UNIVERSIDADE DE SÃO PAULO
Instituto de Ciências Matemáticas e de Computação

Etiquetagem morfossintática multigênero para o português do Brasil segundo o modelo Universal Dependencies

Emanuel Huber da Silva

Dissertação de Mestrado do Programa de Pós-Graduação em Ciências de Computação e Matemática Computacional (PPG-CCMC)

ICMC USP
SÃO CARLOS

Emanuel Huber da Silva

Etiquetagem morfossintática multigênero para o português do Brasil segundo o modelo Universal Dependencies

Dissertação apresentada ao Instituto de Ciências Matemáticas e de Computação – ICMC-USP, como parte dos requisitos para obtenção do título de Mestre em Ciências – Ciências de Computação e Matemática Computacional. *EXEMPLAR DE DEFESA*

Área de Concentração: Ciências de Computação e Matemática Computacional

Orientador: Prof. Dr. Thiago Alexandre Salgueiro Pardo

USP – São Carlos
Abril de 2023

Ficha catalográfica elaborada pela Biblioteca Prof. Achille Bassi
e Seção Técnica de Informática, ICMC/USP,
com os dados inseridos pelo(a) autor(a)

H877e

Huber da Silva, Emanuel
Etiquetagem morfossintática multigênero para o português do Brasil segundo o modelo Universal Dependencies / Emanuel Huber da Silva; orientador Thiago Alexandre Salgueiro Pardo. -- São Carlos, 2023.
136 p.

Dissertação (Mestrado - Programa de Pós-Graduação em Ciências de Computação e Matemática Computacional) -- Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, 2023.

1. Etiquetagem morfossintática. 2. Universal Dependencies. 3. Etiquetagem multigênero. I. Alexandre Salgueiro Pardo, Thiago, orient. II. Título.

Bibliotecários responsáveis pela estrutura de catalogação da publicação de acordo com a AACR2:
Gláucia Maria Saia Cristianini - CRB - 8/4938
Juliana de Souza Moraes - CRB - 8/6176

Emanuel Huber da Silva

Multigenre part-of-speech tagging for Brazilian Portuguese according to the Universal Dependencies model

Master dissertation submitted to the Instituto de Ciências Matemáticas e de Computação – ICMC-USP, in partial fulfillment of the requirements for the degree of the Master Program in Computer Science and Computational Mathematics. *EXAMINATION BOARD PRESENTATION COPY*

Concentration Area: Computer Science and Computational Mathematics

Advisor: Prof. Dr. Thiago Alexandre Salgueiro Pardo

USP – São Carlos
April 2023

AGRADECIMENTOS

Agradeço a Deus por me presentear mais uma vez com algo que sei ser fruto de Sua infinita Graça, apesar de eu não merecê-la.

Ao Prof. Dr. Thiago Alexandre Salgueiro Pardo, pela orientação, paciência, amizade, pelas boas conversas e por me ensinar o que é ser um pesquisador, ressaltando a importância das contribuições de nossos artigos, teses e dissertações para a sociedade. Contudo, sempre lembrando da responsabilidade que carregamos com esse título.

À minha esposa, Rebecca, que, no momento em que escrevo esta sentença, está mais uma vez me apoiando, incentivando, sendo compreensiva e promovendo um lar de paz. Sem você, nada disso seria possível.

Aos meus pais, Paulo e Ligia, que foram os maiores exemplos de determinação e perseverança que tive na vida. Almejo um dia ser ao menos uma fração do que vocês são.

Às minhas irmãs, Jéssica e Gabrielle, que sempre me apoiaram, suportaram e amaram. Não posso deixar de mencionar meu cunhado, Cesar e meu sobrinho Henrique, que já me proporcionaram momentos muito felizes.

Aos meus amigos, que me apoiaram, aconselharam e se interessaram pelo meu trabalho. Vocês foram uma peça fundamental na minha vida.

Aos professores que me mostraram caminhos que jamais imaginei serem possíveis de trilhar. Que não apenas me mostraram, mas seguraram em minha mão e me acompanharam. Quando um obstáculo aparecia, vocês me mostraram que eu era capaz de sobrepô-lo. Sem vocês, meus amigos, nada disso seria possível. Em especial, agradeço ao professores Dr. Johannes Von Lochter, Me. André Breda Carneiro e Me. Fábio Lopes Caversan.

Aos professores Dr. Norton Trevisan Roman e Dra. Ariani Di Felippo pelas oportunidades de colaboração em artigos, auxílio nos conhecimentos de estatística e pelas boas conversas que tivemos.

À professora Dra. Magali Sanches Duran pelas contribuições das análises de erros deste trabalho. Além disso, agradeço à Dra. Lucelene Lopes e demais colegas do grupo de pesquisa POeTiSA. Com vocês pude aprender mais sobre linguística computacional e fazer parte de um projeto de grande impacto.

Ao Centro de Inteligência Artificial (C4AI-USP) e o apoio da Fundação de Apoio à Pesquisa do Estado de São Paulo (processo FAPESP #2019/07665-4) e da IBM Corporation.

Este projeto também foi apoiado pelo Ministério da Ciência, Tecnologia e Inovações, com recursos da Lei no 8.248, de 23 de outubro de 1991, no âmbito do PPI-SOFTEX, coordenado pela Softex e publicado Residência em TIC 13, DOU 01245.010222/2022-44.

Ao Instituto de Ciências Computacionais e Matemáticas (ICMC), pela oportunidade e flexibilização durante a pandemia que tornou meu ingresso possível.

À Universidade de São Paulo, que disponibiliza edução de altíssima qualidade para todos.

Ao Programa Pós-Graduação em Ciências de Computação e Matemática Computacional (PPG-CCMC), pela realização do curso e pelo apoio.

Ao centro de inovação CESAR e ao centro universitário Facens, que me disponibilizaram tempo para poder dedicar ao curso de mestrado e, consequentemente, viabilizaram a minha permanência e conclusão do curso.

"(...) Talvez eu esteja começando a conseguir, pois, de repente, parece-me que a destruição do que não deveria haver, isto é, a destruição do que vocês chamam de mal, é menos justa e desejável do que a conversão desse mal naquilo que vocês chamam de bem."

(Isaac Asimov, As Cavernas de Aço)

"Porque o juízo será sem misericórdia sobre aquele que não fez misericórdia; e a misericórdia triunfa do juízo."

(Tiago 2:13, Bíblia Sagrada)

RESUMO

SILVA, E. H. **Etiquetagem morfossintática multigênero para o português do Brasil segundo o modelo Universal Dependencies**. 2023. 136 p. Dissertação (Mestrado em Ciências – Ciências de Computação e Matemática Computacional) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos – SP, 2023.

A etiquetagem morfossintática é um dos primeiros níveis de estruturação linguística. Encontrando-se entre a morfologia e a sintaxe, busca-se identificar as classes gramaticais de cada palavra ou *token*. A tarefa é necessária para desambiguação morfossintática e, consequentemente, para a criação de ferramentas e métodos de Processamento de Língua Natural mais robustos. Nessa linha, existe uma variedade de trabalhos para o português do Brasil utilizando córpus de gênero jornalístico com diferentes conjuntos de etiquetas. O formalismo *Universal Dependencies* (UD) é a teoria linguística que tem sido mais adotada por córpus na área, o que permite a padronização entre diferentes línguas e gêneros textuais, inclusive, do conjunto de etiquetas morfossintáticas. Apesar de existirem trabalhos de etiquetadores para o português do Brasil baseados em diversos formalismos, existem poucos trabalhos que se baseiam na UD. Além disso, há escassez de trabalhos que abordam córpus com variedade de gêneros textuais. Cada gênero textual possui diferentes características linguísticas e, consequentemente, apresenta desafios para os métodos de etiquetagem. Nesse projeto, foi realizada a investigação de métodos de etiquetagem morfossintática para o português do Brasil adotando o formalismo UD. Destaca-se a análise no contexto multigênero com textos jornalísticos, acadêmicos e Conteúdo Gerado por Usuário (CGU).

Palavras-chave: Etiquetagem morfossintática, Universal Dependencies, Etiquetagem multigênero.

ABSTRACT

SILVA, E. H. **Multigenre part-of-speech tagging for Brazilian Portuguese according to the Universal Dependencies model**. 2023. 136 p. Dissertação (Mestrado em Ciências – Ciências de Computação e Matemática Computacional) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos – SP, 2023.

Part-of-Speech tagging is one of the first levels of linguistic structuring. Lying between morphology and syntax, and seeks to identify the grammatical classes of each word or token. The task is necessary for morphosyntactic disambiguation and, consequently, for the creation of more robust Natural Language Processing tools and methods. In this line, there is a variety of work for Brazilian Portuguese using journalistic genre corpus with different sets of tags. The Universal Dependencies (UD) formalism is the linguistic theory that has been most adopted by corpora in the area, which allows standardization across different languages and textual genres, including the set of morphosyntactic tags. Although there are works on taggers for Brazilian Portuguese based on several formalisms, there are few works based on UD. Furthermore, there is a dearth of works that address corpus with a variety of textual genres. Each text genre has different linguistic characteristics and, consequently, presents challenges for tagging methods. In this project, we investigated morphosyntactic tagging methods for Brazilian Portuguese adopting the UD formalism. Notably, the analysis in the multigenre context with journalistic, academic and User-Generated Content (UGC) texts.

Keywords: Part-of-Speech tagging, Universal Dependencies, multigenre tagging.

LISTA DE ILUSTRAÇÕES

LISTA DE QUADROS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

AM	Aprendizado de Máquina
AT	*Adversarial Training*
BERT	*Bidirectional Encoder Representations from Transformers*
Bi-LSTM	*Bidirectional Long Short-Term Memory*
CGU	Conteúdo Gerado por Usuário
CRF	*Conditional Random Fields*
DANTE	*Dependency-ANalised corpora of TwEets*
FN	Falso Negativo
FP	Falso Positivo
GRU	*Gated Recurrent Unit*
HMM	*Hidden Markov Model*
LAN	*Label Attention Network*
LSTM	Long-Short-Term Memory
NGB	Nomenclatura Gramatical Brasileira
OOV	*Out-of-Vocabulary*
PLN	Processamento de Língua Natural
RNR	Rede Neural Recorrente
UD	*Universal Dependencies*
VN	Verdadeiro Negativo
VP	Verdadeiro Positivo

SUMÁRIO

CAPÍTULO

1

INTRODUÇÃO

1.1 Contextualização e Motivação

A área de Processamento de Língua Natural (PLN) busca realizar a compreensão e geração de língua natural através de ferramentas computacionais (JURAFSKY; MARTIN, 2009). Algumas das áreas de aplicação de PLN são: análise de sentimentos, tradução automática e sumarização de textos, entre outras. Para a construção destas aplicações, é possível utilizar ferramentas que permitem extrair características linguísticas, como a estrutura sintática, atributos morfológicos e, em particular, as etiquetas morfossintáticas. Essas informações podem auxiliar no processamento de textos para solucionar a tarefa alvo, por exemplo, um corretor gramatical pode utilizar a estrutura sintática para encontrar erros.

As ferramentas e aplicações de PLN são utilizadas para processar todo tipo de gênero textual, como científico e literário, entre outros. Historicamente, a área iniciou os estudos em córpus que seguiam a norma culta da língua, como os gêneros jornalístico e acadêmico. Atualmente, o gênero de Conteúdo Gerado por Usuário (CGU) ou *User-Generated Content* vem sendo amplamente utilizado, sendo encontrado em fóruns da *web*, redes sociais e blogs, entre outros. Dada a abundante quantidade de dados disponíveis em CGU, este gênero textual se popularizou como objeto de estudo na área de PLN. Os textos CGU são caracterizados por falta de adesão à norma culta da língua, uso de artifícios textuais pertencentes à rede social (como *"hashtags"*), uso frequente de abreviações, criação de palavras e termos próprios, entre outros. O Quadro 1 apresenta uma sentença extraída da rede social Twitter[1]. É possível observar alguns fenômenos linguísticos que podem dificultar a compreensão automática por ferramentas de PLN, sendo eles: abreviações, uso de língua estrangeira e elementos metalinguísticos (*hashtag*, menção), entre outros.

Os métodos de PLN clássicos construídos visando outros gêneros textuais não possuem

[1] Disponível em: <https://twitter.com/>.

Quadro 1 – Exemplo de tweet em português

#VALE5 é #VENDA? rsss #DEAL! #DEAL! #DEAL! ’16 de março às 12:12’ após vencto das opções podem puxar na... http://t.co/4mOMj1Om7d

Fonte – (Di FELIPPO *et al.*, 2021).

bons resultados quando processam CGU, justamente pelo fato das características textuais deste gênero se diferenciarem de outros, já que normalmente os métodos clássicos trabalhavam com textos jornalísticos, literários, enciclopédias e artigos científicos, entre outros. Não obstante, a maior quantidade de conteúdo escrito na humanidade está disponível na internet, em grande parte na forma de CGU. Sendo assim, o processamento automático de CGU é necessário para criação de estudos e aplicações que auxiliem a processar e compreender esse gênero textual.

O PLN possui diversas aplicações no gênero CGU, por exemplo, a análise de opinião automática, onde busca-se extrair sentimentos associados a uma avaliação de um produto ou serviço. As pessoas frequentemente emitem opiniões espontâneas acerca de um produto ou serviço que utilizaram nas redes sociais, dessa forma, com ferramentas de PLN é possível identificar automaticamente a opinião do usuário em relação ao produto ou serviço. Naturalmente, esta informação possibilita que a empresa responsável conheça automaticamente as opiniões de seus consumidores e melhore a experiência do usuário. MARRESE-TAYLOR *et al.* (2020) realizam a análise de opinião em transcrições de vídeos que apresentam avaliações de produtos, permitindo extrair automaticamente os aspectos positivos, negativos e neutros apresentados pelo produtor de conteúdo em relação ao produto utilizado.

Um dos gêneros textuais extensivamente estudado na literatura é o gênero jornalístico. Em contraposição à CGU, este gênero adere à norma culta da língua, possuindo baixa frequência de erros gramaticais, com linguagem clara e objetiva. A Quadro 2 apresenta uma sentença de córpus do gênero jornalístico, onde é possível visualizar as características linguísticas apresentadas.

Quadro 2 – Exemplo de texto do gênero jornalístico

Casos positivos de dengue em Monte Aprazível (38 km de Rio Preto-SP) foram divulgados ontem pelo Ersa (Escritório Regional de Saúde) de Rio Preto.

Fonte – Adaptada de (AFONSO *et al.*, 2002).

O primeiro grande córpus de gênero jornalístico foi criado no fim da década de 1970 (FRANCIS; KUCERA, 1979) e possui aproximadamente 1 milhão de *tokens*. Desde então, diversos córpus foram criados para o gênero jornalístico, como o CETENFolha[2] para o português do Brasil.

Com ferramentas de PLN que se baseiam em texto jornalístico, é possível criar aplicações que encontram aspectos relevantes em notícias e manchetes, entre outros. Por exemplo,

[2] Disponível em: <https://www.linguateca.pt/cetenfolha/>.

(SUNKARA, 2019) apresenta um sistema de PLN para encontrar as 5 questões principais que um artigo jornalístico visa responder, sendo elas: onde, quando, quem, o quê e por quê. A identificação automática desses aspectos permite o uso em sistemas de recomendação de notícias e sumarização automática, entre outros.

Outro gênero textual de grande interesse na literatura é o acadêmico. Este gênero é utilizado para apresentar resultados de pesquisas científicas, que devem ser rigorosos e objetivos, apoiados em evidências e argumentação consistente. Não obstante, este gênero é caracterizado pela adesão à norma culta da língua, dessa forma, não possui os mesmos desafios inerentes ao gênero CGU e se assemelha aos desafios encontrados no texto jornalístico. A Quadro 3 apresenta uma sentença de córpus acadêmico, onde é possível observar a aderência à norma culta da língua e a objetividade do autor.

Quadro 3 – Exemplo de texto do gênero acadêmico

O objetivo deste trabalho é caracterizar as principais estruturas do arcabouço geológico do embasamento aflorante da Bacia de Pelotas e sua correlação com a porção offshore da mesma.

Fonte – Adaptada de (SOUZA; FREITAS, 2022).

O gênero acadêmico foi um dos primeiros a ter córpus digitalmente disponível, sendo o córpus Brown (FRANCIS; KUCERA, 1979) um dos primeiros da categoria com mais de um milhão de *tokens*. Semelhantemente, para o português do Brasil, o gênero também foi um dos primeiros a ser abordado na literatura. O córpus Selva Científica é um subconjunto do córpus Selva (LINGUATECA, 2009) que possui mais de 140 mil tokens. Os documentos utilizados foram capítulos de teses, artigos da Wikipedia[3] e relatórios técnicos.

Os textos acadêmicos normalmente possuem uma estrutura de argumentação clara e concisa, dessa forma, o autor visa realizar uma reivindicação a partir de uma linha argumentativa que possui premissas nas quais os autores baseiam suas ideias. A identificação automática das reivindicações e premissas de linhas argumentativas em textos acadêmicos é um tópico de interesse no PLN, pois permite filtrar, comparar e recomendar artigos científicos com base em suas reivindicações e premissas. ACCUOSTO; NEVES; SAGGION (2021) abordam o problema em textos acadêmicos de linguística computacional e biomedicina, onde, a partir da abordagem proposta pelos autores, é possível identificar as reivindicações e premissas automaticamente.

Para realizar o processamento de CGU e de outros gêneros linguísticos, pode ser necessário utilizar técnicas de PLN para extrair as estruturas linguísticas básicas do texto, como as relações sintáticas e classes gramaticais por meio do uso de analisadores sintáticos e etiquetadores morfossintáticos. A etiquetagem morfossintática ou *Part-of-Speech tagging* busca encontrar as categorias gramaticais de cada palavra em uma sentença. Este processo é um dos primeiros níveis de estruturação de textos para o uso em aplicações finais, encontrando-se entre a

[3] Disponível em: <https://wikipedia.org/>.

morfologia e a sintaxe, sendo assim, a etiquetagem considera a forma e o contexto das palavras e símbolos para associação das classes gramaticais. Um exemplo de sentença com suas respectivas etiquetas morfossintáticas pode ser observado na Figura 1.

Figura 1 – Exemplo de sentença rotulada com classes gramaticais

Fonte: Elaborada pelo autor.

O uso das classes gramaticais funciona como base para várias aplicações de PLN, pois conhecer a classe gramatical das palavras de um texto enriquece a representação computacional, dessa forma, podendo melhorar o desempenho nas tarefas finais. Por exemplo, para analisar a declaração *A noiva casa de branco* é necessário saber que *casa* se refere ao verbo *casar* e não ao substantivo *casa*, e, para isso, o contexto deve ser considerado. Caso o etiquetador morfossintático não considere o contexto, pode-se erroneamente identificar a palavra *casa* como um substantivo, que, neste contexto, não faria sentido. Dessa forma, um sistema que ignore o contexto produzirá etiquetas erradas e, consequentemente, impactará negativamente o desempenho da análise final.

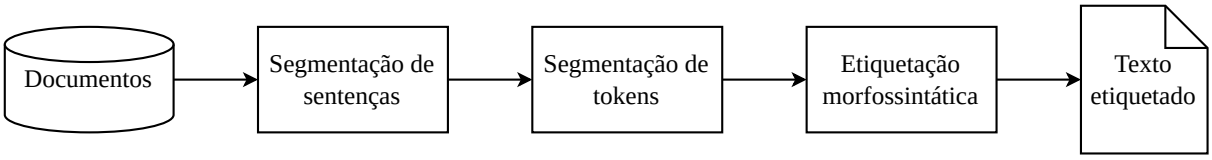
Atualmente, trabalhos na literatura apresentam o uso de etiquetadores morfossintáticos em combinação de técnicas de representação vetorial de textos para melhorar o desempenho na tarefa alvo. LIN *et al.* (2021) apresentam um analisador de opiniões baseado em aspectos que utiliza as classes gramaticais em combinação com as representações vetoriais das palavras. Dessa forma, os autores relatam um aumento consistente no desempenho do modelo ao utilizar as classes gramaticais. Semelhantemente, outros trabalham utilizam a mesma técnica em outras aplicações e apresentam aumento no desempenho, como na tarefa de sumarização automática (ZHAO *et al.*, 2019) e tradução (YANG *et al.*, 2019), entre outros. Não obstante, a construção de etiquetadores morfossintáticos permite que pesquisadores os utilizem para processar textos e realizar estudos dirigidos, por exemplo, GARIMELLA *et al.* (2019) apresentam um estudo entre diferenças gramaticais entre textos escritos por homens e mulheres.

Um sistema de etiquetagem morfossintática realiza uma sequência de passos para processar textos crus até produzir as classes gramaticais de cada *token*[4]. Primeiramente, os documentos devem ser extraídos de sua fonte original. Em seguida, cada documento deve ser dividido em sentenças. Posteriormente, cada sentença deve ser segmentada conforme seus *tokens*, produzindo uma lista de *tokens*. Finalmente, o etiquetador morfossintático analisará a lista de *tokens* e produzirá uma lista de etiquetas gramaticais de mesmo tamanho. O processo descrito é ilustrado na Figura 2.

O etiquetador morfossintático irá produzir etiquetas pertencentes a um sistema pré-estabelecido que dita quais são as classes gramaticais disponíveis. Por muitos anos, diversos

[4] É uma unidade computacional que representa uma palavra, pontuação ou símbolo especial, entre outros.

Figura 2 – Arquitetura para sistemas de etiquetagem morfossintática



Fonte: Elaborada pelo autor.

trabalhos utilizaram ou criaram conjuntos de etiquetas, como (FRANCIS; KUCERA, 1979) para o inglês e (AFONSO *et al.*, 2002) para o português do Brasil, ambos altamente acoplados à língua de origem. A existência de múltiplos conjuntos de etiquetas facilitava a criação de adaptações necessárias para cada língua, contudo, dificultava a adaptação de métodos entre diferentes línguas. Visando esta dificuldade, o projeto *Universal Dependencies* (UD) (NIVRE *et al.*, 2016) apresenta um padrão universal de etiquetas morfossintáticas, além de outras padronizações, como a segmentação de palavras, etiquetas morfológicas e sintáticas. Atualmente, a UD possui *treebanks*[5] para mais de 100 línguas. O conjunto de etiquetas morfossintáticas da UD é formado por 17 rótulos universais que atende todas as línguas contidas no projeto; não obstante, em casos excepcionais, é possível adicionar novas etiquetas a um conjunto separado dos rótulos universais. Apesar de o projeto UD ser recente, diversos córpus anteriores foram convertidos para o formalismo de dependências e adicionados ao projeto, como o *treebank* Bosque (RADEMAKER *et al.*, 2017), que contém textos jornalísticos em português do Brasil.

A busca por um formalismo universal da UD cria algumas dificuldades na rotulação de textos para o português do Brasil. No português, possuímos a classe gramatical *artigo*, que geralmente antecede um substantivo, porém, na UD, esta classe é substituída pela categoria *determinante*. Exemplificando, na frase *Aquelas pessoas estão olhando para você*, *Aquelas* receberia a classe gramatical *pronome*, segundo a gramática brasileira, porém, segundo a UD, a etiqueta correta seria *determinante*. Ao mesmo tempo, em outros contextos, *Aquelas* pode receber a classe gramatical pronome. Sendo assim, é possível concluir que a classe *determinante* inclui mais palavras do que a classe *artigo* e, consequentemente, tornando a tarefa de etiquetagem morfossintática dessa classe mais complexa ao adotar o formalismo UD.

Os sistemas de etiquetagem morfossintática que obtiveram os melhores resultados são baseados em Aprendizado de Máquina, ou seja, utilizam córpus para o aprendizado automático de um modelo que consiga associar as classes gramaticais às palavras de entrada, conforme os rótulos criados previamente por linguistas. Esses sistemas alcançam acurácias entre 97% e 98%, portanto, considera-se a tarefa como resolvida para córpus de gênero jornalístico e acadêmico, porém, o mesmo não ocorre para córpus de gênero CGU, onde a acurácia varia de 86,95% a 93,3%. Além disso, existem relativamente poucos córpus para o português do Brasil com etiquetas morfossintáticas e praticamente nenhum tratando do gênero CGU. É importante

[5] córpus com sentenças com suas estruturas sintáticas.

mencionar que, apesar das altas taxas de acerto dos etiquetadores atuais, deve-se considerar outros aspectos. JURAFSKY; MARTIN (2009) demonstram que cerca de 85% das palavras de um córpus não são ambíguas, ou seja, possuem apenas uma classe gramatical possível. Em contraposição, os 15% restantes têm ocorrência média de 55% nos textos, consequentemente, essa proporção impacta no desempenho final do algoritmo. Aumentos de desempenho substanciais levam a grande redução de erros por sentença, conforme apresentado por KEPLER (2010).

Na literatura, encontram-se diversos trabalhos de etiquetagem morfossintática para o português do Brasil, contudo, existem poucos trabalhos que se baseiam no formalismo UD. Dada a adoção da UD pela comunidade científica, ao criar etiquetadores que aderem ao formalismo, possibilita-se o uso em aplicações que já estão adaptadas a UD, como em analisadores sintáticos. Além disso, o uso de um padrão internacional amplamente aceito na comunidade permite realizar estudos comparativos entre diferentes línguas e córpus. Este trabalho realiza a investigação de etiquetadores morfossintáticos para o português do Brasil com base no formalismo UD, utilizando técnicas recentes da literatura. Além disso, uma das características que permite a versatilidade de etiquetadores é a capacidade de processar textos com gêneros textuais diferentes. Dessa forma, este trabalho realiza a avaliação da etiquetagem no contexto multigênero, considerando os gêneros jornalístico, acadêmico e CGU. Não obstante, é desconhecida a existência de etiquetador morfossintático para o português do Brasil baseado na UD com capacidade multigênero. Por fim, é produzido um etiquetador que proporcionará auxílio a métodos e ferramentas no uso de etiquetas morfossintáticas nas mais diversas aplicações de PLN para o português do Brasil.

1.2 Objetivos e hipóteses de pesquisa

Este projeto tem o ponto focal na investigação de métodos de etiquetagem morfossintática para o português do Brasil com base nas diretrizes do projeto *Universal Dependencies*. Esta avaliação considera os métodos do estado-da-arte para a etiquetagem e abordagens recentes da literatura. Além disso, é realizada a avaliação de etiquetagem no contexto multigênero, considerando os gêneros CGU, jornalístico e acadêmico.

Este trabalho baseia-se em duas hipóteses, sendo a primeira de que a etiquetagem morfossintática para o português do Brasil pode ser realizada com capacidades multigênero, dessa forma, mantendo o desempenho geral para cada gênero. Adicionalmente, hipotetizou-se que o uso do formalismo *Universal Dependencies* é suficiente para atingir resultados do estado da arte na tarefa de etiquetagem morfossintática.

1.3 Organização do texto

O Capítulo 2 apresenta os principais conceitos teóricos relacionados a este projeto de pesquisa, como as características de córpus, conceitos de aprendizado de máquina e modelos de

língua, entre outros. O Capítulo 3 expõe os principais trabalhos relacionados para etiquetagem morfossintática para gênero jornalístico, CGU e multigênero. Em seguida, o Capítulo 4 apresenta os córpus e métodos utilizados nas análises. O Capítulo 5 apresenta em detalhes os experimentos realizados e discute as análises quantitativas e qualitativas. Por fim, o Capítulo 6 apresenta as conclusões do trabalho, ressaltando as principais contribuições e indica direções futuras de pesquisa.

CAPÍTULO

2

FUNDAMENTAÇÃO TEÓRICA

Este capítulo apresenta os conceitos fundamentais relacionados ao desenvolvimento de analisadores morfossintáticos, buscando abordar a definição formal da tarefa, apresentar os métodos de base utilizados na literatura, conjuntos de etiquetas e métricas para avaliação dos algoritmos.

2.1 Análise morfossintática

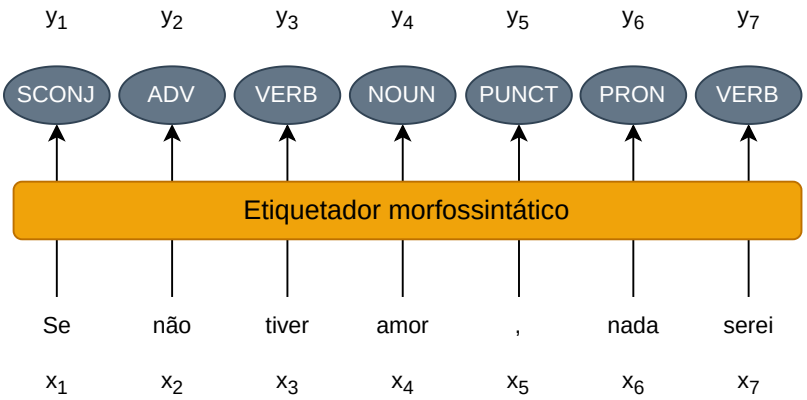
O analisador morfossintático é uma ferramenta computacional de PLN capaz de processar uma sequência de *tokens* e produzir uma sequência contendo as respectivas classes gramaticais. Um *token* é uma unidade que representa uma única palavra, símbolo, número ou pontuação. Além disso, fenômenos linguísticos como as contrações devem ser expandidas previamente, como *do*, que, após a expansão, se torna *de o*. O processo de segmentar uma sentença em uma sequência de *tokens* é denominado *tokenização*, sendo fundamental para a etiquetagem morfossintática, A Figura 2 apresenta a sequência de passos necessária para a realização da etiquetagem morfossintática.

Tendo em vista que a tarefa permeia os níveis linguísticos da morfologia e sintaxe, o analisador deve considerar a forma e o contexto para realizar a associação mais correta possível dos rótulos gramaticais. Uma palavra pode possuir mais de uma classe gramatical e as palavras que a acompanham podem ajudar a solucionar a desambiguação. Conforme (JURAFSKY; MARTIN, 2009), muitas tarefas de PLN são vistas como processos de desambiguação, ou seja, entre todas as possíveis soluções, o algoritmo deve produzir a saída que possui a maior probabilidade de ocorrência. Por exemplo, na sentença *O banco é de madeira*, a palavra *banco* possui baixa probabilidade de ser o verbo *bancar* e esta análise só é possível a partir do contexto completo da sentença.

Na literatura, encontram-se diversas abordagens para construção de etiquetadores, como

algoritmos apoiados por regras pré-definidas por linguistas (KLEIN; SIMMONS, 1963), métodos probabilísticos (KEPLER, 2010), conexionistas (BOHNET *et al.*, 2018), entre outros. Recentemente, os métodos que possuem maior desempenho derivam da abordagem conexionista, que faz o uso de redes neurais artificiais, uma subárea do Aprendizado de Máquina (AM). Na Figura 3, visualiza-se uma arquitetura genérica de etiquetagem morfossintática, indicando as entradas e saídas do sistema. O etiquetador morfossintático recebe como entrada uma lista de *tokens* representada pelo vetor x^n e produz um vetor de saída y^n. Como será realizada a associação entre *tokens* e classes gramaticais, dependerá da abordagem e algoritmo utilizado.

Figura 3 – Etiquetador morfossintático mapeando os *tokens* de entradas $x_{1...n}$ para classes gramaticais $y_{1...n}$



Fonte: Adaptada de (JURAFSKY; MARTIN, 2009).

2.2 Conjunto de etiquetas morfossintáticas

As etiquetas morfossintáticas para o português do Brasil são compostas por 10 classes segundo a Nomenclatura Gramatical Brasileira (NGB) (NOMENCLATURA..., 1959), sendo elas: substantivos, artigos, adjetivos, numerais, pronomes, verbos, advérbios, preposições, conjunções e interjeições. Não obstante, este trabalho baseia-se nas diretrizes do projeto *Universal Dependencies* (UD) (NIVRE *et al.*, 2016), que busca padronizar o conjunto de etiquetas morfossintáticas, atributos morfológicos e relações sintáticas, com base no formalismo de dependências sintáticas. Atualmente, as diretrizes da UD foram amplamente adotadas pela área. Existem cerca de 200 *treebanks* para mais de 100 línguas. Além disso, trabalhos anteriores que utilizavam outros formalismos linguísticos foram convertidos para a UD (RADEMAKER *et al.*, 2017; SILVEIRA *et al.*, 2014). O conjunto de etiquetas definido pela UD é utilizado por todos os *treebanks*, facilitando a adaptação de métodos para o uso multilíngue. Na Figura 4 é possível visualizar as 17 etiquetas morfossintáticas e suas definições.

Observa-se que o conjunto de classes gramaticais das NGB é próximo do conjunto definido pela UD, com exceção das classes gramaticais *artigo*, *preposição* e diferenças pontuais,

Figura 4 – Conjunto das 17 etiquetas morfossintáticas do *Universal Dependencies*

Etiqueta	Descrição	Exemplo
ADJ	Adjetivo	bonito, esplêndido, sábio
ADV	Advérbio	muito, ontem, demais
NOUN	Substantivo	cachorro, livro, violão
VERB	Verbo	programar, andar, levantar
PROPN	Nome próprio	Bayes, São Paulo, USP
INTJ	Interjeição	olá, sim, ufa!
ADP	Adposição	dentro de, ao, graças a
AUX	Auxiliar	devo, posso, poderia
CCONJ	Conjunção Coordenativa	e, ou, mas
DET	Determinante	a, isso, o
NUM	Numeral	um, primeiro, dois
PART	Partícula	que, fora, por
PRON	Pronome	eu, ela, quem
SCONJ	Conjunção Subordinativa	segundo, conforme, desde que
PUNCT	Pontuação	, . (
SYM	Símbolos	#, $, @
X	Outro	nbjf, sadsd, .d.a2

Fonte: Elaborada pelo autor.

como ocorre no *pronome*. Na primeira diferença, o *artigo* é uma subclasse da classe gramatical *determinante*, que é uma etiqueta utilizada para palavras que modificam substantivos, portanto, todo *artigo* é um *determinante*. Não obstante, isto dificulta a tarefa, pois a classe *determinante* inclui mais palavras do que a classe *artigo*. No caso da *preposição*, a UD possui a superclasse *ADP* (adposições) que engloba preposições e posposições. As posposições não existem na língua portuguesa, porém, como é um recurso utilizado por outras línguas, a UD engloba ambos os casos na etiqueta *ADP*. Algumas diferenças entre o formalismo UD e NBG foram apresentadas, porém, não é objetivo deste trabalho discutir as nuances entre eles, portanto, recomenda-se consultar (DURAN, 2021) para informações mais detalhadas.

2.3 Aprendizado de Máquina

Na linguística computacional, existem dois principais paradigmas para abordar os problemas de PLN. Em primeiro lugar, temos a escola Chomskiana que se baseia na ideia de que todas as línguas compartilham um padrão universal de princípios, fundamentando-se na premissa de que todas as línguas foram criadas da mesma forma. Em contraposição, temos a escola empirista que busca compreender os padrões da língua através de dados e experimentos. Atualmente, o principal paradigma utilizado é o empirista, onde normalmente são utilizados métodos de Aprendizado de Máquina (AM) para criação de modelos computacionais capazes de modelar a distribuição dos dados e realizar predições (MANNING; SCHÜTZE, 1999).

O AM é uma subárea da Inteligência Artificial, onde se almeja desenvolver métodos capazes de aprender com base nos dados. Segundo a definição de (MITCHELL, 1997), um programa aprende a partir de um conjunto de experiências *E*, em relação a uma classe de tarefas *T*, com medida de desempenho *P*, se seu desempenho em *T*, medido por *P*, melhora com *E*. No contexto deste trabalho, podemos definir cada um dos elementos da seguinte forma:

- **Tarefa:** Etiquetagem morfossintática
- **Experiência:** Sentenças anotadas automaticamente com etiquetas morfossintáticas
- **Medida de desempenho:** Acurácia, Precisão, Sensibilidade ou Medida-F.

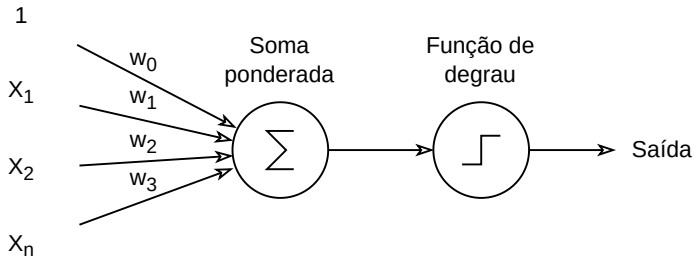
Existem diferentes paradigmas de AM, sendo os principais: aprendizado supervisionado, não supervisionado, semisupervisionado, ativo e por reforço (FACELI *et al.*, 2021). Neste trabalho, foi utilizado o aprendizado supervisionado, que utiliza rótulos previamente conhecidos para treinamento de um modelo. No contexto de análise morfossintática, os rótulos são as classes gramaticais de cada *token*. A partir disso, o etiquetador utiliza os *tokens* e rótulos contidos no córpus na etapa de treinamento para melhorar o desempenho na tarefa.

Com o aprendizado supervisionado, é possível criar métodos probabilísticos (CREŢULESCU *et al.*, 2014) que buscam encontrar as classes gramaticais por meio da estimação de probabilidades para cada *token* de entrada. Além dessa abordagem, existe o aprendizado automático de regras (DOMINGUES, 2011) por meio do algoritmo de BRILL (1995). Atualmente, as abordagens de etiquetagem morfossintática que obtêm resultados do estado-da-arte utilizam as redes neurais artificiais como base; portanto, este trabalho baseia-se na abordagem conexionista.

A partir da iniciativa de buscar compreender o funcionamento do cérebro e tentar modelá-lo matematicamente, surgiram as redes neurais artificiais, uma das subcategorias de métodos no AM. O primeiro modelo de neurônio artificial, apresentado por (MCCULLOCH; PITTS, 1943) e, posteriormente, construído por (ROSENBLATT, 1958), recebeu o nome de *perceptron*. O algoritmo recebe um conjunto de sinais de entrada e os processa, utilizando seus pesos, que são valores numéricos representando os sinais inibitórios ou excitatórios de um neurônio. Por fim, os sinais ponderados são somados, passam por uma função de ativação e são enviados para a saída do *perceptron*. É possível visualizar uma ilustração do *perceptron* na Figura 5.

A partir do *perceptron*, adaptações e extensões foram realizadas, como a adição de múltiplas camadas. O trabalho de (BRYSON; HO, 1969) apresenta como realizar o ajuste dos pesos de forma automática, possibilitando o aprendizado. As redes neurais são aproximadores universais de funções, ou seja, desde que haja uma função de perda diferenciável, é possível utilizar o método para qualquer tarefa. O *perceptron* e suas variações foram utilizados para predições de preços de casas (XU; ZHANG, 2021), análise de imagens (LECUN *et al.*, 1999) e aplicações de PLN, como a tradução automática (BAHDANAU; CHO; BENGIO, 2016;

Figura 5 – O *perceptron* de (ROSENBLATT, 1958)



Fonte: Elaborada pelo autor.

VASWANI *et al.*, 2017), classificação de textos e sentenças (ZHANG; ZHAO; LECUN, 2015; KIM, 2014), entre outros.

Recentemente, as unidades de processamento gráfico se tornaram cada vez mais populares e modernas, sendo capazes de realizar operações matriciais de forma mais eficiente do que as unidades de processamento para uso geral (KAYID; KHALED; ELMAHDY, 2018). Dado que os cálculos realizados em uma rede neural são operações matriciais, este cenário permitiu o uso de redes neurais artificiais com maior número de camadas, como o trabalho de (HE *et al.*, 2015) que atingiu desempenho acima do nível humano na tarefa de classificação de imagens. O uso de redes neurais artificiais com múltiplas camadas popularizou-se pelo nome de aprendizado profundo no início de 2006 (GOODFELLOW; BENGIO; COURVILLE, 2016). O aumento do número de camadas e o uso de novas arquiteturas permitiram grandes avanços na área de PLN, como a criação de modelos com conexões recorrentes Rede Neural Recorrente (RNR), Long-Short-Term Memory (LSTM), *Gated Recurrent Unit* (GRU) e, recentemente, a arquitetura *Transformer*.

Diversos trabalhos na área de PLN foram desenvolvidos a partir do uso do aprendizado profundo e da criação de grandes córpus. Uma das abordagens populares foi a criação de modelos de língua, isto é, modelos computacionais que associam probabilidades a uma sequência de palavras (JURAFSKY; MARTIN, 2009). Os modelos de língua, baseados em aprendizado profundo, utilizam representações distribucionais[6] de *tokens*, ou seja, cada *token* é representado por um vetor em um espaço dimensional de tamanho fixo.

O trabalho de (MIKOLOV *et al.*, 2013a) apresenta uma forma de utilizar redes neurais para representação de palavras em um hiperplano de alta dimensionalidade, buscando fazer com que palavras que possuam significados semelhantes fiquem próximas entre si. Embasado na premissa de que o significado de uma palavra é diretamente relacionado às palavras vizinhas, os autores construíram uma rede neural que busca fazer a predição dos vizinhos (palavras de contexto) de uma palavra (*Skip-gram*), ou de uma palavra dado os seus vizinhos (*Continuous Bag-of-Words*). Este método proporcionou uma representação semanticamente rica, melhorando

[6] Também conhecido por: representação vetorial ou *embedding*.

os resultados em diversas tarefas de PLN (LE; MIKOLOV, 2014; LIU, 2017).

Apesar dos avanços que o trabalho de (MIKOLOV *et al.*, 2013a) trouxe com as representações de palavras no hiperplano, estas não possuem variações conforme o contexto, isto é, a representação vetorial de uma palavra depende unicamente dela mesma. Por mais que o contexto tenha sido utilizado durante a etapa de treinamento, no momento de predição apenas a palavra alvo é utilizada, consequentemente, isso causará ambiguidade na representação vetorial. Ao analisarmos as sentenças presentes na Quadro 4, podemos observar que a palavra *preciso* possui diferentes significados (dependendo do contexto), porém, com o uso da representação vetorial de (MIKOLOV *et al.*, 2013a) em ambas as sentenças, a palavra *preciso* possuirá a mesma representação vetorial, prejudicando as tarefas finais de PLN que precisem realizar alguma desambiguação.

Quadro 4 – Exemplo de sentença com ambiguidade morfossintática

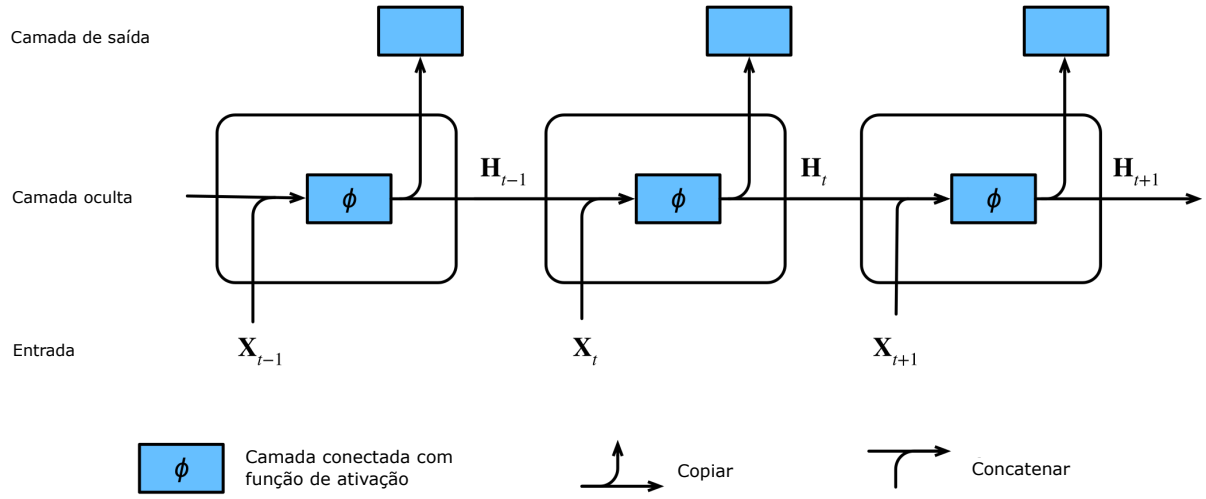
Sentença 1: Navegar é **preciso**
Sentença 2: O cirurgião é **preciso**

Fonte – Elaborada pelo autor.

Tendo em vista a necessidade de representações dependentes de contexto, diversos trabalhos utilizaram o paradigma de RNRs (ELMAN, 1990) para incluir o aspecto contextual. As RNRs trabalham de forma sequencial, isto é, processam uma entrada a cada passo, pois, existe a dependência entre estados anteriores. Um diagrama de uma RNR pode ser visualizado na Figura 6. Inicialmente, o primeiro item da lista de entrada é inserido no modelo ($X_{t=0}$). Em seguida, o cálculo da unidade RNR é realizado considerando o estado oculto ϕ, sendo este dependente da arquitetura selecionada. O resultado é enviado para o processamento da próxima entrada, onde o cálculo considera a entrada atual X_t e o estado oculto anterior H_{t-1}. O processo é repetido até que todos os itens da entrada tenham sido processados pelo modelo.

O uso desta abordagem permite considerar ou desconsiderar passos anteriores e espera-se que, durante o treinamento, a rede aprenda a ignorar os passos irrelevantes e dar maior peso aos passos importantes. (KEPLER, 2010) demonstrou que, no português do Brasil, a palavra *que* é, geralmente, um *PRON* ou *SCONJ*, e o que contribui na distinção é o contexto à esquerda da palavra *que*. Por exemplo, na sentença *Isso é mais que correto*, a palavra *mais* indica a ocorrência da etiqueta *SCONJ* para a palavra *que*. O uso da abordagem de RNR permite que o modelo possa dar mais importância à palavra *mais* do que às outras palavras quando for realizar a predição da palavra *que*. Não obstante, o contexto à direita da palavra alvo também pode ser relevante. Examinando a palavra *a*, é possível resolver a ambiguidade analisando o contexto à direita da palavra: se *a* for seguida de um substantivo, será um *DET*; porém, se for seguida por um *VERB*, provavelmente será um *ADP*. Tendo em vista a necessidade de representar os contextos de uma sentença tanto à direita e à esquerda, os trabalhos baseados em RNRs adotaram a representação bidirecional (SCHUSTER; PALIWAL, 1997).

Figura 6 – RNR com estado oculto



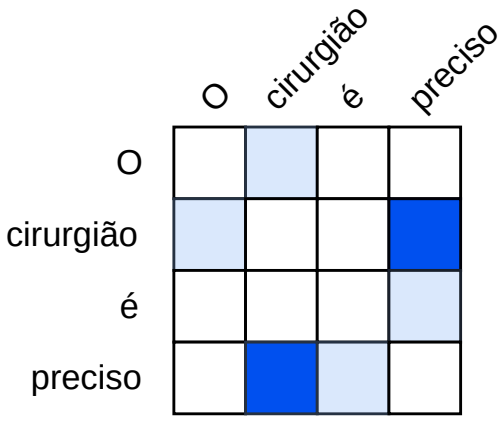
Fonte: Adaptada de (ZHANG *et al.*, 2021).

Apesar das vantagens atribuídas às RNRs, existem dificuldades em manter memórias entre estados muito distantes e maior dificuldade em alcançar a convergência no treinamento dos modelos. (VASWANI *et al.*, 2017) propôs que os benefícios trazidos pelas RNRs podem ser alcançados com uma arquitetura diferente, que não possui as dificuldades apresentadas anteriormente. A arquitetura recebeu o nome de *Transformer*, onde o conceito principal é o uso do mecanismo de auto-atenção para calcular a relevância dos dados de entrada de forma simultânea, sem a necessidade do processamento sequencial das RNRs. Não obstante, o cálculo realizado de forma simultânea permite o uso de técnicas de paralelismo, diminuindo o tempo de inferência e treinamento das redes *Transformer*. O mecanismo de auto-atenção calcula a relevância entre todas as combinações dos itens de entrada. Isto significa que cada palavra associa uma medida de relevância para todas as palavras existentes na sentença. Na Figura 7, é possível visualizar a matriz de atenção, onde a intersecção de duas palavras contém um índice que varia 0 a 1 representando a relevância da conexão. Espera-se que o modelo associe um alto índice de atenção para tuplas de palavras que contribuam para a solução do problema e que o modelo associe um índice baixo para palavras que tenham pouca relevância para o problema.

A matriz de auto-atenção é calculada a partir das matrizes *Key* (K) e *Query* (Q) conforme a Equação 2.1. Os *tokens* de entrada são inicialmente projetados pelas matrizes *Key* (K) e *Query* (Q). Em seguida, a função *softmax* é aplicada para normalizar todos os valores entre 0 e 1. A matriz resultante é multiplicada pela matriz *Value* (V) e dividido pelo fator $\sqrt{d_k}$. Desta forma, a matriz de auto-atenção "pondera" a projeção da representação de cada *token* por meio da matriz *Value*.

$$\text{Attention}\,(Q,K,V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{2.1}$$

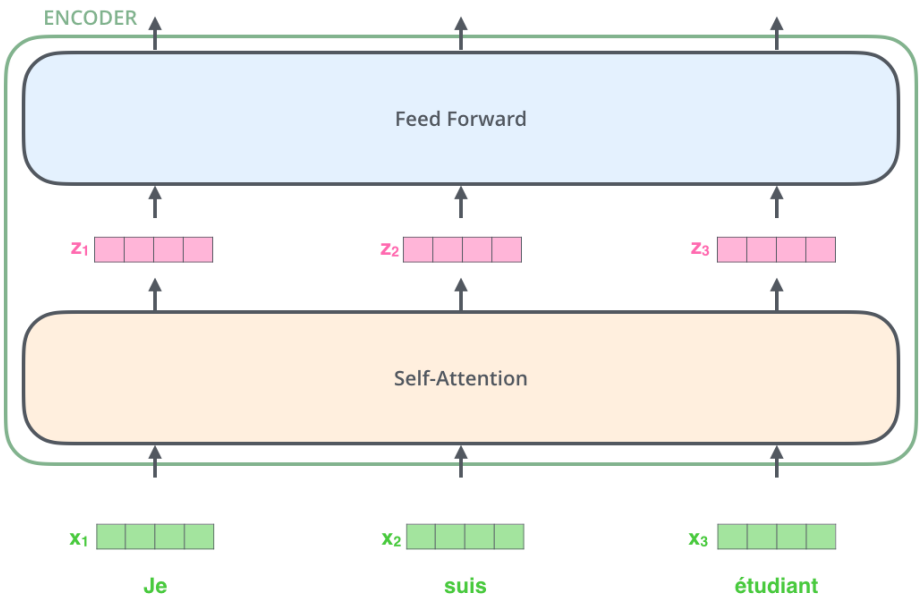
Figura 7 – Exemplo de matriz de auto-atenção



Fonte: Elaborada pelo autor.

Em contraposição às RNRs, a *Transformer* relaxa o viés indutivo do modelo por não assumir o aspecto sequencial dos dados, porém, analisa todas as possíveis combinações entre itens de forma simultânea por meio do mecanismo de auto-atenção. Não obstante, são adicionadas múltiplas matrizes *Q, K, V* para obter maior poder de representação computacional. Esta modificação denomina-se *Multi-Head Attention*. Por fim, a saída é processada por uma camada neural totalmente conectada. A combinação da *Multi-Head Attention* e camada totalmente conectada (*Feed Forward*) compõe o bloco *Encoder* da *Transformer*, conforme ilustrado na Figura 8. O bloco recebe os *tokens* de entrada $X_{1...n}$, então os processa pelo mecanismo de auto-atenção, produzindo os vetores $Z_{1...n}$ e, por fim, são enviados para a camada neural totalmente conectada.

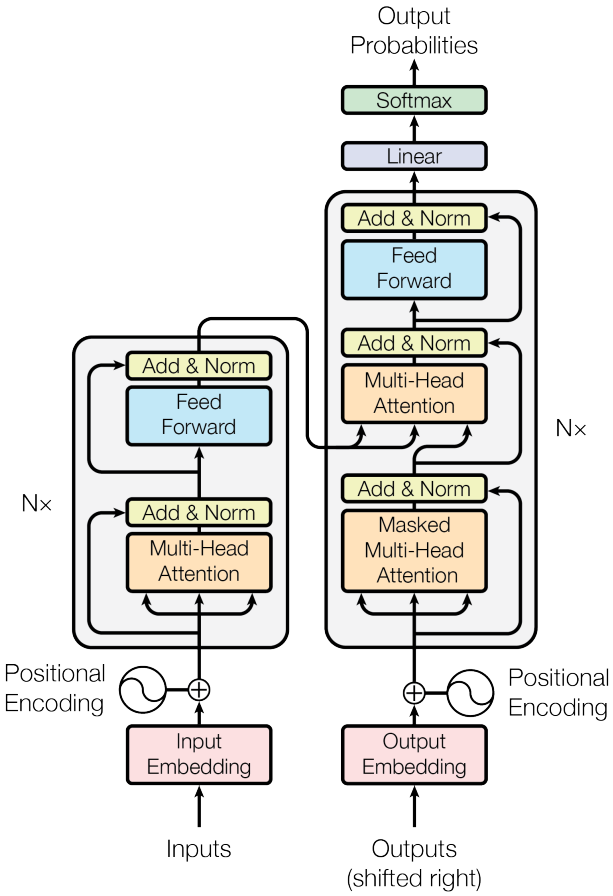
Figura 8 – Bloco *Encoder* da arquitetura *Transformer*



Fonte: (ALAMMAR, 2018).

A arquitetura *Transformer* foi originalmente desenvolvida para a tarefa de tradução automática, portanto, adiciona outras componentes para decodificação de sentenças. A arquitetura completa é apresentada na Figura 9. É possível visualizar o *Encoder* previamente descrito, além do *Decoder* utilizado para a tarefa de tradução automática. Além dos mecanismos descritos anteriormente, a arquitetura também utiliza o sistema de codificação posicional, onde é aplicada uma função em cima das representações vetoriais dos *tokens* de entrada, visando adicionar a informação de posição relativa de cada *token*. Além de alcançar resultados no estado-da-arte na tarefa de tradução, o mecanismo de auto-atenção foi posteriormente adaptado para uso em diversas tarefas.

Figura 9 – Arquitetura *Transformer*



Fonte: (VASWANI *et al.*, 2017).

(DEVLIN *et al.*, 2019) adicionou uma série de modificações na arquitetura *Transformer* e modo de treinamento, produzindo o modelo de língua *Bidirectional Encoder Representations from Transformers* (BERT). O modelo baseia-se no uso do aprendizado auto-supervisionado durante a etapa de pré-treinamento, isto é, a partir de dados não rotulados. Utilizando textos crus, os autores propuseram a criação de duas tarefas para o aprendizado, sendo elas *Next Sentence Prediction* e *Masked Language Model*. A primeira tem o objetivo de receber duas sentenças de entrada e predizer se uma é a continuação da outra. A segunda tarefa substitui *tokens* de uma sentença por uma máscara e o modelo deve predizer qual *token* deve substituir

a máscara para reconstruir a sentença original. Além disso, o BERT é treinado no contexto bidirecional, processando as sentenças de entrada da direita à esquerda e da esquerda à direita simultaneamente.

A partir do treinamento com as tarefas previamente descritas, o modelo aprende a criar representações vetoriais de cada *token* levando todo o contexto da sentença em consideração, por meio do mecanismo de auto-atenção. Após esta etapa, o modelo de língua pode ser ajustado para as tarefas finais, ou seja, o modelo pré-treinado é utilizado como ponto de partida para ajuste fino nas tarefas. Desta forma, todo o poder de representação computacional aprendido durante o pré-treinamento é usufruído no momento do ajuste fino. O ajuste fino é uma forma de realizar a transferência de aprendizado, que busca utilizar o aprendizado de um modelo previamente treinado de uma tarefa T_1 para uma tarefa T_2. A transferência de aprendizado permite que o conhecimento adquirido em T_1 possa ser utilizado em T_2. Isto é particularmente útil em cenários onde se possui um número reduzido de amostras para T_2.

Utilizando a transferência de aprendizado, os autores realizaram o ajuste fino nas tarefas de reconhecimento de entidade nomeada, respostas a perguntas e inferência de conhecimento de senso comum. O uso do mesmo modelo pré-treinado nas três tarefas permitiu alcançar resultados competitivos em relação ao estado-da-arte, mostrando a capacidade de generalização entre diferentes tarefas.

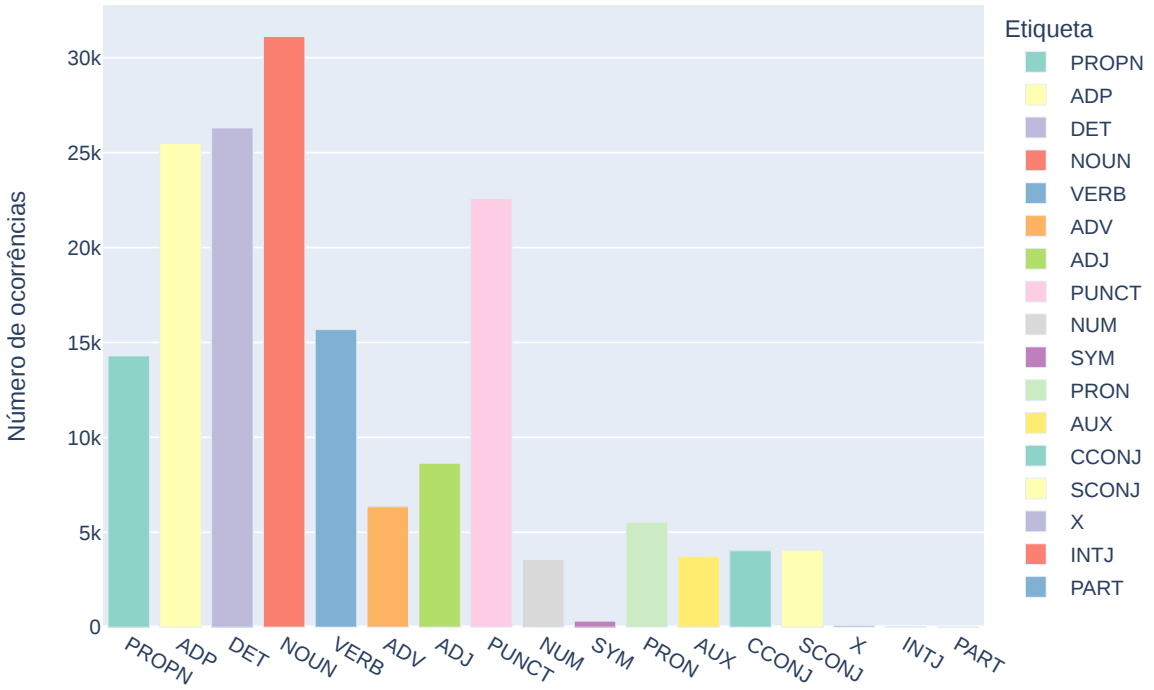
2.4 Avaliação de etiquetadores morfossintáticos

Para avaliação de desempenho de um etiquetador morfossintático, a métrica mais popular é calculada a partir do número de acertos e número total de etiquetas, denominada acurácia, conforme e Equação 2.2. Apesar da versatilidade da acurácia, ela pode ser enganosa caso o córpus esteja desbalanceado. Isto ocorre quando existem diferentes quantidades de ocorrência para cada classe de palavra. A Figura 10 mostra o histograma de etiquetas para o corpus Bosque UD (RADEMAKER *et al.*, 2017), onde é possível visualizar o desbalanceamento no córpus. Segundo o histograma, cerca de 35% de todos os *tokens* são substantivos (*NOUN*). Se um etiquetador sempre associar a classe *NOUN* para todos os *tokens* de entrada, ele apresentará uma acurácia de 35%, porém, a medida pode levar a conclusão enganosa de que esta acurácia abrange todas as classes gramaticais.

$$\text{Acurácia} = \frac{\text{Número de etiquetas corretas}}{\text{Total de etiquetas}} \tag{2.2}$$

Tendo em vista conjuntos de dados desbalanceados, faz-se necessário o uso de métricas que considerem casos de Falso Positivo (FP) e Falso Negativo (FN) separadamente. Os FPs são etiquetas associadas a uma classe diferente da qual deveria ter sido rotulada (um FP também é conhecido como o erro de tipo I). Já os FNs ocorrem quando o modelo não detecta uma

Figura 10 – Histograma das etiquetas morfossintáticas no conjunto de treino do Bosque (RADEMAKER *et al.*, 2017)



Fonte: Elaborada pelo autor.

determinada etiqueta (um FN é denominado como erro de tipo II). Quando um modelo realiza uma predição correta, conta-se como um Verdadeiro Positivo (VP) e, quando deixa de detectar uma etiqueta de forma correta, conta-se como um Verdadeiro Negativo (VN). A partir destas definições, derivaram-se as métricas *Precisão* e *Sensibilidade*, que medem, respectivamente, a taxa de acertos em relação às predições ditas como positivas pelo modelo, e a taxa de acertos em relação a todas as amostras rotuladas como positivas. Ambas as medidas podem ser visualizadas na Equação 2.3 e Equação 2.4.

$$\text{Precisão} = \frac{VP}{VP+FP} \tag{2.3}$$

$$\text{Sensibilidade} = \frac{VP}{VP+FN} \tag{2.4}$$

Outra métrica difundida na literatura é a *Medida-F*, que realiza a média harmônica entre *Precisão* e *Sensibilidade*, podendo desta forma sumarizar o resultado de um algoritmo com uma única métrica. Por se tratar de um problema multiclasses, utilizaremos a métrica *Medida-F* com média macro, ou seja, a métrica será calculada individualmente para cada classe gramatical e

sumarizada pela média aritmética. A Equação 2.5 apresenta a fórmula base para cálculo, sem incluir o cálculo da média.

$$\text{Medida-F} = 2 * \frac{\text{Precisão} * \text{Sensibilidade}}{\text{Precisão} + \text{Sensibilidade}} = \frac{VP}{VP + \frac{1}{2} * (FP + FN)} \tag{2.5}$$

A Figura 11[7] exemplifica o uso das métricas descritas, assim como a seleção dos VPs, FNs e FPs. Inicialmente, verifica-se que os *tokens [humildes, a, sabedoria]* foram corretamente etiquetados, portanto, são 3 VPs. Já os *tokens [Com, os, está]* foram saídas incorretas, sendo assim, contam como 3 FPs. Ao mesmo tempo que houve 3 FPs, o etiquetador não detectou as etiquetas corretas dos *tokens*, portanto, temos 3 FNs. A matriz de confusão apresenta a quantidade de VPs, FPs e FNs. Em seguida, as métricas *Precisão, Sensibilidade e Medida-F* são calculadas para cada classe individualmente e, por fim, a *Medida-F macro* realiza a média aritmética entre todos os valores, resultando no valor $0,325$. Estas métricas são utilizadas para a avaliação de sistemas de classificação multiclasse, onde se encontra a etiquetagem morfossintática. Sendo assim, com essas definições, é possível avaliar e comparar os etiquetadores descritos no próximo capítulo.

[7] A divisão considera o fator ε de $1e-9$

Figura 11 – Exemplo de cálculo de métricas

Sentença original

Sentença analisada

Matriz de confusão

		Saída do sistema			
		ADP	DET	NOUN	AUX
Rótulo	ADP	–	–	1	–
	DET	–	1	–	1
	NOUN	–	–	2	–
	AUX	–	1	–	–

Métricas

	ADP	DET	NOUN	AUX
Precisão	$\frac{0}{0} \simeq 0$	$\frac{1}{2} = 0.5$	$\frac{2}{3} \simeq 0.66$	$\frac{0}{1} = 0$
Sensibilidade	$\frac{0}{1} = 0$	$\frac{1}{2} = 0.5$	$\frac{2}{2} = 1$	$\frac{0}{1} = 0$
Medida-F	0	$2 * \frac{0.5 * 0.5}{0.5 + 0.5} = 0.5$	$2 * \frac{\frac{2}{3} * 1}{\frac{2}{3} + 1} = 0.8$	0

Fonte: Elaborada pelo autor.

CAPÍTULO

3

REVISÃO DA LITERATURA

Este trabalho se baseia nos avanços recentes de métodos para etiquetagem morfossintática, incluindo trabalhos que utilizam os gêneros jornalístico, acadêmico e CGU, no aprendizado em um único gênero e no contexto multigênero. Não obstante, também são apresentados métodos clássicos na área e específicos para o português do Brasil. Além disso, os trabalhos apresentados a seguir são fundamentados nos recentes avanços de modelos de língua por meio do uso de representações contextuais. A estrutura deste capítulo está organizada da seguinte maneira, primeiramente, os métodos de etiquetagem morfossintática são apresentados, em seguida, são apresentados métodos que abordaram a etiquetagem especificamente em textos jornalísticos, CGU e acadêmico. Então, são apresentados os recursos linguísticos (córpus) disponíveis na literatura. Por fim, as considerações finais são expostas.

3.1 Métodos para etiquetagem morfossintática

3.1.1 Métodos baseados em regras

Sistemas baseados em regras podem ser desenvolvidos de várias formas, porém, todos compartilham da premissa de sistemas especialistas, isto é, são projetados com o conhecimento de especialistas da área. Em uma das possíveis abordagens, criam-se regras que refletem o raciocínio de um linguista ao realizar a etiquetagem morfossintática de forma manual. Em contraposição, também existe a possibilidade de realizar o aprendizado automático das regras por meio do aprendizado supervisionado. Em ambos os casos, as regras pré-definidas ou aprendidas são aplicadas e o sistema produz as etiquetas gramaticais de saída.

(KLEIN; SIMMONS, 1963) foi um dos primeiros trabalhos a desenvolver um sistema baseado em regras para a etiquetagem morfossintática em textos extraídos de enciclopédias em inglês. As regras foram desenvolvidas manualmente pelos pesquisadores de forma empírica. Ao fim, o trabalho alcançou acurácia acima de 90% para todas as classes gramaticais. Além disso,

os erros apresentados pelo sistema são causados pelas ambiguidades encontradas no conjunto de dados.

(BRILL, 1992) é um dos principais trabalhos baseados em regras da área. A abordagem foi utilizada para realizar a etiquetagem no córpus Brown (FRANCIS; KUCERA, 1979). O sistema realiza o aprendizado automático de regras por meio do algoritmo *Transformation-based Learning*. Inicialmente, o algoritmo é inicializado com o aprendizado em um grande córpus sem considerar o contexto, dessa forma, cada *token* é sempre classificado com a etiqueta de maior frequência no córpus. Em seguida, a partir de um conjunto de modelos de regras do tipo "Troque a etiqueta **a** pela etiqueta **b** quando a palavra seguinte possuir a etiqueta **z**", o algoritmo aplica os modelos a todas as etiquetas e verifica se o desempenho do etiquetador melhorou, se sim, a regra aprendida é mantida, caso contrário, o algoritmo segue para o próximo modelo. Nos experimentos realizados pelos autores, foi encontrado que o aprendizado automático de 71 regras obteve acurácia acima de 90% no conjunto de testes, sendo equiparável aos sistemas probabilísticos da época. Não obstante, o uso de um conjunto pequeno de regras ocupa menos memória do que grandes tabelas de frequência (utilizadas por métodos estatísticos).

3.1.2 Métodos probabilísticos

Os sistemas que utilizam abordagem probabilística se baseiam na abordagem empiricista, desta forma, utilizam dados rotulados para inferir a probabilidade de ocorrência de uma etiqueta ou de uma sequência de etiquetas. Métodos probabilísticos podem levar o contexto completo ou parcial em consideração. A seguir são apresentados trabalhos baseados na abordagem probabilística.

(KEMPE, 1993) utilizou *Hidden Markov Model* (HMM) no córpus *Penn Treebank* (MARCUS; SANTORINI; MARCINKIEWICZ, 1993a), sendo multigênero e em inglês, contendo textos jornalísticos, manuais técnicos e livros, entre outros. O método HMM possui a premissa de que o cálculo do estado atual, isto é, da etiqueta a ser associada ao *token* de entrada, depende unicamente do estado anterior (previamente calculado)[8]. Sendo assim, o método não considera outros estados anteriores para a predição. O método permite que analisemos o estado atual (*token*) considerando o estado anterior, ou seja, se o HMM inicialmente observar o *token menina* e saber que o estado anterior produziu a classe gramatical *DET*, então *menina* provavelmente possuirá maiores chances de possuir a etiqueta *NOUN*. Com o método que associa probabilidades de transição entre estados (HMM), os autores alcançaram acurácia de 96%.

Conditional Random Fields (CRF) é outra técnica de AM probabilística bastante difundida na literatura. CRFs são modelos gráficos discriminativos probabilísticos, que também baseiam-se na premissa Markoviana. Não obstante, CRFs maximizam diretamente a probabilidade $P(Y|X)$, enquanto HMMs maximizam a distribuição conjunta. EKBAL; HAQUE; BANDYOPADHYAY (2007) aplicam CRFs para o bengali, alcançando acurácia de 90,3%.

[8] Premissa Markoviana

3.1.3 Redes neurais

As redes neurais artificiais são amplamente utilizadas na etiquetagem morfossintática e em outras tarefas de PLN. A capacidade de representação computacional apresentada pelos algoritmos baseados em redes neurais permite alcançar alto desempenho em diversas tarefas. Em especial, na etiquetagem morfossintática tem se popularizado o uso das representações contextuais e estáticas, em combinação de uma camada classificadora *softmax*, porém, diversos trabalhos utilizam as redes neurais em combinação de outras abordagens, como o CRF. A seguir são apresentados os principais trabalhos da área que utilizam métodos baseados em redes neurais artificiais.

UDPipe (STRAKA; HAJIČ; STRAKOVÁ, 2016) é um analisador morfossintático baseado na rede *perceptron*, utilizando atributos extraídos a partir de um conjunto de regras pré-definidas. Em seguida, o UDpipe 2 (STRAKA, 2018) passa a utilizar RNRs com uso da arquitetura *Bidirectional Long Short-Term Memory* (Bi-LSTM). Além disso, a entrada da Bi-LSTM é uma combinação de três representações vetoriais, sendo elas: (1) representação vetorial estática pré-treinada com FastText (JOULIN *et al.*, 2016), (2) representação vetorial inicializada aleatoriamente, e (3) representação vetorial em nível de caracteres utilizando RNRs com células GRU. Posteriormente, ao UDPipe 2.1 (STRAKA; STRAKOVÁ; HAJIC, 2019), foi adicionada a representação contextual por meio do modelo BERT, pré-treinado com dados da Wikipedia[9]. O UDPipe 2 reportou *Medida-F* média de 89.67% para 73 *treebanks* de diferentes línguas, em destaque, 96.37% para o português do Brasil no Bosque (RADEMAKER *et al.*, 2017).

O trabalho desenvolvido por KONDRATYUK; STRAKA (2019) apresentou o modelo Udify, que realiza o treinamento multi-tarefa para a etiquetagem morfossintática, análise de dependências sintáticas, segmentação de *tokens* e análise morfológica. Os autores adicionaram conexões residuais entre cada camada de auto-atenção e a representação final da sentença, com a motivação de que as representações contextuais das primeiras camadas do modelo de língua contêm características sintáticas e morfológicas, portanto, a hipótese é de que o uso de conexões residuais para a representação final pode contribuir para a etiquetação morfossintática e análise de dependências. O Udify alcançou Medida-F média de 93.76% em 124 *treebanks* e, em particular, 97.10% para o português do Brasil no córpus jornalístico Bosque (RADEMAKER *et al.*, 2017).

YASUNAGA; KASAI; RADEV (2018) conjecturaram que o uso de *Adversarial Training* (AT) contribuiria para gerar representações contextuais mais ricas e melhorar o desempenho de etiquetadores morfossintáticos. O AT busca minimizar o impacto de flutuações aleatórias na representação vetorial, isto é, dada uma representação vetorial de um *token*, ao adicionarmos um ruído aleatório a mesma, o desempenho da rede neural deve ter pouca variação, porém, o que geralmente é observado é que pequenas perturbações na representação vetorial variam bastante o desempenho do método. Sendo assim, o AT busca adicionar exemplos contraditórios durante a etapa de treinamento para que o impacto das perturbações aleatórias seja reduzido, funcionando

[9] Disponível em: <https://huggingface.co/bert-base-multilingual-uncased>.

como um método de regularização de aprendizado, consequentemente, aumentando a capacidade de generalização do algoritmo. Para gerar exemplos contraditórios, os autores buscaram encontrar a perturbação de pior caso, por meio da função de perda, conforme demonstrado na Equação 3.1, onde θ são os parâmetros do modelo, s é a representação vetorial da sentença em nível de palavra e caractere, η' é a perturbação aleatória e y é o vetor de rótulos. Desta forma, será gerada uma perturbação aleatória que aumenta o valor da função de custo (direção crescente do vetor gradiente), fazendo com o que o modelo tenha que se adaptar aos exemplos contraditórios para realizar o aprendizado.

$$\eta = \underset{\eta':\|\eta'\|_2 \leq \varepsilon}{\arg\max}\, L\left(\hat{\theta}; s+\eta', y\right) \tag{3.1}$$

Os autores utilizaram a arquitetura LSTM bidirecional em nível de caractere com CRF em conjunto com o AT. Analisando os resultados em 27 *treebanks* da UD, o método obteve Acurácia média de 96.65% e, em especial, 98.07% para o português no Bosque. O uso de AT demonstrou o aumento médio na Acurácia de $0,25\%$. Adicionalmente, os autores demonstraram que o uso de AT teve contribuição para a regularização em *treebanks* de línguas de pouco recurso e melhora o desempenho em sentenças com palavras fora de vocabulário, também conhecidas como *Out-of-Vocabulary* (OOV).

O trabalho de (QI *et al.*, 2020) apresenta um conjunto de ferramentas para processamento de textos e extração de classes gramaticais, atributos morfológicos, árvore sintática, entre outros. Para a etiquetagem morfossintática, os autores utilizaram uma rede neural Bi-LSTM com camada *Biaffine Attention* proposta por DOZAT; MANNING (2016), que foi originalmente utilizada na tarefa de análise sintática automática. O *Biaffine Attention* é uma extensão do mecanismo de atenção tradicional que utiliza uma função de mapeamento bilinear para combinar as representações das palavras em uma matriz de atenção. Com base nesta arquitetura, os autores avaliaram o modelo em todos os *treebanks* disponíveis na versão v2.5 da UD, obtendo acurácia de $97,04\%$ no córpus Bosque.

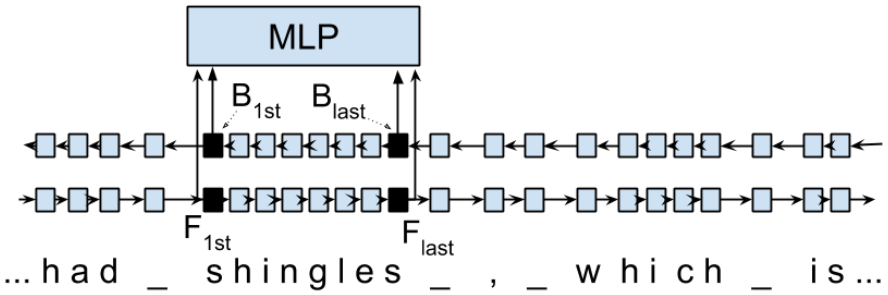
CUI; ZHANG (2019) mencionam a dificuldade na comparação entre os métodos baseados em LSTMs bidirecionais com *softmax* e com CRF. A hipótese apresentada é de que a premissa Markoviana do CRF degrada o desempenho dos modelos. Assim, a arquitetura *Label Attention Network* (LAN) é apresentada, utilizando o número de cabeças da camada de auto-atenção igual ao número de rótulos existentes, buscando que cada cabeça represente uma etiqueta morfossintática. Os autores empregam a arquitetura LSTM bidirecional, utilizando uma combinação da representação RNR com a auto-atenção proposta (LAN). O viés indutivo apresentado permite que o modelo tenha maior poder de representação e, consequentemente, um desempenho maior. Os autores avaliaram a arquitetura proposta em 7 *treebanks* de línguas de alto recurso da UD, obtendo a acurácia média de 96.88% e 98.04% para o português do Brasil.

HEINZERLING; STRUBE (2019) buscam avaliar a contribuição do uso de representa-

ções contextuais e não contextuais para as tarefas de reconhecimento de entidade nomeada e etiquetagem morfossintática no contexto multilíngue, realizando as comparações em 27 línguas, utilizando *treebanks* da UD. A combinação de diferentes representações é realizada pela concatenação dos vetores de cada tipo de representação. Com a execução de experimentos com múltiplas configurações e combinações de diferentes representações (contextuais, não contextuais e em nível de caractere), conclui-se que, em média, o modelo multilíngue BERT (DEVLIN *et al.*, 2019), em conjunto com representações vetoriais em nível de caractere e representações não contextuais BPemb (HEINZERLING; STRUBE, 2018), obteve a melhor *Medida-F* média de 96.8% nos 27 *treebanks* avaliados e, em particular, 98.1% para o português do Brasil. Além disso, nos experimentos realizados em *treebanks* de línguas de baixo recurso, isto é, com poucos recursos linguísticos disponíveis, observou-se que a melhor combinação foi o uso de representações vetoriais estáticas em conjunto de representações em nível de caractere.

BOHNET *et al.* (2018) constroem uma abordagem fundamentada na premissa de que o uso de diferentes representações vetoriais pode contribuir para o desempenho na tarefa. O trabalho propõe o uso de representações vetoriais em nível de caractere e em nível de sentença (baseada em caracteres), onde a segunda é dada pela concatenação dos vetores direcionais para o primeiro e último caractere de cada palavra, conforme a Figura 12. Cada representação vetorial é construída a partir da arquitetura LSTM bidirecional. Em seguida, ambas as representações são utilizadas como entrada de outra LSTM bidirecional, a qual faz a predição das etiquetas morfossintáticas, baseada na concatenação das representações vetoriais antecedentes. Os autores averiguaram que o treinamento separado das representações vetoriais alcançou melhor desempenho em comparação ao treinamento em conjunto. O modelo proposto alcançou resultados do estado-da-arte no córpus *Wall Street Jounal* de 97.96% (acurácia), também reportando resultados para um subconjunto da UD, porém, na tarefa de análise morfológica.

Figura 12 – Representação vetorial baseada em caractere em nível de sentença.



Fonte: (BOHNET *et al.*, 2018).

Os modelos de língua também são utilizados para a etiquetagem morfossintática, em

especial, os modelos baseados na arquitetura BERT (DEVLIN *et al.*, 2019). VRIES *et al.* (2019) realizam o treinamento de um modelo baseado no BERT para a língua alemã e comparam o modelo treinado com o modelo BERT multilíngue na tarefa de etiquetagem. Para utilizar o modelo de língua treinado, os autores adicionam uma camada totalmente conectada à rede neural com o número de neurônios igual à quantidade de etiquetas existentes no córpus. Em seguida, é iniciado o processo de ajuste fino do modelo. Esta etapa inicia um novo treinamento com base no modelo previamente pré-treinado, buscando atualizar os pesos da nova camada adicionada à rede neural. Utilizando a estratégia de ajuste fino, o modelo alcançou 96,6% de acurácia no *treebank* para a língua alemã UD-LassySmall[10], enquanto o modelo multilíngue BERT alcançou acurácia de 92,5%, dessa forma, demonstrando a vantagem de utilizar modelos especializados em uma única língua. Semelhantemente, as análises de VIRTANEN *et al.* (2019) apontam para a mesma evidência na língua finlandesa.

3.2 Etiquetadores para o português do Brasil

AIRES *et al.* (2000) apresentam a avaliação de diferentes modelos para etiquetagem no português do Brasil, sendo eles: Unigram (Treetagger), N-gram (Treetagger), *transformation-based learning* (TBL) e *Maximum-Entropy tagging* (MXPOST). O último obteve o melhor resultado entre os métodos avaliados no córpus NILC[11], alcançando a acurácia média de 88.73%. Adicionalmente, os autores combinaram todos os métodos em um comitê de máquinas com base na moda, isto é, todos os modelos avaliavam os *tokens* de entrada, em seguida, a etiqueta que obteve mais votos era utilizada como saída do comitê de máquinas. Com esta abordagem, os autores alcançaram acurácia média de 89,42%.

Similar às abordagens multilíngues apresentadas anteriormente, SOUSA; LOPES (2019a) utilizam RNRs bidirecionais com representações vetoriais em nível de palavra e caractere. O trabalho avalia a etiquetagem morfossintática no córpus MAC-MORPHO (ALUÍSIO *et al.*, 2003), que contém sentenças do gênero jornalístico. A abordagem proposta alcançou 97,36% de acurácia.

DOMINGUES (2011) apresenta um etiquetador que utiliza o aprendizado baseado em transformações de (BRILL, 1995) para os gêneros jornalístico e acadêmico. Foi adicionado o uso de léxico para tratamento de nomes próprios, regras manuais e a saída de outros dois etiquetadores disponíveis na literatura. O trabalho utilizou os córpus de gênero jornalístico MAC-MORPHO (ALUÍSIO *et al.*, 2003) e Bosque (LINGUATECA, 2009) e, para o gênero acadêmico, a Selva Científica (LINGUATECA, 2009). A avaliação apresentou as acurácias de 98,06%, 98,30% e 98,07%, respectivamente.

É importante ressaltar que nesta seção foram apresentados trabalhos para o português do

[10] Disponível em: <https://universaldependencies.org/treebanks/nl_lassysmall/index.html>.
[11] Disponível em: <http://www.nilc.icmc.usp.br/nilc/tools/corpora.htm>.

Brasil que utilizaram o formalismo UD, córpus multigênero, ou com pelo menos um dos gêneros que fazem parte deste estudo. Além disso, outros trabalhos (AIRES, 2000; BRANCO; SILVA, 2004; KEPLER, 2010) abordaram a etiquetagem morfossintática para o português do Brasil e possuem desempenho similar ou inferior aos métodos citados anteriormente. Não obstante, os métodos citados na subseção anterior apresentam acurácias entre 97% e 98% para o português do Brasil no córpus Bosque, dessa forma, é possível observar que as abordagens apresentadas nesta subseção possuem desempenho similar.

3.3 Etiquetagem em textos jornalísticos, CGU e acadêmico

Os trabalhos apresentados na seção anterior continham córpus de múltiplos gêneros. Tendo em vista a avaliação no contexto multigênero proposta por este trabalho, esta seção apresenta trabalhos voltados aos gêneros textuais jornalístico, acadêmico e CGU, com o objetivo de apresentar diferentes abordagens e o desempenho dos etiquetadores para cada gênero.

Os gêneros jornalístico e acadêmico foram dois dos primeiros gêneros a possuir etiquetadores morfossintáticos na literatura. Para o inglês, o córpus Brown possibilitou a criação de trabalhos baseados na abordagem empiricista. Não obstante, os primeiros trabalhos foram baseados em grandes tabelas contendo a palavra e sua respectiva classe gramatical (KLEIN; SIMMONS, 1963), ou com aprendizado automático das regras (BRILL, 1995). Posteriormente foram desenvolvidas abordagens probabilísticas para o gênero (KEMPE, 1993; EKBAL; HAQUE; BANDYOPADHYAY, 2007). Recentemente, as abordagens de Aprendizado de Máquina utilizando redes neurais elevaram o desempenho dos sistemas ao atual estado da arte (SOUSA; LOPES, 2019a).

Com o advento de córpus CGU com etiquetas morfossintáticas a partir de 2014 (KONG *et al.*, 2014; SILVEIRA *et al.*, 2014), diversos autores adaptaram abordagens já utilizadas no gênero jornalístico para CGU. Não obstante, alguns desafios surgiram com mais frequência e dificuldade no gênero CGU, por exemplo, em córpus CGU é frequente a ocorrência de OOVs, ou seja, palavras às quais o algoritmo não obteve acesso durante o treinamento, porém, que estão presentes no conjunto de testes. Métodos probabilísticos como HMM são prejudicados por não possuírem mecanismos efetivos para atribuir probabilidades a OOVs. Ao mesmo tempo, com o advento de tokenizadores que trabalham em nível de sub-palavra (SENNRICH; HADDOW; BIRCH, 2016; WU *et al.*, 2016), abordagens baseadas em redes neurais melhoraram o desempenho em textos onde existe uma alta ocorrência de OOVs. A seguir são apresentados trabalhos voltados ao CGU, indicando o córpus utilizado e o desempenho alcançado.

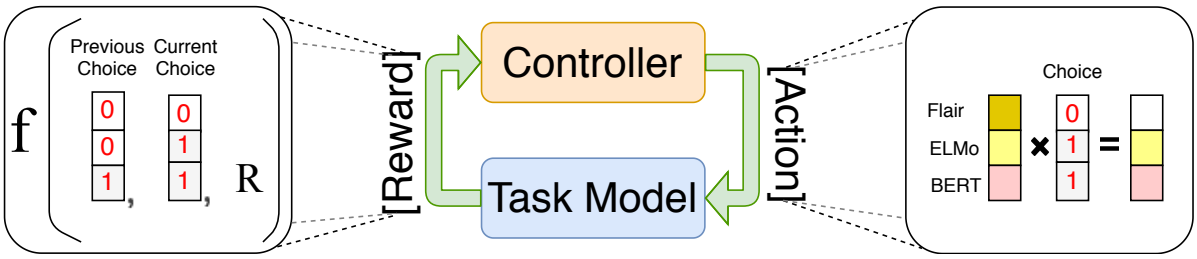
NGUYEN; VU; NGUYEN (2020) realizam o treinamento de um modelo de língua baseado no BERT (DEVLIN *et al.*, 2019), com córpus extraído da rede social *Twitter*. Os autores apresentam uma melhoria consistente de desempenho em relação a outros modelos de língua que

não foram pré-treinados com textos do *Twitter* (comparando os resultados com o BERT). Para a tarefa de etiquetagem morfossintática no *treebank Tweebank v2* (LIU *et al.*, 2018), o método obteve acurácia média de 95.2%.

VIRTANEN *et al.* (2019) também realizam o mesmo procedimento descrito no parágrafo anterior para o finlandês. Os autores compararam o modelo de língua treinado com o BERT multilíngue (DEVLIN *et al.*, 2019). Com a configuração proposta, observou-se acurácia média de 98.23%, uma melhora de 1.26% em relação ao BERT. Os autores concluem que o uso de modelos monolíngue para línguas de baixo recurso são essenciais para obter alto desempenho na tarefa final.

O uso conjunto de diferentes níveis de representações vetoriais aumenta o desempenho dos etiquetadores, porém, não existe consenso entre os pesquisadores sobre quais combinações trazem o melhor resultado para a tarefa. WANG *et al.* (2021) propõem um método baseado em aprendizado por reforço para encontrar a combinação de representações vetoriais que alcance o melhor desempenho. O algoritmo possui uma hipótese (*belief*) de qual é a melhor combinação inicial, por exemplo, a combinação das representações contextuais do BERT (DEVLIN *et al.*, 2019) e ELMo (PETERS *et al.*, 2018). Em seguida, o algoritmo realiza o treinamento de um modelo baseado na concatenação das representações combinadas e retorna a acurácia. A acurácia é utilizada na função recompensa do algoritmo, dessa forma, garantindo maiores recompensas para as melhores combinações de entrada. Então, a hipótese inicial é atualizada com base no algoritmo *policy gradient* (WILLIAMS, 1992). A arquitetura descrita pode ser visualizada na Figura 13, onde é possível visualizar a interação entre o controlador e o modelo da tarefa, além disso, a figura apresenta um vetor de valores binários que representa quais as representações vetoriais foram utilizadas naquele momento e, com base no vetor da iteração anterior e o vetor da atual, a função de recompensa é calculada, realizando a diferença entre o desempenho da escolha atual com a da iteração anterior. Por fim, com o uso da abordagem proposta, o melhor modelo obteve acurácia média de 95,8% no *treebank Tweebank v2* (LIU *et al.*, 2018).

Figura 13 – Arquitetura ACE



Fonte: (WANG *et al.*, 2021).

TSURUOKA *et al.* (2005) apresentam um etiquetador multigênero para os gêneros jornalístico e acadêmico utilizando textos de biomedicina para a língua inglesa. O trabalho baseia-se na arquitetura de redes de dependência cíclica proposta por TSURUOKA *et al.*

(2005), sendo um modelo probabilístico bidirecional. O modelo utiliza como entrada diversos atributos extraídos dos textos de entrada, como o *token* anterior, a etiqueta anterior, prefixos, sufixos, entre outros e produz como saída o conjunto de etiquetas que maximiza a probabilidade de ocorrência das etiquetas no determinado contexto. A avaliação do trabalho consiste em avaliar o desempenho do modelo em córpus de gênero acadêmico com os córpus GENIA (OHTA; TATEISI; KIM, 2002), PennBioIE (KULICK *et al.*, 2004) e jornalístico WSJ (*Wall Street Jounal*) (MARCUS; SANTORINI; MARCINKIEWICZ, 1993b) em diferentes cenários; buscando avaliar a capacidade multigênero do método quando treinado nos gêneros jornalístico e acadêmico, dessa forma, os autores treinaram o modelo isoladamente em cada córpus e realizaram mais um treinamento concatenando os dois córpus, posteriormente, avaliaram se o aprendizado multigênero (córpus concatenados) deteriorou o aprendizado isolado em cada córpus. Isoladamente, o método obteve acurácias de 97,20% no WSJ, 98,55% no GENIA e 97,92% no PennBioIE, em contraposição, ao realizar o treinamento em todas os córpus, foram obtidas as acurácias de 97,20% no WSJ, 98,35% no GENIA e 97,87% no PennBioIE. É possível observar um pequeno decréscimo ao realizar o experimento multigênero nos córpus de gênero acadêmico, contudo, os autores não informam se as diferenças são estatisticamente significativas.

3.4 Recursos linguísticos

Esta seção visa apresentar os principais recursos linguísticos para etiquetagem morfossintática nos gêneros jornalístico, acadêmico e CGU, buscando ressaltar a língua de origem e o conjunto de etiquetas utilizados.

O córpus Brown (FRANCIS; KUCERA, 1967) foi um dos primeiros grandes córpus estruturados para o inglês, possuindo cerca de um milhão de *tokens* rotulados com etiquetas morfossintáticas. O conjunto de etiquetas é composto por 80 classes gramaticais.

Para o português do Brasil, o projeto Floresta Sintática (AFONSO *et al.*, 2002) foi um dos primeiros a criar córpus em larga escala. Dentre eles, encontra-se o Bosque, um córpus estruturado com mais 200 mil *tokens* do gênero jornalístico que contém sentenças em português do Brasil e de Portugal. Inicialmente, o córpus foi rotulado com o formalismo *Constraint Grammar* e, recentemente, foi portado para as diretrizes da UD (RADEMAKER *et al.*, 2017). O Bosque é uma coletânea de textos jornalísticos rotulados com informações morfossintáticas, sintáticas e morfológicas.

Semelhantemente, na Floresta Sintática encontra-se o subconjunto denominado Selva Científica, sendo uma coletânea de textos acadêmicos extraídos da Wikipedia, relatórios dos bancos centrais europeu e do Brasil. O córpus é composto por 6.200 sentenças e cerca de 125.000 palavras. Além disso, as sentenças são textos em português do Brasil e de Portugal e, até o momento, não foi convertido para a UD.

Dados os trabalhos de gênero jornalístico e acadêmico (BRILL, 1992; SOUSA; LOPES,

2019a; AIRES *et al.*, 2000; DOMINGUES, 2011) apresentados anteriormente, diversos trabalhos de PLN foram desenvolvidos com base nesses gêneros. Não obstante, com o crescente número de usuários na internet, o CGU se tornou objeto de estudo na área. O CGU apresenta desafios que não estavam presentes no gênero jornalístico, como a falta de pontuação e uso de elementos metalinguísticos (*hashtag*), entre outros. Desta forma, a literatura vem apresentando a criação de córpus do gênero CGU com rótulos morfossintáticos e outros atributos linguísticos.

(SANGUINETTI *et al.*, 2018) apresenta a criação de um córpus de CGU baseado no formalismo UD utilizando dados da rede social *Twitter* para o italiano. O córpus contém 6.712 *tweets* rotulados com etiquetas morfossintáticas, dependências sintáticas e atributos morfológicos. Os autores apresentam o problema da segmentação de sentenças de *tweets*. Esta segmentação é geralmente realizada pela delimitação do símbolo . (ponto), porém, os *tweets* apresentam diversos casos de falta de pontuação. Dessa forma, os autores optaram por não segmentar *tweets* em diferentes sentenças, sendo assim, cada amostra no córpus é um único *tweet*, sendo que ele pode conter mais de uma sentença.

Tweebank v2 (LIU *et al.*, 2018) é o córpus de *tweets* rotulado com o formalismo UD para o inglês. O córpus possui ao todo 55.067 *tokens* e 3.550 *tweets*. Este trabalho segue a diretriz de (SANGUINETTI *et al.*, 2018) e utiliza um *tweet* como unidade de análise, ou seja, não é segmentado por sentenças, mas, por *tweets*. Não obstante, os autores compararam três métodos de etiquetagem disponíveis na literatura, onde o etiquetador Stanford CoreNLP (OWOPUTI *et al.*, 2013) obteve a melhor acurácia de 94,6%.

Na Tabela 1 é possível visualizar os principais córpus de gênero CGU para línguas[12] estrangeiras. Além disso, é indicada a língua de origem, se utiliza o formalismo UD e a fonte de origem, sendo o Twitter a mais comum entre os trabalhos apresentados.

Tendo em vista a quantidade de recursos disponíveis para línguas estrangeiras que possuem uma variedade de gêneros textuais, o projeto Porttinari (PARDO *et al.*, 2021) apresenta a criação de um grande córpus multigênero para o português do Brasil. O trabalho agrega textos jornalísticos extraídos do jornal *Folha de São Paulo* e do córpus MAC-MORPHO (ALUÍSIO *et al.*, 2003), *tweets* do mercado de ações com o córpus *Dependency-ANalised corpora of TwEets* (DANTE) (SILVA; ROMAN; CARVALHO, 2020) e avaliações de usuários em *e-commerce* e livros (REAL; OSHIRO; MAFRA, 2019; BELISáRIO; FERREIRA; PARDO, 2020). O trabalho segue as diretrizes da UD, o que torna este projeto compatível com *treebanks* que já utilizam o formalismo, como o Bosque (RADEMAKER *et al.*, 2017).

Di FELIPPO *et al.* (2021) apresentam o córpus DANTEStocks com etiquetas morfossintáticas, baseado no trabalho de (SILVA; ROMAN; CARVALHO, 2020). O córpus conta com

[12] AR: Árabe, HI: Híndi, EN: Inglês, AAE: Inglês Afro-Americano, MAE: Inglês Americano Convencional, IT: Italiano, FR: Francês, FI: Finlandês, TR: Turco, DE: Alemão, AA: Árabe Argeliano, SgE: Inglês de Singapura, ZH: Chinês.

Tabela 1 – Visão geral de *treebanks* CGU, juntamente com algumas informações básicas sobre a fonte de dados, os idiomas envolvidos e se eles são baseados no formalismo UD ou não. Em *treebanks* não UD, ‡ e ⋆ indicam, respectivamente, uma representação sintática de constituinte ou dependência

Nome	**Referências**	**Origem**	**Língua**	**UD**
ATDT (UD)	(ALBOGAMY; RAMSAY, 2017)	Twitter	AR	sim
Hi-En-CS	(BHAT *et al.*, 2018)	Twitter	HI/EN	sim
TwitterAAE (TAAE)	(BLODGETT; WEI; O'CONNOR, 2018)	Twitter	AAE, MAE	sim
TWITTIRÒ-UD (TWRO)	(CIGNARELLA; BOSCO; ROSSO, 2019)	Twitter	IT	sim
DWT	(DAIBER; GOOT, 2016)	Twitter	EN	não⋆
W2.0	(FOSTER *et al.*, 2011)	Twitter, sport fora	EN	não‡
Foreebank (Frb)	(KALJAHI *et al.*, 2015)	fórum técnico	EN, FR	não‡
Tweebank (Twb)	(KONG *et al.*, 2014)	Twitter	EN	não⋆
Tweebank2 (Twb2)	(LIU *et al.*, 2018)	Twitter	EN	sim
TDT	(LUOTOLAHTI *et al.*, 2015)	variado	FI	sim
xUGC	(ALONSO; SEDDAH; SAGOT, 2016)	variado	FR	sim
ITU	(PAMAY *et al.*, 2015)	n.a.	TR	não⋆
tweeDe	(REHBEIN; RUPPENHOFER; DO, 2019)	Twitter	DE	sim
PoSTWITA-UD (Pst)	(SANGUINETTI *et al.*, 2018)	Twitter	IT	sim
FSMB	(SEDDAH *et al.*, 2012)	Twitter, Facebook fórum	FR	não‡
Narabizi (NBZ)	(SEDDAH *et al.*, 2020)	fórum de notícias	AA	sim
EWT	(SILVEIRA *et al.*, 2014)	variado	EN	sim
MoNoise (MNo)	(GOOT; NOORD, 2018)	Twitter	EN	sim
STB	(WANG *et al.*, 2017)	fórum	SgE	sim
CWT	(WANG *et al.*, 2014)	Twitter, Sina Weibo	ZH	não⋆
GUM	(ZELDES, 2017)	variado	EN	sim

Fonte: Adaptada de (SANGUINETTI *et al.*, 2020).

4.517 *tweets* contendo menções às ações da bolsa de valores B3[13]. Os autores optaram por seguir o trabalho de (SANGUINETTI *et al.*, 2018) e não segmentar em sentenças, mas utilizar cada *tweet* como uma unidade mínima de anotação.

Semelhantemente, visando aumentar a quantidade e diversidade de recursos linguísticos disponíveis para o português do Brasil, o projeto interinstitucional Petrolês[14] apresenta a criação de diversos córpus, incluindo o PetroGold SOUZA *et al.* (2021). Este córpus é do gênero acadêmico e contém teses, dissertações e monografias da área de óleo e gás das quais foram extraídas as sentenças para construção de um *treebank*. Os autores seguem as diretrizes da UD e alcançam o total de 253.640 *tokens* com etiquetas morfossintáticas, anotações sintáticas,

[13] Disponível em: <https://www.b3.com.br/pt_br/>.
[14] Disponível em: <https://petroles.puc-rio.ai/>.

Tabela 2 – Síntese de etiquetadores morfossintáticos e suas características. ‡ se refere à acurácia e ⋆ à *Medida-F*

Método	Gênero textual	Desempenho	Abordagem	Idioma
HMM (KEMPE, 1993)	multigênero	96,00% ‡	Probabilísitca	Inglês
CRF (EKBAL; HAQUE; BANDYOPADHYAY, 2007)	Jornalístico	90,30% ‡	Probabilísitca	Bengali
UDPipe 2 (STRAKA, 2018)	multigênero	89,67% ⋆	RNR	Multilíngue
Udify (KONDRATYUK; STRAKA, 2019)	multigênero	93,76% ⋆	BERT + softmax	Multilíngue
Stanza (QI *et al.*, 2020)	multigênero	92,49% ⋆	RNR + *Biaffine Attention*	Multilíngue
AT (YASUNAGA; KASAI; RADEV, 2018)	multigênero	96,65% ‡	BiLSTM char/word	Multilíngue
LAN (CUI; ZHANG, 2019)	multigênero	96,88% ‡	BiLSTM + LAN	Multilíngue
BERT-Bpemb (HEINZERLING; STRUBE, 2019)	multigênero	96,80% ‡	BERT + BPEmb char/word	Multilíngue
Meta-BiLSTM (BOHNET *et al.*, 2018)	Jornalístico	97,96% ‡	Meta-BiLSTM	Inglês
MXPOST (AIRES *et al.*, 2000)	Prosa	89,42% ‡	Comitê de máquinas	Português
BiLSTM (SOUSA; LOPES, 2019a)	Jornalístico	97,36% ‡	BiLSTM char/word	Português
TBL (DOMINGUES, 2011)	multigênero	98,14% ‡	*Transformation-based learning*	Português
Rede Cíclica (TSURUOKA *et al.*, 2005)	multigênero	97,81% ‡	*Cyclic dependency network*	Inglês
BERTweet (NGUYEN; VU; NGUYEN, 2020)	CGU	95,20% ‡	BERT + softmax	Inglês
FinBERT (VIRTANEN *et al.*, 2019)	CGU	98,23% ‡	BERT + softmax	Finlandês
Bertje (VRIES *et al.*, 2019)	multigênero	96,60% ‡	BERT + softmax	Alemão
ACE (WANG *et al.*, 2021)	CGU	95,08% ‡	Aprendizado por reforço	Inglês
Bi-LSTM (BHAT *et al.*, 2018)	CGU	90,53% ‡	Bi-LSTM	Hindi-Inglês

Fonte: Elaborada pelo autor.

atributos morfológicos, entre outros.

Além dos córpus apresentados, encontram-se córpus para o português com etiquetas morfossintáticas que utilizam conjuntos de etiquetas diferentes da UD, como o MAC-MORPHO (ALUÍSIO *et al.*, 2003) que possui mais de um milhão de palavras rotuladas a partir de notícias do jornal Folha de São Paulo[15], utilizando conjunto de etiquetas próprio. De tamanho similar, o córpus CINTIL (CINTIL..., 2011) contém textos do português de Portugal e possui um subconjunto menor com textos do português do Brasil. Além disso, o córpus possui conjunto de etiquetas próprios.

3.5 Considerações finais

Foram apresentados métodos para etiquetagem morfossintática, além de incluir métodos específicos para o português do Brasil e que abordaram o contexto multigênero, considerando os gêneros jornalístico, acadêmico e CGU. Além disso, foram apresentados os córpus disponíveis na literatura, considerando o formalismo UD. A Tabela 2 sintetiza as características dos trabalhos apresentados neste capítulo. Observa-se que a maioria dos trabalhos voltados a gêneros jornalísticos e outros gêneros que, não incluam CGU, alcançam altíssimo desempenho, porém, nos trabalhos contendo texto CGU, apenas (VIRTANEN *et al.*, 2019) alcança acurácia acima de 98%.

Em métodos baseados em redes neurais, observa-se com frequência o uso de representações vetoriais em nível de caractere e palavra, onde a combinação dessas representações (ou mais) melhoram o desempenho final do etiquetador (SOUSA; LOPES, 2019a; HEINZERLING;

[15] Disponível em: <https://www.folha.uol.com.br/>.

STRUBE, 2019; YASUNAGA; KASAI; RADEV, 2018). Não obstante, não há consenso sobre quais tipos de representação devem ser combinados, portanto, é importante experimentar diferentes representações para verificar a melhor configuração (WANG *et al.*, 2021). Outro fator importante é que, em línguas de baixo recurso, o uso de representações vetoriais estáticas apresentam melhor resultado (HEINZERLING; STRUBE, 2019), porém, realizar o pré-treinamento de modelos de língua em texto do domínio alvo aparenta resolver este problema (VIRTANEN *et al.*, 2019; TSURUOKA *et al.*, 2005), sendo assim, ambas as abordagens devem ser consideradas quando possível, porém, o custo computacional pode ser muito alto, inviabilizando a experimentação de diversas combinações.

O contexto multigênero na etiquetagem morfossintática já foi estudado por outros autores em contextos diferentes, como o trabalho de (DOMINGUES, 2011), que apresenta um método de etiquetagem multigênero para textos acadêmicos e jornalístico para o português do Brasil. Contudo, o uso de gêneros que aderem à norma culta da língua (jornalístico e acadêmico) e CGU apresenta maior contraste nas diferenças de estilo de escrita, tornando a realização da etiquetagem automática mais desafiadora. Atualmente, existem trabalhos investigando a etiquetagem automática multigênero para outras línguas, como (BEHZAD; ZELDES, 2020), que apresenta um estudo voltado ao inglês, porém, não se sabe da existência de pesquisas para o português do Brasil no contexto apresentado.

CAPÍTULO

4

MATERIAIS E MÉTODOS

Este trabalho tem o objetivo de investigar métodos de etiquetagem morfossintática multigênero, considerando os gêneros textuais jornalístico, acadêmico e Conteúdo Gerado por Usuário (CGU), para o português do Brasil, utilizando o formalismo *Universal Dependencies* (UD).

Como apresentado no Capítulo 1, o processamento multigênero contribui para a criação de métodos robustos nos gêneros contemplados. Dessa forma, aplicações que possuem intersecção do gênero CGU, jornalístico ou acadêmico, como a detecção de notícias falsas (MONTEIRO *et al.*, 2018), podem se beneficiar de um único método com bom desempenho em todos os três gêneros ou em um subconjunto dos gêneros. Contudo, é importante ressaltar métodos multigênero podem ser aplicados em gêneros não contemplados e obtenham resultados satisfatórios. Não obstante, possuir as etiquetas gramaticais possibilita que as ambiguidades em nível morfossintático possam ser resolvidas, sendo assim, permitindo a criação de ferramentas e métodos de PLN mais robustos.

Os trabalhos iniciais na área de etiquetagem morfossintática apresentavam o córpus e um conjunto de etiquetas específico, sendo este conjunto dependente da língua de origem e formalismo adotado. Contudo, com o crescimento da área, diversos córpus para múltiplas línguas foram criados e o uso de diferentes conjuntos de etiquetas e formalismos torna a tarefa mais complexa. Tendo em vista a necessidade de padronização entre córpus, o projeto *Universal Dependencies* cria diretrizes universais e com possibilidade de extensão para etiquetagem morfossintática, morfológica e sintática. Atualmente, a UD é amplamente adotada pelos pesquisadores da área, havendo mais de 200 *treebanks*[16]; dessa forma, este trabalho adota o uso da UD, em especial, o conjunto de etiquetas morfossintáticas.

Atualmente, encontram-se grandes córpus não rotulados para o português do Brasil (SOUZA; NOGUEIRA; LOTUFO, 2020), possuindo mais de 100 milhões de *tokens*, porém,

[16] Disponível em: <https://universaldependencies.org/>.

existem poucos córpus rotulados, em especial, com etiquetas morfossintáticas. Dos córpus que adotam as diretrizes da UD, o português do Brasil conta com quatro córpus manualmente anotados [17] , sendo eles: Bosque[18] (RADEMAKER *et al.*, 2017), Porttinari-base (LOPES *et al.*, 2022), DANTEStocks[19] (Di FELIPPO *et al.*, 2021) e PetroGold (SOUZA *et al.*, 2021). Apesar da existência dos córpus mencionados, o português carece da investigação de métodos de etiquetagem multigênero, em especial, não se sabe da existência de métodos para a etiquetagem multigênero que contemple os gêneros CGU, jornalístico e acadêmico.

As hipóteses deste trabalho, previamente descritas no Capítulo 1 são:

1. É possível obter um método com capacidade multigênero mantendo o desempenho individual em cada gênero utilizado. Esta hipótese é apoiada por trabalhos na literatura que realizam aprendizado multigênero em outros gêneros textuais (MARCUS; SANTORINI; MARCINKIEWICZ, 1993b).

2. O uso do formalismo *Universal Dependencies* é suficiente para métodos de etiquetagem morfossintática alcançarem resultados do estado da arte. Esta conjectura é suportada por outros trabalhos que alcançaram os melhores desempenhos baseados na UD (STRAKA, 2018; KONDRATYUK; STRAKA, 2019).

As próximas seções deste capítulo descrevem os córpus e métodos utilizados e, por fim, descrevem a metodologia de experimentação para validação das hipóteses apresentadas.

4.1 Córpus

Conforme apresentado no Capítulo 1, o primeiro passo para realizar a etiquetagem é acessar os documentos, ou córpus. Este trabalho aborda os gêneros textuais jornalístico, acadêmico e CGU para o português do Brasil, portanto, faz-se necessário possuir um ou mais córpus essas características.

O córpus utilizado para o gênero textual CGU é o DANTEStocks (Di FELIPPO *et al.*, 2021)[20], contendo sentenças extraídas da rede social Twitter por (SILVA; ROMAN; CARVALHO, 2020) e, posteriormente, realizada a inclusão dos rótulos morfossintáticos e segmentação de *tokens*. O DANTEStocks possui 4.517 *tweets* que mencionam índices presentes na bolsa de

[17] Os córpus CINTIL (BRANCO *et al.*, 2022), PUD (ZEMAN *et al.*, 2017) e GSD (MCDONALD *et al.*, 2013) também estão disponíveis no formalismo UD, contudo, estes trabalhos adotaram diretrizes de anotação distintas das adotadas pelos córpus selecionados, dessa forma, estes córpus não foram selecionados para experimentação.

[18] O Apêndice A apresenta a análise inicial dos resultados do etiquetador BERTimbau nos córpus Bosque e DANTEStocks

[19] O Apêndice B apresenta a descrição das diretrizes iniciais de anotação do córpus DANTEStocks

[20] Foi utilizada a versão 16 de novembro de 2022.

valores B3, totalizando 81,048 *tokens*. Não obstante, o córpus foi rotulado com o formalismo da UD.

Baseado na extração de textos jornalísticos, o córpus de SANTANA (2019) foi rotulado com etiquetas morfológicas e morfossintáticas, dando origem ao córpus Porttinari-base (LOPES *et al.*, 2022)[21]. Este trabalho adota as diretrizes de rotulação do projeto UD. Ao total, existem 8,420 sentenças (168,400 *tokens*) extraídas do jornal Folha de São Paulo[22], correspondente ao período de janeiro de 2015 até setembro de 2017.

Contemplando o gênero acadêmico, o córpus PetroGold (SOUZA *et al.*, 2021)[23] apresenta uma coletânea de textos da área de óleo e gás rotulados seguindo o formalismo UD. Ao total, o córpus possui 250,605 *tokens* e 4,048 sentenças rotuladas com etiquetas morfossintáticas, atributos morfológicos e relações sintáticas, entre outros. Os textos contidos neste córpus provêm de uma variedade de teses, dissertações e monografias.

A Tabela 3 sumariza a quantidade e sentenças para cada córpus apresentado. É possível observar que o córpus DANTEStocks tem uma quantidade menor de *tokens* quando comparado aos córpus Porttinari-base e PetroGold, consequentemente, pode ser um fator que deteriore o aprendizado multigênero, a depender de outros fatores experimentais. Não obstante, os córpus DANTEStocks e Porttinari-base originalmente não possuem a divisão em conjunto de treino, validação e teste; dessa forma, para fins de avaliação e comparação entre métodos, foi realizada a divisão com a amostragem aleatória utilizando a proporção de 10% para validação e 20% para o conjunto de teste.

Tabela 3 – Descrição dos córpus utilizados contendo o gênero textual e quantidade de sentenças

Córpus	Gênero	Treino	Validação	Teste	Total de sentenças	Total de *tokens*
DANTEStocks	CGU	2,833	413	802	4,048	81,048
Porttinari-base	Jornalístico	5,894	585	1,668	8,420	168,400
PetroGold	Acadêmico	8,054	447	445	8,946	250,905

Fonte: Elaborada pelo autor.

Outro fator importante a ser analisado é o número de *tokens* por sentença e, em específico ao córpus DANTEStocks, o número de *tokens* por *tweets*. A Tabela 4 sumariza esta quantidade para cada córpus considerado. É possível visualizar que os córpus DANTEStocks e Porttinari-base possuem médias e desvios padrões similares, contudo, a comparação direta não é possível, pois o primeiro córpus segmenta por *tweets* e o segundo é segmentado por sentença; consequentemente, um *tweet* pode conter mais de uma sentença. Não obstante, o córpus PetroGold apresenta média e desvio padrão maiores. Isso dá-se por textos acadêmicos conterem sentenças

[21] Foi utilizada a versão 26 de setembro de 2022.
[22] Disponível em: <https://www.folha.uol.com.br/>.
[23] Foi utilizada a versão disponível no dia 27 de novembro de 2022.

longas e muito curtas como, por exemplo, em uma lista enumerada onde cada elemento da lista é segmentado como um item da lista.

Tabela 4 – Quantidade média e desvio padrão de *tokens* por sentença/*tweet* nos córpus selecionados

Córpus	Treino	Validação	Teste	Total
DANTEStocks	20.09 ± 7.74	20.24 ± 7.64	19.65 ± 8.22	20.02 ± 7.83
Porttinari-base	19.97 ± 7.74	19.97 ± 8.02	20.11 ± 8.04	20.00 ± 7.77
PetroGold	28.02 ± 17.78	28.56 ± 17.37	27.36 ± 17.61	28.01 ± 17.74

Fonte: Elaborada pelo autor.

O Quadro 5 apresenta exemplos de sentenças e *tweets* presentes nos córpus Porttinari-base, DANTEStocks e PetroGold. Os exemplos do córpus Porttinari-base apresentam sentenças objetivas e diretas e seguem a norma culta da língua É possível observar que os *tweets* utilizam mais numerais, símbolos, terminologias do mercado de ações, abreviações e índices da bolsa de valores. Por fim, o córpus PetroGold apresenta um exemplo de texto acadêmico, com uma linguagem mais técnica e especializada, caracterizada pelo uso de terminologias específicas da área de óleo e gás, além de usar elementos comuns em textos acadêmicos, como o uso de referências bibliográficas.

Quadro 5 – Exemplos de sentenças dos córpus Porttinari-base, DANTEStocks e PetroGold

Córpus	Exemplo
Porttinari-base	A população não poderia ter acesso a relatórios que explicassem, por exemplo, os motivos exatos de atrasos em obras de linhas e estações.
	O país onde estou protegido é aqui, afirmou ele, negando que tivesse a intenção de fugir de o país.
DANTEStocks	@Live_Trade assim que vc tiver um tempo comenta alll3. valew
	INTRADAY PETR4: Suportes 13,08 e 13,25 e resistências 13,66 e 13,90 INTRADAY VALE5: Suportes 27,78 e 27,94 e resistências 28,35 e 28,60
PetroGold	Estes produzem distorções locais e nos elementos do campo magnético da Terra.
	O Escudo Sul-Rio-Grandense localiza-se na porção meridional da Província Mantiqueira (Almeida et al. 1981, Hasui et al. 1985), englobando o Orógeno Dom Feliciano, corresponde à área do Estado do Rio Grande do Sul que é marcada pela ocorrência de rochas ígneas, metamórficas e sedimentares pré-paleozóicas, cuja origem é relacionada aos ciclos Transamazônicos (Paleoproterozóico) e Brasiliano/Pan-Africano (Neoproterozóico).

Fonte – Elaborada pelo autor.

4.2 Modelos

Visando construir um etiquetador morfossintático de alto desempenho para o português do Brasil, além da seleção de córpus, também é necessário selecionar os métodos a serem avaliados. Atualmente na literatura encontram-se diversas abordagens para etiquetadores com

resultados do estado da arte. Neste contexto, foram selecionados diferentes métodos disponíveis na literatura que apresentam alto desempenho em diversas línguas e gêneros, visando realizar experimentos para encontrar a abordagem que possui a maior acurácia no português do Brasil nos córpus selecionados.

A seleção de métodos foi baseada no Capítulo 3, onde foram encontrados métodos categorizados em dois grupos: (1) baseado em modelo de língua e (2) baseado em Rede Neural Recorrente (RNR). É importante ressaltar que ambas as categorias utilizam representação computacional contextual, isto é, não são puramente baseadas em representações insensíveis ao contexto da sentença apresentada. A seguir são apresentados os métodos separados por suas respectivas categorias.

4.2.1 Métodos baseados em Modelos de Língua

Conforme relatado no Capítulo 3, foram encontrados na literatura métodos baseados em modelo de língua pré-treinados. Estes modelos são previamente treinados em grandes quantidades de dados para uma ou múltiplas línguas por meio do aprendizado auto-supervisionado. Este método de aprendizado utiliza córpus sem rótulos e cria rótulos automaticamente para o treinamento como, por exemplo, LIU *et al.* (2019) realizam a tarefa de Modelagem de Língua Mascarada (MLM), isto é, o modelo recebe um texto de entrada e alguns dos *tokens* são aleatoriamente substituídos por um *token* especial que indica a omissão daqueles *tokens*; então, o modelo deve encontrar por meio da classificação quais são os *tokens* ocultos. Dessa forma, o modelo aprende a criar representações robustas dos textos de entrada, dado que a tarefa força-o a realizar predições sem conhecer todos os *tokens* de entrada. Alguns trabalhos na literatura que apresentam alto desempenho na tarefa de etiquetagem morfossintática, em textos jornalístico, acadêmico e CGU, utilizam modelos de língua pré-treinados em conjunto da técnica de ajuste fino (VRIES *et al.*, 2019; VIRTANEN *et al.*, 2019; NGUYEN; VU; NGUYEN, 2020). Esta técnica adiciona uma camada de rede neural totalmente conectada à arquitetura do modelo e por meio de uma nova etapa de treinamento, o modelo se especializa na tarefa alvo. A seguir são apresentados os Modelos de Língua selecionados para realização do ajuste fino na tarefa de etiquetagem morfossintática.

XLM-R (CONNEAU *et al.*, 2019) é um modelo de língua multilíngue baseado no modelo RoBERTa (LIU *et al.*, 2019). Os autores treinaram o modelo para mais de 100 línguas, incluindo o português do Brasil, contendo aproximadamente 49 GB de dados em português extraídos da web, consequentemente, incluindo diversos gêneros textuais. O modelo apresenta melhorias consistentes em relação ao modelo BERT multilíngue (DEVLIN *et al.*, 2019) nas tarefas de classificação sequencial, como, por exemplo, o reconhecimento de entidades nomeadas.

DeBERTa-v3 (HE; GAO; CHEN, 2021) é uma variante do modelo de linguagem BERT (DEVLIN *et al.*, 2019), treinado em uma ampla gama de línguas, recorrendo ao mesmo córpus utilizado no XLM-R. Os autores realizam algumas alterações como a modificação da tarefa

de pré-treinamento, alteração do mecanismo de codificação de posição relativa e aumentam a quantidade de parâmetros treináveis. A combinação destes fatores resultou em um modelo consistentemente melhor que o XLM-R, em média um aumento de $3,6\%$ no córpus de avaliação multilíngue XNLI (CONNEAU *et al.*, 2018).

Em contraposição aos modelos apresentados anteriormente, o BERTimbau (SOUZA; NOGUEIRA; LOTUFO, 2020) é um modelo específico para o português do Brasil. Diversos trabalhos na literatura apresentam evidências de que modelos especializados em uma única língua (VRIES *et al.*, 2019; VIRTANEN *et al.*, 2019; SOUZA; NOGUEIRA; LOTUFO, 2020) possuem desempenho superior aos modelos multilíngues em uma variedade de tarefas. Os autores utilizam a arquitetura BERT (DEVLIN *et al.*, 2019) sem modificações e, como córpus de treinamento, utilizam a coletânea de textos brWaC (FILHO *et al.*, 2018) que possui cerca de $2,68$ bilhões de *tokens* extraídos de $3,53$ milhões de documentos.

Os três modelos de língua citados possuem diferentes versões que se distinguem na quantidade de parâmetros treináveis. Os modelos possuem uma versão com menor quantidade de parâmetros denominada *base* com 110 milhões de parâmetros e uma versão maior denominada *large* com 340 milhões de parâmetros[24]. Geralmente, as versões *large* possuem resultados melhores do que a versão *base*, contudo, possuem um custo computacional maior e, consequentemente, dificultam o processo de experimentação. Optou-se por utilizar a versão *base* dos três modelos para obter agilidade na experimentação e por restrições de hardware, dado que nem todos os modelos na versão *large* couberam nas GPUs disponíveis para treinamento.

Por fim, é importante ressaltar que os modelos mencionados não são etiquetadores morfossintáticos em seus trabalhos originais, dessa forma, a técnica de ajuste fino será aplicada para que os modelos aprendam a classificar as etiquetas morfossintáticas automaticamente. Não obstante, os modelos foram selecionados devido a evidências na literatura que apresentam o alto desempenho de modelos de língua pré-treinados na tarefa de etiquetagem morfossintática por meio da técnica de ajuste fino (VRIES *et al.*, 2019; VIRTANEN *et al.*, 2019; NGUYEN; VU; NGUYEN, 2020).

4.2.2 Métodos baseados em Redes Neurais Recorrentes

Os métodos apresentados a seguir utilizam a arquitetura de Redes Neurais Recorrentes (RNRs), ou seja, são baseados em redes neurais com a característica de criar representações contextuais por meio de estados ocultos e conexões recorrentes. Existem diversas arquiteturas de RNRs disponíveis na literatura e as diferenças serão destacadas para cada método selecionado. Diferentemente da seção anterior, nesta encontram-se exclusivamente métodos monolíngue, não obstante, alguns métodos são avaliados no contexto multilíngue. Contudo, os autores realizaram o treinamento no contexto monolíngue e reportaram os resultados para cada língua avaliada,

[24] O modelo DeBERTa-v3 possui 276 milhões de parâmetros na versão *base* e 435 milhões na versão *large*.

dessa forma, não caracterizam o método como multilíngue. Além disso, todos os métodos são específicos para a etiquetagem morfossintática, diferentemente dos modelos de língua apresentados anteriormente.

O primeiro trabalho selecionado provém do projeto Stanza[25] (QI *et al.*, 2020), o qual tem um conjunto de ferramentas de PLN para diversas línguas, que permitem realizar o treinamento para córpus e línguas diferentes. Para a etiquetagem morfossintática, os autores utilizam uma RNR com conexões bidirecionais denominada *Bidirectional Long Short-Term Memory* (Bi-LSTM) em combinação com o mecanismo *Biaffine Attention* (DOZAT; MANNING, 2016).

Baseando-se na arquitetura Bi-LSTM, BOHNET *et al.* (2018) utilizam três modelos baseados nesta arquitetura para a tarefa, onde o primeiro utiliza representação em nível de caractere, a segunda utiliza a representação em nível de sentença e, por último, o terceiro modelo utiliza a combinação das representações anteriores para realizar a predição das classes gramaticais.

Semelhantemente, o trabalho CNCSR (HEINZERLING; STRUBE, 2019) realiza a avaliação de combinações de diferentes tipos de representações, como o uso de representações estáticas e contextuais. Não obstante, os autores também apresentam a avaliação da combinação de diferentes representações com Modelos de Língua e concluem que o uso das representações em nível de caractere e *token* apresentam melhores resultados para línguas de baixo recurso. Para fins experimentais, a configuração utilizada foi a que apresentou os melhores resultados na avaliação para o português do Brasil, sendo esta a combinação de representações em nível de sub-palavra e de caractere.

O modelo UDPipe 2 (STRAKA, 2018) utiliza uma combinação de representações contextuais para realizar a tarefa de etiquetagem morfossintática, sendo estas representações em nível de caractere e *token* e extraídas de modelos de língua. Apesar do trabalho utilizar representações de modelos de língua, o modelo ainda utiliza uma Bi-LSTM para combinar as três representações mencionadas e baseia-se em mais de um tipo de representação contextual. Dessa forma, optou-se por apresentar este modelo nesta seção, contudo, seria possível categorizá-lo como um modelo híbrido entre RNRs e modelos de língua.

A seleção de métodos descrita apresentou resultados competitivos e, em alguns casos, do estado-da-arte, a depender do córpus de avaliação. Todos os métodos são baseados em Bi-LSTMs, porém, com variações do número de redes, representações computacionais de entrada e variações na arquitetura. Dessa forma, a seleção apresenta os melhores métodos da categoria encontrados na literatura e com diversidade nas abordagens, desta forma, podendo-se realizar uma avaliação robusta na tarefa de etiquetagem morfossintática.

[25] Disponível em: <https://stanfordnlp.github.io/stanza/>.

4.3 Metodologia

Esta seção descreve a metodologia utilizada para avaliação da etiquetagem morfossintática para o português do Brasil e da avaliação multigênero no mesmo contexto. As avaliações experimentais utilizam os córpus e modelos apresentados nas seções anteriores deste capítulo. A seguir, é apresentada a metodologia utilizada para comparação dos métodos selecionados para a etiquetagem morfossintática em córpus do português do Brasil.

4.3.1 Avaliação de etiquetadores morfossintáticos em córpus jornalístico

Inicialmente, as abordagens selecionadas são avaliadas no córpus Porttinari-base, visando identificar o método que possui o melhor desempenho neste conjunto de dados para, posteriormente, realizar a avaliação no contexto multigênero. O córpus de gênero jornalístico foi selecionado para esta análise tendo em vista os quesitos de disponibilidade, custo computacional para execução dos experimentos multigênero e frequência de uso em trabalhos da área. No momento da execução destes experimentos, apenas o córpus Porttinari-base tinha estava disponível, onde os demais córpus estavam em desenvolvimento. Além disso, os métodos de etiquetagem disponíveis para o português do Brasil são, em sua grande maioria, baseados em texto do gênero jornalístico, dessa forma, a análise se baseia em um gênero amplamente estudando na área. Além disso, realizar esta análise em um gênero e posteriormente realizar o experimento multigênero reduz o custo computacional total, pois, apenas um método será avaliado no contexto multigênero. Exemplificando, ao considerar 10 execuções para os 7 modelos e avaliando-os nos 3 córpus apresentados, calcula-se o total de 210 experimentos[26].

O procedimento experimental para a avaliação de cada modelo dá-se pela realização de 10 execuções de treinamento no conjunto de treino do córpus Porttinari-base, para então, realizar a comparação entre os modelos e realização de testes de hipótese para identificar diferenças estatisticamente significativas. O teste Anova (FISHER, 1992) com *post hoc* de Tukey (TUKEY, 1949) foi selecionado para realizar esta avaliação. O teste Anova avalia se existem diferenças significativas entre as médias de dois ou mais grupos, então, se identificada a diferença significativa entre os grupos, o teste de Tukey é aplicado para identificar quais os grupos que possuem médias significativamente distintas entre si.

Cada abordagem possui o seu conjunto de hiper-parâmetros, podendo estes serem específicos para a arquitetura do modelo, como, por exemplo, o tamanho da dimensão de estado oculto para aqueles baseados em RNRs e número de cabeças para modelos de língua. A variação de hiper-parâmetros pode impactar no desempenho, dessa forma, optou-se por utilizar os valores reportados pelos autores que obtiveram a maior acurácia em córpus do português do Brasil, não

[26] O tempo de execução depende do modelo utilizado, sendo o menor, em torno de 20 minutos e o maior, aproximadamente 12 horas.

obstante, quando o trabalho não utilizou córpus em português, foram utilizados os valores que obtiveram o melhor resultado geral, ou seja, sem distinguir a língua. O fator alterado entre as execuções foi a semente aleatória, que é utilizada internamente pelas bibliotecas de treinamento para inicialização dos parâmetros de redes neurais e outros processos que necessitem de geração de números aleatórios, dessa forma, execuções com sementes aleatórias diferentes irão gerar modelos com desempenhos diferentes.

A partir da avaliação descrita, será identificado o modelo que possuir o melhor desempenho no contexto apresentado e, dessa forma, possibilitar a verificação da segunda hipótese deste trabalho, sendo ela a de que o uso do formalismo *Universal Dependencies* é suficiente para atingir resultados do estado da arte na tarefa de etiquetagem morfossintática para o português do Brasil. A próxima subseção descreve a metodologia utilizada para avaliação do modelo no contexto multigênero.

4.3.2 Avaliação do aprendizado multigênero

Um dos objetivos deste trabalho é avaliar a capacidade de aprendizado multigênero para o português do Brasil, considerando os gêneros jornalístico, acadêmico e CGU. Dessa forma, a metodologia adotada inclui o uso do modelo que obteve o melhor desempenho na etapa anterior. Em sequência, a avaliação é realizada a partir de diferentes cenários de treinamento. Para avaliar o aprendizado multigênero, é necessário avaliar o modelo isoladamente em cada córpus e, também, utilizando múltiplos córpus com gêneros diferentes para avaliação. Cada cenário consiste no treinamento de um ou mais córpus e avaliação em todos os córpus, por exemplo, o modelo será treinado no córpus jornalístico Porttinari-base e avaliado nos três córpus disponíveis. Da mesma forma, considerando o cenário multigênero, o modelo será treinado em dois ou três córpus e avaliado da mesma forma que o cenário anterior. Naturalmente, existirão sete cenários diferentes, sendo eles, treinado no córpus (1) Porttinari-base, (2) DANTEStocks, (3) PetroGold, (4) Porttinari-base e DANTEStocks, (5) Porttinari-base e PetroGold, (6) DANTEStocks e PetroGold e, por fim, (7) Porttinari-base, DANTEStocks e PetroGold.

Com os cenários experimentais definidos, o modelo é treinado em cada um dos cenários e avaliado em todos os córpus. Semelhantemente a etapa anterior, cada cenário é executado 10 vezes, variando a semente aleatória e utilizando os hiper-parâmetros da fase anterior. Em seguida, é realizado o teste estatístico ANOVA (FISHER, 1992) com *post hoc* de Tukey (TUKEY, 1949) para verificação de diferenças estatisticamente significativas. Dessa forma, se houver diferença significativa entre o modelo multigênero treinado em todos os córpus em relação aos cenários individuais, será averiguado que o aprendizado multigênero deteriorou o aprendizado isolado, da mesma forma. Caso não haja diferença encontrada, o modelo demonstrará capacidade de aprendizado multigênero. A partir da metologia apresentada, será possível verificar a primeira hipótese apresentada neste trabalho, sendo ela a de que é possível obter um método com capacidade multigênero mantendo o desempenho individual em cada gênero utilizado para o

português do Brasil.

O próximo capítulo apresenta os resultados e análises realizadas a partir da seleção de córpus, modelos e metodologia apresentados neste capítulo.

CAPÍTULO

5

RESULTADOS E DISCUSSÃO

Este capítulo apresenta os resultados e discussões relacionadas aos experimentos descritos no Capítulo 4. Inicialmente, são apresentados os resultados da avaliação de etiquetadores morfossintáticos para o português do Brasil considerando o gênero jornalístico no córpus Porttinari; em seguida, são apresentados os resultados da avaliação multigênero. Por fim, é apresentada uma análise manual dos erros encontrados na avaliação dos córpus Porttinari-base, PetroGold e DANTEStocks, buscando avaliar hipóteses de erros sistemáticos.

5.1 Avaliação de métodos para etiquetagem morfossintática no português do Brasil para o gênero jornalístico

Cada modelo utilizou um conjunto diferente de hiper-parâmetros para a etapa de treinamento, dessa forma, os mesmos são descritos abaixo, além da indicação do código-fonte utilizado.

O modelo UDPipe 2 (STRAKA, 2018) foi treinado a partir do código disponível no repositório oficial do projeto[27], onde nenhuma alteração foi realizada. O tamanho de lotes (*batch size*) de 128 amostras, sendo realizado o treinamento com um total de 16 épocas, onde, nas primeiras 8 épocas, é utilizada a taxa de aprendizagem de $1e-3$, e de $1e-4$ nas 8 épocas restantes. Além disso, o modelo utiliza representações de modelos de língua extraídas a partir do modelo BERTimbau (SOUZA; NOGUEIRA; LOTUFO, 2020).

Para a realização dos experimentos com o modelo Stanza (QI *et al.*, 2020), foi utilizado o código disponível no repositório oficial da ferramenta[28], onde não foram realizadas alterações. Foi utilizado o tamanho em lotes padrão de 5.000, taxa de aprendizagem de $1e-3$ e número

[27] Disponível em: <https://github.com/ufal/udpipe/tree/udpipe-2>.
[28] Disponível em: <https://github.com/stanfordnlp/stanza>.

máximo de atualizações de etapas de gradiente de 1.000, onde cada etapa é contabilizada quando um lote (*batch*) passa pela rede e os parâmetros são atualizados.

Meta-BiLSTM (BOHNET *et al.*, 2018) foi treinado utilizando o repositório oficial[29], sem alterações no código-fonte principal, apenas com ajustes no arquivo *config.json* para alteração dos hiper-parâmetros. Foi utilizado o tamanho de lotes de 40.000 para o modelo em nível de palavras e 80.000 para o modelo em nível de caracteres. Além disso, a taxa de aprendizagem é de $2e-3$. Os modelos em nível de caractere, palavra e meta (que combina os modelos anteriores) utilizam camada oculta com 400 neurônios. Além disso, os dois primeiros modelos possuem 3 camadas ocultas e o modelo meta possui uma. Não obstante, foi utilizado o critério de número mínimo de 50 etapas para interromper o treinamento, dessa forma, se o treinamento não melhorar durante 50 etapas seguidas, o processo será interrompido. Além disso, o modelo utiliza representações estáticas em nível de palavra e, para isso, foram utilizadas os vetores de palavras de HARTMANN *et al.* (2017), baseados no modelo Skip-gram (MIKOLOV *et al.*, 2013b), com tamanho 300.

Os experimentos com o modelo CNCSR (HEINZERLING; STRUBE, 2019) foram realizados a partir do código fonte oficial[30], onde foi realizada uma alteração para permitir a gravação em disco das predições do modelo quando realizada a avaliação no conjunto de dados de teste[31]. O modelo foi treinado com tamanho de lotes de 64, número de épocas mínimo de 50 e máximo de 1000, taxa de aprendizagem de $1e-4$, tamanho de vocabulário 100.000 e taxa de *dropout* de $0,2$. Além disso, foram utilizadas as representações em nível de caractere e sub-palavra, sendo elas combinadas por meio de uma rede RNR meta. O modelo em nível de caractere possui representação vetorial de tamanho 50 e camada oculta com 256 neurônios; da mesma forma, os modelos de sub-palavra e meta possuem o mesmo número de neurônios na camada oculta.

Para realizar os experimentos com os modelos de língua BERTimbau, DeBERTa-v3 e XLM-R, foi desenvolvido o código de treinamento com base na biblioteca *transformers*[32]. A biblioteca possui os modelos pré-treinados e permite carregá-los para realizar o ajuste fino na tarefa de interesse do usuário. Para os três modelos, foram utilizados os seguintes hiper-parâmetros: máximo de 30 épocas, taxa de aprendizagem de $2e-5$ e *weight decay rate* de $0,01$, que é um parâmetro do otimizador AdamW (LOSHCHILOV; HUTTER, 2017). Não obstante, os modelos BERTimbau e XLM-R utilizaram tamanho de lotes de 32 e, para o DeBERTa-v3, foi utilizado tamanho 16.

A Tabela 5 sumariza os hiper-parâmetros utilizados para cada experimento. É importante ressaltar que apenas os hiper-parâmetros que foram alterados em relação à implementação original (artigo do(s) autor(es), ou valor padrão da biblioteca), foram apresentados nesta seção.

[29] Disponível em: <https://github.com/google/meta_tagger>.
[30] Disponível em: <https://github.com/bheinzerling/subword-sequence-tagging>.
[31] A alteração está disponível em <https://github.com/huberemanuel/subword-sequence-tagging>
[32] Disponível em: <https://github.com/huggingface/transformers>.

Tabela 5 – Conjunto de hiper-parâmetros utilizados no córpus Porttinari-base

Modelo	*Batch Size*	Épocas	Taxa de aprendizado	Outros
UDPipe 2	128	8;8	1e-3;1e-4	Embeddings de modelo de língua: Bertimbau
Stanza	5000	-	1e-3	Máximo de passos: 1000
Meta-BiLSTM	40000;80000	-	2E-3	hidden_char_size: 400 hidden_word_size: 400 hidden_meta_size: 400 num_layers_chars: 3 num_layers_words: 3 num_layers_meta: 1 early_stopping_steps: 50 embeddings estáticas: skip_s300
CNCSR	64	50 a 1000	1e-4	Representações: caractere, sub-palavra Modelo: Meta-RNR Melhor tamanho de vocabulário: Verdadeiro char-embd-dim: 50 char-nhidden: 256 bpe-nhidden: 256 meta-nhidden: 256 dropout: 0,2
XLM-R	32	30	2e-5	weight_decay: 0,01
DeBERTa-v3	16	30	2-5	weight_decay: 0,01
BERTimbau	32	30	2E-5	weight_decay: 0,01

Fonte: Elaborada pelo autor.

Tabela 6 – Acurácia em nível de *tokens* no conjunto de testes do córpus Porttinari-base

Modelo	Abordagem	Acurácia média (%)	Medida-F macro média (%)
BERTimbau	Modelo de língua	**99,0710 ± 0,0305**	**96,3924 ± 0,3186**
DeBERTa-v3	Modelo de língua	$99,0157 \pm 0,0452$	$95,8095 \pm 0,3890$
XLM-R	Modelo de língua	$98,9979 \pm 0,0426$	$96,3589 \pm 0,4201$
Meta-BiLSTM	RNR	$98,4726 \pm 0,0609$	$94,8859 \pm 0,2779$
Udpipe 2	RNR	$98,0112 \pm 0,0344$	$93,1330 \pm 0,5435$
Stanza	RNR	$98,2172 \pm 0,0538$	$94,5968 \pm 0,2696$
CNCSR	RNR	$98,0953 \pm 0,0674$	$94,0351 \pm 0,2987$

Fonte: Elaborada pelo autor.

A Tabela 6 apresenta os resultados da avaliação da etiquetagem morfossintática em córpus do gênero jornalístico. São apresentadas as acurácias em nível de *tokens* médias e Medida-F Macro médias das 10 execuções de experimentos para cada abordagem avaliada, além dos respectivos desvios padrões. É possível observar que os métodos baseados em Redes Neurais Recorrentes (RNRs) possuem desempenho inferior aos métodos baseados em modelos de língua com ajuste fino, tanto em termos de acurácia e Medida-F macro. É possível observar que a abordagem BERTimbau possui o maior valor absoluto médio para acurácia e Medida-F Macro, não obstante, os modelos DeBERTa-v3 e XLM-R possuem valores similares.

Ao realizar o teste estatístico ANOVA na medida de acurácia, observou-se a estatística $Z = 890,0529$ e p-valor de $p = 6,7058 - 59$, dessa forma, pode-se rejeitar a hipótese H_0 de que

Tabela 7 – P-valores resultantes da análise *post hoc* de Tukey no córpus Porttinari-base

	BERTimbau	DeBERTa-v3	XLM-R	Meta-BiLSTM	UDPipe 2	Stanza	CNCSR
BERTimbau	-	0,1801	0,0249	0,0010	0,0010	0,0010	0,0010
DeBERTa-v3	0,1801	-	0,9000	0,0010	0,0010	0,0010	0,0010
XLM-R	0,0249	0,9000	-	0,0010	0,0010	0,0010	0,0010
Meta-BiLSTM	0,0010	0,0010	0,0010	-	0,0010	0,0010	0,0010
UDPipe 2	0,0010	0,0010	0,0010	0,0010	-	0,0010	0,0048
Stanza	0,0010	0,0010	0,0010	0,0010	0,0010	-	0,0010
CNCSR	0,0010	0,0010	0,0010	0,0010	0,0048	0,0010	-

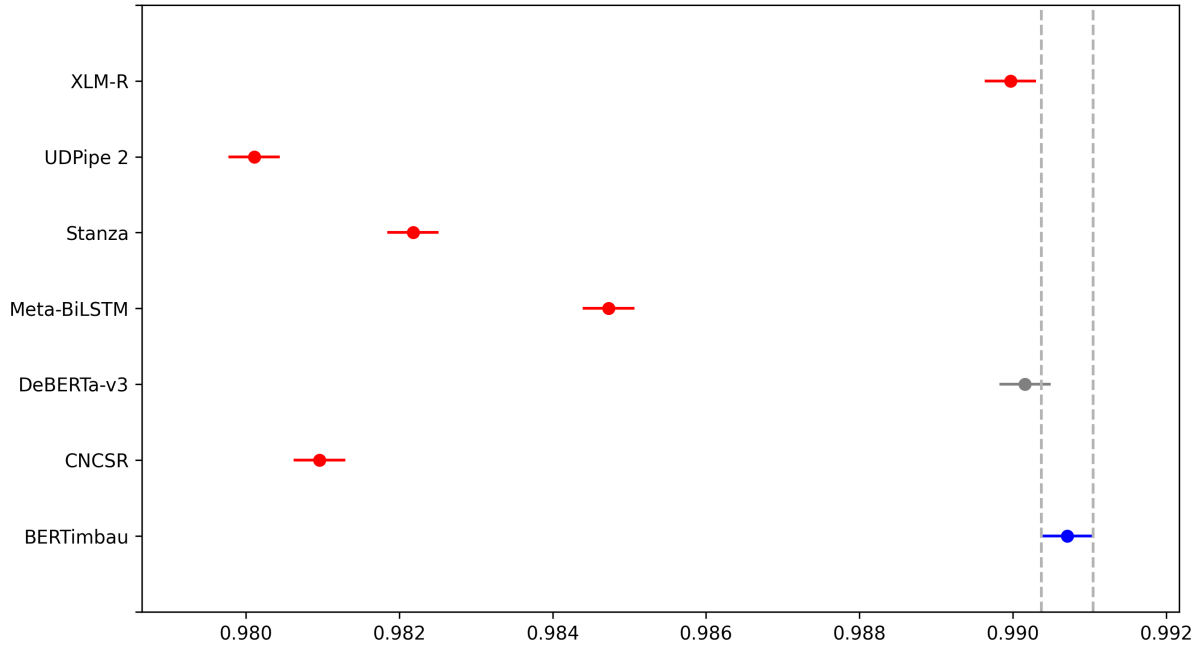
Fonte: Elaborada pelo autor.

as médias das populações sãos iguais com confiança de 95% ($\alpha = 0.05$). Em seguida, o teste de Tukey foi aplicado para identificar as diferenças ao comparar modelo a modelo. A Tabela 7 apresenta a matriz de p-valores ao executar o teste, onde são identificados pela cor vermelha os casos onde foi possível rejeitar a hipótese H_0 de que os dois modelos possuem a mesma média. Por exemplo, ao analisarmos a tupla Meta-BiLSTM e DeBERTa-v3, encontra-se o p-valor de $0,0010$ em vermelho, ou seja, existe diferença significativa com $\alpha = 0.05$. Semelhantemente, ao analisarmos o modelo DeBERTa-v3 e XLM-R, observa-se o p-valor de $0,9000$, neste caso, o teste falha em rejeitar a hipótese de que as médias são iguais. É possível observar que os modelos que não apresentaram médias suficientemente diferentes são os modelos BERTimbau em relação ao DeBERTa-v3 e o modelo XLM-R em relação ao DeBERTa-v3.

A Figura 14 apresenta os intervalos universais de confiança com destaque da cor azul para o modelo BERTimbau, que possui o maior valor absoluto de acurácia e Medida-F no experimento realizado. A linha vertical tracejada em cinza indica o intervalo de confiança com $\alpha = 0.05$, isto é, se houver intersecção com o intervalo de confiança de outra abordagem, indicará que o teste não conseguiu identificar diferenças significativas nas médias, onde estes casos são identificados pela cor cinza na figura. Dessa forma, pode-se observar que não foi possível identificar diferença significativa entre o modelo BERTimbau e DeBERTa-v3, porém, foi possível identificar diferença entre BERTimbau e todos os outros modelos. Conforme apresentado na Tabela 7, os modelos BERTimbau e XLM-R apresentaram p-valor de 0.0214, que é abaixo do $\alpha = 0.05$, contudo, é possível observar que ele está bem mais próximo do intervalo de confiança do BERTimbau em relação às abordagens baseadas em RNR.

Outro fator importante para a análise dos métodos é a comparação de desempenho em sentenças que contenham palavras fora de vocabulário (OOV). É conhecida a dificuldade dos métodos de PLN em obter alto desempenho quando existem *tokens* que não estavam no conjunto de treinamento. Dessa forma, a Tabela 8 apresenta as acurácias em nível de *tokens* segmentada por palavras contidas ou não (OOV) no conjunto de treinamento do córpus Porttinari-base. É possível observar que o método BERTimbau continua com os melhores resultados absolutos em ambas as acurácias. Além disso, os métodos baseados em RNRs possuem uma queda média de $5,35\%$ na acurácia de OOVs, em contraposição, as abordagens que utilizam modelos de língua

Figura 14 – Intervalo de confiança da acurácia em nível de *tokens* média de cada modelo no córpus Porttinari-base.



Fonte: Elaborada pelo autor.

Tabela 8 – Acurácias em nível de *tokens* segmentada por palavras fora (OOVs) e contidas no vocabulário

Modelo	Acurácia em OOVs	Acurácia em tokens do vocabulário
BERTimbau	**96.4264 ± 0.1555**	**99.2990 ± 0.0269**
CNCSR	92.5113 ± 0.2254	98.5769 ± 0.0652
DeBERTa-v3	96.3626 ± 0.2603	99.2446 ± 0.0409
Meta-BiLSTM	94.5308 ± 0.1965	98.8127 ± 0.0721
Stanza	94.7973 ± 0.2057	98.5122 ± 0.0492
UDPipe	91.2462 ± 0.1287	98.5947 ± 0.0291
XLM-R	96.2575 ± 0.1211	99.2342 ± 0.0421

Fonte: Elaborada pelo autor.

possuem uma queda média de $2,91\%$, ou seja, os modelos de língua apresentam uma maior robustez para processar OOVs.

Além da acurácia em nível de *tokens*, a acurácia em nível de sentenças é apresentada na Tabela 9. Esta métrica apresenta a quantidade de sentenças etiquetadas de forma totalmente correta, dessa forma, a métrica é mais restritiva e útil aos usuários de etiquetadores morfossintáticos (MANNING, 2011). Nesta análise, é possível observar uma discrepância maior em relação aos modelos baseados em modelos de língua e RNRs, onde a diferença entre as acurácias dos melhores modelos de cada categoria é de $8,72\%$. Novamente o modelo BERTimbau possui o maior valor absoluto de acurácia, contudo, ao realizar testes de hipótese ANOVA com *post hoc* de Tukey, é possível observar que não houve diferença significativa entre os modelos BERTimbau,

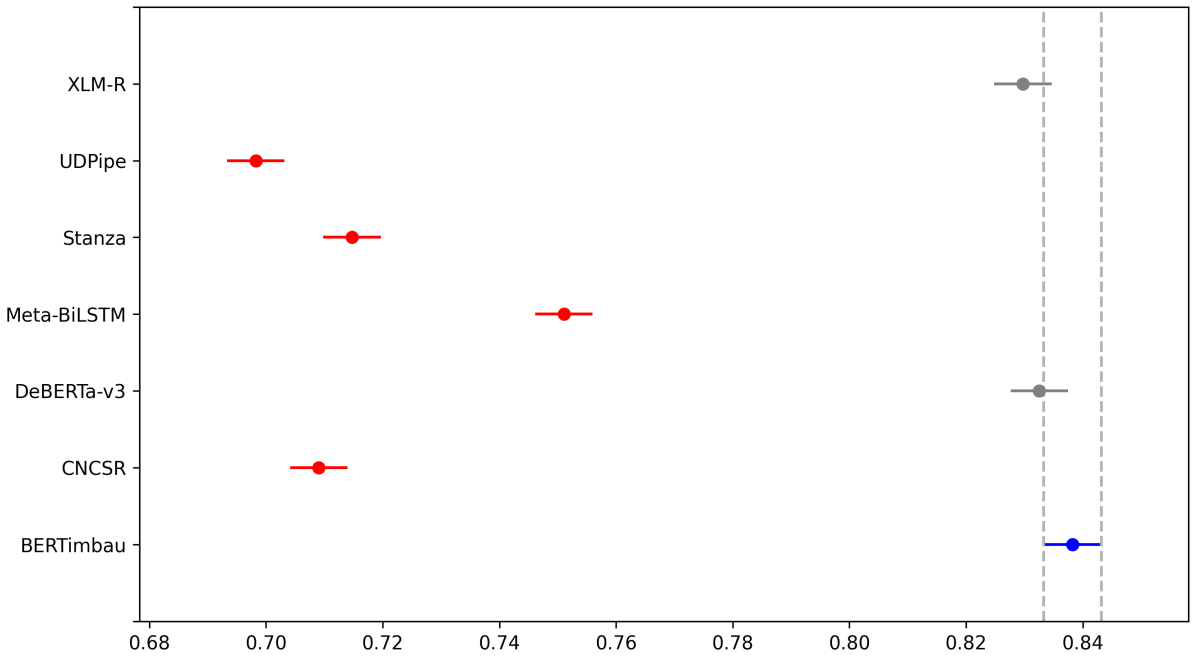
Tabela 9 – Acurácia em nível de sentença no córpus Porttinari-base

Modelo	Acurácia (%)
BERTimbau	**83,8249 ± 0,6829**
DeBERTa-v3	83,2554 ± 0,7161
XLM-R	82,9736 ± 0,7054
Meta-BiLSTM	75,1079 ± 1,0007
UDPipe 2	69,8261 ± 0,5064
Stanza	71,4748 ± 0,6968
CNCSR	70,9053 ± 0,6598

Fonte: Elaborada pelo autor.

DeBERTa-v3 e XLM-R, conforme apresentado na Figura 15.

Figura 15 – Intervalo de confiança da acurácia em nível de sentença média de cada modelo no córpus Porttinari-base.



Fonte: Elaborada pelo autor.

A Tabela 10 apresenta a matriz de p-valores do teste de Tukey aplicado à análise da acurácia em nível de sentença. É possível observar um p-valor maior para a tupla BERTimbau e XLM-R, em comparação à análise realizada em nível de *tokens*, onde se obteve um valor de $0,0249$ e, nesta análise, de $0,1329$. Além disso, é possível observar que os modelos Stanza e CNSCR nesta análise não possuem diferença significativa.

Outro fator importante para a análise de etiquetadores morfossintáticos é a acurácia em nível de *tokens* em sentenças de diferentes tamanhos, pois é relatada na literatura a dificuldade de métodos RNRs processarem sentenças longas, além disso, sentenças pequenas também podem

Tabela 10 – P-valores resultantes da análise *post hoc* de Tukey no córpus Porttinari-base com base na acurácia em nível de sentença

	BERTimbau	DeBERTa-v3	XLM-R	Meta-BiLSTM	UDPipe	Stanza	CNCSR
BERTimbau	-	0,5676	0,1329	0,0010	0,0010	0,0010	0,0010
DeBERTa-v3	0,5676	-	0,9000	0,0010	0,0010	0,0010	0,0010
XLM-R	0,1329	0,9000	-	0,0010	0,0010	0,0010	0,0010
Meta-BiLSTM	0,0010	0,0010	0,0010	-	0,0010	0,0010	0,0010
UDPipe	0,0010	0,0010	0,0010	0,0010	-	0,0010	0,0227
Stanza	0,0010	0,0010	0,0010	0,0010	0,0010	-	0,5676
CNCSR	0,0010	0,0010	0,0010	0,0010	0,0227	0,5676	-

Fonte: Elaborada pelo autor.

apresentar dificuldades aos métodos. A Tabela 11 apresenta as acurácias em nível de *tokens* para todos os modelos com a divisão de sentenças por seus tamanhos em número de *tokens*. O córpus Porttinari-base possui no máximo sentenças com 59 *tokens* no conjunto de teste, dessa forma, optou-se por aferir a acurácia em 6 intervalos de tamanhos iguais, incrementados de 10 em 10 *tokens*. Além disso, são apresentadas no cabeçalho as quantidades de sentenças em cada intervalo e a porcentagem de palavras fora de vocabulário *(OOVs)*.

Tabela 11 – Acurácia em nível de *token* segmentada por tamanhos de sentenças no córpus Porttinari-base

Modelo	1 a 10	11 a 20	21 a 30	31 a 40	41 a 50	51 a 60
Qtde de sentenças	137	926	388	206	8	3
Porcentagem de OOVs	9,8404%	8,1555%	7,9892%	7,1220%	8,2418%	7,6471%
BERTimbau	**97,6950 ± 0,2718**	**99,0847 ± 0,0402**	**99,1689 ± 0,0426**	**99,1406 ± 0,0543**	98,5714 ± 0,3204	99,5294 ± 0,3529
CNCSR	96,7730 ± 0,1644	98,0917 ± 0,0494	98,0722 ± 0,1167	98,3951 ± 0,0928	96,5110 ± 0,3696	99,1765 ± 0,3902
DeBERTa-v3	97,6862 ± 0,2077	99,0050 ± 0,0362	99,1077 ± 0,0568	99,1212 ± 0,0621	**98,7363 ± 0,5095**	**99,7059 ± 0,3946**
Meta-BiLSTM	96,7730 ± 0,1266	98,4799 ± 0,0602	98,4488 ± 0,0821	98,7672 ± 0,0816	98,1044 ± 0,4670	98,7647 ± 0,3168
Stanza	96,6489 ± 0,1625	98,1682 ± 0,0611	98,2839 ± 0,0856	98,4701 ± 0,0804	97,7747 ± 0,2592	99,4118 ± 0,3720
UDPipe	96,4450 ± 0,1007	97,9688 ± 0,0514	98,0681 ± 0,1086	98,2840 ± 0,0374	97,3077 ± 0,2056	98,8824 ± 0,1765
XLM-R	97,3936 ± 0,2545	98,9890 ± 0,0571	99,1160 ± 0,0675	99,1337 ± 0,0655	98,4341 ± 0,2473	99,1765 ± 0,5391

Fonte: Elaborada pelo autor.

É possível observar uma queda nas acurácias em todos os modelos em sentenças com mais de 41 *tokens*, contudo, existem apenas 11 sentenças que se encaixam neste cenário, dessa forma, não é possível tirar conclusões a partir dessa observação. Além disso, observa-se uma queda de acurácia nas sentenças que contêm tamanho de 1 à 10 *tokens*, não obstante, este é o intervalo que possui a maior taxa de palavras fora de vocabulário ($9,84\%$) em comparação aos outros intervalos.

Tratar palavras fora de vocabulário é um grande desafio na área e possui grande impacto no desempenho de ferramentas e aplicações em PLN, dessa forma, a Tabela 12 apresenta a análise de acurácias segmentada por tamanho de sentenças e com distinção entre palavras fora e dentro do vocabulário[33]. Para realização do cálculo, foram segmentados os *tokens* fora de vocabulário e os conhecidos do vocabulário para cada intervalo, em seguida, a acurácia em nível de *tokens* foi calculada para cada caso, onde a coluna OOV com o valor "Sim" indica a acurácia calculada apenas nos *tokens* fora de vocabulário; da mesma forma, quando o valor "Não" é indicado, foram considerados todos os *tokens* que constam no vocabulário. Na coluna que considera sentenças com 1 a 10 *tokens*, é possível observar uma diferença maior das acurácias com e sem OOV, não obstante, essa diferença está presente em todos os modelos. Esta diferença é maior nos modelos baseados em RNR e menor nos baseados em modelos de língua, sendo uma diferença média de $9,58\%$ em modelos RNR e $6,32\%$ em modelos de língua.

[33] Neste contexto, para estar no vocabulário, é necessário que o *token* esteja presente no conjunto de treinamento do córpus; caso esteja apenas no conjunto de testes, então será considerado fora do vocabulário.

Tabela 12 – Acurácia em nível de *token* segmentada por tamanhos de sentenças no córpus Porttinari-base e distinguindo entre palavras dentro e fora de vocabulário

Modelo	OOV	1 a 10	11 a 20	21 a 30	31 a 40	41 a 50	51 a 60
BERTimbau	Sim	91,5315 ± 0,8257	**96,4548 ± 0,2452**	96,9740 ± 0,3130	**96,4717 ± 0,3642**	97,0000 ± 3,7859	**100,0000 ± 0,0000**
	Não	**98,3677 ± 0,2857**	**99,3183 ± 0,0313**	**99,3595 ± 0,0318**	99,3453 ± 0,0427	98,7126 ± 0,2338	99,4904 ± 0,3822
CNCSR	Sim	88,0180 ± 0,7036	92,9340 ± 0,4553	92,4286 ± 0,7114	92,5341 ± 0,3268	90,6667 ± 2,4944	99,2308 ± 2,3077
	Não	97,7286 ± 0,2262	98,5497 ± 0,0587	98,5622 ± 0,0890	98,8445 ± 0,0989	97,0359 ± 0,4531	99,1720 ± 0,4078
DeBERTa-v3	Sim	**92,7928 ± 0,9009**	96,3733 ± 0,2452	**97,0390 ± 0,3664**	95,9649 ± 0,3268	97,0000 ± 2,7689	**100,0000 ± 0,0000**
	Não	98,2203 ± 0,1989	99,2387 ± 0,0428	99,2873 ± 0,0536	**99,3632 ± 0,0564**	**98,8922 ± 0,3555**	**99,6815 ± 0,4273**
Meta-BiLSTM	Sim	88,6486 ± 0,9187	94,5314 ± 0,3166	94,8701 ± 0,3863	95,0097 ± 0,3285	97,0000 ± 1,7951	**100,0000 ± 0,0000**
	Não	97,6598 ± 0,1304	98,8305 ± 0,0833	98,7596 ± 0,0695	99,0553 ± 0,0824	98,2036 ± 0,4828	98,6624 ± 0,3430
Stanza	Sim	90,7207 ± 0,9054	94,8492 ± 0,2023	94,9351 ± 0,4928	94,9318 ± 0,4613	**99,6667 ± 1,0000**	**100,0000 ± 0,0000**
	Não	97,2960 ± 0,2027	98,4629 ± 0,0637	98,5747 ± 0,0695	98,7414 ± 0,0785	97,6048 ± 0,2994	99,3631 ± 0,4028
UDPipe	Sim	84,6847 ± 1,0660	91,2795 ± 0,4250	92,0130 ± 0,5072	91,2281 ± 0,5008	93,0000 ± 1,0000	95,3846 ± 3,7684
	Não	97,7286 ± 0,1027	98,5627 ± 0,0475	98,5938 ± 0,0780	98,8251 ± 0,0342	97,6946 ± 0,1917	99,1720 ± 0,2919
XLM-R	Sim	91,3514 ± 1,6216	96,1369 ± 0,1901	97,0000 ± 0,2987	96,4133 ± 0,2917	99,0000 ± 2,1344	93,0769 ± 2,3077
	Não	98,0531 ± 0,2100	99,2423 ± 0,0530	99,2997 ± 0,0639	99,3423 ± 0,0627	98,3832 ± 0,1986	99,6815 ± 0,5135

Fonte: Elaborada pelo autor.

As análises apresentadas têm o objetivo de apresentar diferentes olhares para o etiquetador morfossintático, dessa forma, permitindo ao usuário final utilizar o etiquetador que atenda melhor as necessidades da sua aplicação. Naturalmente, para seleção do modelo com melhor desempenho no córpus Porttinari-base, a análise realizada na acurácia em nível de *tokens* será utilizada, pois esta análise permite avaliar os etiquetadores de forma holística, além disso, é a métrica mais utilizada na literatura. Foi observado um empate entre os modelos BERTimbau e DeBERTa-v3, onde não foi possível observar diferença significativa entre as acurácias, dessa forma, ambos os modelos poderiam ser selecionados para os experimentos no contexto multigênero. Contudo, outro fator importante para seleção de modelos é o total de parâmetros, pois isso impacta no tempo de treinamento, de inferência e consumo de memória, dessa forma, seguindo o princípio da Navalha de Occam (DOMINGOS, 1999), o modelo que possui menos parâmetros é o BERTimbau, onde, em comparação ao DeBERTa-v3, possui aproximadamente $2,5$ vezes menos parâmetros.

A análise de precisão e sensibilidade por cada tipo de etiqueta gramatical presente no córpus também é relevante e, para realizar estas análises, foi selecionado um dos experimentos executados com o modelo BERTimbau, em particular, foi selecionada a execução que obteve a maior Medida-F macro no córpus Porttinari-base. Não obstante, não existe diferença significativa entre os diferentes experimentos de um mesmo modelo, dado que o único fator que difere os experimentos é a semente aleatória, dessa forma, outro experimento poderia ter sido selecionado e seriam apresentados resultados semelhantes. A Tabela 13 apresenta as medidas de precisão, sensibilidade e Medida-F para cada etiqueta do conjunto UD presente no córpus Porttinari-base, onde é possível observar a falta da etiqueta *PART*, que não teve ocorrências no córpus. Além disso, também é apresentado o número de ocorrências de cada etiqueta no córpus. Dessa forma, é possível observar que algumas etiquetas possuem baixa quantidade de ocorrência, impactando diretamente a métrica, como, por exemplo, na etiqueta *INTJ* (interjeição), que foi avaliada com a sensibilidade de $87,5\%$.

Como esperado, classes gramaticais não ambíguas como *PUNCT* (pontuação) e *SYM* (símbolos) obtiveram métricas iguais ou muito próximas a 100%. No entanto, a precisão da classe *PUNCT* não alcança a métrica máxima por erro do modelo ao etiquetar o *token* "/Juazeiro"[34] como *PUNCT* quando deveria ser *PROPN*, porém, este é um dos casos que ainda necessita de correção no córpus, pois contém dois *tokens* na mesma unidade[35].

A classe *X* apresenta valores baixos para precisão e sensibilidade. Este comportamento é esperado, dado que a etiqueta *X*, segundo o manual da UD, deve ser utilizada pelos rotuladores quando o *token* analisado não se encaixar em nenhuma das outras etiquetas disponíveis. Consequentemente, não existe um padrão bem definido para reconhecimento desta etiqueta, sendo

[34] Sentença completa: "O voo em que foi registrada a maior diferença de preço, uma média de R$ 2.050 entre a empresa mais cara e a mais barata, foi Navegantes (SC)/Juazeiro do Norte (CE).".

[35] O córpus está atualmente em revisão pelo grupo de pesquisa e já existem versões mais recentes com correções de diversos casos.

Tabela 13 – Precisão, sensibilidade e Medida-F para cada etiqueta da UD no córpus Porttinari-base

Etiqueta	Precisão	Sensibilidade	Medida-F	Número de ocorrências
ADJ	95,9977%	96,2209%	96,1092%	1.720
ADP	99,8799%	99,7600%	99,8199%	5.000
ADV	98,4352%	98,2156%	98,3253%	1.345
AUX	98,9669%	99,3776%	99,1718%	964
CCONJ	99,0521%	98,7013%	98,8764%	847
DET	99,2464%	99,7685%	99,5068%	4.752
INTJ	100,0000%	87,5000%	93,3333%	8
NOUN	99,1888%	99,1727%	99,1808%	6.165
NUM	99,4907%	98,4874%	98,9865%	595
PRON	98,6677%	98,3594%	98,5133%	1.280
PROPN	99,4495%	99,1766%	99,3129%	2.186
PUNCT	99,9781%	100,0000%	99,9891%	4.566
SCONJ	96,1945%	97,4304%	96,8085%	467
SYM	100,0000%	100,0000%	100,0000%	96
VERB	98,9130%	98,6872%	98,8000%	3.504
X	76,4706%	73,5849%	75,0000%	53

Fonte: Elaborada pelo autor.

assim, o modelo também não consegue distinguir tão bem estes casos. Analisando manualmente os erros para esta etiqueta, é possível observar que em casos onde ocorre estrangeirismo e os anotadores humanos optaram por rotular com a etiqueta *X*, o modelo não associa a etiqueta *X*, ao contrário, ele busca classificar o *token* com a etiqueta provável, ignorando a regra adotada pelos anotadores. Por exemplo, na sentença "Sim, seria duro aguentar os hermanos campeões em solo carioca, mas Messi merecia."[36], o *token* "hermanos" é rotulado com a etiqueta *X*, porém, o modelo associa a etiqueta *NOUN* (substantivo).

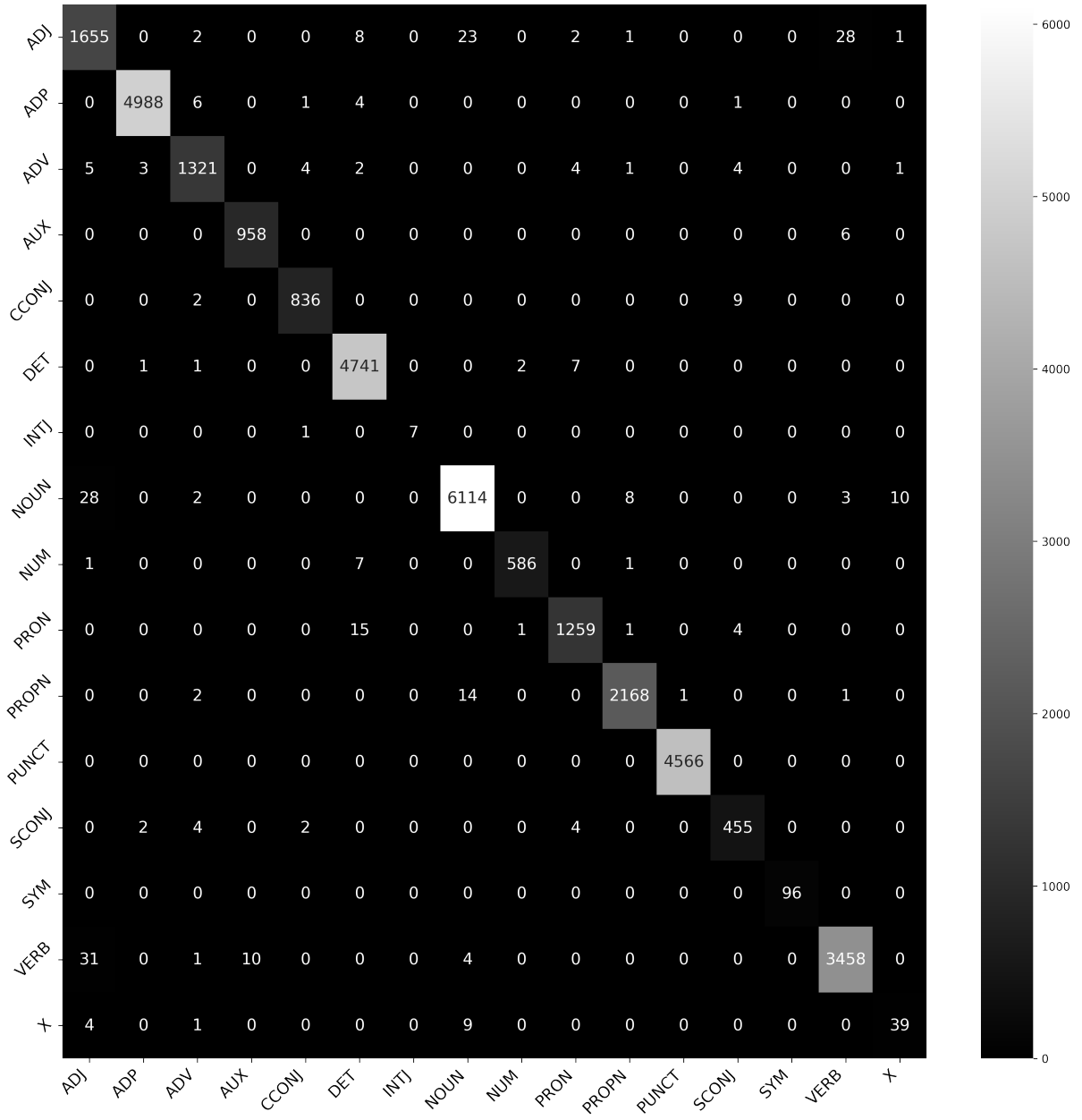
A etiqueta *ADJ* (adjetivo) apresenta valores de precisão e sensibilidade na casa de 96%, sendo menor do que outras etiquetas. Analisando manualmente os casos em que o modelo errou na classificação desta etiqueta, é possível observar que cerca de 35% dos erros são casos de particípio, onde o modelo deveria etiquetar como *ADJ*, contudo, apresenta a etiqueta *VERB* (verbo). Conforme discutido no trabalho de DURAN (2021), os particípios podem ser anotados com *ADJ*, *NOUN* ou *VERB*, consequentemente, tornando a tarefa mais desafiadora para os etiquetadores morfossintáticos.

A matriz de confusão apresentada na Figura 16 auxilia na compreensão dos erros do modelo BERTimbau no córpus Porttinari-base. Cada linha da matriz de confusão representa o rótulo presente no córpus e cada coluna representa a classe associada pelo modelo, dessa forma, na diagonal principal encontram-se os verdadeiros positivos, ou seja, casos onde o modelo

[36] Sentença presente no conjunto de teste do córpus Porttinari-base com identificador *FOLHA_-DOC000073_SENT015*

associou corretamente a etiqueta esperada. A matriz permite a visualização da frequência com que cada classe gramatical foi classificada incorretamente como outra classe, o que auxilia na identificação de quais classes o modelo está tendo dificuldade em distinguir. Por exemplo, a classe *ADJ* possui 65 casos em que o *token* deveria receber a classe *ADJ*, porém, o modelo associou outra classe. Em particular, houve 23 casos onde o modelo classificou como *NOUN*, em *28* vezes a classe *VERB* foi utilizada e em 2 casos a etiqueta *ADV* foi associada. A figura utiliza uma escala de cores para indicar a frequência de ocorrências, variando do preto com 0 casos, até o branco com pouco mais de 6.000.

Figura 16 – Matriz de confusão do modelo BERTimbau no córpus Porttinari-base



Fonte: Elaborada pelo autor.

5.2 Avaliação do modelo BERTimbau para etiquetagem multigênero

A partir do modelo BERTimbau, que foi selecionado com base nos experimentos da subseção anterior, o modelo é avaliado no contexto multigênero, considerando córpus para o português do Brasil dos gêneros jornalístico, acadêmico e Conteúdo Gerado por Usuário (CGU). Conforme descrito no Capítulo 4, o experimento utiliza um modelo avaliado em diferentes contextos, sendo treinado em cada córpus isoladamente e no contexto multigênero, sendo este a combinação de dois ou três gêneros. Para cada cenário de avaliação, são realizados 10 experimentos e a métrica de acurácia em nível de *tokens* e sentença é calculada para cada córpus. Em seguida, são realizados os testes de hipótese para verificar diferenças significativas entre os diferentes contextos. Não obstante, não houve alterações de hiper-parâmetros nas execuções dos experimentos, apenas a alteração da semente aleatória.

Tabela 14 – Acurácias em nível de *tokens* no contexto multigênero

	Acurácia ao nível *tokens* (%)		
Cenário	Porttinari-base	DANTEStocks	PetroGold
Porttinari-base	**99,0709 ± 0,0310**	87,1380 ± 0,5914	96,4603 ± 0,1675
DANTEStocks	96,5491 ± 0,2250	**97,9843 ± 0,0764**	94,9540 ± 0,1958
PetroGold	96,9885 ± 0,1029	84,9592 ± 0,4605	**98,9336 ± 0,0505**
Porttinari DANTEStocks	99,0494 ± 0,0350	97,9136 ± 0,1042	96,5817 ± 0,1606
Porttinari PetroGold	98,9457 ± 0,0643	85,2944 ± 0,3405	98,8480 ± 0,0705
DANTEStocks PetroGold	97,8622 ± 0,0612	97,9856 ± 0,0740	98,9225 ± 0,0536
Porttinari DANTEStocks PetroGold	99,0044 ± 0,0480	97,9174 ± 0,1260	98,8949 ± 0,0609

Fonte: Elaborada pelo autor.

A Tabela 14 apresenta os resultados dos experimentos para todos os cenários, considerando o treinamento nos córpus isolados e no contexto multigênero, onde é exibida a acurácia média em nível de *tokens* no conjunto de teste para cada córpus da avaliação. É possível observar que, para os três córpus, o cenário que obteve a maior acurácia média foi o cenário onde o modelo foi treinado apenas com dados do gênero alvo, por exemplo, o melhor cenário para córpus de gênero acadêmico foi o cenário treinado exclusivamente neste gênero. Contudo, estes modelos possuem baixas acurácias nos outros gêneros, por exemplo, o modelo treinado no córpus PetroGold com acurácia de $98,93\%$ no gênero acadêmico possui acurácia de $84,96\%$ no gênero CGU. É possível observar maior discrepância entre os gêneros textuais que seguem a norma culta da língua em relação ao gênero CGU, que possui características linguísticas diferentes. Quando o cenário PetroGold é avaliado no gênero jornalístico, por exemplo, é possível observar uma acurácia de $96,99\%$, ou seja, quando o modelo é treinado em córpus acadêmico e avaliado no gênero jornalístico, observa-se alta acurácia em comparação ao melhor modelo de gênero jornalístico (uma diferença de $2,08\%$). Já em relação ao gênero CGU, observa-se uma diferença de $13,03\%$ em relação ao melhor modelo treinado no córpus DANTEStocks. O

mesmo é observado no caso inverso, ou seja, quando o modelo é treinado em córpus jornalístico e avaliado no gênero acadêmico.

Constata-se um fenômeno diferente ao analisar os resultados do modelo treinado exclusivamente em gênero CGU, onde o modelo conseguiu resultados consistentes nos córpus jornalístico e acadêmico, apesar de não utilizá-los no treinamento e possuir características linguísticas diferentes. Neste caso, o modelo obteve diferença de $2,52\%$ em relação ao melhor modelo do gênero jornalístico e, para o gênero acadêmico, diferença de $3,98\%$. É importante ressaltar que, apesar do gênero CGU possuir características diferentes dos textos que seguem a norma culta da língua, ainda assim, são encontradas diversas similaridades. Dessa forma, é possível observar que o modelo consegue generalizar ao ponto de capturar as características similares e manter alto desempenho no córpus DANTEStocks, contudo, as limitações nos outros gêneros não deixam de ser relevantes.

Em relação aos quatro cenários multigênero, é possível observar que, em todos os casos, o modelo alcançou desempenhos similares aos melhores cenários de cada gênero e também um acréscimo no desempenho do gênero que não foi contemplado no contexto multigênero. Por exemplo, quando treinado no cenário "PetroGold DANTEStocks", o modelo obteve acurácia de $97,86\%$ no córpus Porttinari-base, em contraposição quando foi treinado exclusivamente no córpus PetroGold, em que foi obtida a acurácia de $96,99\%$, e quando treinado no córpus DANTEStocks, em que foi alcançada a acurácia de $96,55\%$. Dessa forma, houve uma melhora no gênero jornalístico ao realizar o treinamento nos gêneros acadêmico e CGU em relação aos modelos treinados nestes gêneros isolados. Além disso, o modelo treinado em todos os gêneros alcançou desempenho similar aos modelos treinados isoladamente, sendo que a diferença entre as médias possui valor máximo de $0,067\%$.

Para verificar se as diferenças entre os cenários são significativas, os testes de hipótese ANOVA e Tukey foram realizados para cada gênero textual. O teste ANOVA apresentou estatística $1107,7847$ e p-valor $7,3455e-62$, dessa forma, é possível verificar que há diferença significativa entre os cenários com $alpha = 0,05$. A Tabela 15 apresenta a matriz de p-valores ao analisar a acurácia em nível de *tokens* no córpus Porttinari-base, onde são indicados pela cor vermelha os casos significativos, ou seja, onde os resultados dos cenários possuem diferença significativa, por exemplo, o cenário treinado no córpus Porttinari-base em relação ao cenário DANTEStocks possui diferença significativa. Como apresentado na Tabela 14, o cenário que obteve o maior desempenho no córpus Porttinari-base, foi justamente o que foi treinado apenas neste córpus, contudo, é possível observar que não existe diferença significativa deste cenário em relação aos cenários multigênero que utilizaram o córpus de gênero jornalístico em seu treinamento. Por exemplo, o cenário "Porttinari-base + DANTEStocks" possui p-valor de $0,99$ em comparação ao cenário isolado Porttinari-base, dessa forma, é possível afirmar que este contexto multigênero não possui diferença significativa em relação ao cenário Porttinari-base. Assim, pode-se concluir que este cenário multigênero não degradou o desempenho no córpus de gênero

Tabela 15 – Matriz de p-valores da análise de treinamento multigênero avaliando no córpus Porttinari-base

	Porttinari-base	DANTEStocks	PetroGold	Porttinari-base DANTEStocks	Porttinari-base PetroGold	DANTEStocks PetroGold	Porttinari-base DANTEStocks PetroGold
Porttinari-base	-	0,0000	0,0000	0,9992	0,1076	0,0000	0,7728
DANTEStocks	0,0000	-	0,0000	0,0000	0,0000	0,0000	0,0000
PetroGold	0,0000	0,0000	-	0,0000	0,0000	0,0000	0,0000
Porttinari-base DANTEStocks	0,9992	0,0000	0,0000	-	0,2787	0,0000	0,9562
Porttinari-base PetroGold	0,1076	0,0000	0,0000	0,2787	-	0,0000	0,8583
DANTEStocks PetroGold	0,0000	0,0000	0,0000	0,0000	0,0000	-	0,0000
Porttinari-base DANTEStocks PetroGold	0,7728	0,0000	0,0000	0,9562	0,8583	0,0000	-

Fonte: Elaborada pelo autor.

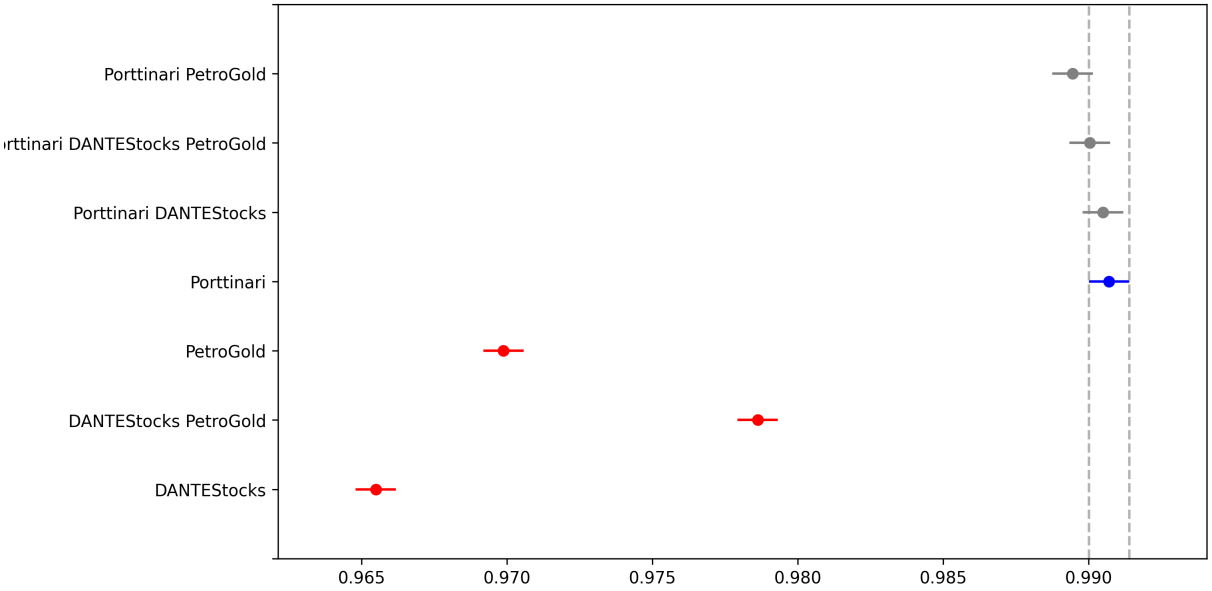
jornalístico. Este comportamento também ocorre nos cenários "Porttinari-base + PetroGold" e "Porttinari-base + DANTEStocks + PetroGold". Já para o cenário multigênero "DANTEStocks + PetroGold", é possível observar que existe diferença significativa, ou seja, o cenário que não incluiu o gênero jornalístico e se apoia nos gêneros acadêmico e CGU, não alcançou desempenho similar ao modelo do gênero jornalístico e esta diferença é significativa. Além disso, na Tabela 14 é possível observar que o cenário "DANTEStocks + PetroGold" alcançou acurácia maior do que os cenários isolados DANTEStocks e PetroGold, onde esta diferença é confirmada pelo teste estatístico, ou seja, o treinamento nos gêneros acadêmico e CGU trouxe uma melhora significativa no desempenho em gênero jornalístico, mesmo que não tenha sido treinado neste gênero.

A Figura 17 apresenta o intervalo de confiança extraído a partir da análise de Tukey. É possível observar que todos os cenários que incluem o córpus jornalístico em seu treinamento estão no intervalo de confiança do cenário isolado com $alpha = 0,05$.

Em seguida, os testes ANOVA e Tukey foram aplicados ao córpus de gênero CGU. O teste ANOVA resultou na estatística de $4171,2253$ e p-valor de $6,7382e-80$, dessa forma, é possível afirmar que existe diferença significativa entre os diferentes cenários aplicados ao córpus DANTEStocks. A Tabela 16 sumariza os p-valores encontrados a partir da análise *post hoc* de Tukey. Semelhante à análise de Tukey aplicada ao córpus de gênero jornalístico, o cenário com melhor desempenho (DANTEStocks) não possui diferença significativa em comparação aos cenários multigênero que utilizaram o córpus DANTEStocks em seu treinamento. Não obstante, o cenário multigênero "Porttinari-base + PetroGold" apresentou diferença significativa, contudo, o desempenho é menor neste cenário. Conclui-se que, para o córpus de gênero CGU, o aprendizado multigênero que contém o gênero CGU no conjunto de treinamento não deteriora o desempenho do modelo, pois não foram encontradas diferenças significativas nos resultados.

A Figura 18 apresenta o intervalo de confiança do teste de Tukey aplicado à avaliação do aprendizado multigênero no córpus de gênero CGU. Novamente, é possível observar que todos

Figura 17 – Intervalo de confiança da acurácia em nível de *tokens* do modelo BERTimbau no córpus Porttinari-base em diferentes cenários.



Fonte: Elaborada pelo autor.

Tabela 16 – Matriz de p-valores da análise multigênero no córpus DANTEStocks

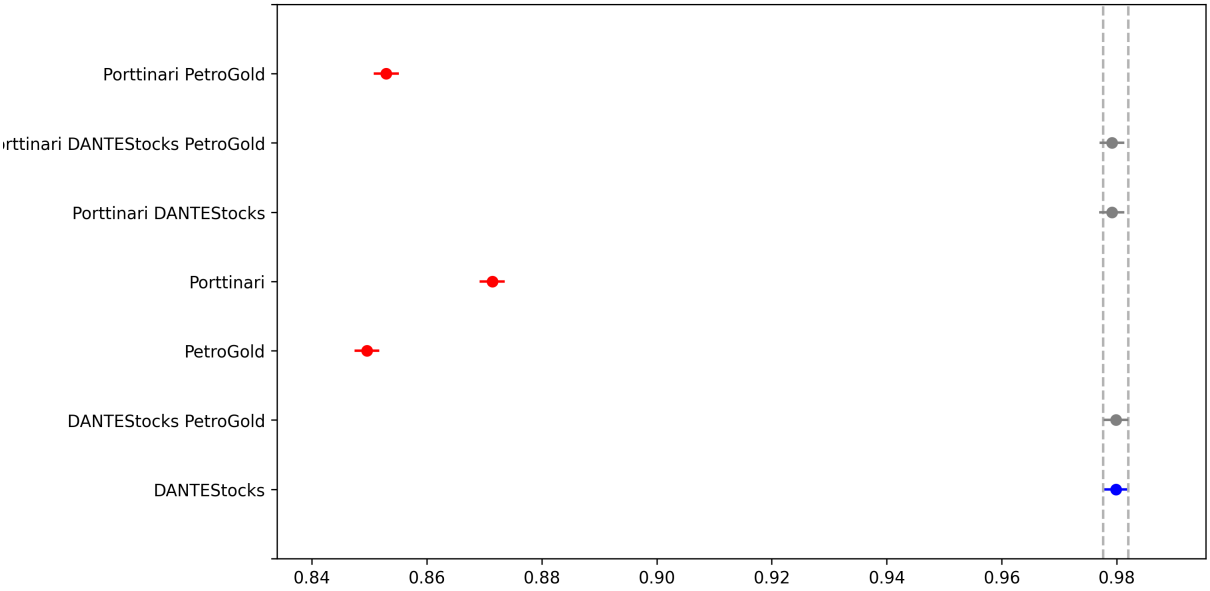
	Porttinari-base	DANTEStocks	PetroGold	Porttinari-base DANTEStocks	Porttinari-base PetroGold	DANTEStocks PetroGold	Porttinari-base DANTEStocks PetroGold
Porttinari-base	-	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
DANTEStocks	0,0000	-	0,0000	0,9988	0,0000	1,0000	0,9992
PetroGold	0,0000	0,0000	-	0,0000	0,2401	0,0000	0,0000
Porttinari-base DANTEStocks	0,0000	0,9988	0,0000	-	0,0000	0,9987	1,0000
Porttinari-base PetroGold	0,0000	0,0000	0,2401	0,0000	-	0,0000	0,0000
DANTEStocks PetroGold	0,0000	1,0000	0,0000	0,9987	0,0000	-	0,9991
Porttinari-base DANTEStocks PetroGold	0,0000	0,9992	0,0000	1,0000	0,0000	0,9991	-

Fonte: Elaborada pelo autor.

os cenários multigênero que possuem o córpus DANTEStocks em seu conjunto de treinamento estão no intervalo de confiança do modelo que foi treinado exclusivamente no contexto CGU. Não obstante, os demais cenários encontram-se distantes dos melhores modelos, conforme o esperado, dadas as características textuais do gênero CGU.

Por fim, o mesmo procedimento foi aplicado a análise do aprendizado multigênero aplicada ao córpus de gênero acadêmico. A execução do teste ANOVA apresentou estatística de $1764,8227$ e p-valor de $3,5136e-68$, ou seja, foi encontrada diferença significativa entre os cenários. A Tabela 17 apresenta os resultados do teste de Tukey, onde é possível observar os p-valores de cada cenário. Novamente, em vermelho encontram-se os casos onde há diferença significativa pelo teste, com $alpha = 0.05$. É possível observar o mesmo comportamento dos

Figura 18 – Intervalo de confiança da acurácia em nível de *tokens* do modelo BERTimbau no córpus DANTEStocks em diferentes cenários.



Fonte: Elaborada pelo autor.

Tabela 17 – Matriz de p-valores da análise multigênero no córpus PetroGold

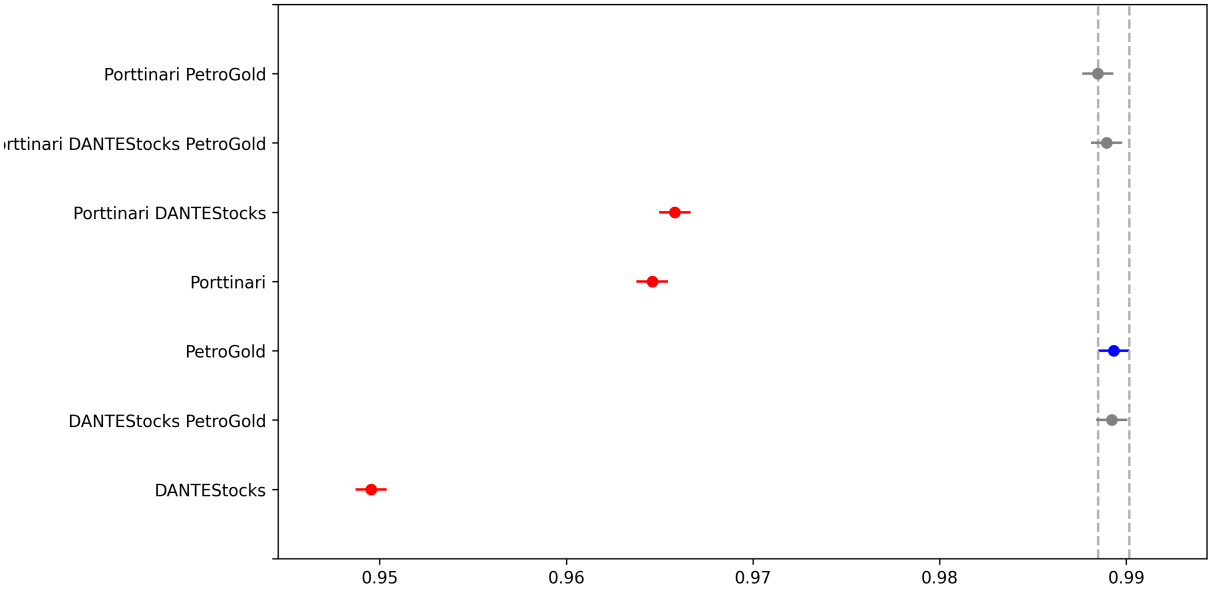
	Porttinari-base	DANTEStocks	PetroGold	Porttinari-base DANTEStocks	Porttinari-base PetroGold	DANTEStocks PetroGold	Porttinari-base DANTEStocks PetroGold
Porttinari-base	-	0,0000	0,0000	0,3081	0,0000	0,0000	0,0000
DANTEStocks	0,0000	-	0,0000	0,0000	0,0000	0,0000	0,0000
PetroGold	0,0000	0,0000	-	0,0000	0,7120	1,0000	0,9921
Porttinari-base DANTEStocks	0,3081	0,0000	0,0000	-	0,0000	0,0000	0,0000
Porttinari-base PetroGold	0,0000	0,0000	0,7120	0,0000	-	0,8242	0,9782
DANTEStocks PetroGold	0,0000	0,0000	1,0000	0,0000	0,8242	-	0,9988
Porttinari-base DANTEStocks PetroGold	0,0000	0,0000	0,9921	0,0000	0,9782	0,9988	-

Fonte: Elaborada pelo autor.

testes em gênero jornalístico e CGU, onde não há diferença significativa entre o cenário exclusivamente treinado no gênero acadêmico em relação aos modelos multigênero que incluem o gênero acadêmico em seu treinamento. Dessa forma, o aprendizado multigênero não deteriorou o desempenho do aprendizado isolado no gênero acadêmico.

Os intervalos de confiança relacionados ao córpus de gênero acadêmico podem ser visualizados na Figura 19. Novamente, é possível visualizar a intersecção entre os intervalos de confiança dos modelos que utilizaram o PetroGold em seu conjunto de treinamento. Não obstante, também é possível observar a intersecção dos intervalos de confiança entre os cenários Porttinari-base e "Porttinari-base + DANTEStocks", dessa forma, o aprendizado em gênero jornalístico e CGU não apresentou melhora significativa em relação ao aprendizado no córpus

Figura 19 – Intervalo de confiança da acurácia em nível de *tokens* do modelo BERTimbau no córpus PetroGold em diferentes cenários.



Fonte: Elaborada pelo autor.

Porttinari-base quando avaliado no córpus PetroGold.

Por fim, o modelo avaliado no contexto multigênero "Porttinari-bae + DANTEStocks + PetroGold" não apresentou diferenças significativas em relação aos modelos treinados isoladamente em cada gênero textual, desse modo, pode-se afirmar que o aprendizado multigênero nos córpus selecionados, utilizando o modelo BERTimbau com a estratégia de ajuste fino e métrica de acurácia em nível de *tokens*, não deteriora o aprendizado nos gêneros individuais. Esta análise permite a verificação da segunda hipótese de pesquisa apresentada no Capítulo 1, onde se conjecturou a viabilidade da criação de um etiquetador morfossintático multigênero para o português do Brasil que não degradasse o desempenho nos gêneros individuais.

Além da análise da acurácia em nível de *tokens*, também é de relevância realizar a análise em nível de sentença, onde é mensurada a porcentagem de sentenças completamente corretas. A Tabela 18 apresenta esta medida para cada cenário dos experimentos realizados. Semelhantemente à análise da acurácia em nível de *tokens*, podemos observar que os cenários que obtiveram o melhor desempenho foram os que consideraram apenas o gênero alvo, contudo, para o córpus de gênero CGU, é observado que o cenário multigênero "Porttinari-base + DANTEStocks" apresentou o melhor desempenho para este gênero. Não obstante, os cenários multigênero novamente obtiveram acurácias similares às obtidas aos cenários isolados. Por exemplo, o cenário "Porttinari-base + DANTEStocks + PetroGold" obteve as diferenças em relação aos melhores cenários para os córpus de gênero jornalístico, CGU e acadêmico de $1,19\%$, $1,01\%$ e $0,70\%$, respectivamente.

Os testes estatísticos foram realizados separadamente para cada gênero textual, seguindo

Tabela 18 – Acurácia média em nível de sentença para cada cenário da avaliação multigênero

	Acurácia em nível de sentença (%)		
Cenário	Porttinari-base	DANTEStocks	PetroGold
Porttinari-base	**83,8249 ± 0,6829**	9,8130 ± 0,7103	45,9326 ± 1,5826
DANTEStocks	54,8681 ± 2,2123	72,4688 ± 0,8055	35,4157 ± 1,1517
PetroGold	57,0024 ± 1,0372	6,2344 ± 0,7573	**78,8539 ± 0,7672**
Porttinari-base DANTEStocks	83,4353 ± 0,5444	**73,3292 ± 1,0587**	48,1798 ± 1,6142
Porttinari-base PetroGold	81,7206 ± 0,8644	5,9975 ± 0,7375	77,3258 ± 1,5693
DANTEStocks PetroGold	67,2422 ± 0,9228	72,3441 ± 0,5049	78,7865 ± 1,2632
Porttinari-base DANTEStocks PetroGold	82,6379 ± 0,8703	72,3192 ± 1,2017	78,1573 ± 0,9639

Fonte: Elaborada pelo autor.

Tabela 19 – P-valores do cenário multigênero "Porttinari-base + DANTEStocks + PetroGold" resultantes do teste de Tukey

	Porttinari-base	DANTEStocks	PetroGold	Porttinari-base DANTEStocks	Porttinari-base PetroGold	DANTEStocks PetroGold
Porttinari-base	0,2466	0,0000	0,0000	0,7047	0,5530	0,0000
DANTEStocks	0,0000	0,9997	0,0000	0,1293	0,0000	1,0000
PetroGold	0,0000	0,0000	0,8957	0,0000	0,7895	0,9334

Fonte: Elaborada pelo autor.

o mesmo procedimento descrito anteriormente, iniciando pelo teste ANOVA para identificar diferença significativa entre todos os cenários e, em seguida, realizando o teste *post hoc* de Tukey. Foram encontradas diferenças significativas nas três análises com o teste ANOVA[37]. Ao realizar o teste de Tukey para cada gênero textual, foi encontrado o mesmo padrão na análise de acurácia em nível de *tokens*, onde os cenários multigênero que continham o gênero alvo não apresentaram diferenças significativas em relação ao cenário isolado. Contudo, houve uma exceção, onde o cenário "Porttinari-base + PetroGold" com acurácia de $81,72\%$ no córpus Porttinari-base apresentou diferença significativa em relação ao modelo isolado, que possui acurácia de $83,82\%$, onde se obteve o p-valor de $0,002$. Dado que não foram encontradas outras diferenças, as três matrizes de p-valores foram sumarizadas na Tabela 19. Esta tabela descreve os p-valores em relação a cada gênero textual, sendo apresentado em cada linha, e em relação aos p-valores da comparação entre o cenário multigênero "Porttinari-base + DANTEStocks + PetroGold" e os demais cenários. Dessa forma, é possível visualizar o resultado da análise para o cenário multigênero que abrange todos os córpus em comparação aos demais cenários. Com esta análise, é possível observar os p-valores para cada cenário isolado em seu respectivo gênero textual, por exemplo, com o p-valor de $0,2466$ para o cenário Porttinari-base, é possível afirmar que o modelo treinado no cenário multigênero não possui diferença significativa em relação ao cenário isolado no gênero jornalístico.

Outro fator importante na análise da acurácia em nível de sentença é a discrepância

[37] Porttinari-base: $Z = 1300,1033$, $p - valor = 4,9687e - 64$
DANTEStocks: $Z = 16757,0885$, $p - valor = 6,7922e - 99$
PetroGold: $Z = 2143,3330$, $p - valor = 7,9996e - 71$

Tabela 20 – Medida-F por classe gramatical considerando o modelo multigênero e os modelos isolados em cada gênero jornalístico

	Porttinari-base			DANTEStocks			PetroGold		
	Jornalístico	Multigênero	Qtde classes	CGU	Multigênero	Qtde classes	Acadêmico	Multigênero	Qtde classes
ADJ	96,22%	96,24%	1720	95,91%	94,65%	591	97,80%	97,44%	705
ADP	99,76%	99,78%	5000	99,74%	99,59%	1.697	99,68%	99,59%	1847
ADV	98,22%	97,70%	1345	96,30%	95,89%	543	98,88%	99,04%	314
AUX	99,38%	99,22%	964	94,68%	94,49%	275	99,66%	99,83%	294
CCONJ	98,70%	98,71%	847	99,05%	99,68%	315	99,64%	99,64%	280
DET	99,77%	99,46%	4752	99,14%	99,07%	1.344	99,91%	99,62%	1597
INTJ	87,50%	77,78%	8	79,41%	82,86%	36	-	-	0
NOUN	99,17%	99,17%	6165	97,66%	97,52%	2.313	98,88%	98,92%	2556
NUM	98,49%	98,82%	595	99,32%	99,49%	875	98,47%	99,08%	326
PART	-	-	0	100,00%	0,00%	1	-	-	0
PRON	98,36%	98,33%	1280	96,11%	95,42%	258	98,28%	97,89%	233
PROPN	99,18%	99,11%	2186	97,49%	97,81%	2.292	95,63%	96,39%	455
PUNCT	100,00%	99,98%	4566	99,76%	99,68%	2.509	99,65%	99,80%	1268
SCONJ	97,43%	95,44%	467	93,08%	94,57%	131	92,62%	94,44%	71
SYM	100,00%	100,00%	96	99,46%	99,28%	829	100,00%	97,87%	23
VERB	98,69%	98,96%	3504	98,20%	97,89%	1.275	99,05%	98,83%	892
X	73,58%	74,16%	53	89,58%	90,63%	408	92,31%	92,31%	7

Fonte: Elaborada pelo autor.

encontrada entre o desempenho em córpus jornalístico, acadêmico e CGU. Enquanto o melhor desempenho em córpus jornalístico alcança acurácia de $83,83\%$, ocorre uma queda de aproximadamente $4,97\%$ em relação ao melhor desempenho no gênero acadêmico. A discrepância é ainda maior no gênero CGU, onde a diferença entre o melhor desempenho de gênero jornalístico e CGU é de $10,50\%$. Dessa forma, é possível observar que o método BERTimbau possui melhor desempenho em sentenças completamente corretas em córpus do gênero jornalístico e seu desempenho decresce no córpus PetroGold (acadêmico) e DANTEStocks (CGU).

Com a verificação das acurácias em nível de *tokens* e sentenças, também é realizada a análise da Medida-F segmentada pelas etiquetas morfossintáticas, dessa forma, possibilitando a visualização do desempenho por classe. No contexto da avaliação multigênero, o foco desta análise é de comparar a Medida-F entre os cenários multigênero e isolados. Dado que o cenário multigênero "Porttinari-base + DANTEStocks + PetroGold" não possui diferença significativa entre os demais cenários multigênero e abrange todos os gêneros, este cenário foi colocado em comparação aos cenários isolados na Tabela 20, onde a coluna "Multigênero" representa o cenário que contém todos os gêneros. A tabela é dividida em três divisões, representando a avaliação em cada córpus. Em cada divisão, encontram-se as Medidas-F para o modelo isolado referente ao gênero da divisão e para o cenário multigênero, além de apresentar a quantidade de etiquetas para a classe correspondente. É importante ressaltar que alguns córpus não possuem algumas classes gramaticais, como o córpus Porttinari-base que não possui etiquetas do tipo *PART* (partícula). Não obstante, a tabela foi construída a partir de uma das execuções, seguindo o mesmo procedimento realizado na análise do córpus Porttinari-base na subseção anterior.

Nos três cenários de avaliação, é possível encontrar valores próximos entre o desempenho do cenário multigênero e isolado. No entanto, em etiquetas que possuem poucas ocorrências,

como a etiqueta *INTJ* (interjeição) no córpus Porttinari-base, existe uma maior variação no desempenho, contudo, este comportamento é esperado dado o baixo número de amostras.

Concluindo, este capítulo apresenta a análise de etiquetadores morfossintáticos para o português do Brasil em córpus jornalístico e a avaliação no contexto multigênero, considerando córpus do gênero jornalístico, acadêmico e CGU. Para a primeira análise, os métodos que obtiveram o melhor desempenho foram os modelos BERTimbau e DeBERTa-v3. Os dois métodos são baseados em modelos de língua pré-treinados, sendo adaptados para a etiquetagem morfossintática a partir da técnica de ajuste fino, obtendo desempenho superior aos etiquetadores do estado-da-arte para o português do Brasil. Além disso, o modelo BERTimbau possui menor custo computacional, dessa forma, este foi o modelo selecionado para a avaliação no contexto multigênero.

A avaliação multigênero abordou diferentes cenários de treinamento do modelo BERTimbau, considerando cenários isolados, onde o modelo foi treinado em apenas um gênero e no contexto multigênero, incluindo dois ou três gêneros. Nas avaliações, foi encontrado que o modelo multigênero, que abrange os três gêneros, alcançou desempenho similar aos cenários que realizaram o treinamento isoladamente em cada gênero. Com a execução de testes de hipótese, verificou-se que não há diferença significativa entre a abordagem multigênero e os modelos isolados, dessa forma, demonstrando a capacidade do etiquetador em manter o alto desempenho nos três gêneros. Assim, espera-se que o etiquetador possa ser utilizado em aplicações de PLN que utilizem diversos gêneros textuais, dessa forma, tornando-as mais robustas e versáteis.

5.3 Análise qualitativa de erros do etiquetador multigênero

Esta seção realiza a análise qualitativa dos erros gerados pelo etiquetador multigênero criado na seção anterior, ou seja, é utilizado o modelo treinado no contexto multigênero ”Porttinari-base + DANTEStocks + PetroGold”. São coletadas as inferências do modelo para todas as amostras do conjunto de testes de cada córpus, em seguida, os erros são destacados e analisados manualmente. Esta análise tem o objetivo de mapear os erros sistemáticos e, quando possível, levantar hipóteses que possam indicar direções para melhorias do etiquetador.

Em primeiro lugar, sabe-se que um dos fatores que contribui grandemente para a degradação do desempenho de etiquetadores morfossintáticos é a ocorrência de OOVs, consequentemente, é importante destacar a quantidades de erros que contém palavras fora de vocabulário. A Tabela 21 apresenta a contagem de etiquetas por classe gramatical para cada córpus da avaliação. Além disso, são apresentadas as quantidades de erros do tipo Falso Positivo (FP) e a quantidade de OOVs presente nesses erros. Os FPs ocorrem quando determinada etiqueta deveria aparecer, porém, o etiquetador classificou com outra etiqueta. Por exemplo, se uma etiqueta do tipo *ADJ* for classificada como *ADV*, isso será contabilizado como um erro FP na linha da etiqueta *ADJ* da

Tabela 21 – Número de ocorrências de etiquetas, falsos positivos e OOVs no conjunto de testes dos córpus Porttinari-base, DANTEStocks e PetroGold

	Porttinari-base			DANTEStocks			PetroGold		
	Ocorrências	Erros	OOVs	Ocorrências	Erros	OOVs	Ocorrências	Erros	OOVs
ADJ	1720	54	26	591	16	9	789	20	4
ADP	5000	11	0	1697	3	0	2088	3	0
ADV	1345	27	1	543	30	10	348	7	0
AUX	964	6	0	275	18	4	317	1	0
CCONJ	847	8	0	315	1	0	320	2	0
DET	4752	22	0	1345	14	0	1798	10	0
INTJ	8	1	1	36	7	5	0	0	0
NOUN	6165	57	20	2313	69	37	2860	47	10
NUM	595	11	0	904	5	4	353	5	3
PART	0	0	0	1	1	0	0	0	0
PRON	1280	17	0	259	8	0	258	8	0
PROPN	2186	22	8	2296	57	20	522	19	8
PUNCT	4566	1	0	2542	12	0	1421	2	0
SCONJ	467	27	0	131	9	0	78	3	0
SYM	96	0	0	829	3	2	36	12	0
VERB	3504	38	14	1277	23	9	981	9	4
X	53	20	15	409	31	15	7	1	0
Total	33548	322	85	15763	307	115	12176	149	29

Fonte: Elaborada pelo autor.

tabela. Na última linha da tabela são sumarizados os totais de ocorrências de etiquetas, erros e OOVs para cada córpus.

Em primeiro lugar, é possível observar que as etiquetas que possuem maior frequência, também possuem menor proporção de erros. Por exemplo, a etiqueta *NOUN* do córpus Porttinari-base possui 57 erros em 6.165 etiquetas, ou seja, menos de 1% das predições dessa classe foram erroneamente classificadas. Em contraposição, a etiqueta *X* do mesmo córpus, com total de 53 etiquetas e 20 erros, possui aproximadamente $37,74\%$ das ocorrências com classificações erradas. Contudo, isso não ocorre apenas por conta da frequência das etiquetas no córpus, por exemplo, a classe *SYM* não possui erros no córpus Porttinari-base, isso se dá por conta de ser uma etiqueta fácil de classificar. Isto não ocorre na classe *X*, por ser uma etiqueta utilizada pelos anotadores quando não foi possível enquadrar o *token* em outra categoria, naturalmente, é uma classe ambígua.

Analisando o número de OOVs para cada classe, é possível observar que algumas etiquetas possuem baixa ou nenhuma ocorrência de erros com OOVs e, em outros casos, possuem uma alta frequência deste fenômeno. Por exemplo, a classe *NOUN* do córpus DANTEStocks, possui cerca de $53,62\%$ *tokens* incorretamente classificados que são OOVs, ou seja, neste caso o modelo apresenta maior dificuldade de classificar substantivos quando são OOVs do que em outras classes, como na etiqueta *AUX* do mesmo córpus, que possui cerca de $22,22\%$ de erros que são OOVs.

Em seguida, foi realizada a análise manual dos erros de cada etiqueta para córpus, bus-

cando identificar erros sistemáticos e levantar hipóteses. As etiquetas que possuíram baixas ocorrências de erros, ou quando não foi possível identificar erros sistemáticos, não são mencionadas nesta análise. Por exemplo, as etiquetas *PUNCT, INTJ, PART* e *SYM* possuem baixa ocorrência de erros em todos os córpus.

Iniciando pela etiqueta *ADJ*, no córpus Porttinari-base foram encontrados 23 casos onde o *token* estavam na forma de particípio. Particípio é uma forma nominal do verbo e pode assumir as etiquetas *ADJ, NOUN* ou *VERB* e é um caso particularmente desafiador para os linguistas (DURAN, 2021). Naturalmente, o mesmo tipo de erro é encontrado ao analisar os erros das classes *NOUN* e *VERB*. Nos córpus DANTEStocks e PetroGold foram encontradas 4 ocorrências em ambos as análises.

Na etiqueta *ADV* foram encontrados 8 casos no córpus Porttinari onde o etiquetador classifica como uma conjunção (*CCONJ, SCONJ*). Este é um caso onde existe divergência entre linguistas, por exemplo, alguns definem ”nem” apenas como conjunção, ignorando a possibilidade de advérbio, contudo, outros defendem que ”nem” pode ser advérbio em alguns contextos. No córpus DANTEStocks e PetroGold não foram identificados casos iguais. Semelhantemente, nas etiquetas de conjunção *CCONJ* e *SCONJ* foram encontrados 15 *tokens* classificados como *ADV* no córpus Porttinari-base. Já no córpus DANTEStocks, foram encontradas 5 ocorrências desse erro.

A etiqueta *NOUN* possui o mesmo erro sistemático encontrado na etiqueta *ADJ* por conta do fenômeno linguístico particípio, em que ambas as etiquetas podem ser usadas para descrever a mesma palavra em diferentes contextos gramaticais. Foram encontradas 6 ocorrências no córpus Porttinari-base e poucos casos nos demais córpus. Além disso, nos córpus PetroGold e DANTEStocks, foram encontrados, respectivamente, 30 e 12 casos onde o modelo classificou como *PROPN* os *tokens* escritos em caixa alta ou com a inicial em maiúsculo.

Para a etiqueta *PROPN*, o fenômeno inverso é observado, onde é possível identificar casos onde o modelo classificou como *NOUN*. No córpus Porttinari-base foram 10 ocorrências, no DANTEStocks 21 e 8 casos no PetroGold. Em especial, no DANTEStocks foi observado que alguns *tweets* continham índices da bolsa de valores sendo classificados com a etiqueta *X*, ao total, foram encontradas 30 ocorrências. O córpus adotou a etiqueta *X* para índices da bolsa que não possuíam função linguística no *tweet* e, quando possuía função, a etiqueta *PROPN*. Dado o número de ocorrências encontrados, é possível observar que o modelo possui alguma dificuldade em realizar esta distinção para realizar a tarefa.

Em seguida, analisando os erros da etiqueta *VERB*, novamente encontramos o fenômeno do particípio e a dificuldade do modelo em desambiguá-lo. No córpus Porttinari-base, foram encontrados 22 casos onde o modelo classificou como *ADJ* ou *NOUN*. Nos outros córpus não foram encontrados casos frequentes.

A etiqueta *DET* possui *tokens* que, além de determinantes, podem ser pronomes e pre-

posições (*ADP*). O modelo apresenta 15 erros na identificação de determinantes que foram classificados como *PRON* no córpus Porttinari-base. No córpus DANTEStocks foram encontrados 6 casos da mesma categoria e 5 erros no PetroGold. Semelhantemente, são encontrados erros na etiqueta *PRON* com o rótulo *DET*.

Finalizando com a etiqueta *X*, que possui *tokens* ambíguos, no córpus Porttinari-base todos os erros encontrados foram casos de estrangeirismo que o modelo tentou associar uma classe gramatical diferente da etiqueta *X*. Este tipo de erro foi encontrado em 11 casos no córpus DANTEStocks e nenhum caso no córpus PetroGold. Não obstante, no córpus DANTEStocks foram encontrados 12 casos onde a etiqueta *PROPN* foi utilizada em *tokens* que representam índices da bolsa de valores.

O Quadro 6 apresenta exemplos de cada erro sistemático encontrado nesta análise. É indicado o córpus a qual a sentença ou *tweet* pertence, além disso, a classe esperada e o texto destacando em vermelho o *token* incorretamente classificado com a etiqueta gerada pelo modelo.

Quadro 6 – Exemplos de erros sistemáticos do etiquetador multigênero nos córpus Porttinari-base, DANTEStocks e PetroGold

Córpus	Classe esperada	Sentença ou *tweet*
Porttinari-base	*ADJ*	O atentado deixou outros 41 feridos(NOUN), muitos em estado grave.
DANTEStocks	*ADJ*	@ferrisss petr4 paradinho(NOUN) em os...
Porttinari-base	*ADV*	Gilmar não participou e nem(CCONJ) o ministro Marco...
Porttinari-base	*NOUN*	...La Mínima volta a os palcos com o premiado(ADJ) "A Noite de os...
PetroGold	*NOUN*	4.1. Pesquisa(PROPN) Bibliográfica
PetroGold	*PROPN*	3.3.3 Parâmetro(NOUN) de a Simulação
DANTEStocks	*PROPN*	...de Açúcar #GPA(X) #CBD(X) #PCAR3(X) #PCAR4(X) até...
Porttinari-base	*VERB*	O chamado(ADJ) "efeito Trump"transcende de longe as...
Porttinari-base	*DET*	Vai garantir dinheiro para campanha, que é o(PRON) mais importante...
Porttinari-base	*X*	O diplomata foi declarado persona(NOUN) non(ADV) grata(ADJ)...
DANTEStocks	*X*	Cadê a #PETRD15? #PETR4(PROPN) #IBOV(PROPN) petrobras

Fonte – Fonte: Elaborada pelo autor.

CAPÍTULO

6

CONCLUSÃO

Neste trabalho, foi realizada a investigação de métodos de etiquetagem morfossintática para o português do Brasil. Utilizando como base as diretrizes do projeto *Universal Dependencies* que possibilitam a padronização de *treebanks* para diversas línguas, dessa forma, alinhando este trabalho com um padrão internacionalmente aceito e estudado por pesquisadores da área. Após a avaliação e seleção do método de maior desempenho, foram realizados experimentos para averiguar a robustez do método em diferentes gêneros textuais e considerando o contexto multigênero.

Este trabalho baseou-se em duas hipóteses de pesquisa. A primeira é de que a etiquetagem morfossintática para o português do Brasil pode ser realizada com capacidades multigênero sem haver deterioração no desempenho individual de cada gênero. Em segundo lugar, hipotetizou-se que o uso do formalismo UD é suficiente para atingir resultados do estado da arte na tarefa de etiquetagem morfossintática.

Foram avaliados sete métodos diferentes de etiquetagem morfossintática para o português do Brasil com avaliação no córpus Porttinari-base. Os métodos baseados em modelos de língua BERTimbau e DeBERTa-v3 alcançaram desempenhos similares com acurácias ao nível de *tokens* de $99,07\%$ e $99,02\%$, respectivamente. Os testes estatísticos não apresentaram diferença significativa entre as duas abordagens, dessa forma, optou-se por utilizar o modelo BERTimbau que possui menor número de parâmetros. Esta avaliação permitiu validar a segunda hipótese de pesquisa, pois os resultados obtidos são similares aos encontrados nos trabalhos do estado da arte na literatura[38].

A partir do método BERTimbau, foi realizada a análise no contexto multigênero. Foram utilizados três córpus dos gêneros: jornalístico (Porttinari-base), acadêmico (PetroGold) e Conteúdo Gerado por Usuário (DANTEStocks). Foi constatado que o modelo multigênero

[38] O Apêndice C apesenta uma breve análise do uso do etiquetador BERTimbau em córpus que não utiliza o formalismo UD.

treinado nos três córpus obteve resultado similar e sem diferença significativa em relação aos desempenhos individuais em cada gênero. Dessa forma, o etiquetador BERTimbau consegue generalizar o aprendizado em diferentes córpus sem perda de desempenho significativa. A partir desta análise foi possível averiguar a primeira hipótese de pesquisa.

A verificação de que o padrão UD é suficiente para atingir alto desempenho na tarefa de etiquetagem morfossintática para o português do Brasil corrobora com outros trabalhos da área, sendo assim, ressaltando a relevância e robustez das diretrizes da UD na tarefa de etiquetagem. Além disso, a investigação de diferentes abordagens demonstrou o alto desempenho dos métodos baseados em modelos de língua para a etiquetagem em português do Brasil no gênero jornalístico. No geral, estes métodos apresentaram maior desempenho em sentenças que possuem palavras fora de vocabulário do que os métodos baseados em RNRs, onde, em média, apresentam acurácias $2,44\%$ maiores do que as RNRs.

Apesar do alto desempenho no contexto multigênero do método BERTimbau, é importante ressaltar que o etiquetador possivelmente terá variação de desempenho ao processar sentenças de outros córpus ou gêneros textuais. Ou seja, o estudo realizado apresenta alto desempenho nos gêneros e córpus analisados e isso não garante que o método apresentará o mesmo desempenho em outros contextos. Para a utilização deste método em outros córpus e gêneros, recomenda-se a avaliação do desempenho do etiquetador e possível ajuste fino incluindo o córpus alvo.

Ambos os métodos baseados em modelos de língua e RNRs possuem maior dificuldade em sentenças com 10 *tokens* ou menos, onde se observa uma queda média de $1,5\%$ na acurácia do córpus Porttinari-base. Contudo, é importante ressaltar que estas sentenças possuem em média $2,03\%$ de OOVs a mais do que sentenças maiores. Não obstante, esta queda pode impactar aplicações em gêneros textuais que possuem muitas sentenças com poucos *tokens*, como no caso de *tweets* em córpus CGU. Para mitigação deste problema, uma série de técnicas podem ser utilizadas, como o uso de aumentação de dados para criação de amostras artificiais, a realização do ajuste fino ou pré-treinamento do modelo de língua em córpus com textos curtos, até a coleta e anotação de novas sentenças.

Outro fator importante a ser considerado é o custo computacional do modelo BERTimbau. O etiquetador possui cerca de 110 milhões de parâmetros e o mecanismo de auto-atenção possui complexidade assintótica quadrática. Dessa forma, o custo de experimentação pode ser elevado e, em alguns casos, proibitivo. Além disso, o tempo de inferência pode ser elevado em processadores limitados. Este trabalho utilizou modelos de língua pré-treinados e pode ser necessário realizar a etapa de pré-treinamento a depender do córpus alvo, contudo, o pré-treinamento é acessível apenas para grandes corporações e poucos laboratórios de pesquisa no mundo. Sendo assim, recomenda-se utilizar modelos de língua já disponíveis na literatura para experimentação com ajuste fino. Adicionalmente, é possível utilizar técnicas de compressão de modelos para reduzir o tamanho e tempo de inferência, como a quantização e poda de rede neural.

Em resumo, este trabalho investigou diferentes métodos de etiquetagem morfossintática para o português do Brasil, utilizando como base as diretrizes do projeto *Universal Dependencies* (UD), e verificou que o uso do formalismo UD é suficiente para atingir resultados do estado da arte na tarefa de etiquetagem morfossintática. O método BERTimbau, apresentou alto desempenho no aprendizado multigênero de textos jornalísticos, acadêmicos e CGU. Em geral, este trabalho contribui para o avanço da pesquisa em etiquetagem morfossintática para o português do Brasil, oferece uma base para estudos futuros na área e disponibiliza o etiquetador morfossintática para a comunidade científica, assim como o repositório[39] de código para a reprodução dos resultados. Além disso, uma aplicação[40] de demonstração do etiquetador foi criada para que o público geral poder etiquetar sentenças e salvar os resultados.

[39] Disponível em: <https://github.com/huberemanuel/porttagger>.
[40] Disponível em: <https://huggingface.co/spaces/Emanuel/porttagger>.

REFERÊNCIAS

ACCUOSTO, P.; NEVES, M.; SAGGION, H. Argumentation mining in scientific literature: From computational linguistics to biomedicine. In: CEUR WORKSHOP PROCEEDINGS. **Frommholz I, Mayr P, Cabanac G, Verberne S, editors. BIR 2021: 11th International Workshop on Bibliometric-enhanced Information Retrieval; 2021 Apr 1; Lucca, Italy. Aachen: CEUR; 2021. p. 20-36**. [S.l.], 2021. Citado na página 25.

AFONSO, S.; BICK, E.; HABER, R.; SANTOS, D. Floresta sintá(c)tica: A treebank for Portuguese. In: **Proceedings of the Third International Conference on Language Resources and Evaluation**. Las Palmas, Canary Island, Spain: [s.n.], 2002. Citado nas páginas 24, 27 e 53.

AIRES, R. V. X. **Implementação, adaptação, combinação e avaliação de etiquetadores para o português do Brasil**. Tese (Doutorado) — Universidade de São Paulo, São Carlos, SP, Brasil, 2000. Citado na página 51.

AIRES, R. V. X.; ALUíSIO, S. M.; KUHN, D. C. S.; ANDREETA, M. L. B.; JR., O. N. O. Combining multiple classifiers to improve part of speech tagging: A case study for brazilian portuguese. **Brazilian Symposium on Artificial Intelligence**, 2000. Citado nas páginas 50, 53, 54 e 56.

ALAMMAR, J. **The Illustrated Transformer**. 2018. <https://jalammar.github.io/illustrated-transformer/>. Acesso em: 28 de fevereiro de 2022. Citado na página 38.

ALBOGAMY, F.; RAMSAY, A. Universal Dependencies for Arabic tweets. In: **Proceedings of the International Conference Recent Advances in Natural Language Processing**. Varna, Bulgaria: [s.n.], 2017. p. 46–51. Citado na página 55.

ALONSO, H. M.; SEDDAH, D.; SAGOT, B. From noisy questions to Minecraft texts: Annotation challenges in extreme syntax scenario. In: **Proceedings of the 2nd Workshop on Noisy User-generated Text**. Osaka, Japan: [s.n.], 2016. p. 13–23. Citado na página 55.

ALUÍSIO, S.; PELIZZONI, J.; MARCHI, A. R.; OLIVEIRA, L. de; MANENTI, R.; MARQUIAFáVEL, V. An account of the challenge of tagging a reference corpus for brazilian portuguese. In: **6th international conference on Computational processing of the Portuguese language**. [S.l.: s.n.], 2003. p. 110–117. Citado nas páginas 50, 54 e 56.

ALUÍSIO, S.; PELIZZONI, J.; MARCHI, A. R.; OLIVEIRA, L. de; MANENTI, R.; MARQUIAFÁVEL, V. An account of the challenge of tagging a reference corpus for brazilian portuguese. In: MAMEDE, N. J.; TRANCOSO, I.; BAPTISTA, J.; NUNES, M. das G. V. (Ed.). **Computational Processing of the Portuguese Language**. Berlin, Heidelberg: [s.n.], 2003. p. 110–117. Citado na página 135.

BAHDANAU, D.; CHO, K.; BENGIO, Y. Neural machine translation by jointly learning to align and translate. **ArXiv**, 2016. Citado nas páginas 34 e 35.

BEHZAD, S.; ZELDES, A. A cross-genre ensemble approach to robust Reddit part of speech tagging. In: **Proceedings of the 12th Web as Corpus Workshop**. Marseille, France: [s.n.], 2020. p. 50–56. Citado na página 57.

BELISáRIO, L.; FERREIRA, L.; PARDO, T. Evaluating richer features and varied machine learning models for subjectivity classification of book review sentences in portuguese. **Information**, p. 1–14, 2020. Citado na página 54.

BHAT, I.; BHAT, R. A.; SHRIVASTAVA, M.; SHARMA, D. Universal Dependency parsing for Hindi-English code-switching. In: **Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)**. New Orleans, Louisiana: [s.n.], 2018. p. 987–998. Citado nas páginas 55 e 56.

BLODGETT, S. L.; WEI, J.; O'CONNOR, B. Twitter Universal Dependency parsing for African-American and mainstream American English. In: **Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**. Melbourne, Australia: [s.n.], 2018. p. 1415–1425. Citado na página 55.

BOHNET, B.; MCDONALD, R.; SIMÕES, G.; ANDOR, D.; PITLER, E.; MAYNEZ, J. Morphosyntactic tagging with a meta-BiLSTM model over context sensitive token encodings. In: **Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Long Papers)**. Melbourne, Australia: [s.n.], 2018. p. 2642–2652. Citado nas páginas 32, 49, 56, 65 e 70.

BRANCO, A.; SILVA, J. R. Evaluating solutions for the rapid development of state-of-the-art pos taggers for portuguese. In: **Proceedings of the 4th International Conference on Language Resources and Evaluation**. Lisboa: [s.n.], 2004. p. 507–510. Citado na página 51.

BRANCO, A.; SILVA, J. R.; GOMES, L.; RODRIGUES, J. A. Universal grammatical dependencies for Portuguese with CINTIL data, LX processing and CLARIN support. In: **Proceedings of the Thirteenth Language Resources and Evaluation Conference**. Marseille, France: [s.n.], 2022. p. 5617–5626. Citado na página 60.

BRILL, E. A simple rule-based part of speech tagger. In: **Third conference on Applied natural language processing**. USA: [s.n.], 1992. p. 152–155. Citado nas páginas 46, 53 e 54.

____. Transformation-based error-driven learning and natural language processing: A case study in part-of-speech tagging. **Computational Linguistics**, p. 543–565, 1995. Citado nas páginas 34, 50 e 51.

BRYSON, A.; HO, Y. **Applied Optimal Control: Optimization, Estimation, and Control**. Blaisdell Publishing Company, 1969. (Blaisdell book in the pure and applied sciences). Disponível em: <https://books.google.com.br/books?id=k_FQAAAAMAAJ>. Citado na página 34.

CIGNARELLA, A. T.; BOSCO, C.; ROSSO, P. Presenting TWITTIRÒ-UD: An Italian Twitter treebank in Universal Dependencies. In: **Proceedings of the Fifth International Conference on Dependency Linguistics (Depling, SyntaxFest 2019)**. Paris, France: [s.n.], 2019. p. 190–197. Citado na página 55.

CINTIL Corpus Internacional do Português. Linguistics Centre, University of Lisbon, 2011. Disponível em: <http://cintil.ul.pt>. Citado na página 56.

CONNEAU, A.; KHANDELWAL, K.; GOYAL, N.; CHAUDHARY, V.; WENZEK, G.; GUZMÁN, F.; GRAVE, E.; OTT, M.; ZETTLEMOYER, L.; STOYANOV, V. Unsupervised cross-lingual representation learning at scale. **ArXiv**, 2019. Citado na página 63.

CONNEAU, A.; RINOTT, R.; LAMPLE, G.; WILLIAMS, A.; BOWMAN, S. R.; SCHWENK, H.; STOYANOV, V. Xnli: Evaluating cross-lingual sentence representations. In: **Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing**. [S.l.: s.n.], 2018. Citado na página 64.

CREŢULESCU, R.; DAVID, A.; MORARIU, D.; VINŢAN, L. Part of speech tagging with naïve bayes methods. In: **18th International Conference on System Theory, Control and Computing**. Sinaia, Romania: [s.n.], 2014. p. 446–451. Citado na página 34.

CUI, L.; ZHANG, Y. Hierarchically-refined label attention network for sequence labeling. In: **Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing**. Hong Kong, China: [s.n.], 2019. p. 4115–4128. Citado nas páginas 48 e 56.

DAIBER, J.; GOOT, R. van der. The denoised web treebank: Evaluating dependency parsing under noisy input conditions. In: **Proceedings of the Tenth International Conference on Language Resources and Evaluation**. Portorož, Slovenia: [s.n.], 2016. p. 649–653. Citado na página 55.

DEVLIN, J.; CHANG, M.-W.; LEE, K.; TOUTANOVA, K. BERT: Pre-training of deep bidirectional transformers for language understanding. In: **Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)**. Minneapolis, Minnesota: [s.n.], 2019. p. 4171–4186. Citado nas páginas 39, 49, 50, 51, 52, 63 e 64.

Di FELIPPO, A.; POSTALI, C.; CEREGATTO, G.; GAZANA, L.; SILVA, E.; ROMAN, N.; PARDO, T. Descrição preliminar do corpus dantestocks: Diretrizes de segmentação para anotação segundo universal dependencies. In: **Anais do XIII Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana**. Porto Alegre, RS, Brasil: [s.n.], 2021. p. 335–343. Citado nas páginas 24, 54 e 60.

DOMINGOS, P. The role of occam's razor in knowledge discovery. **Data mining and knowledge discovery**, p. 409–425, 1999. Citado na página 79.

DOMINGUES, M. L. C. S. **Abordagem para o desenvolvimento de um etiquetador de alta acurácia para o Português do Brasil**. Tese (Doutorado) — Universidade Federal do Pará, Belém, PA, Brasil, 2011. Citado nas páginas 34, 50, 53, 54, 56 e 57.

DOZAT, T.; MANNING, C. D. Deep biaffine attention for neural dependency parsing. **ArXiv**, 2016. Citado nas páginas 48 e 65.

DURAN, M. S. **MANUAL DE ANOTAÇÃO DE POS TAGS**. [S.l.], 2021. Citado nas páginas 33, 80 e 92.

EKBAL, A.; HAQUE, R.; BANDYOPADHYAY, S. Bengali part of speech tagging using conditional random field. In: **Proceedings of Seventh International Symposium on Natural Language Processing**. Pattaya, Chonburi: [s.n.], 2007. p. 67–78. Citado nas páginas 46, 51 e 56.

ELMAN, J. L. Finding structure in time. **Cognitive Science**, p. 179–211, 1990. Citado na página 36.

FACELI, K.; LORENA, A. C.; GAMA, J.; ALMEIDA, T. A.; CARVALHO, A. C. P. L. F. **Inteligência artificial: uma abordagem de aprendizado de máquina**. [S.l.]: Grupo Gen - LTC, 2021. Citado na página 34.

FILHO, J. A. W.; WILKENS, R.; IDIART, M.; VILLAVICENCIO, A. The brwac corpus: a new open resource for brazilian portuguese. In: **Proceedings of the eleventh international conference on language resources and evaluation (LREC 2018)**. [S.l.: s.n.], 2018. Citado na página 64.

FISHER, R. A. **Statistical Methods for Research Workers**. New York, NY: Springer New York, 1992. Citado nas páginas 66 e 67.

FONSECA, E. R.; ROSA, J. L. G. Mac-morpho revisited: Towards robust part-of-speech tagging. In: **Proceedings of the 9th Brazilian Symposium in Information and Human Language Technology**. [S.l.: s.n.], 2013. Citado nas páginas 135 e 136.

FONSECA, E. R.; ROSA, J. L. G.; ALUÍSIO, S. M. Evaluating word embeddings and a revised corpus for part-of-speech tagging in portuguese. **Journal of the Brazilian Computer Society**, 2015. Citado na página 136.

FOSTER, J.; ÇETINOĞLU, Ö.; WAGNER, J.; ROUX, J. L.; NIVRE, J.; HOGAN, D.; GENABITH, J. van. From news to comment: Resources and benchmarks for parsing the language of web 2.0. In: **Proceedings of 5th International Joint Conference on Natural Language Processing**. Chiang Mai, Thailand: [s.n.], 2011. p. 893–901. Citado na página 55.

FRANCIS, W. N.; KUCERA, H. Computational analysis of present-day american english. **Brown University Press**, 1967. Citado na página 53.

____. **Brown Corpus Manual**. [S.l.], 1979. Citado nas páginas 24, 25, 27 e 46.

GARIMELLA, A.; BANEA, C.; HOVY, D.; MIHALCEA, R. Women's syntactic resilience and men's grammatical luck: Gender-bias in part-of-speech tagging and dependency parsing. In: **Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics**. [S.l.: s.n.], 2019. p. 3493–3498. Citado na página 26.

GOODFELLOW, I.; BENGIO, Y.; COURVILLE, A. **Deep Learning**. [S.l.]: MIT Press, 2016. Citado na página 35.

GOOT, R. van der; NOORD, G. van. Modeling input uncertainty in neural network dependency parsing. In: **Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing**. Brussels, Belgium: [s.n.], 2018. p. 4984–4991. Citado na página 55.

HARTMANN, N. S.; FONSECA, E. R.; SHULBY, C. D.; TREVISO, M. V.; RODRIGUES, J. S.; ALUíSIO, S. M. Portuguese word embeddings: Evaluating on word analogies and natural language tasks. In: **Anais do XI Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana**. Porto Alegre, RS, Brasil: SBC, 2017. p. 122–131. Citado na página 70.

HE, K.; ZHANG, X.; REN, S.; SUN, J. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. **2015 IEEE International Conference on Computer Vision**, p. 1026–1034, 2015. Citado na página 35.

HE, P.; GAO, J.; CHEN, W. Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing. **ArXiv**, 2021. Citado na página 63.

HEINZERLING, B.; STRUBE, M. BPEmb: Tokenization-free pre-trained subword embeddings in 275 languages. In: **Proceedings of the Eleventh International Conference on Language Resources and Evaluation**. Miyazaki, Japan: [s.n.], 2018. Citado na página 49.

____. Sequence tagging with contextual and non-contextual subword representations: A multilingual evaluation. **Computing Research Repository**, 2019. Citado nas páginas 48, 56, 57, 65 e 70.

JOULIN, A.; GRAVE, E.; BOJANOWSKI, P.; DOUZE, M.; JÉGOU, H.; MIKOLOV, T. Fasttext.zip: Compressing text classification models. **ArXiv**, 2016. Citado na página 47.

JURAFSKY, D.; MARTIN, J. H. **Speech and language processing : an introduction to natural language processing, computational linguistics, and speech recognition**. [S.l.]: Pearson Prentice Hall, 2009. Citado nas páginas 23, 28, 31, 32 e 35.

KALJAHI, R.; FOSTER, J.; ROTURIER, J.; RIBEYRE, C.; LYNN, T.; ROUX, J. L. Foreebank: Syntactic analysis of customer support forums. In: **Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing**. Lisbon, Portugal: [s.n.], 2015. p. 1341–1347. Citado na página 55.

KAYID, A.; KHALED, Y.; ELMAHDY, M. Performance of cpus/gpus for deep learning workloads. **The German University in Cairo**, 2018. Citado na página 35.

KEMPE, A. **A Probabilistic Tagger and an Analysis of Tagging Errors**. [S.l.], 1993. Citado nas páginas 46, 51 e 56.

KEPLER, F. N. **Modelagem de contextos para aprendizado automático aplicado à Análise Morfossintática**. Tese (Doutorado) — Instituto de Matemática e Estatística da Universidade de São Paulo, São Paulo, SP, Brasil, 2010. Citado nas páginas 28, 32, 36 e 51.

KIM, Y. Convolutional neural networks for sentence classification. In: **Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing**. Doha, Qatar: [s.n.], 2014. p. 1746–1751. Citado na página 35.

KLEIN, S.; SIMMONS, R. F. A computational approach to grammatical coding of english words. **Association for Computing Machinery**, p. 334–347, 1963. Citado nas páginas 32, 45 e 51.

KONDRATYUK, D.; STRAKA, M. 75 languages, 1 model: Parsing Universal Dependencies universally. In: **Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing**. Hong Kong, China: [s.n.], 2019. p. 2779–2795. Citado nas páginas 47, 56 e 60.

KONG, L.; SCHNEIDER, N.; SWAYAMDIPTA, S.; BHATIA, A.; DYER, C.; SMITH, N. A. A dependency parser for tweets. In: **Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing**. Doha, Qatar: [s.n.], 2014. p. 1001–1012. Citado nas páginas 51 e 55.

KULICK, S.; BIES, A.; LIBERMAN, M.; MANDEL, M.; MCDONALD, R.; PALMER, M.; SCHEIN, A.; UNGAR, L.; WINTERS, S.; WHITE, P. Integrated annotation for biomedical information extraction. In: **HLT-NAACL 2004 Workshop: Linking Biological Literature,**

Ontologies and Databases. Boston, Massachusetts: [s.n.], 2004. p. 61–68. Citado na página 53.

LE, Q.; MIKOLOV, T. Distributed representations of sentences and documents. In: **Proceedings of the 31st International Conference on Machine Learning**. Bejing, China: [s.n.], 2014. p. 1188–1196. Citado na página 36.

LECUN, Y.; HAFFNER, P.; BOTTOU, L.; BENGIO, Y. Object recognition with gradient-based learning. In: ____. **Shape, Contour and Grouping in Computer Vision**. [S.l.]: Springer Berlin Heidelberg, 1999. p. 319–345. Citado na página 34.

LIN, Y.; WANG, C.; SONG, H.; LI, Y. Multi-head self-attention transformation networks for aspect-based sentiment analysis. **IEEE Access**, p. 8762–8770, 2021. Citado na página 26.

LINGUATECA. **Floresta Sintá(c)tica**. 2009. <https://www.linguateca.pt/Floresta/corpus.html>. Acesso em: 22 de dezembro de 2022. Citado nas páginas 25 e 50.

LIU, H. Sentiment analysis of citations using word2vec. **ArXiv**, 2017. Citado na página 36.

LIU, Y.; OTT, M.; GOYAL, N.; DU, J.; JOSHI, M.; CHEN, D.; LEVY, O.; LEWIS, M.; ZETTLEMOYER, L.; STOYANOV, V. Roberta: A robustly optimized bert pretraining approach. **ArXiv**, 2019. Citado na página 63.

LIU, Y.; ZHU, Y.; CHE, W.; QIN, B.; SCHNEIDER, N.; SMITH, N. A. Parsing tweets into Universal Dependencies. In: **Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)**. New Orleans, Louisiana: [s.n.], 2018. p. 965–975. Citado nas páginas 52, 54 e 55.

LOPES, L.; DURAN, M. S.; NUNES, M. D. G. V.; PARDO, T. A. S. **CORPORA BUILDING PROCESS ACCORDING TO THE UNIVERSAL DEPENDENCIES MODEL: AN EXPERIMENT FOR PORTUGUESE**. [S.l.], 2022. Citado nas páginas 60 e 61.

LOSHCHILOV, I.; HUTTER, F. Decoupled weight decay regularization. In: **International Conference on Learning Representations**. [S.l.: s.n.], 2017. Citado na página 70.

LUOTOLAHTI, J.; KANERVA, J.; LAIPPALA, V.; PYYSALO, S.; GINTER, F. Towards universal web parsebanks. In: **Proceedings of the Third International Conference on Dependency Linguistics (Depling 2015)**. Uppsala, Sweden: [s.n.], 2015. p. 211–220. Citado na página 55.

MANNING, C. D. Part-of-speech tagging from 97% to 100%: is it time for some linguistics? In: **International conference on intelligent text processing and computational linguistics**. [S.l.: s.n.], 2011. p. 171–189. Citado na página 73.

MANNING, C. D.; SCHÜTZE, H. **Foundations of Statistical Natural Language Processing**. [S.l.]: The MIT Press, 1999. Citado na página 33.

MARCUS, M. P.; SANTORINI, B.; MARCINKIEWICZ, M. A. Building a large annotated corpus of English: The Penn Treebank. **Computational Linguistics**, p. 313–330, 1993. Citado na página 46.

____. Building a large annotated corpus of English: The Penn Treebank. **Computational Linguistics**, 1993. Citado nas páginas 53 e 60.

MARRESE-TAYLOR, E.; RODRIGUEZ, C.; BALAZS, J.; GOULD, S.; MATSUO, Y. A multimodal approach to fine-grained opinion mining on video reviews. In: **Second Grand-Challenge and Workshop on Multimodal Language**. Seattle, USA: [s.n.], 2020. p. 8–18. Citado na página 24.

MCCULLOCH, W.; PITTS, W. A logical calculus of ideas immanent in nervous activity. **The bulletin of mathematical biophysics**, p. 115–133, 1943. Citado na página 34.

MCDONALD, R.; NIVRE, J.; QUIRMBACH-BRUNDAGE, Y.; GOLDBERG, Y.; DAS, D.; GANCHEV, K.; HALL, K.; PETROV, S.; ZHANG, H.; TÄCKSTRÖM, O.; BEDINI, C.; CASTELLÓ, N. B.; LEE, J. Universal Dependency annotation for multilingual parsing. In: **Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)**. Sofia, Bulgaria: [s.n.], 2013. p. 92–97. Citado na página 60.

MIKOLOV, T.; CHEN, K.; CORRADO, G.; DEAN, J. Efficient estimation of word representations in vector space. In: **International Conference on Learning Workshop Papers**. Scottsdale, AZ, USA: [s.n.], 2013. Citado nas páginas 35 e 36.

MIKOLOV, T.; SUTSKEVER, I.; CHEN, K.; CORRADO, G.; DEAN, J. Distributed representations of words and phrases and their compositionality. In: **Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2**. Red Hook, NY, USA: Curran Associates Inc., 2013. p. 3111–3119. Citado na página 70.

MITCHELL, T. M. **Machine Learning**. [S.l.]: McGraw-Hill, 1997. Citado na página 34.

MONTEIRO, R. A.; SANTOS, R. L. S.; PARDO, T. A. S.; ALMEIDA, T. A. de; RUIZ, E. E. S.; VALE, O. A. Contributions to the study of fake news in portuguese: New corpus and automatic detection results. In: **Computational Processing of the Portuguese Language**. Cham: [s.n.], 2018. p. 324–334. Citado na página 59.

NGUYEN, D. Q.; VU, T.; NGUYEN, A. T. BERTweet: A pre-trained language model for English tweets. In: **Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations**. Online: [s.n.], 2020. p. 9–14. Citado nas páginas 51, 56, 63 e 64.

NIVRE, J.; MARNEFFE, M.-C. de; GINTER, F.; GOLDBERG, Y.; HAJIČ, J.; MANNING, C. D.; MCDONALD, R.; PETROV, S.; PYYSALO, S.; SILVEIRA, N.; TSARFATY, R.; ZEMAN, D. Universal Dependencies v1: A multilingual treebank collection. In: **Proceedings of the Tenth International Conference on Language Resources and Evaluation**. Portorož, Slovenia: [s.n.], 2016. p. 1659–1666. Citado nas páginas 27 e 32.

NOMENCLATURA Gramatical Brasileira. Rio de Janeiro: [s.n.], 1959. Citado na página 32.

OHTA, T.; TATEISI, Y.; KIM, J.-D. The genia corpus: An annotated research abstract corpus in molecular biology domain. In: **Proceedings of the Second International Conference on Human Language Technology Research**. San Francisco, USA: [s.n.], 2002. p. 82—-86. Citado na página 53.

OWOPUTI, O.; O'CONNOR, B.; DYER, C.; GIMPEL, K.; SCHNEIDER, N.; SMITH, N. A. Improved part-of-speech tagging for online conversational text with word clusters. In: **Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies**. Atlanta, Georgia: [s.n.], 2013. p. 380–390. Citado na página 54.

PAMAY, T.; SULUBACAK, U.; TORUNOĞLU-SELAMET, D.; ERYIĞIT, G. The annotation process of the ITU web treebank. In: **Proceedings of The 9th Linguistic Annotation Workshop**. Denver, Colorado, USA: [s.n.], 2015. p. 95–101. Citado na página 55.

PARDO, T.; DURAN, M.; LOPES, L.; FELIPPO, A.; ROMAN, N.; NUNES, M. Porttinari - a large multi-genre treebank for brazilian portuguese. In: **Anais do XIII Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana**. Porto Alegre, RS, Brasil: [s.n.], 2021. p. 1–10. Citado na página 54.

PETERS, M. E.; NEUMANN, M.; IYYER, M.; GARDNER, M.; CLARK, C.; LEE, K.; ZETTLEMOYER, L. Deep contextualized word representations. In: **Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)**. New Orleans, Louisiana: [s.n.], 2018. p. 2227–2237. Citado na página 52.

QI, P.; ZHANG, Y.; ZHANG, Y.; BOLTON, J.; MANNING, C. D. Stanza: A Python natural language processing toolkit for many human languages. In: **Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations**. [s.n.], 2020. Disponível em: <https://nlp.stanford.edu/pubs/qi2020stanza.pdf>. Citado nas páginas 48, 56, 65 e 69.

RADEMAKER, A.; CHALUB, F.; REAL, L.; FREITAS, C.; BICK, E.; PAIVA, V. de. Universal dependencies for portuguese. In: **Proceedings of the Fourth International Conference on Dependency Linguistics (Depling)**. Pisa, Italy: [s.n.], 2017. p. 197–206. Citado nas páginas 13, 27, 32, 40, 41, 47, 53, 54 e 60.

REAL, L.; OSHIRO, M.; MAFRA, A. B2w-reviews01 - an open product reviews corpus. In: **XII Symposium in Information and Human Language Technology**. Salvador, BA, Brasil: [s.n.], 2019. p. 200–208. Citado na página 54.

REHBEIN, I.; RUPPENHOFER, J.; DO, B.-N. tweeDe – a Universal Dependencies treebank for German tweets. In: **Proceedings of the 18th International Workshop on Treebanks and Linguistic Theories (TLT, SyntaxFest 2019)**. Paris, France: [s.n.], 2019. p. 100–108. Citado na página 55.

ROSENBLATT, F. The perceptron: a probabilistic model for information storage and organization in the brain. **Psychological review**, p. 386–408, 1958. Citado nas páginas 13, 34 e 35.

SANGUINETTI, M.; BOSCO, C.; CASSIDY, L.; CETINOGLU, Ö.; CIGNARELLA, A. T.; LYNN, T.; REHBEIN, I.; RUPPENHOFER, J.; SEDDAH, D.; ZELDES, A. Treebanking user-generated content: A proposal for a unified representation in Universal Dependencies. In: **Proceedings of the 12th Language Resources and Evaluation Conference**. Marseille, France: [s.n.], 2020. p. 5240–5250. Citado na página 55.

SANGUINETTI, M.; BOSCO, C.; LAVELLI, A.; MAZZEI, A.; ANTONELLI, O.; TAMBURINI, F. PoSTWITA-UD: an Italian Twitter treebank in Universal Dependencies. In: **Proceedings of the Eleventh International Conference on Language Resources and Evaluation**. Miyazaki, Japan: [s.n.], 2018. Citado nas páginas 54 e 55.

SANTANA, M. **Kaggle – news of the brazilian newspaper**. 2019. Acesso em: 15 de março de 2022. Citado na página 61.

SANTOS, C. D.; ZADROZNY, B. Learning character-level representations for part-of-speech tagging. In: **Proceedings of the 31st International Conference on Machine Learning**. Bejing, China: [s.n.], 2014. p. 1818–1826. Citado na página 136.

SCHUSTER, M.; PALIWAL, K. Bidirectional recurrent neural networks. **IEEE Transactions on Signal Processing**, p. 2673–2681, 1997. Citado na página 36.

SEDDAH, D.; ESSAIDI, F.; FETHI, A.; FUTERAL, M.; MULLER, B.; SUÁREZ, P. J. O.; SAGOT, B.; SRIVASTAVA, A. Building a user-generated content North-African Arabizi treebank: Tackling hell. In: **Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics**. Online: [s.n.], 2020. p. 1139–1150. Citado na página 55.

SEDDAH, D.; SAGOT, B.; CANDITO, M.; MOUILLERON, V.; COMBET, V. The French Social Media Bank: a treebank of noisy user generated content. In: **24th International Conference on Computational Linguistics**. Mumbai, India: [s.n.], 2012. p. 2441–2458. Citado na página 55.

SENNRICH, R.; HADDOW, B.; BIRCH, A. Neural machine translation of rare words with subword units. In: **Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**. Berlin, Germany: [s.n.], 2016. p. 1715–1725. Citado na página 51.

SILVA, F.; ROMAN, N.; CARVALHO, A. Stock market tweets annotated with emotions. **Corpora**, p. 343–354, 2020. Citado nas páginas 54 e 60.

SILVEIRA, N.; DOZAT, T.; MARNEFFE, M.-C. de; BOWMAN, S.; CONNOR, M.; BAUER, J.; MANNING, C. A gold standard dependency corpus for English. In: **Proceedings of the Ninth International Conference on Language Resources and Evaluation**. Reykjavik, Iceland: [s.n.], 2014. p. 2897–2904. Citado nas páginas 32, 51 e 55.

SOUSA, R. C. C. de; LOPES, H. Portuguese pos tagging using blstm without handcrafted features. In: NYSTRÖM, I.; HEREDIA, Y. H.; NÚÑEZ, V. M. (Ed.). **Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications**. Cham: [s.n.], 2019. p. 120–130. Citado nas páginas 50, 51, 53, 54, 56 e 57.

____. Portuguese pos tagging using blstm without handcrafted features. In: NYSTRÖM, I.; HEREDIA, Y. H.; NÚÑEZ, V. M. (Ed.). **Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications**. Cham: [s.n.], 2019. p. 120–130. Citado na página 136.

SOUZA, E.; SILVEIRA, A.; CAVALCANTI, T.; CASTRO, M.; FREITAS, C. Petrogold – corpus padrão ouro para o domínio do petróleo. In: **Anais do XIII Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana**. Porto Alegre, Brasil: [s.n.], 2021. p. 29–38. Citado nas páginas 55, 60 e 61.

SOUZA, E. de; FREITAS, C. Polishing the gold–how much revision do we need in treebanks? In: **Procedings of the Universal Dependencies Brazilian Festival**. [S.l.: s.n.], 2022. p. 1–11. Citado na página 25.

SOUZA, F.; NOGUEIRA, R.; LOTUFO, R. Bertimbau: Pretrained bert models for brazilian portuguese. In: CERRI, R.; PRATI, R. C. (Ed.). **Intelligent Systems**. Cham: [s.n.], 2020. p. 403–417. Citado nas páginas 59, 64 e 69.

STRAKA, M. UDPipe 2.0 prototype at CoNLL 2018 UD shared task. In: **Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies**. Brussels, Belgium: [s.n.], 2018. p. 197–207. Citado nas páginas 47, 56, 60, 65 e 69.

STRAKA, M.; HAJIČ, J.; STRAKOVÁ, J. UDPipe: Trainable pipeline for processing CoNLL-U files performing tokenization, morphological analysis, POS tagging and parsing. In: **Proceedings of the Tenth International Conference on Language Resources and Evaluation**. Portorož, Slovenia: [s.n.], 2016. p. 4290–4297. Citado na página 47.

STRAKA, M.; STRAKOVÁ, J.; HAJIC, J. UDPipe at SIGMORPHON 2019: Contextualized embeddings, regularization with morphological categories, corpora merging. In: **Proceedings of the 16th Workshop on Computational Research in Phonetics, Phonology, and Morphology**. Florence, Italy: [s.n.], 2019. p. 95–103. Citado na página 47.

SUNKARA, V. **A DATA DRIVEN APPROACH TO IDENTIFY JOURNALISTIC 5WS FROM TEXT DOCUMENTS**. Dissertação (Mestrado) — University of Nebraska, Lincoln, Nebraska, 2019. Citado na página 25.

TSURUOKA, Y.; TATEISHI, Y.; KIM, J.-D.; OHTA, T.; MCNAUGHT, J.; ANANIADOU, S.; TSUJII, J. Developing a robust part-of-speech tagger for biomedical text. In: **Panhellenic conference on informatics**. Edmonton, Canada: [s.n.], 2005. p. 382–392. Citado nas páginas 52, 53, 56 e 57.

TUKEY, J. W. Comparing individual means in the analysis of variance. **Biometrics**, 1949. Citado nas páginas 66 e 67.

VASWANI, A.; SHAZEER, N.; PARMAR, N.; USZKOREIT, J.; JONES, L.; GOMEZ, A. N.; KAISER, L. u.; POLOSUKHIN, I. Attention is all you need. In: **Advances in Neural Information Processing Systems**. Long Beach, CA, USA: [s.n.], 2017. p. 5998–6008. Citado nas páginas 34, 35, 37 e 39.

VIRTANEN, A.; KANERVA, J.; ILO, R.; LUOMA, J.; LUOTOLAHTI, J.; SALAKOSKI, T.; GINTER, F.; PYYSALO, S. Multilingual is not enough: Bert for finnish. **ArXiv**, 2019. Citado nas páginas 50, 52, 56, 57, 63 e 64.

VRIES, W. de; CRANENBURGH, A. van; BISAZZA, A.; CASELLI, T.; NOORD, G. van; NISSIM, M. Bertje: A dutch bert model. **arXiv preprint arXiv:1912.09582**, 2019. Citado nas páginas 50, 56, 63 e 64.

WANG, H.; ZHANG, Y.; CHAN, G. L.; YANG, J.; CHIEU, H. L. Universal Dependencies parsing for colloquial singaporean English. In: **Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**. Vancouver, Canada: [s.n.], 2017. p. 1732–1744. Citado na página 55.

WANG, W. Y.; KONG, L.; MAZAITIS, K.; COHEN, W. W. Dependency parsing for Weibo: An efficient probabilistic logic programming approach. In: **Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing**. Doha, Qatar: [s.n.], 2014. p. 1152–1158. Citado na página 55.

WANG, X.; JIANG, Y.; BACH, N.; WANG, T.; HUANG, Z.; HUANG, F.; TU, K. Automated concatenation of embeddings for structured prediction. **ArXiv**, 2021. Citado nas páginas 52, 56 e 57.

WILLIAMS, R. J. Simple statistical gradient-following algorithms for connectionist reinforcement learning. **Mach. Learn.**, 1992. Citado na página 52.

WU, Y.; SCHUSTER, M.; CHEN, Z.; LE, Q.; NOROUZI, M.; MACHEREY, W.; KRIKUN, M.; CAO, Y.; GAO, Q.; MACHEREY, K.; KLINGNER, J.; SHAH, A.; JOHNSON, M.; LIU, X.; KAISER, u.; GOUWS, S.; KATO, Y.; KUDO, T.; KAZAWA, H.; DEAN, J. Google's neural machine translation system: Bridging the gap between human and machine translation. **ArXiv**, 2016. Citado na página 51.

XU, X.; ZHANG, Y. House price forecasting with neural networks. **Intelligent Systems with Applications**, p. 200052, 2021. Citado na página 34.

YANG, X.; LIU, Y.; XIE, D.; WANG, X.; BALASUBRAMANIAN, N. Latent part-of-speech sequences for neural machine translation. In: **Conference on Empirical Methods in Natural Language Processing**. [S.l.: s.n.], 2019. Citado na página 26.

YASUNAGA, M.; KASAI, J.; RADEV, D. Robust multilingual part-of-speech tagging via adversarial training. In: **Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)**. New Orleans, Louisiana, USA: [s.n.], 2018. p. 976–986. Citado nas páginas 47, 56 e 57.

ZELDES, A. The gum corpus: creating multilayer resources in the classroom. **Language Resources and Evaluation**, p. 581–612, 2017. Citado na página 55.

ZEMAN, D.; POPEL, M.; STRAKA, M.; HAJIČ, J.; NIVRE, J.; GINTER, F.; LUOTOLAHTI, J.; PYYSALO, S.; PETROV, S.; POTTHAST, M.; TYERS, F.; BADMAEVA, E.; GOKIRMAK, M.; NEDOLUZHKO, A.; CINKOVÁ, S.; JR., J. H.; HLAVÁČOVÁ, J.; KETTNEROVÁ, V.; UREŠOVÁ, Z.; KANERVA, J.; OJALA, S.; MISSILÄ, A.; MANNING, C. D.; SCHUSTER, S.; REDDY, S.; TAJI, D.; HABASH, N.; LEUNG, H.; MARNEFFE, M.-C. de; SANGUINETTI, M.; SIMI, M.; KANAYAMA, H.; PAIVA, V. de; DROGANOVA, K.; ALONSO, H. M.; ÇÖLTEKIN, Ç.; SULUBACAK, U.; USZKOREIT, H.; MACKETANZ, V.; BURCHARDT, A.; HARRIS, K.; MARHEINECKE, K.; REHM, G.; KAYADELEN, T.; ATTIA, M.; ELKAHKY, A.; YU, Z.; PITLER, E.; LERTPRADIT, S.; MANDL, M.; KIRCHNER, J.; ALCALDE, H. F.; STRNADOVÁ, J.; BANERJEE, E.; MANURUNG, R.; STELLA, A.; SHIMADA, A.; KWAK, S.; MENDONÇA, G.; LANDO, T.; NITISAROJ, R.; LI, J. CoNLL 2017 shared task: Multilingual parsing from raw text to Universal Dependencies. In: **Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies**. Vancouver, Canada: [s.n.], 2017. p. 1–19. Citado na página 60.

ZHANG, A.; LIPTON, Z. C.; LI, M.; SMOLA, A. J. Dive into deep learning. **Computing Research Repository**, 2021. Citado na página 37.

ZHANG, X.; ZHAO, J.; LECUN, Y. Character-level convolutional networks for text classification. In: **Proceedings of the 28th International Conference on Neural Information Processing Systems**. Cambridge, MA, USA: [s.n.], 2015. p. 649—657. Citado na página 35.

ZHAO, F.; QUAN, B.; YANG, J.; CHEN, J.; ZHANG, Y.; WANG, X. Document summarization using word and part-of-speech based on attention mechanism. In: **Journal of Physics: Conference Series**. [S.l.: s.n.], 2019. p. 32008. Citado na página 26.

APÊNDICE

A

UNIVERSAL DEPENDENCIES FOR TWEETS IN BRAZILIAN PORTUGUESE: TOKENIZATION AND PART OF SPEECH TAGGING

SILVA, E.; PARDO, T.; ROMAN, N.; FELLIPO, A. Universal dependencies for tweets in brazilian portuguese: Tokenization and part of speech tagging. In: **Anais do XVIII Encontro Nacional de Inteligência Artificial e Computacional**. Porto Alegre, RS, Brasil: [s.n.], 2021. p. 434–445.

Universal Dependencies for Tweets in Brazilian Portuguese: Tokenization and Part of Speech Tagging

Emanuel Huber da Silva[1], **Thiago Alexandre Salgueiro Pardo**[1],
Norton Trevisan Roman[2], **Ariani Di Felippo**[3]

[1]Núcleo Interinstitucional de Linguística Computacional (NILC),
Instituto de Ciências Matemáticas e de Computação - Universidade de São Paulo (USP)

[2]Escola de Artes, Ciências e Humanidades - Universidade de São Paulo (USP)

[3]Núcleo Interinstitucional de Linguística Computacional (NILC),
Departamento de Letras - Universidade Federal de São Carlos (UFSCar)

emanuel.huber@usp.br, taspardo@icmc.usp.br,
norton@usp.br, ariani@ufscar.br

Abstract. *Automatically dealing with Natural Language User-Generated Content (UGC) is a challenging task of utmost importance, given the amount of information available over the web. We present in this paper an effort on building tokenization and Part of Speech (PoS) tagging systems for tweets in Brazilian Portuguese, following the guidelines of the Universal Dependencies (UD) project. We propose a rule-based tokenizer and the customization of current state-of-the-art UD-based tagging strategies for Portuguese, achieving a 98% f-score for tokenization, and a 95% f-score for PoS tagging. We also introduce DANTEStocks, the corpus of stock market tweets on which we base our work, presenting preliminary evidence of the multi-genre capacity of our PoS tagger.*

1. Introduction

In usual Natural Language Processing (NLP) workflows, text preprocessing is an essential procedure. Amongst the numerous strategies, Part of Speech (PoS) tagging is one of the first processes applied to data, being responsible for assigning each word in a sentence its appropriate grammatical role. Being one of the most elementary text analysis and structuring tasks in this workflow, PoS tagging builds the basis for the development of several NLP tools and applications, such as grammar checking and text simplification.

Although having been investigated for some time in the realm of well written texts, such as news for example, where it achieves state-of-the-art results above 97% accuracy in Portuguese (*e.g.*, [Fonseca et al. 2015, de Sousa and Lopes 2019]), the situation is very different when it comes to user-generated content (UGC), such as texts written by users in social networks, which do not strictly follow the rules of standard writing. These texts are sometimes marked by orality and informality, also making use of slangs, abbreviations and media-specific content (*e.g.*, hashtags and at-mentions in Twitter), which pose considerable challenges to their automatic processing.

As a related task necessary to PoS tagging, tokenization provides the elementary units to be tagged. Even though, at first sight, it might appear to be a straightforward task, UGC makes things considerably more difficult, given the above mentioned phenomena. The consequences, however, may endure all along the NLP pipeline, since badly

tokenized text will most certainly have a negative impact on the results of any PoS tagger applied to it and, consequently, on all NLP tasks that depend on this tagger's results.

Consider, for example, the tweet presented in Figure 1, taken from DANTEStocks, along with its tokenization and PoS tagging, as produced by our system. As it can be seen, the text does not comply with the standard rules for writing (specially regarding capitalization, word splitting, punctuation, the presence of slangs and abbreviations etc.), also presenting elements that are characteristic to the platform where they were written (*e.g.* the presence of hashtags and URLs).

Original:
#VALE5 é #VENDA? rsss #DEAL! #DEAL! #DEAL! '16 de março às 12:12' após vencto das opções podem puxar na... http://t.co/4mOMj1Om7d
Tokenized and PoS tagged:
#VALE5/PROPN é/AUX #VENDA/NOUN ?/PUNCT rsss/X #DEAL/NOUN !/PUNCT #DEAL/NOUN !/PUNCT #DEAL/NOUN !/PUNCT '/PUNCT 16/NUM de/ADP março/NOUN a/ADP as/DET 12:12/NUM '/PUNCT após/ADP vencto/NOUN de/ADP as/DET opções/NOUN podem/AUX puxar/VERB em/ADP a/DET .../PUNCT http://t.co/4mOMj1Om7d/SYM

Figure 1. Example of tweet from DANTEStocks, tokenized and PoS tagged.

Recently, initiatives have arisen to build morphosyntactically and syntactically annotated corpora of UGC such as, for instance, the treebank of tweets in English created by [Liu et al. 2018]. Although this is probably the most representative work in the area, many others have emerged, motivating authors like [Sanguinetti et al. 2020] to propose unified strategies to annotate UGC.

Guiding the recent work in morphosyntax and syntax in the area (including the ones cited above) is the Universal Dependencies (UD) initiative[1] [Nivre et al. 2016, Nivre et al. 2020]. UD aims at establishing *universal* tags and syntactical relations for corpus annotation, allowing cross-lingual studies and the reuse of methodologies. Most of the recent work in PoS tagging and syntactical parsing in NLP aligns with such initiative. Currently, the project counts with nearly 200 treebanks in over 100 languages. Amongst these, there are some initiatives for Portuguese (*e.g.* [Rademaker et al. 2017]), but, to the best of our knowledge, none for UGC.

Trying to fulfil this gap, we present in this paper an effort on building tokenization and UD PoS tagging systems for tweets in Brazilian Portuguese. We also introduce DANTEStocks, the corpus of stock market tweets on which we base our work, and which integrates the DANTE (Dependency-ANalised corpora of TwEets) project. To automatically add UD tags to this corpus, we propose a rule-based tokenizer and the customization of current state-of-the-art UD-based tagging strategies for Portuguese. We show that we achieve satisfactory results (98% f-score for tokenization and 95% f-score for PoS tagging), also presenting preliminary evidence of the multi-genre capacity of our PoS tagging system, thereby allowing for the construction of more robust NLP products.

The rest of this article is organized as follows. The next section focuses on briefly

[1] https://universaldependencies.org/

introducing the main related work. Section 3 then presents the details of the DANTE-Stocks corpus and the tokenization and PoS tagging methods that we explore, as well as the achieved results. Finally, our conclusions and final remarks are presented in Section 4.

2. Related work

Recently, PoS tagging and parsing have made their way back into the hot topics in NLP, specially with the advent of the UD project. Several initiatives to build new treebanks or to adapt existent treebanks to the UD formalism have arisen. Formally, a treebank is a corpus that contains sentences paired with their syntactic analyses, usually manually validated. One of the first treebanks in Brazilian Portuguese annotated according to the UD model is Bosque [Rademaker et al. 2017], which comprises well-written sentences extracted from journalistic texts, totaling 9,364 sentences.

More recently, there were also initiatives to annotate UGC texts, such as that of [Liu et al. 2018], which annotated a treebank for tweets in Engligh with 3,550 tweets in total, and that of [Sanguinetti et al. 2018], which built a treebank of 6,738 tweets written in Italian. So far, to the best of our knowledge, there is no such treebank for Portuguese. Given the available treebanks for several languages, PoS tagging and parsing systems have been developed, of which UDPipe [Straka et al. 2016, Straka 2018] is perhaps the most prominent initiative, currently in its second version, with both versions open-sourced.

The first version of UDPipe [Straka et al. 2016] relies on a perceptron network, with pre-computed features from the input text for PoS tagging, along with a bidirectional LSTM (Long Short-Term Memory) network for tokenization. The second version of UDPipe [Straka 2018], in turn, builds on a multi-layer bidirectional LSTM, using contextualized embeddings with a softmax classifier. Its input is a combination of three embedding codifications: (1) embeddings pre-trained on Wikipedia [2], (2) randomly initialized trained embeddings, and (3) character-level word embeddings using bidirectional GRUs.

Besides UDPipe, another popular tool is Udify [Kondratyuk and Straka 2019], which is a multi-task model based on BERT [Devlin et al. 2019], with an additional attention layer that captures the relations between all attention layers presented in the model, and which helps to capture low-level hierarchical information such as syntactic relations for final tasks such as PoS tagging. For tokenization, Udify uses the same word-piece tokenization from BERT [Devlin et al. 2019], where out-of-vocabulary words are split into syllables and their respective embeddings are used.

In what follows, we describe our efforts to build a UD-annotated corpus of UGC and our initiatives for developing appropriate tokenization and PoS tagging systems.

3. Tweet tokenization and PoS tagging

In this work, we build upon the corpus of tweets, written in Brazilian Portuguese for the stock market domain, described in [Vieira da Silva et al. 2020], which is publicly available for download[3]. We refer to the annotated version of this dataset as DANTEStocks, the first corpus to integrate the DANTE (Dependency-ANalised corpora of

[2]https://huggingface.co/bert-base-multilingual-uncased

[3]https://www.kaggle.com/fernandojvdasilva/stock-tweets-ptbr-emotions

TwEets) project. In its current state, DANTEStocks comprehends a total of 2,737 annotated (tokenized and PoS tagged) tweets. By the end of this project, we expect to have annotated all 4,517 tweets presented in the original corpus by [Vieira da Silva et al. 2020].

Based on ideas of in [Hovy and Lavid 2010], tweet annotation started with the grouping of tweets from the original data set into packages, following the order they are in that set. Each package was then automatically tokenized and codified according to the CoNNL-U format[4]. They were then automatically annotated with PoS tags, so as to build the starting point for review by human annotators. Each annotator received a copy of this set, with each copy containing the same tweets, but in a different (random) order.

Using a customized editing tool, each set was annotated by its corresponding annotator. Results from all annotators were then adjudicated[5] by a linguist with experience in NLP. Finally, a final version of the annotated (and adjudicated) package was built and incorporated into the final corpus. The automatic processes (tokenization and PoS tagging) were originally performed by UDPipe, being latter replaced by their customized versions that we describe in this paper.

In our experiments (described in the following subsections), 20% of the tweets from each annotated package were randomly sampled, so as to form our test set. The remaining 80% of the data in each package was then used to composed the training set, totaling 2,189 tweets for training and 548 for testing. So far, we have worked with 8 annotated packages (from a predicted total of 12), whose separation in training and testing sets may be visualized in Table 1.

Package	Total	Training data	Testing data
Subset 0	147	117	30
Subset 1	370	296	74
Subset 2	370	296	74
Subset 3	370	296	74
Subset 4	370	296	74
Subset 5	370	296	74
Subset 6	370	296	74
Subset 7	370	296	74
Total	2,737	2,189	548

Table 1. Number of tweets in the training and testing sets for each package.

3.1. Tokenization

Based on the annotated packages and on the orientations of an expert in Linguistics, we developed a rule-based tokenizer – "DANTE tokenizer" – which uses a set of regular expressions (that encode the rules) to split sentences into tokens. Built on top of the NLTK TweetTokenizer[6], the tokenizer was augmented with specific rules to deal with idiosyncrasies of the Portuguese language and the tweets from the stock market domain.

[4]CoNLL-U format is an already traditional column-based style of encoding UD annotation.

[5]*I.e.,* the cases in which annotators did not agree or did not annotate were marked and a decision was made on the final tag.

[6]https://www.nltk.org/api/nltk.tokenize.html

Within DANTEStocks, our tokenizer was responsible for:

- Removing HTML tags and formatting input texts according to the Unicode NFC standard;
- Decomposing some formal contractions found in Portuguese, such as the mixing of prepositions and articles (*e.g.* "no" → "em" + "o"); prepositions and pronouns (*e.g.* "dele" → "de" + "ele" and "deste" → "de" + "este"); and prepositions and adverbs (*e.g.* "daqui" → "de" + "aqui");
- Splitting up monetary values (*e.g.* "R$300" → "R$" + "300");
- Splitting up clitics (*e.g.* "localiza-se" → "localiza" + "-" + "se"), since individual words form, according to UD, the basic units of annotation; and
- Applying regular expressions to identify usual UGC phenomena in tweets and the DANTEStocks' domain (*e.g.*, hashtags, at-mentions, URLs, emoticons, and stock market codes) and turn them into tokens, given their syntactic role in these tweets.

To illustrate the need for this last task, *i.e.* the setting up of rules to deal with specific phenomena, consider the analysis of "PETR4", the market ticker for Petrobras' preferred stocks. In this case, the original NLTK TweetTokenizer breaks this code into "PETR" and "4", which makes no sense in the stock market domain, since "PETR4" refers to a well specified entity within it. To deal with this problems, a specific rule was added, so as to keep such codes as one single token.

The assessment of the tokenizer's performance was made through traditional Precision, Recall, and Micro F-score[7] measures. However, since the tokenized sentences may be longer than their original counterparts, due to contraction expansion, metric calculations must account for token classes and positions simultaneously. Take, for example, the text spam "da PETR4"[8]. In this case, if the tokenizer outputs "da" + "PETR4", instead of "de" + "a" + "PETR4" (*i.e.* it failed in expanding the contraction "da"), one true positive ("PETR4"), two false positives ("de" and "a"), and one false negative ("da") will be added to the contingency tables. This procedure is illustrated in Figure 2.

Original sentence	O aumento da PETR4 não para! #continuaassim
Predicted sentence	O aumento da PETR4 não para ! #continua assim
True positives	[O, aumento, PETR4, não, para, !]
False positives	[da, #, continua, assim]
False negatives	[de, a, #continuaassim]
Precision	$\frac{tp}{tp+fp} = \frac{6}{6+4} = 0.6$
Recall	$\frac{tp}{tp+fn} = \frac{6}{6+3} \simeq 0.667$
Micro F-score	$2\frac{prec*rec}{prec+rec} = 2\frac{0.6*0.667}{0.6+0.667} \simeq 0.632$

Figure 2. Assessment of the tokenizer's performance.

[7]Counts true/false positives and negatives globally
[8]Of PETR4, or PETR4's.

As a benchmark for comparison, we also tested the rule-based methods NLTK Word Tokenizer [Loper and Bird 2002], spaCy [Honnibal et al. 2020], NLTK TweetTokenizer [Loper and Bird 2002] and Twikenizer[9]. The main difference between them is that while NLTK TweetTokenizer and Twikenizer were tailored to the tokenization of tweets, by adding rules specific to its writing style, NLTK Word Tokenizer and spaCy had their rules derived from other genres, with NLTK Word Tokenizer being based on the Penn Treebank [Marcus et al. 1993] and spaCy being designed for the Bosque [Rademaker et al. 2017] dataset. With the exception of SpaCy, which deals with Portuguese, all tokenizers were designed for the English language.

Results for the tokenizers' evaluation on DANTE's test set can be seen in Table 2. In this table, we show the overall Precision, Recall and Micro F-score, averaged over their individual values at each tweet in the corpus. As it turns out, DANTE Tokenizer outscores its counterparts by at least 17.5% (at recall, against NLTK TweetTokenizer), ranging up to 52.2% (at recall, against spaCy). Regarding precision, gains ranged from 20.2% (against NLTK TweetTokenizer) to 43.3% (against Twikenizer), wheras for micro f-score they ranged from 19.2% (against NLTK TweetTokenizer) to 39.9% (against spaCy).

Tokenizer	Precision	Recall	Micro F-score
NLTK Word Tokenizer	0.7333 ± 0.1482	0.7784 ± 0.1304	0.7516 ± 0.1338
NLTK Twitter Tokenizer	0.8213 ± 0.1531	0.8385 ± 0.1286	0.8275 ± 0.1379
Twikenizer	0.6890 ± 0.2122	0.8286 ± 0.1174	0.7410 ± 0.1668
spaCy	0.7822 ± 0.1390	0.6476 ± 0.1886	0.7051 ± 0.1689
DANTE Tokenizer	0.9873 ± 0.0372	0.9854 ± 0.0447	0.9861 ± 0.0400

Table 2. Tokenization's evaluation results on DANTE's test set.

While looking at these results, however, one must bear in mind that DANTE Tokenizer was tailored to the same genre and domain as the test corpus, whereas others came either from the same genre (*i.e.* tweets), but different domains, as is the case with NLTK TweetTokenizer and Twikenizer; or from different genres and domains, as is the case with NLTK Word Tokenizer and spaCy. This only provides evidence on the low scalability of these tokenizers to other domains.

3.2. PoS tagging

In this work, PoS tagging aims at assigning each token the most probable tag from a subset of 17 possible tags defined by the UD project. At this stage, however, we are not focused on determining syntactic relations between them, a task to be approached in the forthcoming months of the project. Furthermore, since training a PoS tagger usually requires large data sets, which are not available at this stage of our project, we took Bosque as our initial training set, incrementally incorporating tweet packages, as they are being produced by annotators, into it and measuring tagger accuracy in this mixed set.

With this approach, we expect to (i) overcome the problem of data limitation, taking into consideration the acquired "learned knowledge" for standard general language, as brough by Bosque, and progressively incorporating knowledge from UGC; and (ii) evaluate the multi-genre capacity of the trained PoS taggers. Hopefully, as we incorporate

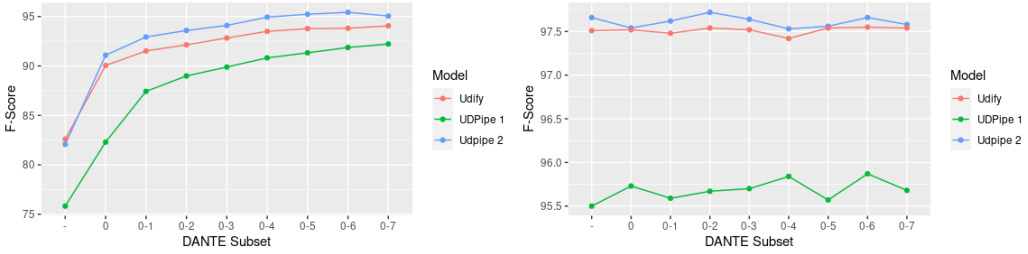
[9]https://pypi.org/project/twikenizer/

tweet information into the training set, taggers will maintain their performance in news texts, while at the same time improving their results in the tweets from DANTEStocks.

Both DANTEStocks' training set and Bosque's training and validation subsets add up to a total of 11,087 training samples, whereas their testing counterpart comprises 1,024 samples. In this work, we tested both version of UDPipe, along with Udify (*cf.* Section 2). To do so, we measured their F-score in the test set as new DANTEStocks' packs were added to the training set. At first, they were trained in Bosque's training set only. Then, after the annotation of DANTEStocks' first pack, the training was repeated with Bosque and this pack. This cycle goes on incrementally, pack by pack, until the last pack available so far. For all these runs, the testing set was kept the same, so as to determine whether there was any improvement along the way.

Results at each run are presented in Table 3. Values under the "DANTE Subset" column refer to the identification number (starting with zero) of the DANTEStocks packs added to the training set, with '-' indicating that no pack was used (*i.e.,* the system was trained only with Bosque). Next, we present the F-score values in the training set, so that differences between training and testing scores are shown. The last two columns report the systems' results when tested separately with DANTEStocks and Bosque test sets.

As expected, the more tweets are incorporated into the training set, the better the results in DANTEStocks' test set, for all tested taggers (Figure 3a). Interestingly, even though we are adding data from a different domain and genre as that of Bosque (which, in turn, might be considered as noise), tagger performance in Bosque's test set does not seem to be affected (Figure 3b). This could be an indication of the multi-genre capability of the tested taggers, even though more in-depth tests are needed to come to such a conclusion.



(a) DANTEStocks test set **(b) Bosque test set**

Figure 3. Taggers performance (F-Score) across the training subsets

As it turns out, the best results were obtained by UDPipe 2 in both test sets, reaching some impressive 95% F-Score in DANTEStocks, specially considering the difficulties of automatically analysing the twitter writing style, as pointed out in Section 1. Although all differences in Bosque were found to be significant[10], only the observed differences between UDPipe 1 and 2 in DANTEStocks were found to be relevant[11], with differences between Udify and both UDPipe 1 and 2 not being of statistical significance[12].

[10] Overall Kruskal-Wallis $(df = 2) = 21.514, p << 0.001$, pairwise Dunn (with Benjamini-Hochberg p-value adjust for multiple testing) $Z = -4.626, p << 0.001$ (UDPipe 1 vs. UDPipe 2), $Z = -2.023, p = 0.043$ (Udify vs. UDPipe 2); and $Z = 2.603, p = 0.014$ (Udify vs. UDPipe 1), at the 95% confidence level.

[11] $Z = -3.059, p = 0.007$

[12] $Z = 1.930, p = 0.080$ (Udify vs. UDPipe 1); and $Z = -1.128, p = 0.259$ (Udify vs. UDPipe 2).

Model	DANTE Subset	train f-score	DANTE test f-score	Bosque test f-score
UDPipe 1	-	98.81%	75.82%	95.50%
	0	99.27%	82.30%	95.73%
	0-1	99.75%	87.43%	95.59%
	0-2	99.73%	88.99%	95.67%
	0-3	99.67%	89.89%	95.70%
	0-4	99.65%	90.82%	95.84%
	0-5	99.67%	91.33%	95.57%
	0-6	99.60%	91.87%	95.87%
	0-7	99.58%	92.22%	95.68%
UDPipe 2	-	99.60%	82.07%	97.66%
	0	99.57%	91.09%	97.54%
	0-1	99.51%	92.93%	97.62%
	0-2	99.45%	93.59%	97.72%
	0-3	99.43%	94.10%	97.64%
	0-4	99.44%	94.94%	97.53%
	0-5	99.41%	95.23%	97.56%
	0-6	99.42%	95.43%	97.66%
	0-7	98.86%	95.05%	97.58%
Udify	-	98.13%	82.59%	97.51%
	0	98.11%	90.05%	97.52%
	0-1	98.01%	91.52%	97.48%
	0-2	97.90%	92.14%	97.54%
	0-3	97.87%	92.83%	97.52%
	0-4	97.81%	93.50%	97.42%
	0-5	97.76%	93.78%	97.54%
	0-6	97.65%	93.81%	97.55%
	0-7	97.62%	94.06%	97.54%

Table 3. PoS tagging results

F-Score results for each tag[13], as produced at the DANTEStocks test set, are presented in Table 4. As shown in the table, results range from 34.29% (with INTJ and Udify) to 99.33% (with CCONJ and UDPipe 2), with worst cases happening with the INTJ and X tags for all taggers. A possible reason for such anomalous values in these two tags, as illustrated in Figure 4, might be the fact that INTJ has fewer occurrences in the corpus, whereas X was used as a "left over" tag, being applied in cases of typos, pre-processing errors, or when annotators could not find a proper tag for the token. Interestingly, token-wise differences between taggers were not significant[14]. The confusion matrix for UDPipe 2 best model can be seen in Figure 5, which elucidates our analysis on INTJ and X classes. The confusion matrices for UDPipe 1 and Udify are available at our Github repository [15].

[13]The PART tag was omitted because there were no occurrences of this tag in DANTEStocks.

[14]Kruskal-Wallis $(df = 2) = 1.351, p = 0.509$

[15]https://github.com/huberemanuel/dante-tagging-eniac2021

Tag	UDPipe	UDPipe 2	Udify
ADJ	87.01	90.46	91.36
ADP	95.04	97.45	97.21
ADV	93.87	94.69	91.10
AUX	88.20	92.39	90.86
CCONJ	99.10	99.33	99.10
DET	97.40	97.70	97.30
INTJ	75.56	44.44	34.29
NOUN	91.63	94.20	92.98
NUM	96.51	97.99	97.27
PRON	88.89	92.10	88.64
PROPN	92.32	95.77	94.78
PUNCT	99.11	99.30	98.88
SCONJ	84.07	89.90	88.67
SYM	97.84	99.13	97.80
VERB	93.22	94.88	94.56
X	74.05	79.06	73.52

Table 4. F-score per class after training with all DANTE subsets (0-7)

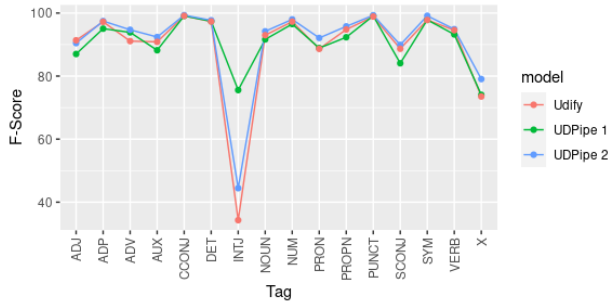


Figure 4. Models' performance at each tag found in the corpus.

4. Conclusion

This paper presented our current effort in building tokenization and PoS tagging services for tweets written in Brazilian Portuguese, inline with the Universal Dependencies international model, and so adding up to the increasing amount of resources devoted to this widely adopted standard for morpho-syntactical annotation. Moreover, we introduced DANTEStocks, a corpus of tweets from the financial market, which served as the basis for our experiments, also showing some preliminary evidence that our PoS tagging strategies have multi-genre capacity, producing good results for tweets while, at the same time, holding their performance in news texts.

This work comes as a part of a larger project that aims at fostering research on syntax and parsing for Brazilian Portuguese: the POeTiSA project[16]. Our final goal in this broader project is to build a large multi-genre treebank for Portuguese, also developing state-of-the-art PoS tagging and parsing systems for this language. Within this context,

[16]https://sites.google.com/icmc.usp.br/poetisa

DANTEStocks comes up as one of the corpora, being the first to integrate the DANTE initiative – a treebank of corpora for tweets, which is itself part of POeTiSA.

As our next steps, we intend to refine our evaluation of the presented taggers' performance, by using the full set of DANTEStocks' annotated data (which was not yet fully annotated during the writing of this article), and to explore other directions for tokenization (*e.g.*, to use sequence models such as the one presented in [Devlin et al. 2019]). We also envision the syntactic annotation of the tweets according the UD guidelines.

As a final remark, it is worth mentioning that, although in this article we have focused on describing how systems were developed and tested, the adoption of the UD model required an extensive linguistic study and adaptation of its guidelines to the Portuguese language, along with the development of strategies and guidelines for the annotation of different genres. These are results that are still under construction, and which will be left for future publications, more centered on the linguistic aspects of the project.

Acknowledgments

The authors are grateful to the USP/IBM/FAPESP Center for Artificial Intelligence (C4AI - grant 2019/07665-4), the Facens University Center, and the researchers and linguists of the POeTiSA project for their hard work on the DANTE treebank.

References

de Sousa, R. C. C. and Lopes, H. (2019). Portuguese pos tagging using blstm without handcrafted features. In *Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications*, pages 120–130.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Fonseca, E., Rosa, J., and Aluísio, S. (2015). Evaluating word embeddings and a revised corpus for part-of-speech tagging in portuguese. *Journal of the Brazilian Computer Society*, 21(2):1–14.

Honnibal, M., Montani, I., Van Landeghem, S., and Boyd, A. (2020). spaCy: Industrial-strength Natural Language Processing in Python.

Hovy, E. and Lavid, J. (2010). Towards a 'science' of corpus annotation: A new methodological challenge for corpus linguistics. *International Journal of Translation*, 22(1):13–36.

Kondratyuk, D. and Straka, M. (2019). 75 languages, 1 model: Parsing Universal Dependencies universally. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2779–2795, Hong Kong, China. Association for Computational Linguistics.

Liu, Y., Zhu, Y., Che, W., Qin, B., Schneider, N., and Smith, N. A. (2018). Parsing tweets into Universal Dependencies. In *Proceedings of the 2018 Conference of*

the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), pages 965–975, New Orleans, Louisiana. Association for Computational Linguistics.

Loper, E. and Bird, S. (2002). Nltk: The natural language toolkit. In *In Proceedings of the ACL Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics. Philadelphia: Association for Computational Linguistics*.

Marcus, M. P., Santorini, B., and Marcinkiewicz, M. A. (1993). Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics*, 19(2):313–330.

Nivre, J., de Marneffe, M.-C., Ginter, F., Goldberg, Y., Hajič, J., Manning, C. D., McDonald, R., Petrov, S., Pyysalo, S., Silveira, N., Tsarfaty, R., and Zeman, D. (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.

Nivre, J., de Marneffe, M.-C., Ginter, F., Hajič, J., Manning, C. D., Pyysalo, S., Schuster, S., Tyers, F., and Zeman, D. (2020). Universal Dependencies v2: An evergrowing multilingual treebank collection. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 4034–4043.

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)*, pages 197–206, Pisa, Italy.

Sanguinetti, M., Bosco, C., Cassidy, L., Çetinoğlu, Ö., Cignarella, A. T., Lynn, T., Rehbein, I., Ruppenhofer, J., Seddah, D., and Zeldes, A. (2020). Treebanking user-generated content: A proposal for a unified representation in Universal Dependencies. In *Proceedings of the 12th Language Resources and Evaluation Conference*.

Sanguinetti, M., Bosco, C., Lavelli, A., Mazzei, A., Antonelli, O., and Tamburini, F. (2018). PoSTWITA-UD: an Italian Twitter treebank in Universal Dependencies. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Straka, M. (2018). UDPipe 2.0 prototype at CoNLL 2018 UD shared task. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 197–207, Brussels, Belgium. Association for Computational Linguistics.

Straka, M., Hajič, 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, Portorož, Slovenia. European Language Resources Association (ELRA).

Vieira da Silva, F. J., Roman, N. T., and Carvalho, A. M. (2020). Stock market tweets annotated with emotions. *Corpora*, 15(3):343–354.

A. Confusion Matrix

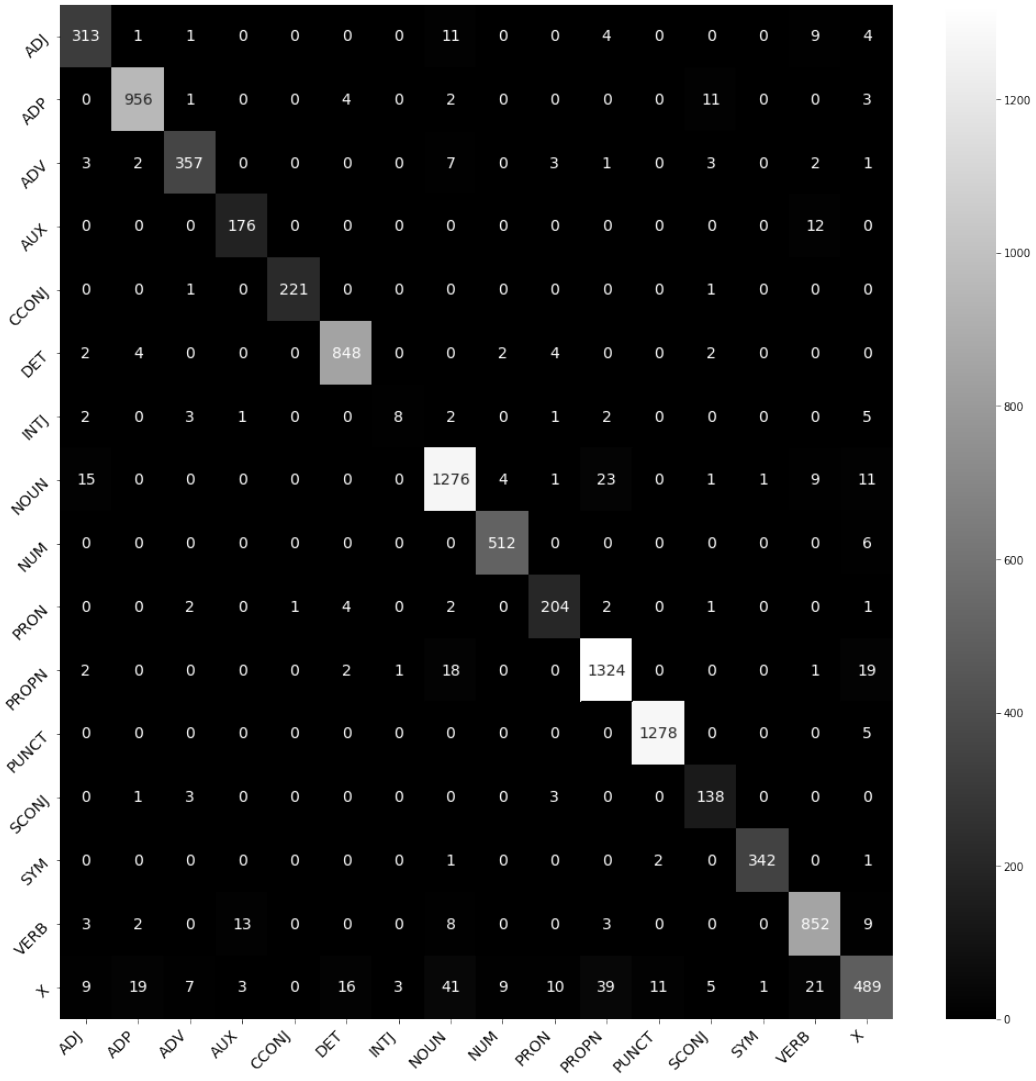


Figure 5. Confusion matrix for UDPipe 2 trained on all DANTE subsets

APÊNDICE

B

DESCRIÇÃO PRELIMINAR DO CORPUS DANTESTOCKS: DIRETRIZES DE SEGMENTAÇÃO PARA ANOTAÇÃO SEGUNDO UNIVERSAL DEPENDENCIES

Di FELIPPO, A.; POSTALI, C.; CEREGATTO, G.; GAZANA, L.; SILVA, E.; ROMAN, N.; PARDO, T. Descrição preliminar do corpus dantestocks: Diretrizes de segmentação para anotação segundo universal dependencies. In: **Anais do XIII Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana**. Porto Alegre, RS, Brasil: [s.n.], 2021. p. 335–343.

Descrição Preliminar do *Corpus* DANTEStocks: Diretrizes de Segmentação para Anotação segundo *Universal Dependencies*

Ariani Di Felippo[1], Caroline Postali[1], Gabriel Ceregatto[1], Laura S. Gazana[1], Emanuel H. da Silva[2], Norton T. Roman[3], Thiago A. S. Pardo[2]

[1]Núcleo Interinstitucional de Linguística Computacional (NILC)
Departamento de Letras – Universidade Federal de São Carlos (UFSCar)
Caixa Postal 676 – 13565-905 – São Carlos – SP – Brasil

[2]Núcleo Interinstitucional de Linguística Computacional (NILC)
Instituto de Ciências Matemáticas e de Computação – Universidade de São Paulo (USP)
Caixa Postal 668 – 13566-970 – São Carlos – SP – Brazil

[3]Escola de Artes, Ciências e Humanidades – Universidade de São Paulo (USP)

ariani@ufscar.br, caroline.postali@gmail.com,
gabriel@ceregatto.admin.br, lauragazana@estudante.ufscar.br,
{emanuel.huber,norton}@usp.br, taspardo@icmc.usp.br

Abstract. *The annotation of informal texts within the Universal Dependencies framework requires two segmentation processes: definition of the relevant unity for syntactic analysis and identification of syntactic words. In this paper, we present the linguistic idiosyncrasies of DANTEStocks, a corpus of tweets from the financial market, written in Portuguese, and the general guidelines for their automatic segmentation. As such, this work contributes to a better understanding of linguistic aspects of tweets and the development of resources and tools for automatic processing of this subgenre of user-generated content.*

Resumo. *A anotação de textos informais segundo a Universal Dependencies requer dois processos de segmentação: delimitação da unidade relevante para a análise sintática e identificação das palavras sintáticas. Neste artigo, apresentam-se as idiossincrasias linguísticas do corpus DANTEStocks, composto por tweets do mercado financeiro, escritos em Português, e as estratégias gerais de segmentação automática. Assim, contribui-se para a descrição de aspectos linguísticos dos tweets e para o desenvolvimento de recursos e ferramentas de processamento automático desse subgênero de "user-generated content".*

1. Introdução

Diante da imensa relevância adquirida na última década, as redes sociais (como *Facebook*, *WhatsApp*, *Twitter*, etc.) são fontes de conteúdo (em inglês, *user-generated content* - UGC) inestimáveis para consumidores, políticos e governos no geral. Com isso, o desenvolvimento de ferramentas e aplicações linguístico-computacionais (como as de análise de sentimento e mineração de opinião) tem se tornado tópico central do Processamento Automático das Línguas Naturais (PLN) [Sanguinetti et al., 2020a].

Nesse cenário, já há vários *taggers* (etiquetadores morfossintáticos) [p.ex.: Owoputi et al., 2013; Lynn et al., 2015; Bosco et al., 2016; Proisl, 2018] e *parsers* (analisadores sintáticos) [p.ex.: Foster, 2010; Petrov, Mcdonald, 2012; Kong et al., 2014 e Liu et al., 2018] relativamente precisos para o processamento de UGCs, sobretudo em inglês. E esse ferramental só foi desenvolvido graças aos *corpora* anotados (*treebanks*) e aos algoritmos de aprendizado de máquina. Grande parte dos *treebanks* de UGC construídos nos últimos anos são compostos exclusivamente por *tweets*. O destaque dos *corpora* de *tweets* (os *tweebanks*) se deve pela facilidade de obtenção dos dados, política do *Twitter* sobre o uso dos dados para fins acadêmicos e relevância para aplicações de PLN. O tamanho desses recursos varia de 500 a aproximadamente 6,700 mensagens [Sanguinetti et al., 2020a].

Os *tweebanks* mais recentes possuem anotação segundo a *Universal Dependencies* (UD) [Nivre, 2015; Nivre et al., 2020], um modelo gramatical que fornece principalmente um conjunto de etiquetas morfossintáticas universais e de relações de dependências sintáticas para anotação de *corpus,* o que possibilita estudos "*cross*-linguísticos" e reuso de metodologias.

A anotação UD de UGC, como os *tweets,* requer inicialmente que a unidade relevante para a análise sintática seja definida. Isso significa decidir se essa unidade será delimitada com base na noção de sentença (como nos textos formais) ou outro critério. Ademais, por se basear em uma visão lexicalista da sintaxe, a anotação UD necessita que as palavras sintáticas[1] sejam identificadas (tokenizadas)[2]. Para tanto, é preciso descrever as características linguísticas (estruturais, ortográficas e lexicais) do *tweets* que compõem o *corpus* que será anotado [Liu et al., 2018; Sanguinetti et al., 2020a,b]. Por exemplo, uma característica geral dos *tweets* é a ocorrência de autocensuras ("m*" → "merda"). Para o reconhecimento das autocensuras como palavras, é preciso prever que o asterisco não seja segmentado, mas reconhecido como parte constitutiva do *token.*

Neste artigo, descrevem-se as características linguísticas do *corpus* construído por Silva et al. (2020) e as decorrentes estratégias automáticas de segmentação (isto é, delimitação da unidade de análise sintática e tokenização) para anotação UD. Denominado DANTEStocks, o *corpus* é composto por *tweets* em português sobre ações do índice Ibovespa e possui anotação de emoções. O DANTEStocks será o primeiro *corpus* de UGC em português com anotação UD. Acredita-se que a anotação UD poderá potencializar o emprego do *corpus* nas investigações sobre análise de sentimentos e ampliar a sua utilidade em outros tipos de pesquisas linguístico-computacionais. Dessa forma, esse trabalho contribui para os estudos descritivos sobre as características linguísticas dos *tweets* e para o desenvolvimento de recursos, ferramentas e aplicações de processamento automático desse tipo particular de UGC.

Nas Seções 2, descreve-se brevemente o modelo UD. Na Seção 3, apresentam-se o *corpus* DANTEStocks, as características estruturais de seus *tweets* e a decorrente delimitação da unidade de análise sintática. Na Seção 4, sistematizam-se os dispositivos linguísticos (lexicais e ortográficos) que caracterizam o *corpus* e discute-se a tokenização de alguns deles. Na Seção 5, apresentam-se as considerações finais sobre o trabalho, destacando suas contribuições e estudos futuros.

[1] Palavra sintática (em inglês, *syntactic word*) é a unidade mínima a que corresponde uma função sintática (https://universaldependencies.org/u/overview/tokenization.html).
[2] Na anotação UD, palavras sintáticas (ou itens lexicais) são sinônimos de *tokens*.

2. O Modelo *Universal Dependencies*

O modelo UD prevê anotação no nível sentencial e diretrizes para tokenização e anotação morfossintática e sintática[3]. Sobre a tokenização, a UD, a partir de uma visão lexicalista da sintaxe, define que uma relação de dependência (*deprel*) ocorre entre palavras de uma sentença e as características morfológicas são representadas por propriedades (ou *features*). Assim, as unidades básicas de anotação são palavras sintáticas. Com isso, os clíticos precisam ser separados de seus hospedeiros ("prepare-se" → "prepare" "se") e tratados como palavras independentes, assim como as contrações precisam ser decompostas ("das" → "de" "as"). Excepcionalmente, o modelo permite a combinação de *tokens* ortográficos em uma única palavra, como é o caso das abreviações (p.ex.: "e.g."). Quanto à anotação linguística, a UD prevê 2 níveis. No nível morfológico, especificam-se 3 tipos de informação: lema, etiqueta morfossintática e traços lexicais/gramaticais (das palavras). No nível sintático, parte-se da premissa de que as deprels são relações binárias e assimétricas [Nivre, 2015; Nivre et al., 2020; Marnefee et al., 2021] e que a representação básica de uma estrutura de dependências é arbórea, na qual uma palavra é o *root* (raiz) da sentença.

Na Figura 1, ilustra-se a anotação UD de uma sentença de um *corpus* jornalístico em português. Em caixa alta, estão codificadas as etiquetas morfossintáticas, como DET para "esse", NOUN para "carro" e VERB para "achado". A versão 2.0[4] da UD dispõe de 17 etiquetas, juntamente com critérios para o emprego de cada uma delas. Logo acima das etiquetas, estão as formas canônicas, por exemplo: "esse", "carro" e "achar" são respectivamente os lemas de "esse", "carro" e "achado". As *deprels* estão indicadas por setas rotuladas que se originam no *head* (cabeça) e se destinam ao dependente. Na Figura 1, "carro", por exemplo, é dependente de "achado" (cabeça) e estes estão conectados pela *deprel* **nsubj:pass** (sujeito nominal da passiva). O verbo "achado" é a raiz da sentença-exemplo. A UD (2.0) fornece 37 relações, juntamente com critérios para o emprego de cada uma delas. A UD também fornece uma lista bastante extensa de traços que codificam propriedades lexicais e gramaticais das palavras. Embora ausentes na Figura 1, "carro", no caso, possui os traços-valores: Gender=Masc e Number=Sing[5].

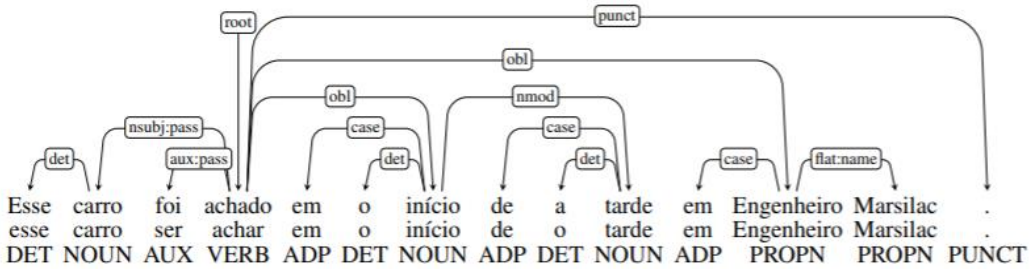


Figura 1. Exemplo de anotação sintática segundo a UD [Rademaker *et al*., 2017].

A seguir, apresenta-se o DANTEStockes para, na próxima seção, descrever as suas particularidades lexicais, ortográficas e estruturais.

[3] Para o português do Brasil, Duran (2021) construiu um manual com diretrizes de tokenização e de anotação morfossintática segundo a UD (especificamente para textos formais, como os jornalísticos).
[4] https://universaldependencies.org/guidelines.html
[5] Essa informação foi recuperada do *corpus UD-Portuguese-Bosque* por meio da plataforma *online* Grew-match (http://match.grew.fr/?corpus=UD_Portuguese-Bosque@2.8).

3. DANTEStocks - Estrutura dos *Tweets* e Definição da Unidade de Análise

O DANTEStocks[6] é um *corpus* de material textual compilado do *Twitter*, que parece uma mescla de rede social e *microblog*[7] [Freitas, Barth, 2015], e cujas principais características são a dinamicidade das interações (sejam comentários ou republicações) e a brevidade das mensagens (restrição de 140 caracteres). Considerado um gênero, o *tweet* parece ser constituído por resquícios de outros gêneros (como notícia, propaganda, bilhete, diário íntimo, etc.), que foram modificados para atender às necessidades de comunicação da rede [Marcuschi, 2008, Freitas, Barth, 2015]. Aliás, esses diferentes gêneros que se entrelaçam nos *tweets* evidenciam a influência da oralidade nessa escrita online. O DANTEStocks engloba especificamente 4.517 *tweets* contendo menção a alguma das ações do índice Ibovespa[8]. As postagens foram coletadas automaticamente em 2014 com base nos *tickers* (códigos) (isto é, cadeias de 4 letras e 1 número que fazem alusão ao nome da empresa e ao tipo de ação, como "PETR4") das ações do índice. O *corpus*, originalmente construído para pesquisas sobre análise de sentimentos, já possui anotação de emoções [cf. Silva et al., 2020].

Quanto à composição estrutural, os *tweets* do *corpus* variam bastante. Há *tweets* formados por uma ou mais sentenças claramente delimitadas, como (1), (2) e (3). Mas há também *tweets* que apresentam, frente às normas da língua padrão, ausência de pontuação (4) ou pontuação equivocada (5). *Tweets* relativamente fragmentados (6) também compõem o *corpus*. Em (4), o *tweet* parece ser composto por duas sentenças ("O #PT conseguiu fazer propaganda eleitoral antecipada" e "O que a @dilmabr tem a dizer sobre isso"). Essa interpretação pode ser corroborada pela capitalização do segundo "o" (negrito). Em (5), o exemplo é de uso inadequado da vírgula, provavelmente em substituição ao ponto de exclamação. O *tweet* (6) exemplifica uma postagem relativamente fragmentada, composta por uma *hashtag* seguida por um sintagma nominal e um *link.* O *tweet* (3), em especial, apresenta alternância de código linguístico (em inglês, *code-switching*) (português-inglês) em nível sentencial.

(1) Sera k petr4 já entrou na baixa?
(2) PETR4 subiu na bolsa 13,50. Muito bem, surpreso com o resultado.
(3) #CSNA3: Está em região de suporte que vem resistindo. Who knows?
(4) O #PT conseguiu fazer propaganda eleitoral antecipada **O** que a @dilmabr tem a dizer sobre isso?
(5) Bom dia Marcos, Alguma previsão para petr4?!
(6) #GGBR4 Suportes e resistências http://t.co/Azw6yIEVI9

Para a anotação sintática de textos formais, a unidade de anotação é comumente a sentença, cuja segmentação automática é tarefa relativamente simples, pois a pontuação pode ser usada como critério [Reynar e Ratnaparkhi, 1997]. Buscando estabelecer certa compatibilidade com os *treebanks* de textos formais, Sanguinetti et al. (2020b) optam por segmentar (automaticamente) somente os *tweets* com sentenças bem delimitadas (como (1), (2) e (3)), podendo utilizar índices para reconstruir, se necessário, as mensagens segmentadas [Rehbein et al., 2019].

[6] Disponível em: https://www.kaggle.com/fernandojvdasilva/stock-tweets-ptbr-emotions/data.
[7] *Microblog* é um tipo de *blog* no qual os usuários fazem atualizações breves de texto (até 200 caracteres), sobretudo veiculando impressões pessoais.
[8] Principal índice da bolsa de valores oficial do Brasil, a B3 (de "Brasil, Bolsa, Balcão").

Em outros trabalhos, o *tweet* é considerado unidade mínima de anotação [Kong et al., 2014; Liu et al., 2018; Sanguinetti et al., 2018]. Embora a não-segmentação dos *tweets* em unidades menores possa levar a um emprego excessivo da parataxis[9] [Sanguinetti et al., 2020], que é a *deprel* usada para relacionar elementos justapostos que não estejam coordenados, subordinados ou em outra relação argumento-predicado, essa foi a opção adotada para o DANTEStocks. Com isso, a anotação UD (morfossintática, por enquanto) do *corpus* está sendo feita no nível do *tweet*.

Essa opção se justifica por algumas razões. Uma delas são os problemas de pontuação, que dificultam a segmentação sentencial automática. Embora haja outros critérios para essa segmentação, como a detecção de estruturas verbo-argumento, estes não foram considerados devido à complexidade de se processar automaticamente os *tweets.* Assim, considerar o *tweet* como unidade única economiza o esforço necessário para desenvolver, manter, adaptar ou realizar o pós-processamento em um segmentador automático. Além disso, considerar o *tweet* como unidade mínima pode ser relevante para pesquisas linguístico-computacionais a respeito desse tipo de UGC. Cignarella *et al*. (2019), por exemplo, destacam que o estudo da correlação entre aspectos sintáticos (via UD) e ironia só foi possível diante dos *tweets* enquanto unidade. Outra razão importante é a interpretação (e consequente anotação sintática) dos *tweets*, que, muitas vezes, depende da mensagem completa. Isso fica evidente diante dos *tweets* que têm certa fragmentação, como (6). Somente é possível interpretar que os níveis de "suporte" e "resistência" (isto é, conceitos de análise gráfica) de interesse são os relativos à ação/*ticker* "GGBR4" com base na mensagem completa. A anotação intersentencial de alternância de código também pode ser considerada mais apropriada no nível do *tweet*.

4. Os Fenômenos UGC do DANTEStocks e a sua Tokenização

A partir de trabalhos como os de Lyddy et al. (2014), Liu et al. (2018) e Sanguinetti et al., 2020a,b), as particularidades ortográficas e lexicais do *corpus* foram sistematizadas em 7 dimensões. As dimensões e os fenômenos estão exemplificados no Quadro 1.

1. **Simplificação de código:** engloba os fenômenos "ergográficos" (em inglês, *ergographic phenomena*), que reduzem o esforço de escrita de um único *token*, como remoção/adição de diacrítico, ausência de hífen, substituição de diacrítico (pela letra "h"), omissão de letras (finais e mediais), erro ortográfico/digitação e fonetização.
2. **Abreviação:** toda sequência de caracteres que representa de forma reduzida várias palavras; a abreviação pode ser do tipo contração (de elementos gramaticais), acrônimo ou inicialismo (do inglês, *initialism*) (isto é, abreviações compostas pelas letras iniciais de palavras comuns ("lp" → "longo prazo") [Lyddy et al., 2014].
3. **Expressão de sentimento:** fenômenos que emulam o sentimento expresso pela prosódia, expressão facial ou gesto na interação via *tweet,* como alongamento grafêmico (sobretudo de vogais), repetição de pontuação, autocensura e emoticons.
4. **Influência de língua estrangeira:** vocábulo formado com base em outra língua; "estopar", por exemplo, baseia-se no verbo em inglês *"stop"* ("parar") (isto é, interromper venda ou compra de um ativo diante de dado preço).
5. **Expressão de oralidade:** toda palavra cuja grafia remonta à comunicação (fala) informal, as quais são, por vezes, empregadas com função humorística.

[9] https://universaldependencies.org/u/dep/parataxis.html

6. **Elemento metalinguístico**: todo elemento que tipicamente ocorre no *Twitter*, como *hashtag*, menção, marca de *retweet*, URL e truncamento lexical (quebra de palavra).
7. **Fenômeno de domínio:** todo fenômeno lexical/gráfico que diferencia os *tweets* do DANTEStocks dos demais *tweets*, a saber: *tickers*, *cashtag*, numerais com parte decimal indeterminada, índices de (des)valorização das ações, substituições lexicais (por símbolo), expressões temporais alfanuméricas e valor monetário aglutinado.

Quadro 1. Exemplo dos fenômenos UGC no DANTEStocks.

Fenômeno	**Exemplo**	**Forma padrão/glosa[10]**
Simplificação de código		
Ausência/adição de diacrítico	**proprio, milhao, Graca, fêz**	*próprio, bilhão, Graça, fez*
Ausência de hífen	**sexta feira, caça níquel**	*sexta-feira, caça-níquel*
Substituição de diacrítico	**eh, neh, tou**	*é, né, tô*
Omissão de letras	**d, n, qdo, tx, ult, pq**	*de, não, quando, taxa, último, porque*
Erro ortográfico/digitação	**comrpa, agradeveis**	*compra, agradáveis*
Fonetização	**k, kd**, **krk, kct**	*que, cadê, caraca, cacete*
Abreviação		
Contração	**oq, pq**	*o que, por que, por favor*
Acrônimo/inicialismo	**BB, cf, lp**	*Banco do Brasil, conselho fiscal, longo prazo*
Expressão de sentimento		
Alongamento de pontuação	Onde a #OIBR4 vai parar**???**	Onde a #OIBR4 vai parar**?**
Alongamento grafêmico	n**oo**s**aaa**, LI**NN**DA	*nossa, linda*
Autocensura	**p**a m***	*puta, merda*
Emoticon	**o.O :) :/**	*surpresa, sorriso (feliz), indecisão*
Influência de língua estrangeira		
Formação verbal	**estopar**	*'parar investimento'*
Marca de oralidade		
Coloquialismo	**guvêrno, bão, ae, péra, vamu**	*governo, bom, aí, espere, vamos*
Expressão cristalizada	**né, daí (dae)**	*'não é', 'de aí'*
Exclamação onomatopeica	**hahaha, hehehe**	*risos*
Elementos metalinguísticos (do *Twitter*)		
Hashtag	**#Petr4**	*'indexadores de tópicos ou assuntos'*
Menção	**@garimpodeacoes**	*'perfil/usuário'*
Marca de *retweet*	**RT** @Ary_AntiPT	*'republicação de um tweet'*
URL	**http://t.co/sROpyWPblN**	*'endereço da web'*
Truncamento (lexical)	Ação sobe **fo...**	*Ação sobe fo(rte)*
Fenômeno do domínio (Ibovespa)		
Ticker	**Petr4**	*'código de uma ação'*
Cashtag	**$LREN3**	*'código de ação precedido por $'*
Indeterminação da parte decimal	De 18,**xx** a 21,00	*'qualquer valor na parte decimal'*
Índice de (des)valorização	**+2,09%**, **-11,42%**	*'percentual de (des)valorização de ação'*
Substituição lexical	... precisam de muito **$**	*... precisam de muito dinheiro*
Expressão (temporal) híbrida	**1T14**	*primeiro trimestre de 2014*
Valor monetário aglutinado	**R$20,00**	*R$ 20,00*

[10] As formas de superfície do *corpus* não foram substituídas pelas formas da linguagem padrão, as quais estão no Quadro 1 apenas como recurso didático fornecido ao leitor para a compreensão dos fenômenos.

Partindo-se da decisão de não normalizar os *tweets* do *corpus* com o objetivo de desenvolver ferramentas e sistemas para o mundo real, foi necessário definir o estatuto de palavra de alguns dos fenômenos sistematizados para a subsequente tokenização.

Quanto aos fenômenos de simplificação de código, ressalta-se que um composto hifenizado (como "caça-níquel") constitui, segundo a visão lexicalista da UD, uma única palavra. Assim, mesmo que a ausência do hífen, como em "caça níquel", resulte na identificação automática de dois *tokens* ("caça" e "níquel"), a anotação UD precisa evidenciar que se trata de um composto, isto é, *token* único. Uma alternativa pode ser a utilização da *deprel* **compound**, como é feito no *corpus* UD_English-EWT[11] em inglês.

As contrações são formas abreviadas de duas palavras funcionais com remoção de espaços e letras. Nessa categoria, no entanto, há diferentes fenômenos de redução, os quais necessitam, por isso, de estratégias distintas de tokenização. A forma superficial "oq" (em "Oq faz?"), por ser constituída por dois pronomes ("o" "que") e ter a função única de pronome, corresponde a um *token* único. Já "pq", ao reduzir duas palavras ("por" "que"), de categorias morfossintáticas diferentes (preposição e pronome, respectivamente), deve ser decomposta em dois *tokens.* Os outros tipos de abreviação, ou seja, acrônimos (que reduzem nomes de entidades), como "BB" ("Banco do Brasil"), e inicialismos (que abreviam expressões compostas por palavras comuns), como "cf" ("conselho fiscal"), são *tokens* únicos, uma vez que desempenham função sintática específica, sendo possível atribuir à forma reduzida a categoria morfossintática do *head*.

Quanto às expressões de sentimento, destaca-se que as autocensuras, como "car*" ("caralho"), e os *emoticons* (";-*" → "beijo") correspondem a palavras sintáticas. No entanto, o adequado reconhecimento destes como tal requer que os sinais de pontuação e os caracteres especiais sejam reconhecidos como elementos constitutivos do *token.* No DANTEStocks, os *emoticons* ocorrem ao final dos *tweets,* não havendo uma ligação clara com a estrutura do *tweet,* a não ser "discursiva".

As marcas de oralidade classificadas como "expressão cristalizada" – "né" e "daí" (ou "dae") – são etimologicamente contrações de "não é" e "de aí". Sendo contrações (ou seja, *tokens* compostos por mais de uma categoria gramatical), elas seriam tokenizadas segundo a UD. No entanto, essas expressões funcionam no *corpus* como uma unidade, sendo o mais adequado, nesse caso, não realizar a decomposição. Atualmente, a categoria gramatical mais adequada a ser atribuída a elas está sob estudo (se advérbio ou interjeição). No nível sintático, no entanto, sabe-se que essas expressões desempenham função discursiva e a anotação via *deprel* precisará evidenciar isso.

Sobre os elementos metalinguísticos, os truncamentos lexicais ocorrem principalmente no fim de um *tweet* devido ao limite de caracteres. Na literatura, eles são tokenizados e, caso as formas completas possam ser recuperadas, os truncamentos são anotados em função delas. No que diz respeito às *hashtags* (e também *cashtags*) e menções, o reconhecimento dos símbolos "$" e "@" como parte constitutiva dos *tokens* parece variar na literatura. No DANTEStocks, esses símbolos foram considerados como tal, compondo um *token* único com a palavra ou expressão que eles precedem.

Quanto aos fenômenos de domínio, os índices de (des)valorização das ações compreendem 3 *tokens* ("+2,09%" → "+" "2,09" "%"). Especificamente, reconhecer "+" como *token* (no caso, um símbolo) justifica-se pela possibilidade de substituí-lo por

[11] https://github.com/UniversalDependencies/UD_English-EWT

outra palavra (como “subiu”). Outra característica de domínio são as formas reduzidas de expressões temporais, como “1T14” (“primeiro trimestre de 2014”). Estas, ao funcionarem como unidade, são consideradas palavras únicas e anotadas com a categoria morfossintática do *head*, como sugerido para os acrônimos e inicialismos. No DANTEStocks, as expressões monetárias podem ocorrer aglutinadas (isto é, sem espaço entre o símbolo monetário e o numeral) (“R$20,00”). Estas, no entanto, são compostas por dois *tokens* (já que “R$20,00” é o mesmo que “vinte reais”) e, por isso, precisam ser tokenizadas.

5. Considerações finais

A caracterização linguística ora apresentada revelou que os *tweets* do DANTEStocks são marcados por convenções e limitações impostas pela plataforma, marcas de informalidade e certos dispositivos linguísticos, alguns deles, aliás, dependentes de domínio. O estudo sobre a estrutura dos *tweets* e a descrição dos dispositivos lexicais e gráficos fundamentaram a segmentação do *corpus* para a anotação UD. A definição do estatuto de *token* dos fenômenos resultou em algumas regras contextuais utilizadas por Silva et al. (2021) para adaptar o tokenizador simbólico de *tweets* do pacote NLTK[12] ao DANTEStocks. Para dar continuidade a este trabalho, pretende-se quantificar os fenômenos no *corpus*, gerando estatísticas de frequência/relevância.

Agradecimentos

Os autores deste trabalho agradecem ao Centro de Inteligência Artificial (C4AI - USP) e o apoio da Fundação de Apoio à Pesquisa do Estado de São Paulo (processo FAPESP #2019/07665-4) e da IBM Corporation.

Referências

Bosco, C., Tamburini, F., Bolioli, A., Mazzei, A. (2016). Overview of the EVALITA 2016 Part of Speech tagging on TWitter for ITAlian task. In: Anas do 5º EVALITA.

Cignarella, A.T., Bosco, C., Rosso, P. (2019). Presenting TWITTIRO-UD: an Italian twitter treebank in Universal Dependencies. In: Anais do 5º Depling, p.190-7. Paris, França, ACL.

Duran, M.S. (2021). Manual de anotação de PoS tags. *Relatório Técnico*, n. 434. NILC-ICMC/USP, 54p. Disponível em: https://sites.google.com/icmc.usp.br/poetisa. Acesso em: 20/09/2021.

Eisenstein, J. (2013). What to do about bad language on the internet. In: Anais do NAACL-HLT, p. 359–369. Atlanta, EUA, ACL.

Foster, J. (2010). “cba to check the spelling”: investigating parser performance on discussion forum posts. In: Anais do NAACL-HLT, p. 381–384. LA, EUA, ACL.

Freitas, E.C.; Barth, P.A. (2015) Gênero ou suporte? O entrelaçamento de gêneros no Twitter. *Revista (Con)Textos Linguísticos*, 9(12), p. 08-26.

Kong, L., Schneider, N., Swayamdipta, S., Bhatia, A., Dyer, C., Smith, N.A. (2014). A dependency parser for tweets. In: Anais do EMNLP, p. 1001–12. Doha, Qatar.

Lyddy, F., Farina, F., Hanney, J., Farrell, L., O'Neill, N.K. (2014). An analysis of language in university students' text messages. *Journal of Computer-Mediated Communication,* 19(3), p. 546-561. Wiley Online Library.

[12] https://www.nltk.org/api/nltk.tokenize.html

Lynn, T., Scannell, K., Maguire, E. (2015). Minority language Twitter: part-of-speech tagging and analysis of Irish tweets. In: Anais do ACL'15 Workshop on Noisy User-generated Text, p. 1–8. July 31. Beijing, China, ACL.

Liu, Y., Zhu, Y., Che, W., Qin, B., Schneider, N., Smith, N.A. (2018). Parsing tweets into Universal Dependencies. In: Anais do NAACL-HLT, p. 965–975. LA, EUA, ACL.

Marcuschi, L.A. Produção textual, análise de gêneros e compreensão. Parábola Ed., 2008.

De Marneffe, M-C., Manning, C.D., Nivre, J. Zeman, D. (2021). Universal Dependencies. In *Computational Linguistics*, 47(2), p. 255-308. ACL. Online ISSN 1530-9312.

Nivre, J. (2015). Towards a Universal Grammar for Natural Language Processing. In: Anais do CICLing 2015. Lecture Notes in Computer Science, vol 9041, p. 3-16, Ed. by A. Gelbukh. Springer, Cham.

Nivre, J. et al. (2020). Universal Dependencies v2: an evergrowing multilingual treebank collection. In: Anais do 12º LREC. P. 4034-4043. Marseille, França. ELRA.

Owoputi, O., O'Connor, B., Dyer, C., Gimpel, K., Schneider, N., Smith, N.A. (2013). Improved part-of-speech tagging for online conversational text with word clusters. In: Anais do NAACL-HLT, p. 380–390. 9-14 de junho. Atlanta, Georgia. ACL.

Petrov, S., Das, D., McDonald, R. (2012). A universal part-of-speech tagset. In: Anais do 8º LREC, p. 2089–2096. 21-27 de maio. Istanbul, Turquia. ELRA.

Proisl, T. (2018). Someweta: A part-of-speech tagger for German social media and web texts. In: Anais do 11º LREC, p. 665–670. May 7-12. Miyazaki, Japão. ELRA.

Plutchik R., Kellerman, H. (eds). 1986. Emotion: theory, research and experience. Nova Iorque: Acad. Press

Rademaker, A.; Chalub, F., Real, L., Freitas, C., Bick, E., Paiva, V. (2017). Universal Dependencies for Portuguese. In: Anais do 4º Depling, p. 197-206. Pisa, Itália.

Rehbein, I., Ruppenhofer, J., Bich-Ngoc, D. (2019). tweeDe – a Universal Dependencies treebank for German tweets. In: Anais do 18º TLT, p. 100-108. Paris, França. ACL.

Reynar, J., Ratnaparkhi, A. (1997). A maximum entropy approach to identifying sentence boundaries. In: Anais do 5º ANLP, p. 16-19. Washington, EUA, ACL.

Seddah, D., Sagot, B., Candito, M., Mouilleron, V., Combet, V. (2012). The French social media bank: a treebank of noisy user generated content. In: Anais do 24º COLING, p. 2441–2458, Mumbai, Índia, ACL.

Sanguinetti, M. et al. (2018). PoSTWITA-UD: An Italian twitter treebank in Universal Dependencies. In: Anais do 11º LREC. p. 1768–75. Miyazaki, Japão. ELRA

Sanguinetti, M. et al. (2020a). Treebanking user-generated content: a proposal for a unified representation in universal dependencies. In: Anais do 12º LREC. p. 5240-50. Marseille, França. ELRA

Sanguinetti, M. et al. (2020b). Treebanking user-generated contente: a UD based overview of guidelines, corpora and unified recommendations. Available in: https://arxiv.org/abs/2011.02063. Access in: 25/09/2021.

Silva, F.J.V., Roman, N.T., Carvalho, A.M.B.R. (2020). Stock market tweets annotated with emotions. In: *Corpora*, 15(3), p. 343-354. Online ISSN: 1755-1676.

Silva, E.H., Pardo, T.A.S., Roman. N.T, Di-Felippo, A. Universal Dependencies for tweets in Brazilian Portuguese: tokenization and part of speech tagging. In: Anais do XVIII ENIAC 2021. 29 de nov. a 3 de dez., 2021. No prelo

APÊNDICE

C

EXPERIMENTO COM CÓRPUS MAC-MORPHO-V2

O Corpus Mac-Morpho (ALUÍSIO *et al.*, 2003) é um córpus composto por cerca de 1 milhão de palavras em português do Brasil. Ele foi criado a partir de textos de jornais e revistas brasileiras, coletados entre no ano de 1994. O córpus contém etiquetas morfossintáticas definidas a partir de um conjunto próprio de etiquetas, onde são encontradas 22 etiquetas-base e 9 complementares. Atualmente, o córpus foi atualizado com correções de sentenças problemáticas e mudanças no conjunto de etiquetas. A versão mais atualizada disponível denomina-se Mac-Morpho v2 (FONSECA; ROSA, 2013) é composta de 23 etiquetas-base e 7 complementares.

Por utilizar um formalismo desenvolvido pelos autores do trabalho, o conjunto de etiquetas é distinto do *Universal Dependencies* (UD). Dadas as diferentes características e o alto desempenho do etiquetador desenvolvido neste trabalho, esta seção realiza experimentos do mesmo etiquetador em um córpus que possui um diferente formalismo. O objetivo dessa análise é verificar se o desempenho do etiquetador no córpus Mac-Morpho v2 é comparável aos trabalhos da área que se baseiam neste córpus.

A metodologia deste experimento consiste em realizar o ajuste fino do modelo BERTimbau no córpus Mac-Morpho v2 e comparar a acurácia ao nível de *tokens* obtida no conjunto de dados de teste em relação a outros trabalhos disponíveis na literatura. Para a experimentação, foi utilizada a versão mais recente disponível do córpus[41]. Além disso, como etapa de pré-processamento, foi realizada a expansão das contrações dos *tokens*. O modelo foi treinado por cerca de 3 épocas com taxa de aprendizagem de $2e-5$ e tamanho de lotes de 32. Os demais hiper-parâmetros foram os mesmos apresentados no Capítulo 5.

Para realização da comparação com trabalhos da área, foram selecionados os seguintes métodos:

[41] Disponível em: <http://nilc.icmc.usp.br/macmorpho/>.

- SCS (FONSECA; ROSA, 2013) é uma rede neural totalmente conectada que utiliza janelas de *tokens* de tamanho fixo para construir representações vetoriais das características da sentença e, em seguida, é realizada a classificação.
- BLSTM (SOUSA; LOPES, 2019b) utiliza a arquitetura de redes neurais recorrentes Bi-LSTM em conjunto da concatenação de representações ao nível de caractere e *token* para realizar a classificação de cada etiqueta.
- FONSECA; ROSA; ALUÍSIO (2015) e SANTOS; ZADROZNY (2014) utilizam a arquitetura CharWNN para a tarefa de etiquetagem morfossintática. Os autores utilizam uma rede profunda que utiliza camadas convolucionais para gerar representações vetoriais ao nível de caractere.

A Tabela 22 apresenta as acurácias ao nível de *token* para cada método selecionado e para o etiquetador BERTimbau. Os resultados apresentados dos métodos selecionados foram extraídos dos artigos originais. É possível observar que o etiquetador BERTimbau obteve a maior acurácia nesta avaliação. Dessa forma, é possível observar que este método obteve resultados comparáveis ao estado-da-arte nos córpus baseados na UD e no córpus Mac-Morpho v2 que possui o conjunto de etiquetas próprio.

Tabela 22 – Acurácia ao nível de *tokens* no conjunto de testes do córpus Mac-Morpho v2

Modelo	Acurácia
SCS (FONSECA; ROSA, 2013)	96,48%
BLSTM (SOUSA; LOPES, 2019b)	97,62%
CharWNN (FONSECA; ROSA; ALUÍSIO, 2015)	97,31%
CharWNN (SANTOS; ZADROZNY, 2014)	97,47%
BERTimbau	**98,36%**

Fonte: Elaborada pelo autor.

Concluindo, este apêndice apresentou um experimento com o etiquetador BERTimbau no córpus Mac-Morpho v2, que utiliza um formalismo linguístico diferente dos córpus baseados na Universal Dependencies. A análise realizada verificou que o desempenho do etiquetador no córpus Mac-Morpho v2 foi comparável aos trabalhos da área que se baseiam neste córpus, bem como aos córpus baseados na UD. Isso indica que o etiquetador BERTimbau pode ser uma alternativa viável para tarefas de etiquetagem morfossintática em diferentes formalismos linguísticos. No entanto, é importante ressaltar que outras análises e experimentos podem ser realizados para verificar a generalização dos resultados apresentados aqui para outros córpus e tarefas linguísticas.


ICMC
USP
SÃO CARLOS