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

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

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

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

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

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

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

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

## 1. Introdução

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

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

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

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

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

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

---

[1] Palavra sintática (em inglês, *syntactic word*) é a unidade mínima a que corresponde uma função sintática (https://universaldependencies.org/u/overview/tokenization.html).

[2] Na anotação UD, palavras sintáticas (ou itens lexicais) são sinônimos de *tokens*.

## 2. O Modelo *Universal Dependencies*

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

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

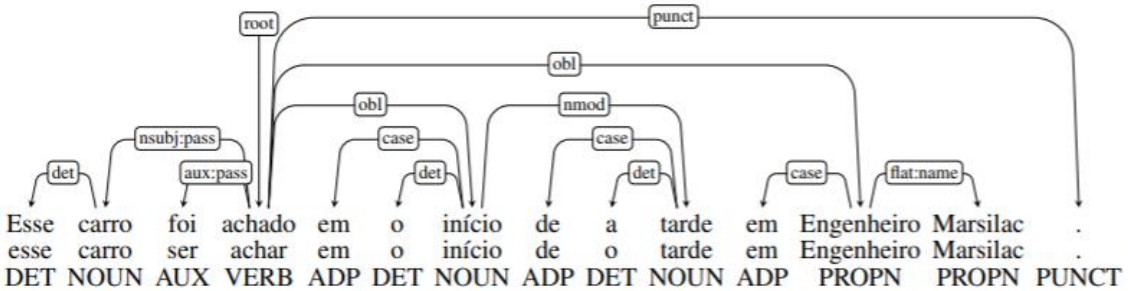


**Figura 1. Exemplo de anotação sintática segundo a UD [Rademaker *et al*., 2017].**

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

---

[3] Para o português do Brasil, Duran (2021) construiu um manual com diretrizes de tokenização e de anotação morfossintática segundo a UD (especificamente para textos formais, como os jornalísticos).

[4] https://universaldependencies.org/guidelines.html

[5] Essa informação foi recuperada do *corpus UD-Portuguese-Bosque* por meio da plataforma *online* Grew-match (http://match.grew.fr/?corpus=UD_Portuguese-Bosque@2.8).

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

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

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

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

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

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

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

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

## 4. Os Fenômenos UGC do DANTEStocks e a sua Tokenização

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

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

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

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

**Quadro 1. Exemplo dos fenômenos UGC no DANTEStocks.**

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

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

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

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

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

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

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

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

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

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

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

## 5. Considerações finais

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

### Agradecimentos

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

## Referências

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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