Porttinari - a Large Multi-genre Treebank for Brazilian Portuguese

Thiago Alexandre Salgueiro Pardo[1], Magali Sanches Duran[1], Lucelene Lopes[1], Ariani Di Felippo[2], Norton Trevisan Roman[3], Maria das Graças Volpe Nunes[1]

[1]Núcleo Interinstitucional de Linguística Computacional (NILC)
Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo

[2]Núcleo Interinstitucional de Linguística Computacional (NILC)
Departamento de Letras, Universidade Federal de São Carlos

[3]Escola de Artes, Ciências e Humanidades, Universidade de São Paulo

taspardo@icmc.usp.br, {magali.duran, lucelene, arianidf}@gmail.com,
norton@usp.br, gracan@icmc.usp.br

Abstract. *This paper presents the project of a large multi-genre treebank for Brazilian Portuguese, called Porttinari. We address relevant research questions in its construction and annotation, reporting the work already done. The treebank is affiliated with the "Universal Dependencies" international model, widely adopted in the area, and must be the basis for the development of state of the art tagging and parsing systems for Portuguese, as well as for conducting linguistic studies on morphosyntax and syntax for this language.*

1. Introduction

Candido Portinari was one of the greatest artists of Brazil. He was born in Brodowski (in the São Paulo state) in 1903 and passed away in 1962 (in the city of Rio de Janeiro, in the Rio de Janeiro state), contributing to Brazilian culture and representing the challenging national reality in his work. His artistic qualities were internationally recognized and, according to the *Projeto Portinari*[1] kept by his son João Candido Portinari, the *Guerra* (War) and *Paz* (Peace) panels created for the United Nations organization are his masterpiece work, considered his "most universal" paintings.

It is not a coincidence that "Porttinari" was chosen to name the initiative that we present in this paper. More than an acronym (Porttinari stands for "PORTuguese Treebank"), it reminds us of the great challenges and equally great contributions that building a large treebank may bring to the Portuguese computational processing and linguistic studies. As defined by Jurafsky and Martin (2008), a treebank is a syntactically annotated corpus, where each sentence is paired with its parse tree, which usually contains the part of speech tag of each word in the sentence and the syntactic structuring of such words, in the form of relationships (in a dependency approach) or their building blocks (the phrases, in a constituency approach).

In the area of Natural Language Processing (NLP), a treebank may be used for developing/training tools as part of speech taggers and syntactic parsers, in charge of automatically uncovering some of the first levels of linguistic structuring of running texts. Such information is useful for several NLP applications, as sentiment analysis

[1] http://www.portinari.org.br/

(where the identification of nouns that represent entities and adjectives that modify nouns is relevant to determine what is being qualified), indexing and summarization (in which noun groups may be used to represent text content), grammar checking (where the related words may have to observe grammatical constraints, as number and gender agreement inside subjects) and machine translation (where predicates and their arguments in the original language must be syntactically ordered in the target language), among many others. More traditional linguistic investigations may also benefit from treebanks, which allow studying usual sentence structuring patterns, possible arguments and adjuncts of verbs and how they happen (and the diathesis alternations) and the behavior of some part of speech tags, among several other interesting research issues.

Research on syntax and parsing in NLP is not new for Portuguese, but, compared to the resources and tools for other languages (e.g., English), Portuguese may be considered a low resource language in this frontier. In order to fulfill this gap, we have proposed the construction of the Porttinari treebank. This paper introduces the treebank and its project details, discussing the relevant issues that a large corpus construction require. Initially aiming to rival in size the English best known reference (the Penn Treebank project of Marcus et al., 1993), other Porttinari distinguishable features include its filiation to the "Universal Dependencies" (UD) international model (Nivre, 2015; Nivre at al., 2020) and its proposal to be a multi-genre corpus in order to foster robust and general-use NLP, bringing relevant linguistic discussions into light and allowing to explore recent machine learning strategies (as transfer learning).

In what follows, we present the main related work in the area for Portuguese. The treebank project is reported in Section 3. Section 4 concludes this paper.

2. Related work

The Portuguese language counts with some treebanks, as *Floresta Sintá(c)tica* (Afonso et al., 2002; Freitas et al., 2008), CINTIL-DependencyBank (Branco et al., 2011), BDCamões DependencyBank (Parts I and II)[2], CORDIAL-SIN (Carrilho and Magro, 2010) and Tycho Brahe (Sousa, 2014), among others. *Floresta Sintá(c)tica* was probably the most representative effort for this language. Particularly interesting, this treebank has a manually revised portion called Bosque (with 9,364 sentences and 210,957 tokens), mapped to UD dependency annotation (Rademaker et al., 2017). Figure 1 shows an example of an UD-annotated sentence, in which one may see the words in their original forms, the syntactic relationships among them (in the level above) and their lemmas and part of speech tags (in the levels below).

The UD model is an attempt to standardize the morphosyntactic and syntactic analyses in the area, proposing an "universal" annotation strategy for all languages, as advocated by Nivre (2015). The model has been widely adopted for tagging and parsing tasks, having a large community of researchers who discuss its issues and contribute to the constant evolution of the model. Currently, there are already over 200 treebanks for more than 100 languages, and UD has become one of the dominant models in the area. For such reasons, we have adopted it as the basis for our treebank. Besides Bosque, which is the most popular treebank for Portuguese, the UD project also makes available

[2] Available at https://portulanclarin.net/

other 3 corpora for this language[3]: PUD (with 1,000 sentences and 21,917 tokens), GSD (with 12,078 sentences and 297,938 tokens) and DHBB (that, at the moment of the writing of this paper, had no available information).

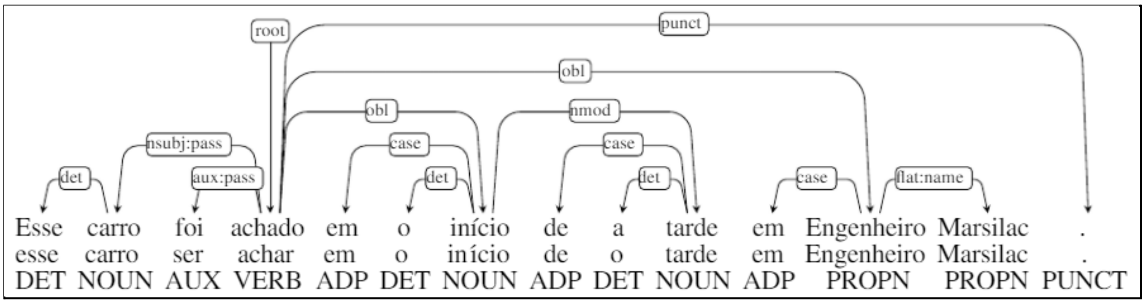
root
punct
obl
obl
nmod
nsubj:pass
case
case
det
aux:pass
det
det
case
flat:name
Esse carro foi achado em o início de a tarde em Engenheiro Marsilac .
esse carro ser achar em o início de o tarde em Engenheiro Marsilac .
DET NOUN AUX VERB ADP DET NOUN ADP DET NOUN ADP PROPN PROPN PUNCT

Figure 1. An example of UD-annotated sentence for Portuguese (reproduced from the work of Rademaker et al., 2017, p. 200)

Overall, the available manually (and, therefore, high quality) syntax annotated resources for Portuguese[4] are far from what other languages have access to. Consider, for instance, the English case, whose worldwide famous Penn Treebank project (Marcus et al., 1993) reports over 4.5 million words. In what follows, we report our efforts to overcome this historic limitation for the Portuguese language.

3. The rise of Porttinari

Porttinari is an ongoing initiative, aiming at growing syntax-based resources and fostering the development of related tools and applications for the Brazilian Portuguese language. The initiative started in 2020, counting with the collaboration of linguists and computer scientists, and is expected to be fully accomplished in the next few years.

Projecting the creation and annotation of a large multi-genre treebank is not a trivial task. There are several issues that must be addressed. Hovy and Lavid (2010) argue that corpus annotation is a science and that careful thought must be given to it. We present in what follows the details of the Porttinari effort according to the 7 research questions that Hovy and Lavid postulate.

3.1. Selecting the data

In order to create a multi-genre treebank, we initially selected two main genres that we consider to be more relevant to current NLP research: news texts, representing the standardized language, and the so called User-Generated Content (UGC), representing the language from web, marked by informality and produced by web users. This selection allows the study of different language genres and the creation of NLP tools and applications that may deal with varied writing styles. Another criterion for selecting material to the treebank was the license. As we expect that the treebank may subsidize other research, "open" licenses are desired, at least for the majority of the corpus.

[3] The interested reader may check them at https://universaldependencies.org/

[4] It is important to add that the other Portuguese corpora with non-UD dependency relations adopt solutions that are almost always not transferable to the UD annotation. Once we filliate to UD, we need to restrict ourselves to the set of part of speech tags and dependency relations that UD provides and to follow its guidelines.

We ended up with the following materials to be the basis for selecting the sentences that will compose the treebank, ordered from the more to the less public ones:

- news texts from *Folha de São Paulo*, from the years 2015 to 2017, made publicly available (license CC0) by Kaggle website[5] (composed by 167,053 texts);
- news texts from the MAC-MORPHO corpus, which is a 1.1 million word corpus originally developed by the Lacio-Web project (Aluísio et al., 2003), made publicly available (license CC-BY);
- stock market tweets from DANTE (Dependency-ANalised corpora of TwEets), publicly available (license GPL 2) by Silva et al. (2020), with 4,277 tweets;
- e-commerce customer reviews from B2W-reviews1 corpus, with more than 130,000 texts (license CC-BY-NC-SA), as described by Real et al. (2019);
- a small but challenging corpus of online book review sentences, as described by Belisário et al. (2020), with 350 texts (with no public license).

Such corpora sum up to nearly 80 million tokens, with almost 4 million sentences. Keeping the not fully public material (as the last two corpora in the list) is interesting for training/development (given the size of B2W-reviews1 corpus) and testing (given the difficulties of book reviews) of systems and language theories and models.

3.2. Instantiating the theory

Before starting the annotation, we investigated the UD model and built the Portuguese annotation guidelines. Although there were already three Portuguese UD corpora (PUD, GSD and Bosque), none of them released an annotation manual containing specific guidelines for Portuguese. At most, examples of the use of each part of speech tag or dependency relation in Portuguese (often examples translated from English) constituted the Portuguese (pt) tab of each item of the UD guidelines.

Directly translating guidelines may cause a detachment from UD theory. In English, for example, there are cases of two non-prepositional objects (as the dative construction "to give somebody something”) that UD annotates as *obj* and *iobj*. Both *obj* and *iobj* are considered core dependents in UD, and both should be able to be promoted to subject position in passive alternation. In Portuguese, however, the recipient of the dative verbs is always prepositional and cannot be converted into a passive voice subject. These cases, according to UD, should be annotated as *obl* and not *iobj* (for example, in English, the recipient of dative verbs introduced by “to” are *obl*: to give something *to somebody*). The translation and the fact that an indirect object is understood in traditional Portuguese grammar as a prepositional object could lead us to erroneously annotate prepositional objects as *iobj*. Aware that the instantiation of theory is much more than a translation task, we have carefully analyzed each of the 17 UD part of speech tags and 37 dependency relations in order to produce Portuguese UD guidelines full of examples and clarifications about our annotation decisions.

Several theoretical challenges have arisen. Just to illustrate a few cases[6], consider the part of speech annotation. UD prescribes that an adjective should always be

[5] https://www.kaggle.com/marlesson/news-of-the-site-folhauol

[6] There are so many challenges that they will be discussed in other papers, as we have space limitations here.

annotated as *ADJ*, even though it may exceptionally head a nominal phrase. In Portuguese, unlike English, many words can be categorized as either adjective or noun, depending on the context (even dictionaries bring both options). For example, in “cuidados médicos” (medical care), “médicos” is an *ADJ*, whereas in “médicos brasileiros” (Brazilian doctors), “médicos” is a *NOUN*. Because of this, we initially adopted context-based annotation: if the word was modified by an adjective, it was considered *NOUN*; if it was modifying a *NOUN*, it was considered *ADJ*. However, after starting annotation, we realized that in some situations[7] there was no clue to decide if a word was an *ADJ* or a *NOUN*. This led us to decide to annotate as *ADJ* the adjectives occurring as *head* of a dependency relation, but never modified by another adjective. For example, the adjective “melhor” (best) is annotated as *ADJ* in “Os melhores serão recompensados” (The best [ones] will be rewarded). Other challenges come from annotating tweets. Even though Sanguinetti et al. (2020) have already proposed a unified scheme for coherent UD treatment of social media in general and for Twitter across different languages, it was necessary to define criteria for certain phenomena typical of tweets mentioning stocks from the Bovespa index. A particular case found in these tweets are the stock codes, which are usually represented by five- or six-character alpha-numerical strings, such as “Petr4” for “Petrobras” and “BBAS3” for “Banco do Brasil”. These codes are so popular among investors that they are commonly used as surrogates for their company names. Because of the relevance in the domain, we annotate the stock codes as *PROPN*. One may also wonder how to deal with multiword expressions. The fact is that UD does not annotate such expressions at the part of speech level, and, at the syntactic level, it only does so for multiword expressions that have no syntactic relation between their tokens. Therefore, the expression “hot dog” is annotated as an ADJ modifying a NOUN. UD, as well, does not address light verb constructions[8], as “take advantage”, which is annotated as a VERB with a NOUN as complement.

Overall, the post-annotation report of the Bosque corpus (Souza et al., 2021) was of great help, as it brings many examples of sentences that generated annotation doubts, which allowed us to foresee problems even before starting our own annotation task.

3.3. Selecting and training the annotators

As mentioned by several authors, the background and training of annotators is an open question, since some researchers claim that they should be experts and others propose training annotators just adequately for the task at hand. Given the UD annotation, the strategy used in our project to select the annotators lies between these two viewpoints. We selected 10 undergraduate students in Linguistics or Letters courses with a reasonably similar grounding in morphosyntax and syntax and offered relatively extensive training based on the Portuguese annotation guidelines.

[7] For cases like this, previous solutions proposed for Portuguese may not work. For instance, the proposal in *Bíblia Florestal* (https://www.linguateca.pt/Floresta/BibliaFlorestal), i.e., keeping the two possible tags separated by a slash, is not feasible within the UD guidelines, which only allow the assignment of one tag.

[8] However, some UD discussion issues do address the possibility of annotating multiword expressions and light verb constructions. There are two current suggestions to take these constructions into account when annotating, involving the inclusion of the related information in a new annotation level or considering it as miscellaneous/additional information.

We did 2 weeks of annotation training before starting the annotation of the corpus. During this period, virtual meetings were held 2-3 times a week, both to discuss the guidelines and to correct annotation divergences, showing the options of each annotator in the annotation tool itself (which we introduce later). From then on, the training meetings became weekly and had as theme the issues that most generated doubts in the previous week. All training meetings were recorded and the corresponding used slides made available in a common access area for the group. We also maintain at Github a issue tracking system where annotators can express their doubts and the project adjudicator (i.e., a chief linguist that is expert on the annotation theory and that evaluates and makes decisions regarding the issues) can provide explanations.

After the initial training, we did 5 weeks of blind annotation with the 10 annotators on the same data and one adjudicator. This phase was rich in the sense that we could evaluate the performance of the annotators simultaneously on the same task, as well as analyze the confusion matrix, which showed the most difficult issues. To deal with the difficulties, we started to release specific studies for such cases, including disambiguation clues and examples of use, which also resulted in improvements in the annotation guidelines.

3.4. Specifying the annotation procedure and workflow

The annotation procedure starts with the automatic annotation of all sentences considered for each corpus of the Porttinari project. This initial annotation is carried out by the UDPipe system (Straka, 2018) trained over Bosque treebank, which produces state of the art results for Portuguese under the UD model. Each set of sentences compose a Repository of Automatic Annotated Sentences (RASS) that is stored using the CoNLL-U file format, which is traditionally used in the area (it is a column-based format in which each column stores the related information of the words in the lines).

The second step is the manual revision of the sentences of each RASS that is made by picking a set of sentences for human analysis. These sets may be randomly chosen or follow some specific criteria, for example, sentences of specific sizes or sentences having interesting patterns such as target tokens or tag sequence patterns. The set of chosen sentences defines a Manual Annotation Package (MAP).

The third step is to assign MAPs to the 10 trained linguists, organized in two to three groups, with each group receiving a shuffle copy of a different MAP. Shuffling MAPs aims at avoiding bias in the annotation and possible information sharing among the annotators. An extra protection comes from the fact that each annotator does not know which other annotators are in his/her group. Each annotator confirms/corrects the annotation of the sentences using a visual annotation tool (see next subsection). The output of this step is a set of CoNLL-U files with the revised annotated sentences.

The fourth step is the adjudication of each MAP by a chief linguist to provide correct and homogeneous annotation. This is done by integrating the CoNLL-U files from each annotator with the original automatic annotation and computing agreement metrics into Package Adjudication Reports (PAR). The PARs are then analyzed and the cases with disagreements are corrected by the adjudicator.

The final step consists of the incorporation of the adjudication into CoNLL-U files that are stored in the Repository of Revised Annotation Sentences (RRAS). This procedure is repeated for as many MAPs as necessary until the selected sentences of the corpora are duly manually revised. Once each adjudication process finishes, the reports with the result of the adjudication are sent to the annotators, so that they can identify the errors and learn from them. This strategy produced different responses: the annotators who were already performing well improved even more, and the annotators who were performing less well did not improve much, perhaps because they had basic deficiencies in grammatical literacy, such as difficulty in identifying passive voice, for example.

Underlying the annotation process is the decision to separately annotate the UD levels in order to simplify each task and to produce better results for each level. This is interesting as the UD annotation has shown to be a highly sophisticated task. We started by reviewing the part of speech tags of the words in each sentence. After that, the morphologic level is semi-automatically reviewed: we use the Unitex-PB lexicon (Muniz, 2004) to retrieve the relevant morphological features of each word and then ask some human annotators to review the difficult cases and those that are not in the lexicon. Finally, the dependency relations must be fully reviewed.

3.5. Designing the annotation interface

As UD annotation is a challenging task, a good annotation interface is very important to help the annotators to clearly and easily find the relevant information and manage it, to have the necessary available functionalities (as tree visualization, searching mechanisms and editing facilities) and to guarantee that the annotated data is saved and stored. We have extended and customized the Arborator-Grew (Guibon et al., 2020) tool to include new functionalities and to correct some bugs, producing a new version of it. The new functionalities include shortcuts for faster annotation, color-based facilities for helping the annotation process, automatic checking of some UD mandatory characteristics and advanced options for project management, among others.

3.6. Choosing and applying the evaluation measures

In order to evaluate the annotation procedure, we assessed the degree to which different annotators agree on their classifications. To do so, and since we were dealing with more than two annotators, we calculate the average agreement agr_i, amongst a set of c annotators, as Artstein and Poesio (2008) propose:

$$\mathrm{agr}_i = \frac{1}{\binom{\mathbf{c}}{2}} \sum_{k\in K} \binom{\mathbf{n}_{ik}}{2} = \frac{1}{\mathbf{c}(\mathbf{c}-1)} \sum_{k\in K} \mathbf{n}_{ik}(\mathbf{n}_{ik}-1)$$

where n_{ik} represents the number of annotators, assigning the label k, from a set of K possible labels, to the same token i. As defined, agreement values range from 0%, representing total disagreement (i.e., each annotator assigns a different label), to 100% (full agreement - all annotators assigned the same label to the token). Following the authors, we take values above 80% to represent significant inter-annotator agreement, with values between 67% and 80% allowing tentative conclusions only. Such effort helps us to evaluate how clear and reproducible the annotation task is and how annotators are understanding it, thereby increasing our confidence in the reliability of the annotation results.

Other interesting evaluation strategies are those presented by Santos and Gasperin (2002) for assessing parsed corpora, which shall be explored in the future.

3.7. Delivering and maintaining the product

The annotated portions of Porttinari must be periodically made publicly available at the project webpage. Once the treebank is ready, we also plan to make it available at the UD webpage and at the PORTULAN CLARIN portal, which is an infrastructure for research on language technologies and already includes NLP products for Portuguese.

For now, the treebank has been maintained by the efforts of the research group, which we expect to keep for the next years. Hopefully, its usefulness for the area will eventually justify its long term maintenance.

4. Final remarks

We have presented and discussed in this paper the procedures and decisions for the project of Porttinari, which shall be a large multi-genre treebank for Portuguese, affiliated with the Universal Dependencies international model. The contributions of this work include the treebank itself, the annotation process (detailed in this paper) and the theoretical issues of UD for Portuguese and the different text genres that we annotate.

So far, we have over 10,000 manually revised sentences for part of speech tagging[9], which, in sequence, were revised for lemmas and morphological features. The available data will be used to train a new tagger for Portuguese, which must produce better quality data to be reviewed, boosting the annotation process. The dependency relation revision must start in the next months.

The interested reader may find more information at the webpage of the POeTiSA project[10] (POeTiSA stands for *POrtuguese processing - Towards Syntactic Analysis and parsing*), where the related resources and tools are available (as the annotation manual and tool, the linguistic studies and the annotated portions of the corpus).

Acknowledgements

The authors are grateful to the Center for Artificial Intelligence of the University of São Paulo (C4AI - http://c4ai.inova.usp.br/), sponsored by IBM and FAPESP (grant #2019/07665-4).

Dedication

In memory of Andréia Gentil Bonfante, who very early attempted to tame the syntax and build a parser for Portuguese.

[9] Ongoing work has been confirming that this was a good decision. Since the correlation between part of speech tags and lemmas, morphological features and dependency relations is high, there has been a significant gain in the annotation.

[10] https://sites.google.com/icmc.usp.br/poetisa

References

Afonso, S.; Bick, E.; Haber, R.; Santos, D. (2002). Floresta sintá(c)tica: um treebank para o português. In Anais do XVII Encontro Nacional da Associação Portuguesa de Linguística, pp. 533-545.

Aluísio, S.M.; Pelizzoni, J.; Marchi, A.R.; Oliveira, L.; Manenti, R.; Marquiafável, V. (2003). An account of the challenge of tagging a reference corpus for brazilian portuguese. In the Proceedings of the 6th International Conference on Computational Processing of the Portuguese Language, pp. 110-117.

Artstein, R. and Poesio, M. (2008). Inter-coder agreement for computational linguistics. Computational Linguistics, Vol. 34, N. 4, pp. 555-596.

Belisário, L.B.; Ferreira, L.G.; Pardo, T.A.S. (2020). Evaluating Richer Features and Varied Machine Learning Models for Subjectivity Classification of Book Review Sentences in Portuguese. Information, Vol. 11, N. 9, pp. 1-14.

Branco, A.; Castro, S.; Silva, J.; Costa, F. (2011). CINTIL DepBank Handbook: Design options for the representation of grammatical dependencies. Technical Report DI-FCUL-TR-2011-03. University of Lisbon.

Carrilho, E. and Magro, C. (2010). A anotação sintáctica do CORDIAL-SIN. In A.M. Brito, F. Silva, J. Veloso and A. Fiéis (eds.), XXV Encontro Nacional da Associação Portuguesa de Linguística. Textos seleccionados, pp. 225-241.

Freitas, C.; Rocha, P.; Bick, E. (2008). Floresta Sintá(c)tica: Bigger, Thicker and Easier. In the Proceedings of the 8th International Conference on Computational Processing of the Portuguese Language, pp. 216-219.

Guibon, G.; Courtin, M.; Gerdes, K.; Guillaume, B. (2020). When Collaborative Treebank Curation Meets Graph Grammars: Arborator With a Grew Back-End. In the Proceedings of the 12th Conference on Language Resources and Evaluation, pp. 5291-5300.

Hovy, E. and Lavid, J. (2010). Towards a 'Science' of Corpus Annotation: A New Methodological Challenge for Corpus Linguistics. International Journal of Translation, Vol. 22, N. 1, pp. 13-36.

Jurafsky, D. and Martin, J.H. (2008). Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics and Speech Recognition. 2a edição. Prentice Hall.

Marcus, M.P.; Santorini, B.; Marcinkiewicz, M.A. (1993). Building a large annotated corpus of English: the penn treebank. Computational Linguistics, Vol. 19, N. 2, pp. 313-330.

Muniz, M.C.M. (2004). A construção de recursos lingüístico-computacionais para o português do Brasil: o projeto de Unitex-PB. MSc Dissertation. Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo. 72p.

Nivre, J. (2015). Towards a Universal Grammar for Natural Language Processing. In the Proceedings of the 16th International Conference on Intelligent Text Processing and Computational Linguistics, pp. 3-16.

Nivre, J.; Marneffe, M-C.; Ginter, F.; Hajič, J.; Manning, C.D.; Pyysalo, S.; Schuster, S.; Tyers, F.; Zeman, D. (2020). Universal Dependencies v2: An Evergrowing Multilingual Treebank Collection. In the Proceedings of the 12nd International Conference on Language Resources and Evaluation, pp. 4034-4043.

Rademaker, A.; Chalub, F.; Real, L.; Freitas, C.; Bick, E.; Paiva, V. (2017). Universal Dependencies for Portuguese. In the Proceedings of the 4th International Conference on Dependency Linguistics, pp. 197-206.

Real, L.; Oshiro, M.; Mafra, A. (2019). B2W-Reviews01 - An open product reviews corpus. In the Proceedings of the XII Symposium in Information and Human Language Technology, pp. 200-208.

Sanguinetti, M.; Bosco, C.; Cassidy, L.; ÇetinoĞlu, Ö.; Cignarella, A. T.; Lynn, T.; Rehbein, I.; Ruppenhofer, J.; Seddah, D.; Zeldes, A. (2020). Treebanking user-generated content: a proposal for a unified representation in universal dependencies. In the Proceedings of the 12th International Language Resources and Evaluation Conference, pp. 5240-5250.

Santos, D. and Gasperin, C. (2002). Evaluation of parsed corpora: Experiments in user-transparent and user-visible evaluation. In the Proceedings of the Third International Conference on Language Resources and Evaluation, pp. 597-604.

Silva, F.J.V.; Roman, N.T.; Carvalho, A.M.B.R. (2020). Stock market tweets annotated with emotions. Corpora, Vol. 15, N. 3, pp. 343-354.

Sousa, M.C.P (2014). O Corpus Tycho Brahe: contribuições para as humanidades digitais no Brasil. Filologia e Linguística Portuguesa, Vol. 16, pp. 53-93.

Souza, E.; Cavalcanti, T.; Silveira, A.; Evelyn, W.; Freitas, C. (2021). Diretivas e documentação de anotação UD em português (e para língua portuguesa). Available at https://nbviewer.jupyter.org/github/comcorhd/Documenta-o-UD-PT/raw/master/Documenta-o-UD-PT.pdf

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