Em Direção à Anotação Sintática – UD de Tweets do Mercado Financeiro

Bryan K. S. Barbosa[1], **Ariani Di Felippo**[1]

[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

bryankhelven@ieee.org, ariani@ufscar.br

Abstract. *Many corpora have recently been built based on the Universal Dependencies (UD) grammatical model, including tweebanks - corpora composed of tweets. Regarding the Portuguese language, there are only guidelines for UD annotation of the general aspects of this language. In this article, guidelines for the syntactic annotation, according to this model, of some aspects or linguistic phenomena identified in financial market tweets are presented. With this, it is sought to contribute to the elaboration of a syntactic annotation manual via UD for tweets and for the construction of the first Portuguese tweebank.*

Resumo. *Muitos córpus têm sido atualmente construídos com base no modelo gramatical Universal Dependencies (UD), inclusive os tweebanks – córpus compostos por tweets. No que diz respeito à língua portuguesa, já há diretrizes segundo esse modelo para anotação de textos que seguem a norma-padrão. Neste artigo, apresentam-se diretrizes para a anotação sintática-UD de alguns fenômenos linguísticos identificados em tweets do mercado financeiro, cuja linguagem se caracteriza pela fragmentação, informalidade e ocorrência de elementos veiculados à plataforma e ao domínio. Com isso, busca-se contribuir para a elaboração de um manual de anotação sintática via UD para tweets e para a construção do primeiro tweebank em português.*

1. Introdução

O Processamento de Línguas Naturais (PLN) tem usado amplamente o modelo gramatical *Universal Dependencies* (UD) [Nivre et al. 2020] na construção de *treebanks*, pois esse modelo estabelece diretrizes e rótulos "universais" para a anotação sintática de córpus em diferentes línguas e domínios. A anotação de córpus para fins de PLN, aliás, apresenta vários desafios. Um dos mais importantes é garantir que fenômenos iguais tenham o mesmo tratamento e que fenômenos distintos recebam etiquetas também distintas, aumentando, assim, a consistência e a qualidade das anotações para processos posteriores de aprendizado automático [Duran et al. 2022].

Embora a pesquisa em PLN envolvendo o português e o modelo UD ainda sejam consideradas incipientes, já há diretrizes para a anotação morfossintática e sintática dos aspectos gerais dessa língua [Duran et al. 2022, Duran 2022]. O mesmo, no entanto, não pode ser dito sobre a construção dos chamados *tweebanks* [Sanguinetti et al. 2017] (*treebanks* compostos por tweets), para os quais há apenas diretrizes de anotação morfossintática [Di-Felippo et al. 2021].

Assim, neste artigo, apresentam-se diretrizes para a anotação sintática-UD de alguns aspectos ou fenômenos linguísticos identificados no DANTEStocks[1], que é um córpus de 4.048 *tweets* do mercado financeiro em português. Entre os fenômenos, estão: URLs, *hashtags*[2], *cashtags*[3], truncamentos de palavras e frases, *emoticons/smileys*, menções, marcas de *retweet* e outros. Esse córpus, aliás, apresenta grandes desafios para o PLN, uma vez que sua linguagem difere muito da norma-padrão cujo processamento tem sido o foco da área. Esse distanciamento se deve ao grau de informalidade, fragmentação, ocorrência de terminologia e de elementos dependentes da plataforma ou meio.

Para tanto, organizou-se o artigo, além desta introdução, em cinco seções adicionais. Na Seção 2, apresentam-se os principais conceitos sobre o modelo UD e o córpus utilizado neste trabalho. Na Seção 3, descrevem-se as etapas metodológicas. Na Seção 4, apresentam-se as estatísticas dos fenômenos particulares identificados no córpus. Na Seção 5, estabelecem-se as respectivas propostas de anotação sintática-UD. Na Seção 6, algumas considerações finais são feitas, destacando as contribuições e limitações dos resultados, assim como enfatizando trabalhos futuros.

2. O modelo *Universal Dependencies* e o córpus DANTEStocks

O modelo UD resulta de um projeto colaborativo que busca desenvolver um modelo gramatical, por dependência, para a construção de córpus anotados em diferentes línguas. Esse modelo captura diversos fenômenos linguísticos de maneira consistente em diferentes línguas, permitindo a comparação e o contraste entre elas [Nivre et al. 2020]. No que tange aos *tweebanks*, a UD tem se tornado um referencial popular de anotação principalmente devido à sua adaptabilidade a diferentes domínios e gêneros. Quanto à anotação, a UD prevê 2 níveis. No nível morfológico, especificam-se 3 informações: lema, etiqueta morfossintática e traços lexicais e gramaticais (*features*). No nível sintático, a anotação se dá por relações de dependência (*deprels*). A representação básica de uma estrutura de dependências é arbórea, na qual uma palavra é o *root* (raiz) da sentença.

A versão atual do DANTEStocks possui apenas anotação semiautomática em nível morfológico segundo a UD. O outro nível de anotação, no qual se explicitam as *deprels*, ainda não foi anotado. Para tanto, este trabalho busca contribuir com as primeiras diretrizes relativas a este nível. Na Figura 1, ilustra-se a anotação-UD completa de um *tweet* do córpus com base em [Sanguinetti et al. 2022] e [Duran et al. 2022], na qual o verbo "indicado" é o *root* da representação. Nessa figura, as etiquetas morfossintáticas (*part-of-speech* ou PoS)) estão em caixa alta, como VERB para "indicado". Abaixo, estão os lemas, como "indicar" para "indicado". As *deprels* estão indicadas por setas rotuladas que se originam no *head* e se destinam ao dependente. Na Figura 1, o numeral "20,05" é dependente do símbolo "R$", os quais estão conectados pela *deprel* NUMMOD (modificador numérico). Os traços não constam na Figura 1, mas, segundo a UD, um verbo no particípio como "indicado", pode ser descrito pelos atributos-valores: VerbForm=Part, Gender=Masc e Number=Sing.

Ressalta-se que o DANTEStocks resulta de um refinamento e da anotação mor-

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

[2]Qualquer palavra/expressão precedida pelo # que funciona como indexador de conteúdo (#Petrobras).

[3]É o símbolo do registro de uma empresa precedido pelo $ ($PETR4). O clique em uma *cashtag* leva o usuário a outros *tweets* sobre esse mesmo símbolo de registro.

fossintática do córpus inicialmente construído por [Silva et al. 2020], cuja compilação foi feita com base na ocorrência de ao menos um *ticker* [4] de uma das 73 ações do IBovespa, que é o principal indicador de desempenho das ações negociadas na B3. Destaca-se também que os 4.048 *tweets* ($\sim$81 mil *tokens*) não foram submetidos a nenhuma normalização e, por isso, sua linguagem se distancia da língua-padrão. Ademais, por ter sido compilado em 2014, os *tweets* têm no máximo 140 caracteres. Quanto à estrutura, o córpus engloba *tweets* com diferentes estruturas internas, podendo apresentar (i) uma ou mais sentenças bem-delimitadas e, (ii) ausência de pontuação, (iii) pontuação equivocada e (iv) fragmentação [Di-Felippo et al. 2021].

3. Metodologia

Este trabalho foi equacionado em 4 etapas, a saber: (i) seleção dos dados de análise, (ii) investigação dos fenômenos, (iii) levantamento estatísticos dos fenômenos e (iv) proposição e exemplificação de estratégias de anotação sintática-UD.

A etapa (i) consistiu em selecionar uma parcela de *tweets* do DANTEStocks para que os primeiros fenômenos não cobertos pelas diretrizes gerais da língua portuguesa pudessem ser manualmente identificados. Para tanto, optou-se por selecionar os *tweets* iniciais do córpus até se obter o conjunto equivalente a 10% do total de *tweets* do DANTEStocks, ou seja, 405 mensagens. Ainda nessa etapa, a parcela do córpus selecionada foi transformada em uma estrutura de dados (*dataframe*) para que pudesse ser facilmente manipulada e analisada, permitindo a aplicação de técnicas de análise de dados, que incluem a filtragem e a quantificação de fenômenos específicos.

A etapa (ii) englobou duas atividades. A primeira foi o estudo de uma sistematização preliminar de alguns fenômenos de Conteúdo Gerado por Usuário (CGU) do DANTEStocks em classes, como expressão de sentimento (*emoticon, smiley* e prolongamento grafêmico), elemento metalinguístico (*hashtag*, marca de *retweet*, URL, menção e truncamento lexical) e fenômeno de domínio (*cashtag*) [Di-Felippo et al. 2021]. O estudo desse trabalho permitiu reconhecer essas particularidades e identificar outras distintas nos 405 *tweets* selecionados para análise. A segunda atividade dessa etapa consistiu em identificar, nos 405 *tweets* do *dataframe*, particularidades de linguagem não cobertas pelo manual de [Duran et al. 2022]. Essa atividade resultou na identificação dos seguintes fenômenos CGU, observados quanto à sua integração ou não à estrutura sintática dos *tweets*: URL, *hashtag*, *cashtag*, menção, *emoticons*, marcas de *retweets* (RT), truncamento de palavra e sentença.

A etapa (iii) consistiu em levantar a estatística de ocorrência dos fenômenos no conjunto de 405 postagens, aplicando filtros ao *dataframe*. Com isso, dá-se início a uma caracterização linguística do DANTEStocks, que será o primeiro *tweebank* com anotação-UD em portugês. As estatísticas de ocorrências estão descritas na próxima Seção (4).

Por fim, na etapa (iv), estratégias de anotação em nível sintático segundo a UD são propostas para os fenômenos CGU. Para cada uma das particularidades, buscou-se, na literatura sobre construção de *tweebanks* em outras línguas, por uma estratégia de anotação de *deprel* já definida e amplamente empregada. Diante de fenômenos CGU não

[4]Combinação composta por quatro letras e um número que refere-se tanto ao nome da empresa quanto ao tipo de ação, como "VALE5"

cobertos pela literatura, estratégias de anotação sintática foram especificamente definidas para o DANTEStocks. Tais propostas são apresentadas e ilustradas na Seção 5.

4. Estatística dos fenômenos CGU no conjunto de análise

Segundo os resultados exibidos nos Quadros 1 e 2, as URLs *standalone*[5] são os fenômenos mais frequentes no córpus de estudo, com 142 ocorrências, seguido pelas *hashtags* integradas à sintaxe, com 98 ocorrências. Enquanto as *hashtags* estão entre os fenômenos mais frequentes, as *cashtags* apresentam frequência relativamente baixa (28 ocorrências no total). A diferença de frequência entre ambas pode estar relacionada ao fato de que as *hashtags* são mais genéricas e populares, ao passo que as *cashtags* são indexadores específicos de empresas/ativos. Ademais, vale ressaltar que (i) marcas de *retweet* (RT), (ii) menções (integradas ou não), e (iii) truncamentos de sentença e de palavra apresentam frequências relativamente similares, com cerca 30 ocorrências de cada no córpus de estudo. As expressões onomatopeicas (no caso, risos) e os *emoticons*, por sua vez, são relativamente raros, com 8 e 6 ocorrências, respectivamente. Uma possível explicação para isso pode ser o fato de que os *tweets* do DANTEStocks, por serem predominantemente informativos, não veiculam muitas expressões de sentimento.

Quadro 1. Frequência de Fenômenos CGU no córpus de estudo.

Fenômeno CGU	Frequência
URL *standalone*	142
URL integrada	47
Hashtag standalone	77
Hashtag integrada	98
Cashtag standalone	26
Cashtag integrada	2
Menção *standalone*	31
Menção integrada	30
Emoticon	6
RT *standalone*	31
Expressão Onomatopeica (riso)	8
Truncamento de sentença	34
Truncamento de palavra	23

Uma vez que os fenômenos foram identificados, passou-se à etapa (iv), em que estratégias para a anotação sintática (isto é, de *deprels*) desses fenômenos foram investigadas na literatura e/ou propostas. Na próxima seção, apresentam-se tais estratégias. Para tanto, cada uma delas é ilustrada com a anotação completa de um *tweet* do córpus no qual o fenômeno correspondente ocorre. A anotação sintática dos aspectos relacionados à norma-padrão contidos nos *tweets*-exemplo foi feita com base no manual para a língua portuguesa de [Duran et al. 2022]. Ademais, a anotação dos *tweets*-exemplo foi revisada por 3 anotadores humanos, os quais utilizaram o Arborator-Grew-NILC[6], que

[5]Neste trabalho, adota-se o termo *standalone* para classificar os fenômenos CGU não-integrados à sintaxe, conforme sugerido em [Sanguinetti et al. 2022].

[6](https://arborator.icmc.usp)

é uma versão expandida e aprimorada da ferramenta web para anotações de sintaxe de dependências de [Guibon et al. 2020].

5. Diretrizes de anotação sintática-UD para os fenômenos CGU

Para a apresentação das diretrizes de anotação, segue-se a ordem de frequência dos fenômenos apresentada no Quadro 1.

5.1. URL

Segundo o Quadro 1, uma *URL* pode ocorrer integrada à sintaxe ou *standalone*. Com base em [Liu et al. 2018], [Sanguinetti et al. 2020] e [Sanguinetti et al. 2022], fenômenos integrados à sintaxe devem ser anotados pela relação de dependência (*deprel*) que representa sua posição ou função sintática. No caso de uma URL integrada, ela pode ocorrer precedida de preposição (1) ou de dois-pontos (2). Em alguns *tweets*, pode-se inferir a ocorrência de uma preposição como em (3). Nesses casos, se o *head* for um verbo, a URL será conectada a ele por **obl**, acrescida da sub-relação url, como na Figura 1. Caso o *head* não seja um verbo (4), a URL será conectada a ele por **parataxis**[7], acrescida de **url**. Mesmo que haja outra opção na literatura, como o uso de *LIST*[8](*nota*) [Silveira et al. 2014], por exemplo, opta-se aqui por também empregar **parataxis** para a anotação de uma URL em ocorrência *standalone*, a qual pode ocorrer após ponto final (5) ou reticências (6), como ilustrado pela anotação do *tweet* (5) na Figura 2.

(1) Publiquei estudo da #HGTX3 no gráfico diário. Rompendo tendência de baixa??? Veja em **http://t.co/oRA4bA8Qye**
(2) Nos últimos 5 pregões #CSNA3 acumulou uma baixa de 17.1% enquanto #USIM5 -14.8% e #VALE5 -10%. Veja o ranking: **http://t.co/XjRazUAN9b**
(3) BBAS3 comprar por R$ 20,05 indicado em 27/02/2014 10:41 [em] **http://t.co/zJRs3Eeyz9**
(4) Macktrader Investimentos: Banco do Brasil On (Bbas3), Gráfico Diário. **http://t.co/9pBbMOk8Nh**
(5) @garimpodeacoes $RSid3 4Q13 (N) Geração Op de Caixa forte, com desalavancagem financeira. Margens se recuperam YoY. http://t.co/CC8MGagICJ
(6) Petrobrás Pn (Petr4), Gráfico Diário. Ação registr... **http://t.co/Zsml5piTaT**

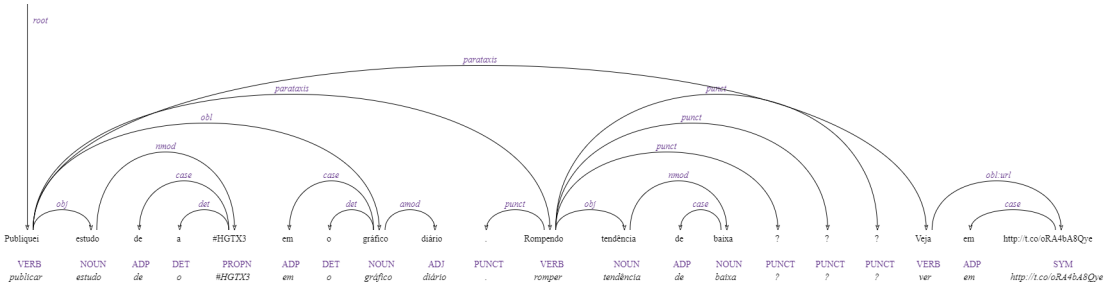

Fig. 1. Exemplo de URL integrada com preposição explícita e anotada com obl:url.

[7] *Deprel* que ocorre entre dois elementos da sentença que poderiam ter relação sintática entre si, porém essa relação não está explicitada.

[8] *Deprel* que ocorre entre os elementos que compõem uma lista.

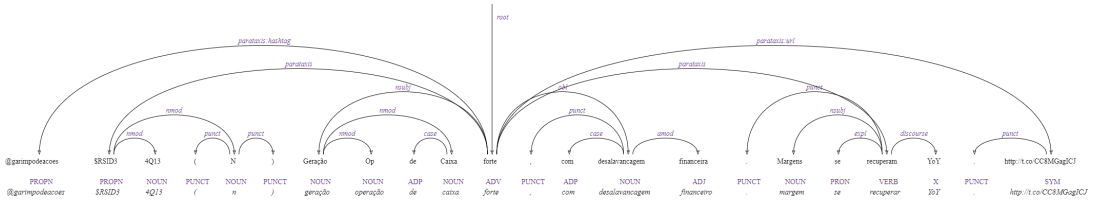

Fig. 2. Exemplo de URL standalone precedida de ponto final e anotada com parataxis:url.

5.2. Hashtag/Cashtag

Assim como as URLs, as *hashtags* e *cashtags* podem ocorrer integradas à estrutura sintática dos *tweets* ou em *standalone*. Quando integradas, devem ser anotadas com base na sua função/posição sintática. Na Figura 3, por exemplo, a *hashtag* "#OIBR4" foi conectada ao *head* por **obl:hashtag** e, na Figura 4, a *cashtag* "$PETR4" foi conectada ao *head* por **nsubj** (nesse caso, assumiu-se que o verbo de cópula está elíptico, isto é, "$PETR4 [está] nesse instante aos 13,32").

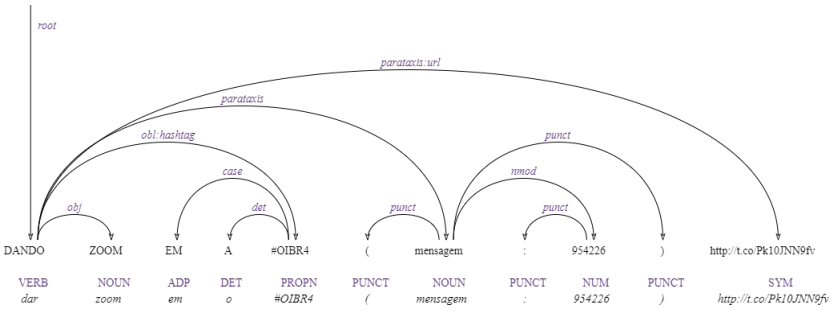

Fig. 3. Exemplo de hashtag integrada anotada com obl:*hashtag*.

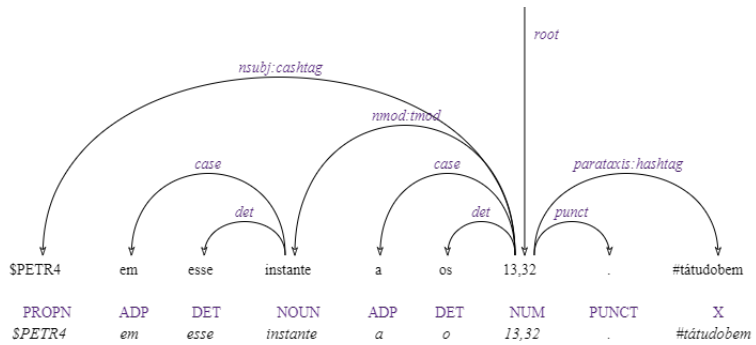

Fig. 4. Exemplo de *cashtag* integrada anotada com nsubj:*cashtag*.

Quando uma *hashtag* ou *cashtag* ocorre de forma *standalone* no final dos *tweets*, como a *hashtag* "#tátudobem" na Figura 4, ela deve ser conectada ao seu head por meio da *deprel* **parataxis**, acrescida da subrelação corresponde, isto é, ***hashtag/cashtag***. As *hashtags* compostas pelos *tickers*, em particular, tendem a ocorrem de forma bastante frequente no início dos *tweets*, compondo um padrão recorrente de mensagem no corpus. Trata-se do padrão: [(<hashtag(Ticker)> <complemento> (mensagem: NNN) <url>], como nos exemplos (7-12). Nesses casos, a (<hashtag(Ticker)> será sempre root e a relação entre ela e o <complemento> depende da natureza deste, podendo ser **nmod, appos, amod, advmod,** etc.

(7) **#vale5** (mensagem: 950904) http://t.co/wfR8HEPu4k (<complemento> vazio)
(8) **#PETR4** - 15 min (mensagem: 951348) http://t.co/7A5UINu9Mu
(9) **#BBAS3** Banco de a Brasil (mensagem: 956467) http://t.co/75T8wtmEXw
(10) **#csna3** semanal (mensagem: 950998) http://t.co/suRkLOSBUz
(11) **#LLXL3** - acima de 1 (um) (mensagem: 952921) http://t.co/11sdL24xTr
(12) **#PETR4** 15 min - acho que nao! (mensagem: 952919) http://t.co/32XqwNSA6Y

5.3. Menção

Outro fenômeno característico das mensagens do *Twitter* são as menções, as quais indicam que a postagem é uma resposta a um *tweet* do usuário mencionado. Anotadas com a PoS *tag* PROPN, elas podem ocorrer integradas à sintaxe do *tweet* ou *standalone*. Quando integradas, devem ser conectadas ao seu *head* em nível sintático pela *deprel* que representa sua função. Na Figura 5, por exemplo, a menção "@petrobras" foi conectada ao *head* "imagem" por **nmod**.

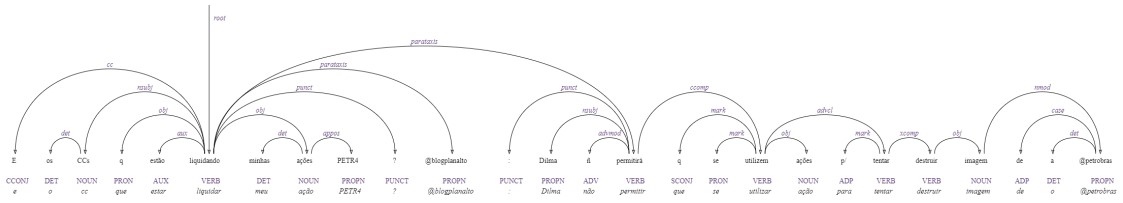

Fig. 5. Exemplo de menção integrada à sintaxe anotada com nmod.

Quando antecedida pela marca de *retweet* (RT), a menção é o dependente da relação nmod que se estabelece com o *token* RT (*head*), como na Figura 6.

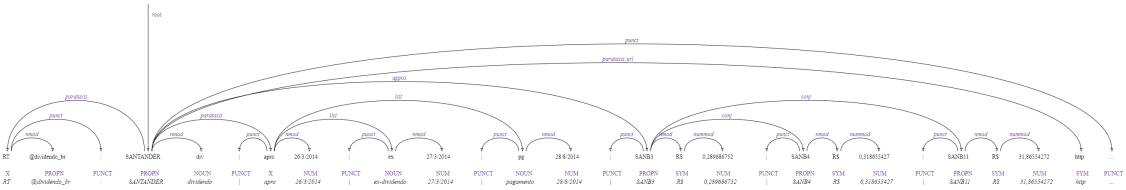

Fig. 6. Exemplo de menção conectada a RT por nmod.

Quando *standalone*, elas são conectadas ao *root* do *tweet* por **parataxis**, acrescida da sub-relação ***mention***, como ilustrado na Figura 7.

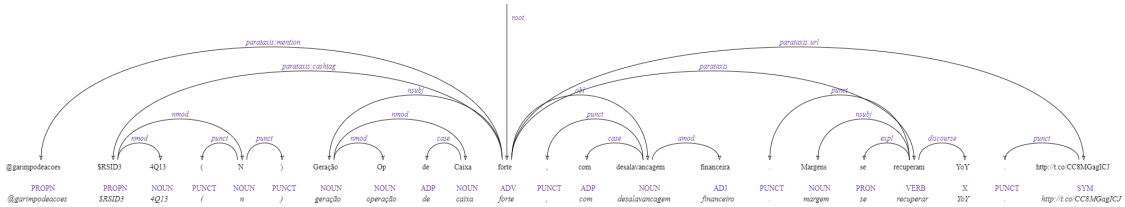

Fig. 7. Exemplo de menção *standalone* anotada com parataxis:mention.

5.4. *Emoticon* e Marcas de expressividade (onomatopeia)

Para as expressões ou fenômenos que indicam sentimentos, como *emoticons* e onomatopeias (de riso), ocorrem apenas em contexto *standalone* no DANTEStocks. Para

esses casos, não há discordância na literatura sobre a *deprel* a ser empregada, que se trata de discourse ([Liu et al. 2018], [Sanguinetti et al. 2022]). Na Figura 8, ilustra-se essa estratégia com a anotação de um *tweet* que contém uma ocorrência de um *emoticon*.

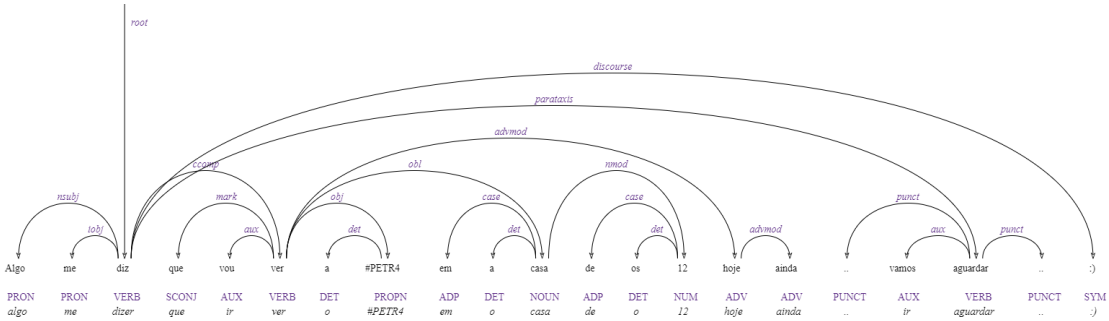

Fig. 8. Exemplo de emoticon anotado com discourse.

5.5. RT (marca de *retweet*)

No DANTEstocks, as marcas de *retweet* ocorrem em contexto *standalone*. Para tanto, a literatura fornece ao menos duas estratégias de anotação sintática: **discourse** [Liu et al. 2018] e **parataxis** [Sanguinetti et al. 2022]. Neste trabalho, opta-se por conectar uma RT ao *root* do *tweet* por meio de **parataxis** (cf. Figura 6) porque se entende que não há uma relação sintática de fato entre a marca de *retweet* e o restante da mensagem.

5.6. Truncamento

Os casos de truncamento, tanto de construções/frases como de palavras, são um desafio para a anotação sintática, uma vez que apresentam a omissão de parte da mensagem/palavra a ser anotada. Embora [Sanguinetti et al. 2020] tenham traçado algumas diretrizes, esse fenômeno tem sido tratado caso a caso no DANTEStocks, uma vez que há uma diversidade grande de ocorrências distintas. No entanto, sempre que possível, busca-se que, diante de um caso de truncamento lexical cuja forma padrão (completa) da palavra foi recuperada (do próprio Twitter ou da web), anotar esse truncamento com base na função sintática da palavra completa no *tweet*. Na Figura 9, por exemplo, "recomend", que é o truncamento de "recomendações", foi conectado ao *token* "permanecem" pela *deprel* obl, acrescida da sub-relação **:wtrunc** (*word truncation*). Os casos de truncamento de construções ou frases são mais complexos e variados, requerendo soluções específicas.

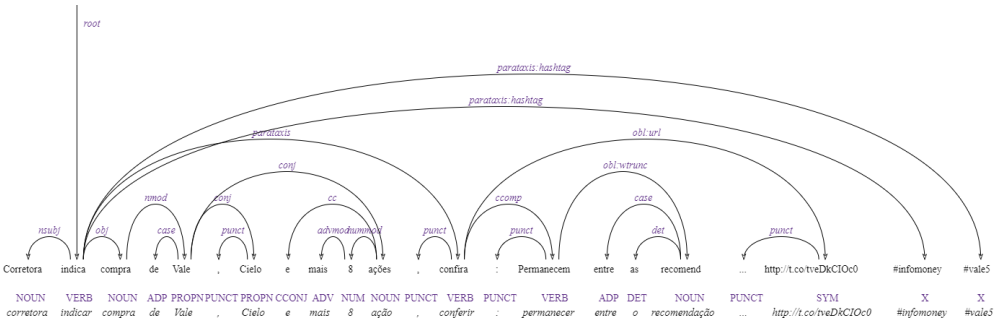

Fig. 9. Exemplo de anotação sintática para truncamentos.

O Quadro 2 sintetiza as estratégias de anotação discutidas/propostas nesta seção.

Quadro 2. Diretrizes iniciais de anotação sintática de fenômenos CGU.

Fenômeno CGU	Diretriz de anotação
URL *standalone*	PARATAXIS
URL integrada	Função sintática exercida. Quando indicar local: OBL
Hashtag standalone	PARATAXIS
Hashtag integrada	Função sintática exercida
Cashtag standalone	PARATAXIS
Cashtag integrada	Função sintática exercida
Menção *standalone*	VOCATIVE
Menção integrada	Função sintática exercida
Emoticon	DISCOURSE
RT *standalone*	PARATAXIS
Express ao Onomatopeica (riso)	DISCOURSE
Truncamento de palavra	Função sintática da palavra truncada caso recuperável, com a sub-relação :wtrunc

6. Considerações finais

Neste trabalho, apresentam-se as primeiras estratégias de anotação sintática segundo o modelo UD para *tweets* em português. Tendo em vista que as estratégias foram discutidas/propostas com base em fenômenos CGU identificados em apenas uma parcela (10%) do córpus DANTEStocks (isto é, 405 *tweets*), pretende-se analisar mais 10% do córpus com o objetivo de validar as propostas de anotação e/ou identificar outros fenômenos ainda não previstos. Uma vez validadas, as estratégias aqui propostas darão origem a um manual de diretrizes anotação sintática-UD que será empregado como material de suporte para a revisão manual da futura anotação automática do córpus DANTEStocks.

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.

References

Di-Felippo, A., Postali, C., Ceregatto, G., Gazana, L., Silva, E., Roman, N., and Pardo, T. (2021). Descrição preliminar do corpus dantestocks: Diretrizes de segmentação para anotação segundo universal dependencies. In *Anais do XIII Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana*, pages 335–343, Porto Alegre, RS, Brasil. SBC.

Duran, M. S. (2022). *Manual de Anotação de Relações de Dependência – Versão Revisada e Estendida*.

Duran, M. S., Oliveira, H., and Scandarolli, C. (2022). Que simples que nada: a anotação da palavra que em corpus de UD. In *Proceedings of the Universal Dependencies Brazilian Festival*, pages 1–11, Fortaleza, Brazil. Association for Computational Linguistics.

Guibon, G., Courtin, M., Gerdes, K., and Guillaume, B. (2020). When collaborative treebank curation meets graph grammars. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 5293–5302, Marseille, France. European Language Resources Association.

Liu, Y., Zhu, Y., Che, W., Qin, B., Schneider, N., and Smith, N. A. (2018). Parsing tweets into Universal Dependencies. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 965–975, New Orleans, Louisiana. Association for Computational Linguistics.

Nivre, J. et al. (2020). Universal dependencies v2: An evergrowing multilingual treebank collection. *Proceedings of the Twelfth International Conference on Language Resources and Evaluation (LREC 2020)*.

Sanguinetti, M., Bosco, C., Cassidy, L., Çetinoğlu, Ö., Cignarella, A. T., Lynn, T., Rehbein, I., Ruppenhofer, J., Seddah, D., and Zeldes, A. (2020). Treebanking user-generated content: A proposal for a unified representation in Universal Dependencies. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 5240–5250, Marseille, France. European Language Resources Association.

Sanguinetti, M., Bosco, C., Cassidy, L., Özlem Çetinoğlu, Cignarella, A. T., Lynn, T., Rehbein, I., Ruppenhofer, J., Seddah, D., and Zeldes, A. (2022). Treebanking user-generated content: a UD based overview of guidelines, corpora and unified recommendations. *Language Resources and Evaluation*, 57(2):493–544.

Sanguinetti, M., Bosco, C., Mazzei, A., Lavelli, A., and Tamburini, F. (2017). Annotating Italian social media texts in Universal Dependencies. In *Proceedings of the Fourth International Conference on Dependency Linguistics (Depling 2017)*, pages 229–239, Pisa,Italy. Linköping University Electronic Press.

Silva, F. J. V., Roman, N. T., and Carvalho, A. M. (2020). Stock market tweets annotated with emotions. *Corpora*, 15(3):343–354.

Silveira, N., Dozat, T., de Marneffe, M.-C., Bowman, S., Connor, M., Bauer, J., and Manning, C. D. (2014). A gold standard dependency corpus for English. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014)*.