UNIVERSIDADE FEDERAL DE SÃO CARLOS

CENTRO DE EDUCAÇÃO E CIÊNCIAS HUMANAS – CECH

DEPARTAMENTO DE LETRAS

GABRIEL CEREGATTO

CARACTERIZAÇÃO MORFOSSINTÁTICA DE UM *CORPUS* DE *TWEETS* E ANÁLISE PRELIMINAR DE ERROS DE *TAGGING*

SÃO CARLOS - SP

2022

UNIVERSIDADE FEDERAL DE SÃO CARLOS

CENTRO DE EDUCAÇÃO E CIÊNCIAS HUMANAS – CECH

DEPARTAMENTO DE LETRAS

GABRIEL CEREGATTO

CARACTERIZAÇÃO MORFOSSINTÁTICA DE UM *CORPUS* DE *TWEETS* E ANÁLISE PRELIMINAR DE ERROS DE *TAGGING*

Trabalho de conclusão de curso apresentado ao Departamento de Letras da Universidade Federal de São Carlos, para obtenção do título de Bacharel em Linguística.

Orientadora: Profª. Drª. Ariani Di Felippo

SÃO CARLOS - SP

2022

UNIVERSIDADE FEDERAL DE SÃO CARLOS
CENTRO DE EDUCAÇÃO E CIÊNCIAS HUMANAS – CECH
DEPARTAMENTO DE LETRAS

Folha de aprovação

GABRIEL CEREGATTO

CARACTERIZAÇÃO MORFOSSINTÁTICA DE UM *CORPUS* DE *TWEETS* E ANÁLISE PRELIMINAR DE ERROS DE *TAGGING*

Orientadora:

_____

Profª. Drª. Ariani Di Felippo

Universidade Federal de São Carlos – UFSCar

Examinador:

_____

MS Jackson Wilke da Cruz Souza

Universidade Federal da Bahia – UFBa

Agradecimentos

Aos meus pais que, mesmo muitas vezes não compreendendo por completo a dimensão das responsabilidades na produção deste trabalho, nunca duvidaram do meu esforço e sempre me apoiaram por completo.

Aos membros do projeto POeTiSA, que me abriram muitas portas e me introduziram ao *corpus* que resultou neste trabalho.

À minha orientadora, Ariani, sempre muito paciente, empática e elucidativa. Este trabalho não existiria sem seu esforço e compreensão inigualável.

RESUMO

A etiquetação morfossintática automática ou *tagging* é um dos primeiros processos para a interpretação de língua natural em sistemas ou aplicações do Processamento Automático das Línguas Naturais (PLN). Sendo definida como a identificação da classe gramatical de cada uma das palavras ou *tokens* de um texto, a tarefa gera conhecimento relevante para os demais processos do sistema/aplicação, como a análise sintática. Dada a relevância das redes sociais, muitas pesquisas sobre *tagging* têm sido conduzidas com vistas ao processamento dos diferentes tipos de "conteúdo gerado por usuário" (CGU). A respeito do arcabouço gramatical e do recurso linguístico, a maioria delas tem se apoiado no modelo *Universal Dependencies* (UD) e na construção de *corpora* anotados de *tweets* (*tweebanks*). Neste trabalho, fez-se primeiramente a caracterização estatística da anotação morfossintática de referência do *corpus* DANTEStocks, que, contendo *tweets* do mercado de ações, é o primeiro *tweebank* com anotação UD em português. Como resultado, verificou-se que (i) as postagens tendem a ser fragmentadas e compostas por usos informais de sinais de pontuação (como a reduplicação), o que é evidenciado pela alta frequência da *tag* PUNCT; (ii) os *tweets* parecem ter uma natureza nominal, uma vez que NOUN e PROPN são muito frequentes, ao contrário de VERB; (iii) as interjeições são raramente empregadas pelos usuários como forma de manifestação de sentimentos e emoções, dada a baixa frequência de INJT, e (iv) a limitação de caracteres imposta pela plataforma parece colaborar para que a estrutura dos *tweets* do mercado de ações não seja complexa, uma vez que CCONJ, SCONJ e AUX possuem baixa frequência. Na sequência, realizou-se uma análise inicial dos erros cometidos pelo método UDPipe 2.1 na anotação PoS de uma parcela do *corpus.* A análise resultou em um conjunto de regras de pós-edição de *tagging,* as quais ainda precisam ser avaliadas, e no levantamento do grau de generalização delas. Ademais, ela permitiu verificar que (i) a grande maioria dos erros de *tagging* se referem a conhecimento de língua geral e não de domínio e (ii) os fenômenos CGU (lexical e/ou ortográficos) do *corpus* parecem pouco interferir no processo de *tagging*, uma vez que uma minoria dos erros é relativa a *tokens* caracterizados por esses fenômenos. Com isso, acredita-se que este trabalho contribui para os estudos linguístico-descritivos e para o PLN.

Palavras-chave: *corpus*; *tweet*; análise morfossintática automática.

ABSTRACT

Part-of-speech tagging is one of the first processes for natural language interpretation in Natural Language Processing (NLP) systems or applications. Being defined as the identification of the grammatical category of each word or token in a text, the tagging task generates relevant knowledge for other processes of the system/application, such as the syntactic analysis or *parsing*. Given the relevance of social networks, many researches on tagging have been developed focusing on processing different types of "user-generated content" (UGC). Regarding the grammatical framework and the linguistic resource, most of the research on tagging has relied on the Universal Dependencies (UD) model, and on the construction of annotated tweet corpora (also called tweebanks). In this work, we first performed the statistical characterization of the *gold-standard* morphosyntactic annotation of the DANTEStocks corpus. Such resource comprises tweets from the stock market domain, and it is the first tweebank with UD annotation in Portuguese. As a result, it was found that (i) the posts in the corpus tend to be fragmented and composed of informal uses of punctuation marks (such as reduplication), which is evidenced by the high frequency of the PUNCT tag; (ii) the tweets seem to have a nominal structure, since NOUN and PROPN are highly frequent, unlike VERB; (iii) interjections are rarely used by users as a way of expressing feelings and emotions, given the low frequency of INJT, and (iv) the character limitation posed by the platform seems to have influence on the simplicity (syntactic) structure of the tweets, avoiding CCONJ, SCONJ and AUX. Next, we carried out an initial analysis of the tagging errors made by UDPipe 2.1 in the annotation of a subset of tweets from DANTEStocks. This analysis resulted in a set of post-editing tagging rules, which still need to be evaluated, and in the classification of the rules according to their degree of generalization. Furthermore, we could found that (i) the vast majority of errors made by the tagging method refer to general language knowledge and not domain knowledge and (ii) the CGU (lexical and/or orthographic) phenomena of the corpus seem to have low influence on the tagging process, since a minority of errors are related to tokens characterized by these phenomena. With that, we believe that this work contributes to the linguistic-descriptive studies and to NLP.

Keywords: *corpus*; *tweet*; *tagging*.

LISTA DE FIGURAS

LISTA DE QUADROS

LISTA DE TABELAS

SUMÁRIO

1 INTRODUÇÃO

No Processamento Automático das Línguas Naturais (PLN), objetiva-se construir sistemas capazes de interpretar e/ou gerar língua natural, mais comumente na modalidade escrita (JURAFSKY, MARTINS, 2022). A tradução automática é uma das aplicações mais tradicionais do PLN (MITKOV, 2005). Com a crescente relevância das mídias sociais, outras aplicações vêm sendo investigadas e desenvolvidas, como a análise de sentimento e a detecção automática de *fake news*.

Todas essas aplicações precisam realizar a etiquetação morfossintática (ou *tagging*)*,* que consiste em identificar a categoria gramatical de cada palavra (ou *token*) de um texto de entrada a partir de um conjunto de *tags* ou etiquetas pré-definido (denominado *tagset*) (JURAFSKY, MARTINS, 2022; MITKOV, 2005). Trata-se de uma das primeiras tarefas a ser realizada pelos sistemas ou aplicações de PLN em direção à interpretação linguística. Assim definida, fica evidente a importância da etiquetação morfossintática para, por exemplo, a tarefa de *parsing* que a sucede*,* uma vez que a identificação das relações sintáticas, sejam elas sintagmáticas ou de dependência, requer a identificação prévia da classe das palavras.

A tarefa de *tagging* para textos formais (jornalísticos, por exemplo) tem sido amplamente investigada. Para o português, Aires *et al.* (2000), Fonseca *et al*., (2015) e Souza e Lopes (2019) evidenciam que várias ferramentas de etiquetação morfossintática ou etiquetadores (em inglês, *part-of-speech* (PoS) *taggers*) têm sido desenvolvidos sob diferentes paradigmas ou abordagens de PLN, como as clássicas, baseadas em regras, ou como as mais recentes, baseadas em Aprendizado de Máquina (AM) que utilizam redes neurais. De um modo geral, a tarefa de *tagging* que lida com textos jornalísticos atingiu acurácia excelente, entre 97% e 98%.

O cenário é diferente quando se trata do desenvolvimento de ferramentas de *tagging* para o processamento do "Conteúdo Gerados por Usuários" da *web* (ou CGU) (do inglês, *User-generated content* - UGC), como os *tweets*[1]), uma vez que o interesse por esse tipo de material linguístico é relativamente recente, sendo fomentado pela relevância crescente das redes sociais e do processamento automático de seu conteúdo.

[1] Entende-se todo conteúdo (seja vídeo, imagem ou texto) postado por usuários em plataformas como *Facebook*, *Twitter*, *blogs*, *chats*, páginas de avaliação etc. Neste trabalho, no entanto, esse termo se restringe a conteúdo em formato textual.

Mesmo com avanços na última década, principalmente para o inglês (p.ex.: LYNN *et al*. 2015; BOSCO *et al*. 2016; PROISL 2018; REHBEIN *et al*. 2018; BEHZAD, ZELDES 2020), a etiquetação morfossintática dos diferentes tipos de CGUs é ainda um desafio. Com acurácia entre 86% e 93% (SILVA, 2022), a linguagem dos CGUs apresenta características de informalidade e outras, típicas da plataforma ou do domínio, que dificultam a identificação automática das classes de palavras.

A maioria dos trabalhos recentes sobre PoS *tagging* no PLN se alinha à iniciativa do projeto *Universal Dependencies* (NIVRE, 2015; NIVRE et al, 2016 NIVRE et al., 2020) e às técnicas de redes neurais artificiais e modelos distribucionais (SANGUINETTI *et al.,* 2020). Em linhas gerais, a UD fornece um esquema "universal" para representar a morfologia e a sintaxe das línguas. Sendo a evolução do esforço colaborativo de uma ampla comunidade de pesquisa, a UD se tornou referência para a construção de *treebanks* (isto é, *corpora* (morfo-)sintaticamente anotados), inclusive de *treebanks* compostos por *tweets* (morfo-)sintaticamente anotados (os chamados *tweebanks*). E isso se deve, sobretudo, à adaptabilidade do modelo às particularidades dos *tweets* em todos os níveis de representação.

Seguindo esse cenário, Silva (2022) realizou as primeiras pesquisas sobre o processo de PoS *tagging* baseado em UD para CGU (no caso, *tweets*) em português. No caso, eles customizaram métodos de *tagging* do estado-da-arte para o português. Para tanto, os autores utilizaram o DANTEStocks, que é um *corpus* de *tweets* do mercado de ações que está sendo pioneiramente anotado segundo o modelo UD (DI FELIPPO *et al*., 2021). Com base nesse recurso linguístico*,* os autores treinaram vários métodos do estado-da-arte com o DANTEStocks, sendo que o de melhor performance foi o UDPipe 2.1, que atingiu 95% f-*score* para a tarefa de *tagging.*

Dada a relevância dos *tweebanks* na literatura geral e a do DANTEStocks para o processamento de CGU em português, apresentam-se, neste trabalho, uma caracterização ou descrição da distribuição das etiquetas PoS na anotação manual desse *corpus* e uma análise preliminar dos erros cometidos pelo UDPipe 2.1 (STRAKA; STRAKOVÁ; HAJIC, 2019). A referida descrição poderá ser comparada à distribuição das etiquetas PoS em um *corpus* de linguagem formal, como a jornalística, contribuindo para a compreensão da linguagem dos *tweets.* A análise dos erros teve como objetivo compreender as confusões (entre classes de palavras) realizadas pelo *tagger* e propor estratégias de pós-edição para futura correção dos erros.

Para apresentar a pesquisa, este relatório está organizado em 5 Seções. Na Seção 2, apresenta-se uma breve revisão da literatura sobre *tagging* segundo o modelo UD, especialmente para o português, destacando o *corpus* DANTEStocks como recurso linguístico fundamental para essa investigação. Na seção 3, apresenta-se o primeiro resultado deste trabalho, que é a caracterização da distribuição estatísticas das *tags* PoS na anotação de referência do *corpus.* Na seção 4, apresenta-se uma análise preliminar dos erros de *tagging* cometidos pelo UDPipe 2.1, que foi o método de *tagging* que obteve o melhor desempenho quando customizado para processar o DANTEStocks, atingindo medida-f de 95%*.* Na Seção 5, por fim, são apresentadas as considerações finais deste trabalho, enfatizando contribuições, limitações e trabalhos futuros.

2 REVISÃO DA LITERATURA

2.1 Etiquetação Morfossintática Automática ou *Tagging*

A etiquetação morfossintática automática ou *tagging*, como mencionado, é a tarefa de identificar as classes das palavras de um texto de entrada. A partir de um conjunto pré-definido de etiquetas ou *tags,* a ferramenta precisa ser capaz de identificar adequadamente a *tag* que representa a classe de cada uma das palavras em contexto, desambiguando-as quando necessário. O conjunto de etiquetas ou *tagset* utilizado nessa tarefa pode variar bastante.

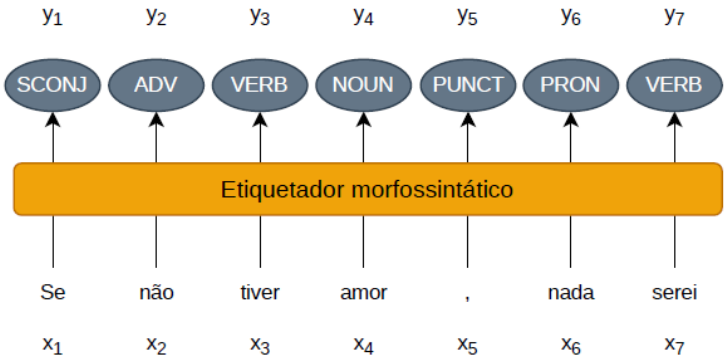
Para que a análise morfossintática de um texto seja feita, o *tagger* precisa realizar dois processos prévios: (i) segmentação sentencial, que é comumente feita com base na ocorrência de sinais de pontuação, e (ii) *tokenização*.

A *tokenização*, em particular, consiste em identificar ou delimitar as palavras que serão etiquetadas. Estas, em português, são frequentemente separadas umas das outras por espaços em branco, mas o espaço em branco nem sempre é suficiente. A depender da base teórica ou mesmo da aplicação de PLN a qual o *tagger* se destina, *Nova York*, por exemplo, pode ser tratada como um *token* individual, embora contenha espaço, enquanto *do*, por exemplo, precisa ser decomposto em *de* e *o.*

Tendo em vista que a tarefa em questão permeia os níveis morfológico e sintático de descrição linguística, o etiquetador pode considerar a forma e o contexto para associar mais adequadamente uma palavra a sua possível classe/*tag* gramatical. Como uma palavra pode ter mais de uma classe gramatical possível, seu co-texto, (isto é, as palavras ao seu redor) pode ajudar na desambiguação, contribuindo, assim, para identificar a etiqueta de maior probabilidade (JURAFSKY, MARTIN, 2022). Para ilustrar, destaca-se que, na sentença *O banco de madeira quebrou*, a ocorrência das demais palavras ao redor de “banco” indica que há baixa probabilidade dessa palavra ser da classe dos verbos (bancar).

Na Figura 1, tem-se uma arquitetura genérica de um método de análise ou etiquetação morfossintática. Nela, vê-se que o etiquetador morfossintático recebe como entrada uma lista de *tokens*, representada pelo vetor *xn*, e produz um vetor de saída *yn*. A forma como se dá a associação entre os *tokens* e as classes gramaticais dependerá da abordagem e do algoritmo de *tagging* empregados na tarefa.

FIGURA 1: ARQUITETURA GENÉRICA DE UM ETIQUETADOR MORFOSSINTÁTICO.



Fonte: SILVA (2022).

Na literatura atual, há métodos de etiquetação morfossintática desenvolvidos segundo diferentes abordagens, como métodos baseados em regras pré-definidas por linguistas (KLEIN; SIMMONS, 1963), métodos probabilísticos (KEPLER, 2010), conexionistas (BOHNET et al., 2018), entre outros.

Para a avaliação dos etiquetadores, utilizam-se algumas métricas, como (i) acurácia, que é obtida pela divisão do número total de etiquetas corretas pelo número total de etiquetas, (ii) precisão e cobertura, que medem, respectivamente, quantos etiquetas corretas foram detectadas em relação a todas que foram detectadas pelo método e quantas etiquetas corretas foram detectadas em relação a todas que deveriam ter sido identificadas, e (iii) medida-f, que é a média harmônica entre precisão e cobertura, reduzindo a avaliação a uma única métrica (JURAFSKY, MARTIN, 2022).

Com base nessas métricas, os autores evidenciam que os métodos mais recentes de melhor desempenho se fundamentam na abordagem conexionista e fazem uso de redes neurais artificiais, atingindo acurácia excelente para textos jornalísticos, entre 97% e 98%.

2.2 Etiquetação Morfossintática Automática de CGUs

Os avanços na última década relativos à etiquetação morfossintática dos diferentes tipos de CGUs, principalmente para o inglês (p.ex.: LYNN *et al*. 2015; BOSCO *et al*. 2016; PROISL 2018; REHBEIN *et al*. 2018; BEHZAD, ZELDES 2020), devem-se principalmente às técnicas mais recentes de PLN, baseadas em AM e redes neurais, e ao desenvolvimento do modelo gramatical UD, juntamente com os *corpora* anotados segundo o modelo.

Embora tenha havido avanços, a acurácia da etiquetação morfossintática para GCU está entre 86% e 93% (SILVA, 2022), que é inferior à obtida para textos jornalísticos. A dificuldade em processar os CGUs decorre de sua linguagem, que apresenta características de informalidade e outras, típicas das plataformas ou dos domínios, as quais dificultam a identificação automática das classes de palavras.

Em uma publicação bastante recente, Sanguinetti *et al.* (2022) lista 30 *corpora* de CGU em diferentes línguas (cf. Anexo 1), construídos entre 2011 e 2020. Nesse conjunto, observa-se que a maioria dos recursos possui anotação segundo o modelo UD. Ademais, observa-se a predominância dos *tweets* como material linguístico constitutivo desses *corpora.*

Seguindo a literatura internacional, Silva (2022) investiga de forma pioneira a tarefa de PoS *tagging* baseada no modelo UD para CGU (no caso, *tweets*) em português. Como mencionado, ele customizou os métodos de *tagging* do estado-da-arte para CGU o português. Especificamente, o autor investigou os métodos UPipe 1 (STRAKA; HAJIČ; STRAKOVÁ, 2016), UDPipe 2.1 (STRAKA; STRAKOVÁ; HAJIC, 2019) e Udify (KONDRATYUK; STRAKA, 2019), os quais utilizaram representações contextuais baseadas no BERT (DEVLIN et al., 2019) e em rede neural.

Para tanto, os autores utilizaram o DANTEStocks, que é um *corpus* de *tweets* do mercado de ações que está sendo pioneiramente anotado segundo o modelo UD (DI FELIPPO *et al*., 2021). Com base nesse recurso linguístico*,* os autores treinaram os vários métodos do estado-da-arte com o DANTEStocks, sendo que o de melhor performance foi o UDPipe 2.1, que atingiu 95% medida-f para a tarefa de *tagging.* Silva (2022) realizou a investigação dos métodos de *tagging* com base na anotação morfossintática de referência do *corpus,* isto é, revisada por humanos.

Após a descrição do *corpus* DANTEStocks na subeção 2.3., apresenta-se, na seção 3, a caracterização da anotação morfossintática de referência do mesmo, assim como apresenta-se, na seção 4, uma análise prévia dos erros cometidos pelo UDPipe 2.1.

2.3. O *corpus* DANTEStocks

Nesta seção, destacam-se a origem do DANTEStocks, o modelo gramatical empregado na construção de sua versão de referência (isto é, cuja anotação foi manualmente revisada) e as diretrizes de pré-processamento e anotação de PoS.

a) O ponto de partida

O DANTEStocks originou-se do recurso compilado por Silva *et al* (2020), que inicialmente continha 4,517 postagens coletadas automaticamente do *Twitter* com base na ocorrência de um código ou *ticker* de pelo menos uma das 73 ações que compõem o índice IBOVESPA, que é o principal indicador de desempenho das ações negociadas na B3. Um *ticker* é um código alfanumérico (normalmente quatro letras e um número) que representa a empresa e o tipo da ação. No exemplo (1), o *ticker* VALE5 corresponde às ações preferenciais de classe A da empresa Vale do Rio Doce. Os *tweets* foram compilados de março a maio de 2014 e, assim, possuem no máximo 140 caracteres. Segundo Silva *et al.,* o corpus já possui anotação manual de emoção, que foi feita com base nos quatro eixos de oposição emocional da teoria de Plutchik (PLUTCHIK, KELLERMAN, 1986) (*joy* vs *sadness, anger* vs *fear, trust* vs *disgust* e *surprise* vs *anticipation*). Por exemplo, o *tweet* “PETR4 só me traz alegrias” recebeu os seguintes rótulos para 3 dos pares emocionais: *joy*, *trust* e *surprise*.

A linguagem dos *tweets* é informal e marcada por dispositivos típicos da plataforma. A estrutura dos *tweets* do DANTEStocks, por exemplo, varia bastante. Há *tweets* formados por uma ou mais sentenças claramente delimitadas como em (1). Mas há também *tweets* que apresentam ausência de pontuação (2) ou pontuação equivocada (3). Em (2), o *tweet* parece ser composto por duas sentenças e essa interpretação pode ser corroborada pela capitalização do segundo “o” após “antecipada”. Em (3), o exemplo é de uso inadequado da vírgula após “Marcos”, em substituição ao ponto de exclamação. *Tweets* relativamente fragmentados (4) também compõem o *corpus*. O exemplo (5) traz um *tweet* cuja interpretação só é possível diante da mensagem completa e de certo conhecimento de domínio.

(1) Parece que VALE5 vai abrir forte. No leilão de abertura tá a 28,16.
(2) O #PT conseguiu fazer propaganda eleitoral antecipada O que a @user tem a dizer sobre isso?
(3) Bom dia Marcos, Alguma previsão para petr4?!
(4) $BBDC3 - Bradesco (bbdc-n1) - Demonstracoes Financeiras De 31/12/2013 (individual) http://t.co/Mh92PHzdlD
(5) @user fala para sua queridinha # #vale5 para 26,20 ??? Ta me irritando...

O *corpus* também apresenta dispositivos ortográficos e lexicais que evidenciam fenômenos como (DI-FELIPPO *et al.,* 2021): (a) *simplificação de código* (que reduzem o esforço de escrita de um *token*); (b) *abreviação informal* (sequência de caracteres que representa de forma reduzida várias palavras), (c) *expressão de sentimento* (que emulam sentimento expresso pela prosódia, expressão facial ou gesto), (d) *influência de língua estrangeira* (vocábulo formado com base em outra língua, como o verbo "estopar" (do inglês *"stop"*), que significa interromper venda ou compra de um ativo), (e) *expressão de oralidade* (palavra cuja grafia remonta à comunicação (fala) informal, as quais são, por vezes, empregadas com função humorística), (f) *elemento metalinguístico* (elemento que tipicamente ocorre no *Twitter*), e (g) *fenômeno de domínio* (fenômeno exclusivo do DANTEStocks).

b) O modelo *Universal Dependencies*

O modelo UD provê um esquema de representação nos níveis morfológico e sintático. A descrição morfológica de uma palavra sintática consiste em três níveis de representação: (i) um lema, que representa o conteúdo semântico da palavra; (ii) uma etiqueta morfossintática (*part-of-speech* ou tag PoS), que codifica categoria gramatical da palavra, e (iii) um conjunto de *features,* que captura propriedades lexicais e gramaticais associadas à palavra. A anotação sintática consiste de relações de dependência (*deprel*) entre palavras. Uma dependência é estabelecida entre uma palavra sintaticamente dependente e outra palavra da qual ela depende (denominada, *head* ou núcleo). A versão atual da UD (v2) (NIVRE *et al.,* 2016), a qual foi utilizada para a anotação do DANTESTocks, engloba 17 tag PoS*s*[2] e 37 relações de dependência (ou *deprels*).

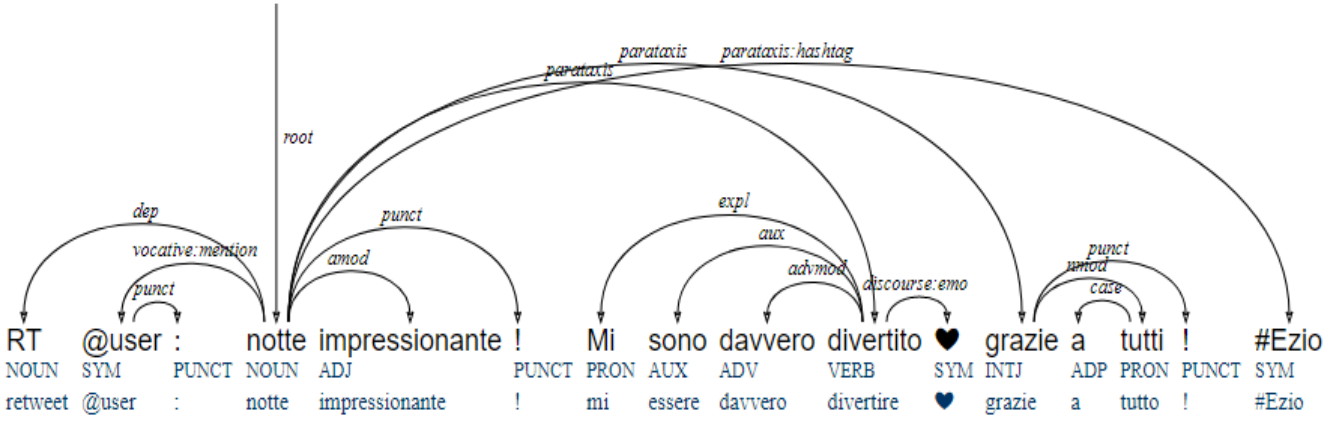
Na Figura 2, ilustra-se a anotação UD de um *tweet* extraído da versão 2.1 do *corpus* em italiano PoSTWITA-UD (SANGUINETTI *et al.,* 2018).

Logo abaixo às formas de superfície que compõem o *tweet,* tem-se as etiquetas de PoS representados por *tags* em letras maiúsculas, como NOUN para *notte* ("noite"). Logo abaixo das *tags*, estão as formas canônicas em letras minúsculas. As *deprels* estão indicadas por setas rotuladas que se originam no *head* e se direcionam ao dependente. Na Figura 1, o adjetivo *impressionante,* por exemplo,

[2] A versão atual da UD (v2), há 17 PoS *tags*, sendo 6 para palavras de conteúdo (ADJ, ADV, INTJ, NOUN, PROPN, VERB), 8 para palavras funcionais (ADP, AUX, CCONJ, DET, NUM, PART, PRON, SCONJ) e 3 artificiais (PUNCT, SYM, X), que não correspondem a categorias morfossintáticas.

é dependente de *notte*, que é seu *head*, havendo entre eles a *deprel* **amod** (*adjectival modifier*) (isto é, modificador adjetival). O nome *notte,* aliás, foi anotado como *root* (raiz) desse *tweet*. Embora as *features* estejam ausentes na Figura 2, *notte*, por exemplo, possui os atributos-valores: *Gender=Masc* e *Number=Sing*.

FIGURA 2: EXEMPLO DE UM *TWEET* EM ITALIANO COM ANOTAÇÃO UD.



Fonte: http://universal.grew.fr/?corpus=UD_Italian-PoSTWITA@2.10#

c) <u>O pré-processamento do *corpus*</u>

Para a anotação-UD do DANTEStocks, o *tweet* foi tomado como unidade de análise sintática (DI-FELIPPO *et al.,* 2021). Isso significa que os *tweets* não passaram por nenhum processo de segmentação em unidades menores, como sentenças ou mesmo sintagmas, o que normalmente ocorre quando se lida com textos de linguagem formal. Essa decisão foi tomada com base na pontuação assistemática verificada das postagens desse *corpus* (cf. exemplos (1)-(5)), que dificultaria a segmentação automática. Dessa forma, os *tweets* foram anotados na íntegra, assim como ocorreram e foram compiladas da plataforma.

Além disso, não se aplicou nenhum tipo de normalização (isto é, conversão da linguagem em um formato "padrão" ou mais formal) aos dados do *corpus.* Essa decisão foi norteada pelo interesse futuro do projeto DANTE/POeTiSA[3] de utilizar o *corpus* para o desenvolvimento de ferramentas e aplicações voltados para dados reais. Assim, o estatuto de palavra sintática dos fenômenos do *corpus* foi analisado e decisões de tokenização foram definidas (Quadro 1). No Quadro 1, a expressão "*no*

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

change” indica que a forma de superfície foi reconhecida como *token.* Já “*split*” indica que a forma de superfície sofreu decomposição em mais de um *token.*

Quanto à simplificação de código, destaca-se que a ausência de hífen e a diretriz de não normalizar os dados originais levaram ao reconhecimento automático de 2 *tokens* (“sexta” e “feira”), os quais foram anotados com suas respectivas categorias de origem (isto é, NOUN e NOUN) segundo o manual de anotação de PoS da língua portuguesa (DURAN, 2021). No entanto, segundo a visão lexicalista da UD, um composto hifenizado como “sexta-feira” constitui uma única palavra sintática (ou *token*) e essa informação precisará estar capturada na anotação. Para esse caso, uma alternativa pode ser a utilização da *deprel* `goeswith`, mas isso só será definido quando a anotação sintática (das relações de dependência) for realizada.

Sobre as abreviaturas, vale ressaltar que as contrações no DANTEStocks são formas abreviadas de duas palavras funcionais com remoção de espaços e/ou letras. Nessa categoria, temos formas superficiais como “oq”, constituída por dois pronomes (“o” “que”), e formas como “pq” (“por” “que”), que reduzem palavras de categorias morfossintáticas diferentes (preposição e pronome, respectivamente). Para ambos os tipos de contrações, a diretriz de tokenização foi a de decompô-las em dois *tokens*. Os inicialismos, por sua vez, consistem em formas compostas pelas letras iniciais de palavra comuns, como “LTA” (“linha de tendência de alta”). Os inicialismos foram considerados *tokens* únicos e, por isso, não sofreram nenhum tipo de decomposição. Isso se deu porque um inicialismo desempenha função sintática específica.

Quanto às expressões de sentimento, as autocensuras, como “m*” (“merda”), e os *emoticons,* como “o.O” (“surpresa”), foram reconhecidos como *tokens* únicos, pois podem desempenhar funções sintáticas específicas. Ademais, sequências de *emoticons* (como “**:):)**”) foram decompostas e, assim, cada *emoticon* foi considerado um *token* individual. O mesmo critério foi adotado para as repetições de pontuação.

As marcas de oralidade, em particular as “expressões cristalizadas”, como “né” e “dae” (“daí”), foram decompostos em 2 *tokens,* uma vez que são originalmente contrações de “não é” e “de aí”, respectivamente.

Sobre os elementos metalinguísticos, os truncamentos lexicais ocorrem no fim de um *tweet* devido ao limite de caracteres. Seguindo a literatura, um elemento truncado foi reconhecido como *token*. No que diz respeito às *hashtags, cashtags* e menções, o reconhecimento dos símbolos “$” e “@” como parte constitutiva dos

tokens varia na literatura. No DANTEStocks, eles foram considerados como tal, compondo *tokens* únicos com as palavras ou expressões que precedem.

QUADRO 1 - OS FENÔMENOS UGC NO DANTESTOCKS E SUA TOKENIZAÇÃO.

Tipo	**Fenômeno**	**Exemplo/Forma padrão**	**Tokenização**	
			No change	*Split*
Lexplificação de *código*	Omissão/adição de diacrítico	proprio (*próprio*), fêz (*fez*)	✓	
	Ausência de hífen	sexta feira (sexta-feira)		✓
	Substituição de diacrítico	eh (*é*), tou (*tô*)	✓	
	Omissão de letras	d (*de*), qdo (*quando*), pq (*porque*)	✓	
	Erro ortográfico/digitação	compra (*compra*)	✓	
	Fonetização	k (*que*), kd (*cadê*), krk (*caraca*)	✓	
Abrev	Contração	oq (o que), pq (*por que*)		✓
	Inicialismo	LTA (*linha de tendência de alta*)	✓	
Expressão / sentimento	Repetição de pontuação	Foi**!!!** (*Foi!*)		✓
	Alongamento grafêmico	LI**NN**DA (*linda*)	✓	
	Autocensura	p**a (*puta*), m* (*merda*)	✓	
	Emoticon	**o.O** (*surpresa*), **:)** (*sorriso*)	✓ (individual)	✓ (sequência)
Marca de oralidade	Coloquialismo	guvêrno (*governo*), bão (*bom*)	✓	
	Expressão cristalizada	né (*não é*), dae (*daí*)		✓
	Exclamação onomatopeica	hahaha	✓	
Elemento metalinguístico	*Hashtag*	#Petr4	✓	
	Menção	@user	✓	
	Marca de *retweet*	**RT** @user	✓	
	URL	http://t.co/sROpyWPblN	✓	
	Truncamento (lexical)	Ação sobe **fo...** (*forte*)	✓	
Fenômeno do domínio (Ibovespa)	*Ticker*	Petr4		
	Cashtag	$LREN3	✓	
	Indeterminação decimal	De 18,**xx** a 21,00	✓	
	Índice de (des)valorização	+2,09%, -11,42%		✓(3 *tokens*)
	Substituição lexical	muito **$** (dinheiro)	✓	
	Expressão temporal híbrida	1T14	✓	
	Valor monetário aglutinado	R$20,00		✓(2 *tokens*)

Influência / outra língua	Formação verbal	estopar	✓	

Fonte: Elaborada pelo autor.

Quanto aos fenômenos de domínio, os índices de (des)valorização das ações foram decompostos em 3 *tokens* (“+2,09%” → “+” “2,09” “%”). O reconhecimento de “+” como *token* (no caso, um símbolo) justificou-se pela possibilidade de substituí-lo por outra palavra (como “a ação subiu 2,09%”). As formas reduzidas de expressões temporais, como “1T14” (“primeiro trimestre de 2014”), funcionam como uma unidade e, portanto, foram consideradas palavras únicas ou *tokens*. No DANTEStocks, as expressões monetárias, quando aglutinadas (isto é, sem espaço entre o símbolo monetário e o numeral) (“R$20,00”), foram decompostas em 2 *tokens* “R$” e “20,00”.

As estratégias sistematizadas no Quadro 1 foram “traduzidas” em regras contextuais e implementadas no tokenizador simbólico de *tweets* do pacote NLTK, o qual, uma vez customizado para os fenômenos do *corpus,* foi utilizado para a tokenização do DANTEStocks. Os resultados desse processo foram manualmente revisados (cf. SILVA, 2022).

d) A anotação de PoS do DANTEStocks

Após a tokenização, realizou-se a anotação do *corpus* segundo o modelo UD. Para a construção do primeiro *tweebank* em português, optou-se por trabalhar separadamente cada um dos níveis de anotação da UD. Isso foi feito com o objetivo de simplificar a tarefa e produzir resultados melhores para cada nível, pois a anotação UD é altamente sofisticada (PARDO *et al.,* 2021). Assim, a construção do DANTEStocks teve início com a anotação semiautomática das etiquetas morfossintáticas ou *tags* PoS.

Para tanto, o total de 4,517 *tweets* foi dividido em 13 pacotes. Com exceção do pacote 0, que tinha 147 *tweets* para treinamento dos anotadores, e o pacote 13, que continha o restante dos *tweets,* os outros 11 continham 370 *tweets* cada (cf. Tabela 1). O *corpus* foi dividido nesses pacotes para que a anotação automática de PoS pudesse ser feita de forma incremental. Tendo em vista que não havia um *tagger* para CGU em português, a anotação do pacote 1 foi feita pelo *parser* UDPipe 2.1 (STRAKA, 2018), o qual até então havia sido treinado para lidar apenas com textos

formais em português (no caso, notícias jornalísticas). Assim, após a revisão manual da anotação automática de PoS do pacote 1, este foi utilizado para retreinar o UDPipe2 de forma a prepará-lo para anotar o pacote 2, e assim incrementalmente até que todos os pacotes tivessem sido anotados e revisados.

A revisão manual de PoS no DANTEStocks foi feita com o auxílio da versão refinada, por Miranda e Pardo (2022), da ferramenta *online* Arborator-Grew (GUIBON *et al.,* 2020), desenvolvida para as tarefas de anotação/revisão de *corpora* anotados segundo a UD.

Ademais, a revisão foi guiada pela consulta a dois materiais de suporte, sendo ambos construídos a partir das 17 etiquetas de PoS da UD. Um deles foi o manual de Duran (2022), que instancia as diretrizes de anotação de PoS da UD para o português, e o outro foi o manual de Di-Felippo *et al.* (2022), que faz essa mesma instanciação só que para os fenômenos típicos dos *tweets* do mercado de ações.

No manual de Di-Felippo *et al.*, é possível identificar, por exemplo, que um truncamento (como “técni” em (6)) deve ser anotado em função da classe de sua forma completa (“técnico”). Caso esta não tenha sido recuperada (e nem inferida), o truncamento deve ser anotado com a *tag* X. Ou ainda que os inicialismos devem ser anotados com a categoria morfossintática do *head* da expressão correspondente, como LTA em (7), anotado como NOUN porque o núcleo da expressão é “linha”.

(6)

Original: Petrobrás Pn (PETR4), Gráfico Diário. Estudo técni... http://t.co/5HHwxBvF6W
Tokenizado: Petrobrás Pn (PETR4) , Gráfico Diário . Estudo **técni** ... http://t.co/5HHwxBvF6W
Anotação PoS: (PROPN, PROPN, PUNCT, PROPN, PUNCT, PUNCT, NOUN, ADJ, PUNCT, NOUN, **ADJ**, PUNCT, SYM)

(7)

Original: lembra da **LTA** da #PETR4 ? Hoje está passando pelos R$18,10 ...o.O
Tokenizado: lembra de a **LTA** de a #PETR4 ? Hoje está passando por os R$ 18,10 ... o.O
Anotação PoS: (VERB, ADP, DET, **NOUN**, ADP, DET, PROPN, PUNCT, ADV, AUX, VERB, ADP, DET, SYM, NUM, PUNCT, SYM)

Destaca-se que cada um dos 13 pacotes de *tweets* foi submetido à revisão de três anotadores humanos e somente os casos de divergência entre eles foram adjudicados por meio de arquivos .xls.

Na Seção seguinte, apresenta-se a distribuição das *tags* PoS na anotação manual do DANTEStocks, isto é, na anotação de referência (ou *gold standard*) do *corpus.* Tal caracterização é o primeiro resultado deste trabalho

3. CARACTERIZAÇÃO DAS *TAGS* POS NO DANTESTOCKS

Uma vez que o corpus foi anotado e revisado, procedeu-se, neste trabalho, à descrição da distribuição (estatística) da frequência geral das etiquetas morfossintáticas no DANTEStocks, isto é, na versão de referência do *corpus.*

Antes, porém, o conjunto de *tweets* passou por um refinamento manual, que consistiu na exclusão de *tweets* repetidos e de postagens que não eram do domínio em questão.

No caso dos *tweets* repetidos, excluíram-se todas as mensagens repostadas (isto é, com marca de RT (*retweet*)), cujas postagens originais fizessem parte do *corpus.* Em (8), por exemplo, o *tweet* descrito em (b) foi excluído, uma vez que o *tweet* descrito em (a) integrava o *corpus.*

Quanto às postagens que não eram do domínio do mercado de ações, ressalta-se que a exclusão destas se mostrou necessária porque a compilação automática pautada na ocorrência de *tickers* extraiu postagens que não expressavam conteúdo sobre ações da Ibovespa. O *tweet* em (9), por exemplo, foi excluído, pois, embora compilado pela ocorrência de “Cruz3”, não se refere à ação ordinária da empresa Souza Cruz.

(8a) @garimpodeacoes #CSNA3 ILG = 0,23 ILC = 0,78 ILS = 0,40 Se achar alguma coisa boa lá, vende !

(8b) RT @ppaulovagner : @garimpodeacoes #CSNA3 ILG = 0,23 ILC = 0,78 ILS = 0,40 Se achar alguma coisa boa lá , vende !

(9) “1Antes de empezar - Santiago Cruz3”

Na Tabela 1, tem-se os dados estatísticos referentes ao processo de refinamento. Nesse quadro, os dados foram organizados em função dos 13 pacotes nos quais os *tweets* foram divididos. Ademais, ressalta-se que, em alguns pacotes, houve intersecção entre os casos de exclusão. No pacote 1, por exemplo, 1 mesmo *tweet* foi excluído simultaneamente por não ser do domínio e por ser um *retweet* (embora, nesse caso, a mensagem original não integrasse o *corpus*).

Observa-se que, ao final, o DANTEStocks passou de 4,517 para 4,048 *tweets,* agora distintos e efetivamente do domínio do mercado de ações.

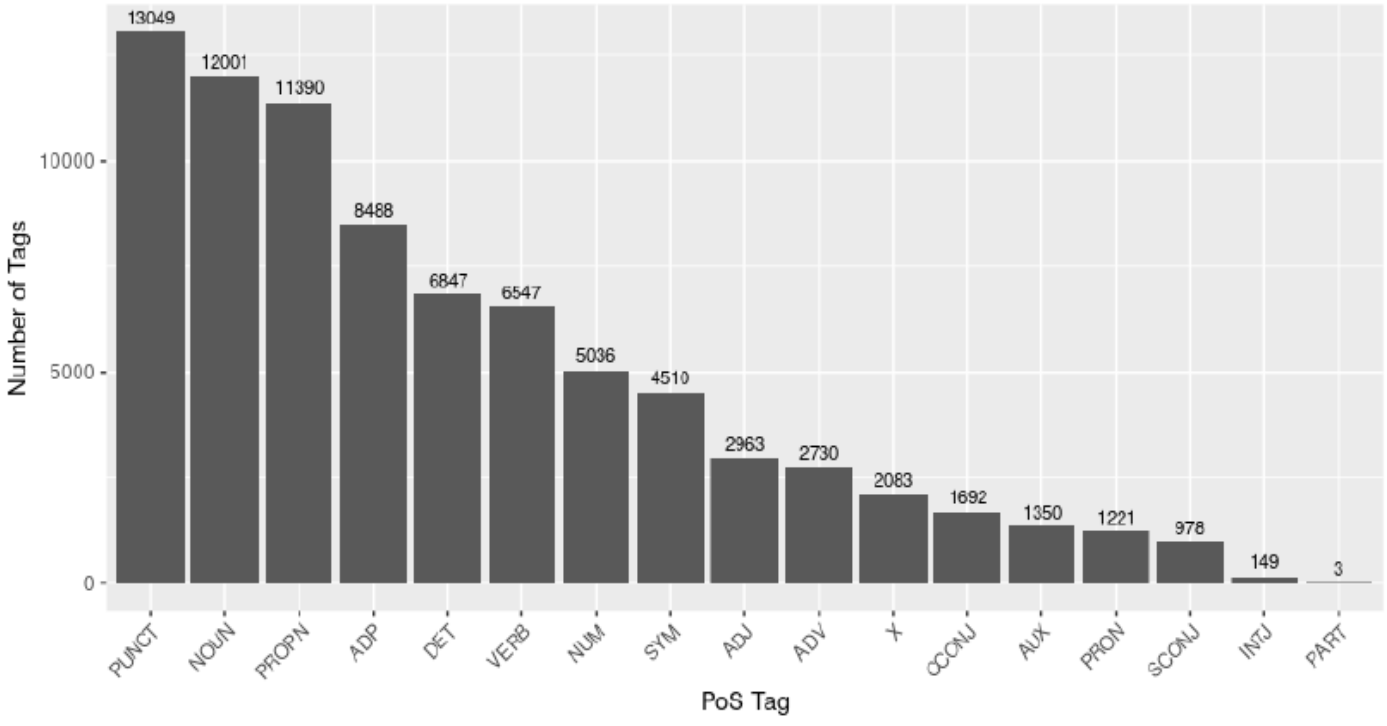
TABELA 1 - ESTATÍSTICA SOBRE O REFINAMENTO DO DANTESTOCKS.

Pacote	Qt. de *tweets*	Qt. de *tweets* *"out-domain"*	Qt. de *tweets* repetidos	Qt. de intersecções	Qt. de *tweets* distintos	Qt. final de *tweets*
0	147	7	4	0	11	136
1	370	30	9	1	38	332
2	370	35	15	3	47	323
3	370	30	9	1	38	332
4	370	25	15	2	38	332
5	370	24	16	2	38	332
6	370	23	7	1	29	341
7	370	29	13	1	41	329
8	370	29	20	2	47	323
9	370	20	12	2	30	340
10	370	32	10	1	41	329
11	370	35	15	2	48	322
12	300	15	9	1	23	277
TOTAL	**4517**	**334**	**154**	**19**	**469**	**4048**

Fonte: Elaborada pelo autor.

Assim, a caracterização das *tags* foi feita a partir do conjunto final, pós-refinamento. Isso quer dizer que a distribuição geral de cada *tag* presente na Figura 3 diz respeito ao total de 4,048 *tweets* do *corpus.*

FIGURA 3 - DISTRIBUIÇÃO DAS *TAGS POS* NO DANTESTOCKS.



Fonte: Elaborada pelo autor.

Com essas na Figura 3, é possível identificar algumas características do corpus. Percebe-se inicialmente que, no DANTEStocks, ocorreram todas as 17 etiquetas de classe de palavra previstas pela v2 da UD.

Destas, as mais frequentes foram PUNCT, NOUN e PROPN.

A alta frequência da etiqueta PUNCT pode ser justificada por duas razões, que são: (i) a própria estrutura fragmentada da maioria dos *tweets*, composta por sequências de elementos muitas vezes separados por sinais de pontuação, e (ii) a diretriz de tokenização referente às sequências repetidas de sinais de pontuação. Em alguns *corpora* de *tweets* da literatura, os autores optaram por delimitar uma sequência de sinais de pontuação repetidos como um *token* individual. Assim, a sequência "!!!!!!" equivale a 1 *token*. Como essa escolha da literatura não foi justificada, optou-se por seguir a diretriz referente aos sinais de pontuação presente no manual para a língua portuguesa, que é a de delimitar cada sinal de pontuação como um *token* individual. Com isso, a sequência-exemplo "!!!!!!" equivale a 6 *tokens* no DANTEStocks.

A etiqueta NOUN tem alta frequência devido à natureza nominal nos *tweets* do DANTEStocks. Isso quer dizer que a maior parte das postagens do *corpus* são compostas por sintagmas nominais, sejam eles longos ou curtos, fragmentados ou não. Essa observação é corroborada pela relativa baixa ocorrência de verbos (VERB e AUX) no *corpus.*

Já a alta frequência da etiqueta PROPN pode ser justificada, em partes, pela diretriz de anotação UD adotada para os *tickers.* Tais *tokens,* como "PETR4" em (6), foram anotados como PROPN, pois o reconhecimento manual dos nomes próprios no DANTESTtocks está intimamente ligado ao conceito de entidade nomeada. Sendo os *tickers* as pistas linguísticas utilizadas para a compilação dos *tweets* na plataforma, estes ocorrem pelo menos uma vez em todas as postagens, o que justifica a ocorrência elevada da *tag* PROPN.

As etiquetas PART e INTJ foram as menos frequentes no *corpus*. A etiqueta PART foi utilizada na anotação de apenas 3 casos. Dois deles dizem respeito à ocorrência (livre) do prefixo "pré", resultante da grafia ou digitação equivocada da palavra "pré-abertura". No caso, a ocorrência de "pré abertura" sem hífen acarretou na identificação de dois *tokens* ("pré" "abertura"), sendo "pré-" anotado como PART. O outro caso diz respeito à ocorrência, também livre, do prefixo "des-", uma vez que o neologismo "(dês)Graça Foster" foi decomposto em 3 *tokens*.

Quanto a INTJ, vale ressaltar que sua baixa frequência no *corpus* pode ser justificada pelo fato de essa *tag* ser usada, segundo a UD, para anotar palavras usadas exclamativamente, as quais expressam uma reação emocional e não estão sintaticamente relacionadas a outros *tokens* que a acompanham. No caso dos *tweets*, e em especial os do mercado de ações, as emoções dos usuários são expressas por meio de outras estratégias ou dispositivos linguísticos, como *emojis* e prolongamentos grafêmicos, o que explicaria a ocorrência relativamente baixa de INTJ.

Ainda sobre as etiquetas com baixa frequência, levanta-se a hipótese de que PRON ocorra pouco devido ao domínio, uma vez que os *tweets* do mercado de ações tendem a ser muito factuais, não contendo ocorrência marcante de certos tipos de pronomes, como o pessoal, reflexivo, interrogativo ou demonstrativo. Acredita-se que a *tag* PRON foi majoritariamente atribuída a pronomes relativos, mas isso ainda precisa ser verificado quando a anotação das *features* gramaticais tiver sido realizada.

A *tag* SCONJ é usada para anotar uma palavra que introduz orações subordinadas substantivas ou adverbiais e a *tag* CCONJ rotula uma palavra que liga palavras ou constituintes maiores sem que um esteja subordinado ao outro. Sendo assim, acredita-se que a baixa frequências destas esteja relacionada à limitação de caracteres das postagens produzidas no *Twitter*. O mesmo pode ser dito para a *tag* AUX. Como os verbos auxiliares fazem parte de construções na voz passiva, mais longa que a ativa, acredita-se que a limitação de caracteres leve o usuário a optar por outras construções, mais curtas.

Por fim, observa-se na Figura 3 que a etiqueta X também é relativamente rara no DANTEStocks. Segundo as diretrizes de anotação da UD, essa *tag* deve ser usada para rotular os casos não cobertos pelas demais etiquetas do modelo. No *corpus* em questão, a etiqueta X cobre exatamente os fenômenos que não podem ser rotulados pelas demais *tags,* como é o caso de (i) *hashtags* e *cashtags* que funcionam apenas como indexadores*,* (ii) símbolos de *retweet* (RT), (iii) truncamentos cujas formas completas não foram recuperadas ou inferidas, (iv) *tokens* compostos por frases ou expressões complexas (p.ex: #EuApoioCPIdaPetrobras e #SardenbergResponde) e (v) *tokens* em língua estrangeira que não fazem parte do domínio do mercado de ações (p.ex.: #Whoknows).

Na próxima seção, apresenta-se a análise preliminar dos erros cometidos pelo UDPipe 2.1, que foi o modelo de melhor desempenho segundo Silva (2022).

4. DESCRIÇÃO E ANÁLISE PRELIMINAR DOS ERROS DE *TAGGING*

O *corpus* utilizado por Silva (2022) para o estudo dos métodos de *tagging* para CGU em português ainda não havia sido refinado, contendo, assim, os 4,517 *tweets* originalmente compilados por Silva *et al.* (2020). Ademais, quando da investigação dos métodos, havia apenas 8 (0-7) pacotes com anotação de referência, isto é, cuja anotação automática havia sido manualmente revisada. Dessa forma, os métodos foram investigados com base em apenas 8 pacotes, totalizando 2,737 *tweets.*

Na Tabela 2, tem-se a medida-f para cada *tag* obtido pelo UDPipe 2.1 com base nos 8 pacotes de *tweets* disponíveis. Na Tabela, as *tags* foram ordenadas de forma descrente em função da medida-f. Com isso, a *tag* de medida-f mais elevada ocupa o topo da lista e a de medida-f mais baixa ocupa a última colocação na tabela.

TABELA 2 – MEDIDA-F DAS *TAGS* OBTIDA PELO UDPipe 2.1.

Tag	UDPipe 2.1	Qt.
CCONJ	99.33	1692
PUNCT	99.30	13049
SYM	99.13	4510
NUM	97.99	5036
DET	97.70	6847
ADP	97.45	8488
PROPN	95.77	11390
VERB	94.88	6547
ADV	94.69	2730
NOUN	94.20	12001
AUX	92.39	1350
PRON	92.10	1221
ADJ	90.46	2963
SCONJ	89.90	976
X	79.06	2083
INTJ	44.44	149

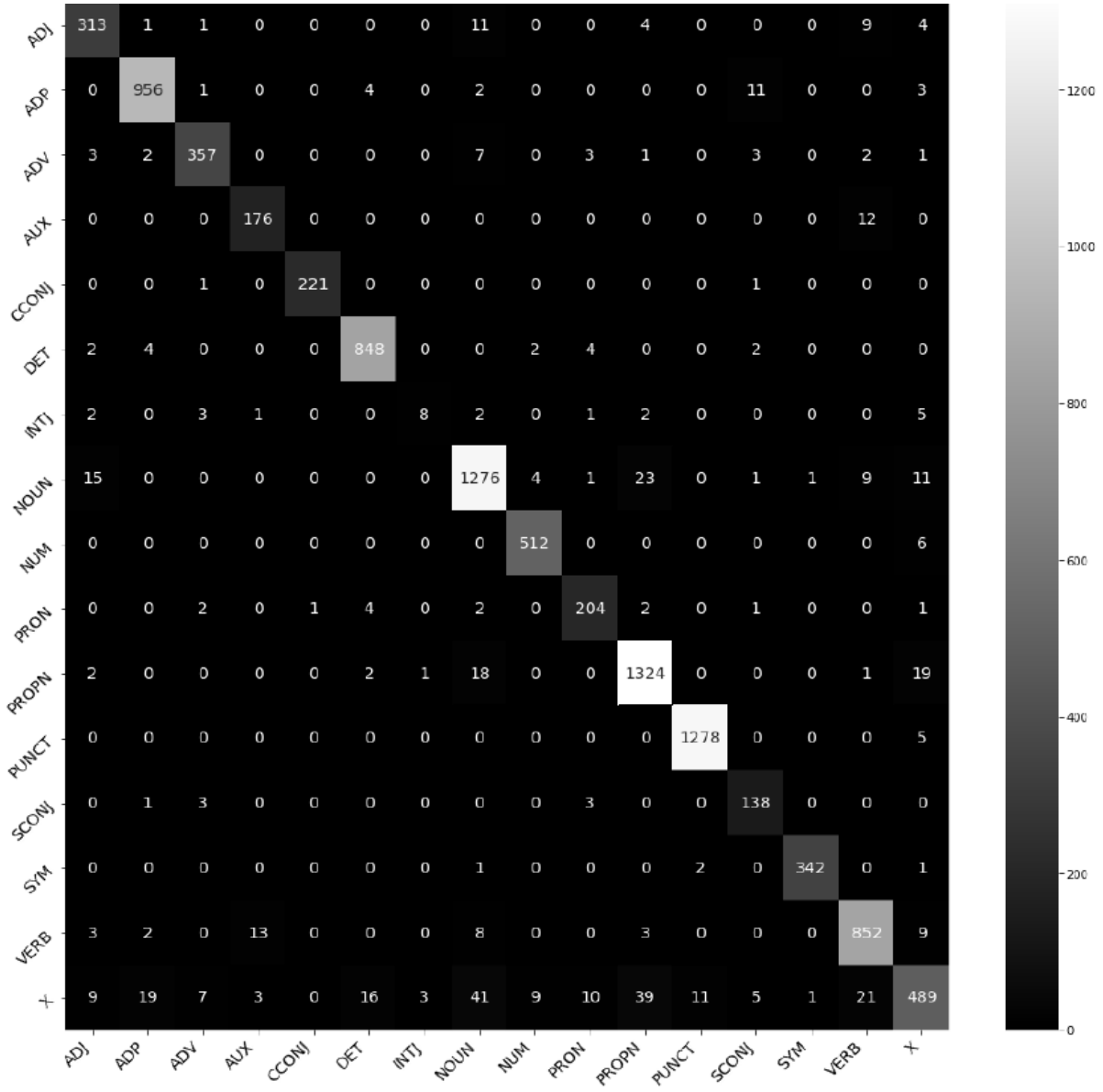
Fonte: Silva (2022)

Com base na Tabela 2, observa-se que o método de *tagging* apresenta resultados que variam de 44,44% (para INTJ) a 99,33% (para CCONJ). Uma possível razão para essa discrepância de valores pode ser o fato de a INTJ ser uma *tag* pouco frequente no *corpus,* havendo, por isso, poucas ocorrências para o aprendizado do método. Como mencionado, as interjeições não perecem ser a forma mais usual de expressar

emoções e sentimentos nos *tweets* sobre o mercado de ações. Embora com uma medida-f mais elevada que INTJ, a *tag* X também possui uma medida-f (79.06%) relativamente mais baixa se comparada a obtida para as outras classes de palavras. Isso pode se justificar pelo fato de que, mesmo com um total de 2.083 ocorrências nos 8 pacotes de *tweets,* a etiqueta X cobre um conjunto variado de fenômenos, o que parece dificultar a identificação automática dessa *tag*.

Na Figura 4, tem-se a matriz de confusão referente aos resultados da Tabela 2. A matriz auxilia na avaliação da classificação feita pelo UDPipe 2.1, uma vez que consiste em uma tabela comparativa das *tags* que o modelo trouxe como predição em relação às *tags* da anotação manual.

FIGURA 4 - MATRIZ DE CONFUSÃO DOS RESULTADOS DO UDPIPE 2.1.



Fonte: Silva (2022).

Na 1ª linha da matriz, vê-se, por exemplo, que, dos 343 *tokens* anotados manualmente como ADJ, o UDPipe 2.1 anotou corretamente 313 e anotou erroneamente 30 *tokens,* sendo 1 deles como ADP, 1 como ADV, 11 como NOUN, 4 como PROPN, 9 como VERB e 4 como X; e isso resultou na medida-f de 90.46%. A Tabela 3 apresenta os números absolutos dos erros e acertos em função de cada *tag*.

TABELA 3 - QUANTIDADE DE ERROS E ACERTOS POR *TAG*.

tag PoS	Qt total de casos	Qt de acertos	Qt de erros
ADJ	343	313	30
ADP	977	956	21
ADV	379	357	22
AUX	188	176	12
CCONJ	223	221	2
DET	862	848	14
INTJ	24	8	16
NOUN	1341	1276	65
NUM	513	512	1
PRON	217	204	13
PROPN	1367	1324	43
PUNCT	1283	1278	5
SCONJ	145	138	7
SYM	346	342	4
VERB	890	852	38
X	656	489	167
TOTAL	**9762**	**9311**	**468**

Fonte: Elaborada pelo autor.

A partir da matriz de confusão, procedeu-se à análise dos erros de *tagging*. Para tanto, a *tag* PROPN, assim como todos os casos de confusão das demais *tags* com PROPN, não foram analisados neste trabalho, pois a anotação manual dos nomes próprios no DANTEStocks estava sendo revisada no momento da realização do trabalho aqui apresentando. O mesmo critério foi aplicado para a exclusão dos erros de AUX e das confusões envolvendo esse *tag*. Assim como PROPN e AUX, a etiqueta X e todos os casos de confusão das demais *tags* com X também foram excluídos da análise dos erros de *tagging*. Essa última exclusão foi feita porque muitos dos diferentes fenômenos rotulados por X no DANTEStocks estavam sob revisão manual, sendo reanotados em função de novas diretrizes de anotação UD.

De forma mais concreta, as linhas referentes às *tags* PROPN, AUX e X da matriz de confusão da Figura 4 foram excluídas para a análise dos erros, assim como todos os casos contidos nas colunas que se referem às etiquetas PROPN, AUX e X. Descrevendo a matriz da esquerda para a direita, a coluna relativa a AUX é a 4ª, a de PROPN é a 11ª e a coluna relativa a X é a última. Tais colunas compreendem os casos com os quais as demais *tags* foram confundidas com AUX, PROPN e X.

Como consequência dessas exclusões, os 4 casos de confusão entre ADJ e PROPN e os 4 casos de confusão entre ADJ e X, por exemplo, foram excluídos do total de erros de ADJ analisados. Assim, do total de 30 erros cometidos pelo UDPipe 2.1. para ADJ (cf. Tabela 3), restaram 22 casos.

Na Tabela 4, apresenta-se o total de erros após a exclusão das etiquetas PROPN, AUX e X e dos casos de confusão entre estas e as demais *tags*. Observa-se com base na Tabela 4 que o total de erros referente às etiquetas NUM e PUNCT foi zerado, pois tais erros consistiam exclusivamente em confusões entre estas e a etiqueta X, excluída da análise.

TABELA 4 – QUANTIDADE DE ERROS PÓS EXCLUSÃO DE PROPN, AUX E X.

tag PoS	Qt total	Qt de acertos	Qt de erros inicial	Qt de erros final
ADJ	343	313	30	22
ADP	977	956	21	18
ADV	379	357	22	20
CCONJ	223	221	2	2
DET	862	848	14	14
INTJ	24	8	16	8
NOUN	1341	1276	65	31
NUM	518	512	6	**0**
PRON	217	204	13	10
PUNCT	1283	1278	5	**0**
SCONJ	145	138	7	7
SYM	346	342	4	3
VERB	890	852	38	13
TOTAL	**7548**	**7305**	**243**	**148**

Fonte: Elaborada pelo autor.

Após a exclusão de PROPN, AUX e X, procedeu-se à identificação dos casos de erros que ocorriam em *tweets* que já não mais integravam o *corpus.* Isso foi necessário porque, como mencionado, o *corpus* utilizado por Silva (2022) para o estudo dos

métodos de *tagging* para CGU em português ainda não havia passado pelo processo de refinamento descrito anteriormente, que consistiu na exclusão de *tweets* repetidos e de postagens que não expressavam efetivamente conteúdo sobre o mercado de ações.

Ao fazer essa verificação, observou-se também que alguns erros envolviam problemas na anotação de referência do DANTEStocks, assim como outros englobavam erros de digitação no co-texto (isto é, nas palavras que ocorrem ao redor da *tag*-problema) (ou de *tokenização*). A Tabela 5 resume esse segundo refinamento dos dados para a análise dos erros. Com base na Tabela 5, todos os casos de erros que envolviam (i) problemas na anotação de referência, (ii) erros de digitação no co-texto ou de *tokenização* e (ii) ocorrência em *tweets* previamente excluídos do *corpus* não foram analisados neste trabalho. Os casos que envolvem o fator (ii) foram especificamente desconsiderados porque erros de digitação no co-texto ou de *tokenização* dificultariam a proposição de regras de pós-edição. Na Tabela 5, a coluna "total (de erro) por *tag*" expressa a quantidade final de casos efetivamente investigados. Do total de 148 erros da Tabela 4, restaram 128 casos para análise. Quanto aos erros de CCONJ, observa-se que, após o segundo refinamento, não restaram erros para analisar

TABELA 5 - QUANTIDADE DE ERROS PARA ANÁLISE APÓS SEGUNDO REFINAMENTO.

***Tag-alvo* (casos)**	***Confusão***	**Qt.**	**Erro no *corpus* de referência**	**Erro de digit./ *token.***	**Exclusão prévia do *corpus***	**Total por confusão**	**Total por *tag***
ADJ (22 casos)	ADP	1	1	0	0	0	19
	ADV	1	0	0	0	1	
	NOUN	11	1	0	0	10	
	VERB	9	1	0	0	8	
ADP (18 casos)	ADV	1	0	0	0	1	14
	DET	4	2	1	0	1	
	NOUN	2	0	0	0	2	
	SCONJ	11	1	0	0	10	
ADV (20 casos)	ADJ	3	0	0	0	3	19
	ADP	2	0	0	0	2	
	NOUN	7	0	0	0	7	
	PRON	3	0	0	1	2	
	SCONJ	3	0	0	0	3	
	VERB	2	0	0	0	1	

CCONJ (2 casos)	ADV	1	1	0	0	0	0
	SCONJ	1	1	0	0	0	
DET (14 casos)	ADJ	2	0	0	0	2	14
	ADP	4	0	0	0	4	
	NUM	2	0	0	0	2	
	PRON	4	0	0	0	4	
	SCONJ	2	0	0	0	2	
INTJ (8 casos)	ADJ	2	0	0	0	2	8
	ADV	3	0	0	0	3	
	NOUN	2	0	0	0	2	
	PRON	1	0	0	0	1	
NOUN (31 casos)	ADJ	15	2	0	1	12	23
	NUM	4	0	0	0	4	
	PRON	1	0	0	0	1	
	SCONJ	1	1	0	0	0	
	SYM	1	0	0	0	1	
	VERB	9	2	2	0	5	
PRON (10 casos)	ADV	2	0	0	0	2	8
	CCONJ	1	1	0	0	0	
	DET	4	1	0	0	3	
	NOUN	2	0	0	0	2	
	SCONJ	1	0	0	0	1	
SCONJ (7 casos)	ADP	1	0	0	0	1	7
	ADV	3	0	0	0	3	
	PRON	3	0	0	0	3	
SYM (3 casos)	NOUN	1	0	0	0	1	3
	PUNCT	2	0	0	0	2	
VERB (13 casos)	ADJ	3	0	0	0	3	13
	ADP	2	0	0	0	2	
	NOUN	8	0	0	0	8	
	TOTAL	**148**	**15**	**3**	**2**	**128**	**128**

Fonte: Elaborada pelo autor.

Na sequência, fez-se uma análise inicial dos 128 erros. Diz-se inicial porque a análise realizada neste trabalho focou apenas em propor possíveis regras de pós-edição de *tagging* e caracterizar linguisticamente os equívocos do método de *tagging* em função da linguagem dos CGU. As regras de pós-edição, no entanto, não foram avaliadas neste trabalho, uma vez que o entendimento dos erros e a proposição de regras se revelaram tarefas mais demoradas e laboriosas que se supôs no início do trabalho.

No Apêndice 1, exibe-se o conjunto total de 93 regras propostas, as quais têm o potencial de serem empregadas após o processo de *tagging,* isto é, podem funcionar como estratégia de pós-edição. Para ilustração, o Quadro 2 exibe apenas as regras propostas para as correções dos erros relativos a *tag* ADJ. Especificamente, para os 19 casos de erros de ADJ da Tabela 5, foram propostas 15 regras. No Quadro 2, a Regra 1, por exemplo, prevê a correção da etiquetação equivocada de “certo” como ADV para ADJ. A Regra 14, por exemplo, prevê a correção da etiquetação equivocada do adjetivo “elevado” (em “risco elevado”) como VERB. A condição para a Regra 14 é a ocorrência da palavra “risco” imediatamente à esquerda do *token* “elevado”, o que é indicado por “se *token* 1E=risco”. Caso essa condição seja satisfeita, a regra prevê a substituição da *tag* VERB por ADJ.

QUADRO 2 – REGRAS DE PÓS-EDIÇÃO PARA OS ERROS REFERENTES A *TAG* ADJ.

Qt	DE	PARA	Condição	Regra	Exemplo	Casos
1	ADV	ADJ	Se token="certo"	Substituir ADV por ADJ	certo/ADJ>ADV	1
2	NOUN	ADJ	Se o token for precedido por DET e NOUN	Substituir por NOUN por ADJ	em/ADP a/DET sexta/NOUN **passada**/ADJ>NOUN	2
3	NOUN	ADJ	Se o token for imediatamente precedido por DET e sucedido por NOUN.	Substituir NOUN por ADJ	a/DET **corrente**/NOUN>ADJ gravidade/NOUN	1
4	NOUN	ADJ	Se o token for imediatamente precedido por ADP e sucedido por NOUN.	Substituir NOUN por ADJ	de/ADP **módico**/NOUN>ADJ repique/NOUN	1
5	NOUN	ADJ	Se o token anterior for NOUN, substituir tag do token "log" por ADJ.	Substituir NOUN por ADJ	escala/NOUN **log**/ADJ>NOUN	1
6	NOUN	ADJ	Se token="casado"	Substituir NOUN por ADJ	**casado/ADJ>NOUN**	1
7	NOUN	ADJ	Se token="coitada", primeiro token da sentença e seguido por ADP	Substituir NOUN por ADJ	**coitada/ADJ>NOUN**	1
8	NOUN	ADJ	Se o token for seguido por PUNCT (/) + ADJ	Substituir NOUN por ADJ	**médio/ADJ>NOUN /\>PUNCT longo/NOUN**	1
9	NOUN	ADJ	Se token="melho"	Substituir qualquer tag PoS por ADJ	**melho/NOUN>ADJ**	1
10	NOUN	ADJ	Se token="voláteis"	Substituir qualquer tag PoS por ADJ	**voláteis/NOUN>ADJ**	1
11	VERB	ADJ	Se token 1E e 2E = ADJ + NOUN	Substituir VERB por ADJ	lucro/NOUN líquido/ADJ **consolidado**/ADJ>VERB	2
12	VERB	ADJ	Se token 1E="trade"	Substituir VERB por ADJ	Trade/NOUN **fechado**/ADJ>VERB	1
13	VERB	ADJ	Se token="lindo"	Substituir qualquer tag PoS por ADJ	lindo/ADJ>VERB	1
14	VERB	ADJ	Se token 1E="risco"	Substituir VERB por ADJ	risco/NOUN **elevado**/ADJ>VERB	1
15	VERB	ADJ	Se tokens 1D, 2D e 3D = ADP + DET + NOUN/PROPN	Substituir VERB por ADJ	para/ADP o/DET bebê/NOUN **abandonado**/ADJ>VERB	3

Fonte: Elaborada pelo autor.

Após a proposição das regras, fez-se uma classificação delas com base no seu grau de generalização. Em outras palavras, buscou-se verificar o número de regras compostas por condições lexicalizadas ou generalizadas (isto é, compostas por *tags*

PoS). As regras 1 e 14 descritas anteriormente são exemplos de regras lexicalizadas, uma vez que suas condições pressupõem a ocorrência de certos itens lexicais ou *tokens*. Um exemplo de regra generalizada (ou com maior grau de abstração) é a Regra 3, cuja condição para a adequada etiquetação de um *token* (como *passada* em "*a sexta passada*") como ADJ é ocorrência de um DET à sua esquerda e de um NOUN à sua direita. Verificando, assim, o grau de generalização, observou-se que há 61 regras lexicalizadas e 32 não-lexicalizadas. Tais dados podem ser verificados no Apêndice 1. Para os erros de ADJ, por exemplo, 9 das 15 regras são lexicalizadas e 6 não são (Quadro 3).

QUADRO 3 – CLASSIFICAÇÃO DAS REGRAS PARA OS ERROS DE ADJ.

Regra	Tipo
1	Lexicalizada
2	Não-lexicalizada
3	Não-lexicalizada
4	Não-lexicalizada
5	Lexicalizada
6	Lexicalizada
7	Lexicalizada
8	Não-lexicalizada
9	Lexicalizada
10	Lexicalizada
11	Não-lexicalizada
12	Lexicalizada
13	Lexicalizada
14	Lexicalizada
15	Não-lexicalizada

Fonte: Elaborada pelo autor.

Na sequência, procedeu-se à caracterização dos erros do UDPipe 2.1 em função da linguagem, buscando verificar se os erros tinham relação com a linguagem dos CGUs e se eles envolviam fenômenos típicos dos *tweets.* Para tanto, os erros foram classificados em língua geral ou de domínio. Com base nos dados do Apêndice 1, observa-se que a grande maioria (cerca de 72%) dos erros são relativos a palavras

ou *tokens* da língua geral e que o restante diz respeito ao léxico do domínio. Para os erros de ADJ, por exemplo, apenas 1 deles foi classificado como sendo relativo ao vocabulário de domínio, como registrado no Quadro 4. Para a caracterização em função dos fenômenos CGU do DANTESTtocks, utilizou-se a tipologia de fenômenos apresentada no Quadro 1 para anotar os erros. Os dados referentes a essa descrição também estão no Apêndice 1. No geral, identificaram-se 18 erros cujos *tokens* sob anotação são marcados por fenômenos típicos dos CGU, sendo 7 erros envolvendo abreviações, 3 erros referentes a coloquialismos ou marcas de oralidade, 1 erro envolvendo palavra estrangeira, 1 erro relativo a símbolos e 6 erros de anotação de palavras truncadas. No caso dos erros de ADJ, por exemplo, apenas 1 dos 15 se caracteriza por ser um truncamento (Quadro 4).

QUADRO 4 – CLASSIFICAÇÃO DOS ERROS EM FUNÇÃO DA LINGUAGEM CGU.

Regra	Qt de casos	Palavra-exemplo	Linguagem	Fenômeno CGU
1	1	certo	Língua geral	
2	2	passada	Língua geral	
3	1	corrente	Língua geral	
4	1	módico	Língua geral	
5	1	*log*	Domínio	
6	1	casado	Língua geral	
7	1	coitada	Língua geral	
8	1	médio	Língua geral	
9	1	melho (melhor)	Língua geral	Truncamento
10	1	voláteis	Língua geral	
11	2	consolidado	Língua geral	
12	1	fechado	Língua geral	
13	1	lindo	Língua geral	
14	1	elevado	Língua geral	
15	3	abandonado	Língua geral	

Fonte: Elaborada pelo autor.

5. CONSIDERAÇÕES FINAIS

Dada a relevância dos *tweebanks* e sobretudo do DANTEStocks para o processamento de CGU em português, este trabalho focou (i) na caracterização estatística da anotação morfossintática de referência do *corpus* e, na sequência, (ii) em uma análise inicial dos erros de *tagging* no DANTEStocks.

A caracterização da anotação de referência do DANTEStocks evidenciou certas peculiaridades desse conjunto de *tweets* do mercado de ações. Particularmente, o levantamento da distribuição da frequência geral das *tags* PoS permitiu que se observasse o seguinte sobre o *corpus*:

a) as postagens tendem a ser fragmentadas e compostas por usos informais de sinais de pontuação (como a reduplicação), o que é evidenciado pela alta frequência da *tag* PUNCT;
b) os *tweets* parecem ter uma natureza nominal, uma vez que NOUN e PROPN são altamente frequentes, ao contrário de VERB;
c) as interjeições são raramente empregadas pelos usuários como forma de manifestação de sentimentos e emoções, já que INTJ é de fato a *tag* de menor frequência (uma vez que PART cobre apenas 3 casos eventuais e não regulares);
d) a limitação de caracteres imposta pela plataforma parece colaborar para que a estrutura dos *tweets* do mercado de ações não seja complexa, uma vez que CCONJ, SCONJ e AUX possuem baixa frequência.

Diante disso, sabe-se mais hoje sobre a linguagem do DANTEStocks. A caracterização da anotação morfossintática também contribuiu especificamente para o refinamento do *corpus*, uma vez que propiciou a identificação e posterior exclusão de *tweets* repetidos e/ou *out-domain.*

A análise dos erros de *tagging* cometidos pelo UDPipe 2.1 para os 8 pacotes de *tweets* até então disponíveis permitiu que se propusesse um conjunto de regras que podem ser empregadas em um processo de pós-edição. Mesmo não tendo sido avaliadas, essas regras podem ser consideradas o ponto de partida para o enriquecimento de métodos de *tagging* (estatísticos e/ou probabilísticos, como o UDPipe 2.1) com conhecimento linguístico. Ademais, a classificação das regras segundo seu grau de generalização/abstração aponta que os erros para os quais as regras lexicalizadas foram propostas talvez possam ser tratados alternativamente por meio de léxicos ou dicionários, uma vez que o erro diz respeito a um *token* particular.

Destaca-se também que a grande maioria dos erros cometidos pelo UDPipe 2.1 se referem a conhecimento de língua geral e não ao léxico do domínio. Por fim, observou-se também que os fenômenos lexical e/ou ortográficos presentes no *corpus* parecem pouco interferir na etiquetação morfossintática automática, já que apenas 18 erros de *tagging* são relativos a *tokens* caracterizados por esses fenômenos, como abreviações, coloquialismos ou truncamentos.

Quanto às limitações deste trabalho, reconhece-se que a não avaliação das regras é uma delas. No entanto, acredita-se que as contribuições geradas são relevantes para as pesquisas no PLN e também para a compreensão do *tweet* como um dos gêneros de CGU e para a compressão da linguagem CGU do domínio do mercado de ações.

O desenvolvimento deste trabalho de conclusão de curso envolveu grandes desafios, como a familiarização com o modelo UD e a anotação manual de um *corpus* de *tweets* segundo esse modelo (o DANTEStocks), e, principalmente, a análise e proposição de regras formais (explícitas e não-ambíguas) de pós-edição.

Como trabalho futuro, salienta-se que, além da avaliação das regras, a distribuição estatística das *tags* PoS do referido *tweebank* pode ser comparada com a distribuição das mesmas em um *corpus* de outro gênero, como o jornalístico. Esse tipo de comparação pode gerar subsídios para a construção de ferramentas e sistemas de PLN que sejam multigênero.

Com isso, acredita-se que este trabalho contribui para os estudos linguístico-descritivos e para o PLN.

6. REFERÊNCIAS BIBLIOGRÁFICAS

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.

BEHZAD, S.; ZELDES, A. A cross-genre ensemble approach to robust Reddit part of speech tagging. In: WEB AS CORPUS WORKSHOP, 12, 2020, Marseille. **Proceedings…** Marseille: ACL, 2020. p. 50-56.

BOHNET, B. *et al*. Morphosyntactic tagging with a meta-BiLSTM model over context sensitive token encodings. In: ANNUAL MEETING OF THE ASSOCIATION FOR COMPUTATIONAL LINGUISTICS, 56, 2018, Melbourne. **Proceedings** [...]. Melbourne: ACL, 2018. p. 2642–2652.

BOSCO, C. *et al.* Overview of the EVALITA 2016 Part Of Speech Tagging on TWitter for ITAlian task. *In*: EVALUATION CAMPAIGN OF NATURAL LANGUAGE PROCESSING AND SPEECH TOOLS FOR ITALIAN, 5, 2016, Naples. **Proceedings** […]. Naples: CEUR, 2016, p. 1-7.

DEVLIN, J. *et al*. BERT: Pre-training of deep bidirectional transformers for language understanding. *In*: CONFERENCE OF THE NORTH AMERICAN CHAPTER OF THE ASSOCIATION FOR COMPUTATIONAL LINGUISTICS: HUMAN LANGUAGE TECHNOLOGIES, 2016, Minneapolis. **Proceedings** […]. Minneapolis, 2019, p. 4171-4186.

DI-FELIPPO, A. *et al*. Descrição preliminar do *corpus* DANTEStocks: diretrizes de segmentação para anotação segundo *Universal Dependencies*. *In*: JORNADA DE DESCRIÇÃO DO PORTUGUÊS, 7, 2021. **Anais** […]. 2021, p. 335-343. (evento *online*)

DI-FELIPPO, A. *et al*. Diretrizes de anotação de tag PoS*s* em *tweets* do mercado financeiro: orientações para anotação em Língua Portuguesa segundo a abordagem *Universal Dependencies.* **Relatório Técnico do ICMC,** 438. ICMC, USP. São Carlos-SP, 24p, 2022.

DURAN, M.S. Manual de anotação de tag PoS*s*: orientações para anotação de etiquetas morfossintáticas em Língua Portuguesa, seguindo as diretrizes da abordagem *Universal Dependencies*. **Relatório Técnico do ICMC**, 434. ICMC, USP, São Carlos, 55p, 2021.

FONSECA, E., ROSA, J., AND ALUÍSIO, S. Evaluating word embeddings and a revised corpus for part-of-speech tagging in portuguese. **Journal of the Brazilian Computer Society**, 21(2), p. 1–14, 2015. Brazilian Computer Society.

GUIBON, G., *et al*. When collaborative treebank curation neets graph grammars: Arborator with a grew back-end. *In*: INTERNATIONAL CONFERENCE ON LANGUAGE RESOURCES AND EVALUATION CONFERENCE, 12, 2012, Istanbul. **Proceedings** […]. Istanbul: ELRA, 2012, p. 5291-5300.

JURAFSKY, D.; MARTIN, J.H. **Speech and Language Processing: an introduction to Natural Language Processing**, Computational Linguistics and Speech Recognition. 3rd ed. Available at: https://web.stanford.edu/~jurafsky/slp3/. Access in: 1 July 2022.

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.

KLEIN, S.; SIMMONS, R. F. A computational approach to grammatical coding of english words. **Association for Computing Machinery**, p. 334–347, 1963.

KONDRATYUK, D.; STRAKA, M. 75 languages, 1 model: parsing Universal Dependencies universally. *In*: CONFERENCE ON EMPIRICAL METHODS IN NATURAL LANGUAGE PROCESSING, 2019, Hong Kong. **Proceedings** […]. Hong Kong: ACL, 2019. p. 2779–2795.

LYNN, T. *et al.* Minority language Twitter: part-of-speech tagging and analysis of Irish tweets. *In*: WORKSHOP ON NOISY USER-GENERATED TEXT, 2015, Beijing. **Proceedings** […]. Beijing: ACL, 2015, p. 1-8.

MIRANDA, L.G.M., PARDO, T.A.S. An improved and extended annotation tool for Universal Dependencies-based treebank construction. *In*: INTERNATIONAL CONFERENCE ON THE COMPUTATIONAL PROCESSING OF PORTUGUESE - DEMO WORKSHOP, 2022. **Proceedings** […]. 2022, p. 1-3. (online event)

MITKOV, R. (Ed.) **The Oxford Handbook of Computational Linguistics** (Oxford Handbooks in Linguistics S.), 1st ed. USA: Oxford University Press, Inc., 2005.

NIVRE, J. Towards a Universal Grammar for Natural Language Processing. *In*: Gelbukh, A. (Eds) Computational Linguistics and Intelligent Text Processing. CICLing 2015. **Lecture Notes in Computer Science**, vol 9041. Springer, p. 3-16. https://doi.org/10.1007/978-3-319-18111-0_1.

NIVRE, J. *et al*. Universal Dependencies v1: a multilingual treebank collection. *In*: INTERNATIONAL CONFERENCE ON LANGUAGE RESOURCES AND EVALUATION,10, 2016, Portorož. **Proceedings** […]. Portorož: ELRA, 2016. p.1659-66.

NIVRE, J., MARIE-CATHERINE. M., GINTER, F. *et al.* Universal Dependencies v2: An Evergrowing Multilingual Treebank Collection. *In*: INTERNATIONAL CONFERENCE ON LANGUAGE RESOURCES AND EVALUATION, 12, 2020, Marseille. **Proceedings** […]. Marseille: ELRA, 2020, p. 4034-4043.

PARDO, T.A.S. *et al*. Porttinari - a large multi-genre treebank for brazilian portuguese. *In*: SYMPOSIUM IN INFORMATION AND HUMAN LANGUAGE, 14, 2021. **Proceedings** […]. 2021. p.1-10. (online event)

PROISL, T. Someweta: a part-of-speech tagger for German social media and web texts. *In*: INTERNATIONAL CONFERENCE ON LANGUAGE RESOURCES AND EVALUATION, 11, 2018, Miyazaki. **Proceedings** […]. Miyazaki: ELRA, 2018, p. 665-670.

PLUTCHIK R., KELLERMAN, H. (Ed). 1986. Emotion: theory, research and experience. Nova Iorque: Acad. Press

REHBEIN, I., RUPPENHOFER, J., BICH-NGOC, D. tweeDe – a Universal Dependencies treebank for German tweets. *In*: INTERNATIONAL WORKSHOP ON TREEBANKS AND LINGUISTIC THEORIES, 18, 2019, Paris. **Proceedings** […]. Paris: ACL, 2019, p. 100-108.

SANGUINETTI, M. *et al.* PoSTWITA-UD: an Italian Twitter Treebank in Universal Dependencies. *In*: INTERNATIONAL CONFERENCE ON LANGUAGE RESOURCES AND EVALUATION, 11, 2018. Miyazaki. **Proceedings** […]. Miyazaki: ELRA, 2018, p. 1768-1775.

SILVA, F. J. V.; ROMAN, N. T.; CARVALHO, A.M.B.R. Stock market tweets annotated with emotions. **Corpora**, 15(3), p. 343-54, 2020. Online ISSN 1755-1676.

SILVA, E.H. *et al*. Universal Dependencies for tweets in Brazilian Portuguese: tokenization and part of speech tagging. *In*: NATIONAL MEETING ON ARTIFICIAL AND COMPUTATIONAL INTELLIGENCE,18, 2021. **Proceedings** [...]. 2021. p.1-12. (online event)

SILVA, E.H. **Etiquetagem morfossintática multigênero para o português do Brasil segundo o modelo *Universal Dependencies***. Qualificação (Mestrado em Ciências de Computação e Matemática Computacional) - Universidade de São Paulo, São Carlos, SP, Brasil, 2022.

SOUSA, R. C.C. de; LOPES, H. Portuguese pos tagging using blstm without handcrafted features. *In*: NYSTRÖM, I.; HEREDIA, Y. H.; NÚÑEZ, V. M. (Eds.). Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications. CIARP 2019. **Lecture Notes in Computer Science**, vol. 11896. Springer. https://doi.org/10.1007/978-3-030-33904-3_11

STRAKA, M.; HAJIČ, J.; STRAKOVÁ, J. UDPipe: Trainable pipeline for processing CoNLL-U files performing tokenization, morphological analysis, POS tagging and parsing. *In*: INTERNATIONAL CONFERENCE ON LANGUAGE RESOURCES AND EVALUATION, 10, 2016, Portorož, **Proceedings** [...]. Portorož: ELRA, 2016. p. 4290-4297.

STRAKA, M.; STRAKOVÁ, J.; HAJIC, J. UDPipe at SIGMORPHON 2019: contextualized

embeddings, regularization with morphological categories, corpora merging. *In*: WORKSHOP ON COMPUTATIONAL RESEARCH IN PHONETICS, PHONOLOGY, AND MORPHOLOGY, 16, 2019, Florence. **Proceedings** [...]. Florence, 2019. p. 95–103.

Name	References	Source	Language	UD-based
ATDT	Albogamy and Ramsay (2017)	Twitter	AR	Yes
Hi-En-CS	Bhat et al. (2018)	Twitter	HI/EN	Yes
TwitterAAE (TAAE)	Blodgett et al. (2018)	Twitter	AAE, MAE	Yes
TWITTIRÒ-UD (TWRO)	Cignarella et al. (2019)	Twitter	IT	Yes
DWT	Daiber and Van Der Goot (2016)	Twitter	EN	No★
W2.0	Foster et al. (2011)	Twitter, sort fora	EN	No‡
Foreebank (Frb)	Kaljahi et al. (2015)	Technical fora	EN, FR	No‡
Tweebank (Twb)	Kong et al. (2014)	Twitter	EN	No★
Tweebank2 (Twb2)	Liu et al. (2018)	Twitter	EN	Yes
TDT	Luotolahti et al. (2015)	Various	FI	Yes
xUGC	Martínez Alonso et al. (2016)	Various	FR	Yes
Estonian Web Treebank (EtWT)	Martínez Alonso et al. (2016)	Various	ET	Yes
ITU	Pamay et al. (2015)	n.a.	TR	No★
WDC	Read et al. (2012b)	Various	EN	No‡
tweeDe	Rehbein et al. (2019)	Twitter	DE	Yes
PoSTWITA-UD (Pst)	Sanguinetti et al. (2018)	Twitter	IT	Yes
FSMB	Seddah et al. (2012)	Twitter, Facebook, discussions fora	FR	No‡
Narabizi (NBZ)	Seddah et al. (2020)	Newspaper fora	DZ/FR	Yes
EWT	Silveira et al. (2014)	Various	EN	Yes
LAS-DisFo (LDF)	Taulé et al. (2015)	Discussion fora	ES	No‡
MoNoise (MNo)	Van Der Goot and van Noord (2018)	Twitter	EN	Yes
STB	Wang et al. (2017)	Discussion fora	SgE	Yes
CWT	Wang et al. (2014)	Twitter, Sina Weibo	ZH	No★
GUM	Zeldes (2017)	Various	EN	Yes
HSE	n.a.	Various	BE	Yes
OOD	n.a.	Various	FI	Yes
TwittIrish (TwIr)	n.a. (Publication forthcoming)	Twitter	GA	Yes
Cadhan (Cdh)	n.a.	Various	GV	Yes
Taiga	n.a.	Various	RU	Yes
IU	n.a.	Various	UK	Yes

APÊNDICE 1 – REGRAS DE PÓS-EDIÇÃO E CARACTERIZAÇÃO DOS ERROS DE *TAGGING*

Qt	DE	PARA	Condição	Regra	Exemplo	Qt	Tipo de regra	Linguagem	Fenômeno CGU
1	ADV	ADJ	Se token="certo"	Substituir ADV por ADJ	certo/ADJ>ADV	1	Lexicalizada	Língua geral	
2	NOUN	ADJ	Se o token for precedido por DET e NOUN	Substituir por NOUN por ADJ	em/ADP a/DET sexta/NOUN **passada**/ADJ>NOUN	2	Não-lexicalizada	Língua geral	
3	NOUN	ADJ	Se o token for imediatamente precedido por DET e sucedido por NOUN.	Substituir NOUN por ADJ	a/DET **corrente**/NOUN>ADJ gravidade/NOUN	1	Não-lexicalizada	Língua geral	
4	NOUN	ADJ	Se o token for imediatamente precedido por ADP e sucedido por NOUN.	Substituir NOUN por ADJ	de/ADP **módico**/NOUN>ADJ repique/NOUN	1	Não-lexicalizada	Língua geral	
5	NOUN	ADJ	Se o token anterior for NOUN, substituir tag do token "log" por ADJ.	Substituir NOUN por ADJ	escala/NOUN **log**/ADJ>NOUN	1	Lexicalizada	Domínio	
6	NOUN	ADJ	Se token="casado"	Substituir NOUN por ADJ	**casado/ADJ>NOUN**	1	Lexicalizada	Língua geral	
7	NOUN	ADJ	Se token="coitada", primeiro token da sentença e seguido por ADP	Substituir NOUN por ADJ	**coitada/ADJ>NOUN**	1	Lexicalizada	Língua geral	
8	NOUN	ADJ	Se o token for seguido por PUNCT (/) + ADJ	Substituir NOUN por ADJ	**médio/ADJ>NOUN />PUNCT longo/NOUN**	1	Não-lexicalizada	Língua geral	
9	NOUN	ADJ	Se token="melho"	Substituir qualquer tag PoS por ADJ	**melho/NOUN>ADJ**	1	Lexicalizada	Língua geral	Truncamento
10	NOUN	ADJ	Se token="voláteis"	Substituir qualquer tag PoS por ADJ	**voláteis/NOUN>ADJ**	1	Lexicalizada	Língua geral	
11	VERB	ADJ	Se token 1E e 2E = ADJ + NOUN	Substituir VERB por ADJ	lucro/NOUN líquido/ADJ **consolidado**/ADJ>VERB	2	Não-lexicalizada	Língua geral	
12	VERB	ADJ	Se token 1E="trade"	Substituir VERB por ADJ	Trade/NOUN **fechado**/ADJ>VERB	1	Lexicalizada	Língua geral	
13	VERB	ADJ	Se token="lindo"	Substituir qualquer tag PoS por ADJ	lindo/ADJ>VERB	1	Lexicalizada	Língua geral	
14	VERB	ADJ	Se token 1E="risco"	Substituir VERB por ADJ	risco/NOUN **elevado**/ADJ>VERB	1	Lexicalizada	Língua geral	
15	VERB	ADJ	Se tokens 1D, 2D e 3D = ADP + DET + NOUN/PROPN	Substituir VERB por ADJ	para/ADP o/DET bebê/NOUN **abandonado**/ADJ>VERB	3	Não-lexicalizada	Língua geral	
16	ADV	ADP	Se token 1D="assim", trocar ADV por ADP	Substituir ADV por ADP	assim/ADV **como**/ADP fizemos/VERB	1	Lexicalizada	Língua geral	
17	DET	ADP	Se token 1E=VERB, substituir DET por ADP	Substituir DET por ADP	que/PRON correspondem/VERB **a**/ADP	1	Lexicalizada	Língua geral	Truncamento
18	NOUN	ADP	Se token 1D="market", substituir qualquer tag que não seja ADP por ADP	Substituir qualquer tag PoS por ADP	o/DET **after**/ADP market/NOUN	1	Lexicalizada	Domínio	Estrang.
19	NOUN	ADP	Se token = "d", substituir qualquer tag que não seja ADP por ADP	Substituir qualquer tag PoS por ADP	42/NUM **d/**ADP	1	Lexicalizada	Língua geral	Truncamento
20	SCONJ	ADP	Se token 1D=SCONJ, substituir SCONJ por ADP	Substituir SCONJ por ADP	perspectiva/NOUN **de**/ADP que/SCONJ	2	Não-lexicalizada	Língua geral	
21	SCONJ	ADP	Se 1E=NOUN e 1D=VERB, substituir SCONJ por ADP	Substituir SCONJ por ADP	Ações/NOUN para/ADP comprar/VERB	2	Não-lexicalizada	Língua geral	
22	SCONJ	ADP	Se 1E=NOUN e 2D for VERB, substituir SCONJ por ADP	Substituir SCONJ por ADP	hora/NOUN **de**/ADP ela/PRON desabar/VERB	3	Não-lexicalizada	Língua geral	
23	SCONJ	ADP	Se NOUN + ADJ em 2E e 1E, e VERB em 1D, substituir SCONJ por ADP	Substituir SCONJ por ADP	fundo/NOUN soberano/ADJ para/ADP comprar/VERB	1	Não-lexicalizada	Língua geral	

24	ADP	ADV	Se token 1D='vc'	Substituir ADP por ADV	**como/ADV** vc/PRON	1	Lexicalizada	Domínio	
25	ADP	ADV	Se token 1D='mesmo'	Substituir ADP por ADV	**até/ADV mesmo/ADV**	1	Lexicalizada	Língua geral	
26	NOUN	ADV	Se token= "tb"	Substituir NOUN por ADV	**tb/ADV**	1	Lexicalizada	Domínio	Abreviação
27	NOUN	ADV	Se token='menos'	Substituir NOUN por ADV	**menos/ADV**	1	Lexicalizada	Língua geral	
28	NOUN	ADV	Se token='avante'	Substituir NOUN por ADV	**avante/ADV**	1	Lexicalizada	Língua geral	
29	NOUN	ADV	Se token='cm'	Substituir NOUN por ADV	**cm/ADV**	1	Lexicalizada	Domínio	Abreviação
30	NOUN	ADV	Se token='msm'	Substituir NOUN por ADV	**msm/ADV**	1	Lexicalizada	Domínio	Abreviação
31	NOUN	ADV	Se token 1E=ADP	Substituir NOUN por ADV	para/ADP **baixo/ADV**	2	Lexicalizada	Língua geral	
32	PRON	ADV	Se token 1D='nojo'	Substituir PRON por ADV	**QUE/ADV nojo/NOUN**	1	Lexicalizada	Língua geral	
33	PRON	ADV	Se token='pouco' antecedido de VERB	Substituir PRON por ADV	há/VERB **pouco/ADV**	1	Lexicalizada	Língua geral	
34	SCONJ	ADV	Se primeiro token='como' for primeiro token da sentença seguido por VERB	Substituir SCONJ por ADV	**Como/ADV diria/VERB**	1	Lexicalizada	Língua geral	
35	SCONJ	ADV	Se 1D for token='que'	Substituir SCONJ por ADV	**tanto/ADV que/SCONJ**	2	Lexicalizada	Língua geral	
36	VERB	ADV	Se token='fora'	Substituir VERB por ADV	**fora/ADV**	2	Lexicalizada	Língua geral	
37	ADJ	DET	se existir DET em 1E	Substituir ADJ por DET	o/DET **mesmo/DET**	2	Não-lexicalizada	Língua geral	
38	ADP	DET	Se 1E e 1D forem NOUN	Substituir ADP por DET	fechamento/NOUN **de/DET** gap/NOUN	2	Não-lexicalizada	Língua geral	
39	ADP	DET	Se 1E=VERB e 1D=NOUN	Substituir ADP por DET	comece/VERB **a/DET** volatil/NOUN	2	Não-lexicalizada	Língua geral	
40	PRON	DET	Se PUNCT em 1E ou 2E	Substituir PRON por DET	./PUNCT **Esse/DET**	2	Não-lexicalizada	Língua geral	
41	PRON	DET	Se DET em 1E ou 2E	Substituir PRON por DET	o/DET **tal/DET**	2	Não-lexicalizada	Língua geral	
42	SCONJ	DET	Se token='que' seguido de token= 'ganhos'	Substituir SCONJ por DET	**que/DET ganhos/NOUN**	1	Lexicalizada	Língua geral	
43	SCONJ	DET	Se token= 'QUE' seguido de token='TAL'	Substituir SCONJ por DET	**QUE/DET TAL/PRON**	1	Lexicalizada	Língua geral	
44	ADJ	INTJ	Se existir PUNCT em 1D	Substituir ADJ por INTJ	**Pronto/INTJ** !/PUNCT	2	Não-lexicalizada	Língua geral	
45	ADV	INTJ	Se token='né'	Substituir ADV por INTJ	**né/INTJ**	1	Lexicalizada	Domínio	Coloquialismo
46	ADV	INTJ	Se token='amém'	Substituir ADV por INTJ	**amém/INTJ**	1	Lexicalizada	Língua geral	
47	ADV	INTJ	Se token='Ah'	Substituir ADV por INTJ	**Ah/INTJ**	1	Lexicalizada	Língua geral	
48	AUX	INTJ	Se existir PUNCT em 1D	Substituir ADJ por INTJ	**É/INTJ ,/PUNCT**	1	Não-lexicalizada	Língua geral	
49	NOUN	INTJ	Se token='blz'	Substituir NOUN por INTJ	**blz/INTJ**	1	Lexicalizada	Domínio	Abreviação
50	NOUN	INTJ	Se token='Ops'	Substituir NOUN por INTJ	**Ops/INTJ**	1	Lexicalizada	Domínio	Coloquialismo
51	PRON	INTJ	Se token='ô'	Substituir AUX por INTJ	**ô/INTJ**	1	Lexicalizada	Língua geral	Coloquialismo

52	ADJ	NOUN	Se token="gasolina".	Então substituir qualquer PoS que não seja NOUN por NOUN.	bolsa/NOUN **gasolina**/ADJ>NOUN	1	Lexicalizada	Língua geral	
53	ADJ	NOUN	Se token="B" e pos do 1º token E for "plano".	Então substituir pos diferente de NOUN por NOUN.	plano/NOUN **B**/ADJ>NOUN	1	Lexicalizada	Língua geral	
54	ADJ	NOUN	Se pos=ADJ.	Então substituir ADJ por NOUN se pos do 1º token E for DET e do 1º token D não for NOUN	a/DET **segunda**/ADJ>NOUN se/PRON	1	Não-lexicalizada	Língua geral	
55	ADJ	NOUN	Se pos=ADJ e pos do 1º token E for VERB e do 1º token à direira for ADP.	Então substituir ADJ por NOUN.	renova/VERB **mínima**/ADJ>NOUN de/ADP	1	Não-lexicalizada	Língua geral	
56	ADJ	NOUN	Se um dos ordinais 2ª, 3ª, 4ª, 5ª ou 6ª=ADJ e pos do 2º token E for ADP e o 1º token E for "esta/DET".	Então substituir ADJ por NOUN.	em/ADP esta/DET **5ª**/ADJ>NOUN	1	Lexicalizada	Língua geral	
57	ADJ	NOUN	Se os tokens "acionista"/"acionistas"=ADJ.	Então substituir ADJ para NOUN.	ser/AUX **acionista**/ADJ>NOUN	1	Lexicalizada	Língua geral	
58	ADJ	NOUN	Se token "estatal" ou "estatais"=ADJ.	Então substituir ADJ por NOUN se pos do 1º token E não for NOUN.	foram/AUX só/ADV **estatais**/ADJ>NOUN	1	Lexicalizada	Língua geral	
59	ADJ	NOUN	Se pos=ADJ.	Então substituir ADJ por NOUN se pos do 1º token E for DET e do 1º token D não for NOUN	minha/DET **querida**/ADJ>NOUN !/PUNCT	1	Não-lexicalizada	Língua geral	
60	ADJ	NOUN	Se sequência de pos=ADJ+NOUN e pos do 1º token E de ADJ for DET e do 1º token D de NOUN for DET ou CCONJ e do 2º D de NOUN for ADJ.	Então inverter as tags (NOUN+ADJ).	a/DET **sexta**/ADJ>NOUN **passada**/NOUN>ADJ uma/DET sub/ADJ	1	Não-lexicalizada	Língua geral	
61	ADJ	NOUN	Se sequência de pos=ADJ+NOUN e pos do 1º token E de ADJ for DET e do 1º token D de NOUN for DET ou CCONJ e do 2º D de NOUN for ADJ.	Então inverter as tags (NOUN+ADJ).	a/DET **alta**/ADJ>NOUN **esquisita**/NOUN>ADJ e/CCONJ repentina/ADJ	1	Não-lexicalizada	Língua geral	
62	PRON	NOUN	Se token "el"=PRON e um dos 2 tokens prévios (E) for "perder"/VERB.	Então substituir PRON por NOUN.	perder/VERB a/DET **el**/PRON>NOUN	1	Lexicalizada	Domínio	Truncamento
63	SYM	NOUN	Se token "abs"=SYM.	Então substituir SYM por NOUN	?/PUNCT **Abs**/SYM>NOUN	1	Lexicalizada	Domínio	Abreviação
64	VERB	NOUN	Se token "bela"=PROPN + token "escolha"=VERB.	Então substituir PROPN por ADJ e VERB por NOUN.	**Bela/PROPN>ADJ escolha/VERB>NOUN**	1	Lexicalizada	Língua geral	
65	VERB	NOUN	Se pos=VERB e pos do 1º token E for PROPN e 1º token D for ":/PUNCT".	Então substituir VERB por NOUN.	#petr4/PROPN **Recompra**/VERB>NOUN :/PUNCT	1	Lexicalizada	Língua geral	
66	VERB	NOUN	Se pos=VERB e pos do 1º token E não for PRON.	Então substituir VERB por NOUN.	@JHF_Oficial/PROPN **erro**/VERB>NOUN	1	Não-lexicalizada	Língua geral	
67	VERB	NOUN	Se token "desculpa"=VERB e pos do 1º token E não for PRON.	Então substituir VERB por NOUN.	peço/NOUN sempre/ADV **desculpa**/VERB>NOUN	1	Lexicalizada	Língua geral	
68	VERB	NOUN	Se token 1E="que".	Então substituir VERB por NOUN.	que/DET **ganhos**/VERB>NOUN	1	Lexicalizada	Língua geral	
69	VERB	NOUN	Se pos=VERB e pos do 2º token E for NOUN e do 1º token E for CCONJ e do 1º token D ser ADP.	Então substituir VERB por NOUN.	baixa/NOUN e/CCONJ **agulhada**/VERB>NOUN de/ADP	1	Não-lexicalizada	Língua geral	
70	VERB	NOUN	Se pos=VERB e pos do 1º token D for ADP e do 2º token D for PROPN.	Então substituir VERB por NOUN.	**repique/VERB>NOUN de/ADP oibr4/PROPN**	1	Não-lexicalizada	Língua geral	
71	ADV	PRON	Se token = "algo"	substituir qualquer tag PoS por PRON	**algo/PRON raro/ADJ**	1	Lexicalizada	Língua geral	
72	ADV	PRON	Se token 1E = "QUE"	Substituir ADV por PRON	QUE/DET **TAL/PRON**	1	Lexicalizada	Língua geral	
73	DET	PRON	Se token = "o" e token 1D = "q"	Substituir DET por PRON	**o/PRON q/PRON**	2	Lexicalizada	Língua geral	
74	DET	PRON	Se 1D = NOUN	Substituir DET por PRON	**Esse/PRON tb/ADV**	1	Não-lexicalizada	Língua geral	

75	NOUN	PRON	Se token = "vcs"	substituir qualquer tag PoS que não seja PRON por PRON	**vcs/PRON>NOUN**	2	Lexicalizada	Domínio	Abreviação
76	NOUN	PRON	Se token = "mim	substituir qualquer tag PoS que não seja PRON por PRON	**mim/PRON>NOUN**	1	Lexicalizada	Língua geral	Abreviação
77	ADP	SCONJ	Se ADV/VERB em 1E, 2E ou 3E e VERB/AUX em 1D, 2D ou 3D	Substituir ADP por SCONJ	antes/ADV **de/SCONJ** ser/AUX	1	Não-lexicalizada	Língua geral	
78	ADV	SCONJ	Se VERB em 2E ou 1E	Substituir ADV por SCONJ	ver/VERB **como/SCONJ**	2	Não-lexicalizada	Língua geral	
79	PRON	SCONJ	Se VERB em 1D, 2D ou 3D	Substituir PRON por SCONJ	**q/SCONJ** vcs/PRON conhecem/VERB	3	Não-lexicalizada	Língua geral	
80	NOUN	SYM	Se token= 'x'	Substituir NOUN por SYM	**x/SYM**	1	Lexicalizada	Domínio	Símbolo
81	ADJ	VERB	Se token= 'regist'	Substituir ADJ por VERB	**registr/VERB**	1	Lexicalizada	Domínio	Truncamento
82	ADJ	VERB	Se token= 'fazen'	Substituir ADJ por VERB	**fazen/VERB**	1	Lexicalizada	Domínio	Truncamento
83	ADP	VERB	Se ADV em 1E	Substituir ADP por VERB	não/ADV **para/VERB**	1	Não-lexicalizada	Língua geral	
84	ADP	VERB	Se houver CCONJ em 1D	Substituir ADP por VERB	**alugado/VERB** mas/CCONJ	1	Não-lexicalizada	Língua geral	
85	ADP	VERB	Se token= 'há'	Substituir ADP por VERB	**há/VERB**	1	Lexicalizada	Língua geral	
86	AUX	VERB	Se não houver VERB em 1D, 2D ou 3D	Substituir AUX por VERB	quem/PRON **for/VERB** a/DET faixa/NOUN	4	Não-lexicalizada	Língua geral	
87	NOUN	VERB	Se token= 'monitoro'	Substituir NOUN por VERB	**monitoro/VERB**	1	Lexicalizada	Língua geral	
88	NOUN	VERB	Se token= 'Lembra'	Substituir NOUN por VERB	**Lembra/VERB**	1	Lexicalizada	Língua geral	
89	NOUN	VERB	Se token= 'renova'	Substituir NOUN por VERB	**renova/VERB**	1	Lexicalizada	Língua geral	
90	NOUN	VERB	Se token= 'Trabalho' sem DET em posição 1D	Substituir NOUN por VERB	**@pagina2/PROPN Trabalho/VERB**	1	Lexicalizada	Língua geral	
91	NOUN	VERB	Se token= 'venda' seguido de tokens 'de' e 'novo'	Substituir NOUN por VERB	**venda/VERB de/ADP novo/NOUN**	1	Lexicalizada	Língua geral	
92	NOUN	VERB	Se token ='peço'	Substituir NOUN por VERB	**peço/VERB**	1	Lexicalizada	Língua geral	
93	NOUN	VERB	Se token='partir'	Substituir NOUN por VERB	**partir/VERB**	1	Lexicalizada	Língua geral	