The Dawn of the Porttinari Multigenre Treebank: Introducing its Journalistic Portion

Magali Sanches Duran, Lucelene Lopes, Maria das Graças Volpe Nunes, Thiago Alexandre Salgueiro Pardo

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

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

Abstract. *This paper introduces the journalistic portion of the Porttinari treebank, which aims to be a multigenre NLP resource for Brazilian Portuguese. We report the construction of the treebank, in particular, the human-revised portion with 8,418 sentences, whose annotation process lasted almost three years and involved more than a dozen trained annotators. The full treebank offers to the Portuguese-speaking NLP community nearly 4 million sentences annotated according to the Universal Dependencies framework.*

1. Introduction

Treebanks, corpora whose sentences are accompanied with their syntactic trees, play important roles for both linguistic and NLP (Natural Language Processing) research. For linguists, treebanks enable detailed analysis of syntax, leading to the discovery of linguistic patterns and a deeper understanding of language structure. In NLP, treebanks can be used for the development or improvement of computational language processing models, especially those that require labeled data with explicit syntax knowledge.

Among the theoretical options to build treebanks, the Universal Dependencies (UD) project (de Marneffe et al., 2021) stands out as an international grammar framework, with more than 140 languages available in more than 240 corpora. Portuguese has four corpora among the UD datasets: Bosque (Rademaker et al., 2017), PUD (Zeman et al., 2017), CINTIL (Branco et al., 2022), and PetroGold (Sousa et al., 2021). Together, these corpora result in less than a million tokens. In order to contribute on this front, we are committed to the construction of a large multigenre treebank for Portuguese, the Porttinari (which stands for "PORTuguese Treebank"). In this sense, this paper introduces the journalistic portion of Porttinari, which is its first annotated genre, composed by three subcorpora with different characteristics and purposes: **Porttinari-base**, a corpus revised in detail to serve as gold standard; **Porttinari-check**, a small corpus structurally similar to Porttinari-base to serve as testbed and to illustrate the contrast between manual and automatic annotation; and **Porttinari-automatic**, a large corpus that was automatically annotated.

The rest of this paper is organized as follows. In Section 2, we briefly present UD, adopted in the treebank. In Section 3, we present the three subcorpora of Porttinari, and, in Section 4, we draw our conclusions and outline future work.

2. The Universal Dependencies framework

UD (de Marneffe et al., 2021) is a language independent initiative originally designed to annotate morphology, part of speech (PoS) and syntactic dependency relations, in an approach inspired by Tesnière (2015) dependency grammar. UD currently has a fixed set of 17 PoS tags and 37 dependency relations, plus a non-fixed set of morphological features. The UD annotation scheme has enabled the training of automatic classifiers and several comparative studies of language typology. As an example, Figure 1 shows an annotated sentence according to UD. Above the sentence, it is possible to see the relations (labeled arcs), pointing from the heads to their dependents; below them, there are the PoS tags of the words, their lemmas and their morphological features.

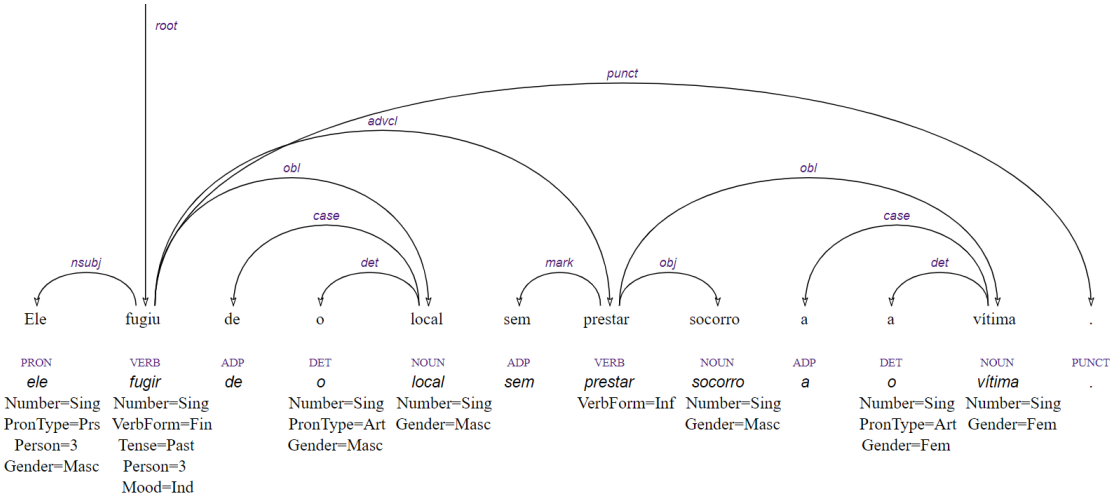

Figure 1. An example of sentence annotated according to UD framework

Much of the annotation in the UD scheme can be accomplished simply by transferring into Portuguese the guidelines described and exemplified in English. However, a number of issues typical of the Portuguese language required linguistic studies and annotation decisions, including phenomena as auxiliary verbs (Duran et al., 2021a), numerals (Duran et al., 2021b) and comparatives (Duran et al., 2023a), among others.

3. The journalistic portion of Porttinari

Porttinari is a large multi-genre treebank (Pardo et al., 2021) and the three subcorpora presented here, containing news texts, are the foundation for the other genres that will follow. Specifically, we used the Folha-Kaggle dataset[1], publicly available, composed of 167,053 news articles extracted from the electronic edition of the Brazilian newspaper Folha de São Paulo published from January 2015 to September 2017.

For preprocessing the corpus, we performed the sentencing and tokenization of each news article. In this phase, Portuguese characteristics had to be taken into account, as the truncated form of infinitive verbs when employed with clitic pronouns (e.g.,

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

"*fazê-lo*" was tokenized into the words "*fazer*" and "*lo*") and the contracted words (e.g., "*na*" was tokenized into the words "*em*" and "*a*"). The ambiguities of some contracted words were of particular interest in this part of the process, and we had to use heuristics to solve them. Some examples are the words "*consigo*" (that can be either an inflection of the verb "*conseguir*" or a contracted word decomposable into "*com*" and "*si*") and "*nos*" (that can be a clitic pronoun or a contraction of the words "*em*" and "*os*").

The preprocessing also assigned an ID to each produced sentence with the generic format FOLHA_DOCxxxxxx_SENTxxx, where DOC has a number between 000001 and 167048, referring to the news article the sentence comes from, and SENT has a number referring to the order of the sentence within the news article (no article has more than 999 sentences). For example, the ID FOLHA_DOC006009_SENT013 corresponds to the thirteenth sentence in the 6,009th news article of the original dataset.

Since 5 documents of the original dataset contained no text, the 167,048 news articles resulted in 3,964,292 sentences and 94,799,734 tokens. The three journalistic subcorpora (Porttinari-base, Porttinari-check, and Porttinari-automatic) were produced from this material. Figure 2 depicts the relative size of the three subcorpora in number of sentences. In what follows, we describe each of them.

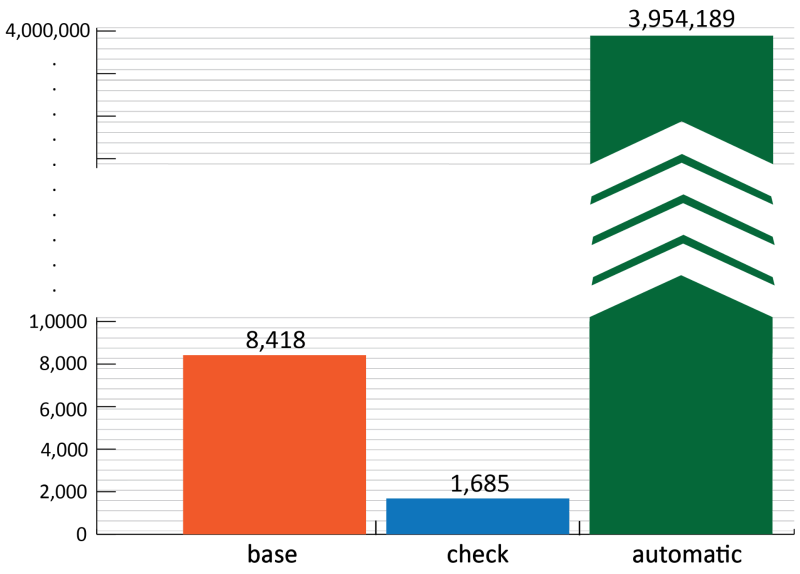

Figure 2. The size of the three subcorpora in terms of number of sentences

3.1. The Porttinari-base subcorpus

The Porttinari-base subcorpus is composed of 8,418 sentences (168,080 tokens - in average, 19.97 tokens per sentence) selected from the initial 5,000 news documents of Folha-Kaggle. The original Folha-Kaggle dataset includes topic classification, as world, economy, education, etc. Therefore, we avoided incorporating full news documents to provide a better diversity of authors and subtopics. The specific choice of sentences was made giving preference for sentences of sizes from 10 to 40 tokens, since small sentences are sometimes too simple and other times are just a juxtaposition of words with no syntax, while too large sentences usually repeat patterns of combined clauses,

either by coordination or subordination. Nevertheless, we also included some sentences out of those bounds, for diversity reasons. Figure 3 shows the distribution of sentences per number of tokens. The resulting subcorpus also has a good lexical diversity of verbal predicates as it has 1,576 different verb lemmas (534 verbs appear only once).

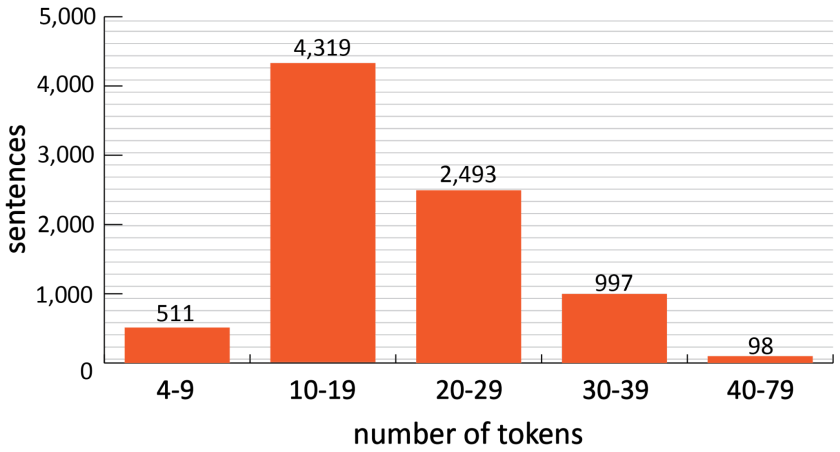

Figure 3. Number of sentences per number of tokens in Porttinari-base

The Porttinari-base annotation process started with an automatic annotation by the parser UDPipe 2 (Straka, 2018) using the Bosque-UD model (Rademaker et al., 2017), which achieves 87% of accuracy (in particular, the Labeled Attachment Score). As the task of revising syntactic trees is complex, we chose to separate the activity into three steps: first we revised the morphosyntactic layer, in which PoS tags are assigned; then we revised the morphological layer, i.e., the lemma and the features of the tokens, in a semi-automated step; and finally we revised the syntactic dependency relations.

The PoS tag manual annotation was performed by trained human annotators, using the Arborator-NILC editor (Miranda and Pardo, 2022) and following strictly the definition made by the PoS tag directives manual (Duran, 2021). The PoS tag distribution achieved is depicted in Figure 4. One of the challenges of PoS tag annotation was the ambiguity of function words, as prepositions and conjunctions, which are sometimes ambiguous with words from other PoS tag classes. Several lexical studies were made to support this task (Lopes et al., 2021; Lopes et al., 2023).

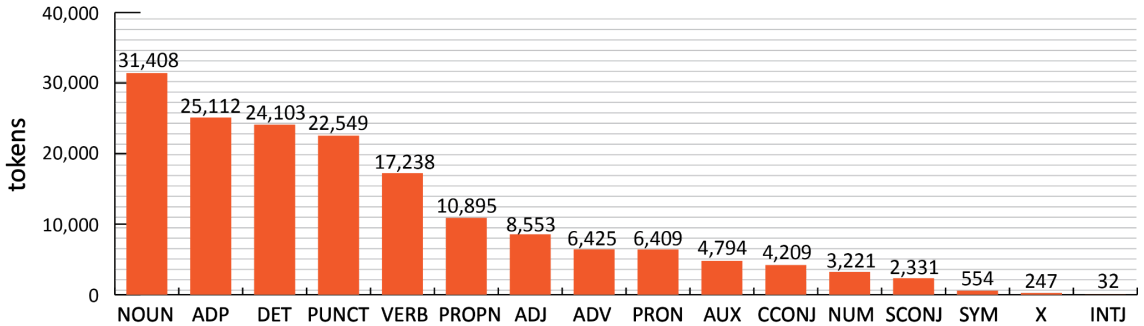

Figure 4. PoS tag distribution in Porttinari-base

The revision of lemmas and morphological features was supported by PortiLexicon-UD (Lopes et al., 2022), which holds the Portuguese forms associated with their PoS tag, lemma, and morphological features using the UD standards (feature names and values). For example, for verbs (either VERB or AUX), the lemma is the infinitive form of the verb and the features include the form of the verb (finite, infinitive, gerund, or participle), and, according to the form, it may include number (singular or plural), gender (feminine or masculine), person (1st, 2nd, or 3rd), mood (indicative, subjunctive, imperative, or conditional), and tense (present, past, future, imperfect, or pluperfect). Human intervention was required to disambiguate cases where more than one combination was possible. For example, the form "*for*" is always a verb, but may have as lemma the verbs "*ir*" or "*ser*".

The revision of dependencies was a more challenging task. The attribution of dependency relations starts with the definition of the root token, and the other relations are defined from it. Functional words (prepositions, conjunctions, and determiners) usually are only dependents of the relations, while content words (nouns, verbs, adjectives, and adverbs) can be either head or dependent of relations. Some phenomena typical of Portuguese are not predicted in the UD, forcing us to decide how to annotate them consistently throughout the corpus.

The dependency relation annotation becomes an even more complex task when using subrelations, a way to subspecify some of the original 37 UD relations. In our corpora, we adopted 10 subrelations (acl:relcl, aux:pass, ccomp:speech, csubj:outer, csubj:pass, flat:foreign, flat:name, nsubj:outer, nsubj:pass, obl:agent). Given that we do not employ 4 of the original 37 relations (clf, compound, dep, goeswith), we have an overall number of 43 dependency relations distributed as depicted by Figure 5.

The revision of the dependency relations was carefully manually executed over all sentences sequentially, but, due to its natural complexity, we performed an additional vertical revision of several linguistic phenomena. Such a vertical analysis step grants more confidence on the homogeneity of the produced data.

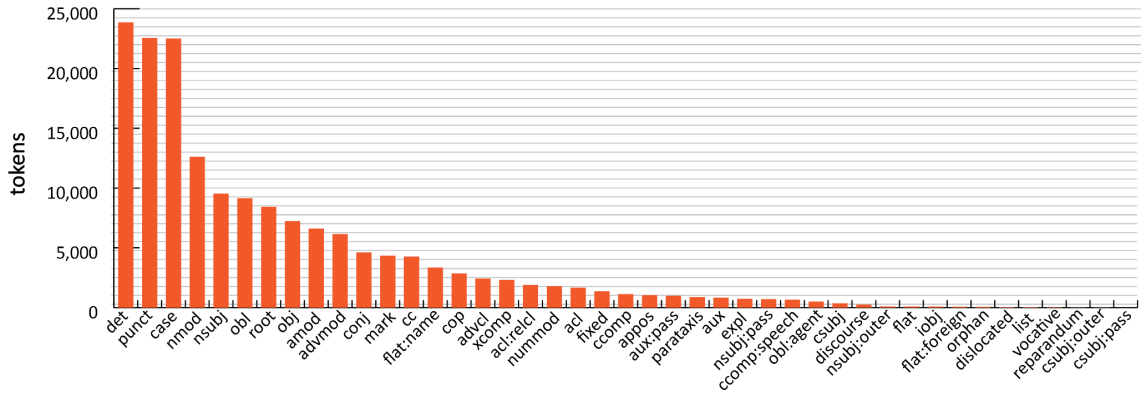

Figure 5. Relation distribution in Porttinari-base

For NLP purposes, Porttinari-base is further subdivided into three subsets: training, validation and test. Such division may be relevant for those interested in training and testing systems using this subcorpus. The subsets have 5,893, 842 and 1,683 sentences in train (70%), dev (10%), and test (20%) files, respectively.

3.2. The Porttinari-check subcorpus

The Porttinari-check subcorpus was developed to provide a controlled dataset to serve as testbed and as a contrast between the careful annotation of the Porttinari-base subcorpus and an automatic annotation. As such, we randomly chose a set of 1,685 sentences (about 20% of the size of Porttinari-base) (consisting of 33,547 tokens, and an average of 19.91 tokens per sentence) with similar characteristics. Specifically, we searched sentences to achieve a proportional distribution of sentence sizes, and similar distributions of PoS tags and dependency relations.

The annotation process of Porttinari-check was fully automatic with UDPipe 2 (Straka, 2018) using Porttinari-base as training set (which, in a preliminary evaluation, showed accuracy results over 98% for PoS tags and 91% for dependency relations).

As a testbed, Porttinari-check was designed to be a complementary evaluation resource. It may be used to check and test (quantitatively or qualitatively) parsing techniques in more varied ways, to complement other evaluation procedures and conclusions (providing varied test sets), to search for and assess specific grammar constructions and to subsidize other studies. As an example, it was already used as the basis for a detailed qualitative evaluation of a parser (Duran et al., 2023b).

3.3. The Porttinari-automatic subcorpus

The annotation of the Porttinari-automatic subcorpus was done entirely automatically, in the same manner as the annotation of Porttinari-check. Although the annotation is completely automatic, careful preprocessing contributes to good results. The quality of the training corpus annotation is critical to ensure that the automatic annotation is consistent.

The purpose of Porttinari-automatic is to produce a very large linguistic resource of NLP for Brazilian Portuguese. It has all the remaining sentences taken from the Folha-Kaggle dataset. As a result, Porttinari-automatic has 3,954,189 sentences, and 94,598,107 tokens, with an average of 23.92 tokens per sentence.

3.4. Overview of distributions in the three subcorpora

We have shown before the distribution of sentence sizes, PoS tags and dependency relations for Porttinari-base only, given it was the seed of the treebank annotation. Now, for comparison purposes, Figures 6, 7 and 8 show such distributions for the three subcorpora.

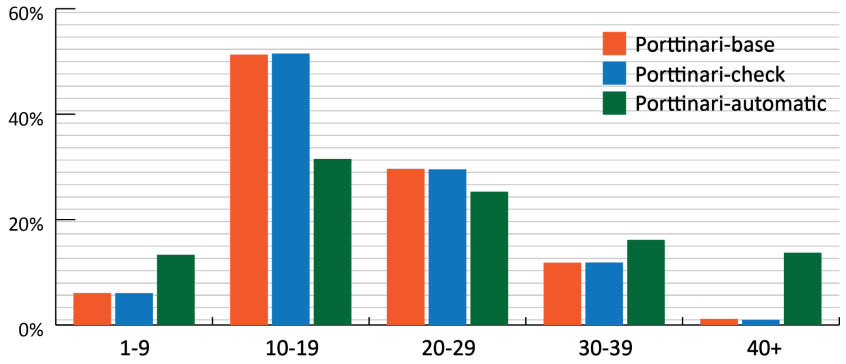

Figure 6. Percentage distribution of sentence sizes in Porttinari-base, Porttinari-check, and Porttinari-automatic

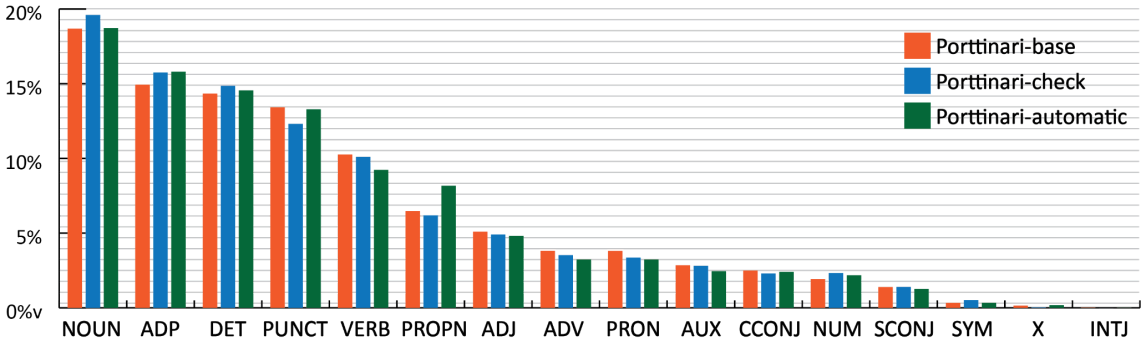

Figure 7. Percentage distribution of PoS tags in Porttinari-base, Porttinari-check, and Porttinari-automatic

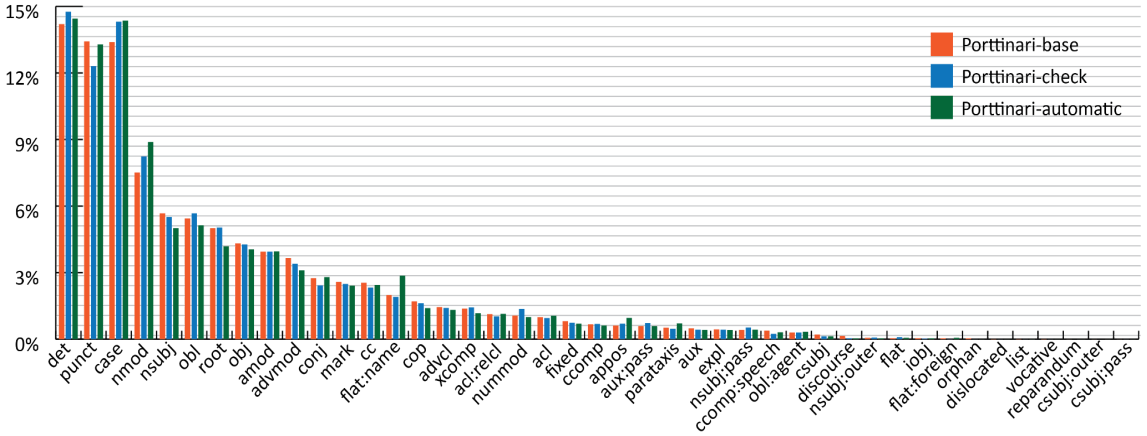

Figure 8. Percentage distribution of dependency relations in Porttinari-base, Porttinari-check, and Porttinari-automatic

It is possible to see that Porttinari-base and Porttinari-check are very similar in their distributions, as Porttinari-check was designed this way. As Porttinari-automatic includes all the remaining sentences of Folha-Kaggle dataset, it diverges from the other subcorpora. Proportionally, it has fewer sentences in the 10-19 size interval and more

sentences in the 40+ interval. Interestingly, it also shows more PROPN PoS tags, and, consequently, more flat:name relations.

4. Conclusion and Future Work

This paper announces the journalistic genre portion that integrates the large multigenre Porttinari treebank, offering the Portuguese-speaking community a treebank of nearly 4 million sentences annotated with dependency syntax in the UD framework. Together, the three subcorpora that compose the journalistic genre add up to 3,964,292 sentences and 94,799,734 tokens, with an average of 23.91 tokens per sentence. To put this resource in perspective, the Bosque-UD treebank, also for the journalistic genre, has 9,364 sentences and 227,825 tokens, with an average of 24.12 tokens per sentence. The addition of Porttinari data in this scenario places the amount of Portuguese UD annotated resources at the same level of well-resourced languages, which opens several possibilities for NLP applications and linguistic studies.

Future work includes (i) the annotation of enhanced dependencies, which, according to Nivre et al. (2018), have proven to be useful for more advanced applications, (ii) the annotation of semantic roles, following the Propbank model (Palmer et al., 2005), and (iii) the exploration of the treebank for developing NLP research products for Portuguese. In the near future, we also envision to announce a new genre in Porttinari, in particular, tweets, whose annotation process is already advanced.

For the interested reader, Porttinari and related materials are publicly available at the webportal of the POeTiSA project (https://sites.google.com/icmc.usp.br/poetisa).

Acknowledgments

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

Branco, A.; Silva, J.R.; Gomes, L.; Rodrigues, J.R. (2022). Universal grammatical dependencies for Portuguese with CINTIL data, LX processing and CLARIN support. In the Proceedings of the Thirteenth Language Resources and Evaluation Conference (LREC), pages 5617-5626.

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

Duran, M.S.; Rassi, A.P.; Pagano, A.S.; Pardo, T.A.S. (2021a). On auxiliary verb in Universal Dependencies: untangling the issue and proposing a systematized annotation strategy. In the Proceedings of the Sixth International Conference on Dependency Linguistics (Depling), pages 10-21.

Duran, M.S.; Lopes, L.; Pardo, T.A.S. (2021b). Descrição de numerais segundo modelo Universal Dependencies e sua anotação no português. In the Proceedings of the VII Workshop on Portuguese Description (JDP), pages 344-352.

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. Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo. São Carlos-SP, Setembro, 55p.

Duran, M.S.; Lopes, L.; Nunes, M.G.V.; Pardo, T.A.S. (2023a). Construções comparativas em português e sua anotação usando a sintaxe de dependências. Revista da ABRALIN. To appear.

Duran, M.S.; Nunes, M.G.V.; Pardo, T.A.S. (2023b). Avaliação qualitativa do analisador sintático UDPipe 2 treinado sobre o córpus jornalístico Porttinari-base. Relatório Técnico do ICMC 442. Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo. São Carlos-SP, Abril, 58p.

Nivre, J.; Marongiu, P.; Ginter, F.; Kanerva, J.; Montemagni, S.; Schuster, S.; Simi, M. (2018). Enhancing Universal Dependency Treebanks: A Case Study. In the Proceedings of the Second Workshop on Universal Dependencies, pages 102-107.

Tesnière, L. (2015). Elements of Structural Syntax. Tradução de OSBORNE, Timothy; KAHANE, Sylvain. Amsterdam: John Benjamins.

Miranda, L.G.M.; Pardo, T.A.S. (2022). An Improved and Extended Annotation Tool for Universal Dependencies-based Treebank Construction. In the Proceedings of the PROPOR Demonstrations Workshop, pages 1-3.

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

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, pages 197-207.

Lopes, L.; Duran, M.S.; Pardo, T.A.S. (2021). Universal Dependencies-based PoS Tagging Refinement through Linguistic Resources. In the Proceedings of the 10th Brazilian Conference on Intelligent System (BRACIS), pages 601-615.

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

Lopes, L.; Fernandes, P.; Duran, M.S.; Inácio, M.L.; Pardo, T.A.S. (2023). Disambiguation of Universal Dependencies Part-of-Speech Tags of Closed Class Words in Portuguese. In the Proceedings of the 12th Brazilian Conference on Intelligent Systems (BRACIS). To appear.

Palmer, M.; Gildea, D.; Kingsbury, P. (2005). The Proposition Bank: A Corpus Annotated with Semantic Roles. Computational Linguistics, 31(1), pages 71-105.

Pardo, T.A.S.; Duran, M.S.; Lopes, L.; Di Felippo, A.; Roman, N.T.; Nunes, M.G.V. (2021). Porttinari - a large multi-genre treebank for brazilian portuguese. In the Proceedings of the XIII Symposium in Information and Human Language (STIL), pages 1-10.

Zeman, D.; Popel, M.; Straka, M.; Hajic, J.; Nivre, J.; Ginter, F.; Luotolahti, J.; Pyysalo, S.; Petrov, S.; Potthast, M.; Tyers, F.; Badmaeva, E.; Gokirmak, M.; Nedoluzhko, A.; Cinkova, S.; Hajic Jr, J.; Hlavacova, J.; Kettnerova, V.; Uresova, Z.; Kanerva, J.; Ojala, S.; Missila, A.; Manning, C. D.; Schuster, S.; Reddy, S.; Taji, D.; Habash, N.; Leung, H.; de Marneffe, M.-C.; Sanguinetti, M.; Simi, M.; Kanayama, H.; Paiva, V.; Droganova, K.; Martinez Alonso, H.; Çoltekin, Ç.; Sulubacak, U.; Uszkoreit, H.; Macketanz, V.; Burchardt, A.; Harris, K.; Marheinecke, K.; Rehm, G.; Kayadelen, T.; Attia, M.; Elkahky, A.; Yu, Z.; Pitler, E.; Lertpradit, S.; Mandl, M.; Kirchner, J.; Alcalde, H. F.; Strnadova, J.; Banerjee, E.; Manurung, R.; Stella, A.; Shimada, A.; Kwak, S.; Mendonca, G.; Lando, T.; Nitisaroj, R.; Li, J. (2017). CoNLL 2017 shared task: Multilingual parsing from raw text to universal dependencies. In the Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies, pages 1-19.

Souza, E.; Silveira, A.; Cavalcanti, T.; Castro, M.; Freitas, C. (2021). PetroGold – corpus padrão ouro para o domínio do petróleo. In the Proceedings of the XIII Symposium in Information and Human Language (STIL), pages 29-38.