PetroGold – Corpus padrão ouro para o domínio do petróleo

Elvis de Souza[1], **Aline Silveira**[1], **Tatiana Cavalcanti**[1], **Maria Clara Castro**[1], **Cláudia Freitas**[1]

[1] Departamento de Letras
Pontifícia Universidade Católica do Rio de Janeiro

{elvis.desouza99,silveira26aline}@gmail.com

tatiana.shc@hotmail.com, mariaclarac@outlook.com

claudiafreitas@puc-rio.br

Abstract. *This paper describes the creation of PetroGold, a gold standard treebank for the oil & gas domain. It is composed of theses, dissertations and monographs, contains 9,127 sentences (253,640 tokens) and has morphosyntactic annotation of dependencies according to the Universal Dependencies approach. We detail some of the linguistic challenges of the domain for syntactic annotation and assess the quality of the corpus through an intrinsic evaluation: using a model created by the UDPipe tool, the corpus leads to 90.65%, 88.53% and 82.88% of correct answers according to the UAS, LAS and CLAS measures, respectively.*

Resumo. *Este trabalho descreve a criação do PetroGold, um treebank padrão ouro para o domínio do óleo & gás. O material é composto por teses, dissertações e monografias, contém 9.127 frases (253.640 tokens) e conta com anotação morfossintática de dependências segundo a abordagem Universal Dependencies. Detalhamos alguns dos desafios linguísticos do domínio para a anotação sintática e verificamos a qualidade do material produzido por meio de uma avaliação intrínseca: utilizando um modelo criado pela ferramenta UDPipe, o corpus leva a 90,65%, 88,53% e 82,88% de acertos conforme as medidas UAS, LAS e CLAS, respectivamente.*

1. Introdução

Um dos requisitos para um Processamento de Linguagem Natural (PLN) eficiente é a existência de recursos linguísticos de qualidade, capazes de oferecer sustentação para as diversas etapas do processamento automático. Embora, para a língua portuguesa, seja possível contar com bons corpora anotados de diversas naturezas – os treebanks do projeto Floresta Sintá(c)tica [Freitas et al. 2008], a Coleção Dourada do HAREM [Freitas et al. 2010], o corpus Summit++ [Antonitsch et al. 2016], o PropBank-Br [Duran and Aluísio 2011] e a quantidade crescente de material para a língua portuguesa associado ao projeto Universal Dependencies (UD) [Nivre et al. 2016], por exemplo – o cenário é menos favorável quando se trata de domínios específicos.

As características linguísticas de domínios de especialidade podem variar bastante quando comparadas a textos considerados de linguagem geral, como corpora jornalísticos.

As diferenças vão muito além do vocabulário, estando presentes também no nível sintático e discursivo. Outro aspecto dependente de domínio é a identificação dos limites das unidades linguísticas frase e palavra, o que acarreta dificuldades para os sistemas de PLN treinados em corpora jornalísticos.

[Thompson et al. 2017] relatam que o desempenho de um parser treinado no Wall Street Journal Treebank tem uma queda de mais de 10% quando aplicado, sem qualquer adaptação, a um corpus do domínio biomédico. Do mesmo modo, [Cohen et al. 2017] informam que sistemas dedicados à resolução de correferência em domínio geral não têm um bom desempenho quando aplicados a um corpus composto por textos acadêmicos.

Neste artigo, apresentamos as etapas de construção do treebank PetroGold, composto por teses, dissertações e monografias (253.640 tokens) relacionadas à indústria do petróleo. A anotação segue a abordagem gramatical do projeto Universal Dependencies (UD).

2. Petrolês e PetroGold

O PetroGold é um subconjunto do Petrolês, que é simultaneamente um corpus e um projeto. Enquanto projeto, tem como objetivo facilitar buscas semânticas em documentos da área; enquanto corpus, trata-se de uma coleção de documentos de fontes públicas de referência na área do O&G, como a Petrobras e a Agência Nacional do Petróleo, Gás Natural e Biocombustíveis (ANP), contendo artigos, documentos acadêmicos, publicações periódicas, notas e estudos técnicos [Gomes et al. 2018].

Um treebank como o PetroGold tem utilidade para diferentes áreas. Do ponto de vista linguístico, permite pesquisas acerca de estruturas sintáticas tendo como base a língua em uso; do ponto de vista do PLN, permite a avaliação e o treinamento de sistemas de anotação sintática e serve ainda como subsídio para outras tarefas de PLN, como a extração de informação aberta [Gamallo et al. 2012].

Para construir um treebank padrão ouro, selecionamos um subconjunto do Petrolês de 19 teses e dissertações (253.640 tokens). Como o objetivo principal do projeto é viabilizar buscas semânticas, que por sua vez dependem da anotação de entidades do domínio (etapa futura do projeto), tomamos como critério para a seleção de documentos a presença de termos candidatos a entidade da área. Nesse primeiro momento, a anotação de entidades restringiu-se apenas à aplicação de um léxico de termos, sem revisão, compilado em [Evelyn 2021]. A Tabela 1 apresenta o PetroGold em termos quantitativos.

PetroGold	
Tokens	253.640
Palavras	223.707
Frases	9.127
Documentos	19

Tabela 1. Características do corpus PetroGold v1

Para a anotação morfossintática, utilizamos o framework do projeto Universal Dependencies, que contém 17 etiquetas para anotação de classes gramaticais e 37 etiquetas para relações sintáticas. Além das diretivas do projeto UD, questões específicas do tipo

de texto presente no Petrolês precisaram ser discutidas entre os anotadores para serem consistentemente aplicadas ao corpus tendo em vista nossos objetivos a médio prazo.

3. Desafios e opções linguísticas do PetroGold

Os desafios linguísticos na criação do PetroGold foram de dois tipos: pré-processamento e anotação morfossintática. Do ponto de vista do pré-processamento, os desafios da conversão dos arquivos PDF em arquivos de texto plano, mantendo as informações linguísticas relevantes para um corpus, foram discutidos em [Silveira et al. 2019]. Neste trabalho, nos concentramos sobre a etapa de segmentação de frases e palavras. Já os desafios morfossintáticos se relacionam ao fato de as diretivas do projeto Universal Dependencies serem genéricas e não contemplarem casos específicos do texto técnico-científico. Nesta seção discutimos alguns desses desafios e relatamos nossas opções linguísticas.

3.1. Pré-processamento

Na etapa inicial da segmentação, as primeiras unidades a serem definidas são as frases. A delimitação dessas unidades segue alguns critérios. O primeiro deles é que apenas ponto final, de exclamação e de interrogação são separadores de frases. Com isso, sinais de pontuação que poderiam ser entendidos como separadores – caso do ponto e vírgula e dos dois pontos, por exemplo – não foram caracterizados dessa forma, nem mesmo em seu uso mais frequente, como em listas e enumerações. Em ambos os casos, mesmo que haja uma quebra de linha decorrente de itemização, o fim da frase só acontece com o ponto final. Uma consequência dessa escolha quanto à segmentação é o alto número de coordenações, que passou a ser uma característica do corpus.

Outro critério de sentenciação é relacionado aos títulos e subtítulos de seções, em situação análoga ao caso das manchetes de jornal. Dado que, em geral, estes não apresentam ponto final, os segmentadores automáticos tendem a colocá-los juntos das frases que os precedem e/ou sucedem, sendo tratados como uma única frase. No entanto, a despeito da ausência de um ponto final, consideramos que os títulos precisam ser segmentados como frases autônomas, caso contrário não seria possível estabelecer uma relação sintática satisfatória entre as partes.

Nesta primeira versão do PetroGold, as frases cuja sentenciação automática fugia às diretivas foram eliminadas. Com essa decisão, eliminamos 10,3% da quantidade de frases que selecionamos inicialmente.

3.2. Anotação morfossintática

No que se refere à morfossintaxe, fenômenos típicos do gênero acadêmico e do domínio de óleo & gás, até então não abordados nas diretivas de anotação, precisaram de um tratamento sistemático, como é o caso das referências bibliográficas.

A anotação de referências bibliográficas apresenta duas exigências: (a) definir a relação entre os termos que constituem uma referência composta como ”McGurk et al. (1990)”, como ilustrado no exemplo (1), e (b) definir a relação entre a referência e o restante da frase, como ilustrado no exemplo (2).

(1) ***McGurk et al. (1990)****, analisando otólitos, encontraram evidências de redução no crescimento de larvas de arenque.*

(2) *Vale lembrar que essa técnica não é permitida dentro de os estuários (**FERREIRA, 2006**).*

Para lidar com a estrutura interna dos elementos que compõem as referências ”McGurk et al. (1990)” e ”FERREIRA, 2006”, temos quatro possibilidades de anotação conforme as diretivas de UD, cada uma com implicações linguísticas diferentes: adjunto adnominal (nmod), coordenação (conj), expressão multi-palavra lexical sem sintaxe (flat) e expressão multi-palavra lexical com alguma sintaxe (compound). Escolhemos *flat*, pois contempla a ideia de que ”McGurk et al.” (sem o ano de publicação) e ”FERREIRA, 2006” são, no contexto acadêmico de referências bibliográficas, um nome único, uma unidade que se refere a um trabalho específico, com uma sintaxe inexistente (ou ao menos uma sintaxe que não nos interessa marcar). Analisamos a relação entre ”1990” e ”McGurk” na frase (1) como *nmod*; quanto à frase (2), a relação sintática entre o núcleo do elemento entre parênteses, ”FERREIRA”, e a raiz da frase, ”Vale”, definimos como de *parataxis*.

Relações entre elementos nominais são muito comuns no domínio de óleo e gás. As convenções de anotação dessas estruturas em UD implicam distinções entre nomes próprios e comuns, o que pode ser extremamente difícil para não especialistas. Além disso, reconhecer qual relação sintática esses termos estabelecem entre si também é uma tarefa complicada. Apesar de a gramática UD dispor de algumas alternativas de classificação para expressões nominais e suas relações, as diretivas apresentam pontos ainda não completamente maduros, como a relação entre substantivos próprios e comuns, que pode ser do tipo adjunto nominal ou aposto, por exemplo[1]. No PetroGold, decidimos utilizar a etiqueta *nmod* (modificador nominal do tipo adjunto adnominal) para a maioria dos casos de nomes que modificam outros nomes. Para os casos que algumas gramáticas tradicionais chamam de *aposto especificativo* evitamos atribuir a etiqueta aposto pela dificuldade de decidir se estamos diante de expressões com dois núcleos – característica típica da etiqueta *appos* em UD – em expressões como ”formação Cidreira”. Assim, também analisamos essas expressões como adjunto adnominal (etiqueta *nmod*), tal como na Figura 1.

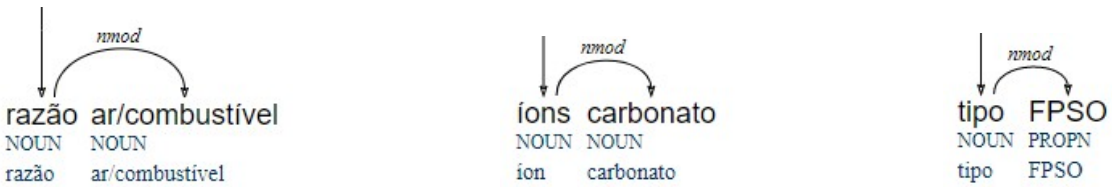

Figura 1. Anotação da relação entre nominais

Uma decisão diferente foi tomada em relação aos compostos químicos, nomes científicos e termos específicos do domínio, como ”alquil glucosídeos” e ”Odontesthes argentinensis”, ”desvio padrão” ou ”efeito estufa”. Todos foram anotados como *compound*.

Outras decisões de anotação do PetroGold estão descritas na documentação do

[1]Para um aprofundamento sobre as discussões, ver `https://github.com/UniversalDependencies/docs/issues/757` e `https://universaldependencies.org/workgroups/newdoc/two_nominals.html`. Acesso em 7 de ago. de 2021.

corpus.

4. Metodologia

Considerando os custos da anotação 100% manual e a capacidade de parsers atuais de produzirem resultados razoáveis, tem sido uma prática comum a criação de corpora padrão ouro por meio da revisão de material anotado automaticamente. Neste caso, o processo de revisão consiste, em grande parte, na busca por inconsistências e/ou erros da anotação automática.

No PetroGold, a revisão da anotação automática foi feita por 4 pessoas já habituadas à abordagem UD. Inicialmente, tendo em vista a familiarização com o gênero e o domínio do Petrolês, alguns documentos foram anotados por todos e as divergências discutidas em grupo. Avaliamos a concordância inter-anotadores utilizando o coeficiente κ (kappa) para cada par de anotadores. O par com concordância mais alta obteve índice de 95,1% na tarefa mais difícil, a análise das dependências sintáticas, e o par com menor concordância alcançou 91,9%.

A revisão do PetroGold foi feita sobre a saída de um modelo customizado treinado no parser Stanza [Qi et al. 2020] utilizando o Bosque-UD [Rademaker et al. 2017] v.2.6, corpus de textos jornalísticos, acrescido de um pequeno material do Petrolês. O processo de revisão demorou 3 meses, com 4 anotadores trabalhando 20 horas semanais.

Na elaboração do PetroGold, seguimos parte da estratégia de revisão utilizada em trabalhos anteriores, seguindo o que chamamos de método IAD (Inter-Annotator Disagreement) e regras linguísticas derivadas do IAD. O método IAD consiste em contrastar duas análises distintas para o mesmo corpus em busca de divergências. Contrastamos as análises fornecidas pelas ferramentas Stanza e UDPipe [Straka et al. 2016] por meio de uma matriz de confusão simplificada (Figura 2), e consideramos a ferramenta com o melhor desempenho – em nosso caso, Stanza – a anotação guia (ou anotador experiente), e UDPipe, a anotação desafiante. Trabalhamos sobre a saída da anotação guia, ou seja, se na comparação entre as duas análises a anotação guia estivesse correta, não era preciso realizar nenhuma modificação. Já se a anotação desafiante ou nenhuma das duas estivesse correta, era necessário editar o arquivo, que corresponde à saída do Stanza.

A estratégia de examinar, por meio da matriz de confusão, as divergências entre análises automáticas como potenciais casos de erro se baseia na hipótese de que se há convergência entre os anotadores, então existe acerto. A visualização das divergências pela matriz de confusão nos permite generalizar e criar hipóteses a partir dos tipos de erros – ou inconsistências – mais comuns, viabilizando a percepção de padrões nos erros e, consequentemente, a elaboração de regras capazes de detectá-los e corrigi-los de forma sistemática.

Por exemplo, a análise dos casos em que os sistemas divergiram entre a classificação de adjuntos adnominais (nmod) e adjuntos adverbiais (obl) nos possibilitou depreender um padrão: quando havia ocorrência de alguns adjetivos transitivos (como ”favoráveis”, na frase (3)), o sistema guia anotava seu complemento (”suprimento”) erroneamente como adjunto adnominal do substantivo (”condições”), e não como complemento do adjetivo (”favoráveis”).

(3) *Ainda segundo os citados autores, a sequência S4 é dominada por depósitos*

sistema	acl	acl:relcl	advcl	advmod	amod	appos	aux	aux:pass	case	cc	ccomp	c
golden												
acl	5070	8	128	0	370	12	1	0	1	0	2	1
acl:relcl	16	2246	60	0	21	4	17	1	0	0	42	0
advcl	305	73	2829	4	55	5	17	0	0	0	59	0
advmod	0	2	1	6549	24	8	3	0	236	35	1	0
amod	139	8	17	41	15970	75	4	2	19	2	12	3
appos	10	5	0	9	44	2389	0	0	15	0	1	5
aux	0	1	4	0	0	0	281	7	0	0	1	0
aux:pass	0	0	1	4	0	0	64	3590	0	0	0	0
case	0	0	3	76	10	24	0	0	45124	68	1	1
cc	0	0	4	117	0	17	0	0	49	7221	0	0
ccomp	5	28	73	5	11	0	4	2	2	0	520	0
compound	1	0	0	6	27	17	0	0	9	0	0	2
conj	69	69	132	50	113	375	20	0	35	0	47	1

Figura 2. Matriz de confusão de etiquetas sintáticas utilizada no método IAD

siliciclásticos que refletem ***condições*** *climáticas úmidas* ***favoráveis*** *ao* ***suprimento*** *sedimentar e desfavoráveis à formação de carbonatos.*

O sistema desafiante, por sua vez, costumava acertar esses casos. Assim, por um lado, a divergência serviu para nos mostrar uma questão complexa para o parser de melhor qualidade; por outro lado, nos indicou uma possível regra linguística para resolver a grande quantidade de erros desse tipo – na presença de certos tipos de adjetivo (que podemos chamar de adjetivos transitivos, como ”favorável”, ”constituinte” e ”existente”), as chances são altas de que o substantivo à direita complemente o adjetivo, e não o substantivo sendo adjetivado.

Como se trata de uma tendência, e não de uma regra determinística, criamos uma ferramenta que nos permite aplicar regras, analisar o resultado e aceitar a alteração apenas nos casos adequados ou realizar quaisquer outras modificações nas frases[2]. Para a análise das matrizes de confusão utilizamos o Julgamento, um ambiente para avaliação de corpora anotados [de Souza and Freitas 2021].

5. Resultado e análise

Os métodos utilizados na revisão do corpus resultaram na análise de 5.107 frases do PetroGold (55,9% do corpus), sendo alterada a anotação de 12.832 tokens (5,7% de todos os tokens). A Tabela 2 quantifica as correções realizadas no corpus por tipo de informação linguística considerando também as interseções (quando um token recebeu mais de uma correção). Além disso, indicamos quantos desses tokens corrigidos o método IAD, que busca por divergências na relação de dependência entre dois anotadores automáticos, identificou.

O método IAD indicou a presença de 25.123 tokens com anotação de relação de dependência divergentes. Desses, 5.656 tiveram alguma anotação corrigida. Por um lado, isso significa que mais da metade das correções realizadas no corpus não foi fruto direto

[2]Disponível em `https://github.com/alvelvis/conllu-merge-resolver`. Acesso em 7 de ago. de 2021.

Anotação	Tokens corrigidos	IAD
Qualquer correção	12.832	5.656
$LEMA$	2.258	768
POS	3.537	1.910
$HEAD$	6.780	2.865
REL	8.206	4.713
$LEMA \cap POS$	924	521
$LEMA \cap REL$	893	523
$POS \cap REL$	2.958	1.783
$HEAD \cap REL$	4.141	2.299

Tabela 2. Tipos de correção realizados no PetroGold

das matrizes de confusão (7.176, ou 55,9% das correções). No entanto, indiretamente as matrizes nos ajudaram a encontrar problemas na anotação das frases, apontando para fenômenos que precisam de atenção, como o exemplo *obl vs. nmod* apresentado na seção anterior. Esses casos foram identificados e corrigidos manualmente pelos revisores ou por meio das 25 regras de correção em lote desenvolvidas durante o processo de revisão.

Por outro lado, é interessante notar como a qualidade do anotador automático que utilizamos como sistema guia no método IAD (o Stanza) produziu resultados superiores aos do sistema desafiante (UDPipe), uma vez que, das 25.123 divergências identificadas, apenas 5.656 (22,5%) foram os tokens que precisaram de correção – nos outros 77,5% dos casos o sistema guia já estava correto. Dos tokens que foram corrigidos pelo método IAD, por sua vez, em 3.525 casos (62,3%) o sistema desafiante estava correto, enquanto que no restante dos tokens nenhum dos sistemas acertou e foi necessária uma terceira análise, humana.

Das 2.258 correções de lema, 1.247 (55,2%) não se associam a erro de qualquer outra informação linguística – são erros apenas de lema, decorrentes sobretudo da falta de familiaridade dos anotadores automáticos com as palavras do domínio, mas que, apesar da falha na lematização, não prejudicaram o parsing. Já os erros de POS parecem se associar diretamente às falhas na classificação da relação de dependência, uma vez que 83,6% dos erros de POS foram também erros de REL. O erro no encaixe de dependências sintáticas também se associa à falha na classificação da relação – 61% dos erros de HEAD também foram de REL. O contrário, no entanto, não é tão expressivo: apenas 50,4% dos erros de REL são também erros de HEAD.

Para uma avaliação intrínseca do PetroGold, separamos o corpus em partições de treino e teste e criamos um modelo utilizando o UDPipe v.1.2.0 sob os parâmetros de treinamento padrões da ferramenta. Como contraste, treinamos também um modelo a partir do Bosque-UD v.2.8, de textos jornalísticos, utilizando os mesmos parâmetros e garantindo a mesma distribuição de frases nas partições de treino e teste, seguindo a proporção de 95% e 5%, respectivamente, resultando em 8.671 frases para treino no PetroGold e 8.328 no Bosque-UD. São, portanto, dois corpora muito próximos em tamanho. Os resultados estão na Tabela 3.

As métricas de avaliação são as do CoNLL 2018 Shared Task [Zeman et al. 2018], onde UPOS avalia os acertos de classe gramatical, UAS avalia os acertos de encaixe

	LEMA (%)	UPOS (%)	UAS (%)	LAS (%)	CLAS (%)
PetroGold	98,48	98,19	90,65	88,53	82,96
Bosque-UD v.2.8	96,95	96,52	85,83	81,59	73,80

Tabela 3. Avaliação intrínseca de modelos treinados no UDPipe

de dependências sintáticas, LAS avalia a classificação das relações de dependência que foram corretamente encaixadas, e CLAS os acertos de LAS para as palavras consideradas de "conteúdo"[3].

Como podemos observar, o modelo treinado a partir do PetroGold apresentou desempenho significativamente superior. A diferença é de aproximadamente 1,5% para a lematização, mais de 1,5% para a anotação de classes gramaticais, mais de 4% no encaixe de dependências, 7% na classificação das relações e 9% na classificação das relações para palavras de conteúdo. Isso indica um grau de consistência interna maior na anotação do PetroGold, já que o modelo parece ter generalizado melhor durante o aprendizado, embora não seja possível estabelecer comparações diretas entre ambos os corpora. O gênero acadêmico, que pode ser bastante formulaico e previsível, também pode ter contribuído para os números mais altos do PetroGold.

6. Considerações finais

Apresentamos a primeira versão do PetroGold, um treebank padrão ouro para o domínio do petróleo. A intenção do material é servir como subsídio para o desenvolvimento de ferramentas de PLN específicas deste domínio, seja como material para treinamento de novos modelos ou para sua avaliação. O corpus está disponível na página do projeto Petrolês[4] e integrará o acervo do projeto Universal Dependencies.

Neste trabalho discutimos os desafios linguísticos relativos ao pré-processamento e à anotação morfossintática específicos de textos técnico-científicos do domínio do qual o Petrolês faz parte. Embora as questões linguísticas sejam específicas de um domínio, o procedimento de identificar as dificuldades no processamento automático e a forma como as resolvemos independem do tipo de texto.

Sugerimos ainda um caminho promissor para a revisão de corpora previamente anotados por ferramentas de PLN, resultando na revisão de mais de 50% das frases do corpus. Por fim, indicamos também a qualidade do material a partir da avaliação intrínseca de um modelo treinado a partir dele, chegando a 98% de acerto de classes gramaticais e 88% de acertos de dependências sintáticas.

Agradecimentos

Este trabalho foi financiado com o apoio da ANP – Agência Nacional de Petróleo, Gás Natural e Biocombustíveis, Brasil, associado ao investimento de recursos oriundos das Cláusulas de P, D & I, por meio de Termo de Cooperação entre a Petrobras e a PUC-Rio.

[3]As siglas significam, respectivamente, do inglês, Universal Part-of-speech Score, Unlabeled Attachment Score, Labeled Attachment Score e Content-Word Labeled Attachment Score.

[4]`https://petroles.puc-rio.ai`. Acesso em 7 de ago. de 2021.

Agradecemos à equipe do Laboratório de Inteligência Computacional Aplicada da PUC-Rio pela geração de modelos de anotação morfossintática customizados, e Elvis de Souza agradece ao Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) pela bolsa de mestrado de processo nº 130495/2021-2.

Referências

Antonitsch, A., Figueira, A., Amaral, D., Fonseca, E., Vieira, R., and Collovini, S. (2016). Summ-it++: an enriched version of the summ-it corpus. In Calzolari, N., Choukri, K., Declerck, T., Goggi, S., Grobelnik, M., Maegaard, B., Mariani, J., Mazo, H., Moreno, A., Odijk, J., and Piperidis, S., editors, *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, pages 2047–2051, Paris, France. European Language Resources Association (ELRA).

Cohen, K. B., Verspoor, K., Fort, K., Funk, C., Bada, M., Palmer, M., and Hunter, L. E. (2017). The colorado richly annotated full text (craft) corpus: Multi-model annotation in the biomedical domain. In *Handbook of Linguistic Annotation*, pages 1379–1394. Springer.

de Souza, E. and Freitas, C. (2021). 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. Association for Computational Linguistics.

Duran, M. S. and Aluísio, S. (2011). Propbank-br: a brazilian portuguese corpus annotated with semantic role labels. In *Proceedings of the 8th Brazilian Symposium in Information and Human Language Technology*.

Evelyn, W. F. D. (2021). Dos termos às entidades no domínio de petróleo. Master's thesis, PPGEL/PUC-Rio.

Freitas, C., Carvalho, P., Oliveira, H. G., Mota, C., and Santos, D. (2010). Second HAREM: advancing the state of the art of named entity recognition in Portuguese. In Calzolari, N., Choukri, K., Maegaard, B., Mariani, J., Odijk, J., Piperidis, S., Rosner, M., and Tapias, D., editors, *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2010)*, pages 3630–3637. European Language Resources Association.

Freitas, C., Rocha, P., and Bick, E. (2008). Um mundo novo na floresta sintá (c) tica–o treebank do português. *Calidoscópio*, 6(3):142–148.

Gamallo, P., Garcia, M., and Fernández-Lanza, S. (2012). Dependency-based open information extraction. In *Proceedings of the joint workshop on unsupervised and semi-supervised learning in NLP*, pages 10–18.

Gomes, D., Cordeiro, F., and Evsukoff, A. (2018). Word embeddings em português para o domínio específico de óleo e gás. In *Proceedings of the 19th Rio oil & gas expo and conference*, page 10.

Nivre, J., De Marneffe, M.-C., Ginter, F., Goldberg, Y., Hajic, J., Manning, C. D., McDonald, R., Petrov, S., Pyysalo, S., Silveira, N., et al. (2016). Universal dependencies v1: A multilingual treebank collection. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1659–1666.

Qi, P., Zhang, Y., Zhang, Y., Bolton, J., and Manning, C. D. (2020). 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*.

Rademaker, A., Chalub, F., Real, L., Freitas, C., Bick, E., and de Paiva, V. (2017). Universal dependencies for portuguese. In *Proceedings of the Fourth International Conference on Dependency Linguistics (Depling 2017)*, pages 197–206.

Silveira, A., de Souza, E., Cavalcanti, T., and Freitas, C. (2019). 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)*.

Straka, M., Hajic, J., and Straková, J. (2016). 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)*, pages 4290–4297.

Thompson, P., Ananiadou, S., and Tsujii, J. (2017). The genia corpus: Annotation levels and applications. In *Handbook of Linguistic Annotation*, pages 1395–1432. Springer.

Zeman, D., Hajic, J., Popel, M., Potthast, M., Straka, M., Ginter, F., Nivre, J., and Petrov, S. (2018). 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*, pages 1–21.