UNIVERSIDADE FEDERAL DE SÃO CARLOS

CENTRO DE EDUCAÇÃO E CIÊNCIAS HUMANAS - CECH

Isabela Santos de Freitas

DESCRIÇÃO DA FRAGMENTAÇÃO ESTRUTURAL DE *TWEETS* DO MERCADO FINANCEIRO VIA RELAÇÃO DE "PARATAXIS" DO MODELO "UNIVERSAL DEPENDENCIES"

SÃO CARLOS - SP

2025

Isabela Santos de Freitas

DESCRIÇÃO DA FRAGMENTAÇÃO ESTRUTURAL DE *TWEETS* DO MERCADO FINANCEIRO VIA RELAÇÃO DE "PARATAXIS" DO MODELO "UNIVERSAL DEPENDENCIES"

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

2025

RESUMO

Uma das características dos *tweets* (atuais *posts* do X) é a fragmentação, pois muitos deles são compostos por sequências de segmentos sem conexão sintática explícita entre eles. Tais segmentos podem ser sentenças, sintagmas curtos e mesmo fragmentos destes que estão apenas justapostos. Quando um *corpus* possui anotação gramatical segundo o modelo *Universal Dependencies* (UD), a fragmentação estrutural dos enunciados é capturada pela relação de dependência (*deprel*) **parataxis**. Em linhas gerais, essa *deprel* é empregada para anotar que dois segmentos justapostos de um mesmo enunciado não possuem relação sintática clara entre si, podendo ser especificada por subetiquetas. Neste trabalho, descreveu-se a fragmentação nos *tweets* do mercado financeiro do *corpus* DANTEStocks via *deprel*. Para tanto, as ocorrências de **parataxis**, com e sem sub-relações, foram extraídas do *corpus* e organizadas de forma semiatuomática em função da frequência simples. Do total de **6.733** ocorrências, 3.840 casos (57%) não possuem sub-relação e 2.893 casos (**43**%) possuem diferentes sub-relações. Na sequência, todas as ocorrências de **parataxis**, com e sem sub-relações, foram organizadas automaticamente em função da frequência das combinações entre as etiquetas morfossintáticas (ou de *part-of-speech* – PoS) do *head* e dependente e da ordem das etiquetas. Com exceção de **parataxis:strunc** e **parataxis:wtrunc**, os demais tipos com sub-relação, isto é, **parataxis:hashtag, parataxis:cashtag** e **parataxis:url**, possuem dependentes lexicais, pois equivalem a apenas uma palavra/*token.* Os dependentes de **parataxis:hashtag** e **parataxis:cashtag** são X ou PROPN e de **parataxis:url**, SYM. Os casos de **parataxis:strunc** e **parataxis:wtrunc** podem apresentar dependentes lexicais ou estruturais (sintagmas ou sentenças truncadas). A ordem do *head* e dependente em **parataxis:hashtag** e **parataxis:cashtag** é variada, ao passo que em **parataxis:url, parataxis:strunc** e **parataxis:wtrunc** é sempre da esquerda para a direita. Sobre as **parataxis** sem sub-relação, as combinações mais frequentes de PoS são VERB (*root*) + VERB/NOUN/PROPN e NOUN (*root*) + VERB/NOUN/NOUN, indicando que a fragmentação dos *tweets* comumente envolve segmentos verbais e nominais justapostos. Com isso, este trabalho contribui para aumentar o arcabouço de conhecimento sobre as características linguísticas do DANTEStocks, que é o primeiro *corpus* composto por *tweets* com anotação UD em português, já tendo permitido o treinamento e avaliação de algumas ferramentas de PLN para CGU.

Palavras-chave: *tweet, Universal Dependencies*, *parataxis*.

ABSTRACT

One of the characteristics of tweets (now X posts) is fragmentation, as many of them are made up of sequences of segments with no explicit syntactic connection between them. These segments can be sentences, short syntagms or even fragments of these that are just juxtaposed. When a corpus is grammatically annotated according to the Universal Dependencies (UD) model, the structural fragmentation of utterances is captured by the **parataxis** dependency relation (deprel). In general terms, this deprel is used to note that two juxtaposed segments of the same utterance have no clear syntactic relationship between them, and can be specified by sub-tags. In this work, we described the fragmentation of financial market tweets in the DANTEStocks corpus via deprel. To this end, the occurrences of **parataxis**, with and without sub-labels, were extracted from the corpus and organized semi-automatically according to simple frequency. Of the total of **6,733** occurrences, 3,840 cases (57%) have no sub-relation and 2,893 cases (**43%**) have different sub-relations. Next, all occurrences of parataxis, with and without sub-relationships, were automatically organized according to the frequency of combinations between the morphosyntactic (or part-of-speech - PoS) labels of the head and dependent and the order of the labels. With the exception of **parataxis:strunc** and **parataxis:wtrunc**, the other types with sub-relation, i.e. **parataxis:hashtag**, **parataxis:cashtag** and **parataxis:url**, have lexical dependents, as they are equivalent to just one word/token. The dependents of **parataxis:hashtag** and **parataxis:cashtag** are X or PROPN and of **parataxis:url**, SYM. The cases of **parataxis:strunc** and **parataxis:wtrunc** can have lexical or structural dependents (syntagms or truncated sentences). The order of the head and dependent in parataxis:hashtag and parataxis:cashtag varies, while in **parataxis:url, parataxis:strunc** and **parataxis:wtrunc** it is always from left to right. Regarding **parataxis** without sub-relation, the most frequent combinations of PoS are VERB (root) + VERB/NOUN/PROPN and NOUN (root) + VERB/NOUN/NOUN, indicating that the fragmentation of tweets commonly involves juxtaposed verbal and nominal segments. With this, this work contributes to increasing the body of knowledge about the linguistic characteristics of DANTEStocks, which is the first corpus composed of tweets with UD annotation in Portuguese, and has already allowed the training and evaluation of some NLP tools for UCG.

Keywords: tweet, Universal Dependencies, parataxis.

LISTA DE FIGURAS

LISTA DE QUADROS

LISTA DE TABELAS

Aos meus pais, que asfaltaram com as próprias mãos o caminho dos meus sonhos.

AGRADECIMENTOS

À minha família, meus pais Jaqueline e Alexander, que fizeram dos meus estudos e dos de meus irmãos a prioridade que movimentou nossas vidas. Agradeço por uma vida toda de sacrifícios e dedicação que me possibilitou todas as oportunidades que a eles foram negadas. Ao meu irmão Augusto, que me acompanha com a fidelidade que só um irmão pode oferecer ao longo de toda a vida e a quem desejo um sucesso infinitamente maior que o meu. Ao meu irmão Eduardo, luz de nossas vidas, por quem lutarei incansavelmente por um mundo que saiba receber de braços abertos toda a sua singularidade.

Agradeço às minhas madrinhas, Flaviane e Lilian, minhas “parentas” do coração, que sempre me amaram, me apoiaram, incentivaram, torceram pelas minhas conquistas, desde quando aprendia uma coreografia quando criança, até a conclusão de uma graduação, e que assim seja para sempre.

Agradeço aos meus amigos Isabella Scuracchio, Camila Amorim, Julia Clápis, Giovanna Quaglio, Mariana Andrade, Guilherme Arcêncio, Vinícius e Adriano, com quem tenho o prazer de dividir a vida há tantos anos por meio de uma amizade incondicional e infinita.

Agradeço aos amigos que fiz na graduação, Giovanna C. Silva, Damonile Arrabaça, Sara Nery, Juliana Macedo, Mavi, Gabriela Assagra, Malik Nasser, Ana Carolina, Bárbara Padilha e Arthur Demasi, que se tornaram um berço familiar ao longo de todos esses anos. Agradeço pelo compartilhamento das melhores experiências da minha vida, acolhimento e fortalecimento nos momentos mais difíceis. Com quem divido a experiência única de ter sido parte do que nos tornamos.

A Ana Carolina Ruiz, com quem dei as risadas mais sinceras da minha vida, por juntas transformarmos nossa casa em um lar, por todas as vivências únicas, todo o cuidado, o amor, a amizade, as noites de curtição, os dias de ressaca,os dias de trabalho, os de estudo, os de faxina, os de desabafo, os de alegria e os de tristeza. Obrigada por ter me escolhido para fazer parte da aventura da sua vida.

Agradeço em especial a Cris Miura, que foi a primeira pessoa a me fazer perceber meu amor pela área da Linguística numa quinta-feira em uma aula de redação, a primeira pessoa que me deu a chance de entrar no mercado de trabalho, mesmo que eu ainda tivesse muito mais a aprender do que a oferecer, e,

principalmente, por ter sido a primeira pessoa que me fez perceber que não preciso mudar ou diminuir quem sou para caber nos espaços onde almejo estar. Meus sinceros agradecimentos.

Agradeço à minha orientadora Ariani Di Felippo, pela paciência, sabedoria e apoio na elaboração dessa pesquisa, e, além disso, agradeço pela disposição e competência em suas aulas durante a graduação, que me possibilitaram o contato com a área do PLN, que tanto me fascina.

Agradeço também ao mestrando do PPGL, Bryan K. Barbosa, que de forma colaborativa construiu o *script* em *Python* utilizado neste trabalho para a seleção dos dados de análise do *corpus.*

Agradeço a todos os profissionais educadores que passaram pela minha vida, que atravessaram os desafios da educação e que me ajudaram a construir cada tijolo da minha trajetória. Lembro do nome de cada um e levo comigo no coração o brilho que te move nessa profissão.

Ainda nisso, não poderia deixar de agradecer mais uma vez à minha mãe, Jaqueline Freitas, a primeira professora da minha vida, que doou tudo de si para a minha criação e de meus irmãos, que repetiu todas as vezes necessárias o som de cada sílaba, a fim de me ensinar o primeiro passo para saber voar: ler!

SUMÁRIO

1. Introdução

No cenário atual do desenvolvimento da área do Processamento de Língua Natural (PLN), *corpus* é um conjunto de textos computacionalmente processável, tendo sido coletado com um propósito e produzido naturalmente. Quando aos dados de *corpus* é adicionada alguma informação linguística explícita, diz-se que foi feita uma "anotação" (Sinclair, 2005, Freitas, 2022).

A utilidade dos *corpora* anotados reside no treinamento e avaliação de modelos para o processamento automático das línguas naturais. Para isso, as anotações são feitas por pessoas ou por máquinas com posterior revisão manual, resultando nos corpora "padrão ouro" (ou *gold standard*). Mesmo com os grandes modelos de linguagem (Large Language Models – LLM) gerativos da atualidade (como os da família GPT), que processam língua sem a necessidade de treinamento em dados rotulados, os corpora anotados são relevantes. Uma das aplicações é exatamente servir de referência para avaliar os resultados dos LLMs.

Na última década, a relevância do "conteúdo gerado por usuários" (CGU) (*user-generated content*, em inglês) de redes sociais tem gerado a demanda por sistemas de PLN específicos, como os de mineração de opinião e análise de sentimentos.

Diante disso, muitos *corpora* anotados de CGU para várias línguas têm sido construídos com o objetivo de desenvolver as ferramentas que comumente compõem os sistemas de PLN e que são responsáveis, por exemplo, pelas tarefas de etiquetação morfossintática (ou *tagging*) e análise sintática (ou *parsing*).

Embora o desenvolvimento dessas ferramentas tenha uma longa tradição no PLN no que tange a textos escritos segundo uma modalidade formal (Jurafsky; Martins, 2024), a linguagem informal dos CGUs, marcada por fenômenos diversos lexicais e estruturais, demanda que essas ferramentas sejam treinadas para "aprender" a lidar com as características dessa linguagem.

A maioria dos *corpora* de CGU construídos desde 2011 é parcial ou inteiramente composta por *tweets* (atuais *posts* da plataforma *X*). Isso se deve particularmente porque os *tweets* abrangem uma diversidade de fenômenos linguísticos que é comum aos diferentes gêneros de CGU (Sanguinetti *et al.,* 2023).

A maioria desses *corpora* de *tweet* possui anotação gramatical segundo o modelo *Universal Dependencies* (UD) (Nivre *et al.,* 2016, Nivre *et al*., 2020). Esse modelo fornece um conjunto de etiquetas morfossintáticas universais e de relações de dependências sintáticas que possibilita estudos "*cross*-linguísticos", com flexibilidade de adaptação também a diferentes gêneros textuais.

Para o português, tem-se o DANTEStocks (Di Felippo *et al.,* 2024a), *corpus* de ~4 mil *tweets* sobre o mercado financeiro anotado segundo o modelo UD. A anotação-UD desse *corpus* foi semiautomática, sendo que, para a revisão manual da anotação automática inicial, os pesquisadores utilizam manuais que contém a adaptação das diretrizes de utilização do modelo UD para a língua portuguesa geral (Duran, 2021, 2022) e para os *tweets* (Di Felippo *et al.,* 2022, 2024c).

Devido à revisão manual, a anotação-UD é tida como "padrão-ouro" e já permitiu o desenvolvimento de um *tagger* para *tweets* (Silva *et al*., 2021) e dois *parsers,* um deles específico para *tweets* (Barbosa, 2024) e outro multigênero (sendo os gêneros *tweet,* jornalístico e científico de divulgação) (Di-Felippo *et al*., 2024b).

Tendo em vista a adaptação das diretrizes gerais de anotação-UD às peculiaridades dos *tweets,* alguns trabalhos têm focado na descrição das características linguísticas gerais dos tweets (p.ex.: Foster, 2010, Seddah *et al*., 2012, Eisenstein, 2013, Sanguinetti *et al.,* 2023).

Sobre o DANTEStocks, Scandarolli *et al* (2023) focaram na sistematização dos fenômenos lexicais-ortográficos, os quais incluem truncamento (lexical), erros de digitação, abreviações informais, etc. Quanto às características estruturais, Di Felippo *et al.* (2021) apontaram que os *tweets* desse *corpus* podem apresentar multiplicidade sentencial, truncamento (de estrutura) e fragmentação.

Neste trabalho, investigamos a fragmentação dos *tweets,* que é capturada pela relação de dependência do modelo UD denominada "parataxis". Especificamente, essa relação é aquela se estabelece entre dois elementos do *tweets* que poderiam ter relação sintática entre si, porém essa relação não está explicitada. Essa relação, aliás, é a quinta mais frequente no *corpus,* com 6.760 casos.

Para apresentar a pesquisa, este relatório está organizado em 5 Seções. Na Seção 2, apresenta-se uma breve revisão da literatura sobre CGU (em particular, o gênero *tweet*), o modelo UD e o *corpus* DANTEStocks. Na seção 3, apresentam-se a seleção dos casos de **parataxis** do *corpus* e a organização dos dados para análise. Na seção 4, apresenta-se a descrição dos dados de **parataxis** no DANTEStocks. Na Seção 5, por fim, são apresentadas as considerações finais deste trabalho.

.

2. Revisão da literatura

2.1. CGU e sua descrição linguística

Segundo Krumm, Davies e Narayanaswami (2008), o termo CGU engloba todo tipo de conteúdo, seja na forma de imagem, vídeo, áudio ou texto, que é postado por usuários de plataformas *online* que agrega conteúdo, como as redes sociais (p.ex.: *Twitter*[1] (atual *X*), *Facebook*, *Whatsapp* e outros), fóruns de discussão, *wikis*, etc. Os autores destacam ainda que os CGUs são marcados pela acessibilidade e natureza colaborativa, formando um contraponto ao conteúdo produzido por meios de comunicação tradicionais. Quanto à rede social *Twitter,* seu CGU consiste em mensagens ou postagens na modalidade escrita.

Segundo Feldman (2013), o *Twitter* é uma rede social amplamente explorada em aplicações de PLN voltadas para a análise de sentimentos e mineração de opinião devido à alta frequência de opiniões expressas em suas postagens. Essa característica, aliás, faz do *Twitter* uma fonte rica para pesquisas em Psicologia e Sociologia, não somente para o PLN (Schwartz *et al.*, 2013). Para Freitas e Barth (2015), o *Twitter* parece uma mescla de rede social e *microblog*, cujas características gerais são a dinamicidade das interações (sejam comentários ou republicações) e a brevidade das postagens. Quanto à brevidade, aliás, vale destacar que o limite inicial de caracteres de uma postagem era 140, o qual passou para 280 em 2017.

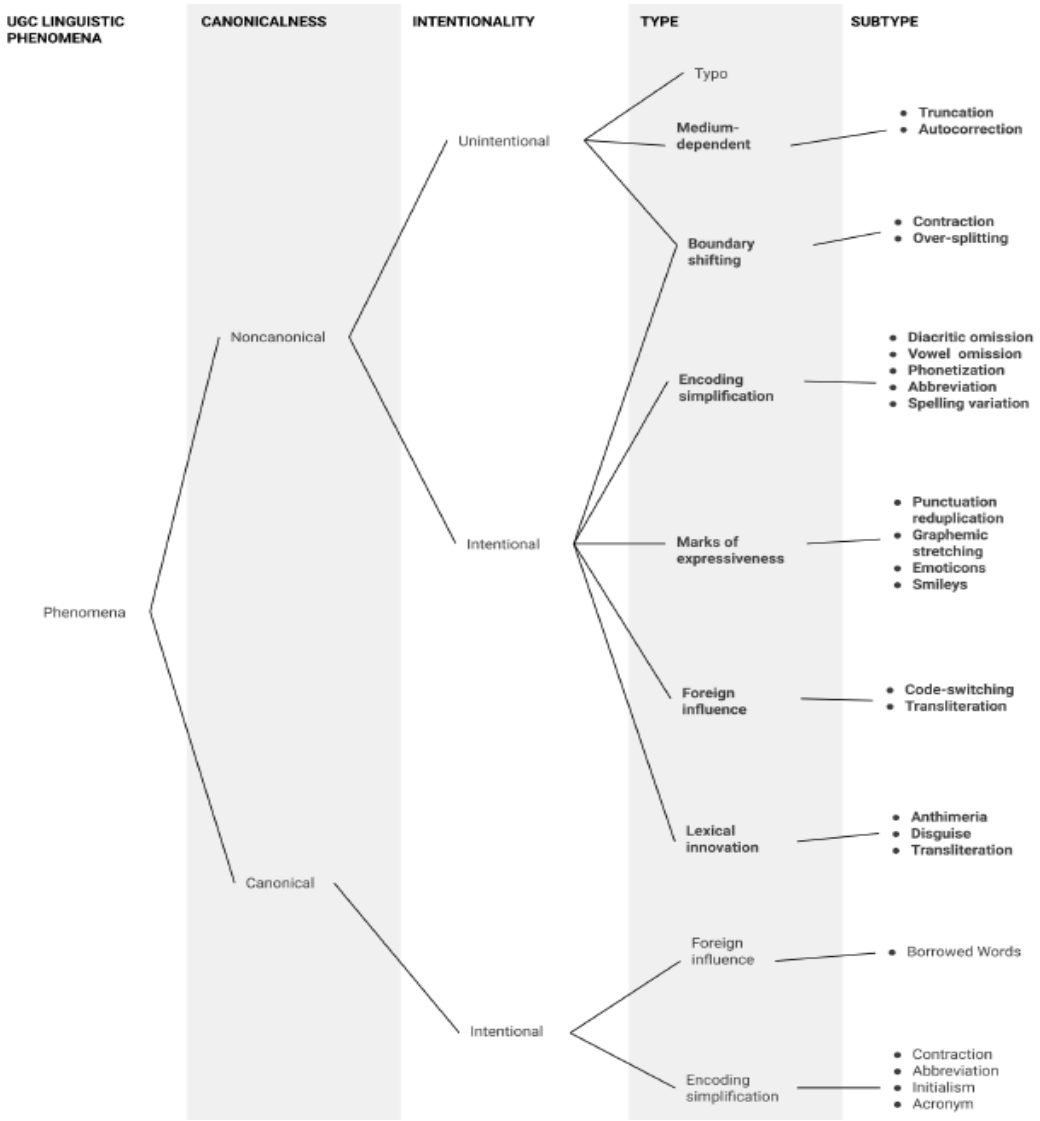
Considerado um gênero, o *tweet* parece ser constituído por resquícios de outros gêneros (como notícia, propaganda, bilhete, diário íntimo, etc.), que foram modificados para atender às necessidades de comunicação da rede. Aliás, esses diferentes gêneros que se entrelaçam nos *tweets* evidenciam a influência da oralidade na plataforma.

Sobretudo visando a proposição de diretrizes de anotação no esquema UD, é essencial compreender as características lexicais e ortográficas específicas dos gêneros CGU (incluindo *tweet*). Assim, alguns trabalhos têm desempenhado um papel crucial ao descrever essas particularidades (p.ex.: Foster, 2010, Seddah *et al*., 2012, Eisenstein, 2013, Sanguinetti *et al.,* 2023).

[1] Embora a plataforma tenha mudado de nome em 2022, assim como suas postagens, optou-se por manter as denominações originais em conformidade com o *corpus* ora descrito, compilado em 2014.

A Figura 1 apresenta a hierarquia dos fenômenos gerais encontrados em diferentes gêneros UGC por Sanguinetti *et al.* (2023). O termo “canonicidade” se refere ao fato de um fenômeno CGU também ser observável em textos de linguagem padrão. Por “intencionalidade”, os autores indicam se um fenômeno linguístico é (ou não) uma produção deliberada (intencional) do usuário[2]. Já o termo “tipo” se refere à variedade do fenômeno, enquanto “subtipo” fornece uma subcategorização de cada “tipo”. Vale ressaltar que uma mesma expressão, e muitas vezes uma única palavra, pode representar múltiplas categorias da tipologia.

Figura 1: Tipologia de fenômenos lexical e ortográficos em corpora de CGU.



Fonte: Sanguinetti *et al.* (2023).

[2] Tendo em vista a incerteza inerente à interpretação, por exemplo, de um *tweet,* a qual se relaciona, aliás, à natureza altamente contextual das postagens, a intencionalidade dos fenômenos só pode ser inferida pelos autores, pois é irreconhecível apenas observando o texto superficial.

A seguir, descrevem-se brevemente os fenômenos não-canônicos da Figura 1.

- *Medium-dependent phenomena* (fenômenos dependentes do meio): engloba os (i) truncamentos lexicais (quebras de palavra), que ocorrem devido ao limite de caracteres dos *tweets*, e (ii) autocorreções, que ocorrem quando uma palavra de uma língua *x*, semelhante a outra de uma língua *y* (comumente o Inglês), é automaticamente "corrigida" para a língua *y* (p.ex: "*coicíse*" ("quinzena") do Irlandês □ "*concise*" ("conciso") do inglês).
- *Boundary shifting* (mudança de limites): tipo de fenômeno que afeta o número de *tokens* (palavras) em comparação com a ortografia padrão; diz respeito à contração (isto é, substituição de vários *tokens* da linguagem padrão por apenas um) (p.ex.: *to go* □ *gonna*) à "superdivisão" (*oversplitting*) (p.ex.: *c'était* □ "*c t"* (era/foi).
- *Encoding simplification* (simplificação de código): engloba questões ergográficas, que reduzem o esforço de escrita, como omissão de diacríticos/vogais (p.ex: *people* □ *ppl* ("pessoas"), fonetização (p.ex.: *to* □ 2 "para"), variação ortográfica (p.ex.: *je sais* □ *je sé* ("eu sei")) e abreviação (p.ex.: *government* □ *govt* ("governo)).
- *Marks of expressiveness* (marcas de expressividade): engloba os subtipos alongamento gráfico (p.ex.: *superrrrrr*), repetição de sinais de pontuação (p.ex.: *Joli !!!!!!* ("lindo")), bem como *emoticons* (p.ex.: <3) e *emojis* (p.ex.: ☹).
- *Foreign language influence* (influência de língua estrangeira): é frequentemente produzido em contextos altamente multilíngues; engloba os subtipos "alternância de código" (isto é, quando uma palavra é substituída por outra estrangeira) (p.ex.: "*non fare la bad girl*" ("não seja uma menina má")) e transliteração, que são palavras novas criadas com base na pronúncia ou forma de outra estrangeira (p.ex.: "*fair play*" (inglês) □ "féar plé" (francês)).
- *Lexical innovation* (inovação lexical): engloba os subtipos disfarce (p.ex.: *shitt* □ *s**t* ("merda") , antimeria (isto é, uso de uma palavra de uma classe gramatical como se fosse de outra, como *tweet* □ *tuitar*), e também os casos de transliteração.

Além dessa tipologia, os autores também listam *hashtags, at-mentions,* URL e marcas de retweet (RT) como dispositivos lexicais típicos da rede social em questão.

Com base na tipologia, Sanguinetti *et al.* propuseram diretrizes para anotar os fenômenos segundo o modelo UD. Na próxima seção, apresenta-se esse modelo gramatical, destacando sua adaptação do português e ao gênero *tweet.*

2.2. O modelo gramatical *Universal Dependencies*

2.2.1. Características gerais

O projeto UD é uma iniciativa colaborativa internacional que busca criar padrões consistentes de anotação para representar a estrutura gramatical de diferentes línguas. Ele prevê diretrizes para a explicitação em *corpus* de informações morfológicas e sintáticas (Nivre *et al.* 2016, Nivre *et al.,* 2020).

No PLN, esse projeto tem desempenhado um papel crucial, pois, ao oferecer um padrão unificado para representar a sintaxe em diferentes línguas, permite a análise comparativa e o compartilhamento de recursos entre elas[3]. Essa padronização é especialmente útil em um cenário global onde as aplicações de PLN precisam lidar com múltiplas línguas e gêneros textuais (incluindo *tweets*). Tal padrão, flexível e adaptável, facilita a criação de recursos linguísticos para línguas com poucos recursos, ampliando a inclusão linguística em aplicações tecnológicas.

O modelo gramatical empregado no projeto é de dependência, também denominado UD. Quanto à anotação, o modelo prevê 2 níveis. No nível morfológico, especificam-se 3 informações: lema, categoria morfossintática e traços lexicais/gramaticais (*features*). No nível sintático, a anotação se dá por relações de dependência (*deprels*), que são binárias e assimétricas. A anotação-UD é codificada no formato CoNLL-U, como o do Quadro 1. Esse arquivo é relativo à sentença em (1), extraída do *corpus* jornalístico em português denominado *Porttinari-base*[4], descrito a seguir.

(1) O jornalista viajou a convite do Festival do Rio.

[3] https://universaldependencies.org/introduction.html
[4] https://www.kaggle.com/datasets/marlesson/news-of-the-site-folhauol

Esse arquivo possui 10 colunas, sendo cada uma delas destinada a uma informação específica, a saber:

1. ID: o identificador da posição do *token* na sentença (índice numérico a partir de 1)
2. FORM: *token* na forma como ocorre na sentença
3. LEMMA: lema ou forma canônica da palavra
4. POS: etiqueta de classe de palavra (ou *part-of-speech (PoS) tag*)
5. XPOS: etiqueta *PoS* específica da língua
6. FEAT: atributos morfológicos do *token*
7. HEAD: ID do *head* da *deprel* cujo *token* (dependente) que está sendo descrito
8. DEPREL: relação de dependência que conecta o *token* ao seu *head*;
9. DEPS: relação de *enhanced dependency* do *token*;
10. MISC: informações adicionais sobre o *token*.

Quadro 1: Exemplo de arquivo no formato CoNLL-U.

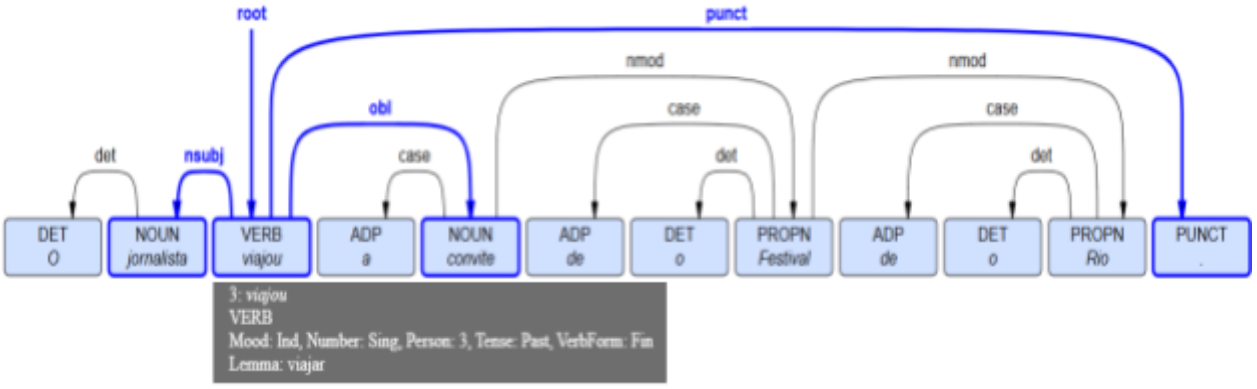
Id	Form	Lemma	Upos Tag	Xpos Tag	Feats	Head	DepRel	Deps	Misc
1	O	o	DET	_	Definite=Def| Gender=Masc| Number=Sing| PronType=Art	2	det	_	_
2	jornalista	jornalista	NOUN	_	Number=Sing	3	nsubj	_	_
3	viajou	viajar	VERB	_	Mood=Ind| Number=Sing| Person=3| Tense=Past| VerbForm=Fin	0	root	_	_
4	a	a	ADP	_	_	5	case	_	_
5	convite	convite	NOUN	_	Gender=Masc| Number=Sing	3	obl	_	_
6-7	do	_	_	_	_	_	_	_	_
6	de	de	ADP	_	_	8	case	_	_
7	o	o	DET	_	Definite=Def| Gender=Masc| Number=Sing| PronType=Art	8	det	_	_
8	Festival	Festival	PROPN	_	_	5	nmod	_	_
9-10	do	_	_	_	_	_	_	_	_
9	de	de	ADP	_	_	11	case	_	_
10	o	o	DET	_	Definite=Def| Gender=Masc| Number=Sing| PronType=Art	11	det	_	_
11	Rio	Rio	PROPN	_	_	8	nmod	_	_
12	.	.	PUNCT	_	_	3	punct	_	_

Fonte: O autor (2025).

A partir de um arquivo CoNLL-U, ferramentas de visualização podem gerar representações gráficas. A Figura 2 ilustra uma árvore sintática gerada por uma dessas ferramentas[5] para a sentença (1).

[5] https://urd2.let.rug.nl/~kleiweg/conllu/

Figura 2: Exemplo de representação arbórea da anotação-UD.



Fonte: O autor (2025).

Nela, vê-se que apenas um *token* é o *root* (raiz) da árvore e que as *deprels* estão indicadas por setas rotuladas que se originam no *head* e se destinam ao dependente. O *token* destacado “viajou” é o *root* e suas informações morfológicas estão no retângulo cinza; ele é o *head* das *deprels* **advcl** (oração adverbial), **ccomp** (complemento oracional fechado) e **punct** (pontuação).

Quanto à morfossintaxe, a UD possui 17 etiquetas de *PoS* (Quadro 2) e 37 relações de dependência (*deprels*) (Quadro 2).

Quadro 2: As 17 *tags PoS* do modelo UD.

ADJ	adjective	ADJETIVO
ADP	adposition	PREPOSIÇÃO
ADV	adverb	ADVÉRBIO
AUX	auxiliary	AUXILIAR
CCONJ	coordinating conjunction	CONJUNÇÃO COORDENATIVA
DET	determiner	DETERMINANTE
INTJ	interjection	INTERJEIÇÃO
NOUN	noun	SUBSTANTIVO
NUM	numeral	NUMERAL
PART	particle	PARTÍCULA
PRON	pronoun	PRONOME
PROPN	proper noun	NOME PRÓPRIO
PUNCT	punctuation	PONTUAÇÃO
SCONJ	subordinating conjunction	CONJUNÇÃO SUBORDINATIVA
SYM	symbol	SÍMBOLO
VERB	verb	VERBO
X	other	OUTRO

Fonte: Adaptado de Nivre *et al.* (2016).

As diretrizes da UD aqui descritas, como mencionado, não são específicas de língua e, por isso, precisam ser instanciadas para a língua cujo *corpus* será anotado. Como

destacado por Duran *et al.* (2022), um fórum mantido pelo projeto UD[6] mostra o quanto essa instanciação é desafiadora, pois centenas de tópicos são discutidos e diversas opiniões concorrem para solucionar as dúvidas.

E o mesmo pode ser dito para os diferentes gêneros. Em outras palavras, as diretrizes do modelo UD precisam ser instanciadas não só para as diferentes línguas, mas também para os diferentes gêneros textuais. Nesse sentido, Sanguinetti *et al* (2023) propuseram estratégias de anotação de PoS e de *deprels* para os fenômenos gerais CGU da tipologia da Figura 1 e para os dispositivos típicos da plataforma.

Quadro 3: As 37 relações de dependência (*deprels*) do modelo UD.

Sigla	Termo em inglês	Tradução
acl	adnominal clause	ORAÇÃO ADNOMINAL
advcl	adverbial clause	ORAÇÃO ADVERBIAL
advmod	adverbial modifier	MODIFICADOR ADVERBIAL
amod	adjectival modifier	MODIFICADOR ADJETIVO
appos	appositional modifier	MODIFICADOR APOSITIVO
aux	auxiliary verb	VERBO AUXILIAR
case	case marking	MARCADOR DE CASO
cc	conjunction	CONJUNÇÃO
ccomp	clausal complement	COMPLEMENTO ORACIONAL
clf	classifier	CLASSIFICADOR
compound	compound	COMPOSTO
conj	conjunct	COORDENADO
cop	copula	VERBO DE CÓPULA
csubj	clausal subject	SUJEITO ORACIONAL
det	determiner	DETERMINANTE
discourse	discourse	DISCURSO
dislocated	dislocated	DESLOCADO
expl	expletive	EXPLETIVO
fixed	fixed expression	EXPRESSÃO FIXA
flat	flat structure	RELAÇÃO PLANA
goeswith	goes with	TOKENS QUE VÃO JUNTOS
iobj	indirect object	OBJETO INDIRETO
list	list	LISTA
mark	marker	MARCADOR DE SUBORDINAÇÃO
nmod	nominal modifier	MODIFICADOR NOMINAL

[6] https://github.com/universaldependencies/docs/issues

nsubj	nominal subject	SUJEITO
nummod	numeric modifier	MODIFICADOR NUMÉRICO
obj	object	OBJETO
obl	oblique nominal	NOMINAL OBLÍQUO
orphan	orphaned dependent	ÓRFÃO
parataxis	parataxis	PARATAXIS
punct	punctuation	PONTUAÇÃO
reparandum	overridden disfluency	DISFLUÊNCIA
root	root	RAIZ
vocative	vocative	VOCATIVO
xcomp	open clausal complement	COMPLEMENTO ORACIONAL ABERTO

Fonte: Adaptado de Nivre *et al.* (2016).

Para ilustrar essas diretrizes, destaca-se um dos fenômenos, o prolongamento grafêmico. Segundo os autores, um caso de prolongamento grafêmico (p.ex.: *superrrrr*) deve ser anotado da seguinte forma: o lema correspondente ao *token* deve ser normalizado (no caso, *super*); a etiqueta de PoS e a *deprel* devem ser definidas pela função do *token* no contexto, e a coluna MISC deve incluir as informações adicionais referentes à `CorrectForm=super` e `NonCan=Strech`.

Para os casos de *at-mentions,* por exemplo, os autores sugerem que a PoS seja sempre PROPN. Se integrada sintaticamente no enunciado, uma *at-mention* deve ser anotada com a *deprel* correspondente à sua função no contexto; caso contrário, será dependente do predicado principal por meio da relação **vocative:mention**.

2.2.2. O modelo UD e o português

Para o português, o site do projeto UD já disponibiliza três *corpora*[7]: o PUD, o GSD e o UD-Portuguese Bosque, este último descrito em Rademaker *et al*. (2017).

Desde 2021, o projeto POeTiSa[8] tem se dedicado à construção de um *corpus* significativo para o processamento automático do português anotado segundo o modelo UD. Trata-se especificamente de um grande *corpus* multigênero para

[7] https://universaldependencies.org/#download
[8] https://sites.google.com/icmc.usp.br/poetisa

fomentar o desenvolvimento de ferramentas e sistemas de análise sintático-semântica. Com isso, o projeto busca contribuir para que o português do Brasil não seja mais classificado como uma “língua pobre de recursos e ferramentas”.

Nomeado Porttinari[9], esse grande *corpus,* em sua versão mais recente, engloba 2 porções de gêneros distintos (jornalístico e CGU) já com anotação-UD.

A porção jornalística contém 167.048 notícias do jornal Folha de São Paulo*,* totalizando 3.964.292 sentenças. Ela é composta por três *subcorpora* com diferentes características e finalidades (Duran *et al.,* 2023): (i) *Porttinari-base*, um *corpus* revisado em detalhe para servir como padrão ouro, (ii) *Porttinari-check*, um pequeno *corpus* estruturalmente similar ao *Porttinari-base* para servir como *testbed* e ilustrar o contraste entre anotação manual e automática, e (iii) *Porttinari-automatic*, um grande *corpus* que foi automaticamente anotado.

A anotação-UD do *Porttinari-base* foi realizada de forma semiautomática. Em particular, o *corpus* foi inicialmente anotado pelo *parser* UDPipe 2 (Straka, 2018), treinado sobre o *corpus* UD-Portuguese Bosque, e os resultados foram posteriormente revisados por humanos. Para a revisão humana, o projeto definiu ou adaptou as diretrizes gerais de anotação de etiquetas PoS e de relações de dependência do modelo UD para a língua portuguesa, publicando os primeiros manuais de anotação para essa língua (Duran, 2021; Duran, 2022).

A anotação padrão-ouro do *Porttinari-base* possibilitou o desenvolvimento do PortParser (Lopes; Pardo, 2024), que alcançou mais de 95% de precisão das *deprels*. Já a combinação do *Porttinari-base* com a porção de CGU do *Porttinari* permitiu também o desenvolvimento do Porttagger (Silva *et al*., 2023), um etiquetador morfossintático de desempenho similar ao PortParser.

2.3. O DANTEStocks

O DANTEStocks corresponde à porção de CGU do *corpus* multigênero *Porttinari.* Trata-se de um *corpus* de *tweets* sobre o mercado financeiro (Di Felippo *et al.,* 2024). O ponto de partida para esse recurso foi a coleção de 4.517 *tweets* de Silva *et al.* (2020), construído para a anotação de emoções.

[9] https://sites.google.com/icmc.usp.br/poetisa/resources-and-tools?authuser=0#h.gxs9bjnlrvma

Compilado automaticamente em 2014, os *tweets* de Silva *et al.* possuem o limite de 140 caracteres. A compilação foi feita pela busca às ocorrência dos *tickers*[10] referentes às 73 ações do índice IBOVESPA na B3[11]. Essa estratégia de coleta resultou em um *corpus* que apresenta uma diversidade de conteúdos ligados ao mercado financeiro, refletindo várias facetas da comunicação nesse mercado (Di-Felippo *et al.*, 2022).

2.3.1. Características linguísticas

Como preparação para a anotação-UD, o *corpus* foi alvo de pesquisas descritivas que buscaram identificar as características da linguagem dos *tweets* do mercado financeiro. Pesquisas essas que fundamentaram a proposição de diretrizes de anotação-UD para as particularidades do DANTEStocks.

Sobre a característica estrutural, Di Felippo *et al.* (2021) apontaram que os *tweets* apresentam uma combinação de linguagem escrita padrão e não padrão. Embora o *corpus* tenha *tweets* compostos por uma ou várias sentenças bem pontuadas (exemplos (1) e (2)), ele também inclui *tweets* com pontuação ausente (3) ou mal-empregada (4). *Tweets* com disfluências (como truncamento estrutural) (5) e com sequências de sintagmas curtos ou fragmentos justapostos (6) também são frequentes no *corpus*. No exemplo (5), em particular, tem-se uma postagem abortada (geralmente indicado pelas reticências), o que resulta em uma estrutura sintática incompleta. No exemplo (6), a postagem é simplesmente composta por uma sequência de fragmentos justapostos.

(1) No momento PETR4 respeita o suporte de R$ 15,42.
(2) Um motivo a menos para a alta da PETR4. Que venha a correção!
(3) O #PT conseguiu fazer propaganda eleitoral antecipada O que a @dilmabr tem a dizer sobre isso?
(4) Bom dia Marcos, Alguma previsão para petr4?!
(5) Petrobrás PN (PETR4), Gráfico Semanal. Estudo das... http://t.co/5bHkUTy8AC
(6) #OIBR4 (mensagem: 956643) http://t.co/VD2ApxqWqR

[10] *Ticker* é um código alfanumérico de 4 letras, que representa a empresa, e 1 número, que indica o tipo da ação; por exemplo, Petr3 é o *ticker* para as ações ordinárias da Petrobras.
[11] B3, em referência às letras iniciais de Brasil, Bolsa, Balcão.

Scandarolli *et al.* (2023) descreveram as idiossincrasias lexicais e ortográficas do DANTEStocks. Eles propuseram uma taxonomia para os fenômenos em duas grandes categorias: (i) “variação da norma padrão” e (ii) “norma inovadora” (cf. Quadro 4).

A primeira categoria se refere às variações ortográficas frente à linguagem padrão. Dada a natureza contextual da interpretação de CGU, já comentada por Sanguinetti *et al.* (2023), os autores não se concentraram na intencionalidade dos desvios, pois ela não pode ser determinada observando apenas o texto de superfície. Para organizar os fenômenos dessa classe em tipos, os autores recorreram a Damerau (1964) e à definição de caractere do padrão *Unicode*. Assim, os desvios podem estar relacionados a caracteres como letra, espaço, hífen e diacrítico.

A “variação da norma padrão” tem 4 tipos: (i) substituição (ou seja, um caractere é substituído por outro), (ii) omissão (ou seja, um caractere está faltando), (iii) inserção (ou seja, um caractere extra é inserido) e (iv) transposição (ou seja, dois caracteres adjacentes são transpostos). A substituição, a omissão e a inserção são subdivididas em dois subtipos, que basicamente se relacionam ao uso inadequado de diacríticos e a outros tipos de caracteres. A classe de transposição não tem subtipos.

A categoria “norma inovadora”, por sua vez, inclui alternativas lexicais a palavras padrão existentes e fenômenos frequentes encontrados na linguagem do Twitter e/ou do domínio do mercado de ações. A categoria é dividida em 6 tipos parcialmente baseados em baseados em Sanguinetti *et al.* (2023), a saber: (i) abreviação, (ii) neologismo, (iii) marca de expressividade, (iv) escrita homófona, (v) *token* dependente do meio e (vi) *token* específico do domínio.

Quadro 4: Taxonomia de fenômenos lexicais no DANTEStocks.

Phenomenon	Type	Subtype	Attested example	Standard form
Standard Norm Variation	**Substitution**	*Diacritic* (cedilla)	**lancamento** das notas	*lançamento*
		Other	**segunda feira**	*segunda-feira*
			Neh?	Né? (não é)
	Omission	*Diacritic*	capital **proprio**	*capital próprio*
		Other	**valu** ferris	*valeu ferris*
	Insertion	*Diacritic*	#PETR4 **fêz** uma Onda 2	*#PETR4 fez uma Onda 2*
		Other	montar um **Streaddle**	*montar um Straddle*
	Transposition	-	vc se manteve na **comrpa**?	*vc se manteve na compra?*
Innovative Norm	**Abbreviation**	*Initialism*	ação de **LP**	*ação de longo prazo*
		Shortening	(eles) falam **q**	(eles) *falam que*
			por **enqt**	*por enquanto*
		Contraction	**pq** será?	*por que será?*
	Neologism	*Agglutination*	44.6k no **Ibolixo**	*44.6 mil no Ibolixo*
		Derivation	**diretassa** do morgan	*diretaça do morgan*
		Foreign influence	#itub4 **estopou**	*#itub4 estopou*
	Expressiveness	*Graphemic stretching*	**choooooram**!	*choram*
		Punctuation repetition	linda!!!!	*linda!*
		Dialectal variation	De **zóio**!	*De olho!*
		Pictogram	:) 🚀 🚀	-
			muito $	*muito dinheiro*
		Capitalization	**LINNDAA**	*linda*
		Disguise	essa **p**a**	*essa puta*
	Homophone Writing	*Phonetization*	é **d+**	*é demais*
		Graphemic substitution	**xatiado**	*chateado*
		Onomatopoeia	**hahaha**	-
	Medium-dependent token	*Hashtag*	Presidente da **#PETR4**	-
		At-mention	né, **@user**?	não é, @user?
		URL	http://t.co/OQ3rDdWilf	-
		RT	**RT** @user...	-
		Truncation	ação sobe **fo**...	*ação sobe forte...*
		Code-switching	E ponto final! **PERIOD**!	-
	Domain-specific token	*Ticker*	**PETR4** subiu	-
		Cashtag	**$PBR** testando	-
		Decimal number	de **18,xx** a 21,00	-
		Valuation rate	ELET6 **+2,09%**	*ELET6 + 2,09 %*
		Temporal expression	**1T14, jun/14**	-
		Monetary value	perdeu só **R$20,00**	*perdeu só R$ 20,00*

Fonte: Adaptada de Scandarolli *et al.* (2023).

Com base no Quadro 4, vê-se que a tipologia de Scandarolli *et al.* (2023) unifica, de certa forma, os fenômenos lexicais e os dispositivos da plataforma mencionados por Sanguinetti *et al* (2023).

Além disso, observa-se pela tipologia que o DANTEStocks apresenta certas particularidades não previstas por Sanguinetti *et al* (2023), como as dependentes do domínio do mercado financeiro. Para ilustrar, destacam-se os *tickers* (códigos abreviados, normalmente compostos por letras maiúsculas e, em alguns casos, números, utilizados para identificar de forma única ações, fundos, criptomoedas ou outros ativos negociados em bolsa) e as *cashtags* (marcação textual utilizada em redes sociais, principalmente no Twitter (X), para referenciar tickers de ativos financeiros, antecedidos pelo símbolo “$”).

Para o tratamento desses fenômenos segundo o modelo UD, diretrizes de anotação específicas precisaram ser definidas. Quanto a um *ticker,* o lema corresponde à forma de superfície, a etiqueta PoS deve ser PROPN e a *deprel* é a correspondente à função do *ticker* no contexto. Para os casos de *cashtag,* o lema corresponde à forma de superfície e a etiqueta de PoS é PROPN quando integrada à sintaxe ou X quando não integrada (em inglês, *standalone*). Se integrada sintaticamente no enunciado, a *cashtag* deve ser anotada com a *deprel* correspondente à sua função no contexto; caso contrário, é dependente do predicado principal por meio da relação **parataxis:cashtag** (Di Felippo *et al.,* 2022; 2024).

2.3.2. A anotação-UD

Seguindo a decisão do projeto POeTiSA, a anotação-UD do DANTEStocks foi fatorada nos níveis morfológico e sintático, pois, segundo Pardo *et al.* (2021), a separação dos níveis produz melhores resultados dado que a tarefa é sofisticada. Antes, porém, da anotação em si, o *corpus* passou um pré-processamento.

a) Pré-processamento

Para a anotação desse recurso segundo o modelo UD, o conjunto inicial de 4.517 *tweets* de Silva *et al.* (2020) passou por um refinamento, consistindo na exclusão de 469 tweets distintos repetidos e/ou não pertencentes ao domínio (Gazana; Di-Felippo, 2022). O refinamento resultou nos 4.048 *posts* que foram efetivamente submetidos à anotação-UD. Ademais, o *tweet* foi tomado como unidade de análise e, com isso, os *posts* não foram segmentados em unidades estruturais menores como

sentenças ou sintagmas, e não se aplicou nenhum processo de normalização da linguagem.

Na sequência, o *corpus* foi tokenizado segundo os pressupostos da UD em um processo semiautomático (Silva, *et al.,* 2021). Como o modelo se baseia em uma visão lexicalista da sintaxe, as unidades básicas da anotação são as palavras sintáticas[12]. Assim, as palavras sintáticas (tokens) foram segmentadas automaticamente por uma versão do NLTK TweetTokenizer, enriquecida com regras específicas para o DANTEStocks (Silva *et al*. 2021). A ferramenta preservou a maioria dos *tokens* delimitados por espaços em branco, incluindo fonetização (por exemplo, “d+” (“demais”), *hashtag, cashtag*, *at-mention*, *emoticon* e URL, e separou *tokens* ortográficos únicos que correspondem a várias palavras (sintáticas), como clíticos, contrações (canônicas e não-canônicas), sinais de pontuação (exceto abreviações), taxas de avaliação das ações na bolsa e valores monetários com ortografia não convencional. Após a revisão manual da saída da ferramenta, o *corpus* totalizou 81.037 tokens.

b) Anotação das informações morfológicas

Como mencionado, as informações morfológicas no modelo UD englobam: lema, etiqueta de PoS e traços gramaticais.

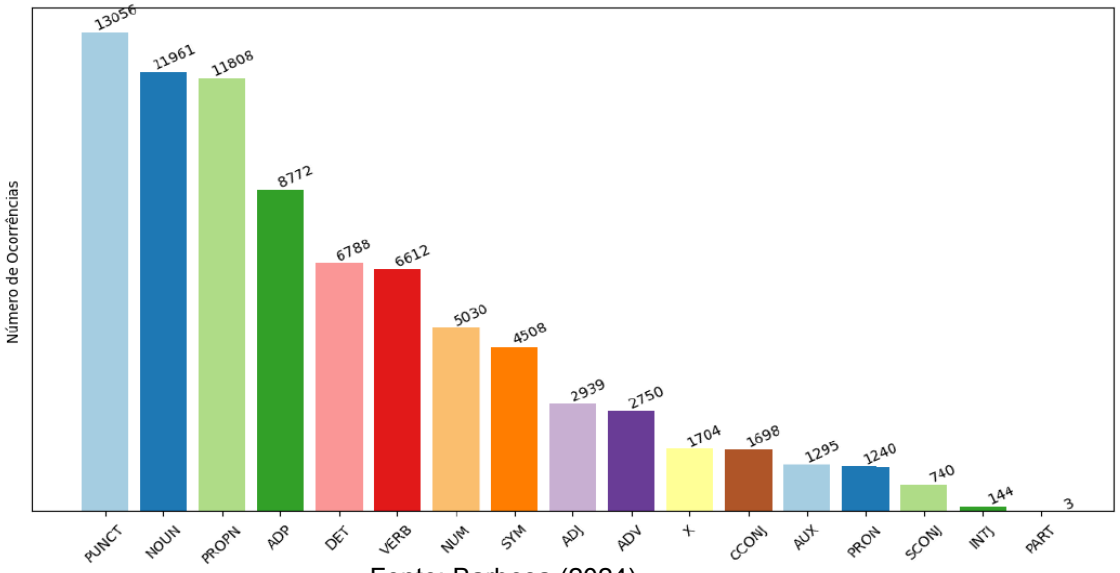
As *tags* PoS foram as primeiras a serem anotadas. De acordo com Silva *et al.* (2021), o processo foi semiautomático. Especificamente, o *corpus* foi submetido ao *parser* UDPipe 2, treinado incrementalmente sobre o UD-Portuguese Bosque e *tweets*. Os resultados do *parser* foram analisados manualmente por três anotadores. Para auxiliá-los nessa tarefa, diretrizes de anotação de PoS para os *tweets* do DANTEStocks foram definidas (Di-Felippo *et al*., 2022), as quais utilizadas em conjunto com as diretrizes adaptadas à língua portuguesa (Duran, 2021). Para os casos de discordância entre os anotadores, foram adjudicados por um linguista sênior com base nos mesmos manuais.

A Figura 3, exibe a distribuição das *PoS* no DANTEStocks. Vê-se que todas as 17 etiquetas UD ocorrem no DANTEStocks. PUNCT é a mais frequente, com cerca de 16% de todos os tokens recebendo essa tag, seguida por NOUN, com

[12] Tradução do termo em inglês *syntactic word*, que é definido como a unidade mínima a que corresponde uma função sintática. Na anotação-UD, palavras sintáticas são sinônimas de *tokens*.

cerca de 15%, e PROPN, correspondendo a aproximadamente 14% de todas as tags. Juntas, essas três etiquetas somam quase metade de todas as *tags* PoS (cerca de 45%).

Figura 3: Frequência das *tags PoS* no DANTEStocks.



Fonte: Barbosa (2024).

De acordo com Di Felippo *et al.* (2024), as lemas e os traços (*features*) gramaticais também foram anotados em um processo semiautomático. Especificamente, eles foram obtidos do PortiLexicon-UD (Lopes *et al*., 2022). Os dados gerados por esse dicionário ou léxico precisam de uma revisão manual relativamente grande devido à alta taxa de palavras/*tokens out-of-vocabulary* (isto é, não previstas em dicionário). Com relação aos traços, o cenário foi bastante diferente. A extração dos traços do PortLexicon foi guiada pelas *tags* PoS e lemas já validados manualmente, o que diminuiu o esforço de revisão manual dos traços. A maioria das correções foi referente a erros decorrentes da ambiguidade dos traços da classe VERB (*VerbFom, Mood, Tense, Genre*, *Number* e *Person*).

c) Anotação sintática

A anotação dos *deprels* no DANTEStocks foi feita em duas etapas semiautomáticas (Barbosa 2024, Di Felippo *et al.* 2024a). A primeira criou um *subcorpus* de referência

e a segunda etapa ajustou um *parser* pré-treinado para *tweets*, usando o *subcorpus* de referência como parte de seu conjunto de treinamento inicial, e anotou o restante do *corpus*. Para tanto, os 4.048 *tweets* foram agrupados em três grandes conjuntos em função do tipo de linguagem/estrutura: linguagem relativamente padrão, padrões estruturais recorrentes e outros (*tweets* que não pertencem aos outros dois conjuntos). Os *tweets* foram agrupados por meio do algoritimo *k-means* e *tf-idf* ("*term frequency–inverse document frequency*" (cf. Barbosa 2024).

A organização dos *tweets* nos referidos conjuntos permitiu selecionar instâncias de cada um deles para compor um *subcorpus* de referência de 1.000 *tweets*, cobrindo, assim, a diversidade estrutural do DANTEStocks. Além disso, a anotação semiautomática do *subcorpus* também foi baseada nessa classificação. Visando consistência na anotação, cada conjunto foi anotado e revisado manualmente em separado, começando pelos *tweets* com linguagem "padrão", os quais foram seguidos pelos *tweets* com padrões estruturais recorrentes e, por fim, pelos outros.

De um modo geral, os 1.000 *tweets* foram submetidos ao UDPipe 2, treinado sobre o UD-Portuguese Bosque. A anotação gerada pelo *parser* foi posteriormente revisada de forma manual por um único especialista. Ao final, obteve-se um *subcorpus* de referência, com anotação padrão ouro.

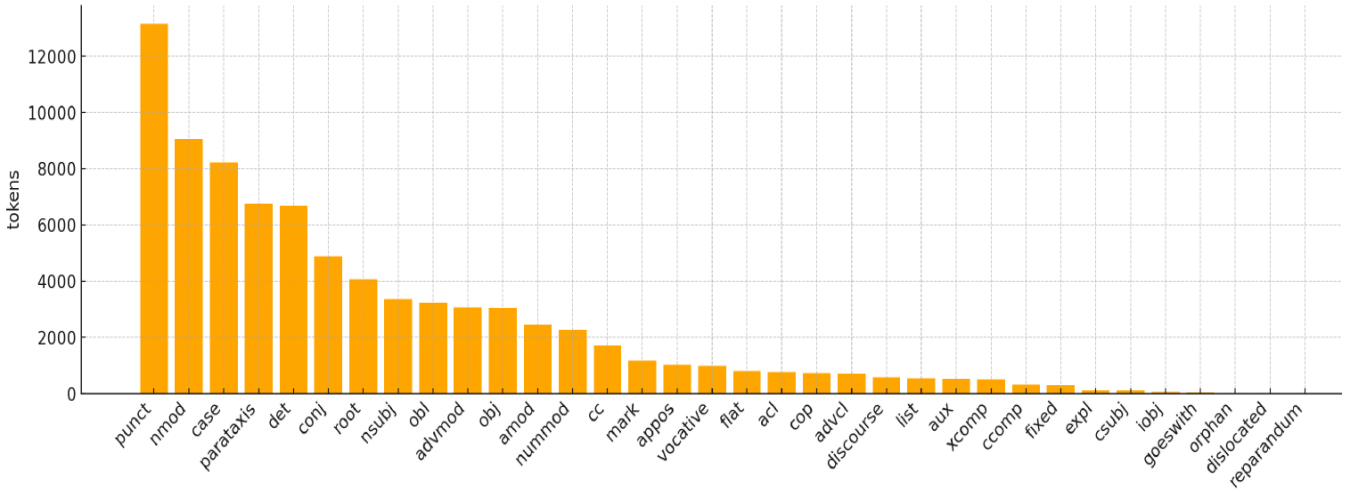
Vale ressaltar que o processo de revisão deu origem a diretrizes específicas para a anotação das *deprels* do modelo UD em *tweets* do domínio do mercado financeiro (Di Felippo *et al*., 2024c). As diretrizes foram usadas para apoiar a revisão manual do restante do *corpus*, que foi feita ao treinar um *parser* de última geração nos *tweets* do DANTEStocks.

O restante do *corpus* foi anotado personalizando o Stanza (Qi *et al*. 2020) para o DANTEStocks. O Stanza é um modelo pré-treinado bem conhecido para o português, tendo a vantagem de ser um *pipeline* amigável para análise de texto. O processo começou com a arquitetura base do Stanza, ajustada no *Porttinari-base* acrescido pelo *subcorpus* de referência. O modelo de *parser* resultante desse treinamento inicial foi usado para anotar um novo pacote de dados (proveniente dos 3.048 *tweets*), que foi revisado manualmente e incorporado ao conjunto de dados inicial, sendo então usado para iniciar uma nova execução de treinamento do Stanza. Esse ciclo continuou de forma incremental até que o último pacote de *tweets*

tivesse sido anotado/revisado. Os pacotes de *tweets* foram adicionados na mesma ordem aplicada na anotação do *subcorpus* de referência: *tweets* de linguagem padrão, *tweets* de padrões estruturais e *tweets* com propriedades lexicais/estruturais variadas.

O desempenho do Stanza foi medido, segundo Di Felippo *et al.* (2024a), com base no *Unlabeled Attachment Score* (UAS) e *Labeled Attachment Score* (LAS). Ao final, ele obteve UAS 95,78% e LAS de 94,62%. Esses resultados são considerados bons, dada a complexidade da tarefa. Figura 4 ilustra a distribuição geral das relações de dependência (sem sub-relações) no DANTEStocks.

Figura 4: Frequência das *deprel* UD no DANTEStocks.



Fonte: Barbora (2024).

A respeito da frequência das relações de dependência do modelo UD no DANTEStocks exibidas na Figura 4, observa-se que 34 das 37 previstas pelo modelo foram empregadas na anotação do *corpus*. As relações não empregadas foram **clf**, **compound** e **dep**. Das 34, ressalta-se que **parataxis** é a quarta mais frequente, com 6.733 ocorrências. Isso comprova aquilo que Di Felippo *et al.* (2021) já haviam observado sobre a alta frequência de *tweets* fragmentados, uma vez que a **parataxis** se estabelece entre dois elementos que poderiam ter relação sintática entre si, porém essa relação não está explicitada.

Diante disso, este trabalho buscou caracterizar a referida fragmentação por meio da descrição da *deprel* parataxis no DANTEStocks. Tal investigação está relatada a seguir.

3. Estudo, seleção e preparação dos dados

Visando atingir o objetivo proposto, este trabalho foi equacionado em 5 etapas metodológicas: (i) estudo da definição de **parataxis** no modelo UD e o emprego dessa etiqueta na anotação do DANTEStocks, (iii) seleção das ocorrências de **parataxis** no *corpus*, (iv) organização das ocorrências em tabelas e (v) descrição da **parataxis** em função de dois critérios, a saber: subrelação e etiqueta PoS do *head.*

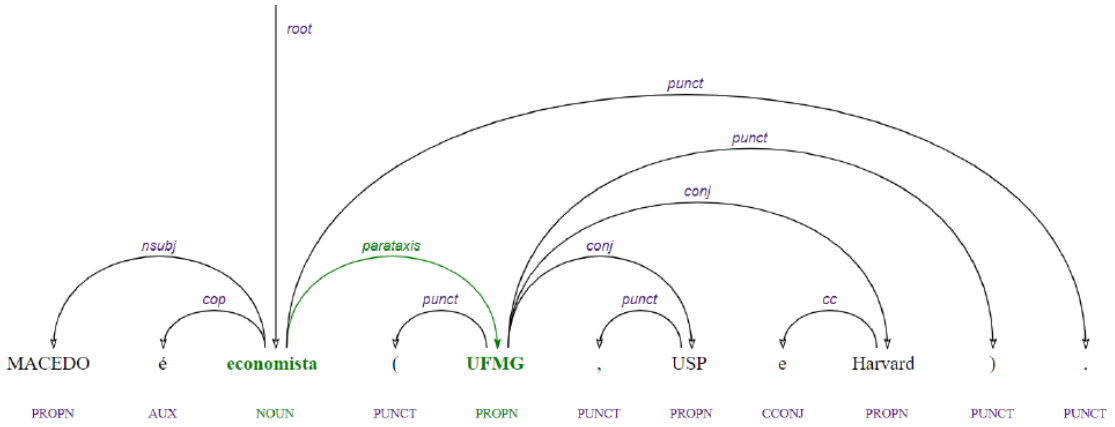
3.1. A *deprel* parataxis segundo o modelo UD

O termo “parataxis” tem origem no grego e significa “colocar lado a lado”. Assim, esse termo, empregado como a etiqueta de uma das 37 relações de dependência sintática no modelo UD (Nivre *et al.,* 2016; Nivre *et al.* 2020), rotula a relação entre elementos de um mesmo enunciado, colocados lado a lado sem qualquer coordenação explícita, subordinação ou relação argumental com a palavra/*token* principal. Em outras palavras, é a relação que ocorre entre dois elementos da sentença/enunciado que não têm conexão sintática explícita. Ademais, salienta-se que a **parataxis** pode ocorrer nos dois sentidos, dependendo de onde se encontra o **root**.

Na língua geral, em especial no português padrão, essa *deprel* é empregada nos casos de: (i) diálogo com o interlocutor, em que a oração que expressa essa interação é anotada como dependente de **parataxis** (p.ex.: **Olha**, assim não vai dar.); (ii) discurso direto, em que o verbo de elocução encaixado entre duas partes do discurso relatado é dependente de **parataxis** (p.ex.: Por enquanto, **diz** uma fonte, o Executivo prefere não contar com isso.) e (iii) modificadores nominais que, uma vez desenvolvidos, seriam **nmod**; esse é o caso, por exemplo, da relação entre um nominal e outro nominal entre parênteses que indica afiliação (p.ex.: Dória (**PSDB**) retirou sua candidatura) (Dória, **do PSDB**) (Duran, 2022).

A Figura 5 ilustra o caso de parataxis descrito em (iii). O exemplo em questão, que consta em Duran (2022), foi extraído do *corpus* Porrtinari-base.

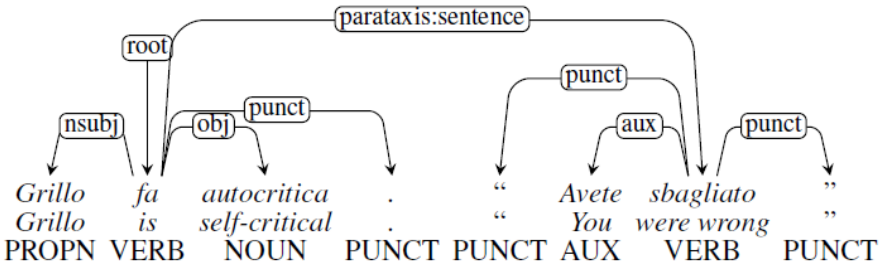
Figura 5: Exemplo de relação de parataxis no português padrão.



Fonte: Duran (2022).

Em CGU no geral, a *deprel* **parataxis** é empregada, segundo Sanguinetti *et al.* (2023), para anotar os casos de *hashtag,* URL e marcas de *retweet* (RT) quando não integrados à estrutura sintática dos *tweets* (isto é, em contexto *standalone*), além de sentenças *side-by-side* (cf. Figura 6)*,* isto é, coladas lado a lado. Para os casos de *hashtag,* URL e multiplicidade de sentenças, os autores sugerem o emprego de sub-relações, a saber: **parataxis:hashtag, parataxis:url** e **parataxis:sentence**.

Figura 6: Exemplo de **parataxis** em caso de sentença *side-by-side*.



Fonte: Sanguinetti et al. (2023).

No DANTEStocks, um conjunto de sub-relações especificamente para os casos de *hashtag, cashtag* e URL em contexto *standalone*, a saber: **parataxis:hashtag, parataxis:cashtag** e **parataxis:url**. A sub-relação **parataxis:sentence**, de Sanguinetti *et al.* (2023), não foi empregada por entender que, excluindo-se os casos de **parataxis** com sub-relação, os restantes são todos do tipo “elemento *side-by-side*”*,* não necessitando, assim, de subespecificação.

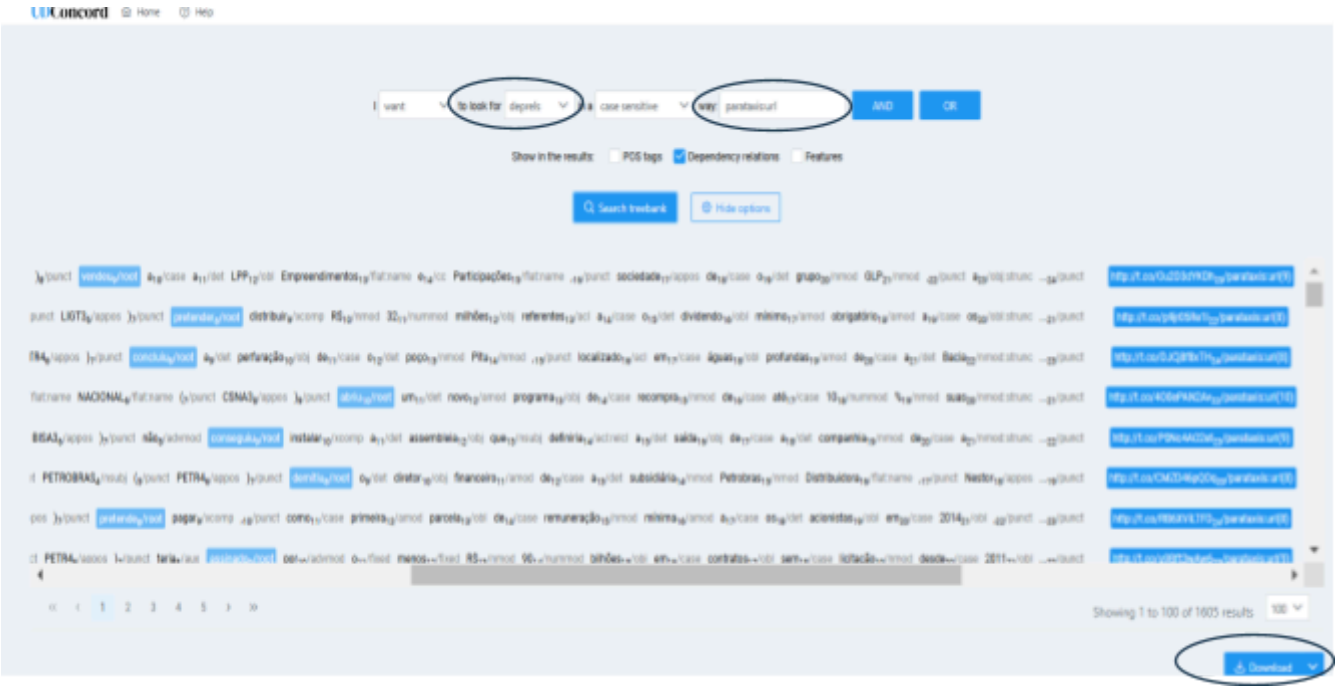
No DANTEStockes, deu-se preferência pela anotação das ocorrências de elementos lado a lado que envolvessem casos de truncamento, lexical ou estrutural, por meio das seguintes sub-relações: **parataxis:wtrunc** e **parataxis:strunc.**

3.2. Seleção dos casos de parataxis e levantamento estatístico

Para o estudo da **parataxis**, foi necessário percorrer o *corpus* com anotação-UD a fim de extrair todas as ocorrências dessa *deprel*. Para isso, utilizou-se o concordanciador[13] *online* UDConcord (Miranda, Pardo, 2022).

Uma vez que os arquivos CoNLL-U de todos os *tweets* do *corpus* tenham sido submetidos à ferramenta, ela permite a busca por diferentes critérios, sendo *deprel* um deles. Dessa forma, o conjunto de ocorrências de cada um dos tipos de **parataxis** foi extraído. A Figura 7 ilustra a busca por **parataxis:url**. A ferramenta permite que o usuário faça o *download* dos resultados da busca em diferentes formatos; no caso, optou-se pelo *.txt.* Ressalta-se mais uma vez que o processo ilustrado na Figura 7 foi repetido para cada um dos tipos de **parataxis**.

Figura 7: Ilustração da extração dos casos de **parataxis** no DANTEStocks.



Fonte: O autor (2025).

[13] Um concordanciador é uma ferramenta de análise de *corpus* que permite encontrar e exibir todas as ocorrências de uma palavra ou expressão em um *corpus*, juntamente com seu contexto imediato.

Após a extração, levantou-se a estatística das relações em questão. Do total de **6.733** ocorrências de **parataxis** da Tabela 1 (com e sem sub-relação), os 3.840 casos sem sub-relação equivalem a **57**%, enquanto as 2.893 ocorrências com as diferentes sub-relações equivalem a **43**%.

Tabela 1:. Frequência de ocorrência de **parataxis** no DANTEStocks.

Deprel	**Qt.**	**%**
parataxis	3840	57%
parataxis:url	1605	43%
parataxis:hashtag	1007	
parataxis:strunc	133	
parataxis:cashtag	111	
parataxis:wtrunc	37	
TOTAL	**6733**	**100%**

Fonte: O autor (2025).

Após a seleção ou extração das ocorrências de cada tipo de **parataxis** do *corpus,* construiu-se um *script* em *Python* que, a partir dos arquivos *txt* de cada tipo de **parataxis**, realizou as seguintes tarefas: (i) identificação das diferentes combinações de *head* e dependente em função das etiquetas PoS, (ii) levantamento estatístico de cada combinação, e (iii) seleção de uma ocorrência-exemplo para cada combinação *head*-dependente. O *script,* que consta como Apêndice 1 deste documento, gera os resultados para cada tipo de **parataxis** em tabelas no formato *.xlsx*, como a ilustrada na Tabela 2. Todas as tabelas constam como Apêndice 2 deste documento.

3.3. Organização dos dados

Cada tabela gerada pelo referido *script* possui 5 colunas. A primeira, denominada *Deprel,* exibe o tipo de **parataxis** sob análise*.* A segunda e a terceira, denominadas *Head* e Dependente, exibem, respectivamente, a etiqueta PoS do *token* na função de *head* e do *token* na função de dependente que se conectam pela **parataxis** em questão, independente da ordem em que esses *token* ocorrem no *tweet*. A quarta coluna apresenta a frequência das ocorrências em função da combinação da PoS do *head* e do dependente. E, por fim, a quinta coluna (ID_exemplo) exibe a

ocorrência-exemplo. Para os restante deste documento, a quinta coluna das tabelas será omitida.

Tabela 2: Exemplo de organização dos dados para descrição.

Deprel	**Head**	**Dependente**	**Qt.**	**ID_Exemplo**
parataxis:cashtag	VERB (root)	X	87	dante_01_466744859461017601
	VERB (root)	PROPN	11	dante_01_472473292683886592l
	NOUN (root)	X	9	dante_01_459737045037232129l
	NOUN (root)	PROPN	2	dante_01_443762116567511040l
	PROPN (root)	PROPN	1	dante_01_442056119322959873l
	VERB	X	1	dante_01_444639828160696320l
		TOTAL	111	--

Fonte: O autor (2025).

Com base nos dados organizados nas tabelas, fez-se uma descrição dos casos de **parataxis** no DANTEStock. Primeiramente, discorrem-se observações a respeito das *deprels* com sub-relação para, na sequência, tratar dos casos de **parataxis** sem sub-relação. Tais observações foram feitas considerando a combinação das etiquetas PoS dos *tokens,* mas também a ordem de ocorrência desses *tokens* na relação. As observações da sobre a ordem de ocorrência dos *token* referente a cada tipo de **parataxis** resultam do estudo das árvores sintáticas das ocorrências.

3.4. Descrição da parataxis no DANTEStocks

3.4.1. Com sub-relação

a) Parataxis:cashtag: essa *deprel* caracteriza os casos em que um *token* precedido pelo símbolo de cifrão ($) é dependente de outro *token*. Vê-se, com base na Tabela 3, que o *head* varia entre VERB, NOUN e PROPN, sendo o dependente X ou PROPN. No *corpus,* essa relação pode ocorrer em ambas as direções, da esquerda para a direita e vice-versa, pois as *cashtags* ocorrem tanto no início como no fim dos *tweets.* Com exceção de apenas 1 caso (exemplo 7) (VERB+X da Tabela 3), os demais têm como *head* o *root* do *tweet.* No exemplo (7), o *tweet* em questão é composto por duas sentenças, sendo "$LIGT3" conectado por **parataxis:cashtag** ao verbo da segunda sentença, que não é o *root*; no caso, o *root* é "arquivou".

(7) A Light S.A. arquivou hoje um Comunicado ao Mercado. Confira: http://t.co/ufz7iCBV92 **$LIGT3**

Tabela 3: Frequência de ocorrência de **parataxis:cashtag** no DANTEStocks.

Deprel	**Head**	**Dependente**	**Qt.**
parataxis:cashtag	VERB (root)	X	87
	VERB (root)	PROPN	11
	NOUN (root)	X	9
	NOUN (root)	PROPN	2
	PROPN (root)	PROPN	1
	VERB	X	1
		TOTAL	111

Fonte: O autor (2025).

b) Parataxis:hashtag: essa *deprel* caracteriza os casos em que um *token* precedido pelo símbolo de *hashtag* (#) é dependente de outro *token.* A Tabela 4 mostra que o *head* pode ter etiqueta PoS variada, sendo o dependente X, PROPN ou NOUN. No *corpus,* essa relação pode ocorrer em ambas as direções, como as *cashtags.* Os 3 casos em que o dependente é NOUN consistem em *hashtags* não compostas por um *ticker,* como em (8).

(8) SBSP3 com um belo sinal de entrada de uma WW **#volume** http://t.co/ONawWUOwE3

Tabela 4: Frequência de ocorrência de parataxis:hashtag no DANTEStocks.

Deprel	***Head***	**Dependente**	**Qt.**
parataxis:hashtag	VERB (root)	X	698
	NOUN (root)	X	138
	VERB (root)	PROPN	74
	NOUN (root)	PROPN	20
	PROPN (root)	X	16
	ADJ (root)	X	13
	ADV (root)	PROPN	10
	PRON (root)	X	6
	PRON (root)	PROPN	5
	ADJ (root)	PROPN	4
	SYM (root)	X	3
	VERB	PROPN	3
	ADV (root)	X	2
	NUM (root)	X	2
	VERB (root)	NOUN	2
	VERB	X	2
	X (root)	X	2
	ADJ	X	1
	AUX (root)	PROPN	1
	NOUN	X	1
	NUM (root)	PROPN	1
	PROPN (root)	NOUN	1
	PROPN (root)	PROPN	1
	SYM (root)	PROPN	1
		TOTAL	1007

Fonte: O autor (2025).

c) Parataxis:url: essa *deprel* caracteriza os casos em que uma URL é dependente de outro *token.* O sentido dessa relação é sempre da esquerda para a direita, uma vez que a URL dependente ocorre sempre ao final dos *tweets.* Segundo o manual de anotação de etiquetas PoS-UD de Di Felippo *et al.* (2022), as URL devem ser sempre etiquetadas como SYM, seguindo, aliás, a própria literatura internacional sobre anotação-UD de CGU (cf. Tabela 5). Assim, os 3 casos de **parataxis:url** em que o dependente é X ou PROPN consistem em equívocos de anotação. Observa-se nas ocorrências exibidas em (9), (10) e (11) que esses casos são

errôneos porque **parataxis:url** foi empregada para conectar 2 *hashtags* e 1 *cashtag* ao *root.*

(9) Várias empresas que anunciam recompra de ações na verdade não recompram absolutamente nada. Não há nenhum controle sobre isso #CSN **#CSNA3_PROPN**

(10) # text = Por tudo que foi abusado da outrora promissora PETR4 **#EuApoioCPIdaPetrobras_X**

(11) Preço da #celulose na #China cai e derruba ações de produtor brasileiro: http://t.co/9jWPoaejZb $FIBR3 **$SUZB3_X**

Tabela 5: Frequência de ocorrência de **parataxis:url** no DANTEStocks.

DEPREL	**Head**	**Dependente**	**Qt.**
parataxis:url	VERB (root)	SYM	778
	NOUN (root)	SYM	688
	PROPN (root)	SYM	37
	PROPN	SYM	32
	ADJ (root)	SYM	22
	SYM (root)	SYM	11
	VERB	SYM	9
	ADV (root)	SYM	7
	NOUN	SYM	3
	X (root)	SYM	3
	ADV	SYM	2
	AUX (root)	SYM	2
	NUM (root)	SYM	2
	PRON (root)	SYM	2
	X	SYM	2
	ADJ	SYM	1
	PRON	SYM	1
	PRON (root)	X	1
	VERB (root)	PROPN	1
	VERB (root)	X	1
		TOTAL	1605

Fonte: O autor (2025).

d) Parataxis:strunc: essa *deprel* codifica um truncamento estrutural, sendo empregada para conectar o *root* do *tweet* a um *token* que é núcleo de um segmento estruturalmente quebrado devido ao limite de caracteres. Assim, essa

deprel ocorre em *tweets* com multiplicidade de segmentos com estrutura sintática e sempre da esquerda para a direita. Como as combinações de etiquetas PoS mais frequentes são VERB (*root*)+NOUN e VERB (*root*)+VERB (cf. Tabela 6), pode-se dizer que essa *deprel* captura a justaposição de uma sentença e um SN ou de duas sentenças (com pontuação adequada ou não).

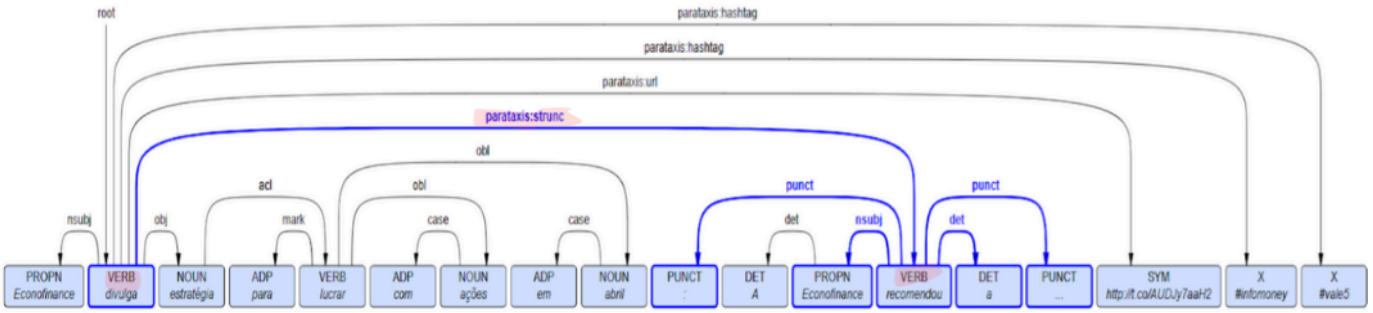
Tabela 6: Frequência de ocorrência de **parataxis:strunc** no DANTEStocks.

Deprel	***Head***	**Dependente**	**Qt.**
parataxis:strunc	VERB (root)	NOUN	46
	VERB (root)	VERB	42
	NOUN (root)	VERB	10
	VERB (root)	PROPN	6
	NOUN (root)	NOUN	5
	VERB (root)	ADV	4
	NOUN (root)	PROPN	3
	PROPN (root)	VERB	2
	VERB (root)	AUX	2
	VERB (root)	X	2
	ADJ (root)	NOUN	1
	ADJ (root)	PRON	1
	NOUN	PROPN	1
	NOUN (root)	X	1
	VERB	ADJ	1
	VERB (root)	ADP	1
	VERB	ADV	1
	VERB	AUX	1
	VERB	NOUN	1
	VERB	X	1
	ADJ	AUX	1
		TOTAL	133

Fonte: O autor (2025).

Na Figura 8, ilustra-se o segundo caso, pois o núcleo (VERB) da segunda “sentença” truncada (“*A Econofinance* recomendou *a* ...”) é dependente por **parataxis:strunc** do *root* (VERB) da predicação principal (“divulga”).

Figura 8: Exemplo de **parataxis:strunc** no DANTEStocks.



Fonte: O autor (2025).

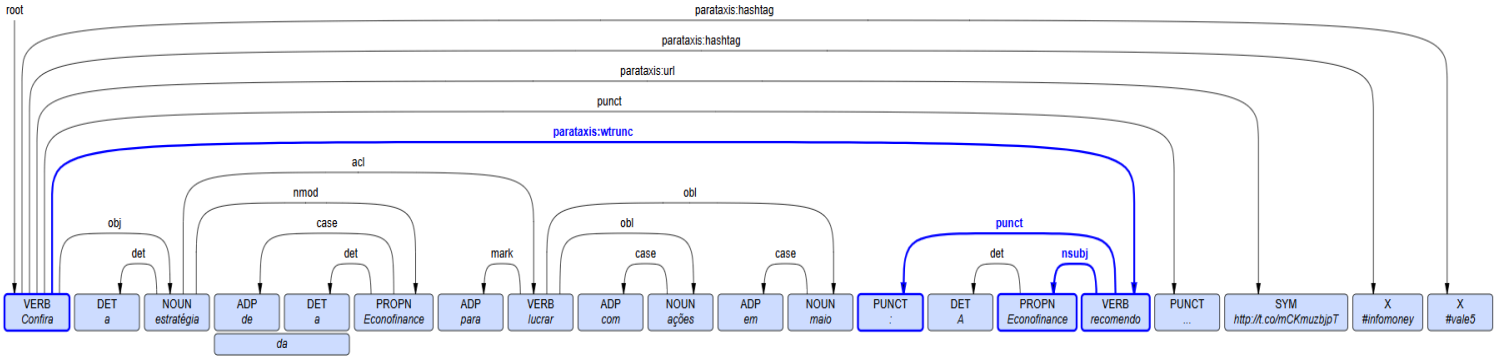
e) Parataxis:wtrunc: essa *deprel* codifica um truncamento lexical, sendo empregada quando a última palavra/*token* do trecho truncado for truncada. Ela comumente codifica a relação em que o dependente possui estrutura sintática e é sempre da direita para a esquerda. A combinação de etiqueta PoS mais frequente é VERB (root)+VERB (cf. Tabela 7). Dessa forma, essa deprel comumente captura a justaposição de duas sentenças (com pontuação adequada ou não), como ilustrado na Figura 9.

Tabela 7: Frequência de ocorrência de **parataxis:wtrunc** no DANTEStocks.

Deprel	***Head***	**Dependente**	**Qt.**
parataxis:wtrunc	VERB (root)	VERB	11
	NOUN (root)	SYM	6
	VERB (root)	NOUN	4
	VERB (root)	PROPN	3
	VERB	VERB	3
	VERB (root)	ADJ	2
	VERB (root)	ADV	2
	VERB (root)	X	2
	NOUN	VERB	1
	NUM (root)	SYM	1
	VERB (root)	AUX	1
	VERB	NOUN	1
	VERB (root)	SYM	1
		TOTAL	38

Fonte: O autor (2025).

Figura 9. Exemplo de **parataxis:wtrunc** no DANTEStocks



Fonte: O autor (2025).

3.4.2. Sem sub-relação

No tocante à *deprel* **parataxis**, sem sub-relação, as combinações de PoS do *head* e dependente são muito variadas, como pode ser observado nas Tabelas 8 e 9. No entanto, fica evidente que as combinações nas quais o *head* é VERB ou NOUN (cf. Tabela 8) são muito mais frequentes que as demais (cf. Tabela 9).

A frequência elevada da combinação VERB (*root*) + VERB do item (a) da Tabela 8 (isto é, 860 casos) indica predominância pela justaposição de duas estruturas sentenciais (ou, no mínimo, envolvendo sintagmas verbais) na composição dos *tweets* com **parataxis**. Essa frequência é muito similar à da combinação VERB (*root*) + NOUN ou PROPN do item (a) da Tabela 9 (isto é, 838 (621+217) ocorrências), indicando que a justaposição de uma estrutura verbal (central) e uma nominal também é relevante no *corpus.* Tais estruturas nominais, nesses casos, são diferentes de *hashtag, cashtag,* URL ou envolvendo truncamentos, uma vez que esses fenômenos são capturados pelas sub-relações já descritas.

Tabela 8: Frequência de **parataxis** com *head* VERB e NOUN no DANTEStocks.

(a)

Head	Dependente	Qt.
VERB (root)	VERB	860
VERB (root)	NOUN	621
VERB (root)	PROPN	217
VERB (root)	X	200
VERB (root)	ADJ	66
VERB	X	35
VERB	VERB	30
VERB (root)	ADV	26
VERB (root)	NUM	23
VERB	PROPN	23
VERB (root)	SYM	19
VERB	NOUN	16
VERB (root)	PRON	15

(b)

Head	Dependente	Qt.
NOUN (root)	NOUN	350
NOUN (root)	VERB	303
NOUN (root)	PROPN	110
NOUN	NOUN	84
NOUN (root)	ADJ	65
NOUN (root)	X	62
NOUN (root)	NUM	31
NOUN	PROPN	27
NOUN (root)	SYM	9
NOUN	VERB	9
NOUN	X	9
NOUN (root)	PRON	8
NOUN	NUM	4

Fonte: O autor (2025).

Fonte: O autor (2025).

Tabela 9: Frequência de **parataxis** com *head* diferente de VERB e NOUN.

Head	Dependente	Qt.
PROPN	PROPN	65
PROPN (root)	VERB	59
PROPN (root)	NOUN	49
PROPN	NOUN	46
PROPN (root)	X	16
PROPN (root)	PROPN	15
PROPN	X	9
PROPN (root)	ADJ	5
PROPN (root)	ADV	5
PROPN	ADV	4
PROPN	VERB	4
PROPN (root)	NUM	3
PROPN (root)	SYM	3
PROPN (root)	AUX	2
PROPN (root)	PRON	2
PROPN	PRON	2
PROPN	AUX	1
PROPN	NUM	1
TOTAL		**291**

Head	Dependente	Qt.
ADJ (root)	VERB	41
ADJ (root)	NOUN	20
ADJ (root)	PROPN	6
ADJ (root)	SYM	4
ADJ	VERB	4
ADJ (root)	NUM	3
ADJ	PROPN	3
ADJ (root)	ADJ	2
ADJ (root)	X	2
ADJ (root)	ADV	1
ADJ	NOUN	1
ADJ (root)	PRON	1
ADJ	X	1
TOTAL		**89**

Head	Dependente	Qt.
SYM	SYM	40
SYM (root)	VERB	11
SYM	PROPN	7
SYM (root)	NOUN	5
SYM (root)	PROPN	5
SYM (root)	X	4
SYM	NOUN	3
SYM (root)	ADJ	2
SYM	X	2
SYM (root)	INTJ	1
SYM (root)	NUM	1
TOTAL		**81**

Head	Dependente	Qt.
ADV (root)	VERB	19
ADV (root)	NOUN	6
ADV	VERB	4
ADV (root)	PROPN	3
ADV (root)	X	3
ADV (root)	ADJ	1
ADV	ADJ	1
ADV (root)	ADV	1
ADV (root)	SYM	1
ADV	X	1
TOTAL		**40**

Head	Dependente	Qt.
PRON (root)	VERB	14
PRON (root)	PROPN	5
PRON (root)	NOUN	3
PRON	VERB	2
PRON (root)	ADJ	1
PRON	PROPN	1
PRON (root)	SYM	1
PRON (root)	X	1
PRON	X	1
TOTAL		**29**

Head	Dependente	Qt.
NUM (root)	VERB	13
NUM	SYM	5
NUM (root)	NOUN	3
NUM (root)	X	2
NUM	X	2
NUM (root)	ADV	1
NUM	NOUN	1
NUM	PROPN	1
NUM (root)	SYM	1
TOTAL		**29**

Head	Dependente	Qt.
AUX (root)	VERB	7
AUX (root)	NOUN	2
AUX (root)	PROPN	2
AUX (root)	X	2
AUX (root)	ADJ	1
TOTAL		**14**

Head	Dependente	Qt.
INTJ	ADJ	1
INTJ (root)	NOUN	1
INTJ (root)	NUM	1
INTJ (root)	VERB	1
TOTAL		**4**

Head	Dependente	Qt.
X (root)	VERB	3
X (root)	NOUN	3
X	NOUN	1
X	PROPN	1
TOTAL		**8**

Fonte: O autor (2025).

Algo semelhante pode ser dito sobre as **parataxis** com NOUN (*root*) do item (b) da Tabela 8. As combinações mais frequentes são com dependentes VERB e NOUN.

Só que, nesse cenário, NOUN é mais frequente (350) que VERB (303). Aliás, ao se considerar também as ocorrências de PROPN como dependente de NOUN (*root*) na **parataxis**, vê-se que a justaposição de estruturas nominais (isto é, NOUN (*root*) + NOUN ou PROPN) é a bem relevante estatisticamente (350 + 110 = 460 ocorrências).

4. Considerações finais

Como menciona Sanguinetti *et al.* (2023), a descrição linguística de um *corpus x* é fundamental para propor diretrizes de anotação consistente para os fenômenos que caracterizam o gênero textual coberto por ele. Ademais, tal descrição também permite buscar compreender o desempenho de ferramentas de análise/anotação do *corpus x*, especialmente no que tange aos erros cometidos por ela.

Assim, conduziu-se a descrição da fragmentação estrutural dos *tweets* do DANTEStocks, que se deve, em partes, pelo fato de o *tweet* ter sido considerado a unidade básica de análise, não sendo submetido a nenhum tipo de segmentação em unidades estruturais menores como orações ou sintagmas. Para tanto, a descrição se baseou na anotação sintática segundo o modelo UD do referido *corpus.*

Tal anotação permitiu verificar que a justaposição de segmentos sem relação sintática clara entre si é bastante frequente no *corpus,* uma vez que a relação de dependência (*deprel*) **parataxis** (com e sem subespecificações), responsável por capturar esse tipo de fenômeno, é a 4ª mais recorrente com 6.733 casos.

Considerando como critério de análise a combinação entre as etiquetas PoS do *head* e dependente, observou-se que a **parataxis** genérica, sem subespecificação, ocorre com combinações variadas, mas com predominância dos seguintes padrões (independente da ordem das etiquetas): VERB (*root*) + VERB/NOUN/PROPN e NOUN (*root*) + VERB/NOUN/NOUN. E isso pode indicar que a fragmentação dos *tweets* comumente envolve segmentos verbais e nominais justapostos. A respeito das etiquetas de **parataxis** com sub-relações, destaca-se que, com exceção de **parataxis:strunc** e **parataxis:wtrunc**, os demais tipos (**parataxis:hashtag, parataxis:cashtag** e **parataxis:url)** possuem dependentes lexicais, pois equivalem a apenas uma palavra/*token.* Embora a PoS do *head* seja variada, o dependente de **parataxis:hashtag** e **parataxis:cashtag** é sempre X ou PROPN e de **parataxis:url**, SYM. Os casos de **parataxis:strunc** e **parataxis:wtrunc** podem apresentar dependentes lexicais ou estruturais (sintagmas ou sentenças truncadas).

Considerando a direção da relação, destaca-se que **parataxis** (genérica)**, parataxis:hashtag** e **parataxis:cashtag** ocorrem em ambas, ao passo que

parataxis:url, parataxis:strunc e **parataxis:wtrunc** ocorre da esquerda para direita.

Diante das observações feitas sobre a **parataxis** (com e sem sub-relação) no DANEStocks, pode-se dizer que se sabe mais atualmente sobre a fragmentação dos *tweets* do mercado financeiro do que se sabia antes da condução deste TCC.

Como limitação, sobretudo de tempo de desenvolvimento de projeto, outros aspectos interessantes a respeito da **parataxis** não foram descritos, como a própria ordem de ocorrência das etiquetas PoS nos padrões combinatórios. Outra limitação foi a respeito do processamento automático dos dados para análise. Mesmo com dedicação e esforço para compreender a linguagem *Phyton,* a construção do *script* demandou a contribuição de um linguista com formação em computação. Por outro lado, esse *script* pode ser visto como resultado de um trabalho colaborativo entre especialistas com conhecimentos complementares.

Por fim, o desenvolvimento deste Trabalho de Conclusão de Curso possibilitou à discente explorar (i) a tarefa de anotação de *corpus* com informação sintática, especificamente segundo o modelo gramatical UD, e (ii) as características linguísticas do *tweet* enquanto gênero CGU.

Embora a anotação sintática de *corpora* (manual e automática) seja um tópico abordado em algumas disciplinas no curso de Bacharelado em Linguística, essa tarefa é comumente tratada de forma mais genérica, isto é, sem dar ênfase a um modelo gramatical específico, e aplicada a textos de linguagem formal ou padrão.

Este TCC, em particular, possibilitou investigar a anotação sintática de um dos gêneros CGU (o *tweet*), sendo que, por causa das peculiaridades da linguagem, essa tarefa é bem diferente da realizada em textos formais/padrão (como o jornalístico). Assim, o desenvolvimento do TCC permitiu ampliar e aprofundar os conhecimentos em PLN, uma vez que a descrição ora apresentada promoveu uma compreensão mais sólida e ampla sobre a anotação sintática de *corpus* ao tratar um gênero diferente daquele estudado na graduação e um modelo gramatical específico e amplamente difundido no PLN hoje.

Além disso, a descrição das características dos *tweets* do *corpus* DANTEStocks, capturadas pela relação de dependência **parataxis** (e suas

sub-relações), indicou o emprego errôneo de **parataxis:url** em alguns casos cujos dependentes não têm a etiqueta SYM. Esses casos já foram reportados aos responsáveis pela construção do *corpus* que devem corrigi-los na próxima versão do *corpus* a ser disponibilizada. Embora sejam apenas 3 casos, pode-se dizer que este trabalho contribuiu para aumentar a consistência da anotação e, com isso, a qualidade do DANTEStocks enquanto recurso para as pesquisas no PLN.

5. Referências bibliográficas

AFONSO, S. et al. Floresta sintá(c)tica: um treebank para o português. In: ENCONTRO NACIONAL DA ASSOCIAÇÃO PORTUGUESA DE LINGUÍSTICA, 17, 2002, Lisboa. **Anais** [...]. Lisboa, 2002. p. 533-545.

BARBOSA, B.K.S. **Descrição sintático-semântica de nomes predicadores em tweets do mercado financeiro em português**, 2024, 208p. Dissertação (Mestrado em Linguística e Língua Portuguesa) - UFSCar, São Carlos, 2024.

BICK, E. **The parsing system "Palavras": automatic grammatical analysis in a constraint grammar framework**. 1st edition. Aarhus University Press, 2000.

BRUCKSCHEN, M. et al. Anotação Linguística em XML do Corpus PLN-BR. **Relatório Técnico do NILC**, NILC-TR-09-08, 2008.

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. *Online.* **Anais** [...]. Porto Alegre, 2021, p. pp. 335-343.

DI FELIPPO, A. et al. Diretrizes de Anotação de PoS Tags em Tweets do Mercado Financeiro: Orientações para anotação em língua portuguesa segundo a abordagem *Universal Dependencies* (UD). **Relatório Técnico do ICMC**, n. 438. ICMC-USP. São Carlos-SP, 24p., 2022.

DI FELIPPO, A. et al. A Dependency Treebank of *tweets* in Brazilian Portuguese: syntactic annotation issues and approach. In: SYMPOSIUM IN INFORMATION AND HUMAN LANGUAGE TECHNOLOGY, 15, 2024. Belém. **Proceedings** [...], Belém, 2024A. P. 192-201.

DI FELIPPO, A. et al. Diretrizes de anotação de relações de dependência em *tweets* do mercado financeiro. **Relatório Técnico do ICMC**, n. 446. ICMC-UDP. São Carlos-SP, 70p., 2024b.

DI FELIPPO, A. et at. Genipapo - a multigenre dependency parser for Brazilian Portuguese. In: SYMPOSIUM IN INFORMATION AND HUMAN LANGUAGE TECHNOLOGY, 15, 2024. Belém. **Proceedings** [...], Belém, 2024c. p. 257-266.

DURAN, M.S. Manual de Anotação de PoS tags: Orientações para anotação de etiquetas morfossintáticas em Língua Portuguesa, seguindo as diretrizes da abordagem *Universal Dependencies* (UD). **Relatório Técnico do ICMC**, n. 434. ICMC-USP. São Carlos-SP, 55p., 2021.

DURAN, M.S. Manual de Anotação de Relações de Dependência - Versão Revisada e Estendida: Orientações para anotação de relações de dependência sintática em Língua Portuguesa, seguindo as diretrizes da abordagem *Universal Dependencies* (UD). **Relatório Técnico do ICMC**, n. 440. ICMC-USP. São Carlos-SP. São Carlos-SP, 166p., 2022.

DURAN, M.S. et al. Manual de anotação como recurso de Processamento de Linguagem Natural: o modelo Universal Dependencies em língua portuguesa. **Domínios de Lingu@gem**, Vol. 16, N. 4, pp. 1608-1643, 2022.

DURAN, M.S. et al. The dawn of the Porttinari multigenre treebank: introducing its journalistic portion. In: SYMPOSIUM IN INFORMATION AND HUMAN LANGUAGE TECHNOLOGY, 14, 2023. Belo Horizonte. **Proceedings** [...]. Belo Horizonte, 2023, p. 115-124.

EISENSTEIN, J. What to do about bad language on the internet. In: NAACL-HLT, 2013. Atlanta, USA. **Proceedings** […]. Atlanta: ACL, p. 359–369. 2013.

FELDMAN, R. Techniques and applications for sentiment analysis. **Communications of the ACM**, v. 56, n. 4, p. 82–89, 2013. DOI: 10.1145/2436256.2436274.

FREITAS. C. **Linguística Computacional** (Linguística para o Ensino Superior, 13). 1ª edição. São Paulo: Parábola Editorial, 2022.

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

FOSTER, J. “cba to check the spelling”: investigating parser performance on discussion forum posts. In: NAACL-HLT, 2010, Los Angeles, USA. **Proceedings** […]. Los Angeles: ACL, p. 381–384, 2010.

JURAFSKY, D.; MARTIN, J. H. **Speech and Language Processing: an introduction to Natural Language Processing, Computational Linguistics and Speech Recognition**. 3rd. ed., 2024. Available at: <https://web.stanford.edu/~jurafsky/slp3/>. Access at: 15 dez. 2024.

KRUMM, J.; DAVIES, N.; NARAYANASWAMI, C. User-generated content. **IEEE Pervasive Computing**, v. 7, n. 4, p. 10–11, 2008.

LIU, P. et al. Pre-train, prompt, and predict: a systematic survey of prompting methods in Natural Language Processing. **ACM Comput. Surv**. Volume 55, Issue 9, p. 1-35, 2023.

LOPES, L. et al. PortiLexicon-UD: a Portuguese Lexical Resource according to Universal Dependencies Model. In: LANGUAGE RESOURCES AND EVALUATION CONFERENCE, 13, 2022, Marseille, France. **Proceedings** […]. Maiselle: ELRA, 2022, p. 6635-6643.

LOPES, L.; PARDO, T.A.S. Towards Portparser - a highly accurate parsing system for Brazilian Portuguese following the Universal Dependencies framework. In: INTERNATIONAL CONFERENCE ON COMPUTATIONAL PROCESSING OF PORTUGUESE, 16, 2023, Santiago de Compostela. **Proceedings** […]. Santiago de Compostela: ACL, 2023, p. 401-410.

MIRANDA, L.G.M.; PARDO, T.A.S. UDConcord: A Concordancer for Universal Dependencies Treebanks. In: PROCEEDINGS OF THE UNIVERSAL DEPENDENCIES BRAZILIAN FESTIVAL (UDFest-BR), 1, 2022, Fortazela, Brazil. **Proceedings** […], Fortaleza: SBC, 2022, p. 1-10.

NIVRE, J. et al. Universal Dependencies v2: an evergrowing multilingual treebank collection. INTERNATIONAL CONFERENCE ON LANGUAGE RESOURCES AND EVALUATION, 12, 2020, Marseille. **Proceedings** […]. Marseille: ELRA, 2020. p. 4034-4043.

NIVRE, J. *et al.* Universal dependencies v1: a multilingual treebank collection. In: INTERNATIONAL CONFERENCE ON LANGUAGE RESOURCES AND EVALUATION, 10, 2016, Portorož, Slovenia. **Proceedings** […]. Portorož: ELRA, 2016. p. 1659–1666.

QI, P. et al. Stanza: A Python natural language processing toolkit for many human languages. In: ANNUAL MEETING OF THE ASSOCIATION FOR COMPUTATIONAL LINGUISTICS: SYSTEM DEMONSTRATIONS, 58, 2020, Seattle, USA. **Proceedings** […]. Seattle, 2020, p. 1-8. Available at: <https://nlp.stanford.edu/pubs/qi2020stanza.pdf>. Access at: 15 dez. 2024.

RADEMAKER, A. et al. Universal Dependencies for Portuguese. In: INTERNATIONAL CONFERENCE ON DEPENDENCY LINGUISTICS (Depling), 4, 2017, Pisa, Italy. **Proceedings** […]. Pisa: Linköping University Electronic Press, 2017, p. 197-206.

SANGUINETTI, M. et al. Treebanking user-generated content: a UD-based overview of guidelines, corpora and unified recommendations. **Language Resources Evaluation**, 57, p. 493–544, 2023.

SCANDAROLLI, C. L. et al. Tipologia de fenômenos ortográficos e lexicais em cgu: o caso dos tweets do mercado financeiro. In: SIMPÓSIO BRASILEIRO DE TECNOLOGIA DA INFORMAÇÃO E DA LINGUAGEM HUMANA, 14, 2023, Belo Horizonte, Brazil. **Anais** [...]. Belo Horizonte: SBC, 2023, p. 240-248.

SEDDAH, D. et al. The French social media bank: a treebank of noisy user generated content. In: INTERNATIONAL CONFERENCE ON COMPUTATIONAL LINGUISTICS, 24, 2012, Mumbai, India. **Proceedings** […]. Mumbai: ACL, 2012 p. 2441–2458.

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, Porto Alegre, Brazil. **Proceedings** [...]. Porto Alegre: SBC, 2021. p. 434-445.

SILVA, E.H. et al. Etiquetagem morfossintática multigênero para o português do Brasil segundo o modelo “Universal Dependencies”. In: SYMPOSIUM IN

INFORMATION AND HUMAN LANGUAGE TECHNOLOGY, 14, 2023, Belo Horizonte, Brazil. **Proceedings** [...]. Belo Horizonte: SBC, 2023, p. 63-73.

SILVA, F.J.V. et al. Stock market tweets annotated with emotions. **Corpora**, v. 15, n. 3, p. 343–354, 2020. ISSN 1755-1676.

SINCLAIR, J. Corpus and text: basic principles. In: WYNNE, M. (Ed.). **Developing linguistic corpora: a guide to good practice**. AHDS Literature Language and Linguistics, 2004, cap. 1, p.1-16. Available at: https://users.ox.ac.uk/~martinw/dlc/chapter1.htm. Access at: 02 jan. 2025.

SOUZA, E. et al. Automatic annotation of enhanced Universal Dependencies for Brazilian Portuguese. In: SYMPOSIUM IN INFORMATION AND HUMAN LANGUAGE TECHNOLOGY, 15, 2024, Belém, Brazil. **Proceedings** [...]. Belém: SBC, 2024, p. 217-226.

STRAKA, M. UDPipe 2.0 prototype at CoNLL 2018 UD shared task. In: ZEMAN, D.; HAJIČ, J. (Ed.). In: CONLL 2018 SHARED TASK: MULTILINGUAL PARSING FROM RAW TEXT TO UNIVERSAL DEPENDENCIES. Brussels, Belgium. **Proceedings** […]. Brussels: ACL, 2018. p. 197–207.

SCHWARTZ, H. A. et al. Personality, gender, and age in the language of social media: the open-vocabulary approach. **PLoS ONE**, v. 8, n. 9, p. e73791, 2013. DOI: 10.1371/journal.pone.0073791.

Apêndice 1 – *Script* em Phyton

```
{
 "cells": [
  {
   "cell_type": "code",
   "execution_count": null,
   "metadata": {},
   "outputs": [
    {
     "name": "stdout",
     "output_type": "stream",
     "text": [
      "      DEPREL             Padrão        Head Dependente  Frequência  \\\n",
      "0  parataxis   ADJ (root) + ADJ  ADJ (root)        ADJ           2   \n",
      "1  parataxis   ADJ (root) + ADV  ADJ (root)        ADV           1   \n",
      "2  parataxis  ADJ (root) + NOUN  ADJ (root)       NOUN          20   \n",
      "9  parataxis         ADJ + NOUN         ADJ       NOUN           1   \n",
      "3  parataxis   ADJ (root) + NUM  ADJ (root)        NUM           3   \n",
      "\n",
      "                    ID_Exemplo  \n",
      "0  dante_01_4529347896672337931  \n",
      "1  dante_01_4492553281281925121  \n",
      "2  dante_01_4459281067053219841  \n",
      "9  dante_01_4644977696136314891  \n",
      "3  dante_01_4615213995686912001  \n"
     ]
    }
   ],
   "source": [
    "import pandas as pd\n",
    "from conllu import parse_incr\n",
    "\n",
    "# Função para buscar o token pelo ID correto\n",
    "def get_token_by_id(sentence, token_id):\n",
    "    for token in sentence:\n",
    "        if token[\"id\"] == token_id:\n",
    "            return token\n",
    "    return None\n",
    "\n",
    "# Função para processar um arquivo .conllu\n",
    "def analyze_parataxis_patterns(conllu_file_path):\n",
    "    data = []  # Lista para armazenar os dados processados\n",
    "\n",
    "    # Abre e processa o arquivo .conllu\n",
    "    with open(conllu_file_path, \"r\", encoding=\"utf-8\") as conllu_file:\n",
    "        sentences = parse_incr(conllu_file)\n",
    "\n",
    "        for sentence in sentences:\n",
    "            sent_id = sentence.metadata.get(\"sent_id\", None)  # Obtém o sent_id\n",
    "            for token in sentence:\n",
    "                if \"deprel\" in token and token[\"deprel\"].startswith(\"parataxis\"):\n",
    "                    # Obtém a relação e o PoS do token parataxis\n",
    "                    deprel = token[\"deprel\"]\n",
    "                    dep_pos = token[\"upostag\"]\n",
    "\n",
    "                    # Obtém o head correto do token\n",
    "                    head_id = token[\"head\"]\n",
    "                    if head_id is not None and head_id > 0:\n",
    "                        head_token = get_token_by_id(sentence, head_id)  # Busca token pelo ID real\n",
    "                        if head_token:\n",
    "                            head_pos = head_token[\"upostag\"]\n",
    "\n",
    "                            # Adiciona \"(root)\" se o head for root\n",
    "                            if head_token[\"deprel\"] == \"root\":\n",
    "                                head_representation = f\"{head_pos} (root)\"\n",
    "                            else:\n",
    "                                head_representation = head_pos\n",
    "\n",
    "                            # Cria o padrão PoS + PoS\n",
    "                            pattern = f\"{head_representation} + {dep_pos}\"\n",
    "\n",
    "                            # Adiciona os dados na lista\n",
    "                            data.append({\n",
    "                                \"DEPREL\": deprel,\n",
    "                                \"Padrão\": pattern,\n",
```

```
   "                            \"Head\": head_representation,\n",
   "                            \"Dependente\": dep_pos,\n",
   "                            \"Frequência\": 1,\n",
   "                            \"Sent_ID\": sent_id\n",
   "                        })\n",
   "\n",
   "    # Converte os dados para um DataFrame\n",
   "    df = pd.DataFrame(data)\n",
   "\n",
   "    # Agrupa por padrões únicos e conta frequências\n",
   "    frequency = df.groupby([\"DEPREL\", \"Padrão\", \"Head\", \"Dependente\"]).agg(\n",
   "        Frequência=(\"Frequência\", \"sum\"),\n",
   "        ID_Exemplo=(\"Sent_ID\", \"first\")\n",
   "    ).reset_index()\n",
   "\n",
   "    # Adiciona uma coluna auxiliar para ordenação lógica\n",
   "    frequency[\"Ordenação\"] = frequency[\"Padrão\"].str.replace(r\" \\(root\\)\", \"\", regex=True)\n",
   "\n",
   "    # Ordena por DEPREL e pela coluna auxiliar\n",
   "    result_df = frequency.sort_values(by=[\"DEPREL\", \"Ordenação\", \"Padrão\"])\n",
   "\n",
   "    # Remove a coluna auxiliar antes de exportar\n",
   "    result_df = result_df.drop(columns=[\"Ordenação\"])\n",
   "\n",
   "    # Exporta para Excel\n",
   "    result_df.to_excel(\"parataxis.xlsx\", index=False)\n",
   "\n",
   "    return result_df\n",
   "\n",
   "# Caminho para o arquivo .conllu\n",
   "conllu_file_path = \"DANTEStocks-V2-full.conllu\"\n",
   "\n",
   "# Processa o arquivo e gera o DataFrame\n",
   "result_df = analyze_parataxis_patterns(conllu_file_path)\n",
   "\n",
   "# Exibe o início do DataFrame\n",
   "print(result_df.head())\n"
  ]
 }
],
"metadata": {
 "kernelspec": {
  "display_name": "base",
  "language": "python",
  "name": "python3"
 },
 "language_info": {
  "codemirror_mode": {
   "name": "ipython",
   "version": 3
  },
  "file_extension": ".py",
  "mimetype": "text/x-python",
  "name": "python",
  "nbconvert_exporter": "python",
  "pygments_lexer": "ipython3",
  "version": "3.9.12"
 }
},
"nbformat": 4,
"nbformat_minor": 2
}
```

Apêndice 2 – Os tipos de parataxis do *corpus* organizadas em função do *head.*

DEPREL	Head	Dependente	Qt.	ID_Exemplo
parataxis	ADJ (root)	VERB	41	dante_01_4489016948026327051
parataxis	ADJ (root)	NOUN	20	dante_01_4459281067053219841
parataxis	ADJ (root)	PROPN	6	dante_01_4416462602617487361
parataxis	ADJ (root)	SYM	4	dante_01_4470995985875722241
parataxis	ADJ	VERB	4	dante_01_4613240875375861771
parataxis	ADJ (root)	NUM	3	dante_01_4615213995686912001
parataxis	ADJ	PROPN	3	dante_01_4464460593726955531
parataxis	ADJ (root)	ADJ	2	dante_01_4529347896672337931
parataxis	ADJ (root)	X	2	dante_01_4419525563703746561
parataxis	ADJ (root)	ADV	1	dante_01_4492553281281925121
parataxis	ADJ	NOUN	1	dante_01_4644977696136314891
parataxis	ADJ (root)	PRON	1	dante_01_4670571677708492801
parataxis	ADJ	X	1	dante_01_4419668972284313601
parataxis	ADV (root)	VERB	19	dante_01_4520933401931571211
parataxis	ADV (root)	NOUN	6	dante_01_4525295801220546561
parataxis	ADV	VERB	4	dante_01_4600348602254295041
parataxis	ADV (root)	PROPN	3	dante_01_4666064584872099841
parataxis	ADV (root)	X	3	dante_01_4488252683090657281
parataxis	ADV (root)	ADJ	1	dante_01_4535162303230976001
parataxis	ADV	ADJ	1	dante_01_4463415821834649601
parataxis	ADV (root)	ADV	1	dante_01_4600348602254295041
parataxis	ADV (root)	SYM	1	dante_01_4463406777908633601
parataxis	ADV	X	1	dante_01_4600348602254295041
parataxis	AUX (root)	VERB	7	dante_01_4459821628743024641
parataxis	AUX (root)	NOUN	2	dante_01_4455577960720793601
parataxis	AUX (root)	PROPN	2	dante_01_4692092393079603201
parataxis	AUX (root)	X	2	dante_01_4558220069830860801
parataxis	AUX (root)	ADJ	1	dante_01_4558220069830860801
parataxis	INTJ	ADJ	1	dante_01_4622881895046266891
parataxis	INTJ (root)	NOUN	1	dante_01_4698885653177425941
parataxis	INTJ (root)	NUM	1	dante_01_4433328934193766401
parataxis	INTJ (root)	VERB	1	dante_01_4433328934193766401
parataxis	NOUN (root)	NOUN	350	dante_01_4567960228076216321
parataxis	NOUN (root)	VERB	303	dante_01_4441935548278579211
parataxis	NOUN (root)	PROPN	110	dante_01_4698095602999255071
parataxis	NOUN	NOUN	84	dante_01_4669395172249067521
parataxis	NOUN (root)	ADJ	65	dante_01_4455277824756981761
parataxis	NOUN (root)	X	62	dante_01_4441935548278579211
parataxis	NOUN (root)	NUM	31	dante_01_4430650762041958401
parataxis	NOUN	PROPN	27	dante_01_4623287867179048961
parataxis	NOUN (root)	SYM	9	dante_01_4449464744856780811

parataxis	NOUN	VERB	9	dante_01_468921386799747072l
parataxis	NOUN	X	9	dante_01_4463417220089978881
parataxis	NOUN (root)	PRON	8	dante_01_4687606721682022401
parataxis	NOUN	NUM	4	dante_01_469214689206206464l
parataxis	NOUN	SYM	4	dante_01_448159009464664064l
parataxis	NOUN	ADJ	2	dante_01_471738020191019010l
parataxis	NOUN (root)	ADV	2	dante_01_4513408434595512321
parataxis	NOUN	ADV	2	dante_01_459398634690600960l
parataxis	NOUN	PRON	2	dante_01_441299691763159040l
parataxis	NOUN (root)	AUX	1	dante_01_445973479092146176l
parataxis	NUM (root)	VERB	13	dante_01_441979084714049536l
parataxis	NUM	SYM	5	dante_01_453157209351143425l
parataxis	NUM (root)	NOUN	3	dante_01_454332774741463041l
parataxis	NUM (root)	X	2	dante_01_454332774741463041l
parataxis	NUM	X	2	dante_01_453157209351143425l
parataxis	NUM (root)	ADV	1	dante_01_4615050656738222081
parataxis	NUM	NOUN	1	dante_01_441673990776823808l
parataxis	NUM	PROPN	1	dante_01_456875129188798464l
parataxis	NUM (root)	SYM	1	dante_01_453151387845951488l
parataxis	PRON (root)	VERB	14	dante_01_443798359951044609l
parataxis	PRON (root)	PROPN	5	dante_01_459718735914598400l
parataxis	PRON (root)	NOUN	3	dante_01_446347773408190464l
parataxis	PRON	VERB	2	dante_01_4687606721682022401
parataxis	PRON (root)	ADJ	1	dante_01_449506024496824320l
parataxis	PRON	PROPN	1	dante_01_448301183678312448l
parataxis	PRON (root)	SYM	1	dante_01_447100219822718976l
parataxis	PRON (root)	X	1	dante_01_453315519886925824l
parataxis	PRON	X	1	dante_01_442042173597687808l
parataxis	PROPN	PROPN	65	dante_01_4462589388128337921
parataxis	PROPN (root)	VERB	59	dante_01_4492505108815749121
parataxis	PROPN (root)	NOUN	49	dante_01_449250510881574912l
parataxis	PROPN	NOUN	46	dante_01_445910769411293184l
parataxis	PROPN (root)	X	16	dante_01_447040378752225280l
parataxis	PROPN (root)	PROPN	15	dante_01_456460101441044480l
parataxis	PROPN	X	9	dante_01_445910769411293184l
parataxis	PROPN (root)	ADJ	5	dante_01_448514697797840896l
parataxis	PROPN (root)	ADV	5	dante_01_451716260364296192l
parataxis	PROPN	ADV	4	dante_01_456498019605565440l
parataxis	PROPN	VERB	4	dante_01_459013948516614144l
parataxis	PROPN (root)	NUM	3	dante_01_445926853204000768l
parataxis	PROPN (root)	SYM	3	dante_01_452113111693402112l
parataxis	PROPN (root)	AUX	2	dante_01_448841050694909952l
parataxis	PROPN (root)	PRON	2	dante_01_441629457304940544l
parataxis	PROPN	PRON	2	dante_01_445910769411293184l
parataxis	PROPN	AUX	1	dante_01_441244662418796544l
parataxis	PROPN	NUM	1	dante_01_459023285372866560l

parataxis	SYM	SYM	40	dante_01_4467670843712061441
parataxis	SYM (root)	VERB	11	dante_01_4430230023396556811
parataxis	SYM	PROPN	7	dante_01_4531573814560317441
parataxis	SYM (root)	NOUN	5	dante_01_4467803318702571531
parataxis	SYM (root)	PROPN	5	dante_01_4491787531898183691
parataxis	SYM (root)	X	4	dante_01_4467803318702571531
parataxis	SYM	NOUN	3	dante_01_4596884916734730241
parataxis	SYM (root)	ADJ	2	dante_01_4493495237612666881
parataxis	SYM	X	2	dante_01_4531573814560317441
parataxis	SYM (root)	INTJ	1	dante_01_4430230023396556811
parataxis	SYM (root)	NUM	1	dante_01_4467803318702571531
parataxis	VERB (root)	VERB	860	dante_01_4480778686447001601
parataxis	VERB (root)	NOUN	621	dante_01_4440984868298670081
parataxis	VERB (root)	PROPN	217	dante_01_4514029019033927691
parataxis	VERB (root)	X	200	dante_01_4480778686447001601
parataxis	VERB (root)	ADJ	66	dante_01_4596988468270243841
parataxis	VERB	X	35	dante_01_4456503877907496961
parataxis	VERB	VERB	30	dante_01_4547261197788160001
parataxis	VERB (root)	ADV	26	dante_01_4614836854273146891
parataxis	VERB (root)	NUM	23	dante_01_4462875776128327681
parataxis	VERB	PROPN	23	dante_01_4441935548278579211
parataxis	VERB (root)	SYM	19	dante_01_4531573814560317441
parataxis	VERB	NOUN	16	dante_01_4415539628570255371
parataxis	VERB (root)	PRON	15	dante_01_4508019785670410241
parataxis	VERB	ADJ	6	dante_01_4717380201910190101
parataxis	VERB	NUM	4	dante_01_4460196038234767361
parataxis	VERB	ADV	3	dante_01_4527267507376087041
parataxis	VERB	SYM	3	dante_01_4476193175770439681
parataxis	VERB (root)	AUX	2	dante_01_4419893776367083521
parataxis	VERB	PRON	2	dante_01_4520815265668464641
parataxis	VERB (root)	CCONJ	1	dante_01_4474733405804953601
parataxis	VERB (root)	INTJ	1	dante_01_4492570486065111041
parataxis	X (root)	VERB	3	dante_01_4577356889902284811
parataxis	X (root)	NOUN	3	dante_01_4470021663988162581
parataxis	X	NOUN	1	dante_01_4607081934006599681
parataxis	X	PROPN	1	dante_01_4666115079888035841

DEPREL	Head	Dependente	Qt.	ID_Exemplo
parataxis:cashtag	VERB (root)	X	87	dante_01_4667744859461017601
parataxis:cashtag	VERB (root)	PROPN	11	dante_01_4724732926838865921
parataxis:cashtag	NOUN (root)	X	9	dante_01_4597370450372321291
parataxis:cashtag	NOUN (root)	PROPN	2	dante_01_4437621165675110401
parataxis:cashtag	PROPN (root)	PROPN	1	dante_01_4420561193229598731
parataxis:cashtag	VERB	X	1	dante_01_4446398281606963201

DEPREL	Head	Dependente	Qt.	ID_Exemplo
parataxis:hashtag	VERB (root)	X	698	dante_01_460775523166461953l
parataxis:hashtag	NOUN (root)	X	138	dante_01_469809560299925507l
parataxis:hashtag	VERB (root)	PROPN	74	dante_01_456479295968272384l
parataxis:hashtag	NOUN (root)	PROPN	20	dante_01_468800572238610432l
parataxis:hashtag	PROPN (root)	X	16	dante_01_456168774593708032l
parataxis:hashtag	ADJ (root)	X	13	dante_01_448901694802632705l
parataxis:hashtag	ADV (root)	PROPN	10	dante_01_446333972562591745l
parataxis:hashtag	PRON (root)	X	6	dante_01_443798359951044609l
parataxis:hashtag	PRON (root)	PROPN	5	dante_01_446347773408190464l
parataxis:hashtag	ADJ (root)	PROPN	4	dante_01_447086599923597312l
parataxis:hashtag	SYM (root)	X	3	dante_01_469549772081475584l
parataxis:hashtag	VERB	PROPN	3	dante_01_446663297103060992l
parataxis:hashtag	ADV (root)	X	2	dante_01_443826529223847936l
parataxis:hashtag	NUM (root)	X	2	dante_01_445973204335882241l
parataxis:hashtag	VERB (root)	NOUN	2	dante_01_453205662068391936l
parataxis:hashtag	VERB	X	2	dante_01_447193082044633089l
parataxis:hashtag	X (root)	X	2	dante_01_457735688990228481l
parataxis:hashtag	ADJ	X	1	dante_01_446446059372695553l
parataxis:hashtag	AUX (root)	PROPN	1	dante_01_445982162874302464l
parataxis:hashtag	NOUN	X	1	dante_01_444448840863997952l
parataxis:hashtag	NUM (root)	PROPN	1	dante_01_461584996340613121l
parataxis:hashtag	PROPN (root)	NOUN	1	dante_01_443063880676159488l
parataxis:hashtag	PROPN (root)	PROPN	1	dante_01_448470560268636160l
parataxis:hashtag	SYM (root)	PROPN	1	dante_01_469834686131605504l

DEPREL	Head	Dependente	Qt.	ID_Exemplo
parataxis:url	VERB (root)	SYM	778	dante_01_4440984868298670081
parataxis:url	NOUN (root)	SYM	688	dante_01_469809560299925507l
parataxis:url	PROPN (root)	SYM	37	dante_01_449250510881574912l
parataxis:url	PROPN	SYM	32	dante_01_443001323957391360l
parataxis:url	ADJ (root)	SYM	22	dante_01_443251532755394560l
parataxis:url	SYM (root)	SYM	11	dante_01_446780331870257153l
parataxis:url	VERB	SYM	9	dante_01_459519562388865026l
parataxis:url	ADV (root)	SYM	7	dante_01_452529580122054656l
parataxis:url	NOUN	SYM	3	dante_01_462328786717904896l
parataxis:url	X (root)	SYM	3	dante_01_457735688990228481l
parataxis:url	ADV	SYM	2	dante_01_442373783211761664l
parataxis:url	AUX (root)	SYM	2	dante_01_447066056914653184l
parataxis:url	NUM (root)	SYM	2	dante_01_454332774741463041l
parataxis:url	PRON (root)	SYM	2	dante_01_444539036682575872l
parataxis:url	X	SYM	2	dante_01_460708193400659968l
parataxis:url	ADJ	SYM	1	dante_01_449506024496824320l
parataxis:url	PRON	SYM	1	dante_01_448858508675796992l
parataxis:url	PRON (root)	X	1	dante_01_450812303794241537l

parataxis:url	VERB (root)	PROPN	1	dante_01_444484995537989632l
parataxis:url	VERB (root)	X	1	dante_01_469133799839002625l

DEPREL	Head	Dependente	Qt.	ID_Exemplo
parataxis:strunc	VERB (root)	NOUN	46	dante_01_449551763301883904l
parataxis:strunc	VERB (root)	VERB	42	dante_01_453534358314315776l
parataxis:strunc	NOUN (root)	VERB	10	dante_01_446787453655871488l
parataxis:strunc	VERB (root)	PROPN	6	dante_01_443205544418234368l
parataxis:strunc	NOUN (root)	NOUN	5	dante_01_445626541658365952l
parataxis:strunc	VERB (root)	ADV	4	dante_01_459058981432881152l
parataxis:strunc	NOUN (root)	PROPN	3	dante_01_454596496378171393l
parataxis:strunc	PROPN (root)	VERB	2	dante_01_459491215151333376l
parataxis:strunc	VERB (root)	AUX	2	dante_01_457996497792274432l
parataxis:strunc	VERB (root)	X	2	dante_01_443009081989541888l
parataxis:strunc	ADJ (root)	NOUN	1	dante_01_441218985657659392l
parataxis:strunc	ADJ (root)	PRON	1	dante_01_446276557812678656l
parataxis:strunc	NOUN	PROPN	1	dante_01_443138797375848448l
parataxis:strunc	NOUN (root)	X	1	dante_01_455484807212838913l
parataxis:strunc	VERB	ADJ	1	dante_01_454616552483323904l
parataxis:strunc	VERB (root)	ADP	1	dante_01_452099943684001792l
parataxis:strunc	VERB	ADV	1	dante_01_447110686561161216l
parataxis:strunc	VERB	AUX	1	dante_01_443479045179600896l
parataxis:strunc	VERB	NOUN	1	dante_01_453552256172847104l
parataxis:strunc	VERB	X	1	dante_01_456089794812973056l
parataxis:strunc	ADJ	AUX	1	dante_01_449506024496824320l

DEPREL	Head	Dependente	Qt.	ID_Exemplo
parataxis:wtrunc	VERB (root)	VERB	11	dante_01_461203818915434496l
parataxis:wtrunc	NOUN (root)	SYM	6	dante_01_452114970592165888l
parataxis:wtrunc	VERB (root)	NOUN	4	dante_01_447139227361234944l
parataxis:wtrunc	VERB (root)	PROPN	3	dante_01_464369509340430337l
parataxis:wtrunc	VERB	VERB	3	dante_01_443853161887838208l
parataxis:wtrunc	VERB (root)	ADJ	2	dante_01_468819730334093313l
parataxis:wtrunc	VERB (root)	ADV	2	dante_01_443040792588738562l
parataxis:wtrunc	VERB (root)	X	2	dante_01_454665164701192192l
parataxis:wtrunc	NOUN	VERB	1	dante_01_446769021795315712l
parataxis:wtrunc	NUM (root)	SYM	1	dante_01_449658348586414080l
parataxis:wtrunc	VERB (root)	AUX	1	dante_01_459046521259184129l
parataxis:wtrunc	VERB	NOUN	1	dante_01_444579970325557248l
parataxis:wtrunc	VERB (root)	SYM	1	dante_01_455822507439030272l