Explorando variações no *tagset* e na anotação *Universal Dependencies* (UD) para Português: Possibilidades e resultados com base no *treebank* PetroGold

Elvis de Souza[1], **Cláudia Freitas**[2]

[1]Departamento de Letras – PUC-Rio
Lab. Inteligência Computacional Aplicada – PUC-Rio

[2]Departamento de Letras – PUC-Rio

elvis.desouza99@gmail.com, claudiafreitas@puc-rio.br

Abstract. *The article analyzes variations in PetroGold, a gold standard treebank. The results show that considering the POS tag of multiword expressions in the annotation of all the words that compose them, as well as simplifying the syntactic tagset of the treebank, produces models with better performance on certain metrics, highlighting the importance of linguistic modeling during annotation for adequate natural language processing (NLP) results. The datasets used in the study are available in a dedicated repository and can be further modified to train better language models.*

Resumo. *O artigo analisa variações no PetroGold, um treebank padrão ouro. Os resultados mostram que considerar a classe gramatical das expressões multipalavras na anotação de todas as palavras que as compõem, assim como simplificar o tagset sintático do treebank, produz modelos com melhor desempenho em algumas métricas, destacando a importância da modelagem linguística durante a anotação para resultados adequados no processamento de linguagem natural (PLN). Os datasets utilizados no estudo estão disponíveis em um repositório dedicado, podendo ser ainda mais modificados para treinar melhores modelos de linguagem.*

1. Introdução

Corpora anotados padrão ouro são recursos de extrema relevância no atual cenário do processamento de linguagem natural, em que modelos de aprendizado de máquina podem se beneficiar dos dados para treinar modelos de predição e para avaliar os resultados dos modelos gerados. Por serem "padrão ouro", há a garantia de que tais recursos passaram por inspeção humana, de tal maneira que as análises linguísticas codificadas na sua anotação são as interpretações humanas dos fenômenos de linguagem. Contudo, o *tagset* e o esquema de anotação de um recurso – quais etiquetas e como serão utilizadas na anotação do *corpus* – pode variar de acordo com os objetivos para os quais o recurso está sendo desenvolvido.

Nesse contexto, partimos de um *corpus* padrão ouro e experimentamos algumas variações no seu *tagset* e na sua anotação com o objetivo de, por um lado, mostrar algumas das muitas possibilidades que um recurso como esse proporciona, fazendo mudanças na anotação que não comprometem a qualidade da informação linguística anotada, e por

outro, mostrar o impacto que essas modificações produzem no aprendizado de máquina, evidenciando o papel da modelagem linguística durante a tarefa de anotação na obtenção de resultados mais adequados para o Processamento de Linguagem Natural (PLN).

O PetroGold é um *treebank* padrão ouro composto por documentos do domínio do petróleo. Foi desenvolvido com o objetivo de gerar bons modelos de anotação morfossintática, e se insere em um cenário de poucos recursos padrão ouro para português – nenhum especificamente para o domínio do petróleo. Com o amplo uso de grandes modelos de linguagem (LLMs), a relevância de materiais customizados para um domínio e/ou língua fica ainda mais evidente, como mostram [Souza et al. 2020, Lewkowycz et al. 2022, Samuel et al. 2023], o que justifica o desenvolvimento de recursos como o PetroGold.

Embora os resultados apontem para a importância do desenvolvimento do esquema de anotação adequado na produção de bons modelos de linguagem, é importante ressaltar que o foco deste trabalho não é a avaliação dos modelos, mas o papel dos *datasets* no seu treinamento. Por isso, realizamos todos os testes utilizando sempre o mesmo algoritmo e hiperparâmetros, tendo como variável apenas os *datasets* com anotação modificada, que estão sendo disponibilizados em um repositório dedicado[1].

As variações incluem alteração das etiquetas de classes gramaticais para expressões multipalavras, simplificação do *tagset* de anotação sintática e mudança na forma de particionar as frases em conjuntos de treinamento, teste e desenvolvimento para o aprendizado automático. Além disso, tendo como pano de fundo a recém publicação da terceira versão do PetroGold, realizamos também uma breve comparação desta com a versão anterior, colocando em evidência o impacto das revisões linguísticas na geração de modelos de linguagem de melhor qualidade.

2. Versões e variações do PetroGold

O PetroGold é um *treebank* composto por teses e dissertações do domínio do petróleo, com as frases na sequência em que aparecem no texto[2]. O *corpus* contém anotação morfossintática padrão ouro no formato do projeto *Universal Dependencies* [de Marneffe et al. 2021] , uma iniciativa que visa tornar consistente a anotação gramatical em diferentes línguas.

A anotação do PetroGold foi obtida automaticamente, utilizando o anotador Stanza [Qi et al. 2020], e foi revista, na maior parte do tempo, por quatro anotadores familiarizados com a abordagem UD e com a ferramenta de busca, edição e avaliação de *corpora* chamada ET [de Souza and Freitas 2021]. Quando submetidos a um teste de concordância interanotadores [Artstein 2017], os anotadores alcançaram um índice *kappa* de até 95,1%, sugerindo a qualidade dos anotadores e, portanto, da revisão empregada.

O PetroGold é publicado em duas versões: uma versão para o projeto Petrolês, com um *tagset* ligeiramente diferente do *tagset* UD, e uma versão para o projeto *Universal Dependencies*[3], que segue as diretivas do projeto e pode ser obtida automaticamente a

[1]Disponível em: *https://github.com/alvelvis/petrogold-stil*.

[2]Do PetroGold foram eliminados apenas elementos como resumo, apêndice e a seção de referências bibliográficas, além de figuras, tabelas e fórmulas matemáticas, uma vez que atrapalham o processamento sintático. Para um detalhamento do material veja-se [de Souza 2023].

[3]Disponível em: *https://github.com/alvelvis/petrogold-stil*. Acesso em 13 de ago. 2023.

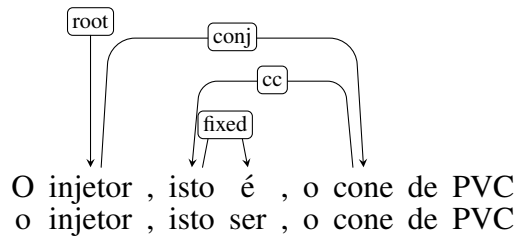

Figura 1. Anotação da MWE "isto é" de acordo com UD

partir da versão Petrolês.

O PetroGold passou por três fases de revisão (versao 1 [de Souza et al. 2021], versão 2 [de Souza and Freitas 2022a] e versão 3, final, apresentada aqui, e cujo processo de construção está detalhadamente descrito em [de Souza 2023]. Esta última versão traz um aumento expressivo no número de revisões da anotação linguística, que foram realizadas para endereçar fenômenos linguísticos anteriormente negligenciados ou para corrigir erros advindos da anotação automática que só puderam ser percebidos com a utilização de novos métodos de revisão semiautomáticos[4]. Entre as modificações realizadas, destacam-se a utilização de um léxico computacional, o PortiLexicon-UD [Lopes et al. 2022], para a revisão de lemas e características morfológicas, além da introdução de três novas etiquetas – *expl:impers*, *expl:pass* e *expl:pv* –, cuja anotação pode ser encontrada em detalhes em [de Souza and Freitas 2023c], e a consolidação de revisões para as expressões multipalavras (MWEs), utilizando três fontes diferentes para obtenção de candidatos a MWE e alinhando os resultados com as diretivas do projeto UD (exemplo na figura 1, e uma descrição sobre a anotação de MWEs no *corpus* pode ser encontrada em [de Souza and Freitas 2023a]).

Em relação ao particionamento das frases dos *datasets* em conjuntos de treinamento, teste e desenvolvimento, as versões Petrolês e UD do PetroGold realizam o procedimento da seguinte forma:

Petrolês O particionamento é realizado de forma *aleatória*, sendo o mesmo para as versões 1, 2 e 3 do *corpus*, garantindo que as versões são comparáveis, e seguindo a proporção de 90% de frases para treinamento, 5% para teste e 5% para desenvolvimento[5].

UD O particionamento é realizado *por documento*, de maneira a manter documentos inteiros em cada partição. Assim, as partições de treinamento, teste e desenvolvimento têm, respectivamente, 15 documentos (80% das frases), 2 documentos (12% das frases) e 2 documentos (8% das frases). Embora esta versão não possa ser diretamente comparada com as versões do PetroGold para o projeto Petrolês, o *corpus* pode ser comparado a versões recentes do Bosque-UD, nas quais essa recomendação de particionamento do projeto já é seguida.

[4]Para uma apresentação e avaliação dos métodos de revisão utilizados no desenvolvimento dessa nova versão, ver [Freitas and de Souza 2023, de Souza 2023].

[5]A proporção 90:5:5 para particionamento das frases foi a escolhida para se alinhar ao Bosque-UD v2.8 que, à época, seguia essa proporção e havia sido utilizado como base para comparação da qualidade do *treebank* [de Souza et al. 2021].

A tabela 1 resume as características de cada versão do Petrogold usada neste trabalho, incluindo o número de correções feitas em cada uma delas. O número de correções corresponde ao número de *tokens* que tiveram alguma das anotações linguísticas modificadas desde a anotação automática original. Para enriquecer a comparação, incluímos os dados do Bosque-UD, até agora o único *treebank* revisto em língua portuguesa que integra o acervo UD.

corpus	**frases**	**tokens**	**correções**
PetroGold-v3	8.946	250.605	30.948
PetroGold-ud-2.11	8.946	250.605	N/A
PetroGold-v2	8.949	250.595	21.634
bosque-ud-2.11	9.357	227.827	N/A

Tabela 1. Características dos *corpora*

Cada um dos 4 *datasets* explorados neste trabalho possui também até 3 variações possíveis no *tagset*: variação "base", variação "mwepos" e variação "simplificado". A variação "base" é a versão padrão dos *datasets*. No PetroGold v3, corresponde a um *tagset* com 5 etiquetas que nem todos os *corpora* disponíveis no projeto UD possuem. Quatro delas (*obl:arg*, *expl:impers*, *expl:pass* e *expl:pv*) são previstas nas diretivas do projeto UD, embora não sejam obrigatórias, e uma delas (*nmod:appos*) é uma criação nossa, para endereçar alguns fenômenos linguísticos específicos relevantes para o projeto Petrolês[6], conforme descrito a seguir:

obl:arg Um subtipo da relação *obl* (sintagmas preposicionados dependentes do verbo), exclusivo para quando o sintagma é argumento do verbo. A anotação já foi discutida anteriormente, em [de Souza and Freitas 2022b], e está exemplificada na figura 2.

subtipos de *expl* Etiquetas *expl:impers*, *expl:pass* e *expl:pv*, empregadas para especificar o pronome "se" (respectivamente, quando há indeterminação do sujeito, voz passiva sintética e verbo pronominal, frases 1, 2 e 3) [de Souza and Freitas 2023c].

nmod:appos A etiqueta foi criada para anotar os fenômenos da frase 4, em que o termo em negrito não é equivalente ao termo do qual depende sintaticamente, mas tem com ele uma relação não explícita que é facilmente interpretada. A etiqueta é empregada também na anotação da frase 5, em que há uma estrutura de hiperonímia, sendo os termos em negrito hipônimos dos termos do qual dependem, e na anotação da frase 6, na qual há uma referência bibliográfica, sendo que o ano de publicação é anotado como *nmod:appos* dependente do núcleo da referência.

1. ***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.
2. ***expl:pass***: Através dos mapas de contorno estrutural do Topo do Embasamento e Topo do Rifte **observou-se** a presença de um adensamento das isolinhas na direção NW-se adjacente ao Lineamento Tibagi interpretado na porção continental.
3. ***expl:pv***: Este estudo **se baseia** nas propriedades magnéticas dos minerais que **se concentram** nas rochas da crosta terrestre.

[6]Quando publicamos o PetroGold no projeto UD, realizamos a simplificação da etiqueta *nmod:appos* para *nmod*, o que está de acordo com as diretivas do projeto.

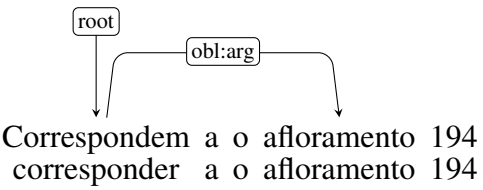

Figura 2. Anotação de argumento verbal introduzido por preposição (*obl:arg*)

4. ***nmod:appos***: Próximo a Presidente Olegário (**MG**) foram escritos em este estudo pacotes siliciclásticos relativamente espessos (até 60 m) pertencentes a esta formação.
5. ***nmod:appos***: Para fluidos Newtonianos, como a **água** e o ar, a viscosidade é independente de a taxa de cisalhamento.
6. ***nmod:appos***: A capacidade de absorção de o solvente é proporcional a a pressão parcial em a unidade de absorção (Gupta, **2003**).

A variação "mwepos" corresponde ao *dataset* quando damos às expressões multipalavras (MWEs) a anotação de classe gramatical correspondente à classe da expressão como um todo. Assim, na figura que vimos (figura 1), todos os *tokens* da expressão "isto é" recebem a etiqueta de POS "CCONJ" (para conjunção coordenativa), no lugar das etiquetas "PRON" e "AUX". A solução não é a adotada pelas diretivas do projeto UD, sendo utilizada neste trabalho apenas para evidenciar a possibilidade e comparar os resultados de aprendizado automático com uma opção linguisticamente motivada, apesar de contrária às diretivas do projeto[7].

A variação "simplificado" corresponde ao *dataset* sem a especificação das etiquetas de relação sintática: *obl:arg* é convertido em *obl*; *expl:*{*impers,pass,pv*} são convertidos em *expl*, e *nmod:appos* é convertido em *nmod*. Com as versões simplificadas, conseguimos comparar as diferentes versões do PetroGold, uma vez que igualamos os *tagsets*, possibilitando visualizar com maior clareza o impacto das revisões linguísticas nos resultados do aprendizado automático. Essa variação é também a que nos permite comparar o PetroGold ao Bosque-UD, uma vez que este não possui o *tagset* tão especificado quanto o do PetroGold, e a outros corpora em UD que não tenham usado essas etiquetas específicas.

A tabela 2 ilustra as diferenças nos *tagsets* e no particionamento das frases de todos os *datasets*. As células em cinza indicam as características dos *datasets* que os deixam em desacordo com as diretivas do projeto UD.[8]

[7] A informação relativa à classe "geral" das MWEs, codificada como "MWEPOS", já havia sido adotada – ainda que de forma assistemática e sem revisão – em versões iniciais do corpus Bosque-UD. Anotar MWEs como uma unidade, e não literalmente, como propõe UD, era (e é) a opção de análise do *parser* PALAVRAS [Bick 2014], responsável pela anotação original do Bosque.

[8] Notamos que o *dataset* "bosque-ud-2.11" encontra-se apenas na variação "simplificada" por dois motivos: primeiramente, porque é apenas nesta variação que podemos compará-lo aos outros *datasets*, e em segundo lugar, porque precisamos de fato realizar uma ligeira modificação no *tagset* do *corpus*, uma vez que continha 4 *tokens* anotados com a etiqueta *expl:pass*, sendo simplificados para *expl*.

variação	*dataset*	***obl:arg*** ***expl:impers*** ***expl:pass*** ***expl:pv***	***nmod:appos***	**MWEPOS**	**partições**
base	PetroGold-v3	x	x		aleatório
	PetroGold-ud-2.11	x			documento
mwepos	PetroGold-v3	x	x	x	aleatório
	PetroGold-ud-2.11	x		x	documento
simplif.	PetroGold-v3				aleatório
	PetroGold-v2				aleatório
	PetroGold-ud-2.11				documento
	bosque-ud-2.11				documento

Tabela 2. Conteúdo dos *datasets*

3. Metodologia

Para avaliar a qualidade das representações linguísticas codificadas nos *datasets*, usamos tanto uma avaliação intrínseca como uma avaliação extrínseca dos *datasets* [Freitas 2023]. Embora, tradicionalmente, avaliações intrínsecas e extrínsecas sejam usadas para verificar a qualidade de modelos/ferramentas, é possível uma mudança de perspectiva: se na avaliação intrínseca "original" verificamos a capacidade do modelo de generalizar a partir dos dados a que foi exposto, na avaliação intrínseca de *datasets* verificamos (indiretamente) o quanto o *dataset* permitiu esta generalização, levando em conta as características do modelo. A partir dessa mudança de perspectiva, quando olhamos para o desempenho de um modelo, vemos também até onde os dados permitiram ir, considerando os limites do modelo, e pressupondo que (i) o material que serviu de treino está bem anotado e que (ii) o modelo gerado tem um desempenho que não é aleatório. Seguindo com a inversão, a avaliação extrínseca verifica se a informação linguística codificada no *dataset* é adequada para as tarefas mais complexas que o *dataset* pretende auxiliar – o que fazemos quando medimos o impacto das mudanças na codificação de POS das MWEs na anotação de dependências sintáticas. Assim, a avaliação intrínseca de *datasets* anotados verifica a consistência da anotação, e a avaliação extrínseca verifica a adequação de uma anotação para uma determinada tarefas [Freitas 2023].

Os modelos treinados, um para cada *dataset* e variação, são gerados utilizando a ferramenta UDPipe [Straka et al. 2016] na versão 1.2.0, configurada com os parâmetros padrões da ferramenta. As métricas de avaliação intrínseca são as da avaliação conjunta do CoNLL de 2018 [Zeman et al. 2018], com enfoque nos resultados de UPOS (avaliação da anotação de classes gramaticais), LAS (avaliação da anotação da relação e do encaixe das dependências sintáticas) e CLAS (avaliação da anotação da relação e do encaixe das dependências sintáticas considerando apenas palavras de conteúdo lexical).

4. Resultados

A tabela 3 mostra os resultados da avaliação do modelo gerado utilizando cada um dos *datasets* como material de treino. Os números em negrito mostram a variação que treinou o modelo com melhores resultados segundo a métrica daquela coluna (UPOS, LAS ou CLAS) e segundo aquele *dataset* sendo avaliado na linha. Para uma uma análise

linguística detalhada da qualidade da anotação das relações sintáticas em termos de LAS e CLAS, bem como uma análise de erros, veja-se [de Souza and Freitas 2023b].

dataset	variação	UPOS (%)	LAS (%)	CLAS (%)
PetroGold-v3	base	**98,63**	89,66	84,66
	mwepos	98,49	89,87	84,79
	simplif.	**98,63**	**90,22**	**85,61**
PetroGold-ud-2.11	base	**98,42**	88,63	83,30
	mwepos	98,23	**89,48**	84,33
	simplif.	**98,42**	89,30	**84,38**
PetroGold-v2	simplif.	98,40	88,82	83,48
bosque-ud-2.11	simplif.	96,52	81,12	73,51

Tabela 3. Avaliação dos modelos gerados utilizando os diferentes *datasets*

Para os *datasets* "PetroGold-v3" e "PetroGold-ud-2.11", as variações que produziram melhores resultados de UPOS foram, empatadas, a "base" e a "simplif.". De LAS, foi a variação "simplif." para a v3 e "mwepos" para a ud-2.11, e de CLAS, foi a variação "simplif." para ambos.

O empate entre "base" e "simplif." na anotação de POS era esperado, uma vez que as variações não contêm nenhuma diferença na anotação de classe gramatical. Em relação à métrica LAS e CLAS, também era esperado que a variação simplificada obtivesse melhores resultados, uma vez que simplificar etiquetas significa necessariamente reduzir o grau de complexidade do que o modelo deve aprender (e para o que deve ser avaliado)[9]. A surpresa, porém, está no fato de que, usando a métrica LAS, os melhores resultados do "PetroGold-ud-2.11" foram da variação "mwepos", e não "simplif.", indicando que, embora a simplificação das etiquetas produza números melhores que a versão "base", nesse caso, modificar o POS das MWEs foi capaz de produzir resultados ainda melhores do que simplificar as etiquetas.

Como esperado, as variações com "mwepos" tiveram desempenho pior que as variações "base" em relação a POS, reforçando que a atribuição de classes de palavras de maneira estática, que não leva em conta o contexto em que as palavras estão inseridas, facilita a generalização das classes. Contudo, para o aprendizado de dependências sintáticas, foco do nosso interesse, vemos uma melhora de até 0,85 p.p. usando a métrica LAS, mostrando que, embora haja perda na anotação de POS, a anotação sintática se beneficia da mudança da classe gramatical das expressões multipalavras, evidenciando o impacto da anotação de um atributo linguístico (classe gramatical) no aprendizado de outro (relação sintática). A mesma tendência ocorre para CLAS, onde a melhora é ainda maior, de até 1,03 ponto percentual.

Considerando que a única diferença entre o "PetroGold-v3" simplificado e o "PetroGold-ud-2.11" simplificado é o modelo de particionamento, já que as etiquetas simplificadas são as mesmas, podemos concluir que, da forma como foram particionados,

[9]Os números devem ser lidos com cautela: embora os melhores resultados sejam os dos *datasets* simplificados, modelos treinados utilizando esses dados não serão capazes, por exemplo, de diferenciar objetos indiretos de adjuntos adverbiais (motivo pelo qual a etiqueta *obl:arg* foi introduzida), de maneira que cabe ao usuário decidir qual *dataset* deseja utilizar no treinamento do seu modelo, conforme seus objetivos.

o *dataset* do projeto Petrolês obteve melhores resultados (85,61% de CLAS) que o do projeto UD (84,38% de CLAS). Embora o particionamento aleatório tenha obtido melhores números, isso não significa necessariamente que seja a melhor forma de particionar um *dataset*, pois é possível que (1) uma outra seleção de frases aleatória obtenha resultados piores, (2) uma outra seleção de documentos por partição obtenha resultados melhores, e (3) a avaliação do modelo considerando frases aleatórias não seja a mais correta, uma vez que somente selecionando documentos inteiros por partição haveria a garantia de que o modelo está sendo confrontado com exemplos de fato inéditos no seu estilo de escrita[10].

Por fim, podemos comparar os *datasets* simplificados em dois grupos: aqueles que têm o particionamento de frases aleatório (modelo do Petrolês) e aqueles que têm o particionamento por documento (modelo do UD). Entre os *datasets* do projeto Petrolês ("PetroGold-v3" e "PetroGold-v2"), vemos que a v3 obtém resultados melhores em todas as métricas em comparação à v2, chegando a até 2,13 p.p. (CLAS) de diferença. Esse é o impacto (positivo) que as revisões linguísticas realizadas nessa nova versão do *treebank* exerceram sobre a geração do modelo de aprendizado automático, considerando que o *tagset* (no caso, simplificado) é o mesmo. Já em relação aos *datasets* que seguem o particionamento do projeto UD ("PetroGold-ud-2.11" e "bosque-ud-2.11"), a diferença é de até 10,87 p.p. (CLAS). Não havendo como confiar em análises especulativas que considerem as diferenças relativas às características dos textos dos *corpora* – o PetroGold é composto por textos do gênero acadêmico, ao passo que o Bosque é composto por textos jornalísticos –, os resultados apenas sugerem que a diferença no desempenho se deve aos vários lotes de revisão pelos quais o PetroGold passou ao longo do tempo, possibilitando uma anotação com maior consistência interna e, portanto, mais facilmente generalizável.

5. Considerações finais

Este artigo explorou diferentes variações na anotação de um *corpus*, visando mostrar as possibilidades que um recurso desse tipo pode oferecer e o impacto dessas modificações no aprendizado de máquina. As variações incluíram alterações na anotação de POS para expressões multipalavras, simplificação do *tagset* de anotação sintática e diferentes estratégias de particionamento para os conjuntos de treinamento, teste e desenvolvimento. Os resultados mostraram que as variações "MWEPOS" e simplificada apresentaram os melhores desempenhos utilizando algumas das métricas, destacando a importância da modelagem linguística durante a anotação para obter resultados mais adequados no processamento de linguagem natural.

Todos os *datasets* testados neste trabalho estão disponíveis em um repositório dedicado[11]. A ideia é que, com os *datasets* (ou com as ideias que deram origem à produção desses *datasets*), futuramente seja possível expandir este trabalho com a avaliação de fato da qualidade dos modelos associados aos *datasets*, e não apenas da mudança de números produzida por materiais com variações de *tagset* e de anotação.

[10]O argumento é um dos apresentados pelo grupo de UD no endereço: *https://github.com/UniversalDependencies/UD_Portuguese-PetroGold/issues/3*. Acesso em 28 de maio de 2023.

[11]Disponível em: *https://github.com/alvelvis/petrogold-stil*. Acesso em 13 de ago. 2023.

Agradecimentos

Os autores agradecem ao CNPq (Conselho Nacional de Desenvolvimento Científico e Tecnológico, processo #130495/2021-2), à FAPERJ (Fundação Carlos Chagas Filho de Amparo à Pesquisa do Estado do Rio de Janeiro, processo #E-26/202.433/2022) e à ANP (Agência Nacional de Petróleo, Gás Natural e Biocombustíveis, Brasil, associada ao investimento de recursos oriundos das Cláusulas de P,D&I, por meio de Termo de Cooperação entre a Petrobras e a PUC-Rio) pelos auxílios concedidos, sem os quais este trabalho não poderia ter sido realizado. Cláudia Freitas atualmente está vinculada ao ICMC/USP e agradece o apoio do Centro de Inteligência Artificial da Universidade de São Paulo (C4AI - http://c4ai.inova.usp.br/), da Fundação de Amparo à Pesquisa do Estado de São Paulo (processo FAPESP #2019/07665-4) e da IBM. Este projeto também foi apoiado pelo Ministério da Ciência, Tecnologia e Inovações, com recursos da Lei N. 8.248, de 23 de outubro de 1991, no âmbito do PPI-Softex, coordenado pela Softex e publicado como Residência em TIC 13, DOU 01245.010222/2022-44.

Referências

Artstein, R. (2017). Inter-annotator agreement. In *Handbook of linguistic annotation*, pages 297–313. Springer.

Bick, E. (2014). PALAVRAS, a constraint grammar-based parsing system for Portuguese. *Working with Portuguese corpora*, pages 279–302.

de Marneffe, M.-C., Manning, C. D., Nivre, J., and Zeman, D. (2021). Universal dependencies. *Computational linguistics*, 47(2):255–308.

de Souza, E. (2023). *Construção e avaliação de um treebank padrão ouro*. Mestrado, PUC-Rio.

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*, pages 35–41, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

de Souza, E. and Freitas, C. (2022a). Polishing the gold–how much revision do we need in treebanks? In *Procedings of the Universal Dependencies Brazilian Festival*, pages 1–11.

de Souza, E. and Freitas, C. (2022b). Still on arguments and adjuncts: the status of the indirect object and the adverbial adjunct relations in Universal Dependencies for Portuguese. In *Proceedings of the Universal Dependencies Brazilian Festival*, pages 1–10, Fortaleza, Brazil. Association for Computational Linguistics.

de Souza, E. and Freitas, C. (2023a). Annotation of fixed multiword expressions (mwes) in a portuguese universal dependencies (ud) treebank: Gathering candidates from three different sources. In *Proceedings of the II Universal Dependencies Brazilian Festival (UDFest-BR)*.

de Souza, E. and Freitas, C. (2023b). Avaliação da anotação automática de dependências sintáticas. *Revista da ABRALIN*.

de Souza, E. and Freitas, C. (2023c). Um pronome com muitas funções: Descrição e resultados da anotação do pronome -se em um treebank segundo o esquema universal

dependencies (ud) para português. In *VIII Jornada de Descrição do Português, STIL 2023*.

de Souza, E., Silveira, A., Cavalcanti, T., Castro, M. C., and Freitas, C. (2021). PetroGold–Corpus padrão ouro para o domínio do petróleo. In *Anais do XIII Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana*, pages 29–38. SBC.

Freitas, C. (2023). Dataset e corpus. In Caseli, H. and Volpe Nunes, M. d. G., editors, *Processamento de Linguagem Natural: conceitos, técnicas e aplicações em Português*, pages –. BPLN.

Freitas, C. and de Souza, E. (2023). A study on methods for revising dependency treebanks: In search of gold. *Language Resources and Evaluation*, (no prelo).

Lewkowycz, A., Andreassen, A., Dohan, D., Dyer, E., Michalewski, H., Ramasesh, V., Slone, A., Anil, C., Schlag, I., Gutman-Solo, T., Wu, Y., Neyshabur, B., Gur-Ari, G., and Misra, V. (2022). Solving quantitative reasoning problems with language models.

Lopes, L., Duran, M. S., Fernandes, P., and Pardo, T. (2022). Portilexicon-ud: a portuguese lexical resource according to universal dependencies model. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 6635–6643.

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*.

Samuel, D., Kutuzov, A., Øvrelid, L., and Velldal, E. (2023). Trained on 100 million words and still in shape: BERT meets British National Corpus. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 1954–1974, Dubrovnik, Croatia. Association for Computational Linguistics.

Souza, F., Nogueira, R., and Lotufo, R. (2020). Bertimbau: Pretrained bert models for brazilian portuguese. In Cerri, R. and Prati, R. C., editors, *Intelligent Systems*, pages 403–417, Cham. Springer International Publishing.

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.

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.