Tipologia de fenômenos ortográficos e lexicais em CGU: o caso dos *tweets* do mercado financeiro

Clarissa Lenina Scandarolli[1,2], **Ariani Di Felippo**[1,2], **Norton Trevisan Roman**[1,3]
Thiago A. S. Pardo[1,4]

[1]Núcleo Interinstitucional de Linguística Computacional – NILC
[2]Departamento de Letras – Universidade Federal de São Carlos – UFSCar
Caixa Postal 676 – CEP 13565-905 – São Carlos – SP – Brasil
[3]Escola de Artes, Ciências e Humanidades –Universidade de São Paulo (USP)
CEP 03828-000 – São Paulo – SP, Brasil
[4]Instituto de CiênciasMatemáticas e de Computação – Universidade de São Paulo (USP)
Caixa Postal 668 –13566-970 – São Carlos –SP –Brasil

clarissa.scandarolli@estudante.ufscar.br, arianidf@gmail.com,
norton@usp.br, taspardo@icmc.usp.br

Abstract. *Twitter is an attractive source of information for several Natural Language Processing (NLP) applications, especially sentiment analysis and opinion mining. In this paper, we present a systematic description of orthographic and lexical phenomena in a corpus of tweets from the stock market domain in Portuguese. As a result, we propose a typology of the phenomena that could support the definition of annotation guidelines for their treatment within the Universal Dependencies framework of syntactic analysis and the development of NLP applications that realize term disambiguation or probabilistic ordering of options, as is the case with suggestions presented to users by spelling checkers.*

Resumo. *Twitter é uma fonte atrativa de informação para várias aplicações do Processamento Automático das Línguas Naturais (PLN), especialmente análise de sentimento e mineração de opinião. Neste artigo, apresenta-se uma descrição de fenômenos ortográficos e lexicais em um corpus de tweets do mercado financeiro em português. Como resultado, propõe-se uma tipologia dos fenômenos que pode auxiliar na definição de diretrizes de anotação segundo o modelo gramatical Universal Dependencies e no desenvolvimento de aplicações de PLN que façam a desambiguação de termos ou a ordenação probabilística de opções, como ocorre com a escolha das sugestões ortográficas apresentadas ao usuário em um corretor ortográfico.*

1. Introdução

O *Twitter* é uma fonte de informações valiosas para diferentes segmentos da sociedade devido principalmente à influência dessas informações. Por conseguinte, aplicações linguístico-computacionais (p.ex.: análise de sentimento e mineração de opinião) que processam o conteúdo gerado pelos usuários (CGU) do *Twitter* têm sido muito desenvolvidas no Processamento Automático das Línguas Naturais (PLN) [Sanguinetti et al. 2022]. E esse desenvolvimento é desafiador devido à linguagem não-padronizada

dos *tweets*, que pode ter sentenças agramaticais, sequências de sintagmáticas curtas, palavras com ortografia não convencional e expressões específicas de domínio. Para o desenvolvimento das aplicações, já há etiquetadores morfossintáticos (*taggers*) e analisadores sintáticos (*parsers*). Tal ferramental, aliás, tem sido construído com base nos *treebanks* ou *corpora* anotados (comumente com informações morfossintáticas e sintáticas) [Sanguinetti et al. 2022]. Os *tweebanks* mais recentes possuem anotação segundo o modelo gramatical *Universal Dependencies* (UD) [Nivre et al. 2016].

Motivados pela necessidade de criação de diretrizes para a anotação-UD de *tweebanks*, autores como Sanguinetti et al. (2022) focaram em descrever as idiossincrasias linguísticas mais gerais dos CGUs, propondo uma tipologia. Isso porque, mesmo o CGU sendo um contínuo de subdomínios textuais que variam de acordo com (i) convenções e limitações específicas impostas pela plataforma utilizada (como blog, fórum de discussão, chat online, microblog, etc.), (ii) grau de "canonicidade" em relação a uma linguagem mais padronizada e (iii) dispositivos linguísticos adotados para transmitir uma mensagem, há fenômenos comuns a esse espectro. Embora haja fenômenos comuns aos diferentes tipos de CGUs, a linguagem pode ser fortemente marcada pelo domínio (ou assunto) do *corpus*.

Assim, apresenta-se aqui a descrição das características ortográficas/gráficas e lexicais do *corpus* DANTEStocks[1], que engloba 4.048 *tweets* em português sobre o mercado financeiro. Acredita-se que a tipologia resultante da sistematização dos fenômenos pode auxiliar no processo de normalização dos *tweets*, desenvolvimento de aplicações multigênero ou de uso geral que não requerem normalização e na definição de diretrizes de anotação segundo o modelo gramatical UD.

2. Trabalhos relacionados

Estudos sobre variantes ortográficas da língua padrão têm longa tradição no PLN, sobretudo devido às aplicações de correção ortográfica. Muitas das pesquisas se baseiam nas 4 categorias de desvios de Damerau (1964) (inserções, exclusões, substituições e transposições de letras). Com o objetivo de verificar se essas classes se aplicavam ao português, Gimenes et al (2014), por exemplo, investigaram um *corpus* de *blogs* de viagens e comentários e, além das 4 categorias de Damerau, identificaram 3 categorias extras: erros no uso de diacríticos, erros no uso da cedilha e erros relacionados à espaço.

Sobre os *tweets* e gêneros similares, Bertaglia (2017), por exemplo, visando à construção de ferramentas de normalização para UGC, investigou um *corpus* em português composto por *tweets*, postagens de um fórum de discussão e análises de produtos. O autor identificou 3.699 palavras distintas que não constavam em um dicionário de referência e anotou essas palavras em função de 8 categorias de desvios da língua padrão: (i) erro ortográfico (e de digitação), (ii) acrônimo, (iii) abreviação, (iv) internetês, (v) estrangeirismo, (vi) unidade de medida, (viii) nome próprio, e (vii) sem categoria (isto é, *tokens* cuja classificação não é clara ou varia conforme o contexto). Sanguinetti et al. (2022) propuseram uma sistematização das particularidades identificadas em *corpora* majoritariamente compostos por *tweets* com base em 2 dimensões: canonicidade e intencionalidade. Por "canonicidade", entende-se a

[1] https://drive.google.com/file/d/1wr9M4czkPgkUj1--U9GT9h8ncXc6rzv4/view?usp=sharing

propriedade de um fenômeno ocorrer na língua padrão ou não. "Intencionalidade" se refere ao fato do fenômeno ter sido produzido deliberadamente ou não. Na hierarquia dos autores, "marcas de expressividade", por exemplo, são um tipo de fenômeno não-canônico e intencional, com os subtipos: (i) reduplicação de pontuação ("!"→"!!!"), (ii) alongamento grafêmico ("linda"→"linnda"), (iii) *emoticons* (":-)") e (iv) *emojis* ("❤").

A seguir, apresenta-se o *corpus* DANTEStocks[2], que foi alvo deste trabalho.

3. O *corpus* DANTEStocks

O DANTEStocks é um *corpus* de UGC em português composto por *tweets* sobre o mercado financeiro. Ele resultou do refinamento e da anotação morfossintática do *corpus* de Silva et al. (2020), cuja compilação se baseou na ocorrência de menos um *ticker*[3] de uma das 73 ações do IBovespa (principal indicador de desempenho das ações negociadas na B3). Atualmente, o DANTEStocks possui 4.048 *tweets* (~81 mil *tokens*), os quais não foram submetidos nenhuma normalização e, por terem sido compilados em 2014, têm no máximo 140 caracteres. Quanto à estrutura, o *corpus* engloba *tweets* com diferentes constituições internas, podendo apresentar (i) uma ou mais sentenças bem delimitadas (1) e (2), (ii) ausência de pontuação (3) ou pontuação equivocada (4), (iii) fragmentação (5), e (iv) colagens de manchetes de outras fontes (6) [Di-Felippo et al. 2021].

(1) Sera k petr4 já entrou na baixa?
(2) PETR4 subiu na bolsa 13,50. Muito bem, surpreso com o resultado.
(3) #PT conseguiu fazer propaganda eleitoral antecipada O que a @user[4] 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.

O DANTEStocks possui anotação de emoção, realizada manualmente com base nos 4 eixos de oposição emocional da teoria de *Plutchik* [Plutchik e Kellerman 1986] (*joy* vs *sadness*, *anger* vs *fear*, *trust* vs *disgust* e *surprise* vs *anticipation*) [Silva et al. 2020]. O *tweet* (1), por exemplo, recebeu os seguintes rótulos para 3 dos pares emocionais: *joy*, *trust* e *surprise*. O DANTEStocks também possui anotação semiautomática em nível morfológico segundo a UD, na qual se especificaram o lema, a etiqueta morfossintática e os traços lexicais/gramaticais (*features*) das palavras. O outro nível de anotação, no qual se explicitam as relações sintáticas de dependência (*deprels*), ainda não foi anotado. Na Figura 1, ilustra-se a anotação-UD completa de um *tweet* do *corpus* com base em Sanguinetti et al. (2022). Nessa figura, as etiquetas morfossintáticas (*part-of-speech* ou *PoS*)[5] estão em caixa alta, como NOUN para "acordo". Acima, estão os lemas, como "voo" para "voos". As *deprels* estão indicadas por setas rotuladas que se originam no *head* e se destinam ao dependente. Na figura, "acordo" é dependente de

[2] https://drive.google.com/file/d/1wr9M4czkPgkUj1--U9GT9h8ncXc6rzv4/view?usp=sharing
[3] Em (1), por exemplo, o *ticker* "petr4" indica ações preferenciais da Petrobras.
[4] As menções aos usuários do *Twitter* foram anonimizadas.
[5] A versão 2.0 da UD dispõe de 17 *tags* de *PoS* e de critérios para o emprego/anotação de cada uma delas.

"assinou" e estes estão conectados pela *deprel*[6] *obj* (objeto direto[7]). O verbo "assinou" é o *root* dessa representação. Os traços não constam na Figura 1, mas, segundo a UD, "acordo", por exemplo, tem os traços-valores: Gender=Masc e Number=Sing.

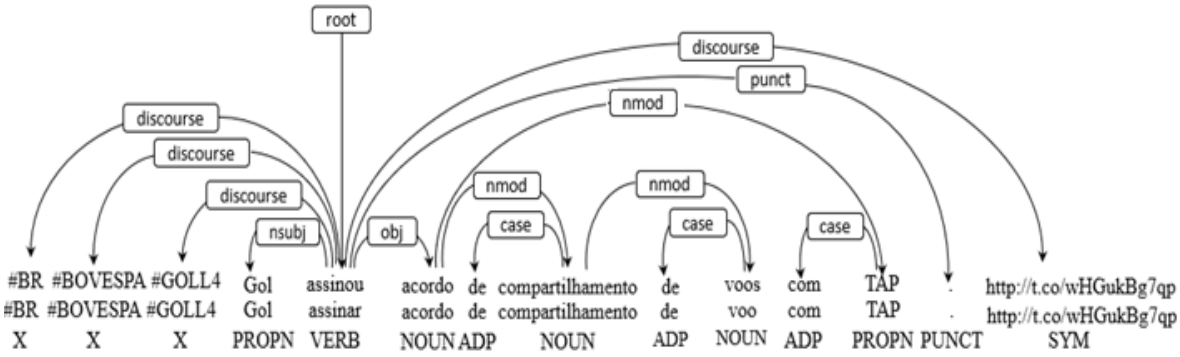

Figura 1. Exemplo de *tweet* do DANTEStocks com anotação-UD.

3. Procedimento metodológico

A identificação das particularidades gráficas/lexicais foi feita a partir da anotação semiautomática de PoS. Em linhas gerais, o *corpus* foi automaticamente anotado pelo *parser* UDPipe2 [Straka 2018] e, na sequência, submetido à revisão manual de 3 anotadores humanos diferentes, sendo que os casos de divergência entre os anotadores foram adjudicados por uma linguista sênior. Especificamente, a revisão manual das etiquetas foi feita em duas etapas. Na primeira, os anotadores humanos identificaram apenas as classes gramaticais das palavras conhecidas (isto é, pertencentes ao vocabulário da língua geral) com base em manuais que contém diretrizes para a anotação UD do português e para os fenômenos típicos dos *tweets*, e assinalaram, com a etiqueta genérica *Typo=Yes,* todos os *tokens* que possuíam algum tipo de variação de forma frente à grafia padrão ou que não estavam presentes em dicionários da língua geral. Na segunda etapa, todos os casos de *Typo=Yes* foram analisados e anotados com suas respectivas etiquetas PoS. Isso foi feito porque, para a maioria dos casos de *Typo=Yes*, ainda não havia diretrizes de anotação-UD e estas precisaram ser identificadas na literatura ou desenvolvidas para o *corpus* em questão.

Assim, a identificação das particularidades gráficas e lexicais do DANTEStocks foi feita com base nos 1.363 *tokens* anotados com *Typo=Yes.* Esses casos foram organizados em uma tabela no formato .xls e cada caso analisado individualmente, buscando-se identificar classes ou categorias de fenômenos.

4. A tipologia de fenômenos ortográficos e lexicais

A natureza dos fenômenos presentes nos 1.363 *tokens* levou à identificação de 2 dimensões: "Norma[8] Padrão" e "Norma Inovadora". A Figura 2 exibe a organização hierárquica das idiossincrasias do DANTEStocks nessas duas dimensões.

---

[6] A UD 2.0 provê 37 *deprels* e critérios para o emprego de cada uma delas.

[7] Relação entre o predicado verbal e o segundo argumento *core* do verbo (o primeiro é *nsubj*).

[8] Por "norma", entende-se "o conjunto de fatos linguísticos que caracterizam o modo como normalmente falam as pessoas de certa comunidade" [Faraco 2008, pág 40].

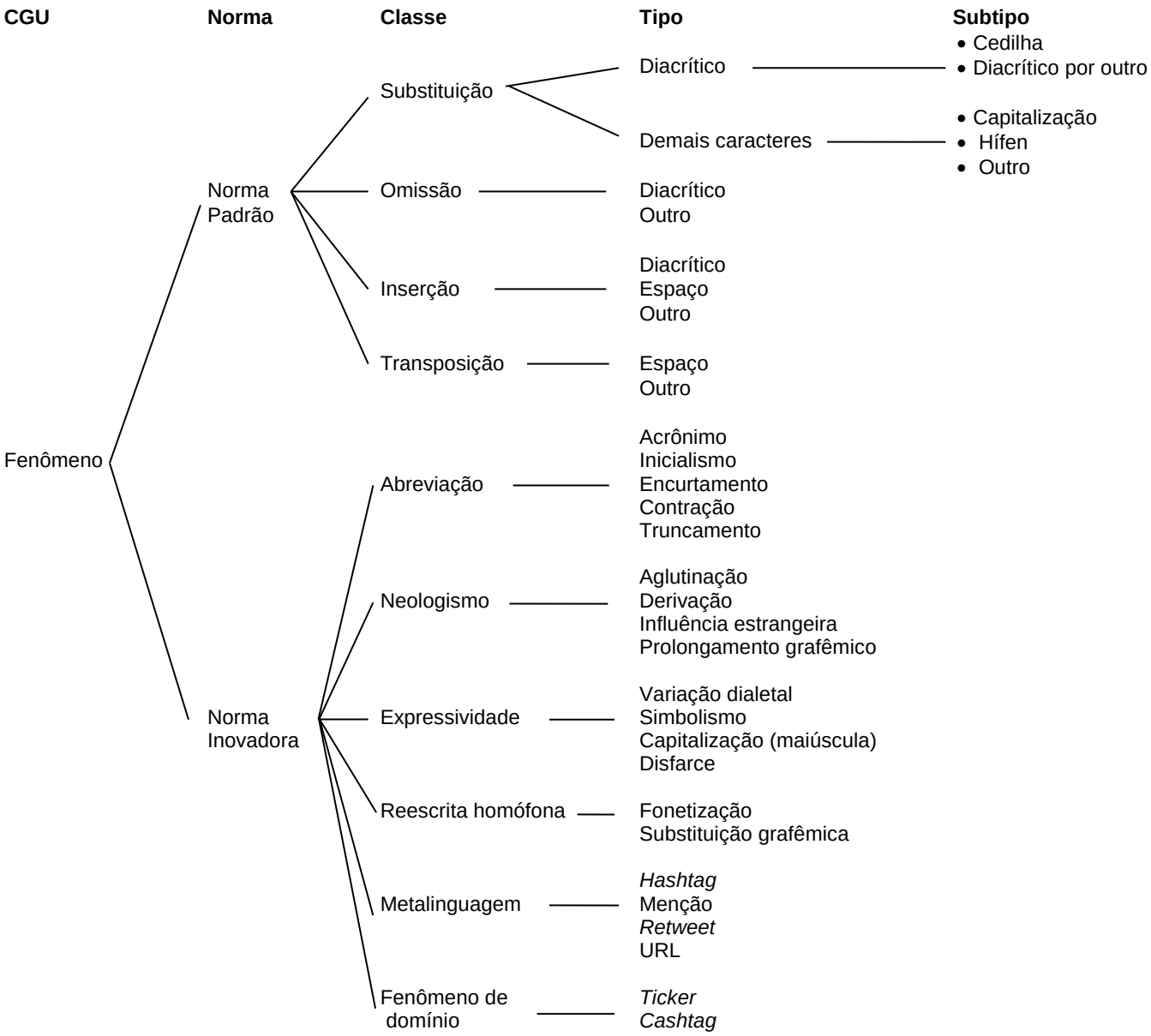

Figure 2. Tipologia de idiossincrasias gráficas/lexicais do DANTEStocks.

A "Norma Padrão" engloba fenômenos considerados desvios da norma-padrão por diversos motivos (como desconhecimento da ortografia, influência do meio e dispositivo, influência de novas regras fonéticas, etc.). As categorias dessa dimensão foram definidas com base no conceito de "caractere" do padrão Unicode[9,10]. Nesse padrão, letras com diferentes capitalizações ("a" e "A"), diferentes acentuações ("o" e "ó"), diacríticos (em isolado) ("~") e o próprio espaço são caracteres diferentes e, por isso, representados por códigos únicos. 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 "á" pode ser concebido como a composição de 2 *code points* (p. ex.: ao se digitar "´" + "a", obtém-se o "á"), o qual, nos algoritmos de normalização do Unicode, é

[9] http://www.unicode.org/standard/WhatIsUnicode.html

[10] 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.

convertido para um *code point* único. Sendo assim, a adoção do caractere permite identificar e classificar as variações gráficas com base em um critério genérico, abrangente e concreto, além de amplamente empregado na Computação. Aplicando o conceito de caractere às categorias de Damerau, os fenômenos dessa dimensão foram organizados em classes, tipos e subtipos. Ressalta-se que um mesmo *token* pode apresentar mais de um fenômeno da Norma Padrão. A "Norma Inovadora" engloba fenômenos que são empregados de modo a concorrer com outras palavras da linguagem-padrão para expressar um mesmo conceito ou de modo a expressar um conceito novo. De certa forma, essa norma se relaciona às "variantes linguísticas" de uma comunidade de fala, as quais, necessariamente, não estão contempladas na norma-padrão, pois são resultados da utilização de recursos ortográficos de forma criativa e inovadora.

1) Norma Padrão

- *Substituição:* ocorre quando ao menos um caractere (diácritico ou não) de um *token* é substituído por outro, ocasionando um erro da ortografia padrão. A substituição de diacrítico pode ser de 2 tipos: (i) cedilha, como "acougue" ("açougue"), pois, embora haja um *code point* único para o diacrítico do cedilha, este é indissociável da letra "c", e (ii) diacrítico por outro, como "mâe" ("mãe"). Os demais casos de substituição envolvem (i) capitalização (maiúscula e minúscula), como "dilma" ("Dilma"), (ii) substituição de hífen por espaço (e vice-versa), como "cruz credo" ao invés de "cruz-credo", ou (iii) outro caractere, como ocorre em "hirário" ao invés de "horário".
- *Omissão:* ocorre quando um caractere deixa de ser expresso. Uma omissão pode ser relativa a (i) diacrítico, pois o usuário deveria ter digitado duas ou mais teclas para compor o caractere, mas não o fez, como "esta" ("está"), ou (ii) demais caracteres, como a ausência do *s* plural no final de "açõe" ("ações").
- *Inserção:* ocorre quando um caractere é inserido da palavra. Uma inserção pode ser relativa a (i) diacrítico, como "Petrobrás", quando o correto seria "Petrobras", (ii) espaço, como "a final" ao invés de "afinal", e (iii) outro caractere, como "Streaddle", quando o correto é "Straddle".
- *Transposição:* ocorre quando um caracter é trocado de ordem com outro. Uma transposição pode ser de (i) espaço, como "meua migo", quando o correto é "meu amigo", ou de (ii) demais caracteres, como "acrodo" ao invés de "acordo".

2) Norma Inovadora

- *Abreviação:* fenômeno que gera um *token* mais curto do que a palavra ou expresão que lhe deu origem, podendo ser: (i) acrônimo, isto é, *token* composto pelas letras iniciais ou sílabas de uma multi-palavra e que tem pronúncia de palavra única, como "Cemig" ("Companhia Energética de Minas Gerais)"; (ii) inicialismo, que se observa em um *token* composto pelas letras iniciais de uma multi-palavra e que é prounciado letra por letra, como"lp" ("longo prazo"); (iii) encurtamento, isto é, ausência das letras finais de *token*, como "q" ("que"), (iv) contração, observado em um *token* com letras intermediárias ausentes, como "enqt" ("enquanto"), e (v) truncamento, isto é, *token* quebrado que, no caso do DANTEStocks, ocorre no final do *tweet,* comumente seguido por reticências, e que se deve ao limite de caracteres, como "divu" ("divulgou") no exemplo (6).

- *Neologismo:* resulta em uma palavra nova ainda não institucionalizada (isto é, não abonada e incluída em dicionário), podendo ser de 3 tipos: (i) aglutinação, que se observa em um *token* resultante da junção de 2 palavras, como "Ibolixo" ("Ibovespa" + "lixo"); (ii) derivação, que resulta da adição de um afixo a uma radical já existente, como "diretassa" ("direta" + "-assa (-aça)") e (iii) influência estrangeira, que se observa em uma palavra formada com base em outra língua, como "estopar", que provém do verbo em inglês *"stop"* ("parar") e significa "interromper venda ou compra de um ativo diante de dado preço".
- *Expressividade:* fenômeno que majoritariamente simula sentimento expresso pela prosódia, expressão facial ou gesto na interação direta, podendo ser: (i) prolongamento grafêmico, como "noossaaa" ("nossa"), (ii) variação dialetal, como "malmita" ("marmita"), (iii) simbolismo, isto é, ocorrência de um caractere simbólico (seja *emoticon*, *emoji*, *smiley* ou outro) em substituição a uma palavra ou parte dela, (iv) capitalização, como "FEIO" no *tweet* (7) "#btow3: eita papel FEIO. #goll4 de olho na média móvel", e (v) disfarce, que é a substituição de uma ou mais letras por um caractere especial para indicar autocensura, como "m*" ("merda").
- *Reescrita homófona:* refere-se a uma variação gráfica motivada pela fonética ou pela simplificação de diacríticos, podendo ser: (i) fonetização, que é a representação da fala na escrita, como "krai" ("caralho"), e (ii) substituição grafêmica, que é o uso de uma letra a mais em substituição a um diacrítico, como "neh" ("né") e "tou" ("tô").
- *Metalinguagem*: corresponde a todo *token* que tipicamente ocorre no *Twitter* e que, por isso, não está previsto em dicionários, como (i) *hashtag* (pe.x.: #PT em (3)), (ii) menção, como se observa em (3), (iii) marca de *retweet* (RT), como no *tweet* (8) "Região 24,60 a 24,65 RT @Live_Trade: Fibr3 observo p/ compra", e (iv) URL, que se observa nos *tweets* (5) e (6).
- *Fenômeno de domínio:* todo *token* que ocorre recorrentemente em *tweets* do mercado financeiro, como os *tickers* (p.ex.: em (1), (2) e (6)) e *cashtags* como no *tweet* (9) $PETR3 - Petrobras (petr) - Comunicado http://t.co/mHuCIyQmFi.

5. Considerações finais

Atualmente, tem-se definido o conjunto de diretrizes e *tags* para a anotação dos fenômenos da Figura 2. Tendo em vista a adoção do modelo UD, as *tags* estão sendo propostas em inglês.

Para os fenômenos da dimensão denominada Norma Padrão, objetiva-se adotar a *tag* `Typo=Yes` na coluna FEATS (destinada a atributos morfológicos) do formato CoNLL-U[11] como indicado pela própria UD e, na coluna MISC (reservada para demais informações e cujas *tags* podem ser definidas para um *treebank* específico), poder-se-á empregar uma *tag* adicional, como [`SNorm`:standard norm], com os seguintes valores possíveis: [`Sub`=substitution, `Om`=omission, `In`=insertion, `Tr`=transposition, `Ot`=other]. A coluna MISC pode conter ainda outra *tag* para esses fenômenos, como [`Type`:type], a qual especificaria o desvio. Essa etiqueta poderia ter os valores [`Ced`=cedilla, `Dia`=diacritic, `Cap`=capitalization, `Hif`=hifen, `Spc`=space, `Ot`=other].

Quanto a classe das abreviações (Norma Inovadora), objetiva-se seguir a diretriz geral da UD que prevê o emprego da *tag* `Abbr=Yes` na coluna FEATS. Ademais,

[11] https://universaldependencies.org/format.html.

pretende-se empregar, na coluna MISC, uma *tag* adicional como [`INorm`: innovative norm], cujos valores possíveis correspondem aos 5 tipos de abreviações, que são: [`Acr`=acronym, `Init`=initialism, `Short`=shortening, `Cont`=contraction, `Trunc`=truncation]. Para os demais fenômenos da Norma Inovadora, pretende-se anotá-los por meio da *tag* `INorm` na coluna MISC, cujos valores possíveis representam os 16 tipos de fenômenos, a saber: [`Aggl`=agglutination, `Der`=derivation, `Fgn`=foreign, `Ext`=graphemic stretching, `Dial`: dialectal variation, `Sym`: symbolism, `Upp`: uppercase, `Dis`: disguise, `Fon`: fonetization, `Subst`: graphemic substitution, `Hasht`: hashtag, `Me`: mention, `Ret`: retweet, `URL`: URL, `Tic`: ticker, `Casht`: cashtag].

Uma vez que o conjunto de etiquetas estiver de fato definido, o DANTEStocks será inteiramente anotado, revisando os casos iniciais que deram origem à tipologia, e identificando outros que porventura não estavam na lista inicial. Na sequência, pretende-se fazer um levantamento estatístico dos fenômenos/*tags*, gerando uma caracterização do domínio/*corpus*. Validar a taxonomia em outro *corpus* de *tweets* é uma possibilidade de trabalho futuro. Por fim, ressalta-se que a descrição dos fenômenos ora apresentada não só pode contribuir para a definição de diretrizes de anotação-UD, mas também para que aplicações de PLN possam levar em conta a distribuição desses fenômenos, seja de forma geral ou em algum gênero ou domínio específico, de modo a permitir, por exemplo, a desambiguação de termos, ou a ordenação probabilística de opções, como ocorre com a escolha das sugestões ortográficas apresentadas ao usuário em um corretor ortográfico (p.ex.: [Gimenes et al. 2015]).

Agradecimentos. Este trabalho foi realizado no âmbito do Centro de Inteligência Artificial da Universidade de São Paulo (C4AI - http://c4ai.inova.usp.br/), com o apoio da Fundação de Amparo à Pesquisa do Estado de São Paulo (processo FAPESP #2019/07665-4) e da IBM. Este projeto também foi apoiado pelo Ministério da Ciência, Tecnologia e Inovações, com recursos da Lei N. 8.248, de 23 de outubro de 1991, no âmbito do PPI-Softex, coordenado pela Softex e publicado como Residência em TIC 13, DOU 01245.010222/2022-44

Referências

Bertaglia, T.F.C. (2017). Normalização textual de conteúdo gerado por usuário. Dissertação, Instituto de Ciências Matemáticas e de Computação, USP, São Carlos.

Damerau, F. J. (1964). A technique for computer detection and correction of spelling errors. Communications of the ACM, 7(3):171–176.

Di-Felippo, A.; Postali, C.; Ceregatto, G.; Gazana, L.S.; Silva, E.H.; Roman, N.T.; Pardo, T.A.S. (2021). Descrição preliminar do *corpus* DANTEStocks: diretrizes de segmentação para anotação segundo Universal Dependencies. In the Proceedings of the 7th Workshop on Portuguese Description (JDP), p. 335-343.

Faraco, C. A. (2008). Norma culta brasileira: desatando alguns nós. SP: Parábola Editorial.

Gimenes, P., Roman, N. T., Carvalho, A. M. B. R. (2015). Spelling error patterns in Brazilian Portuguese. Computational Linguistics, 41(1): 175–183.

Luotolahti, J., et al. (2015). Towards universal web parsebanks. In the Proceedings of the 3rd Depling 2015, p. 211–220. Uppsala University.

Nivre, J. et al. (2016). Universal Dependencies v1: a multilingual treebank collection. In the Proceedings of the 10th LREC, p.1659-66. Portorož. ELRA

Plutchik R., Kellerman, H. (ed.) (1986) Emotion: theory, research and experience. NY: Acad. Press.

Sanguinetti, M., Bosco, C., Cassidy, L., Çetinoğlu, Ö., Cignarella, A.T., Lynn, T., Rehbein, I. Ruppenhofer, J., Seddah, D., Zeldes, A. (2020). Treebanking user-generated content: a proposal for a unified representation in universal dependencies. In the Proceedings of the 12th LREC. p. 5240-50. Marseille, France. ELRA

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

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