DANTE – A Treebank of Tweets in Brazilian Portuguese

Ariani Di-Felippo[1,2], **Norton Trevisan Roman**[3]

[1]Language and Literature Department (DL/UFSCar), São Carlos – Brazil

[2]Interinstitutional Center for Computational Linguistics (NILC/USP) São Carlos – Brazil

[3]School of Arts, Sciences and Humanities (EACH/USP), São Paulo – Brazil

ariani@ufscar.br, norton@usp.br

Abstract. *In this article, we present the current state of DANTE (Dependency-ANalysed corpora of TwEets), a treebank focusing on tweets (currently X posts) written in Brazilian Portuguese, featuring four subcorpora varying in size, domain, and level of annotation under the Universal Dependencies framework. This article aims to inform the Natural Language Processing and Linguistic communities about currently available corpora, as well as what to expect from the treebank in the near future, while also bringing together the tools and language resources developed within the DANTE project.*

1. Introduction

When it comes to user-generated content[1] (UGC), the existence of annotated treebanks[2] becomes an invaluable resource, as they enable the modelling of highly variable, noisy, and domain-specific language typical of social media [Sanguinetti et al. 2023]. Among the theoretical frameworks for the construction of treebanks, the Universal Dependencies (UD) project [de Marneffe et al. 2021] stands out as a widely adopted international initiative that provides a unified scheme for morphosyntactic and syntactic annotations across languages, with over 150 languages available in more than 200 corpora.

Among the resources available in the UD repository, Portuguese is represented by seven corpora, comprising five general-domain resources (Bosque [Rademaker et al. 2017], GSD[3], PUD [Zeman 2017], CINTIL [Branco et al. 2022], and Porttinari-base [Duran et al. 2023]), one domain-specific corpus in the oil and gas domain (PetroGold [Souza et al. 2021]), and one UGC corpus (DANTEStocks [Di-Felippo and Roman 2025]). Porttinari-base is a carefully revised corpus designed to serve as a gold standard within the multigenre treebank Porttinari ("PORTuguese Treebank")[4]. In addition to this component, the treebank also includes Porttinari-check, a smaller corpus used as a testbed for comparing manual and automatic annotations, and Porttinari-automatic, a large corpus annotated automatically.

Similarly to Porttinari-base, DANTEStocks is part of a larger treebank – DANTE (Dependency-ANalysed corpora of TwEets), also integrating Porttinari. This resource is a pioneering tweet-based treebank, currently comprising data from two domains, along with

[1]UGC is any form of content, such as images, videos, text and audio, that has been posted on online platforms such as social media, review pages, blogs and wikis [Krumm et al. 2008].

[2]A corpus in which every sentence is annotated with a parse tree [Jurafsky and Martin 2026].

[3]https://universaldependencies.org/treebanks/pt_gsd/index.html

[4]https://sites.google.com/icmc.usp.br/poetisa/porttinari-3-0

UD annotation. The treebank is organised into four subcorpora with different characteristics and stages of development: DANTEStocks, DANTEStocks-Large, DANTEShots and DANTEShots-Extended. The first two were drawn from the stock market domain, while the latter two focus on posts by Brazilian politicians regarding COVID-19 vaccination. This article describes then the current state of DANTE, outlining what is already available and what is to be expected for the near future. In addition, it situates DANTEStocks – the only subcorpus with morphosyntactic (Part-of-Speech – PoS) and syntactic gold-standard annotations under the UD framework – within the broader set of UGC resources available in the UD project.

The rest of this article is organised as follows. Section 2 provides an overview of the currently available UGC corpora in the UD project, discussing the role of Portuguese within this landscape, with DANTEStocks as its representative resource. Section 3, in turn, focuses on the DANTE treebank, describing its structure, subcorpora, and annotation layers. In Section 4, some NLP tools and language resources developed based on DANTE's annotations are listed, with Section 5 presenting our final considerations, also giving a brief overview of our future plans for extending each subcorpus.

2. Treebanks featuring UGC in the UD project

In a recent overview of the UGC treebanks available in the literature, [Sanguinetti et al. 2023] have found 30 resources published between 2011 and 2020. This study was based on a semi-systematic Search on Google Scholar for open-access papers and complemented by additional known resources. Unlike this overview, we focus exclusively on corpora standardised according to the UD annotation scheme and integrated into the official repository of the UD project[5], as summarised in Table 1.

The table reports the number (#) and genre of each corpus, their data sources as well as the total number of annotated sentences and tokens. Currently, the UD repository comprises 22 corpora covering 18 languages. Based on this table, the amount of data varies considerably across languages, from South Levantine Arabic, with 789 tokens, to Russian, with nearly 1,8 million tokens. Being represented by DANTEStocks, with 4,042 tweets from the stock market, Portuguese ranks 10th among languages. With 77,576 tokens, it occupies a mid-range coverage in comparison to the other UGC corpora of the UD initiative.

Six out of the 22 corpora outlined in the table are either entirely (TwittIrish, PoSTWITA, TWITTIRO, DANTEStocks, and TueCL) or partially (OOD) composed of Twitter data. This suggests a moderate prevalence of Twitter-based resources, with the platform remaining as a widely used source for short, informal text. Within the subset of corpora exclusively based on Twitter data, Portuguese ranks 2nd in terms of size, with 77,576 tokens. It is surpassed only by Italian, which collectively add up to 147,718 tokens (PoSTWITA and TWITTIRO). Despite the limited resource diversity for Portuguese, its single available corpus exhibits a relatively high token density within this subdomain. Although DANTEStocks provides a non-negligible volume of annotated data, its linguistic coverage remains nevertheless restricted to a narrow communicative setting (to wit, the stock market domain).

[5] https://universaldependencies.org/

Table 1. Overview of UGC corpora in the UD project.

Language	#	Corpus	Genre	UGC source	Sents.	Tokens
Armenian	1	ArmTDP	Various	Web, social	5,527	103,846
Bavarian	1	MaiBaam	Various	Social	1,070	14,678
Belarusian	1	HSE	Various	Web, social	25,231	305,109
Czech	1	PDTC	Various	Web, social	5,527	103,846
English	4	GUM	Various	Reddit	13,263	229,851
		EWT	Various	Web, social	16,662	251,489
		GENTLE	Various	Web, social	1,334	17,619
		GUMReddit	Various	Reddit	859	15,958
Estonian	1	EWT	Various	Web, social	7,190	60,583
Finish	1	OOD	Social	Various (Twitter)	2,122	19,364
Irish	1	TwittIrish	Social	Twitter	2,596	47,790
Italian	2	PoSTWITA	Social	Twitter	6,712	119,334
		TWITTIRO	Social	Twitter	1,424	28,384
Manx	1	Cadham	Various	Web, social	2,336	18,580
Persian	1	Seraji	Various	Web, social	5,997	151,627
Polish	1	LFG	Various	Web, social	17,246	130,967
Portuguese	1	DANTEStocks	Social	Twitter	4,042	77,576
Romanian	1	TueCL	Social	Twitter	270	4,417
Russian	1	Taiga	Various	Social media	121,967	1,758,937
South Levantine Arabic	1	MADAR	Social	Spoken	100	789
Ukrainian	1	IU	Various	Social	7,092	122,701
Western Armenian	1	ArmTDP	Various	Social	6,644	121,432
Total	22				255,211	3,704,877

3. DANTE's Current State

As for the present moment, DANTE comprises four corpora (Table 2), with its first corpus - DANTEStocks - being the richest one in terms of annotations added, also being the only fully available to the community[6]. Altogether, the four corpora feature more than 146,000 tweets, with over 137,000 of them already annotated with some UD layer, either manually or automatically, along with some semantic information, such as named entities, emotions, sentiment and stance, depending on the corpus. In what follows, we present each corpus in more detail, along with their current development stage.

3.1. DANTEStocks

DANTEStocks is the first tweebank in Brazilian Portuguese annotated within the UD framework. The corpus consists of tweets from the stock market domain, collected in 2014 based on mentions to Ibovespa-listed companies, and reflects typical Twitter constraints until 2017 (*e.g.* the 140-character limit). The original resource that gave rise to DANTEStocks was already annotated with emotions following Plutchik's model [Plutchik 2001]. DANTEStocks extends this original resource via standoff annotation, adding morphological (PoS, lemmas and features) and dependency layers following UD guidelines, all curated and validated by experts. The UD-annotation of the corpus

[6]Within the POeTiSA project: `https://sites.google.com/icmc.usp.br/poetisa`

Table 2. UGC corpora in the DANTE treebank.

Subcorpus	Domain	Annotation	Status	Tweets	Tokens
DANTEStocks	Stock market	UD (complete) Emotion Named entity	Gold-standard Gold-standard Gold-standard	4,042	77,576
DANTEStocks-Large	Stock market	UD PoS Emotion	Automatic Automatic	126,579	3,932,604
DANTEShots	Covid-19 vaccines	UD PoS Stance	Automatic Gold-standard	7,056	296,859
DANTEShots-Extended	Covid-19 vaccines	Stance Sentiment	Gold-standard Gold-standard	9,045	
Total				146,722	4,307,039

adopts tweets as the basic unit of analysis, avoiding sentence-level segmentation and normalization [Di-Felippo et al. 2021].

Tokenization was carried out in a semi-automatic way, by using DANTE's Tokenizer [Silva et al. 2021], which is a rule-based system adapted to UGC phenomena. At the morphological layer, the PoS tags were generated using UDPipe 2 [Straka et al. 2016] pre-trained on the Bosque corpus [Rademaker et al. 2017], and iteratively refined through human review and retraining [Di-Felippo et al. 2022, Felippo et al. 2023]. Lemmas and morphological features, as described in [Di-Felippo and Roman 2025], were added semi-automatically using PortiLexicon-UD [Lopes et al. 2022]. Syntactic annotation relied on a hybrid strategy, starting from a manually revised gold-standard subcorpus[7] and incrementally training a parser[8] with iterative human validation [Di-Felippo et al. 2024a, Barbosa 2024], supported by guidelines specifically designed for Portuguese [Duran 2021] and for Twitter data [Di-Felippo et al. 2024b].

The corpus was later annotated with Named Entities [Zerbinati et al. 2024] following the Second HAREM's taxonomy [Mota and Santos 2008]. This annotation was then revised and augmented [Piai et al. 2025], by including a fine-grained, linguistically motivated annotation, tailored to the financial domain of tweets in Portuguese, extending HAREM's traditional categories to better capture entities such as stock market instruments (e.g., tickers), financial indicators, and other market-specific expressions. The annotation was carried out manually by a single annotator, following detailed guidelines and iterative refinement to ensure consistency, including explicit rules for delimiting entity boundaries in noisy data (*e.g.* orthographic variation and truncation) and context-based disambiguation strategies to ensure consistent classification of financial entities.

Recently, [Ceregatto and Felippo 2025] have extended the corpus with a subset of 1,128 distinct tweets manually annotated according to the Abstract Meaning Representation (AMR) framework [Banarescu et al. 2013, Banarescu et al. 2019]. Additionally, [Barbosa 2024] provided argument structure for 1,756 instances (and 1,218 tweets) of

[7]Obtained by automatically annotating 1,000 tweets with UDPipe 2 [Straka et al. 2016] trained on Bosque [Rademaker et al. 2017], followed by manual revision.

[8]Our approach was to begin with the Stanza [Qi et al. 2020] base architecture, fine-tuned on Porttinari-base and our reference subcorpus.

146 predicate nouns[9] in DANTEStock, following NomBank [Meyers 2007] for English, which is a parallel project to PropBank [Palmer et al. 2005, Pradhan et al. 2022].

3.2. DANTEStocks-Large

DANTStocks-Large was collected following the same methodology as the corpus that originated DANTEStocks (described in [da Silva et al. 2020]). Beyond collecting a considerably larger set of tweets, the corpus' goal was also to add date and time information to them, an information that was missing in DANTStocks. In total, over 126 thousand tweets were collected from August, 2022 to September, 2023. These were then automatically annotated with emotions, following the Plutchik's Wheel of Emotions model [Plutchik 2001]. To do so, we relied on classifiers developed and tested in the corpus by [da Silva et al. 2020], as presented in [de Carvalho and Roman 2025]. Besides emotions, we have also adapted the tagger presented in [Silva et al. 2023] and used it to automatically tokenise and annotate the corpus with PoS under UD's framework. The resulting corpus comprises then over 126 thousand tweets annotated with emotions, with almost 4 million tokens automatically annotated with UD-PoS. Before officially presenting the corpus to the community, however, we intend to carry out a human adjudication of the annotations made, both for emotion and PoS.

3.3. DANTEShots

DANTEShots builds on an existing corpus of tweets related to COVID-19 vaccines [Barberia et al. 2023], as posted by political candidates during 2020 and 2021, annotated with stance towards vaccination. This corpus was then automatically annotated with UD-PoS tags, by using an adapted version of the tagger presented in [Silva et al. 2023] (the same used in DANTEStocks-Large), which resulted in a corpus of 7,056 tweets and 296.859 annotated tokens. To assess the annotation's quality, a sample of 1,000 tweets was randomly[10] taken from the corpus, being currently under review by a set of three human adjudicators. The results of the adjudication will be later used to evaluate the tagger's performance, also giving researchers an estimate of labels' accuracy.

3.4. DANTEShots-Extended

The same research group responsible for the corpus of COVID-19 vaccine posts used as the basis for DANTEShots (*cf.* [Barberia et al. 2023]), later on released a larger corpus, this time with 9,045 tweets annotated for relevance, stance and sentiment towards COVID-19 vaccines and vaccination during the pandemics [Barberia et al. 2025]. This new corpus extends the previous one by applying a new annotation scheme and increasing the corpus' size so as to include posts from 2020 to 2022 by local political elites. Since DANTEShots was based on the smallest version of this base corpus, we also decided to augment it with this new corpus, thereby giving birth to DANTEShots-Extended. This is, however, by far the most embryonic of all corpora within DANTE, and efforts are just in their initial stages to automatically annotate the corpus with PoS and syntactic structure. Once finished, the corpus will allow for the connection between morphosyntax, syntax, stance and sentiment. Although we have no fixed deadline to this task at this moment, we expect it to happen sometime next year.

[9] A predicative noun is one that selects an argument structure (valency), associating semantic participants with its lexical meaning.

[10] From a uniform distribution.

4. NLP Tools and Lexical Resources

DANTEStocks has already led to the development of various NLP tools and resources for Brazilian Portuguese[11]. One of these tools is the DANTE tokenizer [Silva et al. 2021], which is a rule-based system developed for stock-market tweets, built on top of the NLTK TweetTokenizer and extended with linguistically motivated regular expressions. Another tool is its UGC Parser [Di-Felippo et al. 2024a, Barbosa 2024], a UD parsing model derived from Stanza [Qi et al. 2020]. The model was initially trained on a combination of Porttinari-base and a gold-standard reference subcorpus, and subsequently improved through an incremental training strategy in which newly annotated tweet packages were automatically parsed, manually revised, and incorporated into successive training cycles. DANTEStocks' gold-standard syntactic (UD-based dependencies) annotation also supported the development of Genipapo [Di-Felippo et al. 2024c], a multi-genre (tweets, news, and academic texts) parser, achieving better or competitive performance in relation to genre specific trained parsers.

Regarding lexical resources, the semantic annotation of a subset of tweets from DANTEStocks inspired in NomBank gave rise to NounBank.DS [Barbosa and Felippo 2025] – an online repository that provides argument structure for instances of predicate nouns[12]. More recently, a descriptive study on DANTEStocks' vocabulary [Di-Felippo et al. 2026] resulted in a lexicon of orthographic variations, innovative forms, and medium- and domain-dependent tokens, which has already been proved to reduce the out-of-vocabulary rate produced by different embedding models. This vocabulary (LexDANTE) is available for online access and download through a dedicated repository.[13]

5. Final Remarks and Directions for Future Research

This paper outlined our current efforts in building DANTE, a treebank focusing on tweets from different domains, as a part of a larger effort to produce Porttinari, a multi-genre treebank for written texts in Brazilian Portuguese. Connecting morphosyntactic, syntactic and, sometimes, semantic information, at various levels and in different domains, we hope that DANTE can substantially contribute to improve the NLP community's ability to understand and process tweets/posts in Brazilian Portuguese.

Currently, efforts within the DANTE project focus on domain diversification, with the development of corpora from different domains (*e.g.* stock market and vaccination), aiming at their future integration into the UD repository and broader availability to the NLP community. As such, we expect to release DANTEShots by the end of this year, with the automatically generated UD annotations fully revised by human annotators. As a next step, we plan to apply a UGC Parser [Di-Felippo et al. 2024c] to automatically generate dependency annotations, thereby extending the corpus with a syntactic layer. Similar developments are planned for DANTEStocks-Large and DANTEShots-Extended, although DANTEShots-Extended is scheduled for release next year only.

[11] Available from `https://sites.google.com/icmc.usp.br/poetisa`

[12] `https://bryankhelven.github.io/NounBank.DS/`

[13] `https://bryankhelven.github.io/LexDANTE/`

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.

Limitations

When upgrading the SRL component to a higher-quality neural model, we re-evaluate the ablation-based rule analysis under the updated intermediate representation, in order to assess the robustness of the rule hierarchy to changes in semantic role prediction quality.

Ethics Statement

All tweets in DANTE are public, and can be publicly accessed through X's website by using their identification numbers. Still, profile identification was removed from all data sets so as to make it harder to determine personal information of tweet authors.

Disclosure Regarding the Use of AI Tools

AI was used in some parts of the article as a way to check the appropriateness of some grammatical structures used in the text.

References

Banarescu, L., Bonial, C., Cai, S., Georgescu, M., Griffitt, K., Hermjakob, U., Knight, K., Koehn, P., Palmer, M., and Schneider, N. (2013). Abstract meaning representation for sembanking. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pages 178–186, Sofia, Bulgaria. Association for Computational Linguistics.

Banarescu, L., Bonial, C., Cai, S., Georgescu, M., Griffitt, K., Hermjakob, U., Knight, K., Koehn, P., Palmer, M., and Schneider, N. (2019). Abstract meaning representation (amr) 1.2.6 specification. Technical report, AMR Project. Version 1.2.6.

Barberia, L., Schmalz, P., Roman, N. T., Lombard, B., and de Sousa, T. M. (2025). It's about what and how you say it: A corpus with stance and sentiment annotation for covid-19 vaccines posts on x/twitter by brazilian political elites. In Hämäläinen, M., Öhman, E., Bizzoni, Y., Miyagawa, S., and Alnajjar, K., editors, *Proceedings of the 5th International Conference on Natural Language Processing for Digital Humanities (NLP4DH 2025)*, pages 365–376, Albuquerque, USA. Association for Computational Linguistics.

Barberia, L. G., de Santana Schmalz, P. H., and Roman, N. T. (2023). When tweets get viral - a deep learning approach for stance analysis of covid-19 vaccines tweets by brazilian political elites. In *Proceedings of the 14th Symposium in Information and Human Language Technology (STIL 2023)*, Belo Horizonte, MG - Brazil.

Barbosa, B. and Felippo, A. D. (2025). Nounbank.ds: a lexical repository of nominal frames from stock market tweets in brazilian portuguese. In *Anais do XVI Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana*, pages 29–41, Porto Alegre, RS, Brasil. SBC.

Barbosa, B. K. S. (2024). Descrição sintático-semântica de nomes predicadores em tweets do mercado financeiro em português. Dissertação (mestrado em linguística), Universidade Federal de São Carlos, São Carlos, SP.

Branco, A., Silva, J. R., Gomes, L., and António Rodrigues, J. (2022). Universal grammatical dependencies for Portuguese with CINTIL data, LX processing and CLARIN support. In Calzolari, N., Béchet, F., Blache, P., Choukri, K., Cieri, C., Declerck, T., Goggi, S., Isahara, H., Maegaard, B., Mariani, J., Mazo, H., Odijk, J., and Piperidis, S., editors, *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 5617–5626, Marseille, France. European Language Resources Association.

Ceregatto, G. and Felippo, A. D. (2025). Dantestocks-amr em construção: Avanços e desafios na anotação semântica de tweets financeiros. In *Anais do XVI Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana*, pages 608–617, Porto Alegre, RS, Brasil. SBC.

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

de Carvalho, W. P. and Roman, N. T. (2025). Classifying emotions in tweets from the financial market: A bert-based approach. In *Proceedings of the 15th International Conference on Recent Advances in Natural Language Processing (RANLP 2025)*, pages 218–226, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

de Marneffe, M.-C., Manning, C. D., Nivre, J., and Zeman, D. (2021). Universal Dependencies. *Computational Linguistics*, 47(2):255–308.

Di-Felippo, A., das Graças Nunes, M., and Barbosa, B. (2024a). A dependency treebank of tweets in brazilian portuguese: Syntactic annotation issues and approach. In *Anais do XV Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana*, pages 192–201, Porto Alegre, RS, Brasil. SBC.

Di-Felippo, A., Nunes, M. d. G. V., and Barbosa, B. K. d. S. (2024b). Diretrizes de anotação de relações de dependência em tweets do mercado financeiro. Technical Report n. 446, Relatório Técnico – ICMC, USP, São Carlos.

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.

Di-Felippo, A., Postali, C., Ceregatto, G., Gazana, L. S., and 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. Technical Report n. 438, Relatório Técnico – ICMC, USP, São Carlos-SP.

Di-Felippo, A. and Roman, N. T. (2025). Dantestocks: A multi-layered annotated corpus of stock market tweets for brazilian portuguese. *Revista Brasileira de Linguística Aplicada*, 25(1).

Di-Felippo, A., Roman, N. T., Barbosa, B. K. S., Oliveira, G. P. d., and Scandarolli, C. L. (2026). Lexical and orthographic variation in Portuguese financial tweets: Annotation, analysis, and implications for embedding models. In Souza, M., de Dios-Flores, I., Santos, D., Freitas, L., Souza, J. W. d. C., and Ribeiro, E., editors, *Proceedings of the 17th International Conference on Computational Processing of Portuguese (PROPOR 2026) - Vol. 1*, pages 628–637, Salvador, Brazil. Association for Computational Linguistics.

Di-Felippo, A., Roman, N. T., da Silva Barbosa, B. K., and Pardo, T. A. S. (2024c). Genipapo – a multigenre dependency parser for brazilian portuguese. In *Proceedings of the 15th Brazilian Symposium in Information and Human Language Technology (STIL 2024)*, pages 257–266, Belém - PA, Brazil.

Duran, M., Lopes, L., Nunes, M. d. G. V., and Pardo, T. A. S. (2023). The dawn of the Porttinari multigenre treebank: introducing its journalistic portion. In *Proceedings of the XIV Brazilian Symposium in Information and Human Language Technology (STIL)*, pages 115–124, Porto Alegre, RS, Brasil. SBC.

Duran, M. S. (2021). Manual de anotação de relações de dependência: orientações para anotação de relações de dependência sintática em língua portuguesa, seguindo as diretrizes da abordagem universal dependencies (ud). Technical Report n. 435, Relatório Técnico – ICMC, USP, São Carlos.

Felippo, A. D., Roman, N. T., Pardo, T. A. S., and de Moura, L. P. (2023). The dantestocks corpus: an analysis of the distribution of universal dependencies-based part-of-speech tags. *Revista da Abralin*, 22(2):249–271.

Jurafsky, D. and Martin, J. H. (2026). *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition with Language Models*. 3rd edition. Online manuscript released January 6, 2026.

Krumm, J., Davies, N., and Narayanaswami, C. (2008). User-generated content. *IEEE Pervasive Computing*, 7(4):10–11.

Lopes, L., Duran, M., Fernandes, P., and Pardo, T. (2022). PortiLexicon-UD: a Portuguese lexical resource according to Universal Dependencies model. In Calzolari, N., Béchet, F., Blache, P., Choukri, K., Cieri, C., Declerck, T., Goggi, S., Isahara, H., Maegaard, B., Mariani, J., Mazo, H., Odijk, J., and Piperidis, S., editors, *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 6635–6643, Marseille, France. European Language Resources Association.

Meyers, A. (2007). Annotation guidelines for nombank: Noun argument structure for propbank. Technical report, NomBank Project, S.l.

Mota, C. and Santos, D. (2008). *Desafios na avaliação conjunta do reconhecimento de entidades mencionadas: O Segundo HAREM*. Linguateca.

Palmer, M., Gildea, D., and Kingsbury, P. (2005). The proposition bank: An annotated corpus of semantic roles. *Computational Linguistics*, 31(1):71–106.

Piai, L., Di-Felippo, A., and Roman, N. T. (2025). Named entities in stock market tweets: A fine-grained and linguistically-motivated annotation. In *Proceedings of the 16th Brazilian Symposium in Information and Human Language Technology: 10th Portuguese Description Journey (JDP 2025)*, pages 654–663, Fortaleza, CE, Brazil.

Plutchik, R. (2001). The nature of emotions: Human emotions have deep evolutionary roots, a fact that may explain their complexity and provide tools for clinical practice. *American Scientist*, 89(4):344–350.

Pradhan, S., Bonn, J., Myers, S., Conger, K., O'Gorman, T., Gung, J., Wright-Bettner, K., and Palmer, M. (2022). Propbank comes of age—larger, smarter, and more diverse. In Nastase, V., Pavlick, E., Pilehvar, M. T., Camacho-Collados, J., and Raganato, A., editors, *Proceedings of the 11th Joint Conference on Lexical and Computational Semantics*, pages 278–288, Seattle, Washington. Association for Computational Linguistics.

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

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

Sanguinetti, M., Bosco, C., Cassidy, L., and et al. (2023). Treebanking user-generated content: a ud based overview of guidelines, corpora and unified recommendations. *Language Resources Evaluation*, 57:493–544.

Silva, E., Pardo, T., Roman, N., and Fellipo, A. (2021). Universal dependencies for tweets in brazilian portuguese: Tokenization and part of speech tagging. In *Anais do XVIII Encontro Nacional de Inteligência Artificial e Computacional*, pages 434–445, Porto Alegre, RS, Brasil. SBC.

Silva, E. H., Pardo, T. A. S., and Roman, N. T. (2023). Etiquetagem morfossintática multigênero para o português do brasil segundo o modelo "universal dependencies". In *Proceedings of the 14th Symposium in Information and Human Language Technology (STIL 2023)*, Belo Horizonte, MG - Brazil.

Souza, E., Silveira, A., Cavalcanti, T., Castro, M., and Freitas, C. (2021). Petrogold – corpus padrão ouro para o domínio do petróleo. In *Anais do XIII Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana*, pages 29–38, Porto Alegre, RS, Brasil. SBC.

Straka, M., Hajič, J., and Straková, J. (2016). UDPipe: Trainable pipeline for processing CoNLL-U files performing tokenization, morphological analysis, POS tagging and parsing. In Calzolari, N., Choukri, K., Declerck, T., Goggi, S., Grobelnik, M., Maegaard, B., Mariani, J., Mazo, H., Moreno, A., Odijk, J., and Piperidis, S., editors, *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 4290–4297, Portorož, Slovenia. European Language Resources Association (ELRA).

Zeman, D. e. a. (2017). CoNLL 2017 shared task: Multilingual parsing from raw text to Universal Dependencies. In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 1–19, Vancouver, Canada. ACL.

Zerbinati, M. M., Roman, N. T., and Felippo, A. D. (2024). A corpus of stock market tweets annotated with named entities. In *Proceedings of the 16th International Conference on Computational Processing of Portuguese (PROPOR 2024)*, pages 276–284, Santiago de Compostela, Galicia/Spain.