Proposta recebida em Agosto 2023 e aceite para publicação em Outubro 2023.

Recursos linguísticos para o PLN específico de domínio: o Petrolês

NLP resources for the oil & gas domain: Petrolês

Cláudia Freitas
PUC-Rio

Elvis de Souza
PUC-Rio

Maria Clara Castro
PUC-Rio

Tatiana Cavalcanti
PUC-Rio

Patricia Ferreira da Silva
Petrobras/CENPES

Fábio Corrêa Cordeiro
Petrobras/CENPES
FGV/EMAp

Resumo

Muitas organizações têm dificuldade em recuperar e extrair informações dos seus repositórios de documentos técnicos, em especial operadoras de óleo e gás que há várias décadas acumulam relatórios e documentos geocientíficos. No entanto, a maior parte dos recursos linguísticos para o processamento de linguagem natural é extraída de páginas da internet em inglês. Neste artigo, apresentamos os recursos linguísticos desenvolvidos ao longo do projeto Petrolês, com ênfase no PetroNer, *corpus* padrão ouro anotado com entidades do domínio, dependências sintáticas, e alinhado a uma ontologia de conceitos geológicos. Relatamos o processo de construção do PetroGold, *treebank* padrão ouro usado na geração de um modelo customizado para anotação de dependências sintáticas, e detalhamos o processo de anotação de entidades no PetroNer, realizado por meio de regras. Também realizamos um estudo sobre a aplicação das regras no *corpus* e, por fim, descrevemos características linguísticas do material que compõe o Petrolês, comparando-o com um *corpus* de textos jornalísticos.

Palavras chave

entidades mencionadas, ontologia geológica, dependências sintáticas, *universal dependencies*, *corpus* padrão ouro

Abstract

Many organizations struggle with retrieving and extracting information from their repositories of technical documents, particularly oil and gas operators with decades of accumulated geoscientific reports and documents. However, the majority of linguistic resources for natural language processing are derived from internet pages in English. In this article, we present the linguistic resources developed throughout the Petrolês project, with an emphasis on PetroNer, a gold standard corpus annotated with domain entities, syntactic dependencies, and aligned with an ontology of geological concepts. We report the construction process of PetroGold, a gold standard treebank used in generating a customized model for syntactic dependency annotation, and we detail the entity annotation process in PetroNer, carried out through the creation of linguistic rules. We also conduct a study on the application of rules in the corpus, and finally, we describe linguistic characteristics of the material comprising Petrolês, comparing it with a corpus of journalistic texts.

Keywords

named entities, geology ontology, syntactic dependencies, universal dependencies, gold standard portuguese *corpus*

1. Apresentação

Um dos requisitos para um Processamento de Linguagem Natural (PLN) bem-sucedido é a existência de recursos linguísticos de qualidade, capazes de oferecer sustentação para as diversas camadas do processamento automático de textos. Quando a tarefa envolve domínios de especialidade, como a área de petróleo, nosso foco, a necessidade de recursos de qualidade é ainda maior.

Modelos dedicados à resolução de correferência, tarefa e etapa importante na identificação de informação em textos, apresentam queda no desempenho quando aplicados a textos acadêmicos (Cohen et al., 2017). Já na análise sintática, um modelo treinado em um *corpus* composto por textos jornalísticos tem uma queda de mais de 10% em seu desempenho quando utilizado na anotação de um *corpus* do domínio biomédico (Thompson et al., 2017).

Mesmo com a popularização de grandes modelos de linguagem (LLMs), a existência de conjuntos de dados linguísticos de qualidade, pro-

DOI: 10.21814/lm.15.2.412

duzidos originalmente na língua de interesse e específicos para um domínio continuam recursos valiosos. Quanto mais cuidado na preparação dos dados, melhor é a qualidade das predições, com a vantagem de serem necessários menos dados para atingir bons resultados (Souza et al., 2020; Lewkowycz et al., 2022; Samuel et al., 2023).

Petrolês é simultaneamente um *corpus* e um repositório de artefatos de PLN especializados no domínio de petróleo em Português, cujo objetivo é servir como referência para os grupos de pesquisas em inteligência artificial e empresas atuantes nesse domínio. Atualmente, o repositório Petrolês[1] conta com conjuntos de dados linguísticos, como *corpora*, e modelos vetoriais pré-treinados especializados no domínio (Gomes et al., 2021). Neste artigo, apresentamos os *corpora* padrão ouro produzidos e disponibilizados pelo projeto, com especial atenção ao PetroNer, um *corpus* com anotação de entidades do domínio, criado para auxiliar extração de informação em documentos técnicos.

Ao longo de 4 anos de Petrolês foram produzidos o *corpus* Petrolês, composto por documentos acadêmicos (monografias, dissertações e teses), boletins e relatórios técnicos em formato texto simples (Cordeiro, 2020); o PetroTok, um subconjunto padrão ouro no que se refere às etapas de pré-processamento textual, especialmente sentenciação (Cavalcanti et al., 2021); o PetroGold, um *treebank* — *corpus* com anotação sintática — com anotação padrão ouro relativa a dependências sintáticas (de Souza et al., 2021; de Souza, 2023), Petro1 e Petro2, pequenos *treebanks* também padrão ouro, e o PetroNer, que além de dependências sintáticas, contém anotação padrão ouro relativa às entidades de interesse do domínio.

Para a anotação morfossintática, utilizamos o *framework* do projeto *Universal Dependencies* (de Marneffe et al., 2021), e com isso também buscamos alinhar os *treebanks* do Petrolês a *treebanks* de outras línguas, contribuindo para a inserção da língua portuguesa no contexto do processamento multilíngue. A anotação das entidades contou com a supervisão de especialistas da área, e foi realizada por meio da aplicação de regras linguísticas. Todo o procedimento de anotação semântica se inspirou no corte-e-costura (Mota & Santos, 2009; Santos & Mota, 2010), ferramenta de auxílio à anotação semântica dos *corpora* do projeto AC/DC (Santos & Sarmento, 2003; Santos, 2011).

[1] https://petroles.puc-rio.ai/

A construção de *corpora* anotados — em nosso caso, motivada primeiramente pelas demandas do PLN — envolve a tomada de decisões variadas, que incluem a compilação de material que seja representativo daquele para o qual a aplicação foi pensada, e variado com relação aos fenômenos (linguísticos) que contém, consistente com relação às anotações codificadas, bem documentado, e com tamanho que permita treinar e avaliar modelos de aprendizado de máquina ou, pelo menos, avaliar modelos e ferramentas.

No Petrolês, a construção de cada recurso anotado foi acompanhada de algum tipo de estudo experimental. A anotação morfossintática do Petro1, Petro2 e PetroGold possibilitou um estudo sobre métodos de revisão de *treebanks* (Freitas & de Souza, 2023; de Souza, 2023) e um estudo sobre o impacto de representações do conhecimento linguístico no desempenho de modelos de aprendizado de máquina (de Souza & Freitas, 2023). A construção do PetroNer, por sua vez, permitiu um estudo sobre o desempenho de um anotador baseado em regras na anotação de entidades do domínio, apresentado aqui.

Neste artigo, detalhamos as etapas de construção do PetroNer, que partiu de um léxico produzido por especialistas da área, e foi anotado com base em regras linguísticas. Na seção 2 apresentamos *corpora* criados com objetivos semelhantes aos do PetroNer; na seção 3 relatamos brevemente a construção do PetroGold, que foi utilizado como material de treino para o modelo de dependências customizado que anotou o PetroNer. Na seção 4 detalhamos a construção do PetroNer, um *corpus* multicamadas e padrão ouro quanto a entidades mencionadas do domínio. Ainda na seção 4, apresentamos o PetroNer como um *benchmark*, e simulamos o desempenho de um anotador baseado em regras na anotação de entidades. Na seção 5, discorremos sobre as características linguísticas do Petrolês e seu possível impacto no desenvolvimento de modelos de linguagem. Por fim, na seção 6, fazemos nossas considerações finais.

2. Trabalhos relacionados

Dois *corpora* nos serviram de inspiração para o PetroNer: o *corpus* (e projeto) GENIA (Kim et al., 2003; Thompson et al., 2017) e o *corpus* CRAFT (Cohen et al., 2017). O GENIA é um *corpus* da área biomédica, composto por 2 mil resumos/400 mil palavras e quase 100 mil anotações de termos biológicos. Possui uma anotação multicamadas, com POS (*part-of-speech*, ou classes gramaticais), sintaxe, entida-

des, eventos e relações, e foi anotado manualmente. Para a anotação dos termos biomédicos, a equipe do GENIA criou, a partir de *corpus*, sua própria ontologia, e a partir dela foi feita a anotação. Ao longo dos anos, as anotações foram continuamente enriquecidas, fazendo com que o GENIA se tornasse um *corpus* de referência no treinamento e avaliação de diversos sistemas do domínio biomédico.

O CRAFT (Colorado Richly Annotated Full Text) é um *corpus* composto por artigos de biomedicina, com 560 mil *tokens*/21 mil frases, e que também conta com diversas camadas de anotação – morfossintaxe, entidades mencionadas e correferência. Possui as mesmas motivações do GENIA e foi inspirado por ele. No entanto, o CRAFT surgiu da necessidade de se trabalhar com artigos completos, e não apenas com resumos de artigos, após o reconhecimento, segundo os autores, de que o corpo do texto trazia mais informações e que a estrutura textual dos resumos era diferente daquela encontrada nos artigos completos. Do ponto de vista da anotação de entidades biomédicas, a motivação para o esquema de anotação criado pelo CRAFT foi ampliar as classes semânticas utilizadas, circunscritas inicialmente aos genes e produtos a eles relacionados, a fim de possibilitar pesquisas relacionadas a outras classes de entidades.

Exceto pelo CRAFT, boa parte dos *corpora* anotados específicos de domínio, como o *corpus* GENIA, é composta apenas por resumos ou por extratos de parágrafos (por exemplo, Gábor et al. (2018) e Augenstein et al. (2017)).

Para a língua portuguesa, temos os *corpora* de domínios criados por Lopes & Vieira (2013), que foram anotados em formato XML pelo *parser* PALAVRAS, e utilizados sobretudo para estudos relativos à extração de terminologias. Dentre esses *corpora* está o GeoCorpus, um *corpus* com anotação de entidades específico para Bacia Sedimentar Brasileira (Amaral et al., 2017). No entanto, o material tem um escopo de entidades mais restrito, e o processo de anotação não pôde contar com a participação direta e intensa de especialistas. O material possui 5.275 frases e, na versão revista,[2] contém 6.126 anotações de entidades.

Nos *corpora* do Petrolês, especificamente PetroGold e PetroNer, todo conteúdo textual está preservado,[3] e as frases estão disponibilizadas na sequência em que foram escritas. No *treebank* PetroGold, foram excluídos dos documentos apenas informações consideradas irrelevantes para os objetivos da anotação e do projeto, como sumário, agradecimentos, folha de aprovação (no caso de teses e dissertações), lista de siglas e a seção de referências bibliográficas. No PetroNer, composto por boletins e relatórios técnicos, os documentos foram processados integralmente.

Na comparação com seus "pares" GENIA, CRAFT e GeoCorpus, o PetroNer contém 615 mil *tokens*/500 mil palavras e quase 19 mil entidades anotadas/27 mil anotações (uma entidade pode ser composta por mais de um *token*), em um processo de anotação cuidadoso que será descrito em 4.1.

3. Sintaxe e o PetroGold

No PLN, a relevância da análise sintática vai além da sintaxe propriamente, uma vez que este tipo de informação pode auxiliar a extração de relações semânticas, como demonstrado em Nooralahzadeh et al. (2018). Além disso, a atribuição de papeis semânticos, também considerada relevantes na identificação de conteúdo em textos, costuma ser feita a partir de *treebanks* de qualidade (Gildea & Jurafsky, 2000). De um ponto de vista metodológico, a existência de informação morfossintática de qualidade é capaz de otimizar outros tipos de anotação humana, como a anotação de entidades. Poder contar com generalizações linguísticas relativas a elementos coordenados, núcleos e modificadores é de grande ajuda na busca e na revisão semântica, como veremos.

O PetroGold é composto por teses e dissertações, e totaliza cerca de 250 mil *tokens*/9 mil frases. Levando em conta os objetivos do projeto, utilizamos como critério de seleção a presença de palavras candidatas à entidade por documento. A variedade lexical, medida pela distribuição *type/token* por documento, foi usada como critério complementar.

A anotação sintática utilizou a abordagem *Universal Dependencies* (UD) (de Marneffe et al., 2021). Escolhemos UD devido à sua crescente utilização na comunidade PLN, o que traz como benefícios não apenas o desenvolvimento e disponibilização de uma série de ferramentas associadas, como anotadores automáticos e ferramentas de auxílio à anotação e revisão de *treebanks*, mas também a possibilidade de alinhamento a outros *treebanks*, tanto de língua portuguesa quanto de outros idiomas, viabilizando estudos multilíngues e multigêneros.

A construção do *treebank* envolveu diferentes fases, e foi feita a partir da revisão de uma anotação automática. A revisão foi feita

[2] http://github.com/bsconsoli/GeoCorpus-V3

[3] Exceto para os casos de frases com problemas graves de tokenização e sentenciação, que foram excluídas.

inicialmente por 4 anotadores, já familiarizados com a abordagem UD. Na primeira fase de anotação/revisão, alguns documentos foram anotados por todos e as divergências discutidas em grupo, com consulta à documentação do próprio projeto UD. No entanto, a anotação de um *corpus* de domínio e gênero novos trouxe desafios linguísticos novos, discutidos nessa primeira etapa. Após a decisão sobre a melhor solução para cada caso, e sua respectiva documentação, a anotação propriamente começou. A concordância interanotadores foi medida utilizando a métrica Cohen's Kappa (Artstein, 2017). O melhor par na tarefa de anotação de relações sintáticas 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%.

Todo o processo de revisão e avaliação foi feito por meio da ferramenta ET (de Souza & Freitas, 2021), uma estação de trabalho para busca, edição e avaliação de arquivos no formato CoNLL-U.[4] A revisão foi feita no ambiente *Interrogatório* da ET, e a avaliação foi feita no ambiente *Julgamento*.

3.1. Procedimentos de revisão

Uma vez que já é possível contar com uma anotação sintática automática de qualidade razoável para o português (veja-se os resultados apresentados em Zeman et al. (2018) para a língua portuguesa), o processo de anotação do PetroGold foi, como indicado, um processo de revisão.

A literatura sobre métodos de revisão de *corpus* é mais rica quando se trata da revisão de *treebanks*, provavelmente devido à tradição deste tipo de anotação, mas também devido à sua dificuldade. Nossa principal preocupação nesta etapa esteve na pesquisa e desenvolvimento de métodos que permitissem uma revisão capaz de encontrar erros ou inconsistências sem precisar analisar todas as palavras do *corpus*. Isto porque, se já partimos de uma anotação de qualidade média, não precisamos passar por todas as palavras do *corpus* em busca de erros, uma vez que o reconhecimento de certas formas como "artigos", "verbos" ou "advérbios" não costuma trazer dificuldades para a análise automática. Além disso, uma mesma frase pode conter erros de diferentes naturezas, o que tem como consequência dificuldade em manter o foco e a consistência, podendo tornar o processo de revisão mais suscetível a erros e mais demorado. Wallis (2003), por exemplo, recomenda trocar uma revisão linear, *token* a *token*, por uma revisão transversal, guiada pelo tipo de fenômeno linguístico, que permitiria ver os fenômenos em questão de forma ampliada e garantiria uma revisão consistente.

A revisão do PetroGold utilizou quatro estratégias para detecção de erros e inconsistências — anotações variantes, discordância entre anotações, revisão guiada por regras e revisão guiada por léxico —, descritas a seguir.

1. A estratégia de anotações variantes (ou n-gramas variantes) se baseia nos trabalhos de Dickinson & Meurers (2003a) e Dickinson & Meurers (2003b). Usada inicialmente na revisão de POS, passou a ser também aplicada na revisão de sintaxe (Boyd et al., 2008; Dickinson, 2015; de Marneffe et al., 2017). Em termos gerais, a estratégia busca inconsistências na anotação, e parte da ideia de que palavras idênticas (ou n-gramas idênticos) anotadas de maneira diferente são candidatas à inconsistência – o que nem sempre é verdade, dada a ambiguidade da língua. Levada para a anotação de dependências sintáticas, a procura por inconsistências de anotação busca detectar (i) pares de palavras idênticos que (ii) possuam uma relação (de dependência) entre eles, mas que (iii) esta relação seja diferente em cada elemento do par.

2. A estratégia de discordância entre anotações também aposta na detecção de inconsistências, mas procura inconsistências não entre sequências de palavras idênticas em um mesmo *corpus*, mas entre análises automáticas de um mesmo *corpus*, dando continuidade à estratégia aplicada em Freitas & de Souza (2023). Trata-se de uma abordagem de revisão que se inspira no procedimento humano de adjudicação das análises na anotação, quando iremos lidar com análises divergentes, e por isso a batizamos de discordância entre anotações. No entanto, substituímos análises humanas por análises automáticas, e comparamos as análises por meio de uma matriz de confusão simplificada, que nos mostra apenas divergências quanto à anotação de relações de

[4] O formato CoNLL-U é uma adaptação do formato CoNLL-X desenvolvida pelo projeto *Universal Dependencies* com o objetivo de codificar os *treebanks* que integram o projeto. Neste formato, os metadados estão indicados no início de cada sentença e, em cada sentença, os *tokens* são dispostos em sequência, um por linha. A cada *token* – em cada linha – está associada informação linguística (anotação) em 10 campos separados por tabulação, tal como lema, classe gramatical, características flexionais, relação sintática etc. Para mais informações sobre o formato, veja-se: `https://universaldependencies.org/format.html`. Acesso em 6 de nov. 2023.

dependências sintáticas. Na revisão do *corpus*, comparamos análises fornecidas por duas ferramentas capazes de produzir bons resultados — Stanza (Qi et al., 2020) e UDPipe (Straka et al., 2016) — e trabalhamos sobre a saída da ferramenta com o melhor desempenho, Stanza, chamada "anotação guia". Isto é, se na comparação entre as duas análises, a anotação guia estiver correta, não precisamos fazer nada. Se a anotação "desafiante" estiver correta, ou se nenhuma das anotações estiver correta, precisaremos efetuar a correção. A estratégia de examinar, por meio da matriz de confusão, as divergências entre análises automáticas como potenciais casos de erro traz ainda a vantagem de permitir generalizar e criar hipóteses a partir dos tipos de erros — ou inconsistências — mais comuns, facilitando a percepção de padrões de erros, o que por sua vez (i) acelera a correção (erros de um mesmo tipo tendem a ter correções parecidas); (ii) permite o desenvolvimento de regras para auxiliar a detecção e correção, e (iii) contribui para o aperfeiçoamento da documentação, caso os erros sejam decorrência de lacunas das diretrizes de anotação. Além disso, cada pessoa responsável pela revisão pode selecionar um grupo de divergências (ou de confusões) para rever, o que dá mais agilidade ao processo e menos chances para inconsistências. Por fim, esta abordagem se baseia na hipótese de que duas ferramentas não cometerão os mesmos erros, ou seja, se há convergência, é porque existe acerto, o que nem sempre se verifica.[5]

3. A revisão guiada por regras linguísticas utiliza desde as regras de validação disponibilizadas pela equipe do projeto UD[6] até, e principalmente, regras relacionadas a fenômenos mais específicos e relevantes apenas para a língua portuguesa, que criamos ao longo do processo de revisão. 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 da anotação automática, e (iv) a exploração de erros detectados pelos outros métodos, e que puderam se transformar em regras de detecção de erros. A lista de regras desenvolvidas certamente não é exaustiva e padrões atípicos nem sempre são erros na anotação, sendo necessária verificação humana para corrigir os erros identificados. O conjunto de regras inclui regras que buscam eventuais erros formais introduzidos pelos anotadores durante a revisão do *corpus*, e regras específicas da estrutura da língua portuguesa. Como exemplo do primeiro tipo, temos uma regra que busca por ciclos na árvore sintática, quando um *token* é dependente de si mesmo. É um erro grave, que inutiliza a árvore sintática da frase, mas que pode ser introduzido sem que a pessoa responsável pela anotação perceba. Como exemplo do segundo tipo, temos uma regra que sinaliza quando há diferença entre os traços morfológicos de um adjetivo e do substantivo que ele modifica. Ao longo da construção do PetroGold, foram criadas 64 regras, que podem ser aplicadas na correção de outros *treebanks* de português que sigam a abordagem UD.[7]

4. A revisão guiada por léxico utilizou o PortiLexicon-UD (Lopes et al., 2022), léxico disponibilizado pelo projeto POeTiSA, e foi aplicada na revisão de lemas, anotação de POS e de características morfológicas. O PortiLexicon-UD é um recurso público e inclui 1.221.218 entradas (palavras ambíguas foram contabilizadas como entradas diferentes quando tinham classificações distintas) com informações morfológicas de acordo com a gramática UD. No processo de revisão, comparamos todas as entradas do léxico com todos os *tokens* do PetroGold. Como esperado, nem todas as palavras do PetroGold existiam no léxico, devido sobretudo à presença de terminologias da área. Quando havia pareamento entre formas do PortiLexicon-UD e formas do PetroGold, verificamos se alguma das anotações do PetroGold divergia do léxico e, se fosse um erro, corrigimos o *corpus*. Na comparação, desconsideramos as palavras cuja POS fosse PROPN, NUM ou X (nomes próprios, numerais e palavras estrangeiras, respectivamente), pois poucas dessas palavras existiam no léxico. Esta forma de revisão foi utilizada apenas na preparação da terceira e última versão do *corpus*.

[5] Buscamos, dentre os *tokens* corrigidos do *corpus*, quantos estavam invisíveis na matriz de confusão porque correspondiam a convergências entre as anotações. O resultado desta análise indicou que as convergências correspondiam a acertos em 94,7%, 95,3%, 98,4% nos casos de identificação do elemento que governa a relação sintática (dephead), do tipo de relação sintática (deprel) e de POS, respectivamente.

[6] As regras de validação UD estão em https://github.com/UniversalDependencies/tools/blob/master/validate.py.

[7] As regras criadas para o PetroGold estão em <https://github.com/alvelvis/ACDC-UD/blob/master/validar_UD.txt>

3.2. Criação do PetroGold

A primeira fase de elaboração do *treebank* padrão ouro produziu um pequeno *corpus*, chamado Petro1. No Petro1, que tem 22.288 *tokens*/652 frases e é composto por resumos e introduções de documentos que compõem o *corpus* Petrolês, todos os *tokens* foram revistos. Em seguida, fizemos uma segunda rodada de anotação, com um conjunto de cerca de 5 mil *tokens*, chamado Petro 2, no qual também todas as frases foram revistas.

Tanto Petro1 quanto Petro2 são pequenos para treinar um modelo de dependências sintáticas, mas podem ser utilizados para avaliação. Assim, este material viabilizou a realização de testes em busca do melhor modelo de anotação sintática, fundamental no processo de revisão do PetroGold.

A segunda fase de revisão já operou sobre o *corpus* PetroGold, que foi anotado automaticamente com um modelo customizado do anotador Stanza treinado com um material que combinava o *corpus* Bosque-UD (Rademaker et al., 2017), composto por textos jornalísticos, e os *corpora* Petro1 e Petro2. O material passou por uma revisão cuidadosa e deu origem ao PetroGold v1.

A terceira fase de revisão enfatizou o tratamento de fenômenos linguísticos específicos, decorrentes sobretudo de mudanças nas diretrizes do projeto UD (PetroGold v2), e a última fase de revisão também foi pautada pela revisão de fenômenos linguísticos pontuais, dando origem à versão final do *corpus*, o PetroGold v3. A contribuição de cada método de revisão, considerando a versão final do *corpus*, está na Figura 1, e o processo de construção do PetroGold v3, bem como uma avaliação sobre a contribuição de cada método de revisão, estão detalhadamente descritos por de Souza (2023).

3.3. Avaliação

Ao longo de cada versão fomos medindo — indiretamente — a consistência interna da anotação por meio de uma avaliação intrínseca, usando sempre a mesma versão da ferramenta UDPipe. Ao longo das versões 2 e 3, foram extensivamente revistas etiquetas (estruturas linguísticas) consideradas opcionais em UD.

Por exemplo, a etiqueta *obl:arg* não é obrigatória em UD, uma vez que é uma especificação da etiqueta *obl*, destinada a elementos nominais preposicionados associados a verbos – no caso do subtipo *obl:arg*, os elementos são aqueles que nossa tradição gramatical convencionou chamar de "objetos indiretos" (como a palavra "sinergias" no exemplo (1)).

(1) *obl:arg*: Aprimoramentos nesta tecnologia **resultarão** em **sinergias** entre a tecnologia proposta e aumento de recuperação de petróleo, sendo uma recomendação para futuros desenvolvimentos.

Fizemos o mesmo para as etiquetas *expl:pv*, *expl:impers* e *expl:pass*, que são especificações da etiqueta mais geral *expl*, atribuída, entre outros casos, ao pronome expletivo *–se*. Dada a tradição gramatical do português de especificar o tipo de *–se* entre índice de indeterminação do sujeito (*expl:impers*, frase (2), pronome apassivador (*expl:pass*, frase (3) e partícula integrante do verbo pronominal (*expl:pv*, frase 4), optamos por incluir as especificações no *treebank*.

(2) *expl:impers*: A princípio, **trabalhou-se** com a hipótese de que, quanto maior o percentual de esmectita de uma argila, maior seria sua afinidade pelo metal.

(3) Somente ao misturar as duas fases é que **se adiciona** o agente modificador.

(4) *expl:pv*: Este estudo **se baseia** nas propriedades magnéticas dos minerais que **se concentram** nas rochas da crosta terrestre.

Se, por um lado, a introdução dessas etiquetas granulares foi motivada pelo tipo de informação linguística que codificam, que consideramos relevante para o processamento do conteúdo de textos, por outro lado, dificulta a comparação entre as versões do *corpus*, e por isso também medimos os resultados levando em conta o que chamamos de versões simplificadas, nas quais as etiquetas granulares são convertidas nas respectivas etiquetas mais gerais (por exemplo, *expl:pv*; *expl:impers* e *expl:pass* são convertidos em *expl*, a única etiqueta obrigatória). Os resultados estão na Tabela 1, onde se pode perceber um aumento de até 1,39 p.p. na métrica CLAS no que se refere ao desempenho do anotador automático treinado no PetroGold v3 com a simplficação das etiquetas.[8]

[8] A métrica UAS (*unlabeled attachment score*) avalia o acerto nas dependências sintáticas, sem levar em conta o tipo da relação sintática, a métrica LAS (*labeled attachment score*) avalia o acerto nas dependências e no tipo de relação, e a métrica CLAS (*content labeled attachment score*) avalia o acerto nas dependências e no tipo de relação, mas considera apenas as relações que se estabelecem entre palavras de conteúdo.

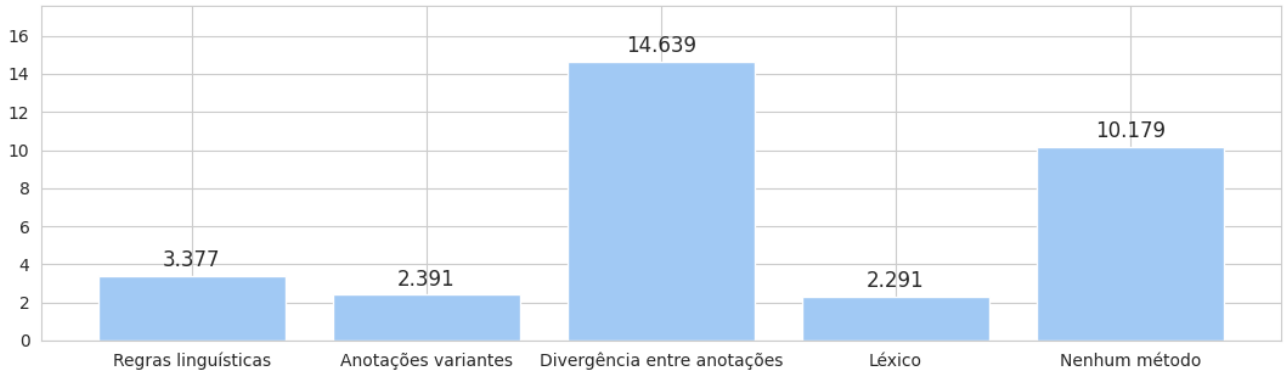

Figura 1: Distribuição da aplicação dos métodos de revisão no PetroGold v3, por *token*, em números brutos.

PetroGold	*Tokens* revistos	UPOS/simp.	UAS/simp.	LAS/simp.	CLAS/simp.
v1	—	98,19/—	90,65/—	88,53/—	82,96/——
v2	8.802	98,40/98,40	90,92/90,66	89,09/88,82	84,07/83,48
v3	9.314	98,63/98,63	91,04/92,02	89,36/90,92	84,22/85,61

Tabela 1: Evolução da anotação sintática do *corpus* PetroGold.

4. Entidades mencionadas e o PetroNer

Diferentemente do PetroGold, que contém documentos acadêmicos, o PetroNer é composto por boletins técnicos. A anotação de entidades está codificada na 9ª coluna do arquivo *CoNLL-U*,[9] e segue o formato de anotação IOB (Inside–Out–Beginning).

Frequentemente, entidades mencionadas são nomes próprios, mas como sinalizam Cohen et al. (2017) e Thompson et al. (2017), nomes próprios são menos relevantes e informativos em domínios técnicos, e é justamente a especificidade da terminologia a responsável por dificultar o processo de anotação. No PetroNer, nomes próprios continuam relevantes, mas, não são suficientes, e entidades podem ser nomes próprios, nomes comuns ou adjetivos.

No reconhecimento e classificação de entidades mencionadas, o principal desafio está na própria definição do que seja uma entidade do domínio – ou seja, o reconhecimento. Em um estudo realizado no âmbito do ACE (*Automatic Content Extraction*), uma das primeiras avaliações conjuntas sobre entidades mencionadas, uma equipe de anotadores experientes obteve concordância de apenas 82.8% na tarefa de identificação de entidades (Maynard et al., 2003). Para a língua portuguesa, no contexto do primeiro HAREM, um estudo de Mota (2007) verificou que, no que se refere à identificação (feita por pessoas), a concordância quanto às classes atribuídas ficou em torno de 45% (e 55% considerando apenas nomes próprios). Já na classificação, a concordância ficou pouco acima de 70%. Mais recentemente, na anotação do material para a tarefa 10 do SemEval (2017) (tarefa de identificação, classificação e relacionamento entre ternos-chave em publicações científicas das áreas de Ciência da Computação, Ciência de Materiais e Física), feita por estudantes de graduação e professores das referidas áreas, o índice de concordância quanto à identificação do que seriam os termos-chave do domínio ficou entre 45% e 85%, sendo a concordância igual ou maior a 60% em metade dos casos (Augenstein et al., 2017).

No PetroNer, a dificuldade na identificação foi contornada com a utilização de classes de entidades e suas instâncias definidas por um grupo de especialistas da Petrobras. Entidades são conceitos ou categorias usadas para agrupar objetos que possuem características próprias em comum. Já as instâncias são os objetos em si, que possuem as características definidas por uma entidade. Por exemplo, a entidade CAMPO corresponde, no domínio, aos campos de petróleo, que são áreas que delimitam estratos geológicos portadores de hidrocarbonetos passíveis de serem extraídos comercialmente. As instâncias *Campo de Marlim* e *Campo de Albacora* são dois exemplos de indivíduos agrupados sob o conceito de CAMPO.

[9] Originalmente esta coluna é dedicada às *Enhanced Dependencies*, mas está inutilizada no PetroNer.

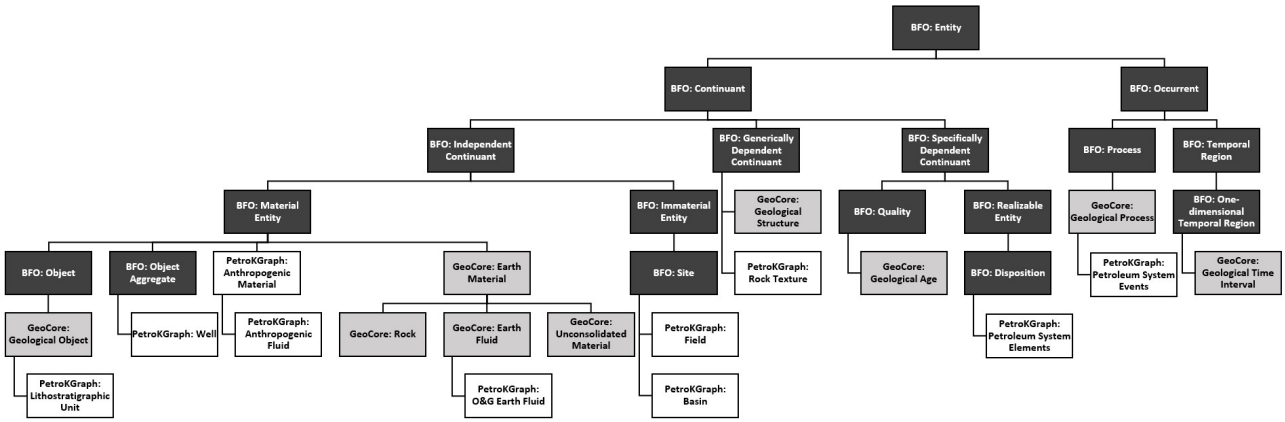

Figura 2: Árvore de relações hierárquicas do PetroKGraph.

Primeiramente, os especialistas indicaram quais tipos de entidades são importantes quando um geocientista busca por documento técnicos. Essas entidades foram então formalizadas em uma ontologia de aplicação denominada PetroKGraph, que importa e estende conceitos da *Basic Formal Ontology* (BFO)[10] e da GeoCore (Garcia et al., 2021). A BFO e a GeoCore são respectivamente ontologias de topo e de domínio (ou *core*) que possibilitam a formalização de complexos conceitos geológicos. Sempre que possível, os conceitos definidos na PetroKGraph buscaram ser compatíveis com outras ontologias desenvolvidas para o domínio de óleo e gás (Cicconeto, 2021; Cicconeto et al., 2020; Silva, 2022; Qu et al., 2023). A Figura 2 apresenta a taxonomia is_a da PetroKGraph.

Além das classes das entidades, os especialistas também compilaram listas de instâncias para cada classe. As listas não pretendiam ser exaustivas, mas abrangentes o suficiente para auxiliar no processo de anotação. Cada entidade anotada no *corpus* também recebe um identificador único referente à instância (codificado no campo *misc* do arquivo CoNLL-U, e atribuído ao primeiro *token* da entidade), além da anotação da classe da entidade, já mencionada. A Figura 3 ilustra um trecho do arquivo anotado, com destaque para a informação que codifica os identificadores de instâncias. Vale destacar que grafias alternativas (ou erros de digitação ou digitalização) de uma mesma instância recebem o mesmo identificador único, o que é valioso para tarefas de desambiguação e povoamento de grafos de conhecimento. Os trabalhos de formalização da ontologia e a construção de grafos do conhecimento oriundos dos documentos do Petrolês ainda estão em andamento e serão divulgados posteriormente.

A anotação de entidades foi feita com base em regras que utilizam informação morfossintática. Para tanto, utilizamos um modelo de anotação de dependências sintáticas gerado pela ferramenta Stanza (Qi et al., 2020) e treinado no PetroGold v2 completo.[11] Como todo o material do PetroGold foi utilizado no treinamento do modelo, para avaliar a qualidade da anotação sintática no PetroNer realizamos a revisão manual de 50 frases (1.658 *tokens*), selecionadas por terem muitos verbos (e supostamente serem mais complexas). Os resultados foram 98,37% (UPOS); 94,51% (UAS); 92,58% (LAS), e 87,23% (CLAS).[12]

O léxico inicial fornecido por profissionais de engenharia de petróleo e de geologia foi aplicado ao *corpus*, e criamos regras tanto para eliminar os casos errados como para incluir anotações que faltavam, produzindo uma anotação padrão ouro. Assim, a utilização de regras linguísticas, além de otimizar o processo de revisão, possibilitou a identificação de novas instâncias para as classes de interesse, enriquecendo o léxico inicial.

O processo de revisão do *corpus* e de criação de regras foi feito por duas linguistas (alunas dos anos finais de graduação em Letras) e contou com a supervisão direta de especialistas da área (geologia e engenharia de petróleo). O trabalho durou cerca de 8 meses e resultou na anotação de quase 20 mil entidades no PetroNer. Todo o trabalho foi auxiliado pelo ambiente *Interrogatório*, também utilizado na construção do PetroGold.

4.1. A anotação do PetroNer

A anotação automática das entidades é realizada em etapas. A primeira delas consiste na anotação do *corpus* com o léxico compilado por especialistas. O léxico contém 18 classes, com en-

[10] Basic Formal Ontology 2.0 Specification and User's Guide, https://github.com/BFO-ontology/BFO/raw/master/docs/bfo2-reference/BFO2-Reference.pdf

[11] A versão 3 ainda não estava disponível.

[12] Estas 50 frases com anotação padrão ouro estão disponibilizadas com o arquivo do PetroNer.

```
# text = Membro Mucuri, Eocretáceo da Bacia do Espirito Santo..
# sent_id = boletins-000001-7
1       Membro  Membro  PROPN   _       Gender=Masc|Number=Sing 0       root    B=UNIDADE_LITO  end_char=501|grafo=#membro_010|start_char=495
2       Mucuri  Mucuri  PROPN   _       Gender=Masc|Number=Sing 1       flat:name       I=UNIDADE_LITO  start_char=502|end_char=508
3       ,       ,       PUNCT   _       _       4       punct   0       start_char=508|end_char=509
4       Eocretáceo      Eocretáceo      PROPN   _       Gender=Masc|Number=Sing 1       conj    B=UNIDADE_CRONO|end_char=520|grafo=#LowerCretaceous|start_char=510
5-6     da      _       _       _       _       _       _       _       start_char=521|end_char=523
5       de      de      ADP     _       _       7       case    0       _
6       a       o       DET     _       Definite=Def|Gender=Fem|Number=Sing|PronType=Art        7       det     0       _
7       Bacia   Bacia   PROPN   _       Number=Sing     4       nmod    B=BACIA|end_char=529|grafo=#BASE_CD_BACIA_270|start_char=524
8-9     do      _       _       _       _       _       _       _       start_char=530|end_char=532
8       de      de      ADP     _       _       7       flat:name       I=BACIA _
9       o       o       DET     _       Definite=Def|Gender=Masc|Number=Sing|PronType=Art       7       flat:name       I=BACIA _
10      Espirito        Espirito        PROPN   _       Number=Sing     7       flat:name       I=BACIA|start_char=533|end_char=541
11      Santo   Santo   PROPN   _       Number=Sing     7       flat:name       I=BACIA|start_char=542|end_char=547
12      .       .       PUNCT   _       _       1       punct   0       start_char=547|end_char=548
```

Figura 3: Codificação das informações no *corpus* PetroNer.

tidades do tipo BACIA e UNIDADE LITOESTRATIGRÁFICA, e 383.168 instâncias distribuídas por essas classes (nem todas foram encontradas no *corpus*). Mesmo em um domínio técnico e lidando com terminologias, a ambiguidade está presente, e por isso a fase de revisão é fundamental. Neste ponto, a tarefa de anotação pode ser vista como uma tarefa de desambiguação. A palavra *bioturbação*, por exemplo, pode pertencer à classe ESTRUTURA FÍSICA ou à classe POROSIDADE; a palavra *água* pode ser uma entidade do tipo FLUIDO DA TERRA DE INTERESSE DA INDÚSTRIA (como no caso de *água de formação*), ou ainda FLUIDO ANTROPOGÊNICO (por exemplo, em *água destilada*), ou não ser uma entidade, e apenas o contexto (ou o conhecimento especializado) será capaz de indicar a anotação correta.

Em seguida, buscamos no *corpus* — já anotado automaticamente com dependências sintáticas — a distribuição, por lemas, para cada uma das palavras anotadas com alguma das 18 classes de entidades. Este passo buscava verificar, caso a caso, o que havia sido anotado com cada etiqueta. Após a análise da lista de lemas, (i) organizamos as palavras conforme seus contextos sintáticos, criando subgrupos de revisão; (ii) identificamos as colocações associadas a cada palavra anotada, e (iii) eliminamos as etiquetas daquelas que não eram entidades. Por exemplo, as buscas pelas colocações da palavra *campo* incluíram sequências como *campo magnético*, *campo de tensões*, ou *trabalho de campo*, contextos em que a palavra "campo" não é considerada entidade.

Também criamos regras para a identificação de falsos negativos, como a busca por palavras terminadas em *-iano* ou *-oceno* que não haviam recebido a etiqueta UNIDADE CRONOESTRATIGRÁFICA, mas que deveriam, e buscas por elementos coordenados ou em relação de aposição com entidades, mas que não haviam sido anotados, como no exemplo 5, que ilustra uma entidade não anotada (*Paraná*), coordenada a uma entidade anotada. Após as análises, as devidas correções foram realizadas por meio de regras.

(5) Entre 1981 e 1990 fez parte da equipe de avaliação de perfis e teste nas *Bacias de Campos*[BACIA] e do **Paraná**, dedicando-se à área de hidrodinâmica e hidroquímica.

Classes com (i) palavras muito frequentes e polissêmicas, como *água* e *óleo*, bem como (ii) entidades do tipo propriedades, como *pioneiro*, *especial* ou *de extensão*, foram inteiramente anotadas por meio de regras — dispensando a fase inicial de aplicação do léxico. No primeiro caso, as regras tornam a anotação menos custosa, uma vez que eliminar os casos errados seria mais trabalhoso. No segundo caso, identificamos os elementos nominais modificados pelos adjetivos (ou nomes modificadores) de interesse e apenas nesses contextos os adjetivos eram anotados como entidade.

Durante todo o processo de revisão, os casos duvidosos foram resolvidos por especialistas da área. Para facilitar a análise, usamos a anotação sintática e organizamos as candidatas a entidade segundo seus perfis lexicográficos, isto é, a palavra candidata à entidade associada a seus núcleos e/ou modificadores. Este procedimento de agrupamento linguístico das palavras candidatas também acelera o processo de criação de regras de revisão, explicitando contextos em que etiquetas devem ser eliminadas, incluídas ou modificadas. A Figura 4 ilustra o processo de construção do PetroNer, e a Tabela 2 traz a totalização de *tokens* e instâncias distintas anotadas para cada entidade encontrada no *corpus*.[13]

[13] Foram também anotadas as entidades TIPO_POROSIDADE, POÇO_T, POÇO_Q e POÇO_R que representam, respectivamente, o tipo de porosidade encontrado nas rochas, o tipo do poço, a classificação do poço segundo sua finalidade (e.g., *estratigráfico* ou *pioneiro*), e o papel do poço no desenvolvimento do campo de petróleo (*produtor*). Essas entidades apresentaram quantidade pouco significativa de *tokens* anotados e, por isso, não foram formalizadas na ontologia nem utilizadas para treinamento de modelos de PLN.

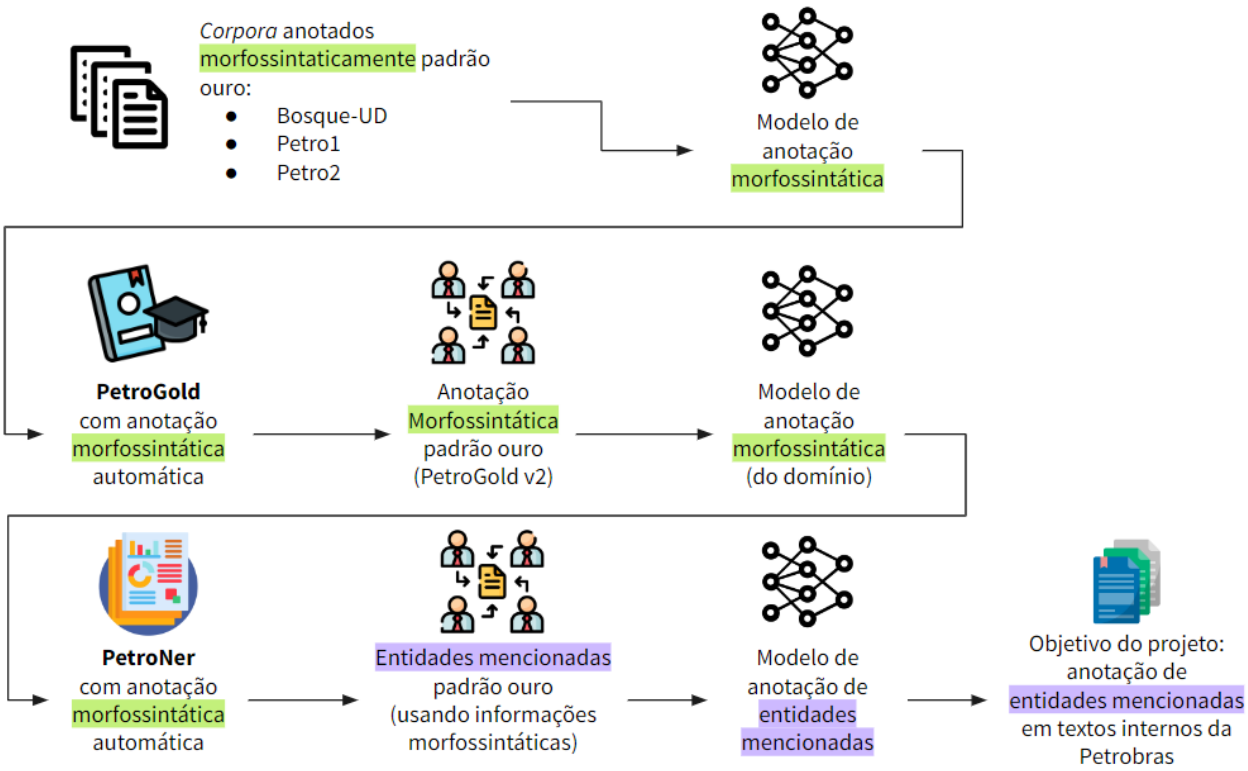

Figura 4: Fluxograma de construção do PetroNer.

4.2. Anotador baseado em regras

O processo de anotação de entidades tem início com a anotação dos arquivos CoNLL-U com as palavras vindas do léxico inicial. Nesta fase, são necessários alguns ajustes no nível de pré-processamento, como a remoção de acentos e a normalização, passando todas as palavras do léxico para minúsculas. Assim, se uma entrada do léxico é "BACIA DE CAMPOS", no texto do arquivo CoNLL-U tanto a ocorrência "bacia de Campos" quanto "BACIA DE CAMPOS" serão anotadas conforme a classe indicada no léxico, sendo a primeira palavra do tipo "B" (*begin*) e as demais do tipo "I" (*in*). Caberá às regras de correção, em outra etapa, corrigir estes casos específicos em que pode ter ocorrido algum erro. Além disso, em algumas palavras, como *Campos* e *Santos*, foi necessário manipular as lematizações produzidas pelo modelo de anotação morfossintática para que não fossem transformadas em *campo* e *santo*, respectivamente.

O léxico disponibilizado pelos especialistas de domínio é abrangente: contém um total de 382.120 entradas, distribuídas em 16 classes, e que serão aplicadas no *corpus*, como mencionamos, sem nenhuma restrição de contexto. A ideia é que a anotação via léxico seja a mais abrangente possível, abarcando todos os possíveis casos de entidades mencionadas do *corpus* — o que, inevitavelmente, irá incluir falsos positivos. É nesse contexto que se justifica a utilização de regras de anotação e de revisão linguisticamente motivadas, desenvolvidas, por um lado (e minoritariamente), para complementar os léxicos a partir das estruturas linguísticas que podem indicar novas entidades não previstas nos léxicos, e por outro lado (e majoritariamente), para realizar a revisão da anotação inicial proveniente dos léxicos, eliminando os falsos positivos. São as regras, portanto, que garantem a precisão da anotação de entidades mencionadas.

As regras de anotação e revisão foram organizadas em blocos conforme o tipo de ação que executam no processo de anotação, e para cada bloco é feita uma nova iteração em todo o *corpus*, tornando mais difícil que alguma regra deixe de ser aplicada devido à ordem em que está disposta no código. Existem 6 blocos (ou seis tipos de ação de anotação), descritos a seguir. Todas as regras foram criadas e testadas no *corpus* com a ferramenta *Interrogatório*, e só então foram incluídas em algum dos blocos.

1. Correções sintáticas — são as primeiras regras aplicadas, pois algumas das regras de anotação semântica dependem dos outros níveis linguísticos. São regras que corrigem erros da anotação morfossintática e, embora sejam as primeiras na ordem de aplicação,

Classe na ontologia	Entidade no PetroNer	*Tokens* no PetroNer	Instâncias distintas no PetroNer
PetroKGraph: Basin	BACIA	4.268	66
GeoCore: Geological Age and Geological Time Interval	UNIDADE_CRONO	3.104	134
GeoCore: Rock	ROCHA	2.772	79
GeoCore: Geological Structure	ESTRUTURA_FÍSICA	2.041	64
PetroKGraph: Lithostratigraphic Unit	UNIDADE_LITO	1.496	217
PetroKGraph: O&G Earth Fluid	FLUIDODATERRA_i	1.378	7
PetroKGraph: Well	POÇO	1.234	194
GeoCore: Unconsolidated Material	NÃOCONSOLID	1.035	11
PetroKGraph: Field	CAMPO	707	111
PetroKGraph: Petroleum System Events	EVENTO_PETRO	257	3
PetroKGraph: Earth Fluid	FLUIDODATERRA_o	246	1
PetroKGraph: Petroleum System Elements	ELEMENTO_PETRO	214	3
PetroKGraph: Athropogenic Fluid	FLUIDO	175	1
PetroKGraph: Rock Texture	TEXTURA	140	23

Tabela 2: Distribuição de entidades no PetroNer.

podem ser criadas a qualquer momento do processo de anotação, sempre que um erro de anotação do modelo de dependências sintáticas for encontrado. Este bloco de regras torna o PetroNer um material parcialmente revisto no nível morfossintático.

2. Adição de entidades novas — este bloco destina-se às entidades descobertas que não estavam no léxico inicial.

3. Expansão — regras que procuram candidatas a entidades a partir de certas estruturas linguísticas, como coordenação e aposto.

4. Revisão — regras que corrigem erros derivados da anotação inicial dos léxicos ou de outras regras anteriores.

5. Regras de mudança de entidade — regras aplicadas apenas no final e que têm como condição alguma palavra já anotada como entidade, isto é, já com alguma classificação semântica específica.

6. Regras de limpeza final (acabamentos) — regras que eliminam erros detectáveis pelo formato da anotação IOB. Por exemplo, uma regra que elimina a anotação da entidade I (*intermediate*) se o *token* anterior tem anotação de entidade de uma classe diferente (aplicada em casos como Bacia_B=BACIA de_I=CAMPO Campos_I=CAMPO, no qual a entidade do primeiro *token* é diferente das demais).

Ao final do processo, foram anotadas quase 20 mil entidades e foram "descobertas" 299 novas instâncias de entidades, enriquecendo 10 das 18 classes — por "descobertas", nos referimos às entidades que não haviam sido previstas nos léxicos especializados mas que, por meio da estrutura linguística das frases, conseguimos identificar como potenciais entidades relevantes para

o domínio, sendo posteriormente validadas pelos especialistas.

As classes que mais foram enriquecidas com dados do *corpus* foram *unidade cronoestratigráfica*, que recebeu 100 novas palavras, *unidade litoestratigráfica*, que recebeu 64, *poço* (56) e *bacia* (48).[14]

4.3. PetroNer como um *benchmark*

Dispondo de um *corpus* padrão ouro, e uma vez que a anotação foi inteiramente realizada por meio de regras, investigamos o comportamento das regras no *corpus* e simulamos o desempenho de um anotador baseado em regras.

A limitação de ferramentas de anotação baseadas exclusivamente em regras linguísticas quando aplicadas a textos inéditos é a impossibilidade de prever exatamente o que vai acontecer em uma boa parcela dos dados. Para anotar cerca de 20 mil entidades foram criadas quase 2 mil regras, e mais da metade delas (56.2%) foi aplicada apenas uma vez.

A tabela 3 traz a distribuição da frequência de aplicação das regras menos frequentes. A análise dos casos permite algumas observações sobre a abrangência das regras: (1) regras que foram aplicadas apenas uma vez (56% delas) são aquelas que realizam correções locais, em frases específicas, de erros da anotação via léxico que não puderam ser transformados em regras de correção gerais; (2) 33% das regras desenvolvidas, por sua vez, foram aplicadas mais de 4 vezes no *corpus*, indicando que são regras para fenômenos que se repetem com maior frequência no PetroNer e que, portanto, são potencialmente replicáveis em outros *corpora* do mesmo tipo, seja para revisar a anotação dos léxicos especializados, seja para encontrar novas entidades a partir da estrutura linguística dos textos.

Freq. aplicação	Qtd. regras
<= 4	1504 (77%)
<= 3	1427 (73%)
<= 2	1316 (68%)
= 1	1091 (56%)

Tabela 3: Distribuição da frequência de aplicação das regras menos frequentes na anotação do *corpus* PetroNer.

A fim de simular o desempenho de um anotador baseado apenas em regras, e, portanto, com poder limitado de generalização, anotamos o PetroNer utilizando as regras aplicadas com frequência maior ou igual a três. Os resultados estão na Tabela 4, assim como para regras com frequências próximas, para comparação. É importante mencionar, entretanto, que as regras foram criadas sem a preocupação sistemática de evitar redundância; o foco estava em corrigir problemas evitando produzir novos erros. Assim, é possível que uma análise cuidadosa das regras diminuísse a quantidade de regras com frequência de aplicação 1 e 2.

Tipo de regra	Precisão	Abrangência
freq. >= 4	98,1%	97,7%
freq. >= 3	98,4%	98,2%
freq. >= 2	98,9%	98,6%

Tabela 4: Distribuição da frequência das regras menos frequentes na aotação do *corpus* PetroNer.

As mesmas regras foram aplicadas a um conjunto de documentos que não podemos tornar públicos, a fim de verificar possibilidades de generalização das regras. O conjunto tem cerca de 250 mil *tokens* e foi anotado com o mesmo modelo de anotação de dependências usado no PetroNer. No entanto, em diversos documentos a qualidade do texto (originalmente PDFs transformados em texto simples) era ruim, prejudicando todo o fluxo de anotação. Neste contexto, a anotação baseada em regras conseguiu 94,1% de precisão e 85% de abrangência. Das 1.939 regras derivadas da criação do PetroNer, 1.594 não foram aplicadas nenhuma vez (eram regras de correção de anotação direcionadas para frases específicas que só existiam no PetroNer) — e 1.123 novas regras foram criadas ao longo do processo de tornar este material um pequeno *corpus* padrão ouro para avaliação.

Os resultados mostram que, embora uma grande quantidade de regras não tenha sido aplicada neste novo material — de fato, a maioria das regras —, aquelas que foram aplicadas, porque apresentam um alto grau de generalização, foram suficientes para produzir um outro *corpus* anotado com alta precisão (94,1%) e abrangência (85%).

[14] Todo processo de aplicação das regras, bem como as regras em si, estão disponíveis em https://github.com/alvelvis/Regras-PetroNer.

5. O idioma Petrolês

Embora pertençam ao gênero técnico-científico, PetroGold[15] e PetroNer têm origens distintas: o primeiro contém teses e dissertações; o segundo, boletins e relatórios técnicos, textos mais próximos daqueles para os quais o projeto Petrolês foi criado. Na busca por uma caracterização linguística do "idioma" Petrolês, verificamos o quanto o PetroNer se aproxima linguisticamente do PetroGold, *corpus* do qual "deriva" sintaticamente, o que traria consequências para a utilização de modelos treinados no segundo e aplicados ao primeiro, e o quanto ambos se aproximam (ou distanciam) de um *corpus* jornalístico como o Bosque, tendo em vista especialidades da linguagem técnica. A Tabela 5 apresenta características dos *corpora*.

Como podemos observar, a média de orações por frase é bem próxima entre os três *corpora*. Os números mais altos do PetroGold se devem a decisões de sentenciação que fizemos no pre-processamento, unindo itemizados como uma única frase caso o separador fosse vírgula ou ponto e vírgula. O PetroNer, ainda que também contenha listas e itens, não passou por um tratamento textual tão cuidadoso, e o Bosque, por sua vez, quase não contém este tipo de estrutura devido à natureza dos textos jornalísticos. Importante mencionar que na contagem de orações foram descartados os verbos que participam de expressões multipalavras, como em "ou *seja*", "a *partir* de" ou "*visto* que". Já quanto à quantidade de frases com pelo menos uma oração, os números se distanciam, e os números destacadamente mais baixos no PetroNer se devem à manutenção, neste *corpus*, de referências bibliográficas — estrutura linguística que costuma ser escassa em verbos — listadas ao final de cada relatório ou boletim. O mesmo motivo explica a alta proporção de frases sem oração no PetroNer. No PetroGold, apesar das referências bibliográficas terem sido excluídas, os títulos de capítulos, seções e subseções – estruturas que também costumam ser escassas em verbos, e que também estão presentes no PetroNer – explicam a diferença numérica para o Bosque. Corroborando essa prevalência verbal do Bosque, está a sua frequência relativa de verbos, que é de 10,5%, superior às frequências do PetroGold (9,2%) e do PetroNer (7,0%), como indica a tabela 6, complementar à Tabela 1. Por fim, assim como a baixa ocorrência de orações, a voz passiva — elemento capaz de impessoalizar textos — aparece como um elemento que contrasta o texto técnico-científico do jornalístico, sendo típico do primeiro.

A Tabela 6 traz a distribuição das classes de palavras mais frequentes em cada *corpus*, e a Tabela 7 a distribuição das relações sintáticas mais frequentes. Como podemos observar na Tabela 6, as principais diferenças estão na frequência dos nomes próprios, mais alta no PetroNer, como esperado, e na frequência dos verbos, mais alta no Bosque, como já discutido. Na comparação entre as relações sintáticas mais frequentes, a diferença mais visível está na classe *flat:name*, usada para os nomes próprios compostos, e por isso mais alta no PetroNer. Também é interessante constatar que, por um lado, a distribuição das classes no PetroGold e no PetroNer é muito próxima, e um dos pontos que os diferencia do Bosque é a frequência mais alta, neste último, da relação *nsubj*, utilizada para anotar sujeitos de orações ativas. Como já comentamos, a impessoalização aparece como uma característica de textos técnico-científicos, o que leva a uma diminuição na frequência de sujeitos explícitos neste tipo de material. Outra diferença está na alta frequência — 6ª posição — tanto no PetroGold quanto no PetroNer, de elementos coordenados, indicados pela relação *conj*. No Bosque, *conj* ocupa a 10 posição.

De um ponto de vista lexical, a comparação entre os adjetivos usados no PetroGold e no Bosque mostrou que, em relação a um total de 3.858 adjetivos, 653 (16,9%) eram compartilhados entre ambos os *corpora*, 1.242 eram exclusivos do PetroGold, e 1963 exclusivos do Bosque. Analisando os 50 adjetivos mais frequentes dentre os 1.242 adjetivos exclusivos do PetroGold, verificamos que 62% deles correspondem a adjetivos terminológicos, como "sedimentar", "deposicional" ou "estratigráfico", sugerindo que os termos específicos do domínio compõem a maior parte dos adjetivos que não são compartilhados pelo Bosque.

Na comparação entre os 50 verbos mais frequentes no PetroNer e no PetroGold, encontramos 72% de convergência. PetroGold e Bosque, no entanto, compartilham apenas 32% dos verbos, e PetroNer e Bosque apenas 30%.

O terceiro verbo mais frequente no Bosque, "dizer", sequer aparece entre os 50 mais frequentes do material Petrolês. O mesmo para o verbo "afirmar", também típico de discurso relatado, muito presente no jornalismo, com posição 16 no Bosque, e quase inexistente no Petrolês (posições 251 no PetroGold e 350 no PetroNer). Verbos típicos do Petrolês, por sua vez, como "observar"

[15] Todas as comparações foram feitas utilizando o PetroGold v2 pois o modelo de anotação sintática do PetroNer foi treinado nele.

	PetroGold v2	**PetroNer**	**Bosque-UD v2.12**
Número de orações	22.278	38.699	21.491
Frases com pelo menos uma oração	7.623 (85,2% de 8.949 frases)	14.267 (59,4% de 24.035 frases)	8.611 (92,0% de 9.357 frases)
Média de orações por frase	**2.9**	**2.7**	**2.5**
Número de frases sem oração	1.326 (14,8% de 8.949 frases)	9.768 (40,6% de 24.035 frases)	756 (8,0% de 9.357 frases)
Número de orações na voz passiva	4.245 (19% de 22.278 orações)	5.771 (14,9% de 38.699 orações)	1.681 (7,8% de 21.491 orações)

Tabela 5: Características dos *corpora*.

	PetroGold v2			**PetroNer**			**Bosque-UD**		
#	**upos**	**freq.**	**(%)**	**#**	**upos**	**freq.**	**(%)**	**#**	**upos**	**freq.**	**(%)**
1	NOUN		26,1	**1**	NOUN		21,8	**1**	NOUN		21,0
2	ADP		19,7	**2**	ADP		17,2	**2**	DET		17,7
3	DET		16,5	**3**	PROPN		16,1	**3**	ADP		17,1
4	VERB		9,2	**4**	DET		13,9	**4**	VERB		10,5
5	ADJ		7,7	**5**	ADJ		8,1	**5**	PROPN		9,5
6	PROPN		5,4	**6**	VERB		7,0	**6**	ADJ		5,8
7	NUM		3,3	**7**	NUM		4,5	**7**	ADV		4,3
8	AUX		3,0	**8**	CCONJ		2,7	**8**	PRON		3,8
9	CCONJ		2,9	**9**	ADV		2,6	**9**	SCONJ		2,7
10	ADV		2,8	**10**	PRON		2,2	**10**	CCONJ		2,7
11	PRON		2,6	**11**	AUX		1,8	**11**	AUX		2,5
12	SCONJ		0,8	**12**	SCONJ		0,8	**12**	NUM		2,4

Tabela 6: Distribuição das classes de palavras mais frequentes nos corpora.

(posição 4 no PetroGold e 5 no PetroNer) e "ocorrer" (posição 7 no PetroGold, e 3 no PetroNer), ocupam as posições 196 e 76 do Bosque.

Considerando os 50 substantivos comuns mais frequentes, há somente 54% de convergências entre PetroNer e PetroGold. Esta queda é resultado da constituição dos *corpora*, e se explica, novamente, pela presença de referências bibliográficas no PetroNer, mas não no PetroGold. Quando analisamos apenas as 50 palavras mais frequentes classificadas como nomes próprios, a diferença aumenta, e temos apenas 26% de convergência. Além das referências bibliográficas (que incluem nomes próprios de pessoas e de locais), a própria natureza dos boletins e relatórios, com mais entidades da área, explica a diferença.

Apesar da diferença nas classes nominais, PetroGold e PetroNer têm muitas semelhanças, por um lado, e divergências com um *corpus* jornalístico, por outro. A semelhança contribui para a boa performance do modelo de anotação morfossintática aplicado no PetroNer, e as diferenças entre petrolês e texto jornalístico ajudam a entender diferenças de desempenho entre analisadores automáticos preparados para lidar com um ou outro tipo de texto.

6. Considerações finais

Apresentamos aqui alguns recursos para o PLN de língua portuguesa, desenvolvidos ao longo do projeto Petrolês e sumarizados na Tabela 8.

Além de criarem condições para identificação e classificação de entidades de um domínio, recursos como o PetroNer e PetroGold permitem avançar com pesquisas na área, ajudando a responder questões como (i) se e quanto a incorporação de *embeddings* do domínio, como PetroVec (Gomes et al., 2021), facilita a tarefa de identificação de entidades; (ii) se e quanto a incorporação de dependências sintáticas facilita a tarefa de identificação de entidades; (iii) se e quanto a incorporação de *embeddings* do domínio facilita a tarefa de dependências sintáticas; (iv) se e quanto modelos gerais de língua têm o desempenho piorado quando aplicados a textos de um domínio específico.

PetroGold v2			PetroNer			Bosque-UD		
#	deprel	freq. (%)	#	deprel	freq. (%)	#	deprel	freq. (%)
1	case	17,7	**1**	case	15,5	**1**	det	17,5
2	det	16,1	**2**	det	13,4	**2**	case	16,6
3	nmod	11,4	**3**	flat:name	11,2	**3**	nmod	9,5
4	amod	6,6	**4**	nmod	10,4	**4**	nsubj	5,5
5	obl	5,4	**5**	amod	7,1	**5**	obl	5,1
6	conj	4,1	**6**	conj	6,6	**6**	obj	5,0
7	root	4,0	**7**	root	4,8	**7**	amod	4,8
8	flat:name	3,4	**8**	obl	4,3	**8**	root	4,7
9	nsubj	3,3	**9**	nsubj	2,7	**9**	advmod	4,0
10	cc	2,9	**10**	cc	2,7	**10**	conj	3,3
11	obj	2,9	**11**	obj	2,4	**11**	flat:name	2,9
12	advmod	2,5	**12**	advmod	2,4	**12**	cc	2,7
13	nummod	2,3	**13**	nummod	2,3	**13**	mark	2,7
14	acl	2,1	**14**	appos	2,3	**14**	xcomp	1,6
15	aux:pass	1,6	**15**	acl	1,8	**15**	appos	1,6

Tabela 7: Distribuição das relações sintáticas mais frequentes nos corpora.

Corpus	***Tokens***	**Frases**	**Anotação**	**Anotação padrão ouro**
PetroTok	38.472	1.139	não se aplica	tokenização e sentenciação
Petro1	22.288	652	lema, pos, morf, sintaxe	lema, pos, morf, sintaxe
Petro2	5.248	166	lema, pos, morf, sintaxe	lema, pos, morf, sintaxe
PetroGold	250.605	8.946	lema, pos, morf, sintaxe	lema, pos, morf, sintaxe
PetroNer	615.418	24.035	lema, pos, morf, sintaxe, entidades	entidades

Tabela 8: Características dos *corpora* do projeto Petrolês.

Durante o desenvolvimento dos recursos, investimos na dimensão metodológica da criação de recursos, e medimos *modos de fazer*. A construção do PetroGold permitiu investigar maneiras de buscar erros de anotação no *corpus* (e construir *treebanks* de maneira eficiente) e criou uma série de regras para detecção de erros em *treebanks* de língua portuguesa que sigam o formato e a gramática UD; a construção do PetroNer permitiu um estudo inicial sobre aplicação de regras, e propôs medidas que podem funcionar como *baseline* da tarefa de anotação de entidades.

Com a onipresença de LLMs (*Large Language Models*), pode parecer antiquado o trabalho de preparação de recursos como esses. No entanto, quando aplicados a áreas de especialidade cujos conteúdos não estão facilmente acessíveis, os modelos gerais tendem a ter um fraco desempenho, que por sua vez pode ser ajustado/customizado desde que existam os recursos adequados. Mas o próprio desempenho dos modelos só pode ser avaliado se existem meios para isso.

Agradecimentos

Esse trabalho foi realizado com o apoio da Petrobras e da Agência Nacional do Petróleo e Gás Natural e Biocombustíveis (ANP).

Referências

Amaral, Daniela, Sandra Collovini, Anny Figueira, Renata Vieira, Renata Vieira & Marco Gonzalez. 2017. Processo de construção de um corpus anotado com entidades geológicas visando. Em *11th Brazilian Symposium in Information and Human Language Technology*, 63–72.

Artstein, Ron. 2017. Inter-annotator agreement. Em *Handbook of linguistic annotation*, 297–313. Springer. 10.1007/978-94-024-0881-2_11.

Augenstein, Isabelle, Mrinal Das, Sebastian Riedel, Lakshmi Vikraman & Andrew McCallum. 2017. SemEval 2017 task 10: ScienceIE – extracting keyphrases and relations from scientific publications. Em *11th International*

Workshop on Semantic Evaluation (SemEval-2017), 546–555. 10.18653/v1/S17-2091.

Boyd, Adriane, Markus Dickinson & W Detmar Meurers. 2008. On detecting errors in dependency treebanks. *Research on Language and Computation* 6(2). 113–137. 10.1007/s11168-008-9051-9.

Cavalcanti, Tatiana, Aline Silveira, Elvis de Souza & Cláudia Freitas. 2021. Os limites da palavra e da sentença no processamento automático de textos. *Revista Brasileira de Iniciação Científica* 8. e021033.

Cicconeto, Fernando. 2021. *GeoReservoir: An ontology for deep-marine depositional system description*: UFRGS. Tese de Mestrado.

Cicconeto, Fernando, Lucas Valadares Vieira, Mara Abel, Renata dos Santos Alvarenga & Joel Luis Carbonera. 2020. A spatial relation ontology for deep-water depositional system description in geology. Em *XIII Seminar on Ontology Research in Brazil and IV Doctoral and Masters Consortium on Ontologies (ONTOBRAS)*, 35–47.

Cohen, Kevin Bretonnel, Karin M. Verspoor, Karën Fort, Christopher S. Funk, Michael Bada, Martha Palmer & Lawrence E. Hunter. 2017. The colorado richly annotated full text (craft) corpus: Multi-model annotation in the biomedical domain. Em *Handbook of Linguistic Annotation*, 1379–1394. Springer. 10.1007/978-94-024-0881-2_53.

Cordeiro, Fábio Corrêa. 2020. Petrolês-como construir um corpus especializado em óleo e gás em português. PUC-Rio. Monografia para obtenção do título de Especialização.

Dickinson, Markus. 2015. Detection of annotation errors in corpora. *Language and Linguistics Compass* 9(3). 119–138. 10.1111/lnc3.12129.

Dickinson, Markus & Detmar Meurers. 2003a. Detecting errors in part-of-speech annotation. Em *10th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, 107–114.

Dickinson, Markus & W Detmar Meurers. 2003b. Detecting inconsistencies in treebanks. *IEEE Transactions on Learning Technologies* 3. 45–56.

Freitas, Cláudia & Elvis de Souza. 2023. A study on methods for revising dependency treebanks: in search of gold. *Language Resources and Evaluation* 1–21. 10.1007/s10579-023-09653-4.

Gábor, Kata, Davide Buscaldi, Anne-Kathrin Schumann, Behrang QasemiZadeh, Haïfa Zargayouna & Thierry Charnois. 2018. SemEval-2018 task 7: Semantic relation extraction and classification in scientific papers. Em *12th International Workshop on Semantic Evaluation*, 679–688. 10.18653/v1/S18-1111.

Garcia, Luan Fonseca, Mara Abel, Michel Perrin & Renata dos Santos Alvarenga. 2021. The GeoCore ontology: A core ontology for general use in geology. *Computers & Geosciences* 135. 104387. 10.1016/j.cageo.2019.104387.

Gildea, Daniel & Daniel Jurafsky. 2000. Automatic labeling of semantic roles. Em *38th Annual Meeting on Association for Computational Linguistics (ACL)*, 512–520. 10.3115/1075218.1075283.

Gomes, Diogo da Silva Magalhães, Fábio Corrêa Cordeiro, Bernardo Scapini Consoli, Nikolas Lacerda Santos, Viviane Pereira Moreira, Renata Vieira, Silvia Moraes & Alexandre Gonçalves Evsukoff. 2021. Portuguese word embeddings for the oil and gas industry: Development and evaluation. *Computers in Industry* 124. 103347. 10.1016/j.compind.2020.103347.

Kim, J.D., T. Ohta, Y. Tateisi & J. Tsujii. 2003. GENIA corpus–semantically annotated corpus for bio-textmining. *Bioinformatics* 19(1). i182–i182. 10.1093/bioinformatics/btg1023.

Lewkowycz, Aitor, Anders Andreassen, David Dohan, Ethan Dyer, Henryk Michalewski, Vinay Ramasesh, Ambrose Slone, Cem Anil, Imanol Schlag, Theo Gutman-Solo, Yuhuai Wu, Behnam Neyshabur, Guy Gur-Ari & Vedant Misra. 2022. Solving quantitative reasoning problems with language models. ArXiv [cs.CL]. 10.48550/arXiv.2206.14858.

Lopes, Lucelene, Magali Sanches Duran, Paulo Fernandes & Thiago Pardo. 2022. PortiLexicon-UD: a portuguese lexical resource according to universal dependencies model. Em *13th Language Resources and Evaluation Conference (LREC)*, 6635–6643.

Lopes, Lucelene & Renata Vieira. 2013. Building domain specific parsed corpora in Portuguese language. Em *Encontro Nacional de Inteligência Artificial e Computacional (ENIAC)*, 1–12.

de Marneffe, Marie-Catherine, Matias Grioni, Jenna Kanerva & Filip Ginter. 2017. Assessing the annotation consistency of the Univer-

sal Dependencies corpora. Em *4th International Conference on Dependency Linguistics (DepLing)*, 108–115.

de Marneffe, Marie-Catherine, Christopher D Manning, Joakim Nivre & Daniel Zeman. 2021. Universal dependencies. *Computational linguistics* 47(2). 255–308. 10.1162/coli_a_00402.

Maynard, Diana, Kalina Bontcheva & Hamish Cunningham. 2003. Towards a semantic extraction of named entities. *Recent Advances in Natural Language Processing (RANLP)* 257–263.

Mota, Cristina. 2007. Estudo preliminar para a avaliação de REM em Português. Em *Reconhecimento de entidades mencionadas em português: Documentação e actas do HAREM, a primeira avaliação conjunta na área*, 19–34. Linguateca.

Mota, Cristina & Diana Santos. 2009. Corte e costura no AC/DC: auxiliando a melhoria da anotação nos corpos. Relatório técnico. Linguateca. http://hdl.handle.net/10400.26/20540.

Nooralahzadeh, Farhad, Lilja Øvrelid & Jan Tore Lønning. 2018. SIRIUS-LTG-UiO at SemEval-2018 task 7: Convolutional neural networks with shortest dependency paths for semantic relation extraction and classification in scientific papers. Em *12th International Workshop on Semantic Evaluation*, 805–810. 10.18653/v1/S18-1128.

Qi, Peng, Yuhao Zhang, Yuhui Zhang, Jason Bolton & Christopher D. Manning. 2020. Stanza: A Python natural language processing toolkit for many human languages. Em *58th Annual Meeting of the Association for Computational Linguistics (ACL)*, 101–108. 10.18653/v1/2020.acl-demos.14.

Qu, Yuanwei, Michel Perrin, Anita Torabi, Mara Abel & Martin Giese. 2023. GeoFault: A well-founded fault ontology for interoperability in geological modeling. *Computers & Geosciences* 105478. 10.1016/j.cageo.2023.105478.

Rademaker, Alexandre, Fabricio Chalub, Livy Real, Cláudia Freitas, Eckhard Bick & Valeria De Paiva. 2017. Universal dependencies for Portuguese. Em *4th International Conference on Dependency Linguistics (DepLing)*, 197–206.

Samuel, David, Andrey Kutuzov, Lilja Øvrelid & Erik Velldal. 2023. Trained on 100 million words and still in shape: BERT meets British National Corpus. Em *Findings of the Association for Computational Linguistics (EACL)*, 1954–1974. 10.18653/v1/2023.findings-eacl.146.

Santos, Diana. 2011. Linguateca's infrastructure for Portuguese and how it allows the detailed study of language varieties. *OSLa: Oslo Studies in Language* 3(2). 113–128.

Santos, Diana & Cristina Mota. 2010. Experiments in human-computer cooperation for the semantic annotation of Portuguese corpora. Em *7th International Conference on Language Resources and Evaluation (LREC)*, 1437–1444.

Santos, Diana & Luís Sarmento. 2003. O projecto AC/DC: acesso a corpora/disponibilização de corpora. *XVIII Encontro Nacional da Associação Portuguesa de Linguística (APL)* 705–717.

Silva, Patricia Ferreira da. 2022. *ResRiskOnto: an application ontology for risks in the petroleum reservoir domain*: PUC-Rio. Tese de Mestrado.

de Souza, Elvis. 2023. *Construção e avaliação de um treebank padrão ouro*: PUC-Rio. Tese de Mestrado. 10.17771/PUCRio.acad.62693.

de Souza, Elvis & Cláudia Freitas. 2021. ET: A workstation for querying, editing and evaluating annotated corpora. Em *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 35–41. 10.18653/v1/2021.emnlp-demo.5.

de Souza, Elvis & Cláudia Freitas. 2023. Explorando variações no tagset e na anotação universal dependencies (UD) para português: Possibilidades e resultados com base no treebank petrogold. Em *XIV Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana (STIL)*, 125–134.

de Souza, Elvis, Aline Silveira, Tatiana Cavalcanti, Maria Castro & Cláudia Freitas. 2021. Petrogold – corpus padrão ouro para o domínio do petróleo. Em *XIII Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana*, 29–38. 10.5753/stil.2021.17781.

Souza, Fábio, Rodrigo Nogueira & Roberto Lotufo. 2020. BERTimbau: Pretrained BERT models for Brazilian Portuguese. Em *Intelligent Systems*, 403–417. 10.1007/978-3-030-61377-8_28.

Straka, Milan, Jan Hajic & Jana Straková. 2016. UDPipe: trainable pipeline for processing CoNLL-U files performing tokenization,

morphological analysis, pos tagging and parsing. Em *10th International Conference on Language Resources and Evaluation (LREC)*, 4290–4297.

Thompson, Paul, Sophia Ananiadou & Jun'ichi Tsujii. 2017. The GENIA corpus: Annotation levels and applications. Em *Handbook of Linguistic Annotation*, 1395–1432. Springer. 10.1007/978-94-024-0881-2_54.

Wallis, Sean. 2003. Completing parsed corpora. Em *Treebanks: Building and Using Parsed Corpora*, 61–71. Springer Netherlands. 10.1007/978-94-010-0201-1_4.

Zeman, Daniel, Jan Hajič, Martin Popel, Martin Potthast, Milan Straka, Filip Ginter, Joakim Nivre & Slav Petrov. 2018. CoNLL 2018 shared task: Multilingual parsing from raw text to Universal Dependencies. Em *CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, 1–21. 10.18653/v1/K18-2001.