A Dependency Treebank of Tweets in Brazilian Portuguese: Syntactic Annotation Issues and Approach

Ariani Di Felippo[1,2], Maria das Graças V. Nunes[1], Bryan K. da Silva Barbosa[1,3]

[1]Núcleo Interinstitucional de Linguística Computacional (NILC)

[2]Departamento de Letras, Universidade Federal de São Carlos (UFSCar)
Caixa Postal 676 -- 13565-905 -- São Carlos -- SP -- Brazil

[3]Programa de Pós-Grad. em Linguística, Universidade Federal de São Carlos (UFSCar)
Caixa Postal 676 -- 13565-905 -- São Carlos -- SP -- Brazil

ariani@ufscar.br, gracan@icmc.usp.br, bryan42@estudante.ufscar.br

Abstract*. We broaden Brazilian Portuguese (BP) dependency parsing to handle "user-generated content" by developing and annotating the first BP treebank of tweets (actual X posts) within the Universal Dependencies framework. DANTEStocks has a size of 4,048 tweets from the stock market domain already annotated with PoS tags and morphological features from UD. In this paper, we describe our standards for dealing with Twitter- and domain-specific properties of the corpus in the dependency annotation process. The enriched version of DANTEStocks with dependency relations from UD and the annotation guidelines are already publicly available.*

Resumo. *Amplia-se a análise de dependência do português brasileiro (pt-br) para lidar com "conteúdo-gerado por usuários" ao desenvolver e anotar o primeiro treebank de tweets (atuais posts do X) em pt-br segundo o modelo Universal Dependencies. O DANTEStocks possui 4,048 tweets do mercado financeiro e anotação-UD de tags PoS e traços morfológicos. Neste artigo, descreve-se a estratégia de anotação sintática adotada para lidar com as idiossincrasias do Twitter e do domínio desse corpus. A versão do DANTEStocks enriquecida com as relações de dependência-UD e as diretrizes de anotação já estão publicamente disponíveis.*

1. Introduction

The *Universal Dependencies* (UD) [Nivre *et al.* 2020] project specifies a complete morphological and syntactic representation with the goal of facilitating multilingual tagger and parser development [Nivre 2016]. The morphology of a word consists of 3 levels of information: PoS tag, lemma, and features. Syntactic annotation consists of typed dependency relations (*deprels*) between words. Currently, the model has 17 PoS tags and 37 *deprels*, plus a non-fixed set of morphological features. Figure 1 shows an example of an annotated tweet in DANTEStocks. In a dependency tree, one word is the head of the utterance (`root`) and all other words are dependent on another word. The labeled arcs represent the *deprels*, pointing from heads to their dependents. The PoS tag and the lemma of each word are displayed below the text. The morphological features are not included in this figure. However, the token "acordo" ("agreement"), for example, has the following *features* and values according to UD: `Number=Sing, Gender=Masc`.

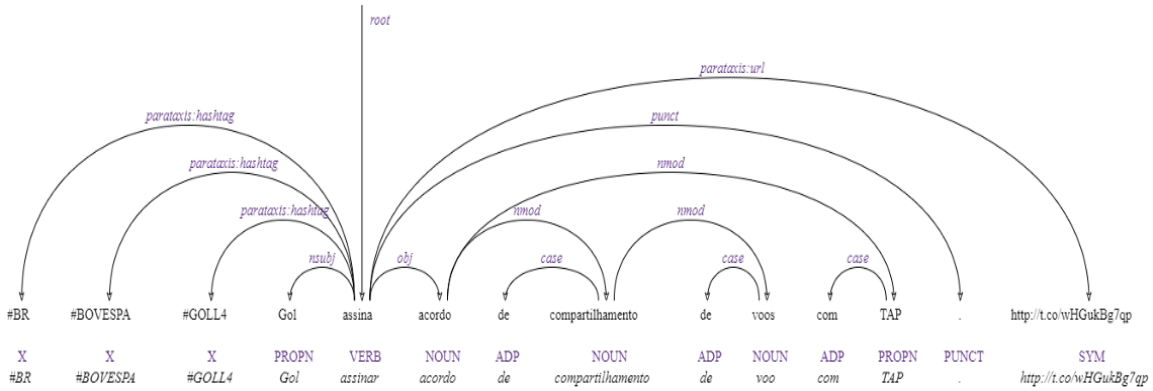


Figure 1. UD annotation of *"#BR #BOVESPA #GOLL4 Gol assina acordo de compartilhamento de voos com TAP - http://t.co/wHGukBg7qp"*[1].

Motivated by UD, treebanks for new domains, genres and language varieties have been recently built. Among the treebanks featuring *user-generated content* (UGC) created from 2014 onwards, a significant number is either partially or entirely made up of Twitter data, whose language diverges from standard written texts in several ways, posing significant challenges for building UD-based treebanks. These challenges include non-standard spelling, capitalization, punctuation, syntax, platform conventions, and creative language use, which often introduce many unknown words. Promoting cross-linguistic consistency, UD guidelines for UGC annotation have been provided [*e.g.* Sanguinetti *et al.* 2022], however, when it comes to a technical domain, specific strategies are required. Due to the variety and complexity of the language, adequate treatment of the phenomena by means of an already existing model, such as UD, is a non-trivial task.

We report the syntactic annotation of DANTEStocks within UD framework. First, we briefly describe the segmentation, tokenization, and the previous PoS annotation of the corpus (§2). Then, we present the annotation guidelines for the UD-*deprels* (§3). In (§4), we detail the semi-automatic approach for annotating the dependency relations, including data organization, creation of a reference subcorpus, and training a state-of-art parsing model on tweets. In (§5), we report a small-scale evaluation of the syntactic annotation. Finally, we put our work into context and outline future work (§6).

2. The DANTEStocks Corpus

DANTEStocks is a corpus comprising 4,048 tweets (with 140-character limit) from the stock market domain. It was automatically collected by fetching posts containing a *ticker*[2] of one of the 73 stocks that compose the Ibovespa[3]. Considering the entire tweet as a basic unit for syntactic analysis, the DANTEStocks' tweets are not segmented into smaller units (sentences, clauses or phrases). This decision saved the effort to conduct a manual segmentation or do revision of an automatic process. Additionally, the corpus was not normalized to preserve its diversity, as the goal was to develop multigenre applications. Although focusing on syntax, we outline the previous segmentation and morphological UD-annotation because they contextualize some annotation decisions.

[1] "#BR #BOVESPA #GOLL4 Gol signs flight sharing agreement with TAP - http://t.co/wHGukBg7qp"
[2] It is a five or six-character alphanumerical string that represents a specific type of stock from a company, such as "PETR4" for Petrobras' preferred stock.
[3] It is the benchmark indicator of B3 ("*Brasil, Bolsa, Balcão*"), which is the main financial exchange in Brazil.

Following the lexicalist view of syntax of UD, the syntactic words[4] (tokens) were automatically segmented by a version of the NLTK TweetTokenizer[5], augmented with specific rules for UGC [Silva *et al.* 2021]. The tool preserves most white-space-delimited tokens, including phonetization (*e.g.* "*d+*" > "*demais*"), hashtag, cashtag[6], at-mention, emoticon, and URL, and splits off single orthographic tokens that correspond to multiple (syntactic) words, such as clitics, contractions (canonical and non-canonical), punctuation marks (except for abbreviations), and valuation rates and monetary values with unconventional orthography. After the manual revision of the tool output, the corpus ends up with a total number of 81,037 tokens.

The morphological annotation was also conducted semi-automatically[7] [Silva *et al.* 2021]. The PoS tags generated by the UDPipe 2 parser [Straka 2018], trained incrementally over UD-Portuguese Bosque [Rademaker *et al*. 2017] and tweets, were manually analyzed by three annotators, and the cases of disagreement among them were adjudicated by a senior linguist based on guidelines tailored for standard texts in BP [Duran 2021] and tweets [Di-Felippo *et al.* 2022]. All 17 UD-tags can be found in DANTEStocks. PUNCT, NOUN, and PROPN are the most frequent, with around 16%, 15% and 14% of all the tags, respectively. Lemmas and grammatical features were semi-automatically obtained by using the PortiLexicon-UD lexicon [Lopes *et al.* 2022]. Major manual adjustments were required for lemmatization due to the high rate of out-of-vocabulary words. Regarding grammatical features, the scenario was quite different. The features extraction was guided by the already validated PoS tags and lemmas, which decreased the manual revision effort. Most of the corrections was related to errors arising from ambiguity about VERB class features (*VerbForm*, *Mood*, *Tense*, *Gender*, *Number* and *Person*). The manual revision also focused on checking *Typo*, *Abbr*, and *Foreign*, which are features that can be associated to words belonging to all PoS classes.

While many syntactic structures of tweets could be quite straightforwardly annotated using the general guidelines adapted for Portuguese [Duran 2022], many of them needed specific choices. In the next section, we discuss the main challenging issues for annotation decisions related to dependency relations (*deprels*).

3. Syntactic Annotation Issues

3.1. Medium- and domain-dependent (lexical) phenomena

Mostly following the recommendations of Sanguinetti *et al.,* tokens classified as orthographic variation from standard norm by [Scandarolli *et al.* 2023] were annotated with their actual syntactic roles, since they are always syntactically integrated. These variations include user-generated content phenomena such as substitution, omission, insertion, and transposition of characters (e.g., letters, spaces, hyphens, and diacritics). A good example is the token "*nao*" (instead of "*não*") ("no") in (1) "*VALE5 nao passa de 29,9*"[8], which has a case of diacritic omission. In the example, "*nao*" was related to the `root` "passa" by `advmod`, since it is an adverb that modifies a predicate.

[4] It is the basic annotation unit that plays a syntactic function in an utterance.
[5] https://www.nltk.org/api/nltk.tokenize.html
[6] It was specifically designed to track financial instruments (*e.g.*, $PETR4).
[7] The version of the corpus containing PoS and features annotation is publicly available at: https://sites.google.com/icmc.usp.br/poetisa/resources-and-tools.
[8] "VALE5 does not exceed 29,9"

The same strategy was adopted for treating most of the phenomena classified as "innovative norm"[9] by [Scandarolli *et al.* 2023] (*i.e.*, abbreviation, neologism, mark of expressiveness and homophone writing), since they are also always syntactically integrated. Pictogram (emoticon/emoji), which is a mark of expressiveness, is the only one that occurs non-syntactically integrated (standalone), being attached to the **root** by `discourse`. The other two types of innovative norm's phenomena required annotation guidelines when standalone and syntactically integrated (Table 1). For the medium-dependent devices, the treatment given to the at-mentions when preceded by the RT mark is only that differs from the recommendation of Sanguinetti *et al.* Instead of considering the at-mention as standalone and attaching it to the main predicate with `vocative`, we treat it as a syntactically integrated token attached to the RT mark by `nmod`. This is due to our interpretation of an elliptical preposition "*de*" ("of") ("RT **de** @user"), indicating an attributive relationship between the RT/SYM and the @user/PROPN. Also differently, all the cases of `parataxis` involving a UGC phenomenon in DANTEStocks are annotated with a corresponding subrelation, not only for URL and hashtags.

Table 1. UD-dependency guidelines for Twitter- and domain-specific issues.

UGC issue	Subtype	Syntactic integration	Standard syntactic role	Other
Medium-dependent token	*Hashtag*	No		`parataxis:hashtag`
		Yes	✓	
	At-mention	No		`parataxis:mention`
		Yes	✓	`nmod` (of the RT)
	URL	No		`parataxis:url`
		Yes	✓	
	RT	No		`parataxis:rt`
		Yes	✓	
	Truncation	Yes	✓ (`:wtrunc` or `:strunc`)	
	Code-switching (intra)	Yes	✓ (if known)	`flat:foreign` (if unknown)
Domain-specific token	*Ticker*	Yes	✓	
	Cashtag	No		`parataxis:cashtag`
		Yes	✓	

3.2. Unconventional syntax

Besides all the linguistic issues previously mentioned, the complexity of the UD-annotation also rises from the highly contextual nature of Twitter, and the high level of fragmentation that seems to be typical in UGC from stock market domain. This provides a rich context for ambiguities and ellipses, resulting in unconventional syntactic structures whose most appropriate UD analysis depends on the interpretation of the tweet content. One example is `nsbuj:pass` without the `aux:pass`. To recommend attaching "*#cyre3*" to the `root` "*postado*" by `nsbuj:pass` in the tweet of Figure 2, we assumed that the auxiliary verb is elided. In Figure 2, we also assumed an elliptical preposition ("*a*") preceding "*+1,78*" to connect "*1,78*" by `obl`. Since the syntactic function of "(+)*1,78*" is ambiguous (i.e., `obl` of *"postado"* or `nmod` of "*abertura*"), the choice of "*1,78*" as dependent on the `root` by `obl` illustrates annotation decisions based on the interpretation of domain experts.

[9] They are lexical alternatives to existing standard words and frequent linguistic devices that are found in the Twitter and/or stock market domain language **[**Scandarolli *et al.* 2023**]**.

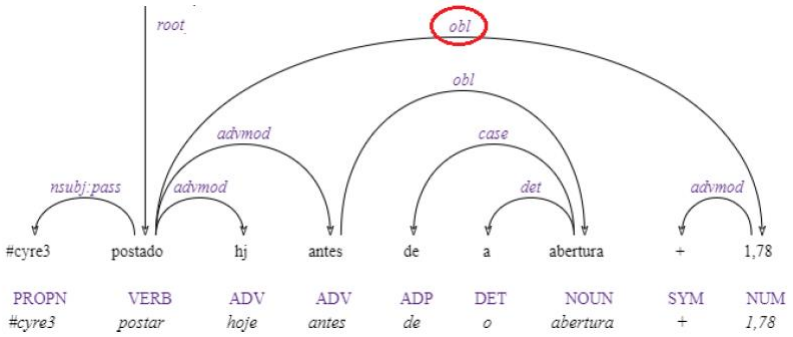


Figure 2. Syntactic ellipsis in the fragment *“‘#cyre3 postado hj antes da abertura +1,78”*[10].

3.3. Structural patterns

Besides the UGC (lexical) phenomena and unconventional syntax issues, we also identified 22 recurring structural patterns among the tweets in DANTEStocks. Such patterns correspond to almost 1,000 instances of the corpus, *i.e.* unique tweets. For each pattern, we created a template for guiding the annotation of the pattern instances in the corpus. The 22 templates also compose the dependency annotation guidelines for the DANTEStocks corpus, as well as the recommendations for the treatment of the lexical phenomena and unconventional structures [Di-Felippo *et al.* 2024].

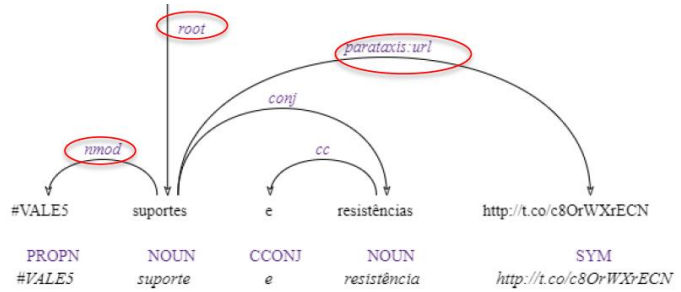
More precisely, a template contains 3 fields: (i) pattern, *i.e.* a mnemonic description, (ii) elements, *i.e.* list of pattern elements and the corresponding annotation guideline within UD, and (iii) example, *i.e.*, at least one attested instance of the pattern from the corpus with its UD-dependency annotation. It is important to mention that, since the patterns usually refer to fragmented and/or full of syntactic ellipsis tweets, the template specification is based on a possible interpretation of the tweets, which was done with the support of stock market´s experts.

For illustration, the Template 11 is shown in Table 2. It corresponds to 20 unique instances in the corpus. Since the pattern of the template represents very fragmented tweets, the domain experts helped us to interpret corpus utterance such as that in Table 2 as being composed by three blocks of information, resulting in the following pattern description: <hashtag-ticker><theme><url>.

The <theme> provides information about a specific stock, codified by the <hashtag-ticker>, and it was considered the main information of the utterance. Since the <theme> is always being introduced by the coordinate expression “support and resistance”[11], the first element of the expression (*i.e.* “suportes”) is the `root`, as indicated in the field “elements”. In the “element” field, it is also indicated that the <hashtag-ticker> is dependent on the `root` with the `nmod` tag, due to interpretation of “#VALE” as a nominal that functionally corresponds to a modifier of another noun (“*suportes*”). Since the `nmod` relation is usually introduced by a preposition (ADP tag) in Portuguese, we assume, to propose the template, that there is an elliptical preposition “*de(+a)*” (*i.e.* “*suportes e resistências* ***da*** *VALE4*”) (“support and resistance” **of** #VALE5). Finally, the <url> is dependent on the `root` with `parataxis:url` because it is a run-on segment.

[10] “#cyre3 posted today before opening +1,78”.
[11] Terms that indicate price levels where a specific stock tends to reject the current trend and reverse, *i.e.*, they indicate potential turning points in a stock’s price.

Table 2 Template for UD-dependency annotation of tweets with structural pattern.

Pattern	**<hashtag-ticker> <theme> <url>**, where:
Elements	a. <hashtag-ticker> is dependent on the **root** with the **nmod** label b. <theme> contains the expression "*suportes e resistências*"; "*suportes*" is the **root** c. <url> is dependent of the **root** with the **parataxis:url** tag
Example	#VALE5 suportes e resistências http://t.co/c8OrWXrECN
	

4. Syntactic Annotation Approach

The dependency-based annotation of DANTEStocks was held in two semi-automatic stages [Barbosa 2024]. The first one aimed at creating a reference subcorpus and the second stage of the annotation focused on fine-tuning a pre-trained parser for tweets by using the reference subcorpus as part of its initial training set. To start the syntactic annotation, all 4,048 tweets were grouped into three major sets, capturing tweets with: (i) relatively standard language, (ii) recurring structural patterns, and (iii) other (tweets that do not belong to any of the first two sets). Tweets were classified through *k-means* clustering [Macqueen 1967] with *tf-idf* ("term frequency–inverse document frequency") [Luhn 1957].

4.1. Creation of a Reference Subcorpus

The organization of tweets into sets as mentioned above allowed us to select a few instances from each set, covering all the lexical and structural diversity of DANTEStocks to compose a reference subcorpus of 1,000 tweets. Furthermore, as an attempt to achieve annotation consistency, particularly given the non-canonical language of the corpus, the semi-automatic annotation of the subcorpus was also based on such classification. This means that the data from each major set was manually reviewed separately.

To create a gold-standard subcorpus we also used the UDPipe 2 parser trained over UD-Portuguese Bosque to annotate the 1.000 tweets. The UD-annotated subcorpus was later manually revised by a single expert. Taking advantage of the previous experience of the expert in UD-annotation of journalistic texts and the training of UDPipe 2 over Bosque, the manual revision started with tweets that present relatively standard language. The next tweets were those with recurring structural patterns, and finally the tweets with a variety of lexical and structural characteristics. During the revision process, the challenging issues described in Section 3 were discussed, and the annotation decisions gave rise to the guidelines for the treatment of tweets from the stock market domain within the UD framework [Di Felippo *et al.* 2024]. The guidelines were used to support the manual revision of the rest of the corpus, which was done when training a state-of-art parser on the tweets from DANTEStocks. After the revision of the subcorpus, we ended up having a gold-standard subset of 1,000 syntactically annotated tweets.

4.2. Parsing model training

The rest of the corpus was annotated by customizing Stanza [Qi *et al.* 2020] for DANTEStocks. Stanza is a well-known pre-trained model for Portuguese, having the advantage of being a user-friendly pipeline for text analysis. The procedure began with the Stanza base architecture, fine-tuned on Porttinari-base [Duran *et al.* 2023], which is a journalistic corpus composed of 8,418 sentences (168,080 tokens) manually annotated with UD, and the reference subcorpus. For the first run of Stanza, comprising Porttinari-base and the reference subcorpus as initial training dataset, was applied the same distribution of data found in Porttinari-base[12], resulting in a dataset of 9,893 samples, being 70% for training, 10% for validation, and 20% for testing. The resulting parser was used to annotate a new package of data (out the remaining 3,048 tweets), which was manually revised and incorporated to the previous data set, being then used to start a new training run of Stanza. This cycle continued incrementally until the last package of tweets was annotated/revised. Besides the first training iteration, we carried out five training runs, adding packages of 203, 300, 400, 400, and 1233 tweets per iteration, respectively (totaling 2,536 tweets). The resulting model of the 6th (final) run was used to annotate the remaining 512 tweets. The tweet packages were added in the same order as the manual revision of the reference subcorpus: standard language tweets, structural pattern tweets, and tweets with varied lexical/structural properties.

For each of the five runs, we kept, whenever possible, the same distribution of data for training, validation and testing used in the first iteration, and computed Stanza's performance based on the *Unlabeled Attachment Score*[13] (UAS) and *Labeled Attachment Score*[14] (LAS). The UAS accuracy increased from 94.46% at the first run to 95,78% in the last (6th) iteration, becoming 1,32% better. For LAS, the final accuracy (6th run) achieved 94,62%, increasing 0,76% from the first run accuracy of 93,86%. The increase of the dependency relation measures indicates that the model's ability to capture the syntactic structures of the tweets has improved as we incorporate news tweet into the training sets. For comparison purposes, the accuracy of the best model for journalist texts in Portuguese was also around 96% (UAS) and 95% (LAS) [Lopes and Pardo 2024]. Figure 3 depicts the overall distribution of the dependency relations (without subrelations) in DANTEStocks.

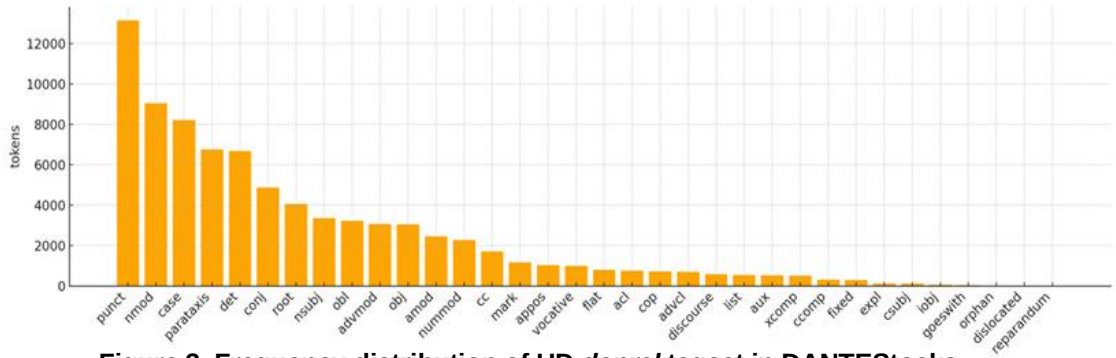


Figure 3. Frequency distribution of UD *deprel* tagset in DANTEStocks.

[12] The 8,418 sentences were split into training, development, and test sets, with 70% (5,893 sentences), 10% (842 sentences), and 20% (1,683 sentences) of the corpus, respectively.

[13] UAS indicates the accuracy of the *head* ignoring the relation's name (*deprel*) [Nivre and Fang, 2017].

[14] LAS evaluates the output of a parser by considering how many words have been assigned both the correct syntactic *head* and the correct label, ignoring subrelations [Nivre and Fang 2017].

5. Reliability of Annotation

To provide a reliability measure of the annotation of DANTEStocks, a second NLP expert (also with UD-annotation experience) manually reviewed the automatic annotation of 100 random tweets based on the same guidelines [Duran 2022; Di Felippo *et al.* 2024]. The dependency-trees analyzed by the additional annotator could be from the reference subcorpus or generated by Stanza in one of its interactions. The *Inter-Annotator Agreement* (IAA) score was calculated by using the *Kappa* coefficient [Cohen, 1960; Carletta, 1996] in two different settings [Barbosa 2024]. In the first, the focus was to evaluate the annotation of *head* and *deprel* separately. The *Kappa* results for *head* and *deprel* were 0.96 and 0.97, respectively. In the second setting, the evaluation aimed at the combination of *head* and *deprel,* obtaining the *Kappa* score of 0.95. The IAA per *deprel* was measured by using the *total agreement* score [Sobrevilha Cabezudo 2015]*,* since *Kappa* is not appropriate given the unbalanced distribution of the relations. We obtained the *total agreement* of 100% for more than half of the 46 different *deprels* (including subrelations) that occur in the sample of 100 tweets. Out of the 1.743 annotated relations, there are 42 cases of disagreement. The most frequent conflict was between *obl* and *nmod.* Some of them were caused by different but potential interpretations about the functional role of the prepositional phrase (in bold) in structure like "*arrisque vd* ***em #petr4***" ("risk selling in #petr4"). While one annotator attached "petr4" to the verb via `obl`, functioning as a non-core (oblique) argument or adjunct, the other assumed that "petr4" is a modifier of the noun "*vd*" ("*venda*"), being attached to it by `nmod`. It is also interesting that, among the 22 *deprels* with *total agreement* different from 100%, 12 of them contain subrelations, indicating that the annotation is more complex when using language-specific relations. Even though a small-scale evaluation, the results indicate that the overall IAA was otherwise quite high, especially for the challenging task. This might be due to the large and detailed recommendations of our guidelines for the syntactic annotation of the tweets.

6. Final Remarks and Future Work

We described our effort on building the first BP treebank for Twitter microtext, annotated within the framework of UD. The contributions are the treebank itself, the instantiation of the UD guidelines for stock market tweets in BP, and the customization of a current state-of-the-art parser for tweets. Our main difficult was interpreting the tweets, due to the medium- and domain-lexical phenomena and uncommon constructions. Thus, despite the constant help of domain experts, we can say that the dependency annotation of many tweets in DANTEStocks (especially those with fragmentation, *e.g.,* aborted text) represents potential syntactic analysis of the tweets. Currently, the two annotators involved in this work are analyzing the disagreements to assign a consensual *deprel* for each case and to make the treebank available soon. The guidelines for the syntactic UD-based annotation of DANTEStocks and the treebank itself (*beta version*) are available at the POeTiSA project webpage (https://sites.google.com/icmc.usp.br/poetisa/).

Acknowledgements. This work was carried out at the Center for Artificial Intelligence of the University of São Paulo (C4AI - http://c4ai.inova.usp.br/), with support by the São Paulo Research Foundation (FAPESP grant #2019/07665-4) and by the IBM Corporation. The project was also supported by the Ministry of Science, Technology and Innovation, with resources of Law N. 8,248, of October 23, 1991, within the scope of PPI-SOFTEX, coordinated by Softex and published as Residence in TIC 13, DOU 01245.010222/2022-44.

References

Barbosa, B. K. S. (2024). Descrição sintático-semântica de nomes predicadores em tweets do mercado financeiro em português. Dissertação de Mestrado. Programa de Pós-graduação em Linguística, Universidade Federal de São Carlos, São Carlos/SP, 208p.

Carletta, J. (1996). Assessing agreement on classification tasks: The kappa statistic. In *Computational Linguistics*, Volume 22, Number 2, pages 249–254. MIT Press.

Cohen, J. (1960). A coefficient of agreement for nominal scales. In *Educational and Psychological Measurement*, Volume 20, Issue 1, pages 37-46.

Di-Felippo, A.; Postali, C.; Ceregatto, G.; Gazana, L. S.; Roman, N. T. (2022). Diretrizes de anotação de PoS tags em tweets do mercado financeiro: orientações para anotação em língua portuguesa segundo a abordagem *Universal Dependencies*. *Relatório Técnico do ICMC 438.* Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo. São Carlos-SP, 24p.

Di-Felippo, A., Nunes, M. G. V., Barbosa, B. K. S. (2024). Diretrizes de anotação de relações de dependência em tweets do mercado financeiro. *Relatório Técnico do ICMC 446*. Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo. São Carlos-SP, Abril, 70p.

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

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

Duran, M. S., Lopes, L., Nunes, M.G.V., Pardo, T. A. S. (2023). The Dawn of the Porttinari Multigenre Treebank: Introducing its Journalistic Portion. In *Proceedings of the 14th Symposium in Information and Human Language Technology*, pages 115-124. Belo Horizonte/MG. SBC.

Krumm, J., Davis, N. Narayanaswami, C. (2009). User-Generated Content. In *IEEE Pervasive Computing*, Volume 7, Issue 4, pages. 10 – 11, IEEE, 2009.

Lopes, L., Duran, M. S.; Fernandes, P. H. L.; Pardo, T. A. S. (2022). PortiLexicon-UD: a Portuguese Lexical Resource according to Universal Dependencies Model. In *Proceedings of the 13th International Conference on Language Resources and Evaluation* (LREC), pages 6635 6643, Marseille, France. ELRA.

Lopes, L.; Pardo, T. A. S. Towards Portparser - a highly accurate parsing system for Brazilian Portuguese following the Universal Dependencies framework. In *Proceedings of the 16th International Conference on Computational Processing of Portuguese* (PROPOR), pages 401-410, Santiago de Compostela, Galiza. ACL.

Luhn, H.P. (1957). A statistical approach to mechanized encoding and searching of literary information. In *IBM Journal of Research and Development*, Volume 1, Issue 4, pages 309-317. ISSN 0018-8646. doi:10.1147/rd.14.0309

Macqueen, J. (1967) Some methods for classification and analysis of multivariate observations. In *Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability*. [S.l.], v. 1, n. 14, p. 281–297.

Nivre, J., Fang, C.-T. (2017). Universal Dependency evaluation. In *Proceedings of the NoDaLiDa 2017 Workshop on Universal Dependencies* (UDW 2017), pages 86–95, Gothenburg, Sweden. ACL.

Nivre, J., *et al*. (2016). Universal dependencies v1: A multilingual treebank collection. In *Proceedings of the 10th International Conference on Language Resources and Evaluation* (LREC), pages 1659–1666, Portorož, Eslovênia. ELRA.

Nivre, J. *et al.* (2020). Universal Dependencies v2: an evergrowing multilingual treebank collection. In *Proceedings of the 12th International Conference on Language Resources and Evaluation Conference* (LREC), pages 4034-4043. Marseille, França. ELRA.

Qi, P., Zhang, Y., Zhang, Y, Bolton, J., Manning, C. D. (2020). Stanza: A Python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (ACL) (System Demonstrations), pages 101-108. Online. ACL.

Rademaker, A., Chalub, F., Real, L., Freitas, C., Bick, E., Paiva, V. de. (2017). Universal Dependencies for Portuguese. In *Proceedings of the 4th International Conference on Dependency Linguistics* (Depling), pages 197–206, Pisa, Italy. Linköping University Electronic Press.

Sanguinetti, M. *et al*. (2023). Treebanking user-generated content: a UD based overview of guidelines, corpora and unified recommendations. In *Lang Resources & Evaluation*, Volume. 57, Issue 2, pages 493–544. Springer-Verlag, Berlin, Heidelberg.

Silva, E.H.; Pardo, T.A.S.; Roman, N.T.; Di Felippo, A. (2021). *Universal Dependencies* for tweets in Brazilian Portuguese: tokenization and Part-of-Speech tagging. In *Proceedings of the 18th National Meeting on Artificial and Computational Intelligence* (ENIAC), pages. 434-445, Online. SBC.

Scandarolli, C. L., Di-Felippo, A., Roman, N. T., Pardo, T. A. S. (2023). Tipologia de fenômenos ortográficos e lexicais em CGU: o caso dos tweets do mercado financeiro. In *Anais da VIII Jornada de Descrição do Português* (JDP) (Evento integrante do *XIV Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana* -STIL), p. 240-248, Belo Horizonte/MG, Brasil. SBC.

Sobrevilla Cabezudo, M.A., Maziero, E.G., Souza, J.W.C., Dias, M.S., Cardoso, P.C.F., Balage Filho, P.P., Agostini, V., Nóbrega, F.A.A., Barros, C.D., Di Felippo, A., Pardo, T.A.S. (2015). Anotação de sentidos de verbos em textos jornalísticos do *corpus* CSTNews. In *Revista de Estudos da Linguagem* (RELIN), Volume 23, Número 3, p. 797-832.

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