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

GABRIELA PINHEIRO DE OLIVEIRA

# ANOTAÇÃO E CARACTERIZAÇÃO DE FENÔMENOS LÉXICO-ORTOGRÁFICOS EM TWEETS DO MERCADO FINANCEIRO

SÃO CARLOS - SP
2025

**Agradecimentos**

Gostaria, primeiramente, de expressar minha profunda gratidão à minha professora e orientadora, Ariani, pela paciência, dedicação e constante apoio ao longo de todas as etapas do desenvolvimento deste trabalho.

Agradeço, também, às amizades que sempre estiveram ao meu lado durante a realização do curso. Partindo daquelas que, desde a infância, ainda hoje tenho o prazer de compartilhar a vida, até as incríveis pessoas que me foi concebida a honra de conhecer em São Carlos e levarei comigo para sempre.

Por fim, deixo meus sinceros agradecimentos à minha família, que acreditou no meu desenvolvimento e crescimento pessoal. Sem seu incentivo e suporte, não teria sido possível realizar o sonho de me tornar a primeira pessoa de toda a família a ingressar e graduar em uma universidade pública.

**RESUMO**

Com o crescimento das mídias digitais e redes sociais, a importância do "conteúdo gerado por usuário" (CGU) aumentou consideravelmente, demandando ferramentas e aplicações de Processamento de Língua Natural (PLN) capazes de lidar com a linguagem primordialmente não canônica dos diferentes gêneros CGU. Para tanto, os *corpora* anotados são recursos essenciais, assim como a descrição e análise de suas características linguísticas. Entre eles, destacam-se os *corpora* compostos por *tweets/posts*, dada a relevância da plataforma Twitter/X para vários segmentos da sociedade. Neste Trabalho de Conclusão de Curso (TCC), deu-se prosseguimento à anotação manual de fenômenos léxico-ortográficos no *corpus* de *tweets* do mercado financeiro denominado DANTEStocks. Tal anotação, que atingiu cerdca de 75% do total de *tweets,* permitiu uma caracterização preliminar do *corpus* em função de uma tipologia hierárquica de classes, tipos e subtipos, que busca capturar fenômenos criativos e de variação da norma padrão.

**Palavras-chave:** *tweets*, anotação de *corpus*, fenômenos léxico-ortográficos

***ABSTRACT***

With the rise of digital media and social networks, the importance of “user-generated contente” (UGC) has increased significantly, requiring Natural Language Processing (NLP) tools and applications capable of handling the predominantly non-canonical language found in various UGC genres. In this context, annotated corpora are essential resources, as are the description and analysis of their linguistic features. Among these, corpora composed of tweets (now, posts) stand out, given the relevance of the Twitter/X platform across multiple sectors of society. This undergraduate thesis continued the manual annotation of lexico-orthographic phenomena in the financial market tweet corpus known as DANTEStocks. This annotation, which covered about 75% of the total tweets, enabled a preliminary characterization of the corpus based on a hierarchical typology of classes, types, and subtypes, designed to capture creative phenomena and variation from the standard norm.

***Keywords****: tweets, corpus annotation, lexical-orthographic phenomena*

## LISTA DE FIGURAS

## LISTA DE QUADROS

# Sumário

# 1. Introdução

No Processamento de Língua Natural (PLN), os *corpora* anotados, isto é, coleções de textos autênticos com informações linguísticas explícitas (SINCLAIR, 2005; MCENERY *et al*., 2006), são recursos essenciais, sobretudo os *gold standard* (ou de referência). Diz-se isso pela consistência das anotações, que podem ser manuais ou semiautomáticas (DURAN; PARDO, 2024).

Os *corpora* anotados são essenciais porque servem tanto como fonte de conhecimento explícito para a formulação de regras linguísticas, quanto como base para o treinamento supervisionado de modelos de Aprendizado de Máquina (AM), especialmente as atuais redes neurais profundas (*Deep Neural Network*) e representações vetoriais das palavras (*word embeddings*).

A popularização das mídias sociais e a crescente importância de seu conteúdo para a sociedade, por um lado, e a necessidade de desenvolver ferramentas de PLN capazes de lidar com a linguagem não canônica pela qual esse conteúdo é veiculado, por outro lado, motivaram a construção de *corpora* anotados de "conteúdo gerado por usuário" (CGU) em várias línguas. A maioria deles é composta por *tweets* (agora *posts*) (SANGUINETTI *et al.*, 2023).

As anotações mais frequentes nos *corpora* de *tweets* são a morfossintática (ou *Part-of-Speech* ou PoS) e a sintática, que têm subsidiado a investigação, respectivamente, das tarefas de *tagging*[1] e *parsing*[2] para CGU, sobretudo em inglês (SANGUINETTI *et al.*, 2023). Enriquecidos com essas informações linguísticas explícitas, os *corpora* passam a ser classificados como *tweebanks.*

As anotações morfossintática e sintática em *tweebanks* geralmente seguem o modelo gramatical *Universal Dependencies* (UD) (NIVRE et al., 2020), amplamente adotado no PLN. Para aplicá-lo à linguagem CGU, foi necessário definir diretrizes específicas com base em estudos linguísticos prévios sobre as características lexicais e estruturais dos tweets, já que as diretrizes originais do UD foram elaboradas para textos formais (p.ex.: SANGUINETTI *et al.*, 2023).

[1] *Tagging* é o processo de atribuir rótulos a cada *token* de uma sentença que indicam categorias gramaticais ("partes do discurso") (JURAFSKY; MARTIN, 2025).
[2] *Parsing* é a tarefa de reconhecer a estrutura sintática de uma sentença, isto é, determinar sua estrutura sintagmática ou as relações de dependência (JURAFSKY; MARTIN, 2025).

Para o português, tem-se o *corpus* DANTEStocks, composto por 4.048 *tweets* do domínio do mercado financeiro (DI-FELIPPO; ROMAN, 2025). Esse *tweebank* possui algumas camadas de anotação *gold standard,* a saber: (i) morfossintática e sintática segundo o modelo UD; (ii) emoções segundo o modelo *Wheel´s of Emotions* de Plutchik e Kellerman (1986) e (iii) entidades nomeadas segundo as categorias genéricas do Segundo HAREM (MOTA; SANTOS, 2008).

Além dessas anotações acerca do DANTEStocks, outras têm sido desenvolvidas com o suporte do trabalho de Scandarolli *et al*. (2023), que mapeou as particularidades lexicais e ortográficas do *corpus*, que, por não ter passado por nenhum tipo de normalização lexical ou segmentação em unidades estruturais menores, apresenta uma série de fenômenos lexicais (e estruturais) desafiadores tanto para a anotação manual quanto para o processamento automático.

Com base na análise (e anotação) manual de 1.069 *tweets* do total de 4.048, os autores propuseram uma taxonomia de fenômenos léxico-ortográficos, que permitiu, por exemplo, definir novas diretrizes de anotação de PoS seguindo o modelo UD e, por conseguinte, o desenvolvimento do primeiro *tagger*-UD para *tweets* em português do mercado de ações (DI-FELIPPO; ROMAN, 2025).

Neste Trabalho de Conclusão de Curso (TCC), deu-se prosseguimento à anotação manual dos fenômenos léxico-ortográficos em mais 2.000 *tweets* distintos do DANTEStocks com base na taxonomia de Scandarolli *et al*. Somando os novos 2.000 tweets anotados à anotação anterior (1.069), obteve-se um total de 75.82% de anotação do corpus (3.069 de 4.048) e apresenta-se aqui uma caracterização linguística do *corpus* DANTEStocks em função dos fenômenos mapeados pela taxonomia.

Acredita-se que, uma vez concluída a anotação do corpus completo, o conhecimento explícito dos fenômenos léxico-ortográficos poderá contribuir para a redução de erros causados por formas linguísticas atípicas em tarefas supervisionadas, como *tagging*, *parsing* e reconhecimento de entidades nomeadas. Isso pode permitir que esse tipo de anotação auxilie os modelos na compreensão da diversidade textual, evitando que variações sejam automaticamente interpretadas como erros. Além disso, esse tipo de anotação

pode auxiliar no diagnóstico de falhas dos métodos, ao tornar possível identificar quais tipos de fenômenos afetam negativamente o desempenho.

Para apresentar a pesquisa, este relatório está organizado em 6 Seções. Na Seção 2, apresenta-se uma breve revisão da literatura sobre o subgênero CGU “tweet” e sua caracterização linguística. Na Seção 3, descreve-se o *corpus* DANTEStocks, com especial atenção às decisões de pré-processamento, sobretudo as diretrizes de tokenização UD, que têm relação direta com os fenômenos léxico-ortográficos presentes no recurso. Na seção 4, apresenta-se o trabalho de anotação de *corpus* realizado neste TCC efetivamente, dando destaque à taxonomia de fenômenos e a metodologia empregada na anotação. Na seção 5, discorre-se sobre a caracterização linguística do *corpus* em função da distribuição estatística das categorias, tipos e subtipos dos fenômenos anotados. Na Seção 6, por fim, são apresentadas as considerações finais deste trabalho, enfatizando contribuições, limitações e trabalhos futuros.

## 2. Revisão da literatura

### 2.1. Conteúdo gerado por usuário (CGU)

O avanço da comunicação digital, especialmente a partir do surgimento e crescimento das redes sociais, *blogs*, fóruns e outras mídias interativas, fomentou uma transformação profunda na forma como os indivíduos produzem e compartilham conteúdo. Nesse cenário, cunhou-se o termo *user-generated content* (UGC), isto é, conteúdo gerado por usuários (CGU), para se referir a todo tipo de conteúdo criado e publicado por usuários da *web* na forma de texto, vídeo, imagem ou áudio (KRUMM *et al.*, 2008).

No que tange ao conteúdo textual, o termo CGU recobre um *continuum* de subgêneros segundo Sanguinetti *et. al*. (2023), o qual, no geral, caracteriza-se pelo uso da linguagem do cotidiano nas mídias digitais. A variação dos subgêneros textuais de CGU depende, em grande medida, das convenções e limitações impostas pelo meio ou plataforma em que são veiculados. Embora possuam fenômenos característicos gerais e reconhecidos, conforme apontado por diversos autores (p.ex.: FOSTER, 2010; EISENSTEIN, 2013; SANGUINETTI *et al.*, 2023), a natureza informal e a amplitude dos subgêneros tornam o seu processamento automático uma tarefa complexa.

A crescente importância do conteúdo veiculado pelos diferentes subgêneros CGU e a linguagem não canônica que o caracteriza motivaram a construção de *corpora* anotados de CGU em várias línguas. No período de 2011 a 2019, Sanguinetti *et al.* (2023) identificaram 30 *corpora* de CGU com anotação sintática de referência construídos para diversas línguas europeias, inglês americano, árabe, chinês, híndi e outras (Figura 1).

**Figura 1.** Lista de *corpora* compostos por GCU.

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

**Fonte:** Sanguinetti *et. al*. (2022).

A maioria desses *corpora* é composta, parcial ou totalmente, por *posts* extraídos do Twitter/X e segue o modelo gramatical UD (indicado na coluna “UD-based”). Além do alcance das opiniões veiculadas na plataforma, outras razões para a proeminência dos *tweebanks* foram a facilidade de obtenção dos dados via *Application Programming Interface* (API) e a política de uso dos dados para

fins acadêmicos adotada até muito recentemente pela plataforma. Até 2023, pesquisadores acadêmicos tinham acesso gratuito à API para coletar grandes volumes de dados, sendo possível acessar/compilar todos os tweets públicos desde 2006, sem limite de tempo.

### 2.2. Subgênero textual "tweet"

O Twitter, renomeado para X em 2023, é uma rede social e plataforma de *microblogging*, que permite a publicação de mensagens curtas[3] chamadas *tweets* (agora, *posts*). Desde seu lançamento em 2006, o Twitter se consolidou como um importante canal de comunicação em tempo real. Segundo o DataReportal[4] – plataforma *online* que oferece relatórios detalhados e *insights* sobre o cenário digital global –, o Twitter atingiu 586 milhões de usuários ativos em janeiro de 2025, ocupando a 7ª posição no *ranking* mundial das mídias sociais. No Brasil, há 16 milhões de usuários, sendo a 9ª mais popular[5].

Um *tweet* é uma unidade discursiva autônoma ou parte de uma sequência (*thread*), podendo conter texto, imagens, vídeos, *links* e outros elementos interativos (como *hashtags*[6], menções, *emojis,* etc.). Diante disso, diz-se que, enquanto (sub)gênero textual, ele é uma forma breve, multimodal e altamente contextualizada de comunicação digital. Embora inclua características de outros gêneros (como notícia, *blog*, SMS (*short message service*), conversa informal, bilhete, citação, etc.), o *tweet* possui características particulares que o distinguem dos demais gêneros digitais, sobretudo pela limitação de espaço, dinamismo e uso frequente de recursos interacionais e hipertextuais. Autores como Eisenstein (2013), Zappavigna (2012) e Cardoso (2019) destacam que o *tweet* é um gênero adaptado ao ritmo e às práticas da cultura digital, que se insere em um ecossistema comunicativo altamente interacional e efêmero.

A combinação entre esses gêneros menos formais e as características da plataforma favorece a predominância da informalidade no Twitter. Como destacam Freitas e Barth (2015), ainda que muitos *tweets* apresentem traços da

[3] A partir de 2017, o limite de caracteres para cada *tweet* passou de 140 para 280, podendo ser maior com assinaturas ou em *threads.*
[4] https://datareportal.com/reports/digital-2025-global-overview-report
[5] https://datareportal.com/reports/digital-2025-brazil
[6] *Hashtag* é uma etiqueta textual precedida pelo símbolo # (cerquilha), usada para marcar palavras-chave ou tópicos em plataformas de mídias sociais como o Twitter/X.

norma culta, o formato reduzido limita o uso de construções mais elaboradas, incentivando a inserção de hipertextos, *links* e construções reduzidas. Soma-se a isso o fato de que a comunicação ocorre em um ambiente em que os interlocutores compartilham um repertório digital comum, o que permite o uso de códigos, gírias e referências específicas sem necessidade de explicitação. Essa informalidade não apenas contribui para a agilidade na troca de informações, mas também reforça uma atmosfera discursiva marcada pela espontaneidade, autenticidade e proximidade entre os usuários.

Em suma, pode-se dizer que o *tweet* apresenta as seguintes características gerais: (i) extensão reduzida; (ii) estrutura multimodal; (iii) interatividade intensa, por meio de curtidas, retuítes, respostas e menções; (iv) marcas de oralidade e informalidade, como pontuação não convencional, vocabulário coloquial e *emojis*; (v) forte hipertextualidade, com uso de *hashtags* e URLs; (vi) organização seriada, por meio de *threads* que ampliam o conteúdo; (vii) temporalidade marcada, com foco em eventos do presente; (viii) ampla variação funcional, incluindo usos informativos, opinativos, promocionais, humorísticos, narrativos e militantes.

### 2.3. Modelo *Universal Dependencies* (UD)

Como mencionado, a maioria dos *tweebanks* possui anotação baseada no modelo gramatical UD (NIVRE *et al.,* 2020), que tem cada vez mais se torna uma referência padrão e popular para anotação de *corpus* devido à sua adaptabilidade a diferentes domínios e gêneros (SANGUINETTI *et. al*., 2023). Trata-se especificamente de um modelo ou esquema de anotação gramatical voltado à representação estruturada e comparável dos elementos morfossintáticos das línguas naturais.

Esse modelo define a descrição em dois níveis: morfológico e sintático.

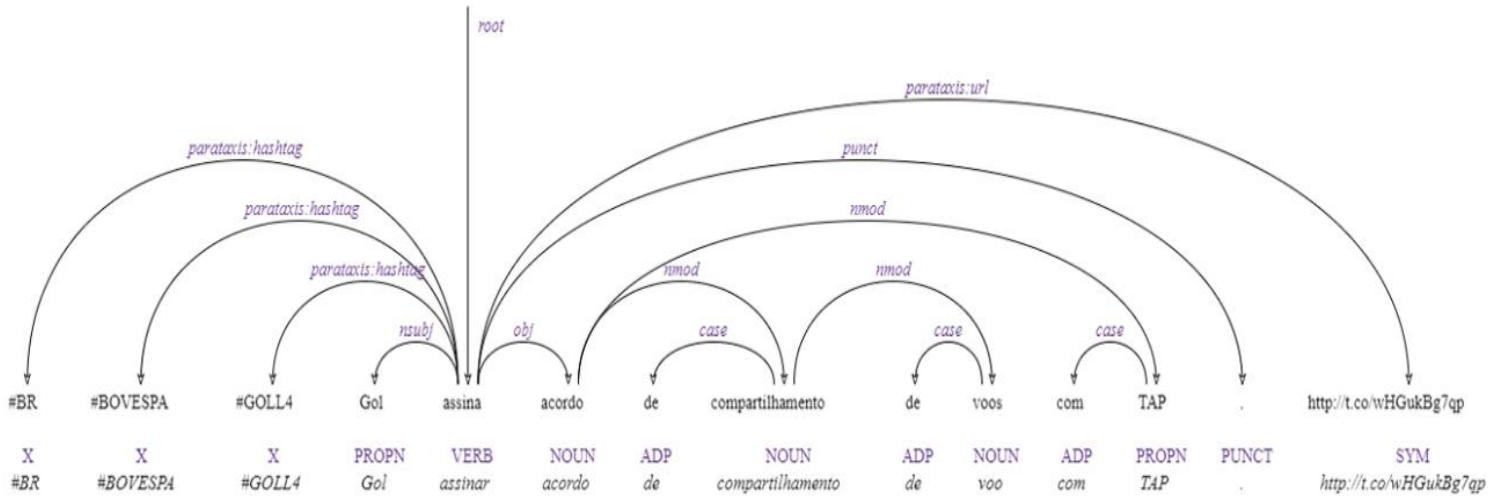
No nível morfológico, cada unidade textual (*token*) é anotada com um lema (ou forma canônica), uma etiqueta de categoria gramatical (ou *tag* PoS) e um conjunto de traços morfológicos (*features*) que codificam informações gramaticais adicionais, como tempo, número, gênero, aspecto, entre outros.

No nível sintático, prevê-se a descrição das relações de dependência (*deprel*) entre palavras/*tokens*. Uma dependência é estabelecida entre uma

palavra sintaticamente dependente e outra palavra da qual ela depende (*head* ou núcleo). A versão atual da UD (v2) engloba 17 *tags* PoS[7] e 37 *deprels*[8].

Na Figura 2, ilustra-se a anotação-UD de um *tweet* do *corpus* DANTEStocks no formato arbóreo. As etiquetas ou *tags* PoS são indicadas em caixa alta, como NOUN para “acordo”. Acima das formas flexionadas, estão os lemas, como “voo” para o plural “voos”. As relações de dependência sintática (*deprels*) são representadas pelas setas que partem do núcleo (*head*) e apontam para o dependente. A palavra “compartilhamento”, por exemplo, é dependente de “acordo” pela *deprel* ***nmod*** (modificador nominal), sendo essa relação mediada pela preposição “de”. A preposição, por sua vez, é dependente de “compartilhamento” e está ligada a ele pela *deprel* ***case***, que marca relações introduzidas por elementos funcionais como preposições. O verbo “assinou” constitui o *root* da estrutura, ou seja, a raiz sintática do *tweet*, servindo como núcleo a partir do qual se organizam as demais relações de dependência.

**Figura 2.** Exemplo de *tweet* com anotação-UD em formato de árvore.




**Fonte:** Barbosa (2024).

Como resultado da anotação-UD, tem-se um arquivo correspondente no formato CoNLL-U. Nele, o enunciado é representado em uma tabela de 10 colunas, em

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

que cada linha corresponde a um *token* do enunciado (no caso, *tweet*) anotado (ZEMAN *et. al*, 2017).

Na Figura 3, tem-se o arquivo CoNLL-U correspondente ao *tweet* da Figura 2. Da esquerda para a direita, a coluna 1 indica a ordem de ocorrência do *token* na sentença (ID) e a coluna 2 exibe o próprio *token* (FORM). As informações morfológicas são descritas nas colunas 3, 4, 5 e 6, sendo, respectivamente, lemas (LEMMA), etiquetas universais de *part-of-speech* (PoS) (classe de palavra) (UPOS), etiquetas PoS específicas da língua[9] (XPOS) e traços gramaticais (FEATS). As informações sintáticas são indicadas nas colunas 7 e 8. A coluna 7 indica o *head* (cabeça) da relação (HEAD), sendo esta, de fato, rotulada na coluna 8 (DEPREL). A coluna 9 fica destinada às chamadas relações *enhanced* (ou enriquecidas)[10] (DEPS), que basicamente buscam explicitar certas relações implícitas entre as palavras. Por fim, a coluna 10 (MISC) pode exibir informações adicionais sobre os *tokens*.

**Figura 3.** Exemplo de anotação de um tweet do DANTEStocks no formato CoNLL-U.

| ID | FORM | LEMMA | UPoS | XPoS | FEATS | HEAD | DEPREL | DEPS | MISC |
|---|---|---|---|---|---|---|---|---|---|
| 1 | #BR | #BR | X | _ | _ | 5 | parataxis:hashtag | _ | _ |
| 2 | #BOVESPA | #BOVESPA | X | _ | _ | 5 | parataxis:hashtag | _ | _ |
| 3 | #GOLL4 | #GOLL4 | X | _ | _ | 5 | parataxis:hashtag | _ | _ |
| 4 | Gol | Gol | PROPN | _ | _ | 5 | nsubj | _ | _ |
| 5 | assina | assinar | VERB | _ | Mood=Ind\|Number=Sing\|Person=3\|Tense=Pres\|VerbForm=Fin | 0 | root | _ | _ |
| 6 | acordo | acordo | NOUN | _ | Gender=Masc\|Number=Sing | 5 | obj | _ | _ |
| 7 | de | de | ADP | _ | _ | 8 | case | _ | _ |
| 8 | compartilhamento | compartilhamento | NOUN | _ | Gender=Masc\|Number=Sing | 6 | nmod | _ | _ |
| 9 | de | de | ADP | _ | _ | 10 | case | _ | _ |
| 10 | voos | voo | NOUN | _ | Gender=Masc\|Number=Plur | 8 | nmod | _ | _ |
| 11 | com | com | ADP | _ | _ | 12 | case | _ | _ |
| 12 | TAP | TAP | PROPN | _ | _ | 6 | nmod | _ | SpaceAfter=No |
| 13 | . | . | PUNCT | _ | _ | 5 | punct | _ | _ |
| 14 | http://t.co/wHGukBg7qp | http://t.co/wHGukBg7qp | SYM | _ | _ | 5 | parataxis:url | _ | SpaceAfter=No |

**Fonte:** A autora (2025).

## 2.4. *Corpora* de *tweets* em português

Para o processamento de CGU em português, tem-se alguns *corpora* compostos por *tweets*, como sistematizado no Quadro 1.

[9] Esses etiquetas não foram contempladas na anotação conduzida no POeTiSA.
[10] O Porttinati-base ainda não possui anotação dessas relações.

O *corpus* de Silva *et al.* (2011) reúne 76.358 *tweets* que mencionam os candidatos à eleição presidencial de 2010 (isto é, Dilma Rousseff e José Serra), configurando um recurso de domínio político. O *Corpus* 7x1 (MORAES *et al.*, 2015) concentra 2.728 *tweets* postados durante as semifinais da Copa do Mundo da FIFA ocorrida em 2014 no Brasil. Outro recurso é o construído por Corrêa Junior *et al.* (2017) para subsidiar especificamente a investigação de métodos automáticos de classificação de polaridade e sentimento. Esse *corpus* reúne mais de 980 mil *tweets*, sendo que 554.623 possuem *emojis* positivos e 425.444 possuem *emojis* negativos. O TweetSentBR (BRUM; NUNES, 2018), por sua vez, agrupa 15.000 *tweets* relacionados a programas de televisão, com foco na extração e classificação de opinião. Por fim, destaca-se o DANTEStocks (DI FELIPPO; ROMAN, 2025) que, como mencionado, reúne 4.048 *tweets* do domínio do mercado financeiro.

Entre os *corpora* compostos por CGU, especialmente *tweets,* o DANTEStocks é o primeiro anotado segundo o modelo UD. Além das anotações de PoS e dependências sintáticas, ele também possui anotação de emoções, advinda do recurso compilado por Silva *et al.* (2020) que deu origem ao *corpus.* Recentemente, o DANTEStocks ganhou mais uma camada de anotação, no caso, de entidades nomeadas (Zerbinati *et al.* 2024).

**Quadro 1.** Lista de *corpora* de *tweets* em português.

| ***Corpus*** | **Referência** | **Fonte** |
|---|---|---|
| *Corpus* Eleições 2010 | Silva *et al.* (2011) | Twitter (político; campanha eleitoral) |
| *Corpus* 7x1 | Moraes *et al.* (2015) | Twitter (esporte/futebol) |
| *Corpus* Emojis | Corrêa Júnior *et al.* (2017) | Twitter (*emojis* positivos/negativos) |
| TweetSentBR | Brum e Nunes (2018) | Twitter (show de TV) |
| DANTEStocks | Di-Felippo e Roman (2025) | Twitter (bolsa de valores) |

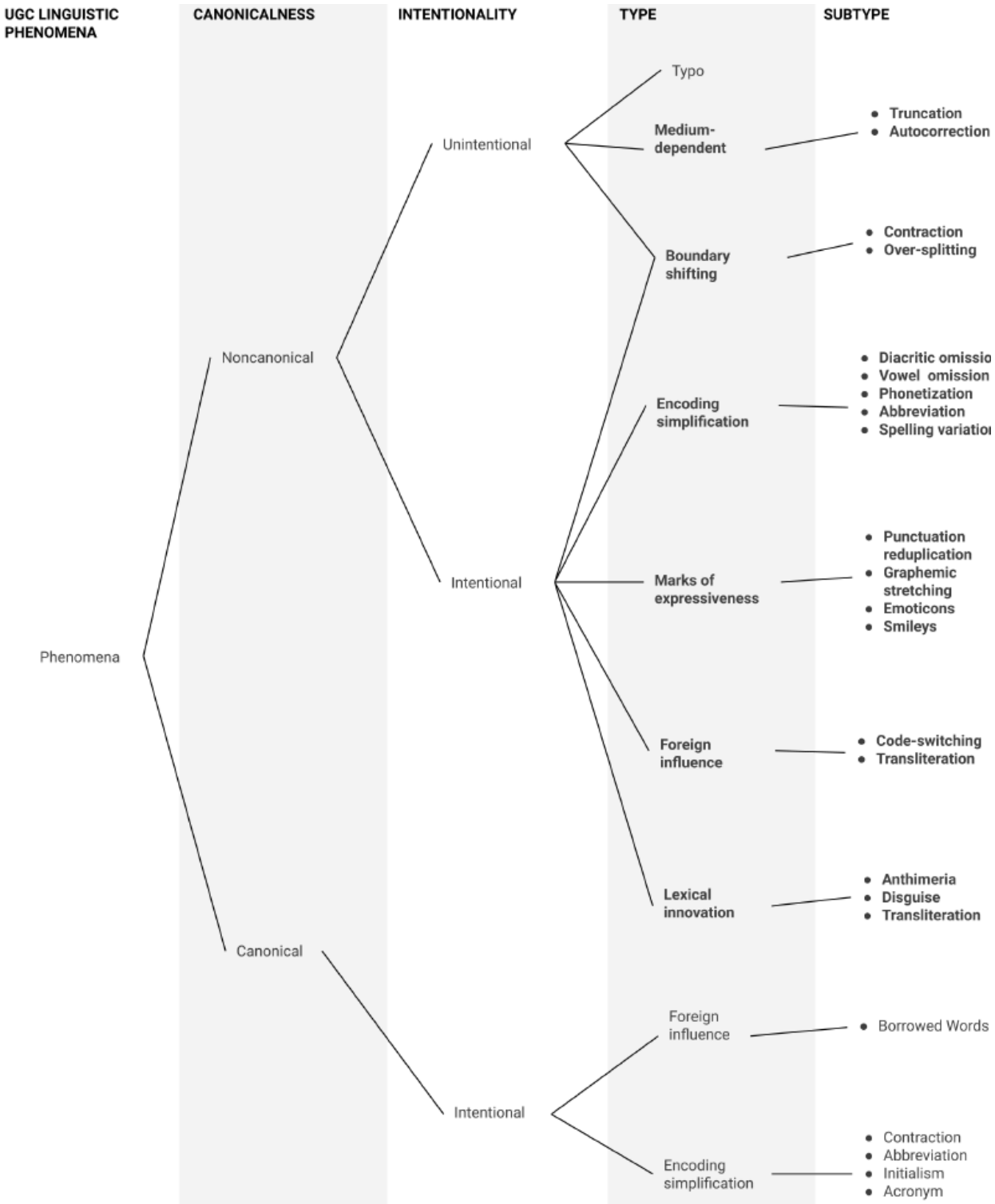
**Fonte:** A autora (2025).

### 2.5. Caracterização linguística de CGU/*tweets*

Diante da necessidade de definir diretrizes de anotação-UD, Sanguinetti *et al.* (2023) organizaram as particularidades lexicais e ortográficas de CGU (especialmente, *tweets*) em uma taxonomia de tipos (variedade do fenômeno) e subtipos (subcategorização de cada tipo) com base em duas dimensões:

canonicidade e intencionalidade. “canonicidade” se refere à ocorrência ou não do fenômeno em textos formais e “intencionalidade” diz respeito ao fato de o fenômeno ter ocorrido de forma inadvertida/acidental ou não (Figura 4).

**Figura 4.** Tipologia de fenômenos léxico-ortográficos de Sanguinetti et al.




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

Sobre a intencionalidade, os autores ressaltam que, em razão da incerteza inerente à interpretação e da natureza contextual dos *tweets*, a categorização do fenômeno quanto à intencionalidade pode ser apenas inferida, já que não é possível confirmá-la apenas pela observação da superfície do texto. Por fim, eles dizem que um único *token* pode apresentar diferentes fenômenos.

Na Figura 5, listam-se exemplos dos diferentes fenômenos sistematizados pelos autores, sem, entretanto, a indicação de intencionalidade e canonicidade.

**Figura 5.** Tabela de exemplos de fenômenos CGU não-canônicos.

| Phenomenon | Lang | Attested example | Standard form | Gloss |
|---|---|---|---|---|
| Encoding simplification | | | | |
| Diacritic omission | GA | *Leigh aris!* | *Léigh arís!* | 'Read again!' |
| | TR | *Istanbuldaki agaclar* | *İstanbul'daki ağaçlar* | 'trees in Istanbul' |
| Vowel omission | EN | *ppl* | *people* | 'people' |
| | TR | **slm** | *selam* | 'hi' |
| Phonetization | EN | *Happy Birthday* **2** *me* | *Happy Birthday to me* | 'Happy Birthday to me' |
| | TR | **1** *az* | *biraz* | 'some' |
| | DE | *k* **1** *Mensch hat so* **1** *Thailandhass* | *kein Mensch hat so einen Thailandhass* | 'nobody has such a hatred of Thailand' |
| Spelling variation | FR | *je* **sé** | *je sais* | 'I know' |
| | GA | **gura** *míle* | *go raibh míle* | 'thank you very much' |
| | FR | *tous mes* **examen** **son** *normaux* | *tous mes examens sont normaux* | 'All my examinations are normal' |
| | IT | **anno** *mangiato* | *hanno mangiato* | '(they) have eaten' |
| Abbreviation | EN | **govt** | *government* | 'government' |
| | DE | **zuggm** | *zugegebenermaßen* | 'admittedly' |
| Boundary shifting | | | | |
| Contraction | FR | **nimp** *quoi* | *n'importe quoi* | 'rubbish' |
| Over-splitting | FR | **c a dire** | *c'est-à-dire* | 'namely' |
| | TR | **gele bilirim** | *gelebilirim* | 'I can come' |
| Marks of expressiveness | | | | |
| Punct. reduplication | FR | *Joli* ***!!!!!!*** | *Joli !* | 'nice!' |
| | IT | *chi* ***?!?!?!*** | *chi?* | 'who?' |
| Case variation | GA | *is* ***BREÁ le daoine*** | *is breá le daoine* | 'people love' |
| Graphemic stretching | EN | *supe***rrrrrrrr** | *super* | 'great' |
| | IT | ***siiiiiiiiiiii*** | *sì* | 'yes' |
| Emoticons/smileys | - | *:-)* *<3* | – | – |
| | GA | *<3 mór* | *Grá mór* | 'Lots of love' |
| Lexical innovation | | | | |
| Disguise | IT | ***caxxo*** | *cazzo* | 'fuck' |
| | TR | ***mok / b.k / b*k*** | *bok* | 'shit' |
| | DE | ***Verfi**t*** lange Reise | *Verfickt lange Reise* | 'fucking long trip' |
| Anthimeria | IT | **tuittare** | *twittare* | 'to tweet' |
| | EN | feel free to *PM* | *personal message* | 'to send a message' |
| | DE | **achtisch** | *EN eightish* | 'about 8 o'clock' |
| Foreign language influence | | | | |
| Transliteration | GA | **áicbheaird** | *amscaí* | 'awkward' |
| | TR | **taymlayn** | *zaman akışı* | 'timeline' |
| Medium-dependent phenomena | | | | |
| Truncation | GA | *thart fa' 53* **nó...** | *thart fa' 53 nóiméad* | 'over 53 mi...(minutes)' |
| Autocorrection | GA | **concise** | *coicíse* | 'fortnight' |

**Fonte:** Sanguinetti *et. al*. (2022).

Os fenômenos canônicos (isto é, que também são observados em textos formais) e intencionais, na parte de baixo da Figura 4, são dos tipos “influência estrangeira” (*foreign influence*), que tem nos empréstimos lexicais seu único subtipo, e “simplificação de código” (*encoding simplification*), que envolve economia de esforço na escrita, como ocorre nos subtipos contração, abreviação, inicialismo e acrônimo.

Os fenômenos não canônicos (parte superior da Figura 4) e não intencionais, isto é, que ocorrem tipicamente em CGU de forma acidental, são de três tipos. Um deles são os erros de digitação *(typo*). Outro tipo são os fenômenos dependentes do meio (*medium-dependent*), isto é, aqueles realizados pela própria plataforma; seus subtipos são truncamento (isto é, quebra de uma palavra pelo limite de caracteres) e autocorreção, que pode consistir na filtragem ou substituição de palavras tabu ou ofensivas. O deslocamento de fronteira (*boundary shifting*) é o único fenômeno dependente do meio que pode ser não intencional ou intencional. Ele se refere a alterações no número de *tokens* comparado à ortografia padrão, sendo que seus subtipos são a contração (p.ex.: a expressão multipalavra do francês “*n’importe*” é contraída para “*nimp*”) e a supersegmentação (*over-splitting*), que ocorre quando um único *token* da língua padrão é desmembrado em vários *tokens* (p.ex.: a expressão em francês “*c’est-à-dire*” (“isto é”/”ou seja”) aparece supersegmentada em “*c a dire*”).

Os fenômenos não canônicos e exclusivamente intencionais são:

a) simplificação de código, cujos tipos são (i) omissão de diacrítico (p.ex.: a expressão irlandesa *Léigh arís!* (“leia de novo”) ocorre como *Leigh aris!*) e vogal (p.ex.: *people* (“pessoa”) para *ppl*), (ii) fonetização[11] (p.ex.: a expressão em inglês *Happy Birthday to me* (“Parabéns pra mim”) ocorre como *Happy Birthday 2 me*), (iii) abreviação (p.ex.: *government* (“governo”) para *govt*) e (iv) variação de grafia (p.ex: *je sais* em francês (“eu sei”) ocorrer como *je sé*).
b) marcas de expressividade (*marks of expressiveness*), que são usadas para indicar emoção ou ênfase; os tipos são (i) prolongamento grafêmico (p.ex.:

[11] Processo de representar sons da fala com símbolos ou grafias que imitam a pronúncia em vez da ortografia padrão.

*yesssss* (“sim”)), repetição de pontuação (p.ex.: ?????) e *emoticons* (como xD para risada intensa) ou *emoji* (p.ex.: 😳).

c) influência estrangeira, cujos tipos são (i) mistura de línguas (*code-switching*) em um único *tweet* (p.ex.: “*non fare la bad girl*” (“não seja a garota má”) substitui a forma padrão “*non fare la cattiva ragazza*”) e (ii) transliteração[12] (p.ex.: “áicbheaird” parece uma forma de imitar foneticamente a palavra inglesa “awkward” em irlandês).

d) Inovação lexical, cujos tipos são (i) disfarce (*disguise*), isto é, formas alternativas, abreviadas ou censuradas para evitar o uso explícito de palavrões, seja por censura, eufemismo ou para suavizar a expressão (p.ex.: *caxxo* (“caralho/porra”) ao invés da forma padrão *cazzo* em italiano), (ii) anthimeria[13] (p.ex.: o verbo *to tweet* adaptado para *tuittare* em italiano) e o já explicado (iii) transliteração.

Além do trabalho de Sanguinetti *et el.* (2023), destaca-se o de Scandarolli *et al*. (2023). Como este fora baseado no *corpus* DANTEStocks e empregado neste TCC, ele é apresentado adequadamente na próxima seção, após a descrição do *corpus.*

## 3. O *Corpus* DANTEStocks

### 3.1. Informações gerais

O DANTEStocks é um *corpus* de CGU com múltiplas camadas de anotação *gold standard* em língua portuguesa. A versão atual é composta por 4.048 *tweets* sobre o mercado financeiro, totalizando aproximadamente 81 mil *tokens*.

Ele integra o *treebank* Porttinari, que está sendo desenvolvido no âmbito do projeto POeTiSA[14] (*POrtuguese processing – Towards Syntactic Analysis and parsing*) da frente de NLP2 (*Natural Language Processing for Portuguese*) do Centro de Inteligência Artificial (C4AI) da Universidade de São Paulo. O projeto POeTiSA visa contribuir para o avanço de recursos e ferramentas de PLN voltados à análise sintática do português por meio da criação do *corpus*

[12] Processo de representar os sons de uma palavra de uma língua usando os caracteres de outro sistema de escrita, preservando sua forma fonética.
[13] Anthimeria é uma figura de linguagem que consiste em usar uma palavra de uma classe gramatical para exercer a função de outra.
[14] https://sites.google.com/icmc.usp.br/poetisa

multigênero Porttinari[15], anotado conforme o modelo UD. Em outras palavras, o DANTEStocks corresponde ao *subcorpus* de CGU que compõe o Porttinari, que atualmente está em sua versão 2.0.

Mais precisamente, o DANTEStocks resulta do refinamento e da anotação morfossintática-UD do *corpus* originalmente compilado por Silva *et. al*. (2020) em 2014. A compilação dos *tweets* foi feita com base na ocorrência de ao menos um *ticker* de uma das 73 ações que compunham o índice Bovespa à época. Um *ticker* é o código alfanumérico (normalmente quatro letras e um número) que representa a empresa e o tipo da ação, como “PETR4”, que representa a ação preferencial da Petrobras. Os *tickers*, aliás, são comumente empregados no mercado financeiro em substituição aos nomes das empresas e organizações. Como a compilação ocorreu em 2014, a extensão máxima dos *tweets* do DANTEStocks é de 140 caracteres.

Ademais, destaca-se que os *posts* em questão estão em sua forma original, isto é, eles não foram submetidos a nenhum processo de segmentação em unidades estruturais menores (sentenças ou sintagmas) ou mesmo de normalização lexical. Com isso, o *corpus* possui uma mistura de linguagem padrão e não-padrão.

Essa mistura pode ser evidenciada pelo fato de que os *tweets* apresentam estruturais internas bastante variadas, incluindo (DI-FELIPPO *et al.*, 2021): (i) uma ou mais sentenças bem delimitadas (1) e (2); (ii) ausência de pontuação ou pontuação empregada inadequadamente (3) e (4); (iii) fragmentação textual (5); e (iv) colagens de trechos jornalísticos ou manchetes de outras fontes (6).

(1) Sera k petr4 já entrou na baixa?
(2) PETR4 subiu na bolsa 13,50. Muito bem, surpreso com o resultado.
(3) #PETR4 #PETROBRAS a R$13,13. Pronto! O #PT conseguiu fazer propaganda eleitoral antecipada O que a @user4 tem a dizer sobre isso?
(4) Bom dia Marcos, Alguma previsão para petr4?!
(5) #GGBR4 Suportes e resistências http://t.co/Azw6yIEVI9
(6) Logística, ex-LLX, anuncia prejuízo de R$ 135,8 milhões em 2013: A Prumo Logística, ex-LLX (LLXL3), divu... http://t.co/LwmlKPqssk.

[15] https://sites.google.com/icmc.usp.br/poetisa/porttinari-2-0

Sobre seus aspectos lexicais, os quais têm relação direta com este trabalho, discorre-se, na sequência, sobre as estratégias empregadas no processo de *toekenização,* posto que algumas das características das palavras no DANTEStocks resultam desse processo. Embora o *corpus* possua várias camadas de anotação, como mencionado, elas não são aqui descritas, pois não se relacionam diretamente com o tópico do trabalho. Mais informações sobre elas podem ser encontras em Di-Felippo e Roman (2025).

### 3.2. *Token*ização

A *tokenização* do *corpus* foi feita com base nos pressupostos do modelo UD, pois esse processo é condição para as anotações de PoS e de dependências sintáticas estabelecidas pelo modelo.

A *tokenização* é o processo responsável por dividir o texto em um conjunto de segmentos significativos, chamados *tokens*. Assumindo a perspectiva lexicalista da sintaxe do modelo UD, as relações de dependência ocorrem entre palavras, o que significa que não há necessidade de segmentar as palavras em morfemas. No entanto, é necessário delimitar as unidades básicas de anotação, denominadas “palavras sintáticas”[16].

Para *tokenizar* os 4.048 *tweets*, aplicou-se uma abordagem semiautomática, ou seja, tokenização automática seguida de revisão manual. Para isso, desenvolveu-se uma ferramenta baseada em regras chamada DANTE Tokenizer (SILVA, E. *et al.*, 2020). Esse *tokenizador* é uma versão do TweetTokenizer[17] do NLTK ampliada com regras específicas para preservar o conteúdo original dos *tweets* e respeitar as diretrizes-UD propostas por Sanguinetti *et al*. (2023).

Com base nos pressupostos da UD, o *tokenizador* do DANTEStocks possui, por exemplo, uma regra para separar sinais de pontuação das palavras adjacentes quando estes formam um único *token* (exceto no caso de abreviações).

[16] Palavra sintática (*syntactic word*) é a unidade mínima a que corresponde uma função sintática (https://universaldependencies.org/u/overview/tokenization.html).
[17] https://www.nltk.org/api/nltk.tokenize.html

No que diz respeito às peculiaridades do português, a visão lexicalista da UD impõe a separação ou desmembramento de clíticos e contrações. Assim, um único *token* ortográfico como “fez-se” foi dividido em três *tokens* individuais (“fez” “-“ “se”), uma vez que corresponde a múltiplas palavras (sintáticas). O mesmo ocorreu com contrações convencionais (p.ex. “na” > “em” “a”) e contrações não canônicas, que são muito frequentes no *corpus* (p.ex.: “pq” > “p” “q” e “oq” > “o” “q(ue)”).

No geral, a *tokenização* do *corpus* buscou seguir a delimitação original com base nos espaços em branco, incluindo casos de fonetização, *hashtags*, menções, *emoticons* e URLs. Para ilustrar isso, considere o uso não dos sinais de pontuação como pictogramas. O TweetTokenizer original do NLTK dividiria ocorrências como “:)” em vários sinais de pontuação (isto é, “:” e “)”). Como os *emoticons* podem substituir palavras da língua padrão, eles não devem ser divididos em mais de um segmento. Assim, uma regra foi adicionada para garantir o reconhecimento correto dessas cadeias de pontuação como *tokens* únicos ou individuais.

Regras semelhantes também foram criadas para garantir o reconhecimento adequado de ocorrências de fenômenos específicos do domínio como *tokens*, incluindo *tickers*, *cashtags*[18], números decimais com parte fracionária indeterminada (p.ex.: “18,xx”) e expressões temporais alfanuméricas. Para os *tickers,* por exemplo, uma regra foi criada para reconhecer cadeias alfanuméricas do tipo “petr4” (1) como um *token* único.

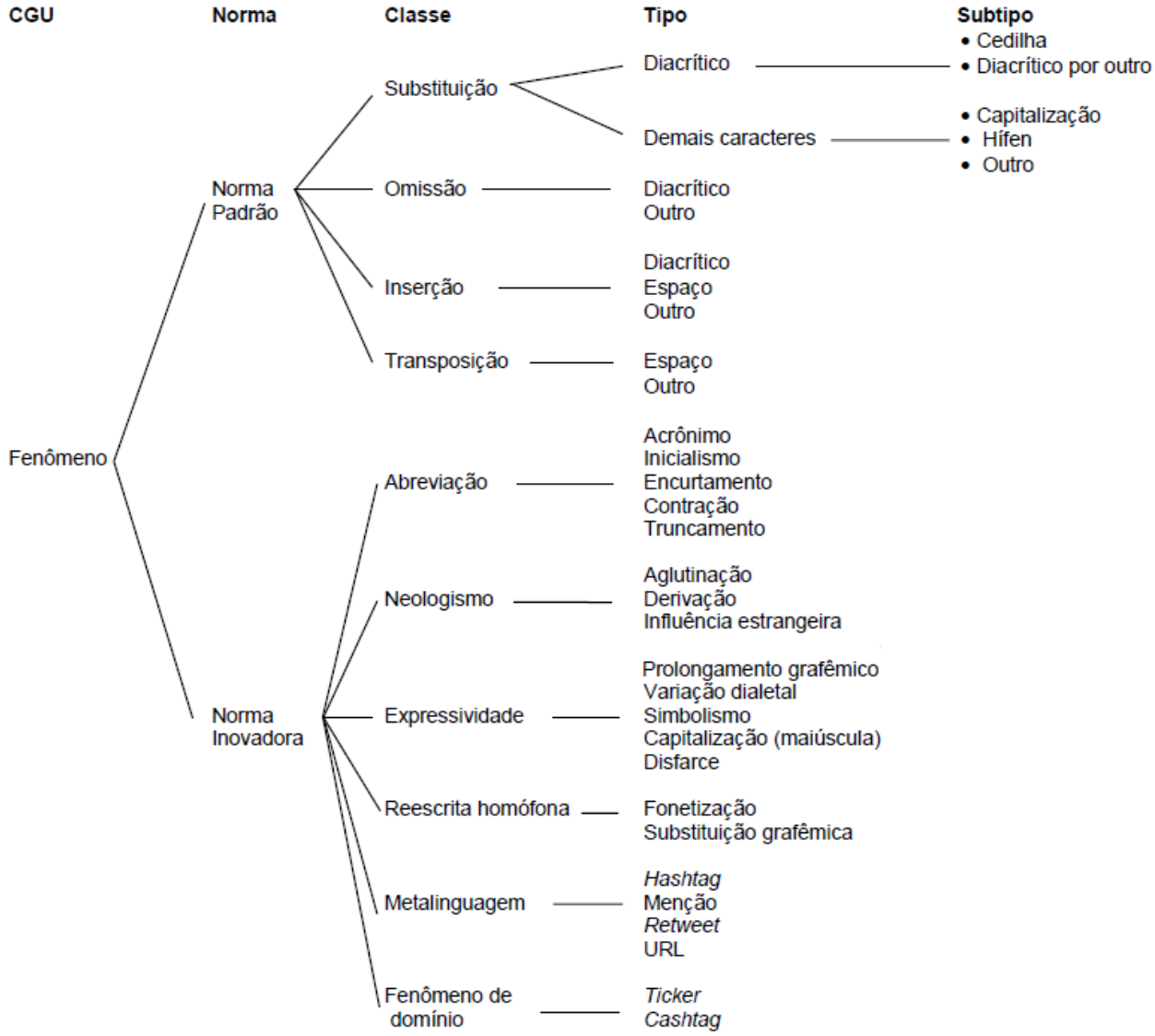
Por outro lado, mudanças de valores das ações (p.ex.: “+2,10%” > “+” “2,10” “%”) e valores monetários com formatos não convencionais (p.ex.: “R$20,00” > “R$” “20,00”) foram divididos em mais de um *token*. Tais ocorrências foram decompostas porque os símbolos matemáticos (como “+” e “-”), símbolo de porcentagem (%) e símbolos de moeda (como “R$”) podem ser substituídos por palavras comuns. Em resumo, a tokenização do DANTEStocks buscou preservar o conteúdo original dos *tweets*, com poucas exceções.

[18] *Cashtag* é um marcador textual que utiliza o símbolo $ seguido do código de uma ação (por exemplo, “$Petr4”). Assim como as *hashtags* (#) agrupam tópicos gerais, as *cashtags* permitem a indexação e rastreamento de menções a ações nos *posts*, facilitando o monitoramento de tendências, análises e opiniões do mercado.

## 4. Fenômenos léxico-ortográficos do DANTEStocks

Embora Sanguinetti *et al*. (2023) tenham sistematizado os fenômenos léxico-ortográficos típicos de CGU/tweets, a intencionalidade permanece como um critério subjetivo e, portanto, passível de questionamento. Além disso, o DANTEStocks, por ser um *corpus* do domínio financeiro, apresenta particularidades que precisavam ser identificadas para fundamentar a proposição de diretrizes específicas para anotação-UD. Nesse sentido, Scandarolli *et al.* (2023) descreveram os fenômenos léxico-ortográficos presentes no DANTEStocks, buscando utilizar um critério menos subjetivo.

**Figura 6.** Tipologia de idiossincrasias léxico-ortográficas do DANTEStocks.




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

A descrição foi feita a partir da análise de um conjunto de 1.363 *tokens* (sendo 1.069 *tweets*) inicialmente marcados como *Typo=Yes* durante a etapa de anotação morfossintática do *corpus.* Essa marcação foi feita exatamente porque tais *tokens* apresentavam ortografia não convencional para a qual era preciso definir a forma mais adequada de anotação-UD. Como resultado, os autores propuseram uma tipologia de duas dimensões: "Norma Padrão" e "Norma Inovadora" (Figura 6). Assim como em Sanguinetti *et al*., considera-se que um mesmo *token* pode apresentar mais de um fenômeno, sejam eles de uma mesma norma ou de normas diferentes.

A Norma Padrão diz respeito às variações gráficas que se afastam das convenções ortográficas tradicionais por razões diversas, como desconhecimento das regras da língua, influência do meio e dispositivo ou influência de novas regras fonéticas. Em outras palavras, tais variações são consideradas desvios da norma-padrão.

Nessa dimensão, os fenômenos foram organizados com base no conceito de *caractere* do padrão internacional de codificação de texto chamado Unicode[19,20]. No padrão Unicode, o termo *caractere* abrange uma ampla variedade de unidades textuais, como letras maiúsculas e minúsculas (como A e a), dígitos, sinais de pontuação e símbolos (como ?, ©, +), espaços e quebras de linha, sinais diacríticos (como acentos, til, cedilha) e pictogramas. A cada *caractere* de praticamente todas as línguas do mundo, o Unicode atribui um código numérico único chamado *code point*. Para ilustrar, a letra minúscula "o" tem o código U+006F e a letra maiúscula "O" é codificada por U+004F. Além disso, um *caractere* como "à" pode ser tratado de formas diferentes dependendo da normalização usada. Na representação pré-composta, ele é tratado como um único *caractere*, codificado por U+00E0. Na representação combinada, o "à" é tratado como a letra "a" seguida por um sinal diacrítico (acento grave), sendo, assim, codificado como dois *code points*: U+0061 U+0300.

Segundo os autores, o conceito de *caractere* permitiu classificar com maior objetividade os fenômenos da Norma Padrão nas categorias propostas por Damerau (1964) em seu estudo clássico sobre variantes ortográficas da língua

[19] http://www.unicode.org/standard/WhatIsUnicode.html
[20] Há *code points* para mais de 1 milhão de caracteres, permitindo que as máquinas representem e manipulem de forma consistente texto de qualquer sistema de escrita.

padrão com base na noção de “letra”. Como pode ser vista na Figura 6, a dimensão Norma Padrão possui as 4 classes (ou categorias) de Demerau (isto é, substituição, omissão, inserção e transposição), as quais possuem tipos, sendo substituição a única com subtipos.

A Norma Inovadora abrange fenômenos empregados tanto para expressar um conceito de modo alternativo à linguagem padrão quanto para comunicar um conceito novo. Assim, tais fenômenos se relacionam às “variantes linguísticas”, as quais não estão contempladas na norma-padrão porque são resultados da utilização de recursos ortográficos de forma criativa e inovadora. As classes dessa dimensão são 6: abreviação, neologismo, expressividade, reescrita homófona, metalinguagem e fenômeno do domínio.

**(a) Norma Padrão**

- Substituição: Acontece quando pelo menos um *caractere* de um *token* é substituído por outro, ocasionando erro ortográfico. Os tipos são (i) diacrítico e (ii) demais caracteres. O tipo “diacrítico” se divide nos subtipos “cedilha” e “diacrítico por outro”. Embora a cedilha tecnicamente seja um diacrítico, os autores optaram por tratá-la separadamente pois, de forma contrária a outros diacríticos (como til e acentos), a cedilha só faz sentido com “c”, tanto é que, em *layouts* de teclado, ela costuma ter tecla própria, refletindo sua associação exclusiva com “c”. Assim, o subtipo “cedilha” captura grafias como “acougue” ao invés de “açougue”. O subtipo “diacrítico por outro”, por sua vez, captura a substituição em casos como “mâe”, em que o til foi substituído pelo acento circunflexo. O tipo “demais caracteres” engloba os subtipos “capitalização” (como “dilma” ao invés de “Dilma”), “hífen” (como “cruz credo” ao invés de “cruz-credo”) e “outro” (como “hirário” em vez de “horário”).
- Omissão: Ocorre quando um *caractere* não é expresso. Seus tipos são (i) diacrítico, como no *token* “esta” (ao invés de “está”) em que o usuário omitiu o acento, e (ii) demais caracteres, como a ausência do “s” final em “açõe”.
- Inserção: Acontece quando um *caractere* é acrescido indevidamente. Seus tipos são (i) diacrítico (como “Petrobrás” ao invés de “Petrobras”), (ii) espaço (como “a final” ao invés de “afinal”) e (iii) outro (como “Streaddle” ao invés de “Straddle”).

- Transposição: Refere-se à troca de ordem de *caracteres*. Seus tipos são (i) espaço, como em “meua migo” (“meu amigo”), e (ii) outro, como “acrodo” ao invés de “acordo”.

**(b) Norma Inovadora**

- Abreviação: Fenômeno que produz um *token* mais curto que a expressão original. Seus tipos são (i) acrônimo, isto é, *token* formado pelas letras iniciais ou sílabas de várias palavras e pronunciado como uma única palavra (p.ex.: “CEMIG” – “Companhia Energética de Minas Gerais”); (ii) inicialismo, ou seja, *token* formado pelas letras iniciais de várias palavras e pronunciado letra por letra (p.ex.: “lp” – “longo prazo”); (iii) encurtamento, que se caracteriza pela ausência das letras finais de uma palavra (p.ex.: “q” – “que”); (iv) contração, que se caracteriza pela ausência de letras intermediárias, (p.ex.: “enqt” – “enquanto”); (v) truncamento, isto é, *token* quebrado no final do *tweet* por limite de caracteres, geralmente seguido de reticências (p.ex.: “divu” – divulgou)
- Neologismo: Palavras novas ainda não institucionalizadas, podendo ser dos tipos (i) aglutinação, isto é, junção de duas palavras em uma só (p.ex.: “Ibolixo” – “Ibovespa” + “lixo”), (ii) derivação, ou seja, adição de afixo a uma raiz existente (p.ex.: “diretassa” – “direta” + “-assa/aça”) e (iii) influência estrangeira, como “estopar” (do inglês “*stop*”).
- Expressividade: Fenômeno que simula sentimento, expressão facial ou gesto, podendo ser dos tipos (i) prolongamento grafêmico, como o alongamento de palavras (p.ex.: “noossaaa”), (ii) variação dialetal, “malmita” (ao invés de “marmita”), (iii) simbolismo, isto é, uso de caracteres simbólicos (*emojis* e *emoticons*) em substituição a palavras ou partes delas, (iv) capitalização, que se caracteriza pelo uso de maiúsculas para expressar ênfase e (v) disfarce, ou seja, substituição de letras por *caracteres* especiais para censura (p.ex.: “m*” ao invés de “merda”).
- Reescrita homófona: Variações gráficas motivadas pela fonética ou simplificação de diacríticos. Seus tipos são (i) fonetização como “krai” (ao invés de “caralho”) e (ii) substituição grafêmica, isto é, uso de letras adicionais para substituir diacríticos (p.ex.: “neh” – “né” e “tou” – tô).

- Metalinguagem: *Tokens* que ocorrem tipicamente no Twitter e não constam em dicionários, como (i) *hashtag* (p.ex.: "#PT"), (ii) menção (p.ex.: "*@Usuário*" [21]), (iii) marca de *retweet* ("RT") e (iv) URL.
- Fenômeno de domínio: *Tokens* comuns em *tweets* do mercado financeiro, como os *tickers* (p.ex.: "LLXL3") e as *cashtags* (p.ex.: $PETR3).

Com base na taxonomia, Scandarolli *et al.* anotaram os fenômenos presentes em 1.069 *tweets* do DANTEStocks (do total de 4.048 *tweets*). A seguir, descreve-se a anotação manual dos fenômenos léxico-ortográficos em mais 2.000 *tweets* distintos do DANTEStocks com base na mesma taxonomia.

## 5. Anotação de *Corpus*

### 5.1. Seleção dos dados e metodologia de anotação

Tomou-se como ponto de partida os 4.048 *tweets* em formato CoNLL-U armazenados em planilhas do *Google Sheets* (Figura 8). Além dos 1.069 *tweets* já anotados em função da taxonomia de Scandarolli *et al.,* selecionou-se um conjunto de 2.000 *tweets* distintos e subsequentes aos 1.069 anotados, que foram organizados em pacotes de 400 *tweets* para controle da tarefa. A anotação manual de cada pacote foi feita de forma linear, *tweet* a *tweet* (e *token a token*).

A anotação durou cerca de um semestre e foi realizada semanalmente, sendo anotados 100 *tweets* por semana, demandando aproximadamente 4 horas em cada anotação.

Na Figura 7, ilustra-se um arquivo CoNLL-U utilizado na anotação. Vê-se que o arquivo original (Figura 3) foi simplificado pela ocultamento de colunas com informações não relevantes à tarefa, a saber: colunas 5 (XPoS), 7 (HEAD), 8 (DEPREL) e 9 (DEPS). Ademais, inseriram-se mais 4 colunas, cada uma delas destinada às informações prevista da tipologia de Scandarolli *et al.,* isto é, norma, classe ou categoria, tipo e subtipo do fenômeno. Na Figura 7, por exemplo, há a ocorrência de dois fenômenos da Norma Inovadora, um observado no *token* "#vale5" e outro em "Obj". O primeiro é da classe metalinguagem (tipo *hashtag*) e o segundo, da classe abreviação (tipo encurtamento, abreviado para "encurt").

[21] Neste exemplo, o usuário foi anonimizado para "*usuário*".

**Figura 7.** Exemplo de arquivo CoNLL-U modificado para a anotação.

| # sent_id = dante_01_446347773408190464l | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| # text = #vale5 quem acreditou no Obj Parabéns!!! | | | | | | | | | |
| **ID** | **FORM** | **LEMMA** | **UPoS** | **FEATS** | **MISC** | **Norma** | **Classe** | **Tipo** | **Subtipo** |
| 1 | #vale5 | #vale5 | PROPN | _ | _ | Inovadora | Metalinguagem | Hashtag | |
| 2 | quem | quem | PRON | PronType=Ind | _ | | | | |
| 3 | acreditou | acreditar | VERB | Mood=Ind\|Number=Sing\|Person=3\|Tense=Past\|VerbForm=Fin | _ | | | | |
| 4-5 | no | _ | _ | _ | _ | | | | |
| 4 | em | em | ADP | _ | _ | | | | |
| 5 | o | o | DET | Definite=Def\|Gender=Masc\|Number=Sing\|PronType=Art | _ | | | | |
| 6 | Obj | objetivo | NOUN | Abbr=Yes\|Gender=Masc\|Number=Sing | FullForm=objetivo | Inovadora | Abreviação | Encurt | |
| 7 | Parabéns | parabéns | NOUN | Gender=Masc\|Number=Plur | SpaceAfter=No | | | | |
| 8 | ! | ! | PUNCT | _ | SpaceAfter=No | | | | |
| 9 | ! | ! | PUNCT | _ | SpaceAfter=No | | | | |
| 10 | ! | ! | PUNCT | _ | SpaceAfter=No | | | | |

**Fonte:** A autora (2025).

Caso um *token* apresentasse fenômenos de normas distintas ou mesmo tipos distintos de uma mesma norma, a anotação foi feita como se ilustra na Figura 8. Nela, observa-se que no *token* “rapáááz (11) ocorrem fenômenos de ambas as normas. Assim, na coluna “norma, tem-se a indicação de ambas na sequência (isto é, “Inovadora, Padrão”. O mesmo ocorre com as respectivas classes expressividade (Exp) e inserção (Ins) e tipos prolongamento grafêmico (“Prolong”) e diacrítico (“Dia”).

**Figura 8.** Exemplo de token com variação de “Norma Inovadora” e “Norma Padrão”

| # sent_id = dante_01_446652807991791616l | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| # text = PETR4, alta de 10% em 3 dias... rapáááaz | | | | | | | | | |
| **ID** | **FORM** | **LEMMA** | **UPoS** | **FEATS** | **MISC** | **Norma** | **Classe** | **Tipo** | **Subtipo** |
| 1 | PETR4 | PETR4 | PROPN | _ | SpaceAfter=No | Inovadora | Domínio | Ticker | |
| 2 | , | , | PUNCT | _ | _ | | | | |
| 3 | alta | alta | NOUN | Gender=Fem\|Number=Sing | _ | | | | |
| 4 | de | de | ADP | _ | _ | | | | |
| 5 | 10 | 10 | NUM | NumType=Card | SpaceAfter=No | | | | |
| 6 | % | % | SYM | _ | _ | | | | |
| 7 | em | em | ADP | _ | _ | | | | |
| 8 | 3 | 3 | NUM | NumType=Card | _ | | | | |
| 9 | dias | dia | NOUN | Gender=Masc\|Number=Plur | SpaceAfter=No | | | | |
| 10 | ... | ... | PUNCT | _ | _ | | | | |
| 11 | rapáááaz | rapaz | NOUN | Gender=Masc\|Number=Sing | CorrectForm=rapaz\|SpaceAfter=No | Inovadora, Padrão | Exp, Ins | Prolong,Dia | |

**Fonte:** A autora (2025).

### 5.2. Recorte tipológico

Da proposta original de Scandarolli *et al.*, optou-se por anotar apenas os fenômenos que tendem a ocorrer com predominância nos *tweets* do mercado financeiro. Com isso, acrônimos (p.ex.: "CEMIG") e inicialismos (p.ex.: "PT" – Partido dos Trabalhadores) convencionais, de ampla circulação em textos formais e de língua geral, não foram incluídos na anotação. Em contrapartida, *tokens* como "ILC" (Índice de Liquidez Corrente) foram consideradas por representarem conceitos específicos do universo da bolsa de valores. O mesmo recorte foi aplicado a abreviações já dicionarizadas, como "vol." (volume) e "tá" (forma reduzida de "estar"), também não anotadas.

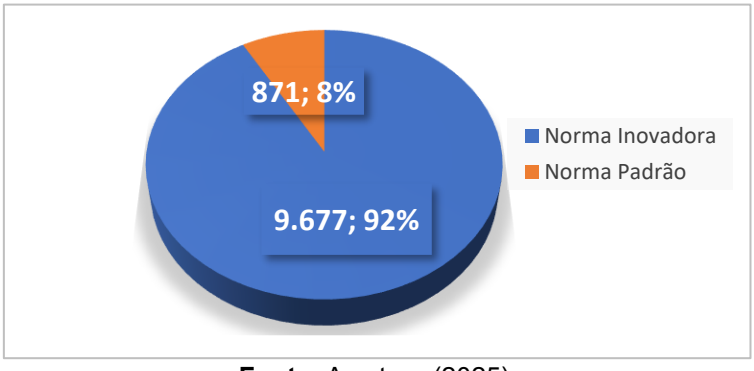
## 6. Caracterização linguística do *corpus*

A presente seção tem por objetivo apresentar a caracterização linguística do *corpus* DANTEStocks com base nos fenômenos léxico-ortográficos anotados via tipologia de Scandarolli *et al.* (2023).

### 6.1. Caracterização geral

No total, anotaram-se 10.548 ocorrências de fenômenos léxico-ortográficos foram anotadas, as quais ocorrem em 10.289 *tokens* distintos. Dos fenômenos identificados, 9.677 (92%) pertencem à dimensão Norma Inovadora, correspondendo a 9.850 *tokens*, enquanto 871 (8%) se enquadram na Norma Padrão, totalizando 709 *tokens*, conforme ilustrado na Figura 9.

**Figura 9.** Ocorrências dos fenômenos em função das Normas Inovadora e Padrão.




**Fonte:** A autora (2025).

### 6.2. Norma Inovadora

Os 9.677 fenômenos da Norma Inovadora se distribuem nos tipos e classes conforme descritos no Quadro 2. Nesse quadro, observa-se que todas as 6 classes ocorreram no *corpus*, sendo que, dos 21 tipos, apenas 20 foram observados. A única exceção diz respeito a acrônimo, da classe abreviação. Quanto às classes, vê-se claramente que os mais frequentes são os fenômenos de metalinguagem, seguidos de perto pelos de domínio. Em contrapartida, a menos frequentes é neologismo, com apenas 27 ocorrências.

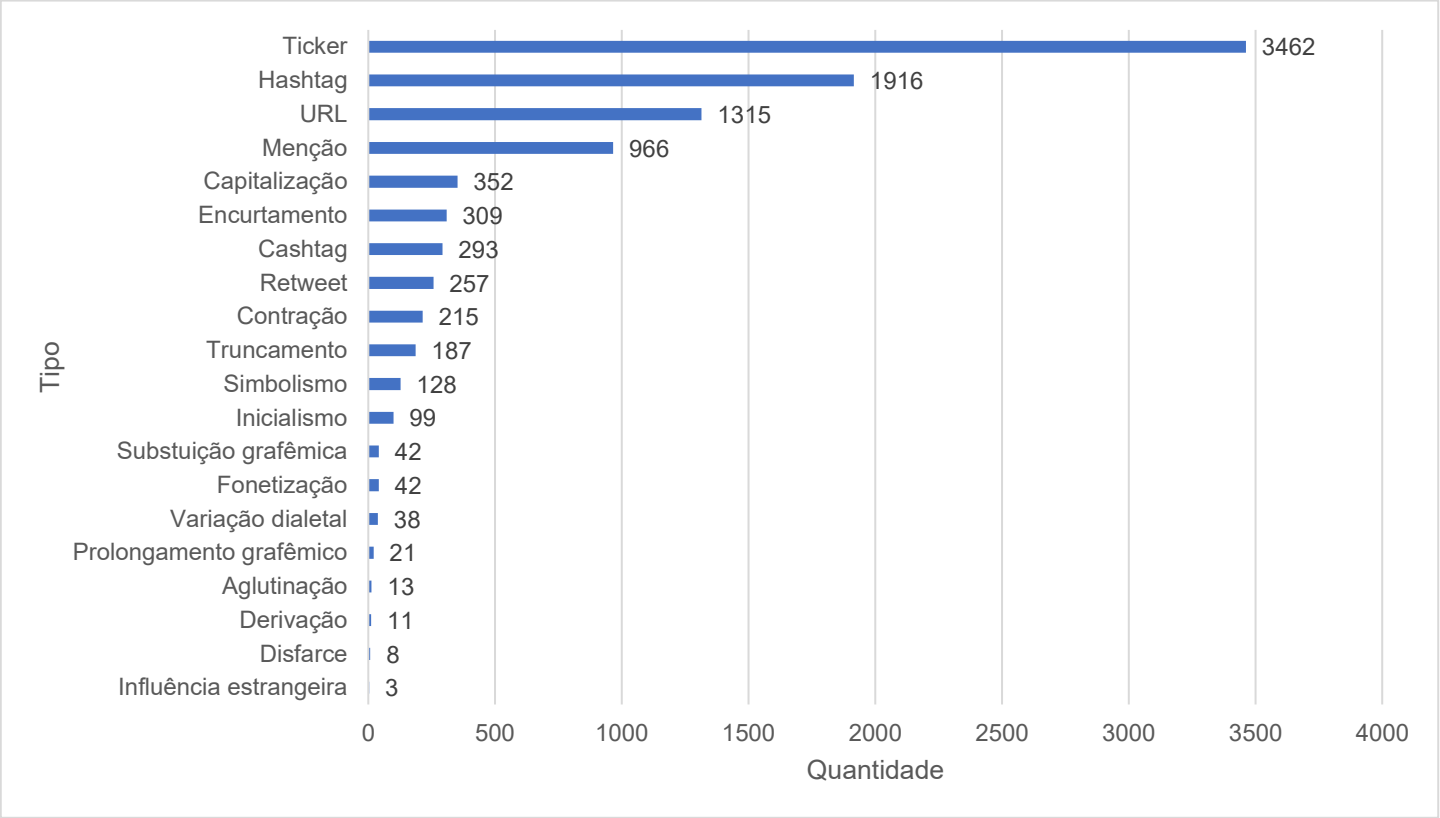
Na Figura 10, apresenta-se a distribuição dos tipos em gráfico. Nela, vê-se que os *tickers* e as *hashtags* são os mais frequentes, com 3.462 e 1.916 casos, respectivamente. A alta frequência de *ticker* pode ser explicada pelo modo de compilação do *corpus*, que se baseou na ocorrência desses elementos nas mensagens. Em contrapartida, há apenas 3 casos de influência estrangeira.

**Quadro 2.** Distribuição das classes da Norma Inovadora.

| Classe | Tipo | Exemplo | Lema | Qt. | Subtotal |
|---|---|---|---|---|---|
| Abreviação | Inicialismo | PB | price-to-book | 99 | 810 |
| | Encurtamento | d | de | 309 | |
| | Contração | Abç | abraço | 215 | |
| | Truncamento | abai | abaixo | 187 | |
| Neologismo | Aglutinação | ibolixo | ibolixo | 13 | 27 |
| | Derivação | diretassa | direto | 11 | |
| | Influência estrangeira | stopando | stopar | 3 | |
| Expressividade | Prolongamento grafêmico | noosaaa | nossa | 21 | 547 |
| | Variação dialetal | ocê | você | 38 | |
| | Simbolismo | =) | =) | 128 | |
| | Capitalização | ALEGRIA | alegria | 352 | |
| | Disfarce | P**a | puta | 8 | |
| Reescrita homófona | Fonetização | hehehe | hehehe | 42 | 84 |
| | Substituição grafêmica | eh | ser | 42 | |
| Metalinguagem | Hashtag | #VALE5 | #VALE5 | 1916 | 4454 |
| | Retweet | RT | RT | 257 | |
| | Menção | @user | @user | 966 | |
| | URL | http://t.co/03Aet66FTO | http://t.co/03Aet66FTO | 1315 | |
| Fenômeno de domínio | Ticker | ABEV3 | ABEV3 | 3462 | 3755 |
| | Cashtag | $AEDU3 | $AEDU3 | 293 | |
| | | | | **TOTAL** | **9677** |

**Fonte:** A autora (2025).

**Figura 10.** Distribuição dos tipos da Norma Inovadora.




**Fonte:** A autora, 2025.

Além dos mais e menos frequentes, os tipos URL (1.315 ocorrências) e a menções (966 ocorrências) também se mostram expressivos, o que reforça a natureza interativa e referencial da linguagem utilizada no Twitter.

Um dado bastante interessante diz respeito à quantidade de tipos de fenômenos da Norma Inovadora que ocorre por *token*, pois isso evidencia as múltiplas operações linguísticas simultâneas que foram feitas em um único *token*, indo além de variações simples, causando intersecções entre tipos. Vale ressaltar que o somatório de *tokens* por norma excede o número de *tokens* distintos, por conta da existência de alguns casos quais ocorrem intersecções entre as normas também.

Com base no Quadro 3, a grande maioria dos *tokens* anotados (9.489 de 9.850) apresenta apenas 1 tipo de fenômeno. No entanto, há 91 deles em que se observou a ocorrência de 2 tipos de fenômenos e 3 em que há a ocorrência de 3 tipos de fenômenos distintos.

**Quadro 3.** Quantidade de tipos da Norma Inovadora por *token.*

| Tipo/*token* | Qt. |
|---|---|
| 1 | 9.486 |
| 2 | 91 |
| 3 | 3 |
| **TOTAL** | **9.850** |

**Fonte:** A autora (2025).

No Quadro 4, os 18 pares de tipos que ocorrem no *corpus* estão exemplificados e quantificados. Pode-se observar que o par “truncamento+URL” é o mais frequente, ocorrendo em 28 *tokens.* A predominância dessa combinação se justifica pelo grande volume de *tweets* terminados em *links* que indicam a fonte da informação veiculada e que, por conta do limite de caracteres da rede, acabam sendo “cortados” ou “truncados”. Na sequência, tem-se o par “encurtamento+substituição grafêmica”, com 24 ocorrências. Um exemplo dessa combinação ocorre no *token* “Ñ”. Essa forma lexical resulta do encurtamento de “não” pela omissão dos caracteres finais (isto é, da vogal nasal “ão”) e pela substituição grafêmica com valor fonético simbólico, uma vez que houve a troca do “n” por **“**ñ**”** (caractere que não pertence ao português, mas que evoca a nasalidade por associação com o til).

No Quadro 5, exibem-se as 2 combinações triplas de tipos encontradas no DANTEStocks. O trio “contração+capitalização+prolongamento_grafêmico” ocorreu apenas 1 vez, especificamente no *token* “MTOOOO” (“muito”). A contração nesse *token* consistiu na supressão dos caracteres iniciais “u” e “i” e o prolongamento grafêmico se deu pela repetição da volgal "o"a, além da capitalização para expressar ênfase. O outro trio observado foi “variação dialectal+prolongamento grafêmico+fonetização”, com dois casos. Em “Buunituuu” (“bonito”), a variação dialetal manifesta-se na troca de “bo” por “bu”, refletindo uma pronúncia típica de certas regiões ou contextos afetivos. A fonetização está presente na tentativa de reproduzir graficamente a pronúncia da fala, com a duplicação do “u” inicial aproximando a sonoridade informal. Já o prolongamento grafêmico, evidente na repetição do “u” final (“uuu”), intensifica a carga expressiva da palavra, simulando entonação alongada e afetiva típica da oralidade. O outro caso diz repito ao *token* “Brigaduu” (“obrigado””. Nele, a variação dialetal aparece na omissão do caractere inicial “o”. A fonetização

consistiu em substituir o “o” final por “u” para refletir o som mais fechado e informal. Por fim, o prolongamento grafêmico, evidenciado na repetição da vogal final (“uu”), intensifica o tom afetivo ou jocoso da expressão, simulando a entonação prolongada característica da oralidade.

**Quadro 4.** Ocorrência concomitante de 2 tipos da Norma Inovadora no *token*.

| # | Pares de tipos | Exemplo | Lema | Qt. |
|---|---|---|---|---|
| 1 | Truncamento, URL | http:// | http:// | 28 |
| 2 | Encurtamento, Substituição grafêmica | Ñ | não | 24 |
| 3 | Variação dialetal, Fonetização | guvêrno | governo | 6 |
| 4 | Inicialismo, Capitalização | PQP | PQP | 4 |
| 5 | Contração, Substituição grafêmica | d+ | demais | 4 |
| 6 | Hashtag, Truncamento | #VAL | #VALE5 | 4 |
| 7 | Contração, Capitalização | HJ | hoje | 3 |
| 8 | Variação dialetal, Substituição grafêmica | Bão | bom | 3 |
| 9 | Encurtamento, Capitalização | Q | quem | 3 |
| 10 | Contração, Fonetização | kgda | cagada | 2 |
| 11 | Variação dialetal, Prolongamento grafêmico | Ôrrrra | ôrrrra | 2 |
| 12 | Capitalização, Prolongamento grafêmico | CHUPAA | chupar | 2 |
| 13 | Aglutinação, Substituição grafêmica | sifu | sifu | 1 |
| 14 | Contração, Influência estrangeira | plz | please | 1 |
| 15 | Derivação, Influência estrangeira | scaplerzinho | scalper | 1 |
| 16 | Derivação, Capitalização | CAPETALIZAÇÃO | capetalização | 1 |
| 17 | Prolongamento grafêmico, Fonetização | ehehehhe | ehehehhe | 1 |
| 18 | Capitalização, Aglutinação | PETROFUMO | PETROFUMO | 1 |
| | | | **TOTAL** | **91** |

**Fonte:** A autora (2025).

**Quadro 5.** Ocorrência concomitante de 3 tipos da Norma Inovadora no *token*.

| # | Trios de tipos | Exemplo | Lema | Qt. |
|---|---|---|---|---|
| 1 | Contração, Capitalização, Prolongamento grafêmico | MTOOOO | muito | 1 |
| 2 | Variação dialetal, Prolongamento grafêmico, Fonetização | Buunituuu | bonito | 2 |
| | | Brigaduu | obrigado | |
| | | | **TOTAL** | **3** |

**Fonte:** A autora (2025).

### 6.3. Norma Padrão

As 871 ocorrências de fenômenos da Norma Padrão distribuem-se entre as classes, tipos e subtipos conforme ilustrado no Quadro 6.

**Quadro 6.** Distribuição dos fenômenos da Norma Padrão.

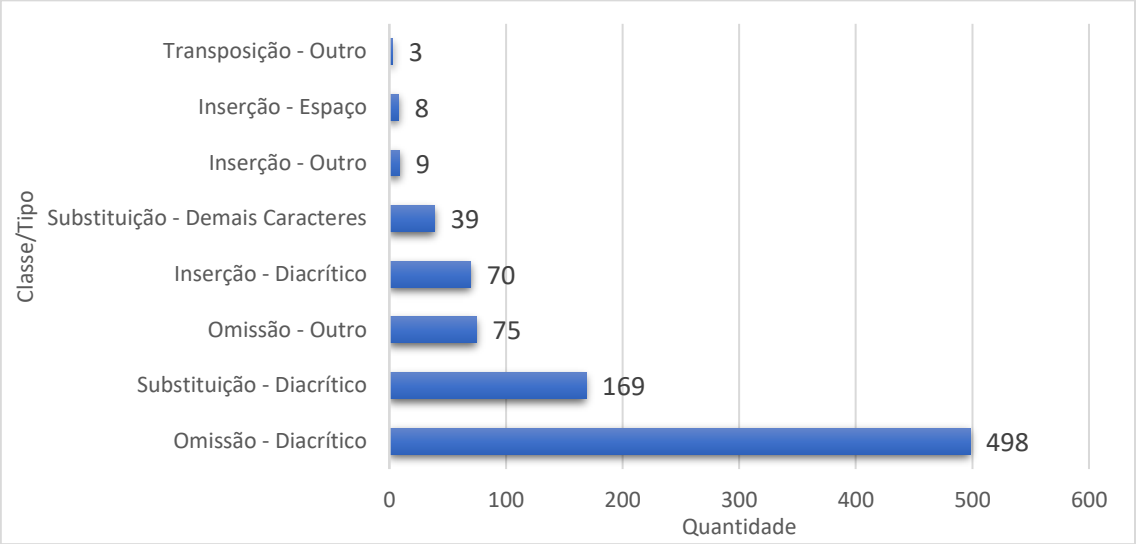
| **Classe** | **Qt.** | **Tipo** | **Qt** | **Subtipo** | **Qt.** | **Exemplo** | **Lema** |
|---|---|---|---|---|---|---|---|
| Substituição | 208 | Diacrítico | 169 | Cedilha | 167 | Abraco | abraço |
| | | | | Diacrítico por outro | 2 | sõ | só |
| | | Demais caracteres | 39 | Capitalização | 16 | MAravilha | maravilha |
| | | | | Hífen | 3 | pré sal | pré-sal |
| | | | | Outro | 20 | Aqwele | aquele |
| Omissão | 573 | Diacrítico | 498 | | | Economicos | econômico |
| | | Outro | 75 | | | Aind | ainda |
| Inserção | 87 | Diacrítico | 70 | | | analisár | analisar |
| | | Espaço | 8 | | | sub onda | subonda |
| | | Outro | 9 | | | nóis | nós |
| Transposição | 3 | Espaço | 0 | | | -- | -- |
| | | Outro | 3 | | | grnade | grande |
| **TOTAL** | **871** | | | | | | |

**Fonte:** A autora (2025).

A respeito das classes, observa-se no Quadro 6 que todas as 4 da dimensão Norma Padrão ocorrem no *corpus* (substituição, omissão, inserção e transposição). Dos 9 tipos, o único que não ocorre é a transposição envolvendo espaço. A classe omissão é de longe a mais frequente, com 573 ocorrências, o que equivale a aproximadamente 65% dos fenômenos da Norma Padrão anotados. A menos frequente é a transposição, com apenas 4 ocorrências.

Na Figura 11, apresenta-se a distribuição dos tipos em gráfico. Por não ser observado no *corpus,* o tipo transposição de espaço não consta do gráfico. Com base nessa figura, o tipo omissão de diacrítico é o mais frequente, com 498 ocorrências. Supõe-se que a frequência elevada desse tipo de fenômeno da Norma Padrão em tweets esteja relacionada à natureza informal e rápida desse meio, no qual a economia de esforço na digitação e a facilidade de compreensão, mesmo sem marcas gráficas, tornam esse tipo de simplificação ortográfica recorrente. Em contrapartida, o tipo transposição de outro *caractere* teve 4 ocorrências. Esses casos são: “deixnado” (“deixando”), “grnade” (“grande”) e “reias” (“reais”).

**Figura 11.** Distribuição dos tipos da Norma Padrão.




**Fonte:** A autora (2025).

Com base no Quadro 7, a grande maioria dos *tokens* anotados (548 de 709) apresenta apenas 1 tipo/subtipo de fenômeno da Norma Padrão. No entanto, há 159 deles em que se observou a ocorrência de 2 tipos distintos e dois subtipos correspondentes. Ademais, há 2 *tokens* em há a ocorrência de 1 tipo de fenômeno e 2 de seus subtipos.

**Quadro 7.** Quantidade de tipos/subtipos da Norma Padrão por *token.*

| **Tipo-Sutipo/*token*** | **Qt.** |
|---|---|
| 1 tipo, 1 subtipo | 548 |
| 2 tipos, 2 subtipos | 159 |
| 1 tipo, 2 subtipos | 2 |
| **TOTAL** | **709** |

**Fonte:** A autora (2025).

Os *token* com 2 tipos distintos e 2 subtipos correspondentes se caracterizam quase que exclusivamente pela ocorrência da combinação entre "substituição de diacrítico do subtipo cedilha" e "omissão de diacrítico". É o que se observa em *tokens* como "producao" (para "produção"), "acao" (para "ação") e "opcao" (para "opção"). À primeira vista, esses casos poderiam ser interpretados como ocorrências de omissão de dois diacríticos. No entanto, segundo a tipologia proposta por Scandarolli *et al*., esses dois fenômenos são categorizados de maneira distinta. A ausência do til é, de fato, classificada como

omissão de diacrítico, mas a troca de “ç” por “c” é considerada uma substituição de *caractere*, mais especificamente do subtipo “cedilha”. Essa distinção se baseia no fato de que “ç” e “c” são tratados como *caracteres* diferentes no sistema de codificação Unicode), e não apenas como uma letra com ou sem sinal gráfico, justificando seu tratamento separado dentro da tipologia. A exceção diz respeito à “pre abertura” (“pré-abertura), em que se observa a “substituição do hífen” (por espaço) e a “omissão do diacrítico” (til).

Os casos em que ocorrem 1 tipo e 2 subtipos são “GIlma” (“Dilma”) e “presal”. No primeiro *token,* “GIlma”, ocorre o tipo “substituição de demais *caracteres*”. Especificamente o *caractere* “D” foi trocado outro (subtipo “outro”), “G”, e “i” foi substituído por “I”, configurando o subtipo “capitalização”. No segundo *token* “presal”, há 2 subtipos de omissão. Um deles se refere à ausência do acento agudo (subtipo “diacrítico”) e outro à ausência do hífen (subtipo “outro”).

### 6.4. Intersecção entre as normas Inovadora e Padrão

Para além da análise restrita ao interior de cada norma, observou-se também a ocorrência de 14 casos em que fenômenos pertencentes a ambas as normas — Padrão e Inovadora — se manifestam simultaneamente, distribuídos em 10 tokens distintos (Quadro 8). Em outras palavras, a investigação contemplou não apenas uma análise intranorma, mas também uma análise internormas, permitindo identificar interações entre os diferentes padrões de variação ortográfica.

No Quadro 8, a classe e tipo de cada dimensão ou norma está descrita, juntamento com os *token* nos quais a concomitância foi observada.

**Quadro 8.** Interseções entre as “Norma Inovadora” e “Norma Padrão”.

| **Norma Inovadora** | **Norma Padrão** | **Qt.** | ***Token*** | **Lema** |
|---|---|---|---|---|
| **Classe-Tipo** | **Classe-Tipo** | | | |
| Neologismo-Aglutinação | Inserção-Diacrítico | 2 | PeTebrás | PeTebrás |
| Expressividade-Prolongamento grafêmico | Inserção-Diacrítico | 1 | rapáááz | rapaz |
| Metalinguagem-*Hashtag* | Inserção-Espaço | 1 | # | # |
| Expressividade-Capitalização | Inserção-Diacrítico | 1 | ELETROBRÁS | Eletrobras |
| | | 4 | PETROBRÁS | Petrobras |
| Expressividade-Capitalização | Omissão-Diacrítico | 4 | AMANHA, DIARIO, MINIMO, PARABENS | amanhã, diário, mínimo, parabéns |

| | | | | |
|---|---|---|---|---|
| Expressividade-Capitalização | Omissão-Espaço | 1 | LONG&SHORT | Long & Short |
| | **TOTAL** | **14** | | |

**Fonte:** A autora (2025).

De modo geral, esses casos de intersecção das normas, apesar de pouco numerosos, evidenciam a tensão entre inovação gráfica e adesão à norma padrão.

## 7. Considerações Finais

O presente trabalho teve como principal objetivo ampliar a anotação de fenômenos léxico-ortográficos no *corpus* DANTEStocks e, com isso, fornecer uma caracterização preliminar desse recurso com base na tipologia de Scandarolli *et al.* (2023).

De modo geral, observou-se que os fenômenos da Norma Inovadora predominam amplamente no *corpus*, refletindo o caráter específico da linguagem presente em CGU, especialmente a usada no Twitter/X. Em contrapartida, os fenômenos associados à Norma Padrão, embora em número reduzido, evidenciam uma recorrência significativa de omissões gráficas, o que pode ser considerado uma "economia de esforço" na escrita, devido à praticidade e rapidez que os *tweets* exigem.

Por fim, o desenvolvimento desse trabalho possibilitou não apenas a consolidação de conhecimentos em PLN, mas também o contato direto com um *corpus* linguisticamente tão rico e complexo como o DANTEStocks. O trabalho com dados autênticos e a anotação/análise dos *tokens* foram marcados por desafios que contribuíram significativamente para a formação acadêmica da autora.

## 8. Referências Bibliográficas


BARBOSA, B. K. S. *Descrição sintático-semântica de nomes predicadores em tweets do mercado financeiro em português*. 2024. 163 f. Dissertação (Mestrado em Linguística) – Universidade Federal de São Carlos, São Carlos, 2024.

CARDOSO, P. C. S*. A conversação pública no Twitter: uma análise enunciativo-discursiva*. 2019. 265 f. Tese (Doutorado em Linguística) – Universidade Federal de Minas Gerais, Belo Horizonte, 2019.

DAMERAU, F. J. A technique for computer detection and correction of spelling errors. *Communications of the ACM*, v. 7, n. 3, p. 171–176, 1964.

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, [evento online]. *Anais* [...]. São Carlos: ICMC-USP, 2021. p. 335–343.

DI FELIPPO, A.; ROMAN, N. T. DANTEStocks: a multi-layered annotated corpus of stock market tweets for Brazilian Portuguese. *Brazilian Journal of Applied Linguistics*, Corpus Linguistics: Studies and Applications, p. 1–23, 2025. To Appear.

DURAN, M. S.; PARDO, T. A. S. Anotação de córpus, um lugar privilegiado de observação linguística: o estudo das aposições do português brasileiro segundo o modelo Universal Dependencies. *In*: ENCONTRO DE LINGUÍSTICA DE CORPUS, 16., 2024, Brasília. *Anais* [...]. Brasília, 2024. p. 118–123. Realizado em 21–24 out. 2024.

EISENSTEIN, J. What to do about bad language on the internet. *In*: CONFERENCE OF THE NORTH AMERICAN CHAPTER OF THE ASSOCIATION FOR COMPUTATIONAL LINGUISTICS – HUMAN LANGUAGE TECHNOLOGIES, 2013, Atlanta. *Proceedings* [...]. Atlanta: ACL, 2013. p. 359–369.

FOSTER, Jennifer. *Automatic Error Correction for Text***.** 2010. Tese (Doutorado em Computação) – Dublin City University, Dublin, 2010.

FREITAS, C. et al. Second HAREM: advancing the state of the art of named entity recognition in Portuguese. *In*: INTERNATIONAL CONFERENCE ON LANGUAGE RESOURCES AND EVALUATION, 17., 2020, Valletta. *Proceedings* [...]. Valletta: ELRA, 2020. p. 3630–3637.

FREITAS, T.; BARTH, F. Gênero textual e redes sociais: o caso do Twitter. *Revista Brasileira de Linguística Aplicada*, v. 15, n. 3, 2015.

JURAFSKY, D.; MARTIN, J. H. *Speech and Language Processing*: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition. 3. ed. Disponível em: https://web.stanford.edu/~jurafsky/slp3/. Acesso em: 1 mai. 2025.

KRUMM, J.; Davies, N.; Narayanaswami, C. User-Generated Content. *IEEE Pervasive Computing*, v. 7, n. 4, p. 10–11, 2008. DOI: 10.1109/MPRV.2008.85.

MARCUSCHI, Luiz Antônio. *Gêneros textuais: definição e funcionalidade*. São Paulo: Cortez, 2008.

McENERY, T. et al. *Corpus-Based Language Studies*: an advanced resource book. London: Routledge, 2006. 408 p.

MOTA, C., SANTOS, D. Desafios na avaliação conjunta do reconhecimento de entidades mencionadas: O Segundo HAREM, Linguateca, 2008.

NIVRE, J. et al. UNIVERSAL DEPENDENCIES v2: an evergrowing multilingual treebank collection. *In*: INTERNATIONAL CONFERENCE ON LANGUAGE RESOURCES AND EVALUATION, 12., 2020, Marseille. *Proceedings* [...]. Marseille: ELRA, 2020. p. 4034–4043.

PLUTCHIK, R.; KELLERMAN, H. (Eds.). *Emotion: Theory, Research and Experience*. Nova Iorque: Academic Press, 1986.

SANGUINETTI, M. et al. Treebanking user-generated content: a UD based overview of guidelines, corpora and unified recommendations. *Lang Resources & Evaluation*, Volume 57, Issue 2, p. 493–544, 2023. Springer-VerlagBerlin, Heidelberg. ISSN:1574-020X

SCANDAROLLI, C. L.; DI-FELIPPO, A.; ROMAN, N. T.; PARDO, T. A. S. 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. *Anais* [...]. Porto Alegre: SBC, 2023. p. 240-248.

SILVA, E. H.; PARDO, T. A. S.; ROMAN, N. T.; DI-FELLIPO, A. Universal Dependencies for *Tweets* in Brazilian Portuguese: Tokenization and Part of Speech Tagging. *In*: ENCONTRO NACIONAL DE INTELIGÊNCIA ARTIFICIAL E COMPUTACIONAL, 18, 2021 [online]. *Proceedings* [...]. Porto Alegre: SBC, 2021. p. 434-445.

SILVA, F. J. V.; ROMAN, N. T.; CARVALHO, A. M. B. R. 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.

WAGNER, R. A.; FISCHER, M. J. The String-to-String Correction Problem. Journal of the ACM, v. 21, n. 1, p. 168–173, 1974.

ZAPPAVIGNA, M. *Discourse of Twitter and social media*: How we use language to create affiliation on the web. London: Continuum, 2012.

ZEMAN, D. et al. CoNLL 2017 Shared Task: multilingual parsing from raw text to Universal Dependencies. *In*: CONFERENCE ON COMPUTATIONAL NATURAL LANGUAGE LEARNING, 2017, Vancouver. *Proceedings* [...]. Vancouver: ACL, 2017. p. 1–19.

ZERBINATI, M. M., ROMAN, N. T., DI-FELIPPO, A. A corpus of stock market tweets annotated with named entities. In: INTERNATIONAL CONFERENCE ON COMPUTATIONAL PROCESSING OF PORTUGUESE, 16, 2024, Santiago de Compostela, Galicia/Espanha. Proceedings [...]. Santiago de Compostela: ACL, 2024. P. 276–284.