources and Evaluation**, Springer, (no prelo), 2023. Citado 6 vezes nas páginas 35, 39, 54, 58, 59 e 125.

FREITAS, C.; ROCHA, P.; BICK, E. Um mundo novo na Floresta Sintá (c) tica–o treebank do Português. **Calidoscópio**, v. 6, n. 3, p. 142–148, 2008. Citado 3 vezes nas páginas 15, 73 e 89.

FREITAS, C.; TRUGO, L. F.; CHALUB, F.; PAULINO-PASSOS, G.; RADEMAKER, A. Tagsets and datasets: some experiments based on Portuguese language. In: SPRINGER. **International Conference on Computational Processing of the Portuguese Language**. [S.l.], 2018. p. 459–469. Citado na página 40.

GALVES, C. The tycho brahe corpus of historical portuguese: Methodology and results. **Linguistic Variation**, John Benjamins Publishing Company Amsterdam/Philadelphia, v. 18, n. 1, p. 49–73, 2018. Citado na página 30.

GARSIDE, R.; LEECH, G.; VÁRADI, T. Manual of Information for the Lancaster Parsed Corpus. **Bergen, Norway: Norwegian Computing Center for the Humanities**, 1995. Citado na página 30.

GERDES, K. Collaborative dependency annotation. In: **Proceedings of the second international conference on dependency linguistics (DepLing 2013)**. [S.l.: s.n.], 2013. p. 88–97. Citado na página 62.

GOMES, D. da S. M.; CORDEIRO, F. C.; CONSOLI, B. S.; SANTOS, N. L.; MOREIRA, V. P.; VIEIRA, R.; MORAES, S.; EVSUKOFF, A. G. Portuguese word embeddings for the oil and gas industry: Development and evaluation. **Computers in Industry**, v. 124, p. 103347, 2021. ISSN 0166-3615. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0166361520305819>. Citado na página 14.

HAJIČ, J. Building a syntactically annotated corpus: The prague dependency treebank. **Issues of Valency and Meaning. Studies in Honour of Jarmila Panevová**, p. 106–132, 1998. Citado na página 31.

HALTEREN, H. van. The detection of inconsistency in manually tagged text. In: **Proceedings of the COLING-2000 Workshop on Linguistically Interpreted Corpora**. [S.l.: s.n.], 2000. p. 48–55. Citado na página 39.

HARDIE, A. Cqpweb—combining power, flexibility and usability in a corpus analysis tool. **International journal of corpus linguistics**, John Benjamins, v. 17, n. 3, p. 380–409, 2012. Citado na página 62.

HEINECKE, J. Conllueditor: a fully graphical editor for universal dependencies treebank files. In: **Proceedings of the Third Workshop on Universal Dependencies (UDW, SyntaxFest 2019)**. [S.l.: s.n.], 2019. p. 87–93. Citado na página 62.

HINRICHS, E. W.; BARTELS, J.; KAWATA, Y.; KORDONI, V.; TELLJOHANN, H. The Tübingen treebanks for spoken german, english, and japanese. In: **Verbmobil: Foundations of speech-to-speech translation**. [S.l.]: Springer, 2000. p. 550–574. Citado na página 36.

HOCKENMAIER, J.; STEEDMAN, M. Acquiring Compact Lexicalized Grammars from a Cleaner Treebank. In: **Proceedings of the Third International Conference on Language Resources and Evaluation**. [S.l.: s.n.], 2002. p. 1974–1981. Citado na página 30.

JÄRVINEN, T. Bank of English and beyond. In: **Treebanks**. [S.l.]: Springer, 2003. p. 43–59. Citado na página 35.

JONES, K. S. Natural language processing: a historical review. **Current issues in computational linguistics: in honour of Don Walker**, Springer, p. 3–16, 1994. Citado 2 vezes nas páginas 20 e 21.

JOOS, M. Description of language design. **The Journal of the Acoustical Society of America**, Acoustical Society of America, v. 22, n. 6, p. 701–707, 1950. Citado na página 25.

KARLSSON, F.; VOUTILAINEN, A.; HEIKKILAE, J.; ANTTILA, A. **Constraint Grammar: a language-independent system for parsing unrestricted text**. [S.l.]: Walter de Gruyter, 2011. v. 4. Citado na página 32.

KAY, M. Acl lifetime achievement award: A life of language. **Computational Linguistics**, v. 31, n. 4, p. 425–438, 2005. Citado na página 16.

KILGARRIFF, A. Gold standard datasets for evaluating word sense disambiguation programs. **Computer Speech & Language**, Elsevier, v. 12, n. 4, p. 453–472, 1998. Citado na página 35.

KUCERA, H.; FRANCIS, W. N. **Computational Analysis of Present-day American English**. [S.l.]: Brown University Press, 1967. Citado na página 29.

LEECH, G. Introducing corpus annotation. In: _____. **Corpus Annotation: Linguistic Information from Computer Text Corpora**. Longman, 1997. (Pearson Education). ISBN 9780582298378. Disponível em: <https://books.google.com.br/books?id=8ewKAQAAMAAJ>. Citado 2 vezes nas páginas 21 e 154.

LOPES, L.; DURAN, M. S.; FERNANDES, P.; PARDO, T. Portilexicon-ud: a portuguese lexical resource according to universal dependencies model. In: **Proceedings of the Thirteenth Language Resources and Evaluation Conference**. [S.l.: s.n.], 2022. p. 6635–6643. Citado na página 55.

LOPES, L.; DURAN, M. S.; PARDO, T. A. Universal dependencies-based pos tagging refinement through linguistic resources. In: SPRINGER. **Brazilian Conference on Intelligent Systems**. [S.l.], 2021. p. 601–615. Citado na página 94.

MANNING, C.; SCHUTZE, H. **Foundations of statistical natural language processing**. [S.l.]: MIT press, 1999. Citado 7 vezes nas páginas 24, 25, 27, 37, 91, 97 e 154.

MANNING, C. D. Probabilistic syntax. **Probabilistic linguistics**, MIT press Cambridge, MA, v. 289341, 2003. Citado 7 vezes nas páginas 24, 25, 26, 27, 80, 89 e 154.

MANNING, C. D. Part-of-speech tagging from 97% to 100%: is it time for some linguistics? In: SPRINGER. **International conference on intelligent text processing and computational linguistics**. [S.l.], 2011. p. 171–189. Citado 6 vezes nas páginas 39, 40, 44, 127, 141 e 156.

MARCUS, M. P.; SANTORINI, B.; MARCINKIEWICZ, M. A. Building a large annotated corpus of english: The penn treebank. **Comput. Linguistics**, v. 19, n. 2, p. 313–330, 1993. Citado na página 30.

MARNEFFE, M.-C. D.; MANNING, C. D.; NIVRE, J.; ZEMAN, D. Universal dependencies. **Computational linguistics**, MIT Press One Rogers Street, Cambridge, MA 02142-1209, USA journals-info . . . , v. 47, n. 2, p. 255–308, 2021. Citado 6 vezes nas páginas 10, 31, 49, 50, 81 e 82.

MCENERY, T.; HARDIE, A. **Corpus Linguistics: Method, Theory and Practice**. Cambridge University Press, 2011. (Cambridge Textbooks in Linguistics). ISBN 9781139502443. Disponível em: <https://books.google.com.br/books?id=3j3Wn_ZT1qwC>. Citado 3 vezes nas páginas 80, 91 e 94.

NETO, J. B. Morfologia: conceitos e métodos. **Colóquios linguísticos e literários: enfoques epistemológicos, metodológicos e descritivos. Teresina: Edufpi**, p. 53–72, 2011. Citado 2 vezes nas páginas 76 e 78.

NEVES, M. H. de M. **Gramática de usos do português**. [S.l.]: Unesp, 2000. Citado na página 93.

NIVRE, J. Treebanks. In: KYTÖ, M.; LÜDELING, A. (Ed.). **Corpus Linguistics: An International Handbook**. Mouton de Gruyter, 2008. p. 225–241. Disponível em: <http://stp.lingfil.uu.se/~nivre/docs/hsk.pdf>. Citado 5 vezes nas páginas 10, 29, 30, 31 e 32.

NIVRE, J.; FANG, C.-T. Universal dependency evaluation. In: **Proceedings of the NoDaLiDa 2017 Workshop on Universal Dependencies (UDW 2017)**. [S.l.: s.n.], 2017. p. 86–95. Citado na página 61.

OEPEN, S.; FLICKINGER, D.; BOND, F. Towards Holistic Grammar Engineering and Testing–Grafting Treebank Maintenance into the Grammar Revision Cycle. In: CITESEER. **IN PROCEEDINGS OF THE IJCNLP WORKSHOP BEYOND SHALLOW ANALYSIS**. [S.l.], 2004. Citado na página 36.

OEPEN, S.; TOUTANOVA, K.; SHIEBER, S. M.; MANNING, C. D.; FLICKINGER, D.; BRANTS, T. The LinGO Redwoods treebank: Motivation and preliminary applications. In: **COLING 2002: The 17th International Conference on Computational Linguistics: Project Notes**. [S.l.: s.n.], 2002. Citado na página 30.

OLIVA, K. The Possibilities of Automatic Detection/Correction of Errors in Tagged Corpora: A Pilot Study on a German Corpus. In: SPRINGER. **International Conference on Text, Speech and Dialogue**. [S.l.], 2001. p. 39–46. Citado na página 36.

OLIVEIRA, C.; GARRÃO, M.; AMARAL, L. A. et al. Recognising complex prepositions Prep+ N+ Prep as negative patterns in automatic term extraction from texts. In: **TIL'2003-1º Workshop em Tecnologia da Informação e Linguagem Humana, evento associado ao XVI Brazilian Symposium on Computer Graphics and Image Processing-(SIBGRAPI)**. [S.l.: s.n.], 2009. Citado na página 97.

OLIVEIRA, C.; NOGUEIRA, C.; GARRAO, M. Locution or collocation: comparing linguistic and statistical methods for recognising complex prepositions. In: **Anais do 2º Workshop em Tecnologia da Informação e da Linguagem Humana**. [S.l.: s.n.], 2004. Citado 3 vezes nas páginas 94, 96 e 97.

PRZEPIÓRKOWSKI, A.; PATEJUK, A. Arguments and adjuncts in Universal Dependencies. In: **Proceedings of the 27th international conference on computational linguistics**. [S.l.: s.n.], 2018. p. 3837–3852. Citado 3 vezes nas páginas 81, 83 e 86.

QI, P.; ZHANG, Y.; ZHANG, Y.; BOLTON, J.; MANNING, C. D. Stanza: A Python natural language processing toolkit for many human languages. In: **Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations**. [s.n.], 2020. Disponível em: <https://nlp.stanford.edu/pubs/qi2020stanza.pdf>. Citado 4 vezes nas páginas 48, 59, 84 e 117.

RADEMAKER, A.; CHALUB, F.; REAL, L.; FREITAS, C.; BICK, E.; PAIVA, V. D. Universal dependencies for Portuguese. In: **Proceedings of the Fourth International Conference on Dependency Linguistics (Depling 2017)**. [S.l.: s.n.], 2017. p. 197–206. Citado 5 vezes nas páginas 48, 59, 61, 84 e 135.

RAMISCH, C. A generic framework for multiword expressions treatment: from acquisition to applications. In: **Proceedings of the ACL 2012 Student Research Workshop**. Jeju, Republic of Korea: ACL, 2012. <https://aclweb.org/anthology/W12-3311>. Citado na página 90.

ROCHA, P. A.; SANTOS, D. CETEMPúblico: Um corpus de grandes dimensões de linguagem jornalística portuguesa. In: NUNES, M. d. G. V. (Ed.). **V Encontro para o processamento computacional da língua portuguesa**. [S.l.], 2000. Citado na página 29.

ROCIO, V.; ALVES, M. A.; LOPES, J. G.; XAVIER, M. F.; VICENTE, G. Automated creation of a Medieval Portuguese partial treebank. In: **Treebanks**. [S.l.]: Springer, 2003. p. 211–227. Citado na página 30.

SAMPSON, G. **English for the Computer: The SUSANNE Corpus and Analytic Scheme**. Clarendon Press, 1995. ISBN 9780198240235. Disponível em: <https://books.google.com.br/books?id=odFt5ZHBri4C>. Citado 3 vezes nas páginas 23, 25 e 29.

SAMPSON, G. **Empirical linguistics**. [S.l.]: A&C Black, 2002. Citado 4 vezes nas páginas 26, 27, 77 e 154.

SAMPSON, G. Thoughts on two decades of drawing trees. In: **Treebanks**. [S.l.]: Springer, 2003. p. 23–41. Citado 2 vezes nas páginas 29 e 154.

SAMPSON, G.; BABARCZY, A. Definitional and human constraints on structural annotation of English. **Natural Language Engineering**, Cambridge University Press, v. 14, n. 4, p. 471–494, 2008. Citado 3 vezes nas páginas 40, 41 e 42.

SANTOS, D. **Avaliação conjunta: um novo paradigma no processamento computacional da língua portuguesa**. [S.l.]: Instituto superior técnico Press, 2007. Citado na página 44.

SANTOS, D. Corporizando algumas questões. **Avanços da Lingüística de Corpus no Brasil**, Editora Humanitas, 2008. Citado na página 22.

SANTOS, D.; BICK, E. Providing internet access to portuguese corpora: the ac/dc project. In: **Maria Gavrilidou; George Carayannis; Stella Markantonatou; Stelios Piperidis; Gregory Stainhauer (ed) Proceedings of the Second International Conference on Language Resources and Evaluation (LREC 2000)(Athens 31 May-2 June 2000)**. [S.l.: s.n.], 2000. Citado na página 62.

SGALL, P.; HAJICOVÁ, E.; HAJICOVÁ, E.; PANEVOVÁ, J.; PANEVOVA, J. **The meaning of the sentence in its semantic and pragmatic aspects**. [S.l.]: Springer Science & Business Media, 1986. Citado na página 30.

SILVA, E. H. **Etiquetagem morfossintática multigênero para o português do Brasil segundo o modelo Universal Dependencies**. Tese (Mestrado) — Universidade de São Paulo, 2023. Citado na página 15.

SILVEIRA, A.; DE SOUZA, E.; CAVALCANTI, T.; FREITAS, C. Do pdf ao txt: Desafios na extração de informação em textos técnico-científicos. In: **VI Workshop de Iniciação Científica em Tecnologia da Informação e da Linguagem Humana (TILic 2019). TILic**. [S.l.: s.n.], 2019. p. 15–18. Citado na página 71.

SIMOV, K. I.; OSENOVA, P.; SLAVCHEVA, M.; KOLKOVSKA, S.; BALABANOVA, E.; DOIKOFF, D.; IVANOVA, K.; SIMOV, A.; KOUYLEKOV, M. Building a Linguistically Interpreted Corpus of Bulgarian: the BulTreeBank. In: **Proceedings of the Third International Conference on Language Resources and Evaluation**. [S.l.: s.n.], 2002. p. 1729–1736. Citado na página 30.

SKUT, W.; KRENN, B.; BRANTS, T.; USZKOREIT, H. An annotation scheme for free word order languages. In: **Fifth Conference on Applied Natural Language Processing**. Washington, DC, USA: Association for Computational Linguistics, 1997. p. 88–95. Disponível em: <https://aclanthology.org/A97-1014>. Citado na página 36.

STRAKA, M.; HAJIC, J.; STRAKOVÁ, J. 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)**. [S.l.: s.n.], 2016. p. 4290–4297. Citado 2 vezes nas páginas 59 e 60.

TAYLOR, A.; KROCH, A. S. The Penn-Helsinki Parsed Corpus of Middle English. **MS. University of Pennsylvania**, 1994. Citado na página 30.

TAYLOR, A.; MARCUS, M.; SANTORINI, B. The Penn treebank: an overview. **Treebanks**, Springer, p. 5–22, 2003. Citado 2 vezes nas páginas 36 e 141.

THOMPSON, P.; ANANIADOU, S.; TSUJII, J. The genia corpus: Annotation levels and applications. In: **Handbook of Linguistic Annotation**. [S.l.]: Springer, 2017. p. 1395–1432. Citado na página 15.

TYERS, F.; SHEYANOVA, M.; WASHINGTON, J. Ud annotatrix: An annotation tool for universal dependencies. In: **Proceedings of the 16th International Workshop on Treebanks and Linguistic Theories**. [S.l.: s.n.], 2017. p. 10–17. Citado na página 62.

VILELA, M.; KOCH, M. I. V. Gramática da Língua Portuguesa: gramática de palavra, gramática de frase e gramática de texto/discurso. **Coimbra: Almedina**, 2001. Citado 3 vezes nas páginas 74, 76 e 77.

ZEMAN, D. Core arguments in Universal Dependencies. In: **Proceedings of the fourth international conference on dependency linguistics (DepLing 2017)**. [S.l.: s.n.], 2017. p. 287–296. Citado 4 vezes nas páginas 81, 82, 83 e 142.

ZEMAN, D.; HAJIC, J.; POPEL, M.; POTTHAST, M.; STRAKA, M.; GINTER, F.; NIVRE, J.; PETROV, S. 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**. [S.l.: s.n.], 2018. p. 1–21. Citado na página 60.

A
Verbos aos quais se associam objetos preposicionados

nunca (74):

acondicionar; aflorar; afundar; aglomerar; apodrecer; ascender; cair; cimentar; coalescer; coexistir; colapsar; colar; competir; condensar; contestar; coprecipitar; crescer; datar; decair; decantar; depletar; desaparecer; desarticular; desfolhar; deteriorar; dissipar; divulgar; dopadar; dopar; emergir; endurecer; entrar; envelhecer; escorrer; especificar; esperar; esquentar; estender; evaporar; evoluir; explodar; fluir; flutuar; gelificar; gotejar; impactar; impermeabilizar; inchar; inflamar; intervir; ler; lixiviar; mergulhar; migrar; minar; molhar; ocorrer; optar; penetrar; perdurar; permear; permutar; persistir; pesquisar; preocupar; prograd ar; recomeçar; reflutuar; sair; solubilizar; subir; terminar; tocar; viver

menos de 30% das vezes (12):

calcinar; circular; escoar; operar; partir; prosseguir; reagir; retornar; seguir; ser; trabalhar; variar

entre 30 e 70% (61):

acontecer; acumular; adsorver; advir; agir; ajudar; anular; aparecer; aquecer; arrancar; atestar; atuar; avançar; bastar; carregar; cessar; chegar; complementar; concordar; continuar; convergir; danificar; decidir; decrescer; delinear; descobrir; diminuir; dissolver; emulsificar; examinar; existir; fechar; ficar; finalizar; focar; funcionar; gastar; gerir; gradar; implementar; incidir; investir; ir; lembrar; manipular; monitorar; passar; permanecer; precipitar; prevalecer; provar; resfriar; romper; secar; servir; somar; suplementar; surgir; truncar; turvar; visualizar

mais de 70% das vezes (111):

acabar; adaptar; aderir; afirmar; alcançar; apontar; apresentar; armazenar; atingir; aumentar; auxiliar; avaliar; buscar; capturar; caracterizar; carrear; causar; classificar; coincidir; começar; comparar; completar; compreender; comprimir; concluir; conseguir; considerar; consistir; constituir; contar; conter; contribuir; corresponder; corroborar; culminar; definir; deixar; demonstrar; desenvolver; deslocar; discriminar; dividir; dizer; dominar; elevar; empregar; encontrar; englobar; estabilizar; estudar; evitar; fazer; formar; gerar; identificar; ilustrar; incluir; influenciar; inibir; iniciar; interagir; interferir; interromper; invadir; lançar; levar; medir; minimizar; modificar; mostrar; mudar; necessitar; notar; observar; obter; parecer; percorrer; perder; permitir; poder; possuir; precisar; prevenir; produzir; propor; provocar; (...)

reconhecer; reduzir; remover; resolver; responder; resultar; resumir; retardar; saber; separar; simular; sugerir; suportar; tender; ter; tornar; transportar; tratar; usar; utilizar; valer; ver; verificar; vir; voltar

sempre (461):

abarcar; abastecer; abordar; abranger; abrigar; abrir; absorver; acarretar; aceitar; acelerar; acomodar; acompanhar; acreditar; acrescentar; adentrar; adequar; adicionar; admitir; adotar; adquirir; aferir; afetar; aglutinar; agregar; agrupar; ajustar; alertar; alimentar; alinhar; almejar; alterar; ampliar; analisar; anteceder; anunciar; aparentar; aplicar; apoiar; aprimorar; aprisionar; aprovar; aproveitar; aproximar; argumentar; arrastar; assinar; assistir; associar; assumir; atender; atenuar; ativar; atrair; atravessar; atribuir; auditar; averiguar; baixar; balancear; balizar; banhar; bloquear; bombear; bordejar; caber; calcular; calibrar; carecer; ceder; certificar; chamar; circundar; citar; cobrir; coletar; colidir; colocar; combinar; compatibilizar; compensar; complicar; compor; comprometer; comprovar; comunicar; conceder; concentrar; concernir; condicionar; conduzir; conferir; configurar; confinar; confirmar; confundir; conhecer; consolidar; constar; constatar; construir; consumir; contabilizar; contaminar; contemplar; contornar; contrabalancear; contrapor; contrariar; contrastar; controlar; converter; correlacionar; corrigir; corroer; corromper; cortar; costumar; crer; criar; cruzar; cumprir; custar; dar; debilitar; deduzir; defender; defletir; deformar; delimitar; demandar; demostrar; denominar; denotar; depender; depositar; desbloquear; descartar; desconsiderar; descontar; descrever; desempenhar; desestabilizar; desidratar; designar; desprender; despressurizar; destacar; destruir; desviar; detalhar; detectar; deter; determinar; dever; diagnosticar; diferenciar; diferir; dificultar; diluir; dimensionar; dinamizar; direcionar; dirigir; discernir; discutir; disponibilizar; dispor; dissertar; distinguir; distorcer; distribuir; divergir; dotar; durar; efetuar; elaborar; eleger; elucidar; emitir; empurrar; encarecer; encerrar; encorajar; enfatizar; enfocar; enfrentar; engolir; enquadrar; enriquecer; ensaiar; entender; enviar; envolver; equilibrar; equivaler; erodir; esboçar; esclarecer; escolher; espessar; estabelecer; esticar; estimar; estimular; estruturar; evidenciar; exceder; excitar; executar; exemplificar; exercer; exibir; exigir; expandir; experimentar; explicar; explorar; expor; exportar; expressar; expulsar; extrair; extrapolar; fabricar; facilitar; falar; favorecer; ferver; fomentar; formular; fornecer; fortalecer; forçar; fragmentar; fraturar; frisar; fritar; fugir; ganhar; garantir; generalizar; gerenciar; girar; habitar; haver; herdar; hidratar; homogeneizar; impedir; implantar; implicar; impor; (...)

impossibilitar; impulsionar; incentivar; incorporar; independer; indicar; individualizar; induzir; inferir; influir; inicializar; injetar; inserir; instalar; integrar; intensificar; interceptar; interligar; interpretar; introduzir; intrudir; inverter; investigar; inviabilizar; irritar; isolar; juntar; justificar; levantar; liberar; lidar; liderar; ligar; limitar; limpar; listar; livrar; lubrificar; manter; mapear; marcar; margear; mariscar; mascar; mascarar; maximizar; melhorar; merecer; misturar; mitigar; modelar; montar; mosaicar; motivar; mover; movimentar; multiplicar; neutralizar; nivelar; nortear; obedecer; objetivar; obrigar; obstruir; ocasionar; ocupar; oferecer; orbitar; organizar; orientar; originar; ostentar; otimizar; participar; perceber; perfazer; perfurar; pertencer; perturbar; pesar; posicionar; possibilitar; postular; potencializar; preceder; predizer; predominar; preencher; prejudicar; prender; preparar; preponderar; preservar; pressurizar; prever; priorizar; proceder; processar; procurar; projetar; prolongar; promover; propiciar; proporcionar; prospectar; proteger; prover; providenciar; provir; purificar; qualificar; quantificar; quebrar; queimar; quimissorver; raspar; rastrear; ratificar; realizar; realçar; reativar; receber; recobrir; recomendar; recuperar; refinar; refletir; reforçar; registrar; regular; reiniciar; reiterar; relacionar; relatar; remeter; remunerar; representar; requerer; residir; respeitar; responsabilizar; ressaltar; restar; restaurar; restringir; reter; retirar; retomar; retrabalhar; retratar; reunir; revelar; reverter; salientar; satisfazer; saturar; seccionar; secionar; selecionar; sequestrar; significar; simplificar; sinalizar; situar; sobrepor; sofrer; solucionar; subdividir; submeter; substituir; subtrair; suceder; sumarizar; superar; superpor; supor; suprimir; suprir; surtir; sustentar; tamponar; tangenciar; tanger; tentar; testar; tolerar; tomar; totalizar; traduzir; transferir; transformar; transmitir; transpor; trazer; traçar; trocar; ultrapassar; unir; validar; vaporizar; varrer; vencer; vender; viabilizar; vincular; visar; viscosificar; vislumbrar; vivenciar

B
Verbos que só ocorrem com complemento oracional, no particípio ou associados ao pronome –se

Verbos com complemento apenas oracional (13):

acanalar; acreditar; aparentar; argumentar; comunicar; costumar; crer; descontar; encarregar; frisar; retratar; salientar; tentar

Verbos que só ocorrem no particípio (324):

abater; acanalar; acentuar; acessar; acinzentar; acionar; aconselhar; acoplar; acrescer; acunhar; adelgaçar; aditivar; administrar; adsorvir; afastar; afinar; agitar; alargar; aliar; alicerçar; alongar; alternar; amalgamar; amarelar; ameaçar; amostrar; amplificar; ancorar; anotar; aplainar; apropriar; apurar; arranjar; arredondar; arrefecer; aspirar; assinalar; assorear; atacar; atrelar; atualizar; auto-suportar; autorizar; avermelhar; basear; batizar; bombardear; borbulhar; canalizar; carbonificar; carbonizar; carburar; catalogar; centralizar; centrar; centrifugar; cercar; cisalhar; colorir; comercializar; compactar; compartilhar; comprar; computar; conceber; conceituar; conectar; confeccionar; congelar; conjecturar; conjugar; consultar; contatar; contratar; coordenar; copolimerizar; credenciar; creditar; cultivar; curvar; declarar; decompor; decorrer; dedicar; degradar; deionizar; delgar; derivar; derramar; derreter; desacelerar; desaluminizar; desconhecer; descontinuar; desejar; desequilibrar; desgaseificar; desligar; desmagnetizar; desobstruir; desordenar; despender; desprover; dessorver; destilar; destinar; destituir; difundir; digerir; digitalizar; dinoflagelar; dispensar; dispersar; dissociar; diversificar; doar; dobrar; dolomitizar; dosar; embasar; empilhar; empreender; emulsionar; encaixar; encaminhar; encarregar; enrolar; enterrar; entregar; entrelaçar; equipar; equivocar; esbranquiçar; escanear; escarpar; escrever; escurecer; espaçar; especializar; espelhar; espiralar; estagnar; estar; estipular; estratificar; estriar; esverdear; etoxilar; evacuar; exacerbar; exagerar; exaurir; explanar; explotar; exprimir; extrapesar; extrudir; falhar; festonar; filtrar; fissurar; fixar; fluidificar; fundamentar; fundar; fundir; galvanizar; georreferenciar; governar; gridar; guardar; hidrolisar; igualar; imbricar; imobilizar; importar; impregnar; imprimir; inalar; inaugurar; inclinar; inconsolidar; indefinir; indeterminar; informar; insaturar; inspirar; intemperizar; intercalar; interconectar; interdigitar; (...)

interessar; interestratificar; interpenetrar; interpolar; isentar; justapor; laminar; lavar; linearizar; localizar; magnetizar; matar; materializar; medida; mencionar; microemulsionar; microprocessar; moderar; moer; morrer; nanoestruturar; nomear; normalizar; numerar; não-afinar; não-associar; não-confinar; não-consolidar; não-inibir; obliterar; ondular; ordenar; pagar; paleomagnetizar; parar; particionar; peneirar; planejar; plotar; polarizar; polimerizar; ponderar; pontilhar; povoar; praticar; preconizar; predeterminar; preferir; prensar; pretender; privatizar; programar; proibir; pronunciar; pré-aquecer; pré-determinar; pré-estabelecer; pré-tratar; publicar; pulverizar; pôr; quantizar; ramificar; re-encaminhar; realocar; reaproveitar; reaquecer; rebaixar; rebater; rechear; reciclar; recircular; recolher; reconectar; recém-fragmentar; redepositar; reescrever; referir; regenerar; regionalizar; reinjetar; rejeitar; renegociar; repetir; reportar; repousar; reservar; respaldar; restabelecer; reticular; reutilizar; revestir; salgar; sanfonar; saponificar; sediar; segregar; semi-refrigerar; silicificar; sintetizar; sobrebalancear; soerguer; solidificar; sombrear; soprar; soterrar; sotopor; suavizar; subarredondar; subdeterminar; subordinar; subsaturar; subverticalizar; sujeitar; sulfurar; super-impor; supracitar; suspender; tabelar; tecer; temperar; termoprogramar; testemunhar; transesterificar; triturar; umedecer; vazar; vedar; vegetar; verter; verticalizar; vestir; zerar

Verbos que só ocorrem associados a –se (77):

acentuar; acionar; acunhar; adelgaçar; afastar; agitar; aguardar; alargar; aliar; alicerçar; aliviar; alojar; alongar; alçar; amplificar; anexar; aprofundar; apurar; assemelhar; assentar; ater; basear; beneficiar; comportar; conectar; curvar; decompor; dedicar; depreender; desejar; destinar; difundir; dispersar; dissociar; efetivar; encaixar; espalhar; esquematizar; estreitar; evacuar; fender; filtrar; fixar; fundamentar; fundir; horizontalizar; imaginar; informar; instabilizar; intemperar; interdigitar; interditar; ionizar; julgar; liquefazer; localizar; manifestar; paralelizar; pautar; pensar; plotar; ponderar; prestar; presumir; pretender; propagar; reajustar; recolher; referir; repetir; reportar; reproduzir; reutilizar; sobressair; tirar; titular; zerar

C
Expressões que foram anotadas tanto como MWEs quanto sintagmas transparentes em contextos distintos no PetroGold

de acordo com	MWE	**De acordo com Luiz & Silva (1995) , frequentemente os lineamentos observados em os mapas magnéticos são paralelos a as direções estruturais como zonas de cisalhamento , falhas , fraturas e dobras .
	transp.	Os valores geométricos foram adotados para simplificação de o problema , em o entanto , estão **de acordo com as características de muitos poços de a região .
até que	MWE	Os glóbulos de asfalto são mantidos separados por o agente emulsificante **até que a emulsão se deposite em a superfície de o terreno , de o pavimento existente ou envolvendo as partículas de agregado .
	transp.	Alguns estudos indicam **até que a reação não acontece sem a presença de um catalisador (VICENTE et al. , 2005) .
é que	MWE	Somente a o misturar as duas fases **é que se adiciona o agente modificador .
	transp.	De acordo com Oda (2000) , o comportamento esperado **é que a adição de nanomateriais torne o ligante mais consistente , resultando em valores mais baixos de penetração , proporcionalmente a o teor de nanomaterial inserido .
quanto a	MWE	Este ensaio tem como objetivo verificar a susceptibilidade de a emulsão asfáltica **quanto a sedimentação , verificando se a emulsão apresenta estabilidade quanto a estocagem (não haver separação de as fases constituintes) .
	transp.	A quantidade de CO2 emitido em a exaustão indica o **quanto a combustão é completa .

sendo assim	MWE	**Sendo assim , a interação entre duas cargas elétricas , a exemplo , seria explicada por uma ação que se propaga entre uma carga e outra através de o éter existente entre elas .
	transp.	A lei de Beer pode ser aplicada somente para soluções diluídas , **sendo assim uma lei limite .
isto é	MWE	O problema de a identificação de os argilominerais presentes em uma argila é relativamente simples quando a amostra é pura , **isto é , contém apenas um argilomineral .
	transp.	**Isto é feito através de modificações químicas ou físicas , cada uma de elas oferecendo vantagens em a mitigação de os efeitos de as patologias apresentadas por os pavimentos asfálticos durante sua vida útil .
sem que	MWE	Os argilominerais têm capacidade de trocar íons , isto é , têm íons fixados em a superfície , entre as camadas e dentro de os canais de a estrutura cristalina , que podem ser trocados , através de reações químicas , por outros íons em solução aquosa , **sem que isso venha trazer quaisquer modificação de a sua estrutura cristalina .
	transp.	O sistema é apropriado para a elevação de petróleos pesados e viscosos e para situações em águas profundas **sem que a instalação de plataforma a menor distância de os poços e inviável , técnica ou economicamente .
a o	MWE	**A o combinarem se as tecnologias para o tratamento de gases com alta pressão parcial de CO2 , a primeira etapa de a separação é realizada por membranas , reduzindo o teor de CO2 em o gás .
	transp.	A área de estudo se estende de o sul de o Estado de São Paulo **a o norte de o Rio Grande de o Sul em a porção emersa e a Bacia de Pelotas em a região offshore , perfazendo um total de 440 530 Km^2 , de estes 278 048 Km^2 estão localizados em a porção continental e 162 482 Km^2 em a parte offshore (Figura 2) .

para tal	MWE	**Para tal é necessário realizar uma análise econômica de esta unidade , o que será feito em este trabalho .
	transp.	Uma interpretação plausível **para tal comportamento seria o efeito somado de o mergulho de o embasamento e a presença de as rochas de baixa magnetização associadas a as supracrustais .
a o que	MWE	A planície de Icapuí foi construída a a medida que os fluxos de matéria e energia proporcionavam a produção de sedimentos e nutrientes , a sua distribuição e deposição a o longo de as unidades ambientais e ecossistemas associados , **a o que se aliaram a as flutuações de o nível relativo de o mar , mudanças climáticas e ação de as energias modeladoras atuais (ondas , marés , ventos , gravidade , pluvial e hidrodinâmica superficial e subterrânea) (MEIRELES , 2012) .
	transp.	Estes domínios (rochas supracrustais) têm um padrão magnético de baixa freqüência que contrasta com as unidades vizinhas , muito semelhante **a o que se esperaria em o caso de o abatimento de blocos de mesma natureza magnética .

D
112 MWEs encontradas no PetroGold e respectiva anotação morfossintática

MWE	POS	MWE POS	REL	FEATS
a a medida em que	ADP DET NOUN ADP PRON	SCONJ	mark	_ Definite=Def | Gender=Fem | Number=Sing | PronType=Art Gender=Fem | Number=Sing _ Gender=Fem | Number=Sing | PronType=Rel
a a medida que	ADP DET NOUN SCONJ	SCONJ	mark	_ Definite=Def | Gender=Fem | Number=Sing | PronType=Art Gender=Fem | Number=Sing _
a as vezes	ADP DET NOUN	ADV	obl	_ Definite=Def | Gender=Fem | Number=Plur | PronType=Art Gender=Fem | Number=Plur
a exemplo de	ADP NOUN ADP	ADP	case	_ Gender=Masc | Number=Sing _
a favor de	ADP NOUN ADP	ADP	case	_ Gender=Masc | Number=Sing _
a fim de	ADP NOUN ADP	SCONJ	mark	_ Gender=Masc | Number=Sing _
a medida que	ADP NOUN SCONJ	SCONJ	mark	_ Gender=Fem | Number=Sing _
a não ser que	ADP ADV VERB SCONJ	SCONJ	mark	_ Polarity=Neg VerbForm=Inf _

a o contrário de	ADP DET NOUN ADP	ADP	case	_ Definite=Def | Gender=Masc | Number=Sing | PronType=Art Gender=Masc | Number=Sing _
a o invés de	ADP DET NOUN ADP	ADP	case	_ Definite=Def | Gender=Masc | Number=Sing | PronType=Art Gender=Masc | Number=Sing _
a o largo de	ADP DET NOUN ADP	ADP	case	_ Definite=Def | Gender=Masc | Number=Sing | PronType=Art Gender=Masc | Number=Sing _
a o longo de	ADP DET NOUN ADP	ADP	case	_ Definite=Def | Gender=Masc | Number=Sing | PronType=Art Gender=Masc | Number=Sing _
a o menos	ADP DET ADV	ADV	advmod	_ Definite=Def | Gender=Masc | Number=Sing | PronType=Art _
a o menos que	ADP DET ADV SCONJ	SCONJ	mark	_ Definite=Def | Gender=Masc | Number=Sing | PronType=Art _ _
a o passo que	ADP DET NOUN PRON	SCONJ	mark	_ Definite=Def | Gender=Masc | Number=Sing | PronType=Art Gender=Masc | Number=Sing Gender=Masc | Number=Sing | PronType=Rel
a o ponto de que	ADP DET NOUN ADP PRON	SCONJ	mark	_ Definite=Def | Gender=Masc | Number=Sing | PronType=Art Gender=Masc | Number=Sing _ Gender=Masc | Number=Sing | PronType=Rel
a o que	ADP PRON PRON	SCONJ	mark	_ Gender=Masc | Number=Sing | PronType=Dem Gender=Masc | Number=Sing | PronType=Rel

a o todo	ADP DET PRON	ADV	obl	_ Definite=Def | Gender=Masc | Number=Sing | PronType=Art Gender=Masc | Number=Sing | PronType=Tot
a partir de	ADP VERB ADP	ADP	case	_ VerbForm=Inf _
a principio	ADP NOUN	ADV	obl	_ Gender=Masc | Number=Sing
a princípio	ADP NOUN	ADV	obl	_ Gender=Masc | Number=Sing
a priori	ADP NOUN	ADV	obl	_ Gender=Fem | Number=Sing
a respeito de	ADP NOUN ADP	ADP	case	_ Gender=Masc | Number=Sing _
a seguir	ADP VERB	ADV	obl	_ VerbForm=Inf
ainda que	ADV SCONJ	SCONJ	mark	_ _
além de	ADV ADP	CCONJ	cc	_ _
além de isso	ADV ADP PRON	ADV	advmod	_ _ Gender=Masc | Number=Sing | PronType=Dem
além de isto	ADV ADP PRON	ADV	advmod	_ _ Gender=Masc | Number=Sing | PronType=Dem
além de o mais	ADV ADP DET PRON	ADV	advmod	_ _ Definite=Def | Gender=Masc | Number=Sing | PronType=Art Gender=Masc | Number=Sing | PronType=Ind
além de o que	ADV ADP PRON PRON	CCONJ	cc	_ _ Gender=Masc | Number=Sing | PronType=Dem Gender=Masc | Number=Sing | PronType=Rel

além de o quê	ADV ADP DET NOUN	ADV	advmod	_ _ Definite=Def | Gender=Masc | Number=Sing | PronType=Art Gender=Masc | Number=Sing
apesar de	ADV ADP	SCONJ	mark	_ _
assim como	ADV ADP	CCONJ	cc	_ _
assim por diante	ADV ADP ADV	ADV	conj	_ _ _
assim que	ADV SCONJ	SCONJ	mark	_ _
assim sendo	ADV AUX	ADV	advmod	_ VerbForm=Ger
até que	ADP SCONJ	SCONJ	mark	_ _
bem como	ADV ADP	CCONJ	cc	_ _
cada vez mais	DET NOUN DET	ADV	advmod	Gender=Fem | Number=Sing | PronType=Tot Gender=Fem | Number=Sing Gender=Fem | Number=Sing | PronType=Ind
caso contrário	SCONJ ADJ	ADV	obl	_ Gender=Masc | Number=Sing
cerca de	ADV ADP	ADV	advmod	_ _
com base em	ADP NOUN ADP	ADP	case	_ Gender=Fem | Number=Sing _
com isso	ADP PRON	ADV	obl	_ Gender=Masc | Number=Sing | PronType=Dem
com relação a	ADP NOUN ADP	ADP	case	_ Gender=Fem | Number=Sing _
com vistas a	ADP NOUN ADP	SCONJ	mark	_ Gender=Fem | Number=Plur _

como relação a	ADP NOUN ADP	ADP	case	_ Gender=Fem | Number=Sing _
como também	ADV ADV	CCONJ	cc	_ _
de acordo com	ADP NOUN ADP	ADP	case	_ Gender=Masc | Number=Sing _
de acordos com	ADP NOUN ADP	ADP	case	_ Gender=Masc | Number=Plur _
de aí	ADP ADV	ADV	obl	_ _
de forma a	ADP NOUN ADP	SCONJ	mark	_ Gender=Fem | Number=Sing _
de forma que	ADP NOUN SCONJ	SCONJ	mark	_ Gender=Fem | Number=Sing _
de maneira a	ADP NOUN SCONJ	SCONJ	mark	_ Gender=Fem | Number=Sing _
de modo a	ADP NOUN ADP	SCONJ	mark	_ Gender=Masc | Number=Sing _
de modo que	ADP NOUN SCONJ	SCONJ	mark	_ Gender=Masc | Number=Sing _
de o que	ADP PRON PRON	SCONJ	mark	_ Gender=Masc | Number=Sing | PronType=Dem Gender=Masc | Number=Sing | PronType=Rel
de tal forma que	ADP DET NOUN SCONJ	SCONJ	mark	_ Gender=Fem | Number=Sing | PronType=Ind Gender=Fem | Number=Sing _
desde que	ADP SCONJ	SCONJ	mark	_ _

devido a	VERB ADP	ADP	case	Gender=Masc | Number=Sing | VerbForm=Part _
em a faixa de	ADP DET NOUN ADP	ADP	case	_ Definite=Def | Gender=Fem | Number=Sing | PronType=Art Gender=Fem | Number=Sing _
em a verdade	ADP DET NOUN	ADV	obl	_ Definite=Def | Gender=Fem | Number=Sing | PronType=Art Gender=Fem | Number=Sing
em direção a	ADP NOUN ADP	ADP	case	_ Gender=Fem | Number=Sing _
em função de	ADP NOUN ADP	ADP	case	_ Gender=Fem | Number=Sing _
em geral	ADP ADJ	ADV	obl	_ Gender=Masc | Number=Sing
em o caso de	ADP DET NOUN ADP	ADP	case	_ Definite=Def | Gender=Masc | Number=Sing | PronType=Art Gender=Masc | Number=Sing _
em o entanto	ADP DET NOUN	CCONJ	cc	_ Definite=Def | Gender=Masc | Number=Sing | PronType=Art Gender=Masc | Number=Sing
em razão de	ADP NOUN ADP	ADP	case	_ Gender=Fem | Number=Sing _
em relação a	ADP NOUN ADP	ADP	case	_ Gender=Fem | Number=Sing _
em relação as	ADP NOUN ADP	ADP	case	_ Gender=Fem | Number=Sing _
em relação á	ADP NOUN ADP	ADP	case	_ Gender=Fem | Number=Sing _
em seguida	ADP NOUN	ADV	obl	_ Gender=Fem | Number=Sing

em separado	ADP NOUN	ADV	obl	_ Gender=Masc | Number=Sing
em termos de	ADP NOUN ADP	ADP	case	_ Gender=Masc | Number=Plur _
em torno de	ADP NOUN ADP	ADV	obl	_ Gender=Masc | Number=Sing _
em vez de	ADP NOUN ADP	SCONJ	mark	_ Gender=Fem | Number=Sing _
enquanto que	ADV SCONJ	SCONJ	mark	_ _
isto é	PRON AUX	CCONJ	cc	Gender=Masc | Number=Sing | PronType=Dem Mood=Ind | Number=Sing | Person=3 | Tense=Pres | VerbForm=Fin
junto a	ADV ADP	ADV	advmod	_ _
já que	ADV SCONJ	SCONJ	mark	_ _
mesmo assim	ADV ADV	ADV	advmod	_ _
mesmo que	ADV SCONJ	SCONJ	mark	_ _
não que	ADV SCONJ	ADV	advmod	_ _
não só	ADV ADV	ADV	advmod	Polarity=Neg _
ou seja	CCONJ VERB	CCONJ	cc	_ Mood=Sub | Number=Sing | Person=3 | Tense=Pres | VerbForm=Fin
para que	ADP SCONJ	SCONJ	mark	_ _
para tal	ADP ADV	ADV	obl	_ _

por causa de	ADP NOUN ADP	ADP	case	_ Gender=Fem | Number=Sing _
por conseguinte	ADP NOUN	ADV	obl	_ Gender=Masc | Number=Sing
por exemplo	ADP NOUN	ADV	obl	_ Gender=Masc | Number=Sing
por exemplos	ADP NOUN	ADV	obl	_ Gender=Masc | Number=Plur
por fim	ADP NOUN	CCONJ	cc	_ Gender=Masc | Number=Sing
por meio de	ADP NOUN ADP	ADP	case	_ Gender=Masc | Number=Sing _
por o menos	ADP DET NOUN	ADV	obl	_ Definite=Def | Gender=Masc | Number=Sing | PronType=Art Gender=Masc | Number=Plur
por parte de	ADP NOUN ADP	ADP	case	_ Gender=Fem | Number=Sing _
por sua vez	ADP DET NOUN	ADV	obl	_ Gender=Fem | Number=Sing | PronType=Prs Gender=Masc | Number=Sing
por vezes	ADP NOUN	ADV	obl	_ Gender=Fem | Number=Plur
quanto a	ADV ADP	ADP	case	_ _
quanto mais	ADV ADV	SCONJ	mark	_ _
sem que	ADP SCONJ	SCONJ	mark	_ _
sempre que	ADV SCONJ	SCONJ	mark	_ _
sendo assim	AUX ADV	ADV	obl	VerbForm=Ger _
sendo que	AUX SCONJ	SCONJ	mark	VerbForm=Ger _

tais como	DET ADP	ADP	case	Gender=Fem | Number=Plur | PronType=Dem _
tal como	ADV ADP	SCONJ	mark	_ _
tanto quanto	ADV ADV	ADV	advmod	_ _
tanto que	ADV SCONJ	SCONJ	mark	_ Gender=Masc | Number=Plur | PronType=Rel
toda vez que	DET ADP SCONJ	SCONJ	mark	Gender=Fem | Number=Sing | PronType=Tot _ _
um pouco	DET ADV	ADV	obl	Definite=Ind | Gender=Masc | Number=Sing | PronType=Art _
uma vez em	ADP NOUN ADP	SCONJ	mark	_ Gender=Fem | Number=Sing _
uma vez que	DET NOUN SCONJ	SCONJ	mark	Definite=Ind | Gender=Fem | Number=Sing | PronType=Art Gender=Fem | Number=Sing _
visto que	VERB SCONJ	SCONJ	mark	Gender=Masc | Number=Sing | VerbForm=Part _
é que	AUX SCONJ	SCONJ	discourse	_ _

E
100 trigramas e quadrigramas mais bem *rankeados* pelo algoritmo *Likelihood-ratio* no Petrolês

Rank	Trigrama	Quadrigrama
1	a partir de	em a produção de
2	a fim de	em a presença de
3	a respeito de	em a formação de
4	de acordo com	em a região de
5	em relação a	em a concentração de
6	de resistência a	em a temperatura de
7	em função de	em a superfície de
8	com relação a	em a análise de
9	de produção de	em a indústria de
10	por meio de	de o processo de
11	em presença de	em a área de
12	em torno de	em a faixa de
13	para produção de	em a reação de
14	com base em	em a determinação de
15	de formação de	em a redução de
16	de concentração de	em a maioria de
17	em termos de	em a composição de
18	de petróleo em	em a remoção de
19	de energia de	em a simulação de
20	para obtenção de	em a etapa de
21	de óleo de	em a direção de
22	de obtenção de	em a forma de
23	de redução de	em a ausência de
24	para determinação de	em a câmara de
25	de energia em	em a década de
26	de óleo em	em a saída de
27	de perda de	em a entrada de
28	de transferência de	de o sistema de
29	de análise de	de o modelo de
30	de água de	de o aumento de
31	no caso de	de o óleo de

32	de carbono em	de o tempo de
33	de água em	de o uso de
34	em direção a	de o método de
35	em comparação a	de o número de
36	de remoção de	de o valor de
37	para análise de	de o tipo de
38	por parte de	de a produção de
39	em condições de	de a presença de
40	de massa de	de a formação de
41	de injeção de	de o fluido de
42	de dados de	de o desenvolvimento de
43	para remoção de	de a equação de
44	de controle de	de o teor de
45	de temperatura de	de o comportamento de
46	em virtude de	de o volume de
47	de variação de	de o ponto de
48	de volume de	de o gás de
49	de geração de	de a concentração de
50	de emissões de	de a temperatura de
51	de tempo de	de o estado de
52	de gás de	de o estudo de
53	de adsorção de	de o consumo de
54	de recuperação de	de o nível de
55	de emissão de	de o fator de
56	em contato com	de o campo de
57	de entrada de	de o coeficiente de
58	em razão de	de o tamanho de
59	de dióxido de	de a região de
60	de gás em	de o fluxo de
61	de pressão de	de a quantidade de
62	de reação de	de a utilização de
63	em massa de	de o preço de
64	em comparação com	de o comprimento de
65	de conversão de	de o bagaço de
66	para avaliação de	de a análise de
67	de distribuição de	de o corpo de
68	em amostras de	de a taxa de
69	de separação de	de a superfície de

70	de operação de	de a indústria de
71	de massa em	de a área de
72	em forma de	de a reação de
73	de saída de	de a água de
74	por unidade de	de a velocidade de
75	de consumo de	de a pressão de
76	de biodiesel em	de a faixa de
77	de fluxo de	de a capacidade de
78	para fins de	de a perda de
79	de crescimento de	de a mistura de
80	de processamento de	de a adição de
81	de transporte de	de a solução de
82	de onda de	de a curva de
83	de tratamento de	de a coluna de
84	de equilíbrio de	de a variação de
85	de oxidação de	de a razão de
86	de conservação de	de a vazão de
87	de degradação de	de a composição de
88	de liberação de	de a estrutura de
89	de hidróxido de	de a aplicação de
90	de medição de	de a distribuição de
91	com exceção de	de a atividade de
92	de mistura de	de a viscosidade de
93	de extração de	de a injeção de
94	de pressão em	de a ordem de
95	como resultado de	de a qualidade de
96	de metano em	de a técnica de
97	de alimentação de	de a massa de
98	de gases de	de a linha de
99	de vida de	de a década de
100	de captura de	de a camada de

F
Pronomes "se" que se associam tanto a verbos pronominais quanto verbos de sujeito indeterminado no PetroGold

lema	expl	exemplo
chamar	pv	Admite-se que o petróleo foi formado há milhões de anos pelo acumulo de diferentes seres vivos como a decomposição de plânctons - seres que são geralmente encontrados na zona costeira, mares, oceanos e estuários - esses seres teriam se acumulados no fundo dos mares, rios e lagos e soterrados pela ação do movimento da crosta terrestre e posteriormente com o passar dos anos transformando-se em uma pasta oleosa que hoje **se chama** petróleo (VAZ, 2011).
	impers	*Em biologia marinha e limnologia, **chama-se** bentos aos organismos que vivem no substrato, fixos ou não, em contraposição com os pelágicos, que vivem livremente na coluna de água.
dever	pv	Podemos ver através de gráfico 4.3 que existe uma relação entre a concentração do cálcio e a do tensotivo, no processo de extração como pode-se ver no gráfico para elevadas concentrações de metal e baixas concentrações de tensoativos, já se pode ver uma moderada extração do óleo, isso **se deve** a fato de o floco carrega na sua estrutura íons do metal interagindo com a parte iônica do tensoativo e o óleo disperso interagindo com a parte lipofílica do tensoativo.
	impers	**Deve-se** notar, também, que a interpretação magnética sugere que a falha da borda sul, e sua calha condicionada, estão deslocadas por um alinhamento de direção NW-SE possivelmente associado às falhas de rejeito direcional (reativadas no Cretáceo/Terciário) mencionadas anteriormente.

G
Pronomes "se" que se associam tanto a verbos pronominais quanto verbos na voz passiva sintética no PetroGold

lema	expl	exemplo
ajustar	pv	A partir do coeficiente de correlação, percebe-se que todos os modelos se **ajustaram**.
	pass	A bomba de água foi acionada com uma freqüência de 30 Hz e então **ajustou**-se a freqüência baseando-se na vazão de água desejada.
apresentar	pv	A osmose reversa opera sob elevadas pressões, a nanofiltração **apresenta**-se como um processo de filtragem de maior potencial de aplicação na remoção de compostos dissolvidos.
	pass	**Apresentam**-se abaixo tecnologias aplicadas na remoção de metais pesados, compostos orgânicos e produtos químicos dissolvidos em áreas offshore.
caracterizar	pv	Uma vez que o escoamento através do meio poroso **caracteriza**-se por velocidades baixas, a viscosidade mais elevada para o fluido preparado com GX, fez com que este fluido encontrasse uma maior dificuldade em atravessar o meio poroso, sendo retido no mesmo.
	pass	A elaboração de isotermas de adsorção é o procedimento mais usual para se **caracterizar** as propriedades dos polímeros em solução e na presença de um adsorvente.
correlacionar	pv	O processo de reativação da região Sudeste do Brasil, durante o Cretáceo, foi acompanhado de um típico e intenso magmatismo toleítico que se **correlaciona** aos derrames da Bacia do Paraná com idades de 137-127 Ma (Turner et al. 1994 em Ferrari 2001) onde teve sua máxima expressão.
	pass	**Correlacionou**-se os dados dos ensaios de aderência da argila ao metal com os de inibição da reatividade natural das argilas.

dar	pv	Veil (2004) explica que, a origem da água produzida se dá nas formações subterrâneas produtoras, que geralmente são permeadas por diferentes fluidos, tais como óleo, gás e água com alto teor de salinidade.
	pass	A movimentação das partículas do óleo e as mudanças que ocorre enquanto esse óleo está em contato com a água, se dá o nome de intemperismo (POSSOBON, 2012).
descrever	pv	Acima das rochas fluviais e flúvio-deltaicas **descreve**-se a associação das fácies St, Sd, Sh, Sr e subordinadamente lamitos.
	pass	Nesta seção observaram-se fácies típicas da Fm. Quricó, **descrevendo**-se fácies heterolíticas, mas com recorrente contribuição arenosa, com lentes e camadas de arenitos.
desenvolver	pv	A presença desta calha no embasamento raso, que produz uma fraca anomalia positiva, apóia a idéia proposta por Zalán e Oliveira (2005) acerca do sistema marítimo de riftes que se **desenvolveu** no Terciário sobre a plataforma continental da Bacia de Santos.
	pass	Para a realização de medidas de ângulo de contato estático no interior de uma tubulação **desenvolveu**-se um aparato constituído por uma secção de tubulação idêntica a utilizada nos experimentos de repartida, uma seringa para injeção de óleo, uma válvula, duas tampas de acrílico, uma haste e uma base.
dividir	pv	Os mecanismos de retenção de polímeros no meio poroso **dividem**-se em: adsorção, aprisionamento mecânico e retenção hidrodinâmica.
	pass	A partir da emulsão preparada, **divide**-se em cinco amostras iguais e, em seguida, adiciona-se tensoativos em concentrações diferentes de 0, 1, 2, 3, 4 % do volume da amostra.
encontrar	pv	Tensoativos anfóteros podem se comporta tanto como tensoativo catiônico ou aniônico, isso vai depender do PH do meio onde ele se **encontra**.

	pass	Não houve, entretanto, preocupação neste momento, de se **encontrar** o significado matemático desta conclusão.
estabelecer	pv	Por outro lado, todos os gráficos indicam que com o aumento do tempo de parada, a duração do transiente inicial, onde o escoamento se **estabelece** e ocorre mistura das fases, tende a se tornar mais longo.*
	pass	Para a realização do ensaio com o sistema argila-água, **estabeleceu**-se a quantidade exata de água necessária para que cada argila apresentasse plasticidade (Tabela III.3).
fazer	pv	Assim, na indústria de petróleo se reconhece a condição do efluente residual gerado de forma exponencial, sabendo que este é nocivo ao meio ambiente, diante desta situação se **faz** necessário desenvolver e aprimorar métodos de tratamento para todos esses resíduos.
	pass	**Faz**-se uso da cal (CaO) para elevar o pH da água fornecendo a alcalinidade necessária e o carbonato de sódio (Na2CO3), fornecendo a alcalinidade para a reação e também os íons carbonatos necessários. (FERNANDES, 1995; HANSEN, 1994).
iniciar	pv	A medida que o petróleo é extraído e transportado desde o poço à superfície, vai perdendo calor para o ambiente, com esta diminuição da temperatura, segundo Morán (2007), mais precisamente no ponto de névoa, cada componente parafínico torna-se menos solúvel até que a cristalização da parafina se **inicie**.
	pass	Após isso, **inicia**-se a obtenção dos sistemas microemulsionados.
introduzir	pv	Em seguida assume uma orientação NNE, mais retilínea, ao largo da Restinga da Marambaia e se **introduz** na área emersa cortando a Serra da Carioca ao longo da cidade do Rio de Janeiro.
	pass	Apenas pela presença de um filme emulsificante **introduz**-se uma barreira que previne o processo de quebra ou separação (Schramm, 1992).

ligar	pv	Segundo seus estudos, o hexacarbonil se **liga** à zeólita por meio de ligação entre o oxigênio do CO e o íon Na+.
	pass	O presente trabalho se propõe a estudar o procedimento de repartida, realizado **ligando**-se somente a bomba de água, de uma linha de escoamento óleo pesado / água em padrão core flow que sofre uma parada inesperada das bombas de óleo e água, a fim de contribuir para a consolidação prática da técnica de core flow.
manter	pv	Portanto, em baixas concentrações, este fluido apresenta-se como um fluido Newtoniano, a viscosidade se **mantém** constante com a variação da taxa de cisalhamento.
	pass	Já para determinar o grau de tixotropia, foram necessários três minuto de experimento, no primeiro minuto, variou-se a taxa de cisalhamento de 5 a 1010 s-1, no segundo minuto **manteve**-se a taxa de cisalhamento em 1010 s-1, e no terceiro minuto variou-se a taxa de cisalhamento de 1010 a 5 s-1.
misturar	pv	Desta maneira, o antigo conceito de que água e óleo não se **misturam** sofreu mudanças significativas.
	pass	A reação de transesterificação é uma reação de equilíbrio e a transformação ocorre essencialmente **misturando**-se os reagentes.
mostrar	pv	Optou-se, no entanto, por representar alguns dos modelos de geometria complexa, carentes de homogeneidade, pela forma suavizada em cores que se **mostrou** mais expressiva para estes casos.
	pass	A seguir **mostra**-se o gráfico de consumo específico.
projetar	pv	Esta zona de fratura se **projetaria** na plataforma continental, na região do Alto de Cabo Frio, coincidindo aí com a Zona de Transferência do Rio de Janeiro.
	pass	Na equação mostrada acima, FCt representa o fluxo de caixa no t -ésimo período, I é o investimento inicial, n é o período para o qual se **projeta** o investimento e i é a taxa de desconto definida, podendo ser a taxa mínima de atratividade do investimento.

propor	pv	O presente trabalho se **propõe** a estudar o procedimento de repartida, realizado ligando-se somente a bomba de água, de uma linha de escoamento óleo pesado / água em padrão core flow que sofre uma parada inesperada das bombas de óleo e água, a fim de contribuir para a consolidação prática da técnica de core flow.
	pass	**Propõe**-se que estes catalisadores sejam utilizados como alternativa aos catalisadores de metais nobres, devido à semelhança das propriedades catalíticas.
reduzir	pv	Depois de atingido um estado estacionário, um aumento na vazão do fluido aumentará também a retenção do polímero, e quando se volta ao valor inicial da vazão, a retenção do polímero se **reduz** ao valor inicial também.
	pass	As principais vantagens apresentadas por não realizar o processamento do gás offshore são: a redução de área ocupada e peso de equipamentos topsides, que são fortes limitações às plataformas e a transferência do processamento para o continente, **reduzindo**-se significativamente o custo das plataformas (diminuindo o CAPEX do pré-sal).