UDFest-BR 2023

2nd Edition of the Universal Dependencies Brazilian Festival

Proceedings of the Conference, Vol. 1

September 25, 2023

About the workshop

The 2nd Edition of the Universal Dependencies Brazilian Festival (UDFest-BR 2023) takes place with the 14th Symposium in Information and Human Language Technology (STIL 2023). It is a forum in which researchers involved with the study and application of the Universal Dependencies (UD) model, along with its adaptation to Portuguese, can meet and discuss best practices and strategies, as well as problems and solutions related to this topic.

UD is a cross-language international proposal for grammatical annotation (morphosyntactic tags, lexical/morphosyntactic features and syntactic dependencies) that resulted from the efforts of an open community of hundreds of researchers. Building upon a common framework, the model captures idiosyncrasies and similarities between different languages, fostering contrastive linguistic studies, along with the development of multilingual technologies for Natural Language Processing (NLP). In recent years, this model has drawn considerable attention from the NLP community. Currently, there are nearly 200 UD annotated corpora available in more than 100 languages.

Although there are some annotated corpora and lexical resources for Portuguese, computational linguistic research involving UD and this language is still limited. As such, this workshop helps to fill in this gap, giving more visibility to current efforts in this area and fostering new initiatives and collaborations.

This edition of the workshop includes ten papers. There are theoretical and practical discussions, reports on corpus annotation and the related challenges, and software for UD-related processing. The papers and their authors are:

- *Lexical noun phrase chunking with Universal Dependencies for Portuguese* (Aleksander Tomaz de Souza, Evandro Eduardo Seron Ruiz)

- *Construções sintáticas do português que desafiam a tarefa de parsing: uma análise qualitativa* (Magali Sanches Duran, Maria das Graças Volpe Nunes, Thiago A. S. Pardo)

- *Annotation of fixed Multiword Expressions (MWEs) in a Portuguese Universal Dependencies (UD) treebank: Gathering candidates from three different sources* (Elvis de Souza, Cláudia Freitas)

- *Verifica-UD: a Verifier for Universal Dependencies Annotation for Portuguese* (Lucelene Lopes, Magali Sanches Duran, Thiago Alexandre Salgueiro Pardo)

- *Enhanced dependencies para o português brasileiro* (Adriana S. Pagano, Magali Sanches Duran, Thiago Alexandre Salgueiro Pardo)

- *A dependency-based study of medicine package inserts in Brazilian Portuguese* (Adriana S. Pagano, André V. Lopes Coneglian, Lucas Emanuel Silva e Oliveira)

- *Um estudo das construções dar + para + V [infinitivo] nas Universal Dependencies* (Marcella M. Lemos Couto, Oto Araújo Vale)

- *Insights into the UD Tagset: Unveiling its Intricacies* (Magali Sanches Duran)

- *Em Direção à Anotação Sintática - UD de Tweets do Mercado Financeiro* (Bryan K. S. Barbosa, Ariani Di Felippo)

- *Universal Dependencies and Language Contact Annotation: Experience from Warao refugees signs in Brazil* (Dalmo Buzato)

Organizing committee

This workshop is promoted by the POeTiSA project, which is a long term initiative that aims at growing syntax-based resources and developing related tools and applications for Brazilian Portuguese language, looking to achieve world state-of-the-art results in this area. The project is part of the Natural Language Processing initiative (NLP2) of the Center for Artificial Intelligence (C4AI) of the University of São Paulo, sponsored by IBM and FAPESP. The workshop is also supported by the Ministry of Science, Technology and Innovation, according Law N. 8,248, of October 23, 1991.

The following researchers organized this edition of the workshop:

Thiago Alexandre Salgueiro Pardo, Magali Sanches Duran, Lucelene Lopes
Instituto de Ciências Matemáticas e de Computação, *Universidade de São Paulo*
São Carlos/SP, Brazil

Program committee

Adriana Silvina Pagano (UFMG)
Alexandre Rademaker (IBM)
Ariani Di Felippo (UFSCar)
Cláudia Freitas (PUC-Rio)
Joakim Nivre (Uppsala University)
Jorge Baptista (*Universidade do Algarve*)
Leonardo Zilio (UFRGS)
Maria das Graças Volpe Nunes (USP)
Norton Trevisan Roman (USP)
Roana Rodrigues (UFS)
Valeria de Paiva (Topos Institute)

Lexical noun phrase chunking with Universal Dependencies for Portuguese

Aleksander Tomaz de Souza[2]
Evandro Eduardo Seron Ruiz[1,2]

[1] Center for Artificial Intelligence (C4AI) – INOVA.USP
Butantã, São Paulo, SP – Brasil

[2]Departamento de Computação e Matemática
Faculdade de Filosofia, Ciências e Letras de Ribeirão Preto (FFCLRP)
Universidade de São Paulo – USP
Ribeirão Preto, SP – Brasil

[aleksander, evandro]@usp.br

Abstract. *Partial parsing retrieves a limited amount of syntactic information from a sentence. This project describes the identification of a specific type of noun phrase, through partial syntactic analysis, defined as a lexical noun phrase (NP_L), in texts written in Brazilian Portuguese, and annotated according to the Universal Dependency (UD) formalism. The Transformation Based Learning algorithm, TBL–Brill, applied as baseline, obtained an accuracy of 87.42% considering the UD dependency relations and 91.44% considering the UD morphosyntactic tags. Two other classifiers, one based on binary trees and the other based on a decision forest, had inferior performance.*

1. Introduction

Partial parsing, or *shallow parsing*, refers to a set of Natural Language Processing (NLP) methods aiming to retrieve a limited amount of syntactic information from a sentence. A peculiar application of *shallow parsing* that seeks to define distinct syntagmatic segments (noun phrases, verb phrases, adjective phrases, prepositional phrases, among others constituents within the text) is called *text chunking* [Hammerton et al. 2002]. Of these phrases, the noun phrases (NP) are relevant for discriminating elements with a substantive meaning and fulfill thematic roles within a sentence, encompassing functions like subject-object or agent-instrument relationships [Silva and Koch 2012].

Considering the conceptual models that categorize noun phrases, the authors Oliveira and Freitas [Oliveira and Freitas 2006] proposed the Lexical Noun Phrase (from now on NP_L), a specific NP that allows substantives, prepositions, adjectival phrases, among others, in its domains [Tjong Kim Sang and Buchholz 2000]. This type of noun phrase is critical in information retrieval and therefore is helpful for document indexing.

Currently, the Universal Dependency grammar formalism [Marneffe et al. 2021], also known as UD, is highlighted in the computational linguistics scenario. The UD is a framework for grammatical annotations across different existing natural languages. In this context, the *Center for Artificial Intelligence* (C4AI)[1], through one of

[1]https://c4ai.inova.usp.br/

its fronts, the *Natural Language Processing to Portuguese* (NLP2), seeks to enhance data and tools to enable high-level NLP of the Portuguese language, such as the Portinari project [Pardo et al. 2021], a corpus in the order of 10 million *tokens* annotated in this UD formalism. With this, in this research, we propose the identification of NP_L using morphosyntax tags (UD PoS Tag) and UD dependency relations. Due to the complexity of these phrases, the composition of rules for identifying the NP_L does not dispense a pattern recognition task through machine learning (ML) [Ramshaw and Marcus 2002]. Thus, this work increases the results obtained by [Souza and Ruiz 2022] in identifying this type of phrase.

2. Theoretical references

Syntax analysis, or parsing, is the process of analyzing and proposing an implicit grammatical structure to a sentence. Through syntactic analysis, one can determine textual patterns and understand the meanings of a sequence of terms of a logical and comprehensible structure [Jurafsky and Martin 2021]. Classical literature establishes two syntactic theories for grammatical annotation, which are: (a) constituent analysis [Chomsky 2009], and; dependency grammar [Tesnière 2015, Hjelmslev 1975]. The distinction between the two types of annotation is because the first is based on the structures of overlapping phrases; while the second is based on dependency relationships, or (in(ter))dependence[2], existing between the terms of a sentence [Pagani 2015]. Dependency grammar emphasizes the idea that linguistic units, such as words, are interconnected and interrelated.

The two theories considered in this research portray similar syntactic structures from different perspectives [Rambow 2010]. We emphasize that in the syntax of dependencies, the natural syntagmatic markings of constituents are absent and, therefore, are not made explicit. The NP_L can be characterized by presenting different syntactic lexical signatures [Souza and Ruiz 2022] and, in its extension, different gradients of complexity [Elhadad 1996].

For a brief description of NP_L one may notice that its identification, as in the example 1 below, can be trivial. However, in specific examples, whereas the NP_L are marked in bold, it is clear this identification can become a complex activity, such as we can see in the examples 2, in which the term **caneta** is a NP_L and **papel** another, and the example 3 which corresponds to a more extensive presentation of these phrases. See the examples below:

1. **A caneta** é esferográfica.
2. **Caneta** e **papel** para escrever.
3. **Caneta esferográfica Montblanc** para escrever em **papel apergaminhado de cor sépia**.

Wherever named entities, such as proper names, names of government entities or institutions, and geographic locations are encountered, they should be considered a single element. See the example 4, below, in which **João Pessoa** should be understood as a unit preserving the coordination with the term **Maceió**, that is, two distinct cores of NP_L:

[2]With the term (in(ter))dependence, we are abbreviating the three possibilities of this type of relationship: (non-reciprocal) dependence, independence (mere concatenation, without dependence on any part) and interdependence (reciprocal).

4. **João Pessoa** e **Maceió** são capitais de estados brasileiros.

The noun core restriction disregards as NP_L segments in which pronouns (example 5) and numerals play the role of central element (see example 6) because they are 'anaphoric reference to another lexical or clause element in the discourse [Oliveira and Freitas 2006].

5. **Elas** são capitais de estados brasileiros.
6. **Os três** são bons livros.

3. Related work

Ramshaw and Marcus used the Transformation-Based Learning methodology, TBL, for shallow parsing in phrase identification. The TBL method obtained precision and recall in the order of 92% for NP and 89% for other types of English phrases. Hammerton and co-authors [Hammerton et al. 2002] showed several different NLP applications that do not dispense phrasal identification. In another study on identifying noun phrases, Tham [Tham 2020] obtained an accuracy of 85.0%, and an F-measure of 85.12%. In addition to the applications on the English language, machine learning methods associated with shallow parsing were used in the Turkish language [Topsakal et al. 2017], for Hindi-English [Sharma et al. 2016], and also for Portuguese-English translation in a project developed at the University of Alicante by Garrido Alenda [Garrido Alenda et al. 2004]. Phrasal segmentation of texts written in Portuguese was also accomplished by Silva [da Silva 2007] as a shallow parser based on finite state automata. According to the researched literature, the work of Ophélie Lacroix [Lacroix 2018] introduced the identification of English NP chunks annotated under the *Universal Dependency* formalism. Following Lacroix's methodology, Souza e Ruiz [Souza and Ruiz 2022] had previously achieved an accuracy of 87.0% for the identification of NP_L and F-measures in the order of 85.3% for texts written in Portuguese in the UD context.

4. Data and methods

Data

As NP_L have their origin in the annotation noun phrases, we selected the Brazilian Portuguese Bosque *corpus* as it is annotated under two contexts: a) in the constituent grammar context, the Bosque 8.0 [Afonso et al. 2002], and; b) in the context of *Universal Dependency*, the UD_Portuguese-Bosque 2.10 [Rademaker et al. 2017]. Considering both corpora, each sentence is annotated under both theories, the constituent grammar and in the UD formalism. This way, we were able to analyze the declared syntactic structures through the hierarchical constituent model (Bosque 8.0), as well as explore their internal structures, typology, and hierarchical topology under the UD model (UD_Portuguese-Bosque 2.10). In that way, the NP_L were manually annotated in an empty field of the CoNLL-X UD file structure, a typical file structure for the *Universal Dependencies* project.

Here, we use a subset of the first 790 sentences annotated in both corpora. Table 1 describes this subset in the following fields: the number of sentences, words/*tokens*, CoNLL fields, UD Relations, and UD PoS Tags existing in the used corpus.

Table 1. The working corpus extracted from Bosque.

# Sentences	790
# Tokens	16,672
# CoNLL fields	10
# UD Relations	38
# UD PoS Tag	16
Data quantity	1020 kB

Table 2 depicts the data available to the algorithms. This table presents the terms categorized in the following fields: *token, deprel, upos, deps, IOB-format*, respectively: the corresponding word/token, the type of dependency relationship projected from the token (*deprel*), its morphosyntactic class (*upos*), its level under the dependency hierarchy (*deps*), and the marker considering the (*IOB-format*).

Table 2. Tokens and corresponding tags under UD and IOB format.

token	*deprel*	*upos*	*deps*	*IOB-format*
Averróis	root	PROPN	0	B
no	-	-	-	I
em	case	ADP	2	I
o	det	DET	3	I
poder	nmod	NOUN	1	I

Methods

Previous approaches that resorted to the use of ML for pattern identification demonstrate significant results in the application of abstract methods for composing rules using tags in the IOB format [Ramshaw and Marcus 2002]. This choice of performing certain language structures with tags that correspond to segments of interest in the text also proved to be pertinent [Ramshaw and Marcus 1999]. Considering the conceptual reflection of Santos [Santos 2021], we use algorithms that represent different mechanisms for searching sequential patterns, such as the algorithm Transformation-Based Learning, (TBL) [Brill 1995], and two classifiers, one based on decision trees and another on random forests, both using boosting, as addressed by Chen and Guestrin [Chen and Guestrin 2016].

Uneson [Uneson 2014] highlights some relevant features of the TBL algorithm, such as i) interpretability of the learned representation, ii) synthesis of the learned representation, iii) representative objective function, iv) resistance to *overtraining*, v) research during training instead of an application, vi) integration of heterogeneous resources and vii) competitive performance.

The TBL is an algorithm focused on pattern analysis that considers the positional aspect as predominant to analyze the attributes of the sentence terms. It also considers

the term's typology, order of occurrence, and place of occurrence. Its execution for composing rules considers a range, a domain, as predefined (templates). This amplitude will only be determined at the end of training. To this algorithm we present, in a first approximation, the UD relations and the markings in the IOB format corresponding to the NP_L. In a further moment, we present the UD PoS Tags with the same IOB tags. That is, we performed two experiments separately, or better, without the influence of one on the other. In the TBL methodology, the idea of learning is to start with some simple solution (initial rules) that identifies the phrases and apply transformations (new rules) that can improve the previous performance of tagging the phrases.

In the case of classifiers based on decision trees and random forests (XGBClassifier and XGBRFClassifier respectively), we emphasize that the presentation of the allowed data typology of these algorithms can be done simultaneously, that is, we can insert the fields as predictive attributes, *deprel, upos and deps*, the latter being preprocessed to extract the morphosyntactic attribute of the hierarchically superior word/*token* of the predicate/argument relation; and also the *IOB-format* tags. This way, they jointly treat independent and dependent attributes. These algorithms are characterized by identifying patterns by processing data in parallel and serial mode, that is, they search for residual patterns in features that are initially excluded by the classifier. These algorithms train under a series of trees or weak forests to obtain an increasingly robust model, in addition to having a shrinking technique that reduces the contribution of each tree in the final model, which decreases the influence of each tree, making the slower-fitting process, but resulting in robust models.

5. Results

The TBL-Brill algorithm obtained an accuracy of 87.42% through the UD relations, -4.02 p.p. below that obtained with the use of UD PoS Tags that reached 91.44% of accuracy. Thus, the TBL-Brill algorithm managed to filter representative rules with regular patterns by using only two templates, which composed six representative rules of NP_L, as summarized in Table 3.

Table 3. Final results (%) for TBL.

	Templates	Rules	Accuracy
UD dependency relations	8	57	87,42%
UD PoS Tag	2	6	91,44%

5.1. TBL rules

For the UD relations markup, we applied eight templates to compose 57 rules identified as NP_L modelers. Some of these rules are represented in Table 4. The existence of a wide range of UD tags allows for a wide range of events identified as NP_L.

As for the UD PoS tags, we noticed a high representation of the NP_L considering only two *templates* that form only the six rules represented in Table 5. Considering these rules, TBL identified 553 sequences, which represents a high performance for the different syntactic lexical signatures of NP_L.

Table 4. Some rules formed by UD relations and IOB labels.

Template	Starting tag	Final tag	Rule
017	'B'	'I'	(*token*[-1], 'det'), (*token*[1], 'flat:name')
009	'B'	'O'	(*token*[-1], 'nsubj')
017	'B'	'I'	(*token*[-1], 'case'), (*token*[1], 'flat:name')
000	'I'	'B'	(*tag*[-1], 'O')
010	'B'	'O'	(*token* [1]), 'acl:relcl')
017	'I'	'B'	(*token*[-1], 'root'), (*token*[1]),'obj')

Table 5. Main rules composed by UD PoS Tag and IOB labels.

Template	Starting tag	Final tag	Rule
017	'O'	'B'	(*token*[-1], 'ADP'), (*token*[1], 'NOUN')
017	'O'	'I'	(*token*[-1], 'DET'), (*token*[1], 'NOUN')
017	'O'	'B'	(*token*[-1], 'ADP'), (*token*[1], 'PROPN')
017	'O'	'I'	(*token*[-1], 'NUM'), (*token*[1], 'NOUN')
001	'O'	'B'	(*tag* [1]), 'I')
017	'O'	'B'	(*token*[-1], 'ADP'), (*token*[1]),'SCONJ')

One may notice that the main rules are identified by a number (either 017 or 001) and that there is a predominance of changes from initial inscription labels ('O') to final inscription labels ('B' or 'I') in patterns composed of one or more elements (*tokens*). Each tuple consists of a token at a specific position and a label corresponding to morphosyntactic markup (e.g.: 'NOUN'). In the algorithm's search for all compositions of NP_L, it found 92 possible templates, of which 18 templates were considered useful. In this specific case, template 017 obtained a score of 91.9%, that is, it was considered one of the most important templates by the model, as it composed 5 rules that represented about 83.3% of the total created rules. Meanwhile, template 001 obtained a score of 0.81%, representing lower importance concerning the other, since it established only one rule that corresponds to 16.7% of the total formed rules.

In Table 5, in its first line, we show, as an example, the sequential use of template 017, considering the tokens immediately before and immediately after the one in analysis. In composing this rule, the algorithm waits for an initial 'ADP' tag and a final 'NOUN' tag that delimits the segment corresponding to an NP_L. Analogous reasoning can be applied to the other rules. In addition, the results of applying the TBL algorithm for classification of NP_L, shown in Table 6, express the metrics of precision, recall, F-measure, and accuracy for the IOB markings conditioned to the POS labels. We see that comparing the TBL performance between the UD dependency ratios and the POS markings, the latter obtained slightly more advantageous results, culminating in an accuracy of 91.4% when considering the POS and 87.4% considering only the UD relations.

Table 6. TBL comparative percentage results for IOB tags.

Tags	Metrics			
	Precision	Recall	F-measure	Accuracy
UD relations				87,42
B	78,66	88,02	83,08	–
I	88,79	76,34	82,09	–
O	88,94	89,47	89,20	–
UD PoS Tag				91,44
B	92,18	93,81	92,99	–
I	90,91	85,89	88,33	–
O	91,05	92,91	91,97	–

5.2. Classifiers based on decision trees and forests with boosting

Remember that we also tested two other classifiers, which are the XGBClassifier and the XGBRFClassifier. The XGBClassifier is a machine learning algorithm applied to structured data. This classifier is an implementation of decision trees with gradient boosting. This implementation is focused on performance gain. XGBoost is the basis of the XGBClassifier classifier and is typically used to train gradient-boosting decision trees. XGBRFClassifier is a version of XGBClassifier trained using random forest.

The XGBClassifier algorithm obtained a precision of 87.19%, that is, a value of only +0.86 p.p. above its peer based on decision forest. Similarly, we found an increment of +0.58 p.p. on revocation, also +0.62 p.p. in the F-measure, and +0.58 p.p. for accuracy. Modest but superior increments.

Following its experiment, we applied another closed cross-validation method. The results showed a mean accuracy of 86.78% for XGBClassifier, a +0.13 p.p increase compared to its previous result; and 86.18% for XGBRFClassifier, a +0.11 p.p. increase when compared to its previous result.

Table 7. Results (%) of classifiers based on decision trees and forests.

Classifiers	Metrics			
	Precision	Recall	F-measure	Accuracy
XGBClassifier	87,19	86,65	86,78	86,65
XGBRFClassifier	86,33	86,07	86,16	86,07

6. Conclusion

Partial parsing proved to be a plausible strategy for the identification of NP_L, in texts written in Portuguese and annotated in the UD formalism, through machine learning tech-

niques and rule abstraction, the latter implemented using IOB tags. Computational learning that resorts to the classification of IOB labels allows the identification of fragments that compose an NP_L and, therefore, its more extensive configurations may have their limits poorly defined or discontinued.

As for the computational treatment with the TBL algorithm, the UD PoS Tags stood out with a percentage of +4.02 p.p. in accuracy (91.44%) when comparing the marks of UD Dependency Relations (87.42%). Considering the other algorithms, those based on decision trees and forests, they obtained an accuracy metric of at least 4.79 p.p. more expressive (86.65% for the XGBClassifier). The cross-validation technique slightly improved the application, achieving +0.13 and +0.11 mean accuracy. The performance of TBL in these tests for the Brazilian Portuguese language does not represent an appropriate result to state that ML applications may have similar performance for other natural languages annotated in the UD formalism since the TBL algorithm, when establishing a sequence for identifying the NP_L, is subordinated to the specific typology of terms in this other language. We also point out that the dependence of many current algorithms on the volume and variety of data for pattern extraction can influence the results.

The research carried out by Oliveira and Freitas [Oliveira and Freitas 2006] to identify NP_L is inserted in another syntactic context, the context of constituents. These researchers brought a new relevant syntagmatic specification to the computational linguistics scenario, the NP_L, and reached an accuracy of 85.9% and an F-measure of 86.2%. Establishing a comparison between this and the research by Oliveira and Freitas would be inappropriate since they treat different data annotated under different grammatical formalisms. However, the proximity between the metrics obtained in these two experiments confirms Rambow's assertion that constituent and dependency grammars bring the same syntactic content from different perspectives [Rambow 2010]. We emphasize that UD morphosyntax, combined with UD dependency relations, are elements that allow the establishment of non-natural syntagmatic segments of the universal dependency grammar.

Finally, we highlight possible future contributions: i) the expansion of the *corpus* with a greater number of annotated sentences to reaffirm or not TBL's performance against such state-of-the-art algorithms; ii) approximations that increase the accuracy and recall achieved so far; iii) the identification of this specific type of phrase in other languages to reaffirm the proposal of the Universal Dependency project, as well as the correlation of the NP_L to other natural languages, and; iv) verify if only with the restricted use of UD dependency relations in another natural language the typological question can be crossed.

Acknowledgement

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 23rd, 1991, within the scope of PPI-SOFTEX, coordinated by Softex and published as Residence in TIC 13, DOU 01245.010222/2022-44.

References

Afonso, S., Bick, E., Haber, R., and Santos, D. (2002). Floresta sintá(c)tica: A treebank for Portuguese. In *Proceedings of the 3rd International Conference on Language Resources and Evaluation (LREC)*, pages 1698–1703, Las Palmas, Spain.

Brill, E. (1995). Transformation-Based Error-Driven Learning and Natural Language Processing: A Case Study in Part-of-Speech Tagging. *Comput. Linguist.*, 21(4):543–565.

Chen, T. and Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD'16, page 785–794, New York, NY, USA. Association for Computing Machinery.

Chomsky, N. (2009). *Syntactic structures*. De Gruyter Mouton.

da Silva, J. R. M. F. (2007). *Shallow processing of Portuguese: From sentence chunking to nominal lemmatization*. PhD thesis, Universidade de Lisboa, Faculdade de Ciências.

Elhadad, M. (1996). Lexical choice for complex noun phrases: Structure, modifiers, and determiners. *Machine Translation*, 11:159–184.

Garrido Alenda, A., Gilabert Zarco, P., Pérez-Ortiz, J. A., Pertusa, A., Ramírez Sánchez, G., Sánchez-Martínez, F., Scalco, M. A., and Forcada, M. L. (2004). Shallow parsing for Portuguese-Spanish machine translation. In *Workshop Notes of TASHA'2003*, pages 21–24, Lisboa, Portugal. Edições Colibri.

Hammerton, J., Osborne, M., Armstrong, S., and Daelemans, W. (2002). Introduction to Special Issue on Machine Learning Approaches to Shallow Parsing. *Journal of Machine Learning Research*, 2:551–558.

Hjelmslev, L. (1975). *Prolegômenos a uma teoria da linguagem*. Perspectiva.

Jurafsky, D. and Martin, J. (2021). *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*, volume 3. Stanford Edu.

Lacroix, O. (2018). Investigating NP-Chunking with Universal Dependencies for English. In *Proceedings of the Second Workshop on Universal Dependencies (UDW 2018)*, pages 85–90, Brussels, Belgium. Association for Computational Linguistics.

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

Oliveira, C. and Freitas, M. C. (2006). Um modelo de sintagma nominal lexical na recuperação de informações. *XI Simpósio Nacional e I Simpósio Internacional de Letras e Linguística (XI SILEL)*, pages 778–786.

Pagani, L. A. (2015). Duas Noções Fundamentais para Gramáticas de Dependência.

Pardo, T., Duran, M., Lopes, L., Felippo, A., Roman, N., and Nunes, M. (2021). Porttinari – a Large Multi-genre Treebank for Brazilian Portuguese. In *Anais do XIII Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana*, pages 1–10, Porto Alegre, RS, Brasil. SBC.

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 Confer-*

ence on Dependency Linguistics (Depling 2017), pages 197–206, Pisa, Italy. Linköping University Electronic Press.

Rambow, O. (2010). The Simple Truth about Dependency and Phrase Structure Representations: An Opinion Piece. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 337–340, Los Angeles, California. Association for Computational Linguistics.

Ramshaw, L. and Marcus, M. (2002). Text Chunking Using Transformation-Based Learning. *Third ACL Workshop on Very Large Corpora. MIT*, pages 157–176.

Ramshaw, L. A. and Marcus, M. P. (1999). Text chunking using transformation-based learning. *Natural language processing using very large corpora*, pages 157–176.

Santos, D. S. M. (2021). Grandes quantidades de informação: um olhar crítico. In *II Congresso Internacional em Humanidades Digitais*, Online. UFRJ.

Sharma, A., Gupta, S., Motlani, R., Bansal, P., Shrivastava, M., Mamidi, R., and Sharma, D. M. (2016). Shallow Parsing Pipeline – Hindi-English Code-Mixed Social Media Text. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1340–1345, San Diego, California. Association for Computational Linguistics.

Silva, M. C. and Koch, I. G. (2012). *Linguística aplicada ao português.* Cortez.

Souza, A. and Ruiz, E. E. S. (2022). Investigating Lexical NP-Chunking with Universal Dependencies for Portuguese. In *Anais do XIX Encontro Nacional de Inteligência Artificial e Computacional*, pages 342–351, Porto Alegre, RS, Brasil. SBC.

Tesnière, L. (2015). *Elements of structural syntax*. John Benjamins Publishing Company.

Tham, M. J. (2020). Bidirectional Gated Recurrent Unit For Shallow Parsing. *Indian Journal of Computer Science and Engineering (IJCSE)*, 11(5):517–521.

Tjong Kim Sang, E. F. and Buchholz, S. (2000). Introduction to the CoNLL-2000 Shared Task Chunking. In *Fourth Conference on Computational Natural Language Learning and the Second Learning Language in Logic Workshop*.

Topsakal, O., Açikgöz, O., Gürkan, A. T., Kanburoglu, A. B., Ertopçu, B., Özenç, B., Çam, I., Avar, B., Ercan, G., and Yildiz, O. T. (2017). Shallow parsing in Turkish. In *2017 International Conference on Computer Science and Engineering (UBMK)*, pages 480–485.

Uneson, M. (2014). When Errors Become the Rule: Twenty Years with Transformation-Based Learning. *ACM Comput. Surv.*, 46(4).

Construções sintáticas do português que desafiam a tarefa de *parsing*: uma análise qualitativa

Magali S. Duran[1], Maria das Graças V. Nunes[1,2], Thiago A. S. Pardo[1,2]

[1]Núcleo Interinstitucional de Linguística Computacional (NILC)
[2]Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo (USP)
magali.duran@uol.com.br,gracan@icmc.usp.br,taspardo@icmc.usp.br

Abstract. *When used to train a parser, an annotated corpus reveals its strengths and weaknesses. Based on a qualitative analysis of the performance of a parser trained on an annotated corpus in the Universal Dependencies scheme, this paper points out some errors motivated by the non-canonical order of constituents in Portuguese: postposed subjects and determiners and anteposed adjectives. By using illustrations of syntactic trees before and after manual correction of these errors, the article aims to highlight the importance of having a reasonable number of sentences with these non-canonical structures in order to increase the probability that the parser learns to analyze them correctly.*

Resumo. *Ao ser usado para treinar um parser, um córpus anotado mostra suas qualidades e suas deficiências. Baseado em uma análise qualitativa do desempenho de um parser treinado em córpus anotado no esquema Universal Dependencies, este artigo discute alguns erros motivados pela ordem não canônica dos constituintes em Português: sujeitos e determinantes pospostos e adjetivos antepostos. Usando ilustrações de árvores sintáticas antes e depois da correção manual desses erros, o artigo tem por objetivo destacar a importância de haver uma quantidade razoável de sentenças com essas estruturas não canônicas a fim de aumentar a probabilidade de que o parser aprenda a analisá-las corretamente.*

1. Introdução

Em tempos em que o Aprendizado de Máquina (AM) é a abordagem dominante nos sistemas de Inteligência Artificial (IA) e de Processamento de Línguas Naturais (PLN), torna-se importante identificar os fenômenos que apresentam dificuldade para os algoritmos de aprendizagem. Em tarefas como *parsing* (ou seja, análise sintática automática), que alcançam precisão expressiva com essa tecnologia, é de se esperar que construções que dependam de algum conhecimento semântico ofereçam maior desafio para os algoritmos, por exemplo, os conhecidos *PP-attachments* e as coordenações em contextos com opções variadas. Entretanto, esses não são os únicos casos. Posições menos frequentes de elementos na sintaxe da língua, como sujeitos pospostos, determinantes pospostos e adjetivos antepostos, também são responsáveis por erros recorrentes.

Este trabalho tem por objetivo discutir problemas de natureza exclusivamente sintática observados durante a análise qualitativa do *parser* UDPipe 2[1] [Straka 2018]

[1] https://ufal.mff.cuni.cz/udpipe/2

treinado sobre o córpus Porttinari-base (que compõe o *treebank* Porttinari [Pardo et al. 2021]), um córpus de 8.418 sentenças (168.080 tokens), extraídas do córpus jornalístico Folha-Kaggle[2], que foram inicialmente anotadas com relações de dependências segundo o modelo *Universal Dependencies* (UD) [de Marneffe et al. 2021] [Nivre at al. 2020] e posteriormente revisadas utilizando-se como parâmetro dois manuais de anotação: o Manual de Anotação de *PoS Tags* [Duran 2021] e o Manual de Anotação de Relações de Dependência [Duran 2022].

A análise qualitativa que revelou os casos aqui relatados teve o objetivo de identificar e relatar erros recorrentes para propor melhorias no córpus de treinamento que pudessem incrementar o aprendizado automático. Mesmo pequenas melhorias são relevantes em larga escala, pois o *parser* resultante deverá ser usado para anotar automaticamente o córpus Folha-Kaggle inteiro (com 3.964.292 sentenças e 84.795,823 tokens), bem como outros córpus, de outros gêneros, visando aumentar os recursos baseados em sintaxe e desenvolver ferramentas e aplicativos para a língua portuguesa do Brasil com vistas a alcançar o estado da arte mundial nessa área. Ao leitor interessado, sugere-se a leitura do relatório contendo a análise qualitativa completa [Duran, Nunes & Pardo, 2023].

Na Seção 2, introduzimos brevemente a UD; na Seção 3, apresentamos a metodologia utilizada; na Seção 4, analisamos os resultados; na Seção 5, concluímos com algumas recomendações e trabalhos futuros.

2. *Universal Dependencies* (UD)

É importante introduzir o esquema de anotação UD, pois é o esquema que utilizaremos para ilustrar graficamente nossas anotações. Em sua versão atual, a UD possui dezessete etiquetas morfossintáticas[3] ou *Part-of-Speech* (PoS*) tags*. Também possui 37 etiquetas de relações de dependência[4] – *deprel* (de ***dep****endency* ***rel****ation*). Uma *deprel* é uma relação que liga dois a dois os elementos (*tokens*) de uma sentença tal que:

- um deles é chamado de *head* (cabeça, governante ou núcleo da relação) e o outro é chamado de **dependente**;
- um *token* pode ser *head* de mais de uma relação;
- um *token* pode ser dependente de uma relação e *head* de outra;
- um *token* **não** pode ser dependente de mais de uma relação;
- o nome da relação está sempre associado à função que o dependente desempenha em relação ao *head*;
- graficamente, uma seta parte sempre do *head* em direção ao dependente da relação;
- um *head* é sempre uma palavra de conteúdo (verbo, substantivo, adjetivo, pronome, numeral e advérbio) – exceções são símbolos que podem ser expressos por palavras, como R$ (reais), % (por cento) e § (parágrafo);

[2] https://www.kaggle.com/datasets/marlesson/news-of-the-site-folhauol
[3] https://universaldependencies.org/u/pos/index.html
[4] https://universaldependencies.org/u/dep/index.html

- palavras funcionais (determinantes, preposições, conjunções) e sinais de pontuação, por sua vez, deverão ser apenas dependentes e nunca *head* de relações;
- algumas relações são permitidas apenas em um sentido, enquanto outras relações são admitidas nos dois sentidos;
- quando o dependente tiver forma oracional, o elemento apontado pela seta será o núcleo do predicado da oração dependente;
- toda sentença tem uma raiz (normalmente o predicado da oração principal), marcada como dependente da *deprel* **root** – a *deprel* **root** é a única que não tem *head*, apenas dependente (é a partir do **root** que é construída a árvore sintática, da qual cada nova *deprel* constitui um galho).

A atribuição de relações de dependência deve observar o princípio da projetividade, ou seja, os arcos das relações *não devem* se cruzar. As diretrizes da UD estão disponíveis em sua *homepage*[5], assim como os mais de 200 córpus já anotados. Essas diretrizes já foram instanciadas para a língua portuguesa [Duran 2021, 2022] e há alguns córpus de português brasileiro revisados manualmente já disponíveis no site da UD, como o Bosque-UD [Rademaker et al. 2017] e o PetroGold [Souza et al. 2021].

A Figura 1 ilustra uma sentença anotada com relações de dependência UD.

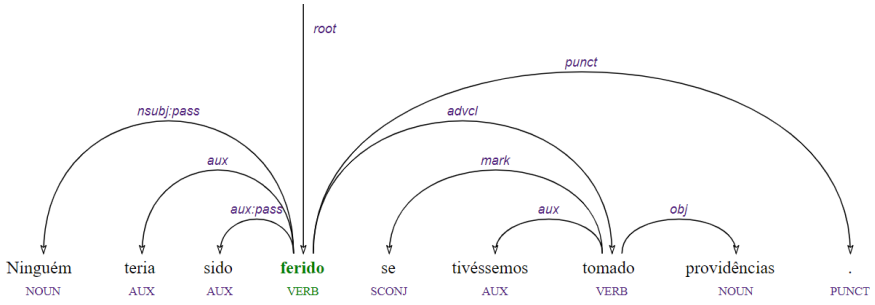


Figura 1 - Exemplo de árvore de dependências anotada com etiquetas da UD

3. Metodologia

Os dados exibidos neste trabalho são parte dos resultados de uma avaliação intrínseca na qual analisou-se o quanto o resultado do *parser* está em conformidade com as diretrizes definidas nos manuais de anotação. O conjunto avaliado consiste de uma amostra aleatória de 600 sentenças, com 12.076 tokens e tamanho médio de 20 tokens (entre 5 e 52 tokens), as quais fazem parte do Porttinari-check, um subcórpus do *treebank* Porttinari, cujos tamanhos de sentenças, *PoS tags* e *deprel* possuem distribuição similar aos do córpus Porttinari-base, sendo, por esse motivo, uma boa amostra desse córpus manualmente revisado. O Porttinari-check foi anotado automaticamente pelo *parser* UD-Pipe treinado sobre o córpus Porttinari-base. O tamanho da amostra (equivalente a 7,13% do córpus Portinari-base) não prejudica a análise qualitativa; pelo contrário,

[5] https://universaldependencies.org/

evidencia que solucionar os problemas recorrentes nesta amostra proporcionará melhorias no treinamento do córpus completo.

Primeiramente, procedeu-se à revisão manual da anotação dessas sentenças, usando a interface do Arborator-NILC[6] [Miranda e Pardo 2022]. Em seguida, foi feita a análise dos erros, buscando agrupá-los de forma lógica para construir uma tipologia de erros. Computou-se um erro para cada alteração feita pelo anotador humano, seja uma mudança só de nome da *deprel*, seja uma mudança só de *head* da *deprel*, ou seja uma mudança de nome e de *head* da *deprel*.

Na avaliação, foram computados 722 erros em 298 sentenças. Das 600 sentenças da amostra analisada, 302 (50%) estavam totalmente corretas. Com base na análise realizada, os erros foram divididos em três categorias: 1) erros provenientes de problemas de pré-processamento (tokenização, lematização, segmentação de sentenças e anotação morfossintatica); 2) erros de escolha de *head* de *deprel* (relacionados à semântica[7]); e 3) erros de natureza puramente sintática, ou seja, aqueles que não são motivados por erros de pré-processamento ou que não são dependentes de informações semânticas.

Na categoria de erros puramente sintáticos, foram identificados e agrupados: erro de identificação de sujeito; erro de **amod** (adjetivo) anteposto; erro de **det** (determinante) posposto; erro de reconhecimento de **fixed** (expressões fixas); erro de reconhecimento de **flat:name** (nomes próprios compostos) e erro de **root**. Os erros de **fixed**, **flat:name** e **root** são exclusivos do esquema de anotação UD. Por esse motivo, decidimos tirá-los do foco das reflexões apresentadas neste artigo. Na próxima seção analisamos os erros que são foco deste artigo, ou seja, aqueles que dizem respeito a ordens de elementos não canônicas na língua portuguesa.

4. Análise dos erros

Discutiremos, a seguir, os erros de sujeito e determinante pospostos e adjetivo anteposto, a fim de levantar subsídios que possam ser úteis àqueles que se dedicam a criar ferramentas automáticas para analisar a língua portuguesa.

4.1 Erro na identificação de nsubj, nsubj:pass e csubj

De modo geral, observa-se que sentenças com seus constituintes na ordem canônica do português - SVO (sujeito, verbo, objeto) - são anotadas corretamente, ao passo que sentenças com elipses de constituintes e/ou inversão da ordem canônica apresentam mais erros.

O problema na identificação do sujeito ocorreu 38 vezes na amostra analisada e compreende três *deprels*: **nsubj** (sujeito da voz ativa), **nsubj:pass** (sujeito da voz

[6] https://arborator.icmc.usp.br/

[7] Só a interpretação semântica pode determinar qual é o *head* de um *token* e, dependendo da *POS tag* do *head*, qual é a relação de dependência. Casos ambíguos incluem **acl** e **advcl** (oração adjetiva e oração adverbial), **advmod** (advérbios simples), bem como **nmod** e **obl** (modificador nominal e oblíquo, respectivamente).

passiva) e **csubj** (sujeito oracional). Para fins de interpretação da relevância desse número, é válido ressaltar que, nas 600 sentenças, houve 599 *deprels* de sujeito atribuídas, onze das quais foram atribuídas incorretamente (não eram sujeitos), ou seja, 588 sujeitos foram atribuídos corretamente. Além disso, houve 27 casos de sujeitos não identificados pelo *parser*. Tanto a identificação incorreta quanto a não identificação do sujeito estão relacionadas, principalmente, à ocorrência do sujeito posposto ao verbo, ou seja, em construções do tipo VS (Verbo Sujeito) ou OVS (Objeto Verbo Sujeito), que, embora sejam ordens não canônicas no português, apresentam frequência significativa no córpus.

Um dos casos comuns de sujeito posposto na voz ativa ocorre com verbos intransitivos, como “estrear”, “faltar”, “sobrar”, “restar”, “ocorrer”, “bastar” e “existir”, e transitivos indiretos, como “caber” e “constar”. Verbos que admitem o papel semântico de tema na posição de sujeito também costumam apresentar sujeitos pospostos, principalmente na negativa, como “não interessa X” e “não importa X”, em que X tem papel de sujeito.

A Figura 2 traz o exemplo de um verbo intransitivo, “sobrar”, e a Figura 3 traz o exemplo de um verbo transitivo indireto, “constar”, em locução verbal com o modal “dever”, que é o *head* no **nsubj**, por ser o verbo que concorda com o sujeito.

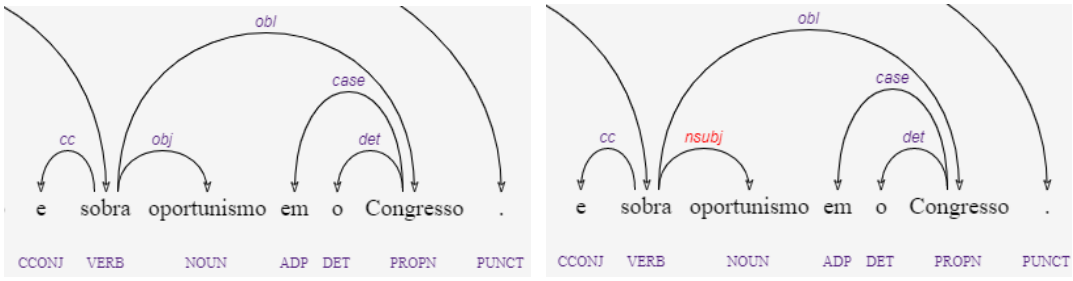


Figura 2 - Sentença com verbo intransitivo anotada pelo *parser* (esq) e corrigida (dir)

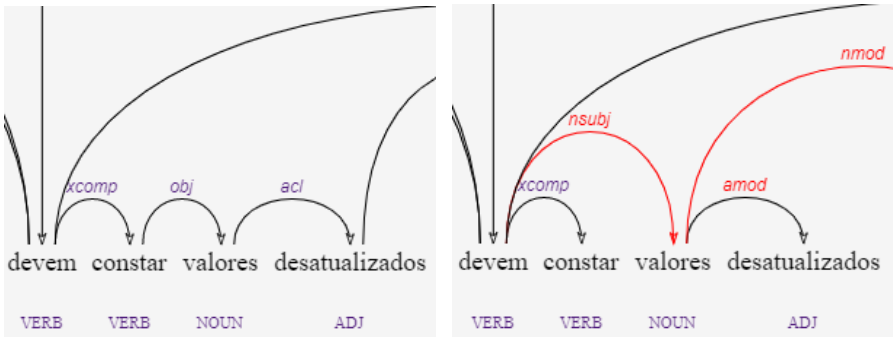


Figura 3 - Sentença com verbo trans. indireto anotada pelo *parser* (esq) e corrigida (dir)

Outro caso muito frequente de sujeito posposto é com verbos na voz passiva analítica, principalmente quando a oração é reduzida e o verbo auxiliar de passiva está elíptico, como em “assim que [for] recebida a denúncia” (Figura 4)[8].

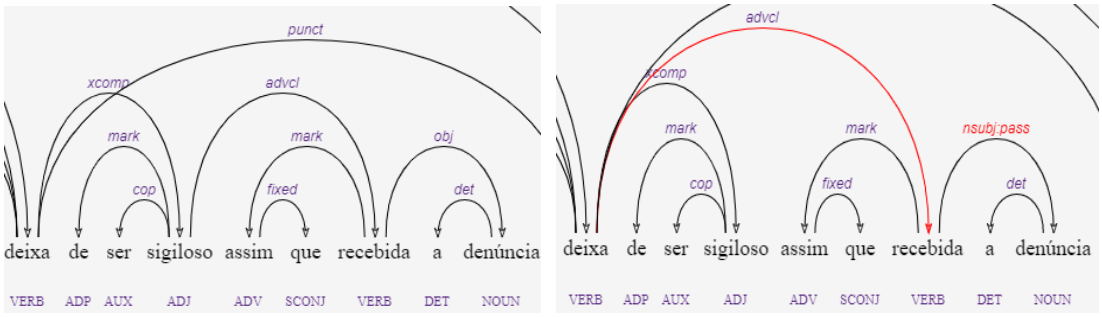


Figura 4 - Sentença na voz passiva analítica anotada pelo *parser* (esq) e corrigida (dir)

Também é frequente o sujeito posposto com verbos na voz passiva sintética, como mostrado na Figura 5.

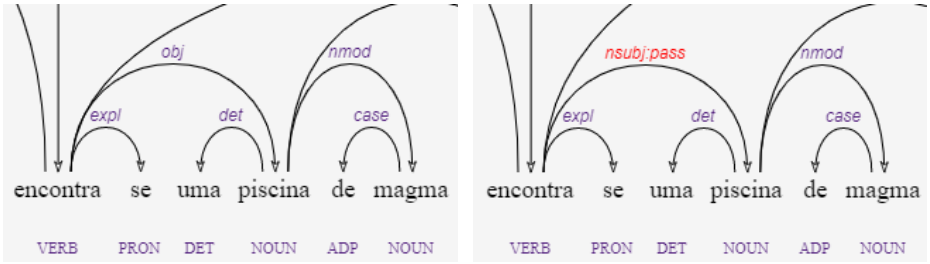


Figura 5 -Sentença na voz passiva sintética anotada pelo *parser* (esq) e corrigida (dir)

Além desses casos de predicados verbais, há também casos de predicados nominais com sujeito posposto. Nesses casos, o verbo de cópula inicia a oração, sendo seguido pelo predicativo e pelo sujeito, na maioria das vezes um sujeito oracional (**csubj**). No geral, o *parser* aprendeu muito bem a identificar esses sujeitos, por isso são raros erros como o ilustrado na Figura 6 e corrigido na Figura 7.

[8] Curiosamente, todos os sujeitos pospostos recorrentes têm papel semântico de tema, papel que também é comumente atribuído ao objeto dos verbos transitivos diretos.

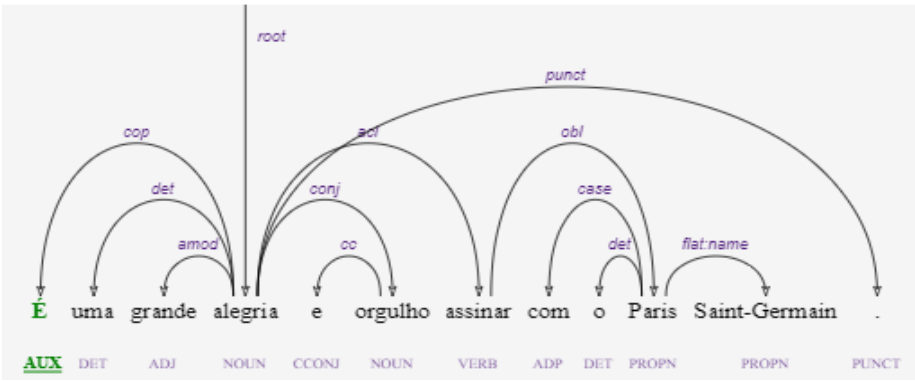


Figura 6 - Sentença com sujeito posposto anotado pelo *parser*

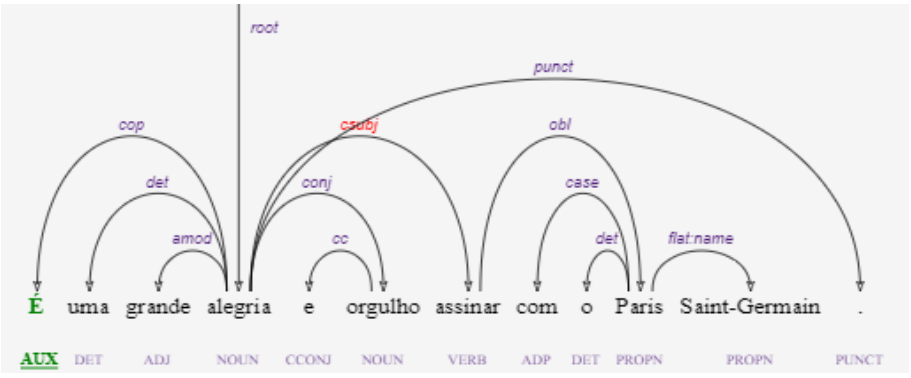


Figura 7 - Sentença com sujeito posposto corrigido manualmente

No português, o *parser* tem que lidar com a possibilidade de o sujeito estar elíptico, ou seja, casos em que a *deprel* **nsubj** não será usada. Quando há um candidato a sujeito à esquerda do verbo, o *parser* quase sempre acerta a atribuição. Entretanto, quando não há um candidato à esquerda e há um candidato à direita, o *parser* se confunde, pois sintagmas nominais à direita podem ser objeto ou sujeito. A Figura 8[9] ilustra uma confusão desse tipo em que o *parser* atribuiu **nsubj** a um token que é **obj**.

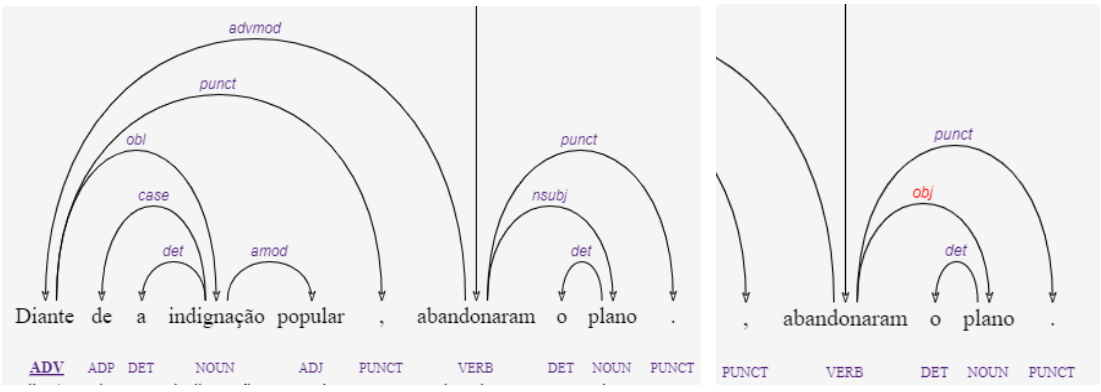


Figura 8 - Sentença com sujeito elíptico, anotada pelo *parser* (esq) e corrigida (dir)

[9] Não se anotou "diante de" como expressão fixa com valor de preposição, mas sim como advérbio predicativo, porque é possível inserir palavras entre as partes (ex: "diante não apenas de x, mas também de y").

4.2 Erro na anotação de modificador amod anteposto

Em português, os adjetivos que modificam substantivos (*deprel* **amod**) podem ocorrer tanto à esquerda quanto à direita (antepostos ou pospostos). Contudo, a ocorrência posposta é extremamente mais frequente. Casos de dois adjetivos, um anteposto e um posposto ao substantivo modificado, costumam confundir o *parser*, mesmo com a anotação correta de *PoS tags*, como pode ser visto na Figura 9, que apresenta um substantivo com terminação típica de adjetivo: “contencioso”.

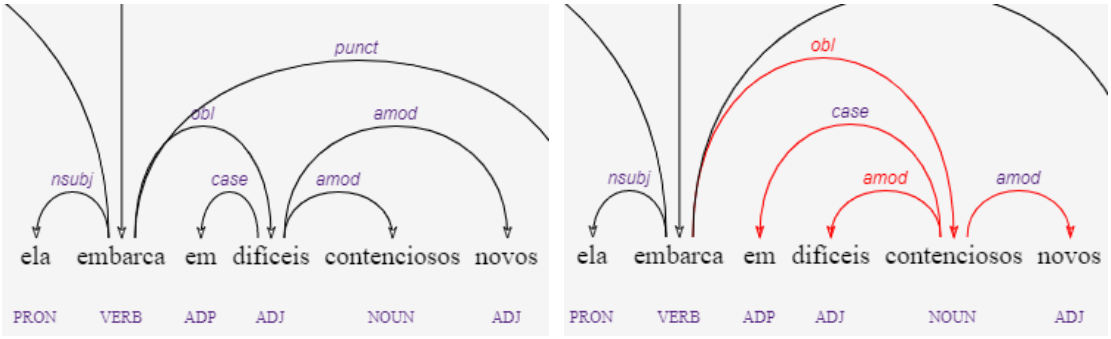


Figura 9 - Sentença com adjetivo anteposto anotada pelo *parser* (esq) e corrigida (dir)

O problema ocorreu pouco na amostra avaliada: seis vezes, incluindo o caso já mencionado de “difíceis contenciosos novos”. São exemplos de **amod** anteposto não identificado pelo *parser* na amostra: “**demasiados** entraves”, “**exímio** pianista”, “**enorme** e **exagerada** reação negativa”, “**impressionante** simulação de violência” e “**legítimo** interesse”. A possibilidade de um ADJ ocorrer anteposto é uma característica lexical e, por isso, enriquecer o córpus com exemplos de ADJ que admitem anteposição pode melhorar o aprendizado automático.

4.3 Erro na anotação de modificador det posposto

Esse tipo de erro é raro (ocorreu uma única vez na amostra), pois em português os determinantes (**det**) ocorrem majoritariamente antes do substantivo. Entretanto, seria interessante ter mais dados de **det** posposto (*deprel* **det** com sentido da esquerda para a direita) para que o *parser* aprenda a anotá-lo, como o caso ilustrado na Figura 10, que apresenta o pronome intensificador "mesmo"[10]. Embora pouco frequente no córpus, é muito normal um **det** posposto, como em “Espelho *meu*”, “proposta *esta* que…”, “ele *próprio*”, “eu *mesmo*” e “eles *todos*”.

[10] Os pronomes que modificam substantivos são anotados como determinantes na UD. A palavra "mesmo", em outros contextos, poderia ser pronome substantivo ou advérbio, situações em que receberia outra anotação.

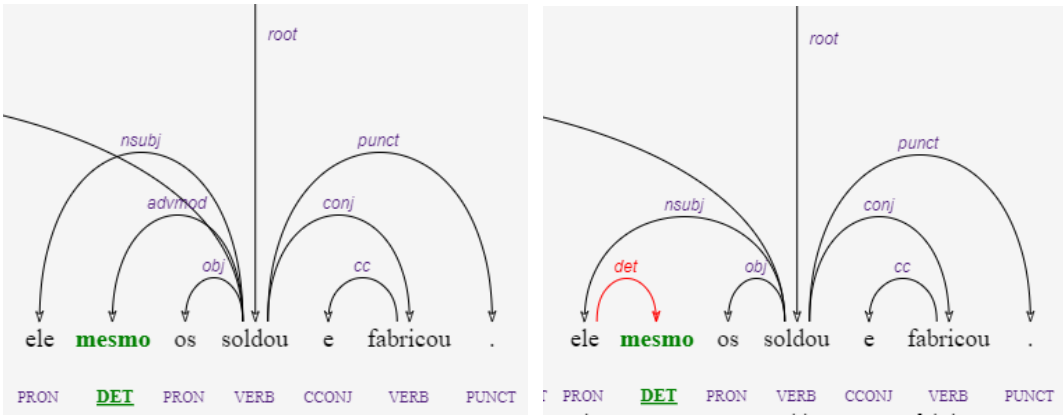


Figura 10. Sentença com det. posposto anotada pelo *parser* (esq) e corrigida (dir)

5. Conclusões

Sob o ponto de vista qualitativo, o *parser* avaliado apresentou excelente desempenho, o que deverá ser ratificado por resultados quantitativos assim que sua versão final for divulgada. Acredita-se que os problemas de natureza exclusivamente sintática possam ser minimizados se o córpus de treinamento vier a ser aumentado usando técnicas recentes de aumento de dados (*data augmentation)* para diminuir a esparsidade de alguns fenômenos [Shorten & Khoshgoftaar, 2019], como sujeitos pospostos (relações **nsubj** da esquerda para a direita), adjetivos antepostos (relações **amod** da direita para a esquerda) e determinantes pospostos (relações **det** da esquerda para a direita). Outro caminho interessante pode ser a proposta de sistemas simbólicos de pós-edição de árvores sintáticas produzidas automaticamente, visando a sua correção.

O estudo apresentado neste artigo deve subsidiar novas iniciativas sintáticas para o processamento computacional do português, mas não apenas isso. Acredita-se que possa também ser a base para outros estudos linguísticos que visem explorar questões de frequência dos fenômenos discutidos ou mesmo explorar tais fenômenos no contexto de ensino da língua portuguesa.

Agradecimentos

Este trabalho foi realizado no âmbito do Centro de Inteligência Artificial da Universidade de São Paulo (C4AI - http://c4ai.inova.usp.br/), com o apoio da Fundação de Amparo à Pesquisa do Estado de São Paulo (processo FAPESP #2019/07665-4) e da IBM. Este projeto também foi apoiado pelo Ministério da Ciência, Tecnologia e Inovações, com recursos da Lei N. 8.248, de 23 de outubro de 1991, no âmbito do PPI-Softex, coordenado pela Softex e publicado como Residência em TIC 13, DOU 01245.010222/2022-44.

Referências bibliográficas

1. 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, Setembro, 55p.
2. Duran, M.S. (2022) “Manual de Anotação de Relações de Dependência –Versão Revisada e Estendida”. Relatório Técnico do ICMC 440. Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo. São Carlos-SP, Outubro, 166p.
3. Duran, M.S.; Nunes, M.G.V.; Pardo, T.A.S. (2023). 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.
4. de Marneffe, M.; Manning, C.; Nivre, J.; Zeman, D. (2021) “Universal Dependencies”, In: Computational Linguistics 47 (2). MIT PRESS, p. 255-308.
5. Miranda, L.G.M.; Pardo, T.A.S. (2022) “An Improved and Extended Annotation Tool for Universal Dependencies-based Treebank Construction”, In: Proceedings of the PROPOR Demonstrations Workshop, p.1-3.
6. Nivre, J.; de Marneffe, M.; Ginter, F.; Hajič, J.; Manning, C.; Pyysalo, S.; Schuster, S.; Tyers, F.; Zeman, D. (2020) “Universal Dependencies v2: An Evergrowing Multilingual Treebank Collection”. In: Proceedings of the 12nd International Conference on Language Resources and Evaluation (LREC 2020), p. 4034-4043.
7. 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: Proceedings of the XIV Symposium in Information and Human Language (STIL 2021), p. 1-10.
8. Rademaker, A.; Chalub, F.; Real, Livy; Freitas, C.; Bick, E.; Paiva, V. (2017) “Universal Dependencies for Portuguese”. In: Proceedings of the Fourth International Conference on Dependency Linguistics. Linköping University Electronic Press, p. 197-206.
9. Shorten, C., & Khoshgoftaar, T. M. (2019). “A survey on Image Data Augmentation for Deep Learning”. In: Journal of Big Data, 6(1), 60.
10. 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. Brussels, Belgium: Association for Computational Linguistics, p. 197-207.

Annotation of fixed Multiword Expressions (MWEs) in a Portuguese Universal Dependencies (UD) treebank: Gathering candidates from three different sources

Elvis de Souza[1], **Cláudia Freitas**[2]

[1]Department of Letters – PUC-Rio
Applied Computational Intelligence Laboratory – PUC-Rio

[2]Department of Letters – PUC-Rio

elvis.desouza99@gmail.com, claudiafreitas@puc-rio.br

Abstract. *Delimiting and correctly annotating multiword expressions (MWEs) is an important task in constructing a gold standard treebank. In this paper, we applied three methods to the PetroGold corpus to identify MWE candidates. The methods include (1) leveraging expressions previously identified by the PALAVRAS annotator, (2) statistical analysis of collocations in Petrolês, a larger non-annotated corpus, and (3) a curated list of co-occurring words from the POeTiSA project. Through extensive filtering and alignment with Universal Dependencies (UD) guidelines, we revised the annotations of 2,467 MWEs in the PetroGold corpus, we tested a new annotation for the part-of-speech (POS) of the words that are part of MWEs and we provide two computationally readable resources to assist other annotators.*

1. Introduction

Multiword Expressions (MWEs) are constructions that can take many forms in a language, such as compound nouns (e.g., "guarda-chuva" [umbrella] and "óleo diesel" [diesel oil]), institutionalized phrases (e.g., "comes e bebes" [food and drinks]), or functional phrases (e.g., "apesar de" [despite], "de acordo com" [according to]). [Ramisch 2012] shows that there is no single definition for MWEs in the literature, and they lie in the gray area between lexicon and syntax, presenting a relevant problem for NLP as they are difficult to handle while being very common in both everyday communication and specialized forms of communication.

Although there is no consensus on a definition, there are some common characteristics of these expressions according to [Ramisch 2012]: (1) they are arbitrary since perfectly grammatical expressions may not be accepted in certain contexts; (2) they are institutionalized, meaning they are part of everyday communication and are accepted and understood by speakers as a conventional way of expressing something; (3) they have limited semantic variation, as they do not undergo the process of semantic compositionality like other language constructions. Therefore, certain parts of an MWE cannot be replaced by any other words or constructions, as the expression is not the result of word composition (nor can the MWE be translated word by word); (4) they have limited syntactic variation, as conventional grammatical rules may not apply to these expressions, making it difficult to determine whether they belong to a speaker's lexicon or grammar (and often they are also extragrammatical, meaning they are unpredictable and difficult

to understand for a language learner who has only learned the grammar of the language), and (5) they are heterogeneous, covering a vast number of language phenomena, each with specific linguistic characteristics, which means that NLP applications should not use a unified methodology to process them.

Delimiting and correctly annotating multiword expressions is an important task in constructing a gold standard treebank. From a machine learning perspective, it is important to ensure consistent annotation of MWEs to avoid providing ambiguous clues about which words, when combined, should be treated as a unit in certain contexts. Without indicating which MWEs will be annotated as locutions (phrasal expressions) in a *corpus*, the morphosyntactic annotation of these phenomena can become inconsistent, with variations in different occurrences, or at worst, it may be impossible to perform any morphosyntactic annotation that makes sense for certain expressions without considering them as multiword units. For example, the expression "isto é" (that is), when used as a conjunctive locution, cannot be annotated as composed of a subject (the pronoun "isto") and a linking verb without jeopardizing the syntactic annotation of the rest of the sentence.

PetroGold v3 is the third version of PetroGold [de Souza 2023], a gold standard treebank for the oil & gas domain in Brazilian Portuguese. The PetroGold was also published in version 2.11 of the Universal Dependencies project, an initiative to provide treebanks for various languages using the same annotation scheme. As a result of this latest PetroGold version, several modifications were made to the annotation of multiword expressions to align with both the guidelines of the UD project and to increase the consistency in the annotation of such expressions.

The approach taken to identify candidates for multiword expressions in Portuguese involved three different sources:

1. **PetroGold:** Expressions previously identified by the PALAVRAS annotator [Bick 2014], which are present in the Bosque-UD *corpus* [Rademaker et al. 2017], and were annotated in previous versions of PetroGold.
2. **Petrolês:** Collocations of the form [PREP (DET)? N PREP], such as "de acordo com" (according to), identified through statistical methods inspired by [Oliveira et al. 2004], reproduced in Petrolês, a much larger non-annotated *corpus* from the oil & gas domain [Cordeiro 2020].
3. **POeTiSA:** A list of words that co-occur without ambiguity, meaning they are always annotated in the same way when they appear together, compiled within the POeTiSA project [Lopes et al. 2021].

These three lists were filtered to adapt them to the very restrictive definition of multiword expressions in the UD project. We also try an alternative annotation to the part-of-speech (POS) tag of the words that compose a MWE, giving each word the POS of the expression as a whole (an annotation which is not recommended by UD), in order to assess the results of a trained model when this alternative annotation takes place.

In the end, the annotations of 2,467 MWEs in the PetroGold *corpus* were revised. We provide, along with this paper, two computationally readable resources that can be useful in other annotation projects: one that represents the union of all the multiword expressions identified using the three methods (amounting to 150 MWEs), and a subset

of this list that includes only the MWEs found in the PetroGold *corpus* (totaling 112 MWEs), along with their corresponding annotations[1]. The UD guidelines to annotate MWEs, the resources we used to find candidates and the results will be presented next.

2. Universal Dependencies take on MWEs

The UD project provides three classes for annotating multiword expressions: (1) "fixed" for fixed expressions that correspond to grammaticalized expressions, which behave as functional or short adverbial words; (2) "flat" for "semi-fixed" expressions, although there is no definition for them except for a list of examples (such as personal names, dates, compound numbers, and foreign phrases); and (3) "compound" for expressions that, unlike the others, have one word functioning as a syntactic head, as in the example "apple pie" (or in our case, "óleo diesel" [diesel oil]).

UD takes an economical approach to what it considers multiword expressions. For example, a very common structure in English is that of two nominals, such as "phone book" and "Natural Resources Conservation Service," or in the PetroGold context, "planta piloto" (pilot plant) and "fase rifte" (rift phase). The guidelines indicate that when there are clear criteria in the language documentation to distinguish compound expressions, the expression can be annotated as an MWE of type "compound," where all words in the expression are annotated as dependents of the main word with the *compound* relation (Figure 1). However, when the criteria are not well established, treebanks should let go of this annotation, tagging them as regular nominal modifiers (nmod).

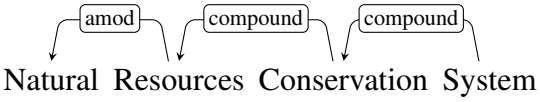


Figure 1. Possible annotation for the structure of "two nominals"

In versions 1 and 2 of PetroGold, we attempted to annotate expressions such as "óleo diesel" (diesel oil) and "meio ambiente" (environment) as *compound* since they would be useful in the oil and gas domain. However, due to the lack of a more comprehensive study on the subject and time limitations of the project, we decided to forego the *compound* label in this third version of the *corpus*, following the guidelines of UD. Instead, we opted for regular nominal modifiers (nmod) as a transparent annotation.

Another type of construction that can be considered an MWE in certain contexts is the one that contains light verbs (or support verbs), as in expressions like "dar um grito" (to scream), "ter em mente" (to have in mind), "tirar um cochilo" (to take a nap), "tomar uma decisão" (to make a decision), or "fazer vista grossa" (to turn a blind eye). In these examples, the verbs *dar/ter/tirar/tomar/fazer* (among others) are verbs "(...) with a greatly depleted meaning that, together with their complement (direct object), form a global meaning, usually corresponding to that of another verb in the language" [Neves 2000, our translation]. Not all expressions with support verbs can be replaced by other verbs in the language, as noted by [Bagno 2012], which justifies the use of the adverb "usually" by Neves. While "dar um grito" (to scream) and "tomar a decisão" (to make the decision) correspond to "gritar" (to shout) and "decidir" (to decide), respectively, "fazer questão"

[1] Available at: *https://github.com/alvelvis/mwe-petrogold-udfest*.

(to insist on) and "soltar balão" (to release a balloon) do not correspond to any verb (and are not compositional either, so these expressions would be filling a gap in the Portuguese lexicon through a phrase).

In any case, UD has defined that such expressions with support verbs should have a transparent annotation, with the noun as the object of the verb[2]. Thus, in sentence 1, "parte" (part) is the direct object of the verb "fazer" (to make), and "Projeto" (Project) is a prepositional object of the same verb, despite "fazer parte" (to be part of) being considered a multiword expression in the Portuguese language.

1. Este levantamento foi realizado em o ano de 1978, **fazendo parte** de o Projeto Aerogeofisico São Paulo – Rio de Janeiro de a CPRM.[3]

3. Methodology

3.1. Obtaining multiword expression candidates

Given that the only annotated MWEs in Portuguese UD are the *fixed* expressions (aside from compound proper names and numbers, respectively *flat:name* and *flat*), we used three methods to obtain candidates for them. The first source of fixed multiword expressions was obtained from the annotation inheritance in PetroGold. The PetroGold *corpus* was originally annotated (before the human inspection) by a model trained on the Bosque-UD *corpus*, which, in turn, is a manually corrected conversion of the PALAVRAS annotation system. Therefore, the first list of expressions analyzed consisted of expressions annotated as units in PetroGold, inherited from PALAVRAS, and subject to specific revisions in the *corpus*.

The second source of multiword expressions was obtained by applying statistical methods, based on [Oliveira et al. 2004], to Petrolês [Cordeiro 2020], a larger text collection, containing 330 academic documents in the field of oil & gas. [Oliveira et al. 2004] dealt with the notions of collocation and multiword expression in order to investigate the cases in which prepositional phrases are both multiword expressions (linguistic units) and collocations (words that frequently co-occur). The analyzed multiword expressions are of the form [PREP (DET)? N PREP], such as "de acordo com" (according to), where there is no determiner, and "no caso de" (in case of), with the determiner.

The method we used to identify collocates was the Likelihood-ratio, which is one of the methods used by [Oliveira et al. 2004]. It measures the probability that events that occurred together are not due to chance. Thus, two hypotheses are calculated: (i) that the words have the same probability of appearing together or separately, and (ii) that the words are more likely to appear together than separately. The metric tells us how much hypothesis (ii) is more likely than (i), and if that's the case, the word sequence is considered a collocation [Manning and Schutze 1999]. We used the NLTK (Natural Language Toolkit) library for the Python programming language to calculate the collocates present

[2]Although they say that each language should define the criteria for annotating *compound*, the UD guidelines indicate that, for English, the "transparent" annotation is the most suitable for "light" constructions (such as "take a decision") and adjective + noun combinations (*hot-dog*). Source: *https://universaldependencies.org/u/dep/compound.html*. Accessed on Mar. 5, 2023.

[3]Transl. "This survey was conducted in the year 1978, **as part of** the São Paulo – Rio de Janeiro Aerogeophysical Project by CPRM."

in the *corpus* and identified which of them would qualify as prepositional phrases (not just collocations) in a sample of the results (the top 40 entries according to the algorithm's evaluation[4]). The results were then manually filtered to find the MWEs in the list.

The third and final source of multiword expressions is not a MWE list itself, compiled within the scope of the POeTiSA project, as part of a series of linguistic resources to improve the quality of POS annotation in a *corpus*. It is a list of words that, when co-occuring, are always unambiguous and should have the same POS annotation. The list was compiled by a linguist during the *corpus* annotation process[5] and the authors note that the entries in the list are not necessarily multiword expressions, so we conducted an analysis to filter out those cases that, according to our criteria, should be removed from the list.

3.2. An alternative Part-Of-Speech (POS) annotation for MWEs

UD prioritizes the annotation of part-of-speech (POS) tags based on the "base" class of the word, regardless of the context in which it is used. Therefore, according to the project, the POS annotation of the words that make up multiword expressions is not different from the annotation of the words if they were not part of an MWE. For example, in the case of "isto é" (meaning "this is"), the first element is a pronoun, and the second element is an auxiliary verb.

During the development of PetroGold, we added extra information to multiword expressions regarding the POS of the entire expression. This information is encoded in the tokens' miscellaneous attribute (tenth column). In the case of the expression "isto é" (meaning "this is"), the first token, "isto" (meaning "this"), annotated as a pronoun in UD, has the information "MWEPOS=CCONJ" in the miscellaneous column, indicating that it is a conjunctive locution. This "alternative" annotation was made with the intention of observing what would happen in syntax if the analysis of MWEs was not done literally at the POS level. To achieve this, we created a variation of the *corpus* where only the POS annotation of all words composing MWEs was replaced with the information from the MWEPOS field. Using the [Straka et al. 2016] tool, we generated two models: one for the modified version with MWEPOS and another for version 2.11 of the UD project. We compared the results using the evaluation metrics presented in [Zeman et al. 2018].

4. Results

The first method to obtain a list of MWEs was the revision of PetroGold, which had initially been annotated by a model trained on the Bosque-UD. The list of fixed MWEs in PetroGold initially contained 148 items. However, after applying the annotation criteria for expressions, only 101 items remained. The second method used was the statistical approach, which had a precision of 70%, considering that only the first 40 entries from the algorithm's results were analyzed (this number would drop significantly if expressions with lower scores were considered). And the third method, using a list compiled by

[4]The limit of 40 was established because, beyond this number, it became very difficult to find prepositional phrases through manual analysis.

[5]We would like to thank Magali Duran and the POeTiSA team for providing the list and all other resources from the project.

POeTiSA, had initially 110 expressions with unambiguous word co-occurrence. After our analysis, 55 MWEs remained from this list.

The three lists – expressions found in the original annotation of PetroGold, extracted through statistical methods from Petrolês, and compiled in the POeTiSA project – have some entries in common and others that are unique, highlighting the efficiency of the strategy used to obtain each list. The diagram in figure 2 shows the number of MWEs found by each method.

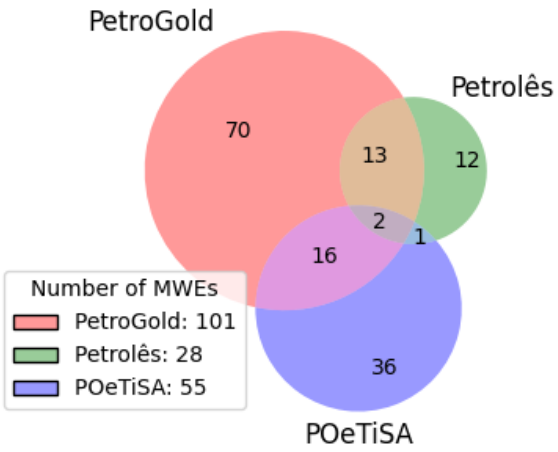


Figure 2. Quantity of multiword expressions obtained by each method

Proportionally, the PetroGold list brings the highest number of MWEs that no other method found – 69.3% of the entries are unique – followed closely by the POeTiSA list (65.45%), and far behind, the Petrolês list (42.85%). Although the statistical method was less efficient in finding MWEs, further investigation could be done in the future to determine if the search performed – for expressions of the type [PREP DET? N PREP] – and the algorithms used are the reason for the low coverage of MWEs returned by the method.

Finally, table 2 compiles the three returned lists without repetitions, containing 150 entries. These 150 entries were applied to the PetroGold *corpus* for the correction of MWEs, and the result is a total of 2,467 occurrences of fixed-type multiword expressions which are now available in the third version of PetroGold. The list containing all 112 MWEs found in PetroGold, along with their corresponding morphosyntactic annotation, can be found in the dedicated repository to this paper[6].

Regarding the part-of-speech (POS) annotation of the words that make up multiword expressions, we also tested an alternative annotation, where each word receives the POS of the entire expression. Thus, if the expression is a conjunction phrase (e.g., "isto é" meaning "that is"), both words were assigned the conjunction POS tag (CCONJ) in this alternative version.

By comparing two models trained on different datasets – the version that follows the UD guidelines (published in version 2.11 of the project) and this variation where the only modification was the POS annotation of MWEs – we observed the results in table 1.

[6]Available at: *https://github.com/alvelvis/mwe-petrogold-udfest*.

In this table, we can see that the POS[7] annotation results worsen, which is not surprising, given the ease with which a model seems to generalize that words should always receive the same POS tag, regardless of the context. However, there is an improvement in syntactic annotation that reaches 0.85 percentage point in the LAS[8] metric, which allows us to reflect on whether the annotation, for certain purposes, may be more appropriate.

UPOS	LAS
98.23% (-0.19 p.p.)	89.48% (0.85 p.p.)

Table 1. Variation in the automatic annotation when using MWEPOS

5. Concluding remarks

In conclusion, our analysis of multiword expressions (MWEs) using different methods and annotation techniques has yielded valuable insights. The results demonstrate that the PetroGold list and the POeTiSA list were the most effective in identifying unique MWEs, with proportions of 69.3% and 65.45% respectively. The Petrolês list, although lagging behind, still managed to uncover a significant number of MWEs at 42.85%. While the statistical method employed in this study proved to be less efficient in identifying MWEs, further investigation could be pursued to determine whether the search performed and the algorithms used are responsible for the lower coverage of MWEs obtained through this approach.

The compilation of the three returned lists, without repetitions, resulted in a total of 150 MWE entries. These entries were then applied to the PetroGold *corpus*, leading to the correction of 2,467 instances of fixed-type multiword expressions in the third version of PetroGold. For a comprehensive list of the 112 MWEs found in corpus, along with their corresponding morphosyntactic annotations, readers can refer to the dedicated repository associated with this article[9].

In terms of part-of-speech (POS) annotation, we also experimented with an alternative approach where each word in an MWE receives the POS tag of the entire expression. The comparison of models trained on both datasets revealed that while the POS annotation results worsened, there was an improvement of 0.85 percentage points in syntactic annotation, as measured by the LAS metric. This finding prompts further reflection on the appropriateness of this alternative annotation for specific purposes.

Overall, this study contributes to our understanding of MWE identification and annotation techniques, opening avenues for future research to enhance the effectiveness and coverage of MWE detection methods while considering the appropriate POS annotation strategies.

Acknowledgments

The authors express their gratitude to CNPq (National Council for Scientific and Technological Development, grant #130495/2021-2), FAPERJ (Carlos Chagas Filho Foundation

[7]POS: Universal Part-Of-Speech, which scores when the tagger finds the correct POS tag.

[8]LAS: Labeled-Attachment Score, which scores when the parser finds the correct attachment for the dependency and the correct label for the relation between governor and dependent.

[9]Available at: *https://github.com/alvelvis/mwe-petrogold-udfest*.

a a medida em que	além de isto	em a verdade	por conseguinte
a a medida que	além de o mais	em direção a	por exemplo
a a toa	além de o que	em face de	por exemplos
a as vezes	além de o quê	em função de	por fim
a cargo de	aos poucos	em geral	por mais que
a despeito de	apesar de	em o caso de	por meio de
a exemplo de	apesar de que	em o entanto	por muito que
a favor de	assim como	em o tocante a	por o menos
a fim de	assim por diante	em razão de	por parte de
a fim de que	assim que	em relação a	por pouco que
a longo de	assim sendo	em relação as	por sua vez
a medida que	até que	em relação á	por vezes
a menos que	bem como	em seguida	por volta de
a não ser que	cada vez mais	em separado	pouco a pouco
a o certo	caso contrário	em termos de	quanto a
a o contrário de	cerca de	em torno de	quanto mais
a o invés de	com base em	em vez de	se bem que
a o largo de	com isso	em vão	sem mais nem menos
a o longo de	com relação a	enquanto que	sem que
a o menos	com vistas a	isto é	sempre que
a o menos que	como relação a	junto a	sendo assim
a o passo que	como também	já que	sendo que
a o ponto de que	de acordo com	logo que	tais como
a o que	de acordos com	mais de o que nunca	tal como
a o todo	de agora em diante	mesmo assim	tanto quanto
a o vivo	de aí	mesmo que	tanto que
a partir de	de forma a	nada que	toda vez que
a ponto de	de forma que	nem a o menos	tudo quanto
a principio	de maneira a	nem mesmo	um a um
a princípio	de maneira que	nem sequer	um por um
a priori	de modo a	não obstante	um pouco
a respeito de	de modo que	não que	uma vez em
a seguir	de o que	não só	uma vez que
a título de	de sorte que	ou seja	visto que
ainda mais que	de tal forma que	para que	volta e meia
ainda que	desde que	para tal	é que
além de	devido a	pelo menos	
além de isso	em a faixa de	por causa de	

Table 2. Final list of multiword expressions obtained by all methods

for Research Support of the State of Rio de Janeiro, grant #E-26/202.433/2022), and ANP (National Agency of Petroleum, Natural Gas, and Biofuels, Brazil), associated with the investment of resources from the Clauses of R,D&I, through a Cooperation Agreement between Petrobras and PUC-Rio, for the provided support, without which this work could not have been accomplished.

References

Bagno, M. (2012). *Gramática pedagógica do português brasileiro*. Parábola Ed.

Bick, E. (2014). PALAVRAS, a constraint grammar-based parsing system for Portuguese. *Working with Portuguese corpora*, pages 279–302.

Cordeiro, F. C. (2020). Petrolês-como construir um corpus especializado em óleo e gás em português. *PUC-Rio, Rio de Janeiro, RJ-Brasil: PUC-Rio*.

de Souza, E. (2023). *Construção e avaliação de um treebank padrão ouro*. Mestrado, PUC-Rio.

Lopes, L., Duran, M. S., and Pardo, T. A. (2021). Universal dependencies-based pos tagging refinement through linguistic resources. In *Brazilian Conference on Intelligent Systems*, pages 601–615. Springer.

Manning, C. and Schutze, H. (1999). *Foundations of statistical natural language processing*. MIT press.

Neves, M. H. d. M. (2000). *Gramática de usos do português*. Unesp.

Oliveira, C., Nogueira, C., and Garrao, M. (2004). Locution or collocation: comparing linguistic and statistical methods for recognising complex prepositions. In *Anais do 2º Workshop em Tecnologia da Informação e da Linguagem Humana*.

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.

Ramisch, C. (2012). A generic framework for multiword expressions treatment: from acquisition to applications. In *Proceedings of the ACL 2012 Student Research Workshop*, Jeju, Republic of Korea. ACL. `https://aclweb.org/anthology/W12-3311`.

Straka, M., Hajic, J., and Straková, J. (2016). UDPipe: trainable pipeline for processing CoNLL-U files performing tokenization, morphological analysis, pos tagging and parsing. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 4290–4297.

Zeman, D., Hajic, J., Popel, M., Potthast, M., Straka, M., Ginter, F., Nivre, J., and Petrov, S. (2018). CoNLL 2018 shared task: Multilingual parsing from raw text to universal dependencies. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual parsing from raw text to universal dependencies*, pages 1–21.

Verifica-UD: a Verifier for Universal Dependencies Annotation for Portuguese

Lucelene Lopes[1], **Magali Sanches Duran**[1], **Thiago Alexandre Salgueiro Pardo**[1]

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

{lucelene,magali.duran}@gmail.com,

taspardo@icmc.usp.br

Abstract. *This paper presents Verifica-UD, a web-based tool to detect problems in Portuguese sentences annotated using Universal Dependencies (UD) standards in the form of a CoNLL-U file. The tool performs three levels of sentence verification: structural (to assess CoNLL-U compliance), morphosyntactic (to assess the part of speech tagging), and syntactic (to assess the parsing information). Verifica-UD also provides detailed help on Portuguese UD annotation directives. The benefits of this tool for reviewing annotated corpora are illustrated with an experiment.*

1. Introduction

The use of Universal Dependencies (UD) [de Marneffe et al. 2021, Nivre et al. 2020] as coding format for annotated corpora brings several advantages in terms of standardization among languages and clarity of concepts for users regardless of language.

CoNLL-U is the standard file format for the annotation of dependency treebanks in UD. Because of that it has been a format of interest for several annotators, and consequently also tool developers. This is the case of web-based visualization tools as Arborator-Grew [Guibon et al. 2020], but also some editors and searchers as UDeasy [Villa 2022] and UDConcord [Miranda and Pardo 2022].

Despite being a compact and relatively simple format, CoNLL-U may be somewhat confusing to be directly handled by human annotators, and often mistakes produced while editing a CoNLL-U file are hard to be spotted using a pure (non-rich) text editor. Some editors provide visualization tools with color tagging of CoNLL-U files, as the VSCode-conllu extension [Grobol 2021], but even such help is not enough when dealing with a large file. Besides the issues of data encoding, UD annotation may be a challenging task, as corpus annotation has shown in recent developments in the area. More than the directives of the original UD project, languages may need additional guidelines to deal with their specific linguistic phenomena.

The UD project offers a validation tool that has general restrictions and language dependent particularities. It is important to highlight that the language particularities included in the UD validation tool tends to include only definitions made based on the available corpora among the UD datasets, and not necessarily following general integrated language directives.

Given this panorama, we propose in this paper a web-based tool to verify CoNLL-U files following the general directives for annotation of Portuguese in UD stated in [Duran 2021, Duran 2022, Lopes et al. 2023a]. This tool, called Verifica-UD, performs the verification of the CoNLL-U file in three levels: structural, Part of Speech (PoS) tagging, and parsing. In fact, the rules applied in our tool reflect the more recent directives considered in the recent discussions of Portuguese annotation in UD and recent resources.

Next section describes the three levels of verification performed by Verifica-UD. The third section briefly presents the online tool usage. The fourth section presents a case study illustrating the potential of the tool to aid in human annotation. Finally, the conclusion section highlights the paper contributions and suggests future work.

2. Verification

This section describes the three levels of verification performed by Verifica-UD:

- Structural verification: analysis of the compliance with the CoNLL-U standards, plus the correct format of the dependency tree (e.g., a single root, connection of all nodes);
- Morphosyntactic verification: analysis of the fields lemma, PoS tag, and morphological features with respect to a lexical resource and general tagging rules;
- Dependency relations verification: analysis of the fields head and dependency relation (DEPREL) with respect to valid connections between head and dependents, as well as the corresponding tagging information of these tokens.

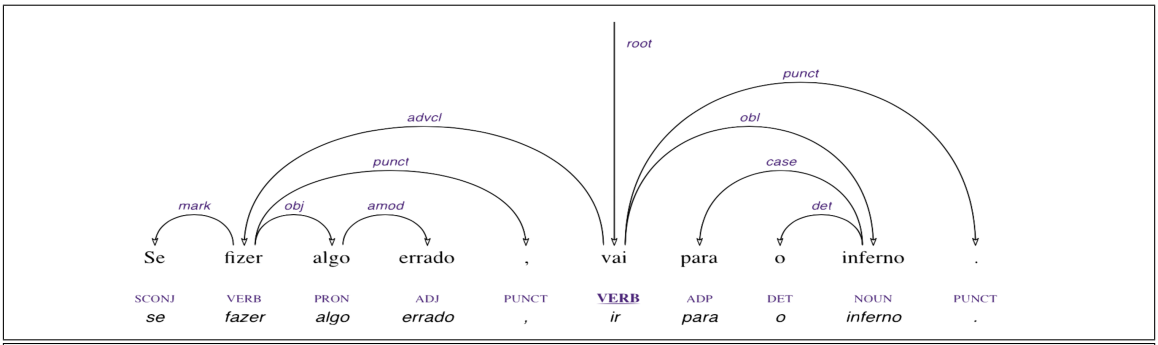
2.1. Structural Verification

The structural verification starts with tests of compliance with the CoNLL-U format, followed by verification of the dependency tree integrity. As such, the structural verification rules are independent of the language, which is not the case of the morphosyntatic and dependency relation rules that are tied to the Portuguese standards. The CoNLL-U format defines that each sentence requires two initial pieces of information:

- the sentence identifier, as `# sent_id = <string>`
- the textual sentence content, as `# text = <string>`

After that, all tokens must be placed in individual lines containing one numbered token per line, each with ten fields separated by tab characters. Contracted words should be split into individual tokens. In the CoNLL-U format, the contracted words must have an individual line to hold it preceding the lines of the split tokens. The ten fields of each token indicate, respectively: ID - the token number identifier; FORM - the token form; LEMMA - the token lemmatized form; POS - PoS tag; XPOS - language specific PoS tag (ignored by Verifica-UD); FEAT - the morphological features; HEAD - the identifier of the token head of the dependency relation; DEPREL - the dependency relation tag; DEPS - enhanced dependency graph (ignored by Verifica-UD); MISC - miscellaneous information. Figure 1 shows an example of a correct sentence[1] with both the CoNLL-U annotation and the corresponding dependency tree.

[1] "Se fizer algo errado, vai para o inferno" (If you do something wrong, you'll go to hell). This sentence may have another interpretation in Portuguese: "Se fizer errado algo" (If you do it wrong); in this case, "errado" (wrong) would be annotated as ADV advmod because it would be modifying the predicate itself and not the object.




```
# sent_id = FOLHA_DOC009367_SENT031
# text = Se fizer algo errado, vai pro inferno.
1   Se  se  SCONJ   _   _   2   mark    _   _
2   fizer   fazer   VERB    _   Mood=Sub|Number=Sing|Person=3|Tense=Fut|VerbForm=Fin    6   advcl   _   _
3   algo    algo    PRON    _   Gender=Masc|Number=Sing|PronType=Ind    2   obj _   _
4   errado  errado  ADJ _   Gender=Masc|Number=Sing|VerbForm=Part   3   amod    _   SpaceAfter=No
5   ,   ,   PUNCT   _   _   2   punct   _   _
6   vai ir  VERB    _   Mood=Ind|Number=Sing|Person=3|Tense=Pres|VerbForm=Fin   0   root    _   _
7-8 pro _   _   _   _   _   _   _   _
7   para    para    ADP _   _   9   case    _   _
8   o   o   DET _   Definite=Def|Gender=Masc|Number=Sing|PronType=Art   9   det _   _
9   inferno inferno NOUN    _   Gender=Masc|Number=Sing 6   obl _   SpaceAfter=No
10  .   .   PUNCT   _   _   6   punct   _   _
```

Figure 1. CoNLL-U example of a sentence and corresponding dependency tree.

The structural verification detects 16 different error kinds. The full list of structural errors detected is available at a previous publication [Lopes et al. 2023b]. Figure 2 illustrates three error examples. The first one is a common error where the token #7 has no morphological features. This token is denoted without the sixth field with the empty indicator ("_") missing. Another error in this figure is in the last token that is incorrectly numbered with #11, instead of the expected #10. The third error in Figure 2 is the dependency tree malformation, since it has two **root** tokens (#2 and #6).

```
# sent_id = FOLHA_DOC009367_SENT031
# text = Se fizer algo errado, vai pro inferno.
1   Se  se  SCONJ   _   _   2   mark    _   _
2   fizer   fazer   VERB    _   Mood=Sub|Number=Sing|Person=3|Tense=Fut|VerbForm=Fin    0   root    _   _
3   algo    algo    PRON    _   Gender=Masc|Number=Sing|PronType=Ind    2   obj _   _
4   errado  errado  ADJ _   Gender=Masc|Number=Sing|VerbForm=Part   3   amod    _   SpaceAfter=No
5   ,   ,   PUNCT   _   _   2   punct   _   _
6   vai ir  VERB    _   Mood=Ind|Number=Sing|Person=3|Tense=Pres|VerbForm=Fin   0   root    _   _
7-8 pro _   _   _   _   _   _   _   _
7   para    para    ADP _   _   9   case    _   _
8   o   o   DET _   Definite=Def|Gender=Masc|Number=Sing|PronType=Art   9   det _   _
9   inferno inferno NOUN    _   Gender=Masc|Number=Sing 6   obl _   SpaceAfter=No
11  .   .   PUNCT   _   _   6   punct   _   _
```

Figure 2. Examples of structural problems, with (a) the sixth field missing error for token #7, (b) last token incorrectly numbered, and (c) malformed dependency tree with two root tokens (#2 and #6).

2.2. Morphosyntactic Verification

Morphosyntactic verification applies to tokens individually. Each token must have consistent information for the fields FORM, LEMMA, POS, and FEAT. In PoS tag annotation for Portuguese, we follow the directives defined at [Duran 2021], that establishes that, from the 17 original PoS tags, only 16 are employed (the PART tag is not used). The PoS

tags ADP, AUX, CCONJ, DET, PRON, and SCONJ are closed classes and, as such, they have, in principle, all possible forms known. The PoS tags ADV and NUM have a defined closed subset, as all primitive adverbs (in Portuguese, adverbs not ending with "*-mente*" - similar to English "-ly") that form a closed subclass, and all numbers written in their extensive form (not with digits). Among the other (open) classes, NOUN, ADJ, VERB, and INTJ, plus the open subset of ADV, have an extensive representation and, together with the closed classes and closed subsets of NUM and ADV, are included in a lexical resource called PortiLexicon-UD [Lopes et al. 2022].

In this way, the morphosyntactic verification is done for each token tagged in the CoNLL-U with tags ADP, ADV (primitive adverbs subset), AUX, CCONJ, DET, NUM (written subset), PRON, and SCONJ, verifying its form against the lexical resource. If present in the lexical resource with the lemma and morphological annotated option, the token is considered correct. Otherwise, general annotation rules are verified, for example, a token tagged as PRON or DET must have a **PronType** feature. If the general annotation rules are valid, a warning is issued stating that the token belongs to a closed class, but it is not present in the lexical resource. However, if the general annotation rules are violated, the token is considered incorrect and an error is issued.

For tokens tagged in the CoNLL-U with tags ADJ, ADV (except primitive ones), INTJ, NOUN, and VERB, the token form is verified against the lexical resource. If present in the lexical resource, the annotation must be one of the options, otherwise a warning is issued. However, if absent in the lexical resource, the general annotation rules are verified and, if the rules are violated, an error is issued, otherwise, a warning is issued.

For tokens tagged in the CoNLL-U with tags PROPN, PUNCT, X, SYM, and NUM (with digits in the FORM field), general annotation rules are verified, and, if the annotation is not one of the expected possibilities, an error is issued.

In Figure 3, three example situations are indicated. An error is indicated for the lemma in token #1 because it mismatches the lexical entry for the SCONJ "*se*", since lemmas of common words must not be capitalized. Another error is found in token #3 that is a known PRON, but it is missing the morphological feature **PronType=Ind**, since all PRON tokens require the **PronType** feature. Token #9, the NOUN "*capetódromo*", is absent in the lexical resource, but, since NOUN is an open class, a warning is issued.

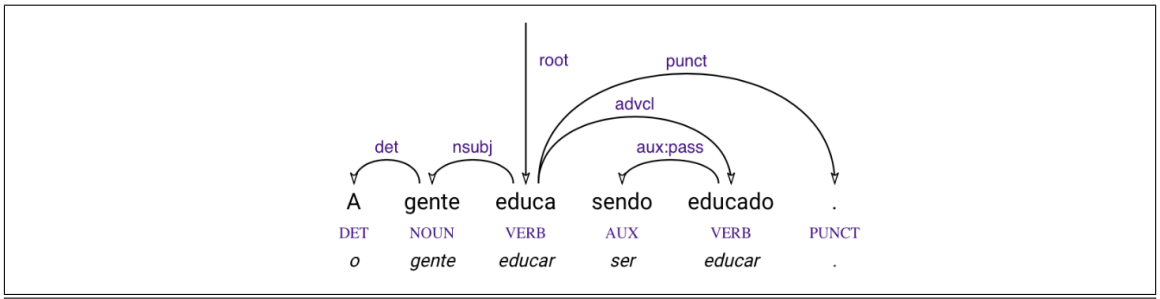
```
# sent_id = FOLHA_DOC009367_SENT031
# text = Se fizer algo errado, vai pro capetódromo.
1   Se  Se  SCONJ   _   _   2   mark    _   _
2   fizer   fazer   VERB    _   Mood=Sub|Number=Sing|Person=3|Tense=Fut|VerbForm=Fin    6   advcl   _   _
3   algo    algo    PRON    _   Gender=Masc|Number=Sing 2   obj _   _
4   errado  errado  ADJ _   Gender=Masc|Number=Sing|VerbForm=Part   3   amod    _   SpaceAfter=No
5   ,   ,   PUNCT   _   _   2   punct   _   _
6   vai ir  VERB    _   Mood=Ind|Number=Sing|Person=3|Tense=Pres|VerbForm=Fin   0   root    _   _
7-8 pro _   _   _   _   _   _   _   _
7   para    para    ADP _   _   9   case    _   _
8   o   o   DET _   Definite=Def|Gender=Masc|Number=Sing|PronType=Art   9   det _   _
9   capetódromo capetódromo NOUN    _   Gender=Masc|Number=Sing 6   obl _   SpaceAfter=No
10  .   .   PUNCT   _   _   6   punct   _   _
```

Figure 3. Examples of morphosyntactic problems, with (a) the wrong lemma for token #1 (it should be "*se*"), (b) the PronType=Ind is missing from FEATS of token #3, (c) token #9 "*capetódromo*" is absent as NOUN in the lexical resource (warning).

In total, the morphosyntactic verification may issue 29 possible errors and 14 possible warnings. The list of rules is available at a previous publication [Lopes et al. 2023b].

2.3. Dependency Relation Verification

Dependency relation verification applies to token relations, therefore, to token sequences belonging to the same branch of the dependency tree, and it concerns primarily the fields HEAD and DEPREL, but also their eventual relations with the fields FORM, LEMMA, POS, and FEAT. In CoNLL-U encoding, each DEPREL tag represents a dependency relation associating a dependent token, where the tag is, to a head token to which the token ID is in the field HEAD. For example, in Figure 4[2], token #2 "*gente*" is the dependent of the dependency relation **nsubj** that has the token #3 "*educa*" as the head of the relation.




```
# sent_id = FOLHA_DOC009367_SENT024
# text = A gente educa sendo educado.
1   A   o   DET _   Definite=Def|Gender=Fem|Number=Sing|PronType=Art    2   det _   _
2   gente   gente   NOUN    _   Gender=Fem|Number=Sing  3   nsubj   _   _
3   educa   educar  VERB    _   Mood=Ind|Number=Sing|Person=3|Tense=Pres|VerbForm=Fin   0   root    _   _
4   sendo   ser AUX _   VerbForm=Ger    5   aux:pass    _   _
5   educado educar  VERB    _   Gender=Masc|Number=Sing|VerbForm=Part|Voice=Pass    3   advcl   _   SpaceAfter=No
6   .   .   PUNCT   _   _   3   punct   _   _
```

Figure 4. CoNLL-U example of a sentence and corresponding dependency tree.

Because of this token relation aspect, the dependency relation rules are of different nature than the previous ones, often establishing restrictions to which kind of PoS tags can be dependent, or head of specific relations, or even establishing the need for some specific morphological feature to some dependency relations. Also at this dependency relation level, some situations are not necessarily errors, but they are unusual, and, as such, may provoke the indication of warnings. For example, the CoNLL-U representation in Figure 5 has three situations indicated. The first token provokes an error as Token #1 "*a*" has PoS tag PRON, but it is the dependent of a **det** relation, and one of the dependency relation rules states that all dependents of **det** relation need to be DET. Token #4 also indicates an error, as a token head of dependency relation **aux:pass** needs to be a VERB holding the morphological features **VerbForm=Part** and **Voice=Pass**. Finally, the last token (#6) provokes a warning as there is a dependency rule stating that final punctuation is usually dependent of a relation **punct** with the head being the sentence's **root**.

The dependency relation rules define 61 errors and 8 warnings. The full list of dependency rules is available at a previous publication [Lopes et al. 2023b].

[2]"A gente educa sendo educado" (We educate being educated or We educate being polite). Depending on the interpretation, "educado" may be a VERB in passive voice or an ADJ, and the AUX "sendo" (being) may be aux:pass ou cop, respectively.

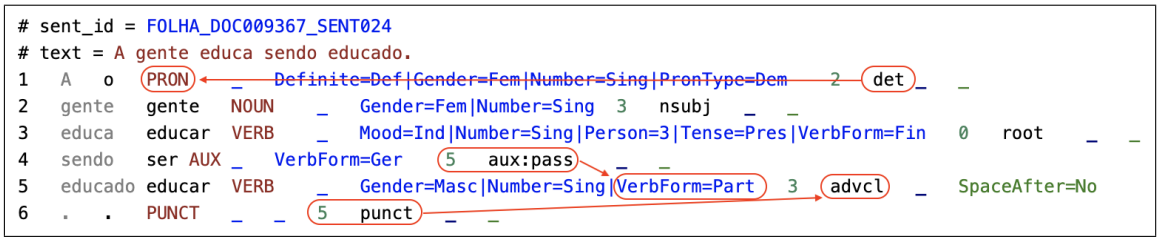


Figure 5. Examples of syntactic problems, with (a) a relation det involving token #1 that has PoS tag PRON, (b) the token #4 is dependent of relation aux:pass that requires as head a token VERB with features VerbForm=Part and Voice=Pass, (c) token #6, the final punctuation, should have relation punct with the root.

3. Verifica-UD Online Tool

Verifica-UD implements the three level rules for errors and warnings as defined in the previous section through a webservice tool[3]. This tool was developed with Python, implementing the server using REST API [Richardson and Ruby 2007], while the web interface was implemented using a HTML/CSS/PHP technology. The basic tool operation consists in uploading a CoNLL-U file (.conllu) that is automatically verified. Once uploaded, the CoNLL-U file is analyzed and a list of errors and warnings grouped by sentences sorted in alphabetical order is displayed as depicted in Figure 6. After performing the verification and displaying the errors and warnings, the tool offers the possibility of exporting a report that lists the sentences, the tokens, and the errors or warnings that were found.

Another feature of the interface is a set of help pages that allows the user to learn more about CoNLL-U annotation for Portuguese according to the directives in [Duran 2021, Duran 2022, Lopes et al. 2023a]. At the help interface, it is also possible to see the full list of errors and warnings currently detected by Verifica-UD grouped by the structural, morphosyntatic, and dependency relation level, as shown in Figures 7 and 8.

4. Evaluation of the Tool

To exemplify the benefits brought by the use of Verifica-UD, we conducted an experiment over a small corpus of 300 sentences (5,818 tokens) automatically annotated using UDPipe 2.0 parser [Straka 2018]. This set of automatically annotated sentences was previously submitted to a human revision that limit the corrections to the fields POS, HEAD, and DEPREL (ignoring LEMMA and FEAT fields of the automatically annotated CoNLL-U). We analyzed the initial corpus and the corpus revised through Verifica-UD to estimate the potential benefits of our tool to improve the human revision.

The human reviewer edited 173 out of the 300 sentences (58%), making amendments to the automatic annotation. We applied Verifica-UD over the original corpus, ignoring all rules that inspected the fields LEMMA and FEAT, as those fields were not contemplated by the human revision. This resulted in Verifica-UD indicating problems for 87 sentences over the 300 sentences (29%). The intersection of the human edited sentence set with Verifica-UD results was of 72 sentences. Observing these 72 edited sentences, we applied Verifica-UD to the human edited version and it turned out that 22

[3]Verifica-UD may be found at the POeTiSA project website (`https://sites.google.com/icmc.usp.br/poetisa`), but also at the links `http://verificaud.icmc.usp.br:24080/verificaud/` and `http://verificaud.icmc.usp.br/`.

Figure 6. Verifica-UD Interface displaying detected errors and warnings.

of those sentences still presented problems according to our rules. We also noticed that 101 sentences were edited by the reviewer without indication of problems by Verifica-UD when applied to the original corpus. In 10 of these 101 edited sentences, the human edition corrected the HEAD error but forgot to adjust the DEPREL, thus generating new errors and warnings by our tool.

The direct benefits of Verifica-UD can be illustrated in 47 out of the 300 sentences:

- the 15 sentences with problems identified by Verifica-UD, but unnoticed by the reviewer;
- the 22 sentences where the reviewer edited, but problems remained, revealing that the problem detected by the tool was not the same as the one corrected by the human annotator; plus
- the 10 edited sentences that resulted in new verification issues.

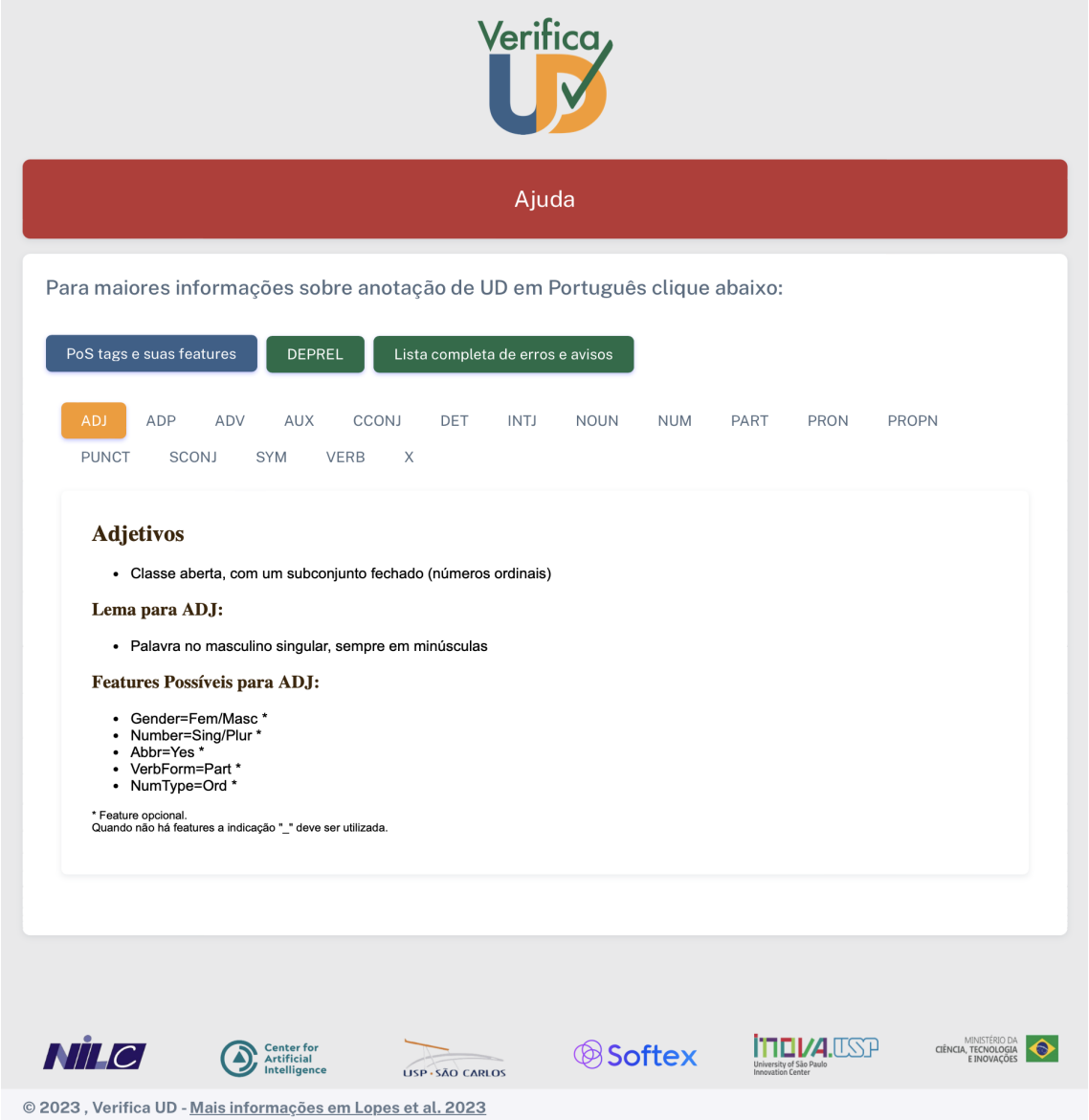


Figure 7. Verifica-UD Interface displaying help topics for tagging level.

Pushing the analysis a little further, we applied Verifica-UD with the full set of rules (including LEMMA and FEAT related ones) over the corpus version edited by the reviewer, and we discovered 81 additional sentences with potential problems. Therefore, in practical terms, using Verifica-UD, the human reviewer can be more productive in correcting these remaining 128 (47+81) sentences with problems pointed out by our tool.

5. Final Remarks

In this paper, we presented Verifica-UD, a web-based tool to verify possible problems in UD-annotated sentences (in CoNLL-U format) according UD guidelines for Portuguese. The tool has the potential to boost the production of annotated corpora in Portuguese, as well as to promote enhancement of the UD annotation in the Brazilian NLP community.

Verifica-UD was developed within the POeTiSA project (`https://sites.google.com/icmc.usp.br/poetisa`) and, according to our experiment, it

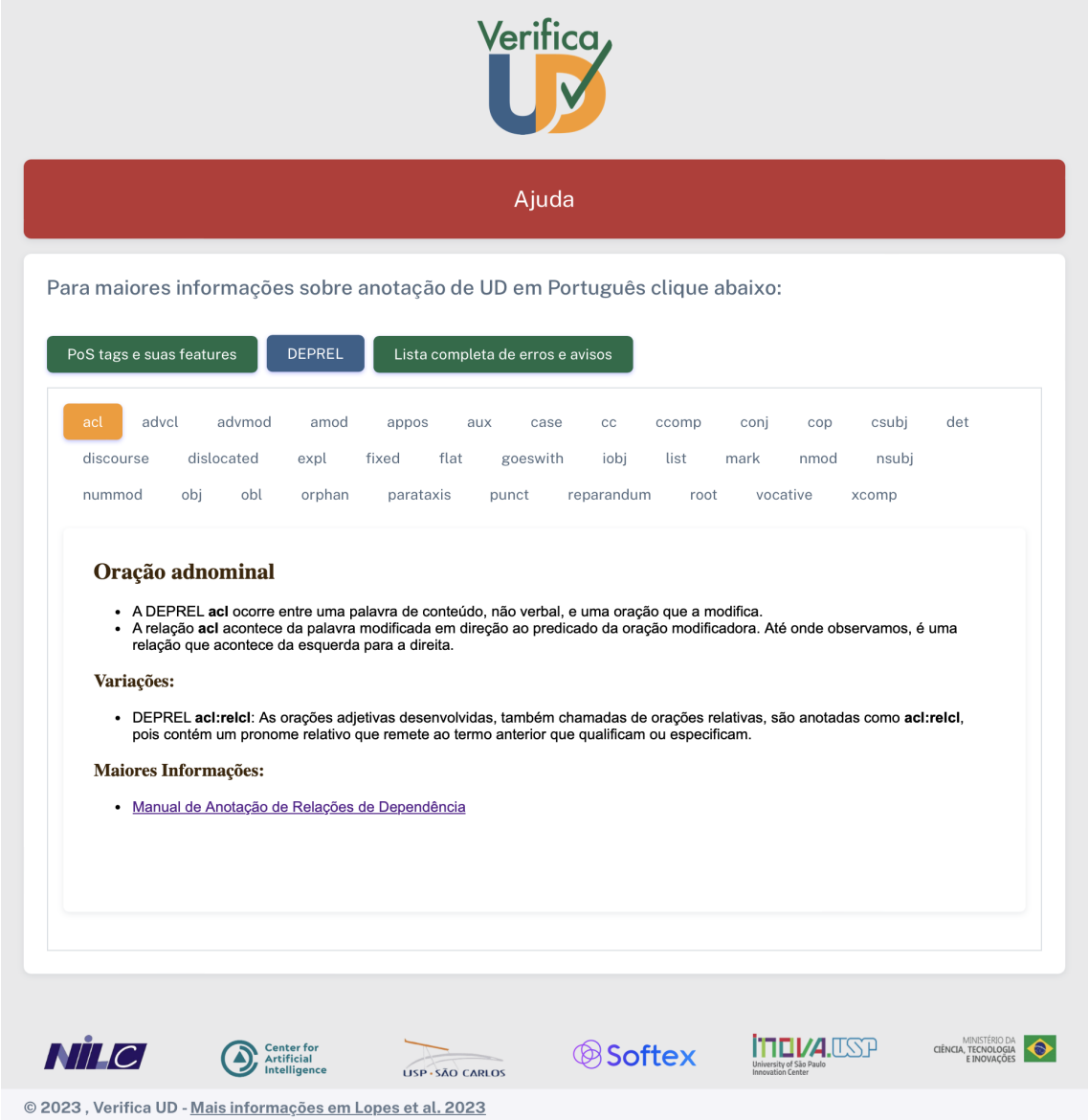


Figure 8. Verifica-UD Interface displaying help topics for parsing level.

presents a good performance in terms of problem detection. Additionally, the availability of detailed help pages facilitates the human correction of the issues.

Future work includes the addition of new verification rules. It is also our plan to extend the tool with new edition and visualization functions.

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

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

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). Technical Report 434, ICMC-USP.

Duran, M. S. (2022). Manual de anotação de relações de dependência: Orientações para anotação de relações de dependência em língua portuguesa, seguindo as diretrizes da abordagem universal dependencies (UD). Technical Report 440, ICMC-USP.

Grobol, L. (2021). VSCode language support for CoNLL-U. `https://github.com/LoicGrobol/vscode-conllu/blob/master/README.md`. Accessed: 2023-06-26.

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

Lopes, L., Duran, M., Fernandes, P., and Pardo, T. (2022). PortiLexicon-UD: a Portuguese lexical resource according to Universal Dependencies model. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference (LREC)*, pages 6635–6643, Marseille, France. European Language Resources Association.

Lopes, L., Duran, M. S., and Pardo, T. A. S. (2023a). Atribuição de lemas e atributos morfológicos seguindo as decisões adotadas na anotação do córpus Porttinari-base dentro das diretrizes da Universal Dependencies (UD). Technical Report -, ICMC-USP. To appear.

Lopes, L., Duran, M. S., and Pardo, T. A. S. (2023b). Verifica-UD - uma ferramenta online para verificação de textos em português anotados no formato CoNLL-U segundo o padrão Universal Dependencies. Technical Report -, ICMC-USP. To appear.

Miranda, L. G. M. and Pardo, T. A. S. (2022). UDConcord: a concordancer for universal dependencies treebanks. In *Proceedings of the Universal Dependencies Brazilian Festival (UDFest-BR)*, pages 1–10. Association for Computational Linguistics.

Nivre, J., de Marneffe, M.-C., Ginter, F., Hajič, J., Manning, C. D., Pyysalo, S., Schuster, S., Tyers, F., and Zeman, D. (2020). Universal Dependencies v2: An evergrowing multilingual treebank collection. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 4034–4043, Marseille, France. European Language Resources Association.

Richardson, L. and Ruby, S. (2007). *RESTful Web Services*. O'Reilly, Beijing.

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.

Villa, L. B. (2022). Udeasy: a tool for querying treebanks in conll-u format. In *Proc. of the Workshop on Challenges in the Management of Large Corpora (CMLC)*, pages 16–19, Marseille, France. European Language Resources Association.

Enhanced dependencies para o português brasileiro

Adriana S. Pagano[1], Magali Sanches Duran [2], Thiago Alexandre Salgueiro Pardo[2]

[1]Faculdade de Letras – Universidade Federal de Minas Gerais (UFMG)
Belo Horizonte – MG – Brasil
[2]Núcleo Interinstitucional de Linguística Computacional (NILC)
Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo (USP)
São Paulo - SP - Brasil

apagano@ufmg.br, magali.duran@uol.com.br, taspardo@icmc.usp.br

Abstract. *This article explores Universal Dependencies' guidelines for annotating the enhanced flavor of dependency relations and presents the main types of enhancement as applied to Brazilian Portuguese. It briefly discusses models for automatically converting basic into enhanced dependency relations with a view to their future implementation to enrich Brazilian Portuguese treebanks.*

Resumo. *Este artigo explora as diretrizes de anotação de "enhanced dependencies" do modelo "Universal Dependencies" e apresenta as principais configurações de anotação em português brasileiro, discutindo a relevância e as aplicações desse tipo de informação. São também discutidos modelos automáticos de conversão das relações de dependência básicas para "enhanced" e tecidas considerações sobre seu uso para o português.*

1. Introdução

As chamadas *enhanced dependencies*[1] (ED) visam o enriquecimento de informações em *treebanks* anotados com relações de dependência básicas da abordagem *Universal Dependencies* – UD (de Marneffe et al., 2021). Esse enriquecimento se mostra útil para aplicações mais avançadas, envolvendo semântica, como a recuperação de informações e a anotação automática de papéis semânticos (Nivre et al., 2018; Ek & Bernady, 2020).

Atualmente, no âmbito da UD, um número ainda limitado, embora crescente, de línguas disponibiliza *treebanks* anotados com ED, como é o caso das línguas inglesa, espanhola, russa, polonesa, finlandesa e holandesa, para citar algumas delas. A língua portuguesa não conta ainda com *treebanks* anotados com ED, lacuna que justifica o esforço aqui apresentado de interpretar as diretrizes da UD quanto à anotação das ED e de ilustrar as explicações com exemplos extraídos de *corpora* de português brasileiro. Busca-se, portanto, estimular e facilitar a implementação das ED em *treebanks* de português.

A seguir, na Seção 2, relatamos sucintamente a evolução da proposta de *enhanced dependencies*. Na Seção 3 discutimos a infraestrutura computacional para anotação e visualização das ED. Na Seção 4 são apresentadas as principais configurações sintáticas passíveis de serem enriquecidas e suas respectivas anotações,

[1] Apesar de ser possível traduzir o termo, optamos pelo termo em inglês, por ser já conhecido no Brasil.

seguindo as diretrizes da UD. Na Seção 5, são discutidos trabalhos que exploram modelos automáticos de conversão de relações UD para relações ED e são tecidas considerações sobre sua possível utilização em *corpora* de português.

2. *Enhanced dependencies*

A UD tem, entre seus precursores, o modelo de dependências de Stanford, que já previa dois tipos de anotação, de acordo com os requisitos das tarefas nas quais eram utilizados: anotação de relações básicas e anotação de relações do tipo *collapsed* (ou *cc processed*) (de Marneffe et al., 2006; de Marneffe & Manning, 2008). As relações *collapsed* permitem anotar cada preposição (classe de palavra relevante enquanto indicadora de papéis semânticos) e cada conjunção coordenativa juntamente com o head das relações de dependência das quais participam (*nmod, obl, conj e advcl*), auxiliando em tarefas como a extração de informação (Schuster et al., 2017).

Schuster & Manning (2016) expandiram as representações do tipo *collapsed* para além de preposições e conjunções, e as adaptaram às diretrizes da UD. As representações ganharam o nome de *enhanced dependencies*. Os autores avaliaram seu uso na extração de informação e detectaram alguns problemas de anotação que não podiam ser resolvidos nem pelas relações básicas, nem pelas relações *enhanced.* Isso os levou a propor um terceiro tipo de anotação, que denominaram *enhanced dependencies* ++, ainda não implementada em português, cuja discussão extrapola o escopo deste artigo..

Para a anotação de ED em arquivos CoNLL-U, que é o formato de arquivo padrão da UD, é reservada a nona coluna, rotulada DEPS. Nesta coluna, temos a relação das arestas que chegam a cada uma das palavras (*tokens*) da sentença. A representação obtida por meio da anotação é um grafo, mas não necessariamente uma árvore, uma vez que podem haver nós vazios, várias arestas chegando a um mesmo nó e ocorrência de ciclos.

Droganova & Zeman (2019) esclarecem que a anotação das ED é opcional em *treebanks*, podendo-se anotar apenas uma, várias ou todas as configurações previstas. Contudo, uma vez decidido incluir um dos tipos de ED, é fundamental que isso seja feito no *corpus* inteiro, por uma questão de consistência.

3. Anotação e Visualização das ED

O estudo das ED e o esforço para instanciá-las em língua portuguesa nos levou a perceber que há uma carência muito grande de ferramentas para anotá-las e visualizá-las, embora para anotar e visualizar as relações básicas da UD existam várias ferramentas[2] . A única ferramenta que encontramos para edição das ED foi o CONLL-U Editor[3], utilizado para construir as árvores das figuras que ilustram este artigo. Para visualização, encontramos ainda as ferramentas Grew-Web[4] e Inception[5].

[2] https://universaldependencies.org/tools.html
[3] https://github.com/Orange-OpenSource/conllueditor
[4] https://web.grew.fr/
[5] https://inception-project.github.io/

Há discussões sobre a melhor forma de visualizar o resultado da anotação das ED[6]. Uma opção é visualizar as ED simultaneamente à visualização das dependências básicas, umas sobrepostas às outras. Isso seria simples se as ED só fizessem acréscimos, porém dois tipos de EDs alteram algumas das relações básicas e, portanto, não há como visualizá-las sob um mesmo plano. Uma alternativa é mostrar as ED abaixo da sentença anotada e as relações básicas acima da sentença anotada, o que é feito pela ferramenta CONLLU-Editor, como pode ser observado nas figuras que ilustram este artigo. A outra alternativa, adotada nas diretrizes da UD, é não visualizá-las simultaneamente. Opta-se por visualizar as relações básicas (anotadas na coluna 8 do CONLLU) ou por visualizar as ED (anotadas na coluna 9 do CONLLU). Em qualquer das alternativas, o recurso de "pintar" as relações não compartilhadas pelas colunas 8 e 9 nos pareceu muito bom para fins de visualização. As instruções da UD acerca da anotação das ED apresentam um "antes" (visualização da coluna 8) e um "depois" da anotação das ED (visualização da coluna 9) e pintam de vermelho as relações básicas não compartilhadas com as ED e de verde as relações das ED não compartilhadas com as relações básicas. Simplificando, o *diff* entre as colunas 8 e 9 aparece em vermelho na visualização da coluna 8 e em verde na visualização da coluna 9 e em ambas visualizações as relações compartilhadas aparecem em preto, sem destaque portanto.

Nas Figuras exibidas na Seção 4, por motivo de economia, só replicamos as relações compartilhadas entre as colunas 8 e 9 no caso em que há inserção de token e muitas mudanças de relações decorrentes dessa inserção (4.2.2). Nas demais figuras, as relações compartilhadas só são exibidas na parte superior, o que não significa que não estejam presentes também na parte inferior (apenas foram ocultadas). Utilizamos a cor vermelha para mostrar, na parte superior, o *diff* das relações básicas em relação às ED e a cor azul para mostrar, na parte inferior, o *diff* das ED em relação às relações básicas.

4. Configurações passíveis de serem anotadas com *enhanced dependencies*

De acordo com as orientações fornecidas pela UD[7], há seis casos previstos de anotação de ED. Em linhas gerais, acreditamos que essas seis ED podem ser agrupadas em duas categorias: aquelas que produzem um acréscimo de informações às dependências básicas (exemplificadas na Seção 4.1) e aquelas que produzem uma cópia modificada das dependências básicas (exemplificadas na seção 4.2).

4.1 *Enhanced Dependencies* de acréscimo

As ED que apenas acrescentam informações às dependências básicas são de quatro tipos: as que reproduzem, nos *heads* das relações *nmod*, *obl*, *conj,* e *advcl*, as preposições e as conjunções que introduzem seus dependentes (4.1.1), as que promovem a propagação de sujeitos de *xcomp* (4.1.2), as que propagam *head* compartilhado de elementos coordenados (4.1.3) e as que propagam dependentes compartilhados por elementos coordenados (4.1.4).

[6] Vide discussão sobre o assunto no Fórum da UD em https://github.com/UniversalDependencies.
[7] https://universaldependencies.org/u/overview/enhanced-syntax.html

4.1.1 Acréscimo de preposições e conjunções

Esta configuração enriquece relações de dependência como *nmod, obl, conj* e *advcl,* acrescentando-lhes uma preposição ou uma conjunção com a qual constroem relações semânticas. A anotação pode inclusive conter especificação do significado construído pela preposição ou conjunção. Se o treebank não tiver sub-relações de *nmod*, *obl, cc* e *advcl* não será possível herdar informações semânticas e, portanto, seria necessário anotá-los manualmente caso sejam de interesse do projeto de anotação de ED. No caso das preposições, constroem-se significados relativos aos casos *gen* (genitivo), *loc* (locativo), *tem* (temporal), *ins* (instrumental), *dat* (dativo), *acc* (acusativo) e outros. O exemplo (7) ilustra esta configuração, especificando nas ED que o verbo "proibiu" apresenta um *obl* introduzido pela preposição "por" com o papel semântico de temporal (*tem*), e outro *obl* introduzido pela preposição "em" com papel semântico de locativo (*loc*). O grafo pode ser visualizado na Figura 1.

(1) O governo proibiu por 120 dias as queimadas em todo o Brasil.

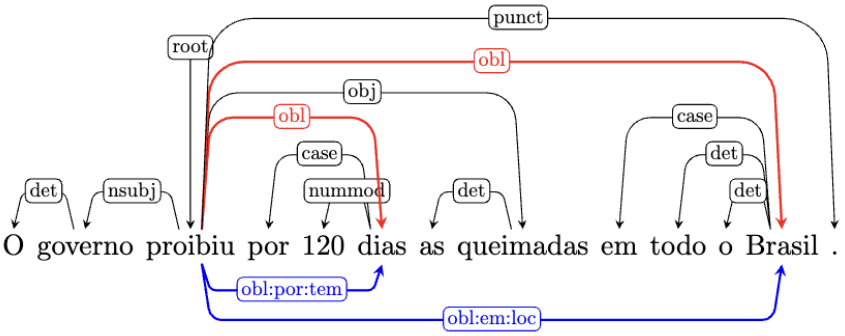


Figura 1. Heads de modificadores enriquecidos com marcadores de caso e papéis semânticos

4.1.2 Propagação de sujeitos de *xcomp*

Esta configuração abrange sentenças com complementos oracionais abertos (*xcomp*), nas quais o sujeito da oração subordinada é nulo (não expresso), mas é controlado pelo sujeito ou pelo objeto da oração matriz. Nas *enhanced dependencies*, anota-se essa relação entre a oração subordinada e o sujeito ou objeto da oração matriz, utilizando a relação *nsubj:xsubj*. O exemplo (2) mostra o sujeito da oração subordinada controlado pelo sujeito da oração matriz, enquanto o exemplo (3) mostra o sujeito da oração subordinada controlado pelo objeto da oração matriz. Os grafos desses dois exemplos podem ser visualizados, respectivamente, nas Figuras 2 e 3.

(2) O governo decidiu proibir as queimadas.

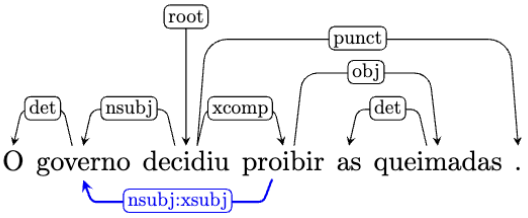


Figura 2. Sujeito da subordinada controlado pelo sujeito da oração matriz

(3) O governo convenceu a oposição a votar no projeto.

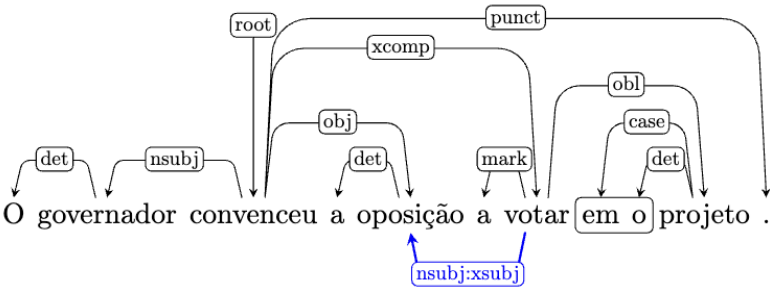


Figura 3. Sujeito da subordinada controlado pelo objeto da oração matriz

4.1.3 Propagação de *head* compartilhado por elementos coordenados

A UD conta separadamente, como dois tipos de ED, o *head* compartilhado por dois elementos coordenados e os dependentes compartilhados por dois elementos coordenados.

A ED de *head* compartilhado ocorre quando há coordenação entre múltiplos elementos que são dependentes de um mesmo *head*, podendo ser vários sujeitos ou objetos de um mesmo predicado ou vários modificadores de um mesmo sintagma nominal. Esses casos são anotados nas dependências básicas com a relação *conj* na direção do *head* para o primeiro elemento coordenado, sendo que os demais elementos coordenados são vinculados ao primeiro. Nas *enhanced dependencies*, estabelecem-se relações entre cada um dos coordenados e o *head* da coordenação. O exemplo (4) ilustra esta configuração, cujo grafo está na Figura 4.

(4) A escolha de Rússia e Qatar para as duas próximas Copas desarranjou a Fifa.

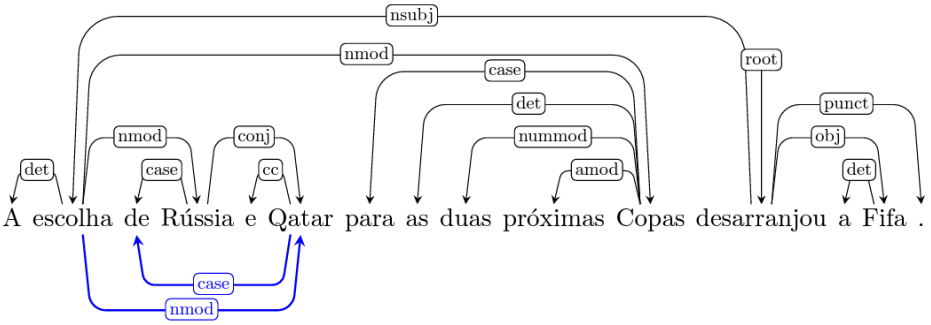


Figura 4. Coordenação de múltiplos elementos dependentes de um mesmo *head*

As relações de dependência básicas na Figura 4 mostram a coordenação de "Rússia" e "Qatar" por meio da relação *conj*, sendo que apenas a primeira está em relação de *nmod* com "escolha". Já nas *enhanced dependencies*, em azul, explicita-se a relação *nmod* entre "Qatar" e "escolha" e a relação de *case* entre "Qatar" e "de".

4.1.4 Propagação de dependentes de elementos coordenados

Já a ED de dependentes compartilhados ocorre quando há dois ou mais elementos coordenados (ligados pela relação *conj*) que possuem um ou mais dependentes em comum. Nas relações de dependência básicas, os dependentes estão ligados a apenas um dos elementos coordenados. Já nas *enhanced dependencies*, os dependentes compartilhados recebem uma ligação para cada um dos elementos coordenados (o que faz com que um mesmo dependente esteja ligado a mais de um *head*). O exemplo (5), cujo grafo pode ser visualizado na Figura 5, ilustra esta configuração, pois o sujeito (*nsubj*) é compartilhado pelos predicados verbais coordenados "pica" e "apresenta".

(5) O inseto pica durante o dia e apresenta fotofobia.

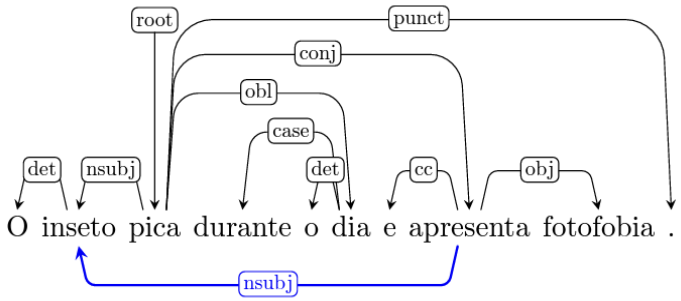


Figura 5. Coordenação de múltiplos predicados com um mesmo dependente

4.2 Enhanced Dependencies de Modificação

As duas ED que modificam as relações da árvore de dependências básicas são: a anotação de referentes de pronomes relativos e a inclusão de predicados elípticos. Essas relações não são tão fáceis de serem anotadas automaticamente e, por isso, demandam maior intervenção humana, seja para anotá-las do zero, seja para revisá-las.

Esses dois tipos de ED justificam algo que vem sendo apresentado por alguns *corpora* que anotam ED e que parece ter se tornado padrão nas diretrizes da UD: a replicação, na coluna 9 (dedicada às ED), de todas as relações compartilhadas com a coluna 8 (dedicada às relações básicas). A seguir explicamos e exemplificamos a anotação das ED de referente de pronomes relativos (4.2.1) e de inserção de token de predicado elíptico (4.2.2).

4.2.1 Anotação do referente dos pronomes relativos

Esta configuração ocorre sempre que há orações relativas, nas quais os pronomes relativos estabelecem uma relação de correferência com o antecedente nominal que retomam. As dependências básicas anotam a função do pronome relativo em relação ao

predicado da oração relativa. Nas ED, o antecedente é vinculado ao pronome relativo que o retoma por meio da relação *ref*. A relação que unia o predicado da oração relativa ao pronome relativo passa ser feita diretamente com o antecedente. O exemplo (6) ilustra esta configuração, cujo grafo pode ser visualizado na Figura 6.

(6) Conheça os 21 livros que o ex-presidente já leu na prisão.

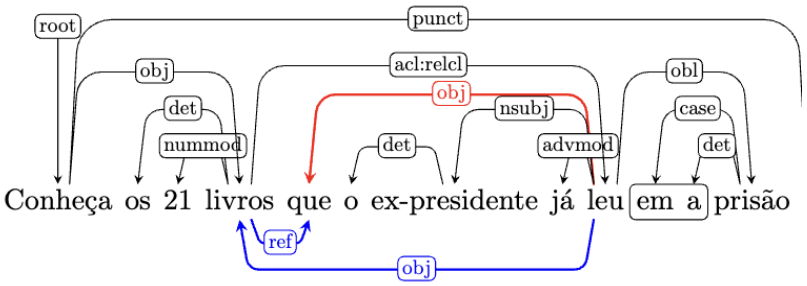


Figura 6. Relação de correferência em construções de orações relativas

4.2.2 Inclusão de predicado elíptico

Esta configuração diz respeito a sentenças nas quais há duas orações, sendo que em uma delas há a realização de um predicado e seu sujeito, enquanto na outra apenas o sujeito é realizado e há elipse do predicado. Nas relações de dependência básicas, sentenças como essas são anotadas com a relação *orphan* para indicar que há um predicado elíptico. Nas *enhanced dependencies*, insere-se um *token* nulo para representar o predicado elíptico e receber as relações a ele devidas. Esse *token* recebe também o lema, a categoria morfossintática e os atributos morfológicos da palavra elíptica. Cumpre destacar que esse tipo de elipse (de predicado) é a única configuração que permite a inserção de um nó vazio na UD. O exemplo (7) ilustra essa configuração de elipse do segundo predicado de uma coordenação. A Figura 7 apresenta uma visualização do grafo dessa sentença, inserindo o token do verbo elíptico "ficam".

(7) A gente fica preso e os bandidos, soltos.

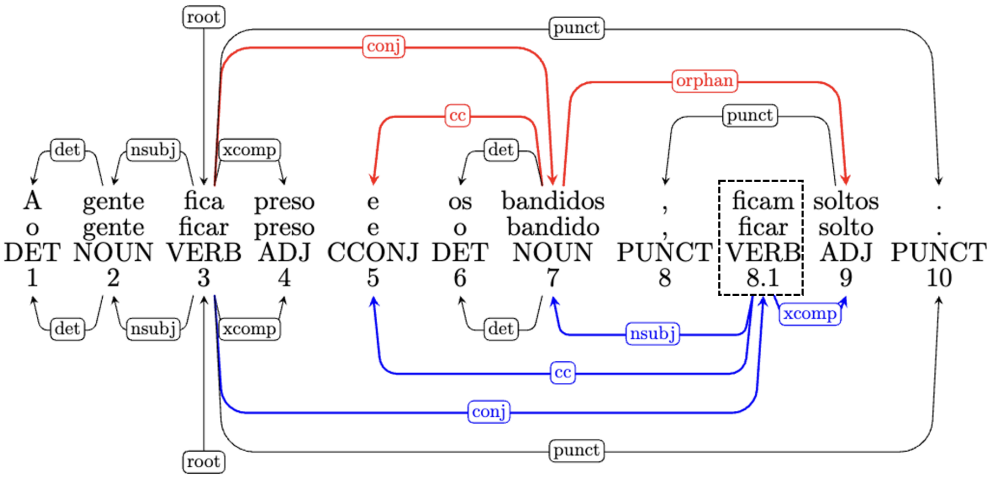


Figura 7. Anotação de predicado elíptico

Na parte de cima da Figura 7, vemos as relações de dependência básicas, nas quais a relação entre "bandidos" e "soltos" é anotada como *orphan*, uma vez que há elipse do verbo. Na parte de baixo, vemos as ED, mostrando a inserção de um *token* no lugar do verbo elíptico, destacado com linha pontilhada na Figura 7. Esse *token* recebe a numeração do *token* anterior (neste caso, 8) juntamente com a indicação de que se trata do primeiro *token* inserido (8.1). Recebe também a forma da palavra que preenche a elipse ("ficam"), além do respectivo lema ("ficar") e a classe de palavra (VERB) com seus atributos morfológicos. As relações que o *token* inserido estabelece com as outras palavras estão destacadas em azul: *head* de *nsubj* com "bandidos", dependente de *conj* com "fica", *head* de *xcomp* com "soltos", e *head* de *cc* com a conjunção "e".
Uma dificuldade adicional a esse tipo de anotação em português é o fato de que, embora o verbo já tenha aparecido na sentença, nem sempre a forma será repetida, pois o preenchimento da elipse pode exigir uma flexão diferente.

5. O enriquecimento de *treebanks* em português com *enhanced dependencies*

Uma vez definidas as configurações de ocorrência de *enhanced dependencies* em português e suas diretrizes de anotação, o próximo passo natural é proceder à anotação das *enhanced dependencies* nos *treebanks* já anotados de acordo com a UD para o português, como o Bosque (Rademaker et al., 2017), o Porttinari (Pardo et al., 2021) e o PetroGold (Souza et al., 2021), entre outros.

A anotação de ED conta com iniciativas automáticas relatadas na literatura da área, valendo-se normalmente da conversão automática de relações de dependência básicas para *enhanced*. Essa alternativa é interessante, pois permite que anotadores humanos avaliem a qualidade da anotação e revisem os casos necessários, como relatado por Nivre et al. (2018). Schuster & Manning (2016) e Grünewald et al. (2021), por exemplo, desenvolveram conversores de relações de dependência básicas para ED para a língua inglesa. Na avaliação deles, embora a conversão tenha gerado resultados com alta acurácia, houve casos nos quais os grafos obtidos construíram significados distintos daqueles construídos pelas sentenças de origem.

Nivre et al. (2018) exploraram duas técnicas de conversão automática aplicadas a múltiplas línguas, uma delas adaptando o conversor de Schuster & Manning (2016) e a outra adaptando o conversor de Nyblom et al. (2013) e avaliaram os resultados como satisfatórios. Heinecke (2020) também relataram bons resultados por meio de uma abordagem híbrida que inclui o *parsing* para a extração de relações básicas e a aplicação de regras linguísticas para a geração de relações *enhanced*.

Portanto, para proceder à anotação de ED para o português, é interessante testar um conversor ou *script* já utilizado por outras línguas e avaliar os resultados. Se forem satisfatórios, passa-se diretamente à revisão manual; se não, procede-se primeiramente à adaptação do conversor utilizando regras específicas para o português, para depois fazer a revisão manual. Como hipotetizamos anteriormente, há relações ED que apresentam maior probabilidade de serem automatizadas do que outras, o que pode indicar a conveniência de se proceder à anotação de um tipo de ED de cada vez. Uma vez concluído esse trabalho, os *treebanks* anotados com *enhanced dependencies* poderão ser

usados para treinar classificadores que realizem essa anotação automaticamente, sem o uso de regras, dedicados à língua portuguesa.

Agradecimentos

Adriana S. Pagano é bolsista de Produtividade em Pesquisa do Conselho Nacional de Desenvolvimento Científico e Tecnológico (Processo CNPq 313103/2021-6). Magali Duran e Thiago Pardo realizaram este trabalho no âmbito do Centro de Inteligência Artificial da Universidade de São Paulo (C4AI - http://c4ai.inova.usp.br/), com o apoio da Fundação de Amparo à Pesquisa do Estado de São Paulo (processo FAPESP #2019/07665-4) e da IBM. Também receberam apoio do Ministério da Ciência, Tecnologia e Inovações, com recursos da Lei N. 8.248, de 23 de outubro de 1991, no âmbito do PPI-Softex, coordenado pela Softex e publicado como Residência em TIC 13, DOU 01245.010222/2022-44.

Referências

Candito, M.; Guillaume, B.; Perrier, G.; Seddah, D. (2017) Enhanced UD Dependencies with Neutralized Diathesis Alternation. In Proceedings of the Fourth International Conference on Dependency Linguistics, pages 42-53.

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.

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. Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo.

Droganova, K.; Zeman, D. (2019) Towards Deep Universal Dependencies. In Proceedings of the Fifth International Conference on Dependency Linguistics, pages 144-152.

Ek, Adam; Bernardy, Jean Philippe. (2020) How much of enhanced UD is contained in UD? Proceedings of the16th International Conference on Parsing Technologies and the IWPT 2020 Shared Task, pages 221–226. Virtual Meeting, July 9, 2020.c2020 Association for Computational Linguistics.

Grünewald, S.; Piccirilli, P.; Friedrich, A. (2021) Coordinate Constructions in English Enhanced Universal Dependencies: Analysis and Computational Modeling. In Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, pages 795-809.

Heinecke, J. (2020) Hybrid Enhanced Universal Dependencies Parsing. In Proceedings of the 16th International Conference on Parsing Technologies and the IWPT 2020 Shared Task on Parsing into Enhanced Universal Dependencies, pages 174-180.

de Marneffe, M.-C.; MacCartney, B.; Manning, C.D. (2006. Generating Typed Dependency Parses from Phrase Structure Parses. In Proceedings of the Fifth International Conference on Language Resources and Evaluation, pages 449-454.

de Marneffe, M.-C.; Manning, C.D. (2008) The Stanford Typed Dependencies Representation. In Proceedings of the workshop on Cross-Framework and Cross-Domain Parser Evaluation, pages 1-8.

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

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

Nyblom, J, Kohonen, S.; Haverinen, K.; Salakoski, T.; and Ginter, F. (2013) Predicting conjunct propagation and other extended stanford dependencies. In Proceedings of theSecond International Conference on Dependency Linguistics (DepLing2013). pages 252–261.

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 Proceedings of the XIII Symposium in Information and Human Language, pages 1-10.

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

Schuster, S.; Manning, C.D. (2016) Enhanced English Universal Dependencies: An Improved Representation for Natural Language Understanding Tasks. In Proceedings of the Tenth International Conference on Language Resources and Evaluation, pages 2371-2378.

Schuster, S., de La Clergerie, É. V., Candito, M. D., Sagot, B., Manning, C. D., & Seddah, D. (2017) Paris and Stanford at EPE 2017: Downstream Evaluation of Graph-based Dependency Representations. In EPE 2017-The First Shared Task on Extrinsic Parser Evaluation, pages 47-59.

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

A dependency-based study of medicine package inserts in Brazilian Portuguese

Adriana S. Pagano[1], André V. Lopes Coneglian[1], Lucas Emanuel Silva e Oliveira[2]

[1]Faculdade de Letras – Universidade Federal de Minas Gerais (UFMG)
Belo Horizonte – MG – Brasil
[2]Pontifícia Universidade Católica do Paraná (PUCPR)
Curitiba - PR - Brasil

{apagano,coneglian}@ufmg.br,lucas.oliveira@pucpr.br

Abstract. *This paper reports on a study of medicine package inserts (MPIs) aimed at verifying to what extent texts addressing patients evidence different morphosyntactic patterns from those addressing HC professionals. To that end, we draw on a corpus of sentences manually retrieved and aligned, which were annotated for dependency syntax following the UD guidelines. Results point to clear distinctive patterns in both sets of MPIs, which are in line with guidelines on simplified language for Brazilian Portuguese.*

1. Introduction

Medicine package inserts (henceforth MPIs) are acknowledged as critical texts in healthcare activities, particularly in countries where people have less access to medical advice to clarify doubts on how to take prescription drugs or resort to over-the-counter medications sold with no prescription at all. MPIs are texts generally regulated by governmental institutions, which dictate standards for pharmaceutical companies to follow. This is the case, for instance, in the European Union and the United States (Pires et al., 2015). In Brazil, MPIs are required to follow a standard by the National Agency of Sanitary Surveillance (ANVISA) both for format and content. Pursuant to Resolution 047 (ANVISA, 2009a), MPIs are required to comprise two separate sections, one addressing patients and another one, healthcare professionals (henceforth HC professionals). ANVISA (2009b) has also published guidelines with best practices regarding the language used in MPIs, including recommendations on accessibility for blind and deaf people.

However, despite pharmaceutical companies' efforts to adhere to ANVISA guidelines, studies on the legibility and understandability of MPIs have shown that texts still pose enormous challenges to patients. Pizzol et al. (2019), for one, carried out a national survey of over 28,000 individuals in Brazil, revealing that although almost 60% of respondents found MPIs relevant texts to read, over 50% of them reported difficulties in reading and understanding them. To make matters worse, respondents with a lower literacy level reported greater difficulty in understanding MPIs.

While some linguistic studies have been carried out on the language of MPIs in Brazil (Amorim et al., 2015), to the best of our knowledge no study has been carried out on texts annotated for dependency relations with a view to comparing MPIs targeting patients and those targeting HC professionals. If text in MPIs is purportedly adapted to a specific target audience (patients vs. HC professionals), differences in language patterns

are expected to be found and these patterns are expected to comply with patterns suggested in text simplification tasks. This paper reports on an exploratory study of the syntax in MPIs aimed at verifying to what extent texts addressing patients evidence different morphosyntactic patterns from those addressing HC professionals. To that end, we draw on a corpus of sentences manually annotated for dependency syntax following the UD guidelines.

The remainder of this paper is structured as follows. Section 2 provides an overview of previous work on syntax complexity indicators and text simplification with a special focus on Brazilian Portuguese. Section 3 describes the corpus compiled and annotated in our study and the steps followed to analyze our data. Section 4 presents the results of our analysis and our findings regarding MPIs comparison. Section 5 discusses our findings with regard to the available literature. Finally, Section 6 presents the main conclusions of our study, its limitations and suggestions for further research.

2. Syntax complexity and text simplification

Drawing on cognitive and psycholinguistic research, studies on text simplification have investigated a number of language indicators for text simplification tasks. These include choices in morphology, lexis and syntax at sentence level as well as in cohesion at discourse level (Siddharthan, 2006). These indicators have been used in different approaches to the simplification task, relying on more or less manual annotation and implementing different solutions such as Phrase-Based Machine Translation, Syntax Based Machine Translation, transformation rules, and methods drawn from other computational tasks, as is text compression (Siddharthan et al., 2014). More recently, studies have begun to explore dependency trees in order to propose rules for simplification. Angrosh, Nomoto & Siddharthan (2014) explore dependency trees to perform lexical and syntactic simplifications. Chatterjee & Agarwal (2021) developed a rule-based tool (DEPSYM) for simplification drawing on dependency trees and focusing on coordinate and subordinate clauses (appositive and relative clauses) and passive-to-active voice conversion.

Likewise, in Brazil, the Interinstitutional Center for Computational Linguistics (NILC) in the State of São Paulo has developed corpora and tools to simplify texts (Aluísio et al., 2008a,b; 2010; Leal et al., 2022) focusing on clause complexes involving coordination and subordination. Hence, subordinate noun clauses functioning as apposition, relative clauses, and adverbial clauses are filtered out and turned into individual sentences. Moreover, passive voice constructions are rewritten into their active voice counterparts. With regards to MPIs, ANVISA itself published guidelines for drafting texts targeting patients (Brasil, 2009). To comply with them, pharmaceutical companies are instructed to use colloquial instead of medical terms, avoid coordinate and subordinate clauses, use verbs rather than abstract nominalizations and prefer active voice constructions to passive voice ones.

To the best of our knowledge, no work has reported on studies exploring dependency syntax for the purposes of text simplification in Brazilian Portuguese. Nor has any study been reported on using dependency syntax to compare texts targeting different readerships with different levels of literacy and domain knowledge. In this respect, MPIs offer a valuable source for corpus compilation of monolingual texts and

their annotation with dependency relations, a fertile approach to gather insights for prospective text simplification tasks.

3. Methodology

In order to carry out our study, we first retrieved MPI texts written and published in Brazil, targeting patients and HC professionals. Sentences representative of each target group were manually extracted in order to compile a comparable corpus. Since not all pieces of information in HC professionals MPIs are included in patient MPIs, for each sentence in patient MPIs we manually selected a counterpart sentence construing a closely analogous meaning in HC professional MPIs. Table 1 shows manually retrieved and aligned pairs of sentences illustrating analogous meaning construed in a patient and a HC professional MPI. An English gloss is provided beneath them.

Table 1. Aligned sentence pairs

	Patient MPI	HC professional MPI
(1)	Se ocorrerem reações cutâneas, como vermelhidão na pele, bolhas e erupções cutâneas, ou qualquer outro sinal de hipersensibilidade, ou ainda piora de problemas de pele já existentes, interrompa o uso do medicamento e procure ajuda médica imediatamente.	O uso do medicamento deve ser descontinuado no primeiro aparecimento de erupções cutâneas ou qualquer outro sinal de hipersensibilidade.
gloss	*If skin reactions, such as redness, blistering and eruptions, or any other sign of hypersensitivity occur or, still, if conditions worsen, stop use of this medication and seek medical attention right away*.	*Use of the drug should be discontinued at the first appearance of skin eruptions or any other sign of hypersensitivity.*
(2)	Não utilize NALDECON DIA juntamente com outros medicamentos que contenham paracetamol.	NALDECON DIA não deve ser usado juntamente com outros medicamentos que contenham paracetamol em sua formulação, devido ao risco de toxicidade hepática.
gloss	*Do not take NALDECON DAY together with other paracetamol-containing products.*	*NALDECON DAY should not be administered with other preparations containing paracetamol in their formulations, due to the risk of hepatotoxicity.*

As the examples in Table 1 show, there is variation in the length of the aligned segments, some of the sentences in patient MDIs being, at times, longer than their counterparts in HC professional MDIs, while at other times the reverse is the case. 200 sentences were selected for each target group, making up a corpus of 400 sentences. Table 2 shows basic statistics of our corpus, which reveal that despite variability in length, sentences in HC professional MPIs have a higher overall number of tokens for the whole set of 200 sentences and a higher average number of tokens per sentence.

Table 2. MPIs corpus

	Patient	HC professional
Total number. of sentences	200	200
Total number of tokens	2762	4816
Average number of tokens per sentence	13.81	24.08

Sentences were parsed using a freely available neural network pipeline for tokenization, tagging, lemmatization and dependency parsing (UDPipe[1]) with a Portuguese language model (Portuguese-bosque-UD-2.10). The output CoNLL-U files were then uploaded into the Arborator Grew NILC[2] tool developed by ICMC/USP and manually revised following the latest annotation guidelines for Brazilian Portuguese (Duran, 2021, 2022).

In order to compare both sets of texts, we computed the total number of tokens and the average number of tokens per sentence. We then computed POS and dependency relation tags and their relative frequencies in order to allow for comparability between the two sets. We focused on tags indicative of coordinated, subordinated and passive constructions in order to verify if texts followed the guidelines available for simplified Portuguese.

4. Results

Figure 1 shows the number of POS tags annotated for each set of MPIs and their relative frequency.

[1] Available at https://lindat.mff.cuni.cz/services/udpipe/
[2] Available at https://arborator.icmc.usp.br/#/

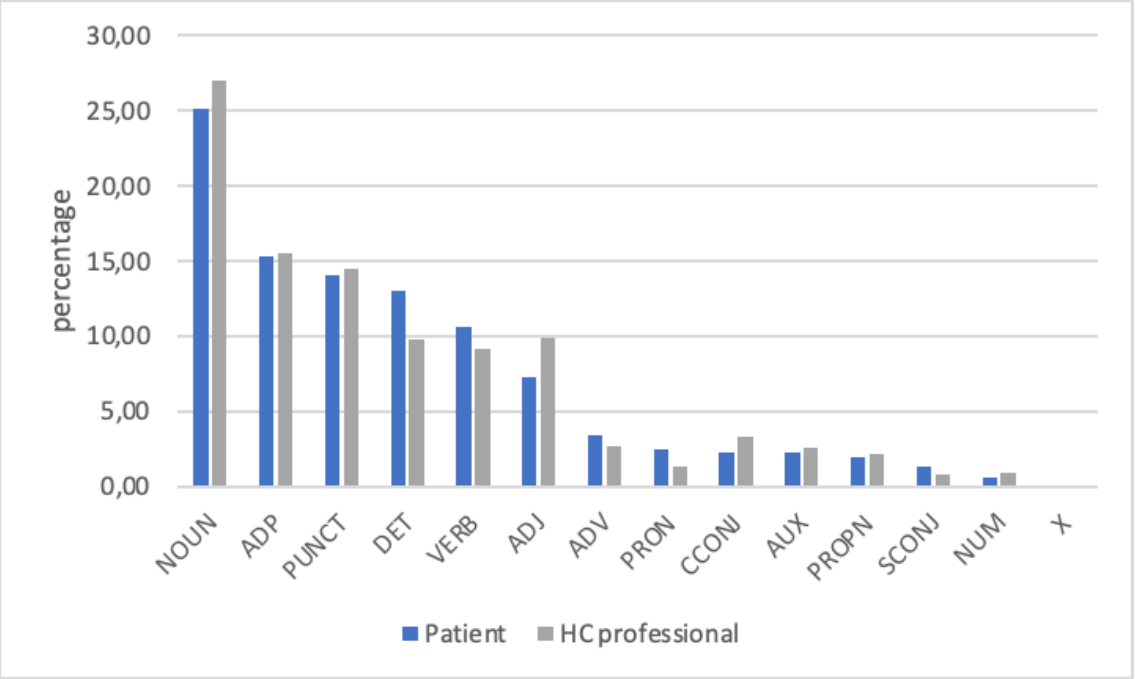


Figure 1. Relative frequency of POS tags

The frequency of POS tags shows that patient MPIs have a higher number of verbs when compared to HC professionals, which, instead, have a larger number of nouns and adjectives. This is in line with simplified language guidelines suggesting patient MPIs make use of less abstract nominalizations. Patient MPIs also have a higher number of pronouns, which can be accounted for by the fact that they address the reader with a second-person singular pronoun ("você"), while HC professional MPIs refer to patients with the noun "patient". The number of coordinating conjunctions is lower in patient MPIs, a finding also in line with guidelines. However, the number of subordinating conjunctions, which would be expected to be lower, is actually higher in patient MPIs. This and other findings need to be interpreted in light of the frequency of dependency relation tags.

Figure 2 shows the number of dependency relation tags annotated for each set of MPIs and their relative frequency.

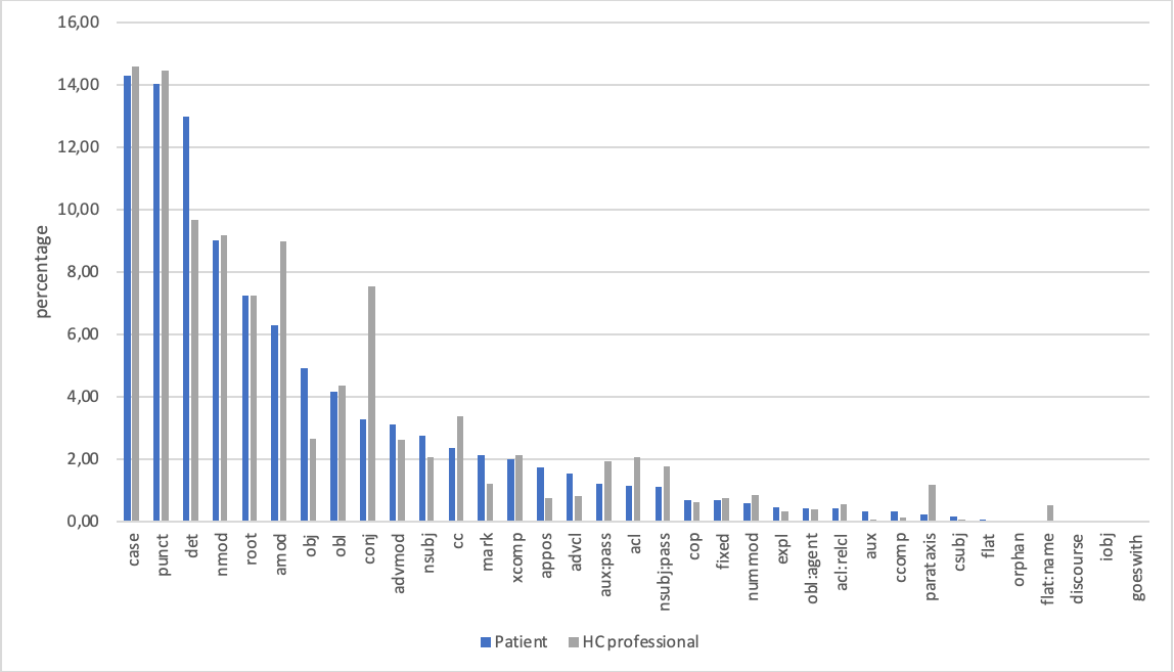


Figure 2. Relative frequency of deprel tags

Figure 2 shows differences for some of the dependency relations in the two sets of texts, many of which can be correlated to frequencies in Figure 1. HC professional MPIs have a higher number of conjuncts (conj) and coordinating conjunctions (cc), a finding that can be related to the higher frequency of the POS coordinating conjunction. Likewise, the higher number of adjectival modifiers in HC professional MPIs can be related to the frequency of the POS adjective. There is also a higher frequency of dependency relations pertaining to passive voice constructions (aux:pass; nsubj:pass) in HC professional PMIs. There is a higher frequency of adjectival clauses (acl) in HC professional MPIs as well. Also remarkable is the higher number of parataxis in HC professional MPIs, a relation established between main clauses and intersected or parenthetical explanations. Regarding patient MPIs, adverbial clauses (advcl) outnumber those in HC professional MPIs, which can account for the higher frequency of markers (mark), i.e., words marking a clause as being subordinate to another clause. This, in turn, may account for the higher number of the POS subordinating conjunction in patient MPIs as seen in Table 1. Adverbial clauses are typically used to construe if-then conditionals.

Sentence 3 in Figure 3[3] is an example of an annotated sentence retrieved from a patient MPI, which evidences the use of active voice and a second-person form of address, in this case, through the use of an imperative form ("consulte" [consult]) and a verb inflected for the form of address to the reader/patient ("tenha" [you have]). It also shows the recurrent use of adverbial clauses to construe a conditional meaning in patient MPIs.

[3] Images were obtained with the tikz-dependency package in a LaTeX editor.

(3) Não use este medicamento caso tenha asma ou úlcera no estômago.
[*Do not use this medication if you have asthma or a stomach ulcer.*]

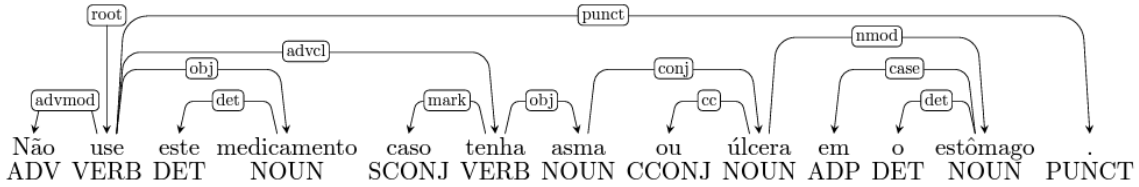


Figure 3. Sample annotated sentence from a patient MPI

Sentence 4 in Figure 4 is a counterpart sentence retrieved from a HC professional MPI.

(4) Este medicamento não deve ser utilizado por pacientes que tenham asma ou úlcera estomacal.
[*This medication should not be used in patients who have asthma or a stomach ulcer.*]

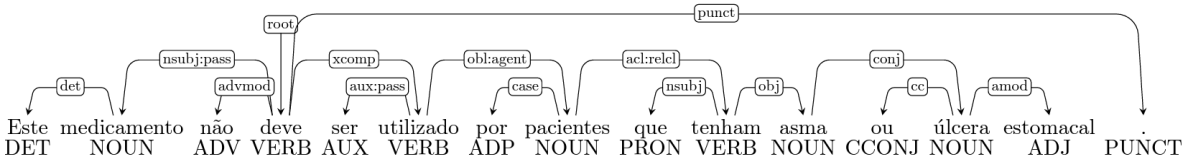


Figure 4. Sample annotated sentence from a HC professional MPI

Unlike (3), sentence 4 shows the use of a passive voice construction, a third-person form of address ("pacientes" [patients]), a defining relative clause ("que tenham asma..." [who have asthma]) and the use of an adjectival modifier ("estomacal" [stomach]).

A further distinction worth noting is the differential use of appositional modifiers in patient and HC professional MPIs. Besides being more frequent in patient MPIs, appositional modifiers (appos) are frequently used to provide synonyms intended to facilitate patient understanding of medical terms. This is illustrated by sentence 5 in Figure 5.

(5) Fissura na retina (rasgo na retina).

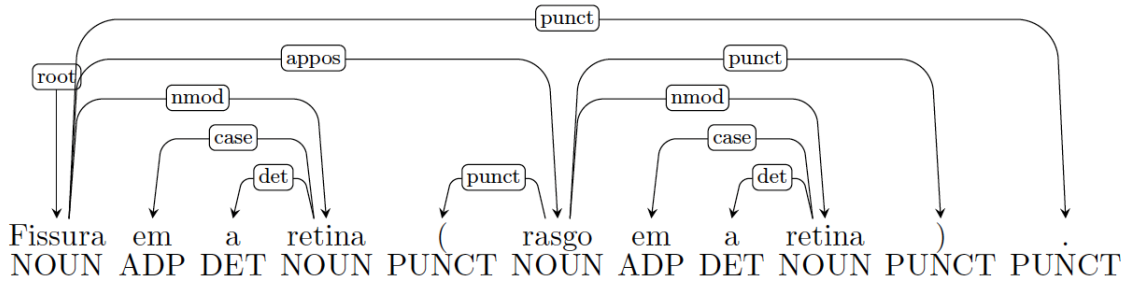


Figure 5. Appositional modifier in annotated sentence from patient MPI

In HC professional MPIs, appositional modifiers are mostly used for abbreviations and acronyms of diseases, as illustrated by example 6 in Figure 6.

(6) Tratamento da hipercalcemia induzida por tumor (HIT).

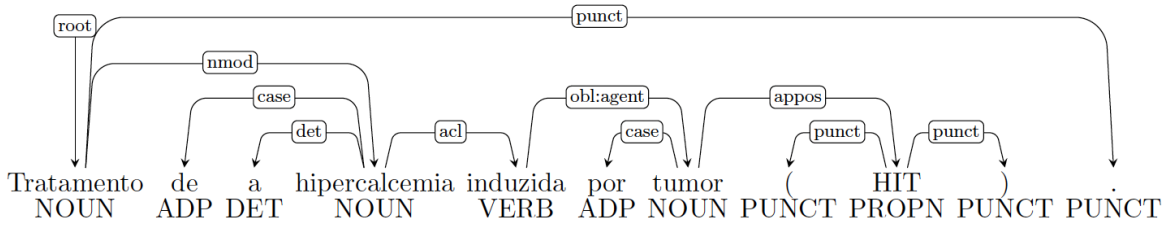


Figure 6. Appositional modifier in annotated sentence from HC professional MPI

5. Discussion

Results for patient MPIs, namely lower frequency of POS tags and dependency relations indicating coordinating constructions and low frequency of passive constructions, suggest that the texts are in compliance with the guidelines for simplified language (Aluísio et al., 2008a,b; 2010; ANVISA, 2009b). So is the use of appositional modifiers to provide synonyms for medical terms. The high frequency of adverbial clauses is accounted for by mostly conditional (if-then) clauses. Simplification guidelines do not advise to split this type of adverbial clause.

Results for HC professional MPIs are also in line with characteristics assumed to pertain to increased text complexity, as they evidence patterns clearly in contrast to those in patient MPIs.

6. Conclusion

This paper reported on a study exploring dependency syntax for the purposes of comparing texts targeting different readerships with different levels of literacy and

domain knowledge. MPIs were found to differ in their morphosyntactic patterns, which are in line with guidelines for simplified text in Brazilian Portuguese.

A corpus of 400 sentences in Brazilian Portuguese manually selected and aligned was compiled and a treebank of 400 sentences annotated for POS and dependency relations following the UD guidelines was developed. Both will be made available for public use.

The UD framework for morphosyntactic annotation proved adequate to retrieve text annotations that can be interpreted in terms of characteristics of simplified texts. Given its potential for comparability, the UD framework is expected to be useful, not only for monolingual aligned sentences, as is the case of our corpus, but also for multilingual sets. Corpora of aligned monolingual texts annotated for dependency relations are useful resources to gather insights for prospective text simplification tasks. In this sense, a further step in our project is to align our corpus with a corpus of MPIs addressing patients and HC professionals written and published in English.

Acknowledgements

The authors would like to thank two anonymous reviewers for their valuable comments. Adriana S. Pagano holds a research productivity grant awarded by Conselho Nacional de Desenvolvimento Científico e Tecnológico (Processo CNPq 313103/2021-6).

References

Aluísio, S.M., Specia, L., Pardo, T.A., Maziero, E.G., & Fortes, R.P. (2008a) Towards Brazilian Portuguese Automatic Text Simplification Systems. In: Proceedings of The Eight ACM Symposium on Document Engineering (DocEng 2008), pages 240-248, São Paulo, Brazil. https://doi.org/10.1145/1410140.1410191.

Aluísio, S.M., Specia, L., Pardo, T.A., Maziero, E.G., Caseli, H. & Fortes, R.P. (2008b) A Corpus Analysis of Simple Account Texts and the Proposal of Simplification Strategies: First Steps towards Text Simplification Systems In: Proceedings of The 26th ACM Symposium on Design of Communication (SIGDOC 2008), pages 15-22.

Aluísio, S.M., & Gasperin, C. (2010). Fostering Digital Inclusion and Accessibility: The PorSimples project for Simplification of Portuguese Texts. North American Chapter of the Association for Computational Linguistics.

Amorim, C. M. da S., Rocha, L. H. P. da, & Costa, M. J. (2015) A linguagem da bula: um estudo de estruturas linguísticas do gênero. Letrônica, 8(2), pages 467–479. https://doi.org/10.15448/1984-4301.2015.2.20401.

Angrosh, M., Nomoto, T., and Siddharthan, A. (2014) Lexico-syntactic text simplification and compression with typed dependencies. In Proceedings of the 25th International Conference on Computational Linguistics: Technical Papers (COLING 2014), Dublin, Ireland, pages 1996–2006..

ANVISA - Agência Nacional de Vigilância Sanitária. (2009a) Resolution RDC nº 47. Brasil: Agência Nacional de Vigilância Sanitária. Available at: http://www.anvisa.gov.br/medicamentos/bulas/rdc_47.pdf. Access on 29 June 2023.

ANVISA - Agência Nacional de Vigilância Sanitária. (2009b) Guia de Redação de Bula Gerência-geral de Medicamentos. GGMED. Brasília. Available at: https://www.gov.br/anvisa/pt-br/setorregulado/regularizacao/medicamentos/bulas-rotulos-e-nome-comercial/arquivos/copy8_of_GuiadeRedaodeBula.pdf. Access on 29 June 2023.

Duran, M. S. (2021) Manual de anotação de PoS tags. Relatório Técnico, n. 434. NILC-ICMC/USP. Available at: https://sites.google.com/icmc.usp.br/poetisa. Access on 29 June 2023.

Duran, M. S. (2022) 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). Relatório Técnico do ICMC 440. Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo. Available at: https://sites.google.com/icmc.usp.br/poetisa. Access on 25 June 2023.

Leal, S.; Duran, M.; Scarton, C.; Hartmann, N.; Aluísio, S. (2022) NILC-Metrix: assessing the complexity of written and spoken language in Brazilian Portuguese. CoRR abs/2201.03445. Available at: https://arxiv.org/abs/2201.03445. Access on 14 August 2023.

Pires, C., Vigário, M., & Cavaco, A. (2015) Readability of medicinal package leaflets: a systematic review. Revista De Saúde Pública, 49. Available at: https://doi.org/10.1590/S0034-8910.2015049005559. Access on 29 June 2023.

Pizzol, T. da S. D., Moraes, C. G., Arrais, P. S. D., Bertoldi, A. D., Ramos, L. R., Farias, M. R., Oliveira, M. A., Tavares, N. U. L., Luiza, V. L., & Mengue, S. S. (2019) Medicine package inserts from the users' perspective: are they read and understood?. Revista Brasileira De Epidemiologia, 22, e190009. https://doi.org/10.1590/1980-549720190009.

Siddharthan, A. (2011) Text Simplification using Typed Dependencies: A Comparison of the Robustness of Different Generation Strategies. In Proceedings of the 13th European Workshop on Natural Language Generation, pages 2–11, Nancy, France. Association for Computational Linguistics. Available at: https://aclanthology.org/W11-2802/. Access on 14 August 2023.

Um estudo das construções dar + para + V [infinitivo] nas Universal Dependencies

Marcella M. Lemos Couto[1] , Oto Araújo Vale[2]

[1] PPGL-UFSCar

[2]Departamento de Letras – UFSCar

{marcella.couto@estudante.ufscar.br, otovale@ufscar.br}

Abstract. *In Natural Language Processing (NLP), the corpus annotation task is the most basic and essential way to improve analyzes that benefit themself from automatic information extractions. This article discusses the challenge of adapting the Universal Dependencies (UD) guidelines to the brazilian Portuguese language. In order to recognize patterns within the potential phenomena and to promote consistency in the annotation process, the construction formed by* ***dar + para + V [infinitive]*** *is analyzed.*

Resumo. *Em Processamento de Linguagem Natural (PLN), a tarefa de anotação de corpus é a forma mais básica e essencial para aprimorar análises que se beneficiam de extrações de informações automáticas. Este artigo discute o desafio de adaptar as diretrizes da Universal Dependencies (UD) para o português brasileiro (PB). Com o objetivo de reconhecer padrões nos fenômenos encontrados e promover consistência nas anotações, analisa-se a construção* ***dar + para + V [infinitivo].***

1. Introdução

A anotação de corpus tem se tornado uma atividade essencial tanto nos recentes estudos linguísticos, quanto para os diversos recursos em Processamento de Língua Natural (PLN). Essa atividade possibilita a descoberta de padrões de uso e tem sido essencial para, por exemplo, o treinamento de algoritmos de aprendizado de máquina. Hovy & Lavid (2010) fazem um apanhado dos diversos níveis de anotação e chegam mesmo a falar de uma "ciência da anotação".

Nesse sentido, um ponto interessante das *Universal Dependencies* - UD - (NIVRE et al., 2016) é que essa abordagem já nasce como um projeto de anotação consistente para diversas línguas. Seu objetivo de facilitar o desenvolvimento de analisadores multilíngues facilita a pesquisa a partir de uma perspectiva de tipologia linguística. Esse modelo é o adotado no projeto *Porttinari-base* (PARDO et al., 2021) que tem por objetivo colaborar para o crescimento de recursos baseados em sintaxe e o desenvolvimento de ferramentas e aplicativos relacionados para o português brasileiro.

Dentro dessa representação, neste trabalho discute-se o desafio de anotação da construção formada pelo verbo **dar** seguido da perífrase verbal formada de **para + V [infinitivo].** A escolha desse tipo de construção se deu a partir de uma série de discussões no processo de anotação do corpus por representar desafios significativos na tarefa de anotação.

2. A problemática dos “argumentos”

Uma noção central na tradição teórica das gramáticas de dependências é a noção de valência que nasce a partir dos estudos de Tesnière (1959) e tem ganhado visibilidade por sua versatilidade. Nivre *et. al.* (2016) explica que, salvo algumas diferenças de um quadro teórico para outro, a valência é frequentemente relacionada à estrutura semântica predicado-argumento e, normalmente, atribuída mais frequentemente ao verbo. Nivre *et al.* (2016) assume que a valência de um verbo inclui apenas argumentos dependentes, mas algumas teorias também permitem que alguns não argumentos obrigatórios sejam incluídos. Além das funções gramaticais tradicionais, (tais como predicado, sujeito e objeto), os papéis semânticos (como, por exemplo, agente, paciente e finalidade) são comumente usados, especialmente nas representações da sintaxe profunda e da semântica.

Nesta seara, o modelo da Gramática de Construções (GC) propõe o pareamento entre forma e significado (FILLMORE, 1988; GOLDBERG, 1995; CROFT, 2001) não existindo uma rigidez entre léxico e gramática. Em outras palavras, a GC defende que a linguagem não é apenas uma lista de palavras com significados fixos e regras gramaticais rígidas para combiná-las. Em vez disso, ela sugere que as construções são unidades fundamentais da linguagem, que consistem em padrões formais (sintáticos) e significados específicos que se associam de maneira flexível. Isso significa que a forma como as palavras são organizadas em uma frase influencia diretamente o significado que é comunicado. Sob essa perspectiva, construções são dotadas de características semânticas não sendo previsíveis a partir de suas partes componentes. De acordo com Goldberg (1995), as construções devem ser analisadas levando em conta tanto as generalizações mais amplas, como também os padrões mais limitados, chegando a considerar, por exemplo, morfemas como construções. Isso quer dizer que, à luz da GC, a análise linguística não pode prescindir do pareamento entre as propriedades formais e as funcionais das construções. Goldberg (1995) coloca em seu foco as construções de estrutura argumental, que são sentenças compostas por um verbo e seus argumentos. Em sua análise, há distinção entre papéis argumentais de papéis participantes. Os papéis argumentais são associados ao verbo e os papéis participantes são previstos pelo constituinte sub-oracional ao qual são associados, e não necessariamente pelo verbo.

Em última análise, pode-se estabelecer que estamos diante de um mesmo desafio para as teorias linguísticas: distinguir o que é dispensável daquilo que é essencial.

3. A complexidade da questão

O verbo **dar** é muito produtivo em português brasileiro e isso reflete uma complexidade maior na análise e classificação de suas construções. Por exemplo, no Dicionário de verbos do português brasileiro, de Francisco Borba (1996), o verbo **dar** indica ação-processo e aparece em muitas entradas distintas. Com o complemento **para + V [infinitivo]**, pode significar (i) doar algo a alguém para alguma finalidade; (ii) ser possível; (iii) ter jeito, vocação ou inclinação; ou (iv) ser suficiente (BORBA, 1996). Coelho e Silva (2014), em um estudo sobre o processo de gramaticalização do verbo **dar** sob uma abordagem de interface entre semântica cognitiva e variação linguística, demonstram como este passou de predicador a auxiliar ao longo dos séculos. Ao perder propriedade lexical e se juntar a uma preposição seguida de outras formas verbais flexionadas no infinitivo, esse verbo forma uma nova construção verbal e passa a

desempenhar, nesse contexto, funções gramaticais relacionadas à expressão da **modalidade** e do **aspecto**. O que interessa é identificar o processo em que o elemento linguístico saiu do nível da criatividade eventual da língua em uso para penetrar nas restrições da gramática, tornando-se mais regular e mais previsível. O verbo **dar** seguido da perífrase verbal em estudo resulta em novos significados que extrapolam o significado do verbo quando ocorre em sua forma plena, por exemplo[1]:

(1) **Dá para fazer** coisas legais aqui.

(2) Todas as verbas somadas **só dariam para construir** nove metros de linha.

(3) E o gigante, que adora dormir, **deu para acordar** à noite, atormentado pelo pesadelo de não poder voltar a dormir tão cedo!

Em (1), o verbo **dar** exprime uma *possibilidade*, a de **se fazer coisas legais aqui**. Já em (2), a construção contém o sujeito **todas as verbas somadas**. Nesse caso, não se trata mais de uma *possibilidade*, mas sim de *ser suficiente*. Em (3), vemos algo que *não acontecia* e *passou a acontecer* ao sujeito de **dar**, marcando assim o *aspecto inceptivo*, isto é, aquele que assinala o início de uma ação. No português europeu, existe a mesma construção, mas a preposição é **em**: **dar + em + V [infinitivo]** (BAPTISTA & MAMEDE, 2020).

Exceto em (1), em que o verbo está na forma impessoal, até aqui, os elementos que antecedem o verbo correspondem aos sujeitos das orações. Entretanto, nem sempre que há um SN antecedendo a construção **dar + para + V [infinitivo]** este será o sujeito da oração, como em:

(4) O torcedor **dá para entender**, mas jornalistas que cobrem futebol torcerem igual é incompreensível.

Em (4), mesmo que haja concordância, o elemento que antecede o verbo principal, **o torcedor**, não é o sujeito, mas sim, o objeto do verbo no infinitivo **entender.** Dito de outra maneira: **dá para entender o torcedor**. Isso fica comprovado em (5), sentença em que não há concordância verbal com o elemento que antecede o verbo:

(05) **Todas essas doenças dá** para curar se a pessoa tratar logo.

Quando o verbo **dar** é empregado em sua construção como verbo pleno (ou distribucional), não há ambiguidade quanto à interpretação do complemento preposicionado seguido de infinitivo, como em:

[1] Todos os exemplos utilizados no artigo foram retirados do corpus Brasileiro (CBRAS), disponível em: https://www.linguateca.pt/acesso/corpus.php?corpus=CBRAS

(6) Ele **nos dará** dinheiro **para manter** a reeleição de governadores e de prefeitos.

Note-se que em (6a) que é possível substituir a preposição **para** pela locução **com a finalidade de**, comprovando assim se tratar de uma oração final.

(6a) Ele nos dará dinheiro **com a finalidade de** manter a reeleição de governadores e de prefeitos.

Outra particularidade do verbo **dar** é possibilidade de construir construções como verbo-suporte. "Particularmente nas construções com verbo-suporte, o predicador central da frase é um nome, enquanto o verbo serve apenas como suporte desse predicador." (RASSI, 2015. p. 18) Acrescenta-se essa informação a título de interesse, já que o escopo deste trabalho é discutir a estrutura **dar + para + V [infinitivo]** como complementador desse verbo. Mais uma confirmação da produtividade desse verbo.

4. O complemento "para + infinitivo"

A perífrase formada de **para + V [infinitivo]**, geralmente, constrói orações adverbiais finais. Neves (2011) aponta para a grande produtividade desse complemento. Explica também que o contexto semântico prototípico é o de que a oração principal tenha um sujeito capaz de exercer controle na final. O que a autora descreve sobre as adverbiais circunstanciais são as que se ligam ao conteúdo preposicionado da oração principal. Também, as orações podem ser clivadas, focalizadas, substituídas por construções nominais do tipo com a finalidade de e podem ser objetos de interrogação.

Borba (1996. p. XVIII) diz que esse "complemento faz parte da estrutura interna de um SN, que se desdobra, portanto, em V + complemento". Por ser indispensável, o complemento faz parte da valência do verbo. O complemento de natureza adverbial com valor semântico de finalidade é expresso pela forma: **para + nome + oração infinitiva/oração conjuncional final.** Para o autor, a subcláusula de finalidade é considerada um complemento essencial.

5. Os desafios e caminhos na atribuição da *deprel*

Qualquer projeto, antes de realizar a anotação segundo as diretrizes da UD, deve realizar as adaptações de acordo com as especificidades de cada língua. Para o PB, já existem duas versões do manual[2] de anotação de relações de dependência do projeto.

Em UD, são previstos dois níveis de anotação: o primeiro, no nível morfológico, contém sete etiquetas morfossintáticas (*PoS tags*); o segundo, no nível sintático, apresenta 37 relações de dependência (`deprel`). A representação da estrutura de dependências é arbórea e uma palavra da sentença é a raiz (`root`) da representação. Nesse segundo nível, a anotação das `deprel` se dá de maneira binária e assimétrica. A representação básica de uma estrutura de dependências é estabelecida entre parte de uma unidade

[2] Versão mais recente disponível em:

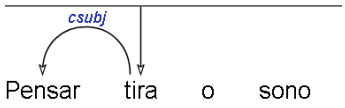
https://drive.google.com/file/d/1ile8Wfxu1qdrZOmLGqkvVuQ4fXvHgVMo/view

lexical que encabeça a relação e outra unidade lexical sintaticamente dependente dela. Os dependentes que não correspondem a argumentos tendem a ser opcionais e podem ocorrer mais de uma vez em um mesmo predicado. Já os dependentes que correspondem a argumentos ocorrem apenas uma vez em cada predicado. (NIVRE et al., 2016)

A UD separa os argumentos dos predicados *em* `core arguments` e `non-core dependents`. Quando o dependente está na forma oracional, como é o caso da perífrase em estudo, o elemento dependente que recebe a seta é o verbo. São três as etiquetas *de* `core arguments` que são usadas para vincular uma palavra principal ao verbo de uma oração dependente (`deprel`: ***csubj, ccomp, xcomp***) e uma em que o predicado é o `head` e o dependente é considerado seu modificador (`deprel`: ***advcl***).

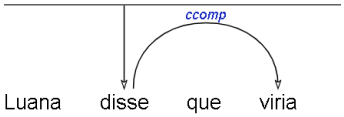
A `deprel` ***csubj*** **(**sujeito oracional) é utilizada para anotar o sujeito constituído de uma oração, como o exemplo abaixo em que a `deprel` ***csubj*** une **tirar**, `head` relação, a **pensar**, dependente da relação.

(7) Pensar tira o sono.



A `deprel` ***ccomp*** (complemento oracional fechado) complementa o sentido do predicado, é um argumento core e é um complemento oracional fechado. Isso significa que esse complemento não pode ter um sujeito controlado, nem pelo sujeito, nem pelo objeto da oração principal. É muito comum e fácil de identificar a relação ***ccomp*** quando há uma oração subordinada completando o sentido do verbo, introduzida por uma conjunção subordinativa. Em paralelo com a gramática tradicional (GT), essa `deprel`, geralmente, contempla o que seria considerada uma oração subordinada substantiva objetiva direta ou objetiva indireta, como em:

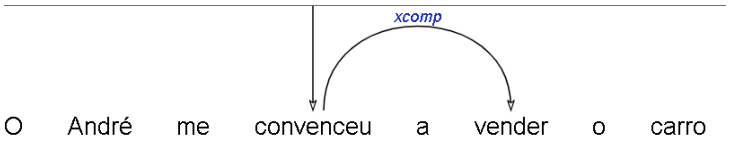
(8) Luana disse que viria.



A última `deprel` que também liga um `head` a um argumento core como as anteriores é a ***xcomp*** (complemento oracional aberto). Muitas vezes, ela pode ser confundida com ***ccomp***, com exceção quanto ao detalhe da diretriz que trata sobre seu sujeito sintático. Para atribuição desta `deprel`, o sujeito do verbo dependente deve ser o mesmo sujeito da oração principal (9) ou o mesmo que o objeto da oração principal (10).

(9) Eu pretendo trabalhar na virada do ano.

(10) O André me convenceu a vender o carro.



A diferença então entre as relações de dependência ***ccomp*** e ***xcomp*** é a de que ***xcomp*** não admite um sujeito explícito na oração subordinada. A `deprel` ***xcomp*** ocorre, por exemplo, com verbos que modalizam, como: dever, poder, precisar. Alguns verbos unem-se a outros para modalizar os enunciados. Esses verbos indicam, principalmente, *modalidade epistêmica* (conhecimento) e *deôntica* **(**dever). Essas modalidades são subdivididas em *necessidade epistêmica* (deve) e *possibilidade epistêmica* (pode); e *necessidade deôntica* (obrigatoriedade) e *possibilidade deôntica* (permissão) (NEVES, 2011).

Retomaremos os exemplos com suas análises daqui em diante. Em (1), o valor de modalidade de **dar para** é o *de possibilidade*. Entretanto, em relação à forma, o verbo está no impessoal.

Para ocorrências tipo em (1), Gorski (2020) considera "para fazer coisas legais" como uma oração subjetiva.

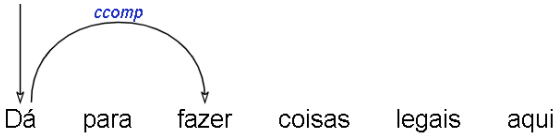
> O fato de INF sujeito ser introduzido por preposição é antigo: "A construção de preposições com o infinitivo tornou-se tão familiar, que, em português, e em outras lingoas romanicas [...], chegam a antepôr-se a infinitivos que exercitam as funções de sujeito [...]" (DIAS, 1970 [1918], p. 217-219 apud GORSKI, 2020, p. 4348)

Embora Gorski (2020) tenha um argumento forte para sua análise, incluindo um teste de paráfrase como uma oração subjetiva: "**É possível fazer** coisas legais aqui", não consideramos a melhor anotação a relação de ***csubj***. Em primeiro lugar, o uso de uma paráfrase é apenas um dispositivo semântico-argumentativo e não pode ser transposto diretamente para a análise sintática. Assim, o fato de o valor semântico da construção **dar + para + V [infinitivo]** ser semelhante ao de *ser possível*, não existe a possibilidade de transposição da construção sintática do adjetivo **possível** (que, na verdade, leva uma oração-sujeito) à construção sintática em análise.

Não se pode perder de vista que a UD é, essencialmente, uma representação de dependência sintática. Dito isso, existe uma diretriz no manual para anotação de dependentes de ***ccomp*** com sujeitos inexistentes. Soma-se a isso o fato de ***xcomp*** não

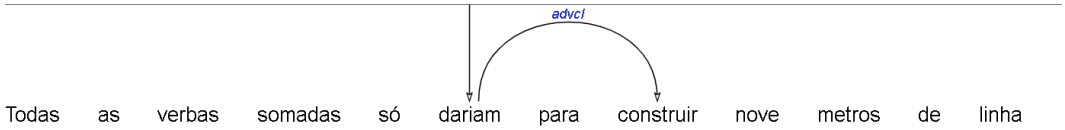
admitir sujeito na oração subordinada. Para (1) é possível existir um sujeito do verbo **fazer**.

(1) **Dá para fazer** coisas legais aqui.



Já (2) expressa algo como *ser suficiente* como também atesta Borba (1996). A relação entre o **dar** e o complemento oracional é a ***advcl***, porque o teste de oração final tem bom resultado.

(2) Todas as verbas somadas só **dariam para construir** nove metros de linha.



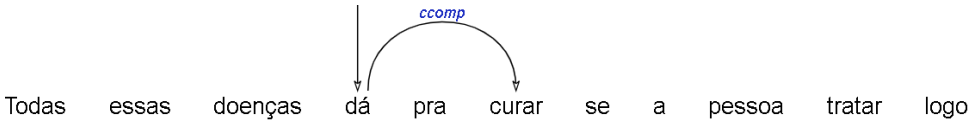
Perceba que o teste da oração final **com a finalidade de** não funciona em (1a) mas funciona na paráfrase em (2a):

(1a) * **Com a finalidade de** fazer coisas legais aqui é possível/dá.

(2a) **Com a finalidade de** construir nove metros de linha, todas as verbas somadas seriam suficientes/dariam.

O exemplo (3) é semelhante ao (1) tendo seu valor semântico girando no eixo da *possibilidade.* Porém, note que não se pode confundir a sequência **Todas essas doenças** com o sujeito da oração. Essa sequência é o objeto de **curar** e está topicalizada.

(3) Todas essas doenças dá pra curar se a pessoa tratar logo.



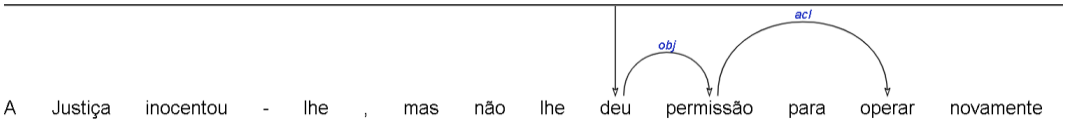
Sobre a noção de *aspecto*, ela aproxima-se da noção de tempo verbal. Comrie (1985) atesta que o tempo linguístico determina o tempo da enunciação. Aspecto, segundo Castilho (1984, p. 14), "é a visão objetiva da relação entre o processo e o estado expressos pelo verbo e a ideia de duração ou desenvolvimento. É, pois, a representação espacial do processo".

(4) E o gigante, que adora dormir, **deu para acordar** à noite.

Não será feita aqui uma exposição do tratamento do assunto nas variadas correntes visto que a intenção é buscar subsídios linguísticos que permitam a realização de anotação consistente. Para fins de discussão do fenômeno, emprega-se a nomenclatura de Neves (2011). Os verbos aspectuais formam perífrases ou locuções indicando: *aspecto inceptivo* (início do evento); *aspecto cursivo* (desenvolvimento do evento); *aspecto habitual* (evento habitual); *aspecto progressivo* (progressão); *aspecto terminativo* ou *cessativo* (término do evento ou cessação); *aspecto resultativo* (resultado do evento); *aspecto frequentativo* (ideia de frequência); sem ideia de frequência, também podem indicar *consecução, intensificação e aquisição do estado*. Por exemplo, na sentença (4), **deu para acordar** indica que **gigante** começou algo e que essa atividade não tem conclusão. **Gigante**, neste caso, é sujeito de **dar** e de **beber**. Por analogia, neste uso aspectual do **dar para + V [infinitivo]**, a `deprel` utilizada entre o **dar** e o verbo no infinitivo é a ***`xcomp`***.

Em construções do verbo **dar** atuando como verbo-suporte, como quem predica não é o verbo e sim um nominal, a relação de dependência parte desse nominal para o verbo. Para esse e outros casos de construções com verbo-suporte, a `deprel` utilizada é ***`acl`***. Essa relação ocorre entre uma palavra de conteúdo não verbal e uma oração que a modifica.

(11) "A Justiça inocentou-lhe, mas não lhe **deu permissão para operar** novamente", diz. (dar permissão = permitir).



7. Conclusão

Como exposto, este estudo propôs-se a apontar caminhos para anotação das construções formadas do verbo **dar** seguidas da perífrase verbal **para + V [infinitivo]** segundo o modelo UD. As propostas de anotação que foram apresentadas levaram em conta a necessidade de consistência na análise e no julgamento dos variados fenômenos linguísticos. O trabalho também contribui com projetos de PLN que utilizem anotações baseadas em relações de dependência, podendo assim, abreviar esforços em anotações sintáticas que seguem o modelo UD. Entende-se que há limitação em relação à análise dos dados, tanto pela grande variedade do fenômeno quanto pela finalidade deste estudo. Relembra-se aqui da diferença entre estudos linguísticos voltados à construção de teorias linguísticas dos estudos para fins de PLN. Espera-se, portanto, que o debate

realizado auxilie nas tarefas de anotação e ofereça informações e dados para pesquisas posteriores. Conforme demonstrado, esse fenômeno é altamente produtivo, por isso, pretende-se avançar os estudos em trabalhos futuros.

Agradecimentos

Registre-se aqui um especial agradecimento a Magali Duran e Maria das Graças Volpe Nunes, que contribuíram com longas discussões a respeito da anotação que deu origem a este trabalho.

Os autores agradecem também às/aos pareceristas anônimos que fizeram importantes sugestões para o aperfeiçoamento deste artigo.

Este trabalho foi realizado no âmbito do Centro de Inteligência Artificial da Universidade de São Paulo (C4AI - http://c4ai.inova.usp.br/), com o apoio da Fundação de Amparo à Pesquisa do Estado de São Paulo (processo FAPESP #2019/07665-4) e da IBM. Este projeto também foi apoiado pelo Ministério da Ciência, Tecnologia e Inovações, com recursos da Lei N. 8.248, de 23 de outubro de 1991, no âmbito do PPI-Softex, coordenado pela Softex e publicado como Residência em TIC 13, DOU 01245.010222/2022-44

Referências

Baptista, J.; Mamede, N. (2020). Syntactic Transformations in Rule-Based Parsing of Support Verb Constructions: Examples from European Portuguese. In *9th Symposium on Languages, Applications and Technologies (SLATE 2020)*. Schloss Dagstuhl-Leibniz-Zentrum für Informatik.

Bechara, Evanildo. *Moderna gramática portuguesa*. Nova Fronteira, 2019.

Borba, Francisco S. *Uma gramática de valências para o português*. São Paulo: Ática, 1996.

Castilho, Ataliba T. *Ainda o aspecto verbal*. Estudos portugueses e africanos, v. 4, 1984.

Coelho, Sueli Maria; De Paula Silva, Silmara Eliza. *Um Estudo Da Variação Linguística No Liame Preposicional Em Construções [Vdar+ Preposição+ Vinfinitivo] No Português Do Brasil*. Revista Diadorim, V. 21, N. 2, P. 125-144.

Comrie, Bernard. *Aspect: An introduction to the study of verbal aspect and related problems.* Cambridge university press, 1976.

Comrie, Bernard. *Tense*. Cambridge university press, 1985.

Croft, William. *Radical construction grammar: Syntactic theory in typological perspective.* Oxford University Press on Demand, 2001.

Duran, Magali Sanches. Manual de anotação de PoS tags. Relatório Técnico, n. 434, 2021.

Duran, Magali Sanches Nunes, Maria das Graças Volpe; Lopes, Lucelene; Pardo, Thiago Alexandre Salgueiro. Manual de anotação como recurso de Processamento de Linguagem Natural: o modelo Universal Dependencies em língua portuguesa. Domínios de Lingu@ gem, v. 16, n. 4, p. 1608-1643, 2022.

Fillmore, Charles J.The *Mechanisms of "Construction Grammar".* Proceedings of the Fourteenth Annual Meeting of the Berkeley Linguistics Society (1988), pp. 35-55. 1988.

Goldberg, Adele E. *Construction grammar: a construction grammar approach to argument structure.* University of Chicago Press, 1995.

Gorski, Edair. *Níveis de integração de cláusulas para + Infinitivo*. In: SEMINÁRIO DO GEL, 17, 1999, Bauru, SP. Estudos Lingüísticos XXIX. Assis/SP: Unesp, 2000. p. 88-102

Gorsk, Edair. *A (não) realização do sujeito e a integração de orações*. Scripta, v. 5, n. 9, p. 161-173, 2001.

Görski, Edair. (2020). Emergência de dar pra/de no domínio funcional da auxiliarização modal deôntica. *Fórum Linguístico*, *17*(1), 4342-4356.

Hovy, Eduard; Lavid, Julia. *Towards a 'science'of corpus annotation: a new methodological challenge for corpus linguistics.* International journal of translation, v. 22, n. 1, p. 13-36, 2010.

Neves, Maria Helena de Moura. *Gramática de usos do português*. São Paulo: Ed. Unesp, 2011.

Nivre, J.; De Marneffe, M.-C.; Ginter, F.; Goldberg, Y.; Hajič, J.; Manning, C. D.; Mcdonal, R.; Petrov, S.; Pyysalo, S.; Silveira, N.; Tsarfaty, R.; Zeman, D.Universal dependencies v1: *A multilingual treebank collection.* In: Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16). 2016. p. 1659-1666.

Nivre, J.; De Marneffe, M. C.; Ginter, F.; Hajič, J.; Manning, C. D.; Pyysalo, S.; … & Zeman, D. . *Universal Dependencies v2: An evergrowing multilingual treebank collection*. arXiv preprint arXiv:2004.10643, 2020.

Pardo, T. A. S.; Duran, S. D.; Lopes, L. Di Felippo, A.; Roman, N. R.; Nunes, M. G. C.. *Porttinari-a Large Multi-genre Treebank for Brazilian Portuguese.* In: Anais do XIII Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana, SBC, 2021. p. 1-10.

Rassi, Amanda Pontes. *Descrição, classificação e processamento automático das construções com o verbo dar em Português Brasileiro*. 2015. Tese (Doutorado em Linguística) – Universidade Federal de São Carlos, São Carlos, 2015. Disponível em: https://repositorio.ufscar.br/handle/ufscar/8278.

Silva, Silmara Eliza de Paula. *A construção verbal v1dar+preposição + v2infinitivo* [manuscrito]: um estudo na interface Sociolinguística e Gramaticalização /Silmara Eliza de Paula Silva. – 2018

Tesnière, Lucien. *Eléments de syntaxe structurale*. Paris, Klincksieck, 1959.

Insights into the UD Tagset: Unveiling its Intricacies

Magali Sanches Duran

[1]Núcleo Interinstitucional de Linguística Computacional (NILC)
Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo (USP)
magali.duran@uol.com.br

Abstract. *This opinion paper explores our inclination to draw on principles of syntactic analysis established in the grammars of our native language when using the Universal Dependencies tagset to assign dependency relation tags. Taking the Portuguese language as a case example, this study argues that a fine-grained comparison of concepts and terms used in traditional grammars of Brazilian Portuguese and those used by Universal Dependencies reveals gaps which lead to different interpretations and, ultimately, to a deviation from the envisaged universality of the dependency relations tagset.*

Resumo. *Este artigo de opinião discute a tendência que temos de projetar, no conjunto de etiquetas de relações de dependência universais, conceitos e termos da análise sintática de gramáticas de nossa própria língua. Tomando a língua portuguesa como exemplo, este estudo argumenta que uma comparação pormenorizada entre os conceitos e termos usados nas gramáticas tradicionais do português do Brasil e os usados pelas Dependências Universais revela lacunas que levam a diferentes interpretações e, em última análise, a um desvio da universalidade prevista para o conjunto de etiquetas de relações de dependência.*

1. Introduction

Using Universal Dependencies (UD) (NIVRE et al. 2020; de MARNEFFE et al, 2021) to annotate a corpus in the Portuguese language and following the discussions that have taken place in the UD Issues on GitHub over the past three years, I have come to realize that there are still some unresolved issues with the assignment of UD dependency relation tags.

For the UD user community, maintaining tagsets unchanged is of paramount importance, as many corpora have already been annotated using them. Every change in a tagset or in the guidelines on how to apply its tags requires rework or, in a worst-case scenario, renders existing corpora non-compliant with the new guidelines. In May 2022, for example, there was a change in the way reported speech is annotated in UD. Although the tagset remained the same, the treebanks had to be adjusted to comply with the new guidelines.

It is also necessary to reach a consensus regarding which phenomena should be annotated with each tag. This is where many discussions start since some phenomena

are more universal than others and the point of view of those who discuss the assignment of UD tags is always influenced by the languages they annotate. Although English is widely used as a lingua franca, it can be challenging to describe phenomena that do not exist in the English language to an English-speaking community.

By applying the UD scheme to Portuguese (PT) corpus annotation, I was able to detect occurrences where we had to make decisions because they were not covered by any tag in the UD tagset. Sharing the process of understanding Universal Dependencies (UD) concepts and identifying gaps in the annotation of certain phenomena may be of interest not only to PT annotators, but also to those who currently use or plan to use the UD annotation scheme in other languages as well.

2. The brief, naive illusion that concepts are familiar

PT annotators may feel comfortable when they first come into contact with the UD tagset because many of the relation tags appear to refer to syntactic functions they are already familiar with. However, not everything is what it seems to be. For example, anyone who sees the tags **nsubj**, **obj** and **iobj** quickly associates them with "subject", "direct object" and "indirect object", which are familiar terms in PT grammars. Our background shapes what we see.

However, this initial mapping of traditional PT syntactic functions onto UD relations is quite misleading. Although **iobj** is the short form for "indirect object" according to the UD guidelines, it is a case of "false friend" in that it does not correspond to the concept of an indirect object in PT (that is, an object introduced by a preposition). In fact, **iobj** is a tag dedicated to a third "core"[1] element (the other two being subject - **nsubj**, and object - **obj**), almost always non-prepositional, which occurs close to the verb and can occupy the subject position in a passive voice alternation. For instance, "Mary" in the English dative construction "John gave **Mary** a book" can be rendered as "Mary was given a book" in one of the two possible passive voice constructions in English. If "Mary" is prepositioned, as in "A book was given to Mary by John", "Mary" is no longer an **iobj**.

UD's decision to annotate **iobj** is supported by comparative studies on "core elements" across languages of various origins (THOMPSON, 1997; ANDREWS, 2007), and it is logical in an approach that aims for universal use. The only problem is the misleading retention of the adjective "indirect" in the tag.

In Portuguese, it seems that there are no cases of a third core element like in English dative constructions. Like other Romance languages, to the best of my knowledge, PT exclusively uses the dependency relation **iobj** to annotate dative pronouns[2] as they are not introduced by prepositions. Although these pronouns cannot assume the subject position in the passive voice, they occur near the verb and meet the criterion of givenness, a key issue in determining core elements.

---

[1] We found no criteria for distinguishing core arguments from other dependents in Portuguese, which is why we used the criteria adopted by the English language.

[2] In Portuguese: *me, te, lhe, se, nos, vos, lhes* (dative pronouns in English are: me, you, him, her, us, them)

The syntactic function known as “indirect object” in PT grammars is annotated as **obl** (oblique) in UD. The relation **obl** is used for both argumental and adjunct modifiers in the form of prepositional noun phrases (PP). This position of UD Guidelines is supported by psycholinguistic studies indicating that the boundary between argumental PP and adjunct PP is not well defined (see Boland & Blodgett (2006) to revisit some of them).

In the first version of UD Guidelines, there was no **obl** relation: all PP modifiers were annotated as **nmod** (nominal modifier). As of version 2, **nmod** is used exclusively for nominal modifiers of NOUN, PROPN and PRON, and the newly created **obl** is now applied to PP modifiers of VERB, ADJ and ADV, both argument and adjunct (UD guidelines do not make a distinction between arguments and adjuncts).

The problem is that, although having the same lexical realization, adjuncts modifying nominals, adjectives, adverbs, and verbs, are now annotated with different dependency relations depending on the part-of-speech (PoS) tag of the head of the dependency relation. For example: in “Rainfall in March is a problem”, “in March” is annotated as **nmod**, because its head “rainfall” is a NOUN, whereas in “It starts to rain in March”, “in March” is annotated as **obl**, because its head “to rain” is a VERB. This difference in classification comes naturally to PT annotators since PT grammars dictate that the former would be classified as “adnominal adjunct” and the latter as “adverbial adjunct”. However, when it comes to arguments, this division of relations into **nmod** and **obl** separates into two groups cases annotated as "*complemento nominal*" (noun complement) in PT. These are PP related to argument-taking nouns, adjectives and adverbs, such as “lack **of confidence**” (**nmod**), “eager **for change**” (**obl**), and “regardless **of nationality**” (**obl**), respectively.

As a result, the relation **obl** corresponds to three syntactic functions in traditional grammars of PT: “nominal complement” of adverbs and adjectives; “indirect object”, and “adverbial adjunct” in PP form. Examples of **obl** (in bold) are:

- *ansioso* ***por novidades*** **[**avid **for news]: obl** modifying an ADJ, corresponding to a “nominal complement” in PT;
- *independentemente* ***da hora*** [regardless **of the time] obl** modifying an ADV, corresponding to a “nominal complement” in PT;
- *reclamar* ***do barulho*** [to complain **about the noise] obl** modifying a VERB, corresponding to an “indirect object” in PT;
- *dormir de noite* [to sleep **at night**] **obl** modifying a VERB, corresponding to an “adverbial adjunct” in PT.

The relation **nmod**, on the other hand, is used to annotate any nominal modifier of a nominal (NOUN, PROPN, PRON), which corresponds to what PT traditional grammars refer to as “*adjuntos adnominais*” (“adnominal adjuncts”) and “nominal complements”. Following this rationale, **nmod** is also used to annotate what is known as a “*aposto especificativo*” (“specifying appositive”) in PT. Examples of **nmod** (in bold):

- *gosto* ***de chocolate*** *na boca* (a taste **of chocolate** in one's mouth) **nmod** corresponding to a “nominal complement” in PT;

- *gosto de chocolate* ***na boca***" (a taste of chocolate **in one's mouth**) **nmod** corresponding to an "adnominal adjunct" in PT;
- o *presidente* ***Lula*** (President **Lula) nmod** corresponding to a "specifying appositive" in PT.

While **obl** and **nmod** are broad relations, **appos** (appositional modifier) has a more restricted usage. The tag is exclusively used for relations that satisfy the following restrictions: occuring after the nominal they modify, having the same referent as the nominal they modify, and being interchangeable with the modified nominal. Examples of **appos** that meet these restrictions (in bold) are:

- *Pelé,* ***o rei do futebol****, morreu no último ano.* (Pelé, the **king** of soccer, died last year.) **appos** corresponding to an appositive in PT;
- *O rei do futebol,* ***Pelé****, morreu no último ano.* (The king of soccer, **Pelé**, died last year.) **appos** corresponding to an appositive in PT;

The modifiers family has three other members: **amod** (adjectival modifier), simple adjectives that modify nominals, as in "an **incredible** landscape"; **nummod** (numeric modifier), numbers that modify a noun indicating a quantity, as in "**three** years"; and **advmod** (adverbial modifier), simple adverbs that modify verbs, adjectives, adverbs and, to a lesser extent, even nouns, as shown in the following examples (**advmod** in bold):

- *falar* ***alto*** (to speak **loudly**) **advmod** modifying a VERB;
- ***extremamente*** *cansado* **(extremely** tired) **advmod** modifying an ADJ;
- ***somente*** *agora* **(only** now) **advmod** modifying an ADV;
- ***só*** *amigos* **(just** friends) **advmod** modifying a NOUN.

3. The search for symmetry

As can be seen, the PoS (Part-of-Speech) tag of the dependent (and sometimes the PoS tag of the head) is used as a criterion for assigning dependency relations. This approach is effective for phrasal dependents, but not for clausal dependents, obviously, because predicates, except for nominal predicates, always implicate a VERB.

Notwithstanding, when we see that a clausal subject is **csubj**, we promptly think: **csubj** is for **nsubj** just as other clause types are for other relations (**obj, iobj, obl, nmod, amod, advmod, appos**), an association similar to that of subordinate clauses in PT (subject subordinate clause, direct object subordinate clause, etc.).

This natural quest for mappings is an individual exercise, since, except for **nsubj/csubj**, the UD guidelines do not associate simple relations (**obj**, **iobj**, **nmod**, **obl**, **advmod**, **amod**, **appos**) with clausal ones (**acl**, **advcl**, **ccomp**, **xcomp**) on a one-to-one basis.

The initial shift that unveils non-symmetric mappings is the existence of two relations for clausal objects: **ccomp** and **xcomp**. The criterion for distinguishing them is related to the subject of the dependent clause. If the subject or object of the parent

clause controls a null subject in the dependent clause, it is an **xcomp**[3]. Otherwise, it is a **ccomp**. In PT, the predicate of a **ccomp** dependent always implicates a finite[4] form and is introduced by a subordinating conjunction or a wh- adverb. The predicate of an **xcomp** always implicates an infinitive form, and is not introduced by subordinating conjunctions. It is relevant to say that the finite form of **ccomp** dependents and the non-finite form of **xcomp** dependents may be "assumed" by the main verb, by an auxiliary or by a copula verb (for nominal predicates), as underlined in the following examples (**ccomp** and **xcomp** dependents in bold)[5]:

- *Ele confirmou que* ***viria***. (He confirmed that he would **come**.) **ccomp**
- *Ele confirmou que havia* ***bebido***. (He confirmed that he had **drunk**.) **ccomp**
- *Ele nos disse que seria o* ***palestrante*** *convidado*. (He told us he would be the invited **speaker**.) **ccomp**
- *Ele quer* ***vir***. (He wants to **come**.) **xcomp**
- *Ele queria ter* ***vindo***. (He wished he had **come**.) **xcomp**
- *Ele pretende ser* ***professor***. (He intends to be a **teacher**.) **xcomp**

However, we encounter a problem when we use **xcomp** in PT: there are many clauses that have all the characteristics of a **xcomp,** but are introduced by a preposition, where in English an infinitive marker is used ("to"[6]):

- *Ele começou* ***a*** *andar.* (He started **to** walk.)
- *Ele esqueceu* ***de*** *fazer isso.* (He forgot **to** do this.)

Since the UD guidelines draw on English-based models and most examples are provided in English, the problem of an **xcomp** introduced by a preposition rarely arises because the infinitive marker "to" is the most usual form. However, when a preposition is used in English, the question arises: is it a marker of an **xcomp**? Ex:

- I'm relying on you to **come**.
- He complained about not being **invited**.

A pending question is: Is there a clausal equivalent to **obl**? That doesn't seem to be the case. Moreover, it seems that there is no one-to-one mapping between phrasal and clausal dependents in UD. And this is not a problem as the ambiguity existing between arguments and adjuncts in phrasal dependents is not present in clausal dependents. Therefore, if a sentence exhibits all the characteristics of an **xcomp**, it should, in my opinion, be annotated as an **xcomp**, regardless of whether it is introduced by a preposition or not.

---

[3] The name and the concept **xcomp** is borrowed from Lexical-Functional Grammar (LFG) (BRESNAN, 1982). Curiously, **xcomp** is used in LFG to distinguish complements from adjuncts, a distinction that UD rejects.

[4] The finite form can be expressed by auxiliary verbs or by copula verbs that modify the predicate, not necessarily by the predicate itself.

[5] Although this is not a criterion for distinguishing **ccomp** from **xcomp** in UD, this characteristic in PT contributes to having less confusion between these two tags in a confusion matrix.

[6] The English infinitive marker "to" is not always translated by a preposition in PT: *Ele quer fazer isso.* (He wants to do this.) *Ele pretende viajar*. (He intends to travel.).

If there is no clausal equivalent for **obl**, what about adverbial adjuncts and arguments of adjectives and adverbs? In PT, arguments of nouns, adjectives and adverbs are classified as "nominal complements", while their clausal correlatives are classified as "*orações completivas nominais*" ("nominal complement clauses"). As UD only explicitly annotates clausal modifiers for nouns as **acl** (adnominal clauses), there is a gap in terms of clausal (argument) modifiers of adjectives and adverbs. We have decided to fill this gap by utilizing our background, specifically by expanding the use of **acl** to clausal modifiers of adjectives and adverbs. This is an advantage from an NLP perspective, because noun, adjective and adverb complement clauses in PT have the same form: they are introduced by a preposition and always implicate an infinitive form[7] or a subjunctive inflected form

Again, in some cases, constructions with prepositions in PT are translated as constructions with prepositions into English, though in others, that is not the case. When an infinitive form in PT is translated as a gerund in English, a preposition may occur. However, when the infinitive in PT is translated as infinitives in English, the preposition is not used because in English the infinitive marker "to" and prepositions do not co-occur.

- *vontade **de** viajar* (desire **to** travel)
- *medo **de** ser demitido* (fear **of** being fired)
- *ansioso **para** viajar* (eager **to** travel)
- *temeroso **de** ser demitido* (afraid **of** being fired)
- *independentemente **de** ter dinheiro* (regardless **of** having money)

The lack of symmetric mappings between phrasal and clausal dependents can also be observed with regard to **amod**. In its clausal form, it corresponds to adnominal clauses - **acl** (the same relation used to annotate the clausal version of **nmod**) and to relative adnominal clauses - **acl:relcl**.

Another question that arises is: What is the clausal version of adverbial adjuncts, annotated with **advmod** or **obl**, depending on whether their PoS tag is an ADV or a NOUN? It seems that they are all covered by the **advcl** dependency relation. According to the UD guidelines[8], **advcl** can modify any predicate, whether it is verbal or nominal. The traditional semantic labels of adverbial clauses, such as temporal, consequence, conditional, and purpose, are mentioned in the guidelines as examples. According to the UD guidelines, the dependent of an **advcl** "must be clausal (or else it is an **advmod**)."

Finally, there are, in PT, clauses classified as "*orações apositivas*" ("appositive clauses"), in which the dependent of the relation is a clause, and "*aposto de oração*" ("apposition of clause"), in which the head of the relation is a clause.

- *Ele só quer isso: que você **venha**.* (He wants only this: that you **come**.) "appositive clause" in PT;

[7] Some nouns and adjectives also allow clausal complements in the form of finite clauses. In this case, the verb takes the subjunctive mood. Ex: *Eu tenho medo de que você se fira.* (I am afraid of you hurting yourself.)

[8] https://universaldependencies.org/guidelines.html

- *Ele propôs irmos de carro,* ***proposta*** *que ninguém aceitou.* (He proposed to go by car, a **proposal** that nobody accepted.) “apposition of clause” in PT, resumptive clause in English;
- *Ele propôs irmos de carro,* ***o*** *que ninguém aceitou.* (literally: He proposed to go by car, which nobody **accepted**.) “apposition of clause” in PT, summative clause in English.

In the UD tagset for dependency relations there is no corresponding clause to **appos**, and **appos** does not allow for either the head or the dependent to be a clause. Therefore, the decision of how to annotate them is up to each language or annotation project. In PT we advocate annotating most of them as **parataxis**, except for one case in which the referent of a clause is another clause, as if it were a relative clause with a clausal antecedent. In this case, the relative clause that modifies another clause could be well represented by **advcl:recl**, since it is adjunctive and modifies a predicate, similar to other **advcl**:

- *Ele esqueceu a chave em casa, o que o* ***fez*** *se atrasar.* (He forgot his key at home, which **made** him late.)

4. Conclusion

By contrasting the UD tagset of dependency relations with the PT set of syntactic functions, we were able to find areas of isomorphism (coincident terms and concepts) and anisomorphism (non-coincident terms and concepts) between the two. This can provide valuable insights for annotators working with different languages who currently use or plan to use the UD tagset. It helps them recognize areas of divergence and prevent the misapplication of concepts from their native language grammars to UD annotation.

This exercise is also beneficial for highlighting gaps in assigning tags to language phenomena. Some examples include:

- clauses introduced by prepositions that clearly function as an **xcomp,** but are not addressed in the UD guidelines;
- clauses introduced by prepositions that serve as complements of argument-taking adjectives and adverbs;
- appositive clauses.

By explicitly defining how to annotate clauses like these, UD would prevent each language or project from filling in the gaps according to its own interpretation. This is crucial for maintaining the universality of the UD tagset.

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 author thanks the support from 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. The author is also profoundly grateful for the two anonymous reviewers' comments and suggestions.

References

Andrews, Avery. (2007). The major functions of the noun phrase. In T. Shopen (Ed.), Language typology and syntactic description (pp. 62-154). Cambridge: Cambridge University Press.

Boland, Julie E.; Blodgett, Allison. (2006) Argument Status and PP-Attachment. Journal of Psycholinguistic Research, 35, pages 385–403. DOI 10.1007/s10936-006-9021-z

Bresnan Joan. (1982) The Mental Representation of Grammatical Relations. MIT Press, Cambridge, Massachusetts. https://doi.org/10.2307/414493

de Marneffe, Marie-Catherine; Manning, Christopher D.; Nivre, Joakim; Zeman, Daniel. (2021). Universal Dependencies. Computational Linguistics, 47(2):255–308. https://aclanthology.org/2021.cl-2.11

Nivre, Joakim; de Marneffe, Marie-Catherine; Ginter, Filip; Hajič, Jan; Manning, Christopher D.; Pyysalo, Sampo; Schuster, Sebastian; Tyers, Francis; Zeman, Daniel. (2020). Universal Dependencies v2: An Evergrowing Multilingual Treebank Collection. In Proceedings of the Twelfth Language Resources and Evaluation Conference (LREC 2020), pages 4034–4043, Marseille, France. European Language Resources Association. https://aclanthology.org/2020.lrec-1.497

Thompson, S. A. (1997). Discourse motivations for the core-oblique distinction as a language universal. In Akio Kamio (editor), Directions in Functional Linguistics, 36, pages 59–82. John Benjamins. https://doi.org/10.1075/slcs.36.06tho

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

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

[1]Núcleo Interinstitucional de Linguística Computacional (NILC)
Departamento de Letras – Universidade Federal de São Carlos (UFSCar)
Caixa Postal 676 — 13565-905 — São Carlos -– SP — Brasil

bryankhelven@ieee.org, ariani@ufscar.br

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

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

1. Introdução

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

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

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

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

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

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

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

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

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

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

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

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

3. Metodologia

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

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

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

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

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

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

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

4. Estatística dos fenômenos CGU no conjunto de análise

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

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

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

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

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

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

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

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

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

5.1. URL

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

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

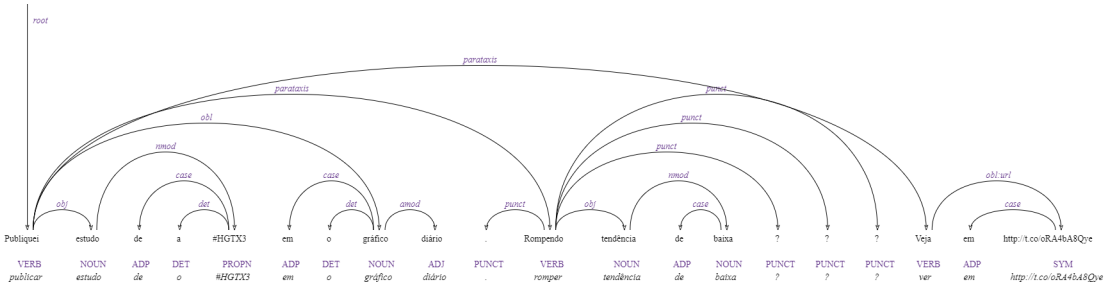


Fig. 1. Exemplo de URL integrada com preposição explícita e anotada com obl:url.

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

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

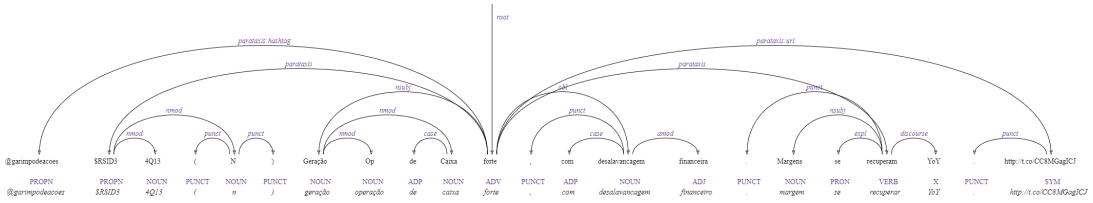


Fig. 2. Exemplo de URL standalone precedida de ponto final e anotada com parataxis:url.

5.2. Hashtag/Cashtag

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

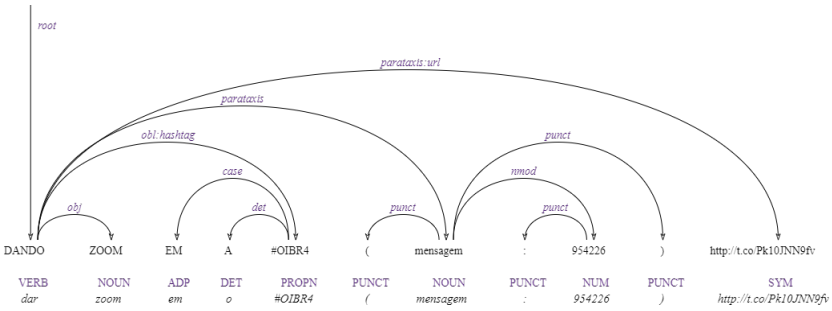


Fig. 3. Exemplo de hashtag integrada anotada com obl:*hashtag*.

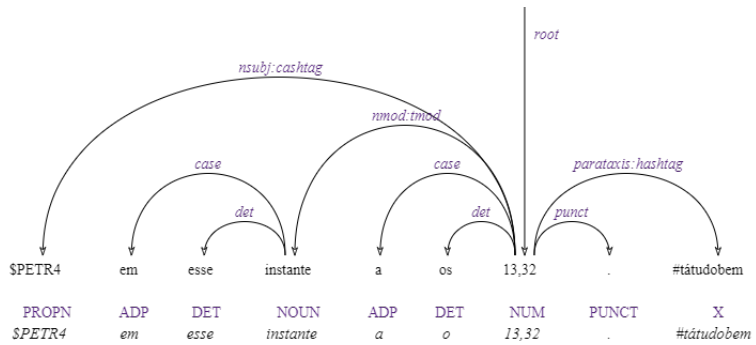


Fig. 4. Exemplo de *cashtag* integrada anotada com nsubj:*cashtag*.

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

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

5.3. Menção

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

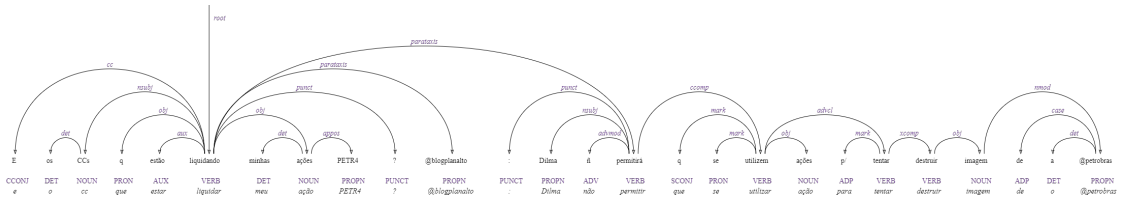


Fig. 5. Exemplo de menção integrada à sintaxe anotada com nmod.

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

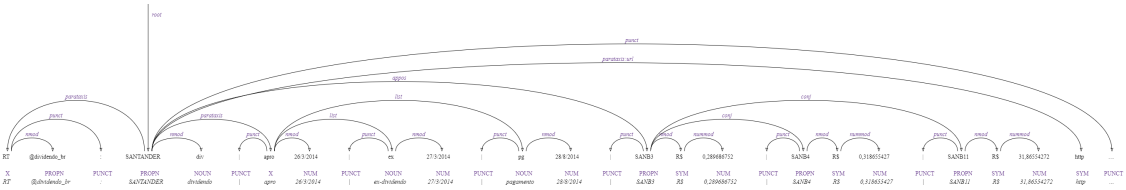


Fig. 6. Exemplo de menção conectada a RT por nmod.

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

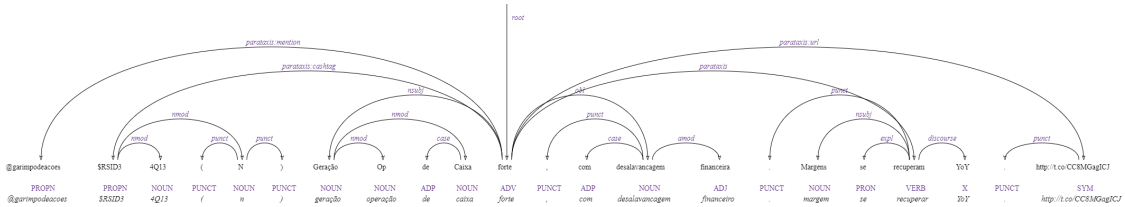


Fig. 7. Exemplo de menção *standalone* anotada com parataxis:mention.

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

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

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

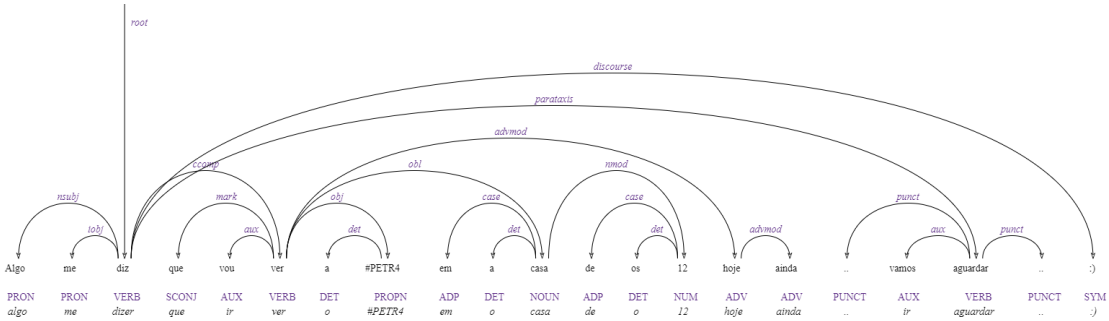


Fig. 8. Exemplo de emoticon anotado com discourse.

5.5. RT (marca de *retweet*)

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

5.6. Truncamento

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

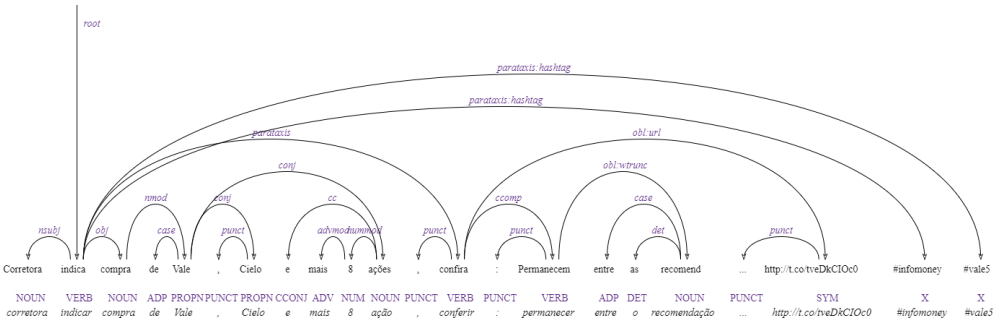


Fig. 9. Exemplo de anotação sintática para truncamentos.

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

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

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

6. Considerações finais

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

Agradecimentos.

Este trabalho foi realizado no âmbito do Centro de Inteligência Artificial da Universidade de São Paulo (C4AI - http://c4ai.inova.usp.br/), com o apoio da Fundação de Amparo à Pesquisa do Estado de São Paulo (processo FAPESP #2019/07665-4) e da IBM. Este projeto também foi apoiado pelo Ministério da Ciência, Tecnologia e Inovações, com recursos da Lei N. 8.248, de 23 de outubro de 1991, no âmbito do PPI-Softex, coordenado pela Softex e publicado como Residência em TIC 13, DOU 01245.010222/2022-44.

References

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

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

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

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

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

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

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

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

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

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

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

Universal Dependencies and Language Contact Annotation: Experience from Warao refugees signs in Brazil

Dalmo Buzato[1]

[1]Federal University of Minas Gerais, Belo Horizonte, Brazil

buzatodalmo@gmail.com

Abstract. *This article aims to present a work in progress that proposes to describe, from the Universal Dependencies (UD) project, the linguistic contact between the Warao language, Venezuelan Spanish, and Brazilian Portuguese on the Warao refugee signs in Brazil. In the present work, in addition to presenting a brief description of the contact between the three languages in the Venezuelan indigenous migration to Brazil, the description of contact languages using the Universal Dependencies is discussed, in addition to initial reflections on methodological choices motivated by the linguistic phenomena observed in the corpus.*

Resumo. *O objetivo do presente artigo é apresentar um trabalho em andamento que se propõe a descrever, a partir do projeto das Universal Dependencies (UD), o contato linguístico entre a língua warao, o espanhol venezuelano e o português brasileiro nas placas de refugiados Warao no Brasil. No presente trabalho, além de apresentar-se uma breve descrição do contato entre os três idiomas na migração indígena venezuelana para o Brasil, discute-se sobre a descrição de línguas de contato valendo-se das UD, além de reflexões iniciais acerca de escolhas metodológicas motivadas pelos fenômenos linguísticos observados no corpus.*

1. Introduction

This article aims to document the ongoing compilation, transcription, and annotation of a treebank consisting of signs written by Warao refugees from Venezuela in Brazil. For this purpose, we rely on the Universal Dependencies (UD) project [Nivre et al. 2016] to describe the linguistic contact between the Warao language, Venezuelan Spanish, and Brazilian Portuguese.

This article will present a brief introduction to the linguistic and sociopolitical context of the Warao migratory uprising from Venezuela to Brazil. Additionally, we will address the relationship between the annotation of contact languages and phenomena (e.g., code-switching, pidgins, creoles, and mixed languages) and the Universal Dependencies project. Finally, we will discuss some initial reflections on the decisions and specific aspects to be taken into consideration in annotating linguistic phenomena in contact contexts for computational purposes.

The Warao are an indigenous ethnic group inhabiting the northeastern region of Venezuela, known as the Orinoco River Delta, as well as some regions in Guyana and Suriname. The Warao people speak a language with the same name, which is an isolated language with no known linguistic relatives. According to [Romero-Figueroa 2020],

[UNHCR 2021a], and [UNHCR 2021b] the Warao constitute the third-largest indigenous population in Venezuela, with approximately 41,000 individuals, making them one of the most prominent and significant indigenous peoples in the country. Anthropological studies suggest that the Warao might be the oldest inhabitants of present-day Venezuela, as they have been residing in the Orinoco Delta region for at least 8000 years.

The humanitarian crisis of the Warao people seems to predate the migratory flow of Venezuelans to Brazil. As reported in [Buzato and Vital 2023] and [GarcÍa-Castro 2006], since the second half of the 20th century, especially from the 1970s onwards, a scenario of subalternity has been observed among the Warao in Venezuela due to the expansion of extractive, agricultural, and industrial activities in the Orinoco delta region. Nomadism was not the traditional way of life for the Warao; however, following the loss of territories to the mentioned activities and in pursuit of better living conditions, the Warao began to migrate to urban centers within Venezuela.

Due to linguistic and cultural differences, the Warao faced numerous challenges in integrating into Venezuelan cities, consequently encountering various hardships such as poverty, violence, discrimination, underemployment, and low educational attainment. These factors could be understood as rendering them subaltern people in their own country after losing their place of origin. The linguistic contact between Warao and Venezuelan Spanish appears to be more enduring, yet its effects are sparsely documented. [Romero-Figueroa 2020] focuses on the lexical effects of this contact.

With the worsening of the political and economic crisis in Venezuela, the migratory flow to Brazil intensified from 2015 onward. Consequently, there was a flow of Warao immigrants to Brazil, initially concentrated in the northern states of the country, particularly in the capitals of Belém and Rondônia. The Warao people, who were already disadvantaged in their country of origin, perceive this condition to be doubly amplified upon migrating to Brazil, now facing an even more distinct and pronounced language barrier.

As a result of the substantial migratory influx from Venezuela to Brazil, the federal government, in collaboration with third-sector organizations, has initiated the ”Operação Acolhida”. This operation entails the internalization of refugees and migrants from Venezuela who arrive in the northern region of Brazil, aiming to enhance their quality of life, facilitate social integration, and mitigate the urban challenges confronted by municipalities near the border.

Despite government and third-sector assistance, a portion of the refugees finds themselves in precarious situations, often relying on the support of the host communities to survive and acquire essential food and daily necessities. In order to do so, refugees resort to signs bearing requests for aid, seeking contributions, typically in monetary form, from the Brazilian population. An overview of these signs and their communicational, textual, and linguistic features can be found in [Mesquita 2020] and [Buzato and Vital 2023].

2. UD e language contact

Linguistic contact occurs when speakers of different languages interact with each other or become part of the same speech community [Crystal 1987]. It is an extremely productive

linguistic phenomenon that has many possible ramifications in linguistic structure. These range from lexical borrowing and code-switching to the emergence of pidgin or mixed and creole languages. The differentiation and distinction among each of these categories are not uniform and continue to be subjects of various discussions in linguistic studies.

Currently, as stated in the Universal Dependencies project documentation, only one Creole language is documented through the framework: the Naija language (Nigerian Pidgin), a contact language spoken in Nigeria. The corpus was constructed from transcriptions of audio recordings collected in 2017 for the *ANR NaijaSyncor* project [Caron et al. 2019]. This oral corpus is also characterized by occasional code-switching to English, as well as to various native Nigerian languages, including Yoruba, Hausa, and Igbo.

However, regarding code-switching studies, there are currently four treebanks dedicated to documenting this phenomenon: *UD Frisian Dutch-Fame* [Braggaar and van der Goot 2021], *UD Maghrebi Arabic French-Arabizi* [Seddah et al. 2020], *UD Turkish German* [Çetinoğlu and Çöltekin 2019], and *UD Hindi English* [Bhat et al. 2018].

The current article appears to contribute to the state of the art in Universal Dependencies by aiming to describe a language that seems to go beyond the boundaries of code-switching, yet is not a Creole language. The language that emerges from the signs produced by refugees appears to surpass the two aforementioned boundaries, constituting a category of emergent languages (e.g., pidgins or mixed languages), which have not been documented within the UD framework up to the present moment.

The discussion about the nature of the language produced by Warao refugees in Brazil is not within the scope of the present article. We agree that for more robust analyses and a more precise definition, a larger amount of data would be necessary, preferably encompassing other communicational situations besides the use of signs for asking for help. Sociolinguistic information about the process of creating these signs, their authors, as well as more detailed profiling of the communities, and the alternation of linguistic uses (for instance, which language they use to communicate among themselves?) would be of great value for a better understanding of the phenomenon and the nature of this emergent language.

Among the documented contact languages in the Universal Dependencies project, *Hindi English* and *Maghrebi Arabic French* are languages that are predominantly documented in written genres. In the case of *Hindi English*, it is documented through tweets, while *Maghrebi Arabic French* has been documented through news articles and comments on Algerian newspaper web forums. This characteristic of other languages seems to support the documentation of contact phenomena in the written modality of language, similar to the case of the Warao refugee signs described in this study. Just as in our case, the written modality of the other treebanks appears to be influenced by the spoken modality, as we will address in the upcoming sections.

3. Data and Initial Discussions

The initially annotated corpus consists of 21 photographs of signs created by refugees, collected by researchers in two medium and large cities in the southeastern region of

Brazil during the months of May 2022 and May 2023. An example of the collected signs can be observed in Figure 1.

Figure 1. Example of a Sign created by refugees

When selecting written texts as the object of analysis, we must consider the specificities of this modality compared to spoken language. However, in our case, the signs are written by speakers who mostly have very low levels of education and formal instruction, as indicated by demographic profiling. We must also take into account the transposition of strategies and phenomena from the oral modality of language into the refugees' textual production. These phenomena are not necessarily derived from linguistic contact, but they certainly influence it in some way. Therefore, choices in transcription related to orality and the absence of formal instruction impact the visualization and understanding of the contact phenomena that we will discuss further.

For example, in Figure 1, despite the absence of uppercase and lowercase letters and spacing between words on the sign, we have chosen to transcribe them separately according to the words present in Brazilian Portuguese. In cases where speakers write words with orthographic deviations, we have decided to retain them due to the possibility of containing contact phenomena. Additionally, if speakers write incomprehensible or nonexistent words in Portuguese, we have chosen to preserve them in that form for potential lexical parallels in the other languages involved in the contact. Lastly, most signs lack any graphical punctuation marks. As we believe that the absence of punctuation usage reveals much about the migrants' formal education level and textual production, we have chosen not to add any punctuation marks to the signs, leaving them with only the possible punctuation marks that each speaker used.

Therefore, the sign we saw above in Figure 1, for instance, was transcribed in our treebank as follows:

(1) boa tarde irmao sou da venezuela preciso ajuda dinheiro amigo para paga luga para comprar roupa gazisa
good afternoon buddy ∅ am from venezuela ∅ need help money friend for payrent for buy clothes cooking gas

If we were to strictly adhere to the rules of grammatical norms, we should reasonably transcribe it as follows:

(2) Boa tarde, irmão! Sou da Venezuela, e preciso de uma ajuda, dinheiro para pagar o aluguel, comprar roupa e gás.
Good afternoon, buddy! I'm from Venezuela, and I need help – money to pay for rent, buy clothes, and get cooking gas.

The content of the photographs was transcribed into a .txt file and automatically annotated using the UDpipe tool [Straka et al. 2016], with a Portuguese language model based on Bosque-UD v. 2.6 [Rademaker et al. 2017]. Afterward, the generated CONLLU files were imported into the annotation tool Arborator-Grew-NILC (https://arborator.icmc.usp). For human annotation and review, we followed the general UD guidelines, as well as the ICMC/USP annotation manual for the Portuguese language [Duran 2021]. In the following sections, we will discuss certain aspects related to the transcription of the signs and the methodological choices made during the process of human review of the content automatically annotated by the model.

The automatic annotation proved to be less effective due to the nature of the input text, along with another factor that significantly influences the written production of refugees: low education levels. From an orthographic perspective, the vast majority of signs lack any punctuation, spacing, or graphical accents. As we are aware, all these factors impact automatic annotation, which certainly explains the low performance of the model with the texts tested here. Furthermore, the writing, strongly influenced by oral aspects, along with anomalous constructions due to the context of linguistic contact and low education levels, also certainly accounts for the substantial need for manual human review and annotation.

For instance, when observing the automatic annotation generated by the model for the sentence provided in (1), shown in Figure 2 below, we notice that the absence of punctuation, combined with certain morphosyntactic characteristics of the signs, led to erroneous annotation by the model. The model mistook the verbal conjugation of the first person singular in the present indicative of the verb precisar (preciso *"I need"*) for the homonymous adjective. This is likely due to the absence of punctuation and the first-person singular pronoun (Eu) before the verb.

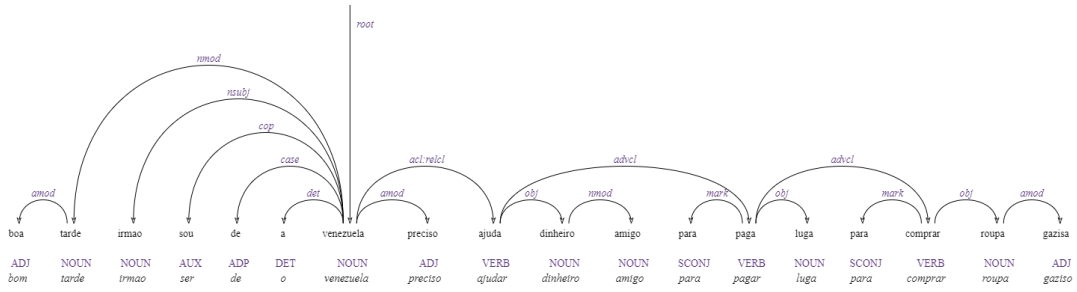


Figure 2. Automatic annotation of transcription (1) generated by the Bosque-UD v. 2.6 model

The model used to automatically annotate the transcriptions was trained on written texts from the journalistic genre, which possibly explains some of the model's difficulties

in annotating more oral-productive phenomena. Examples in transcription (1) can be observed in the model's inadequate annotation of the greeting "boa tarde" (*good afternoon*), an interjection that should be connected to the root with the *discourse* deprel, and the vocative "irmao" (*buddy*), to whom the message is addressed. Written journalistic genres rarely include vocatives or greeting interjections, which could explain the model's struggle in representing them. However, this greeting structure is highly productive in our data, found in almost all signs, along with a high recurrence of structures for farewells and expressions of gratitude (deus abencoe obrigado *"God bless you thanks"*), as seen in Figure 3. Therefore, it is a recurring demand for human correction in our experience.

Figure 3. Example of a Sign created by refugees

Other inconsistencies that can be observed in the automatic annotation of transcription (1) include difficulties in annotating the connections between segments using the *parataxis* and *conj* deprels. Our explanation for this occurrence is that the way elements are connected through these deprels in our transcriptions differs from what is common in Brazilian Portuguese. This is likely influenced by the morphosyntax of the Warao language, as we will discuss later on. Lastly, in using the vocative twice in the passage, seemingly disrupting the linearity of the explanation, as with the word "amigo" (*friend*) between "dinheiro" (*money*) and "para paga" (*to pay*), a phenomenon possibly more common in spontaneous speech, the model annotated "amigo" as a nominal modifier of the word "dinheiro", which lacks semantic consistency. A first revised version of transcription (1) can be found in Figure 4 below.

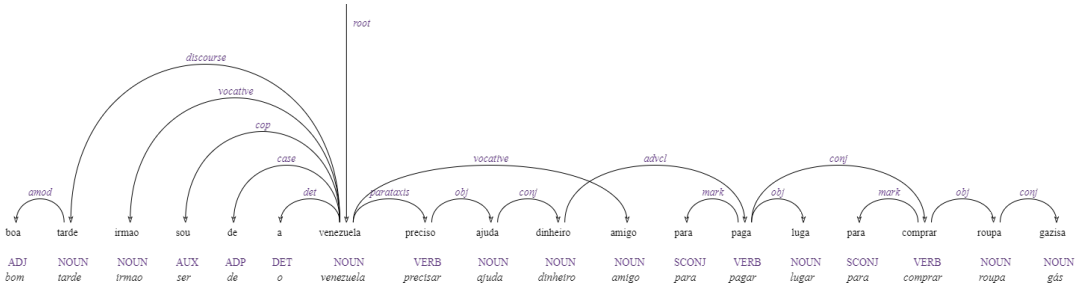


Figure 4. Revised annotation of transcription (1) generated by the Bosque-UD v. 2.6 model

As we can observe in Figure 4, the annotated relationship between the verb "preciso" (*I need*) and the complement "ajuda" (*help*) was labeled with the deprel *obj* (direct

object). According to [Luft 2010], the verb ”precisar” assumes different classifications of verbal transitivity depending on the complement. If the complement is a noun or pronoun, it is common to use a preposition, as exemplified in Figure 5. On the contrary, when the complement is an infinitive verb, the use of a preposition is not required, as in Figure 6.

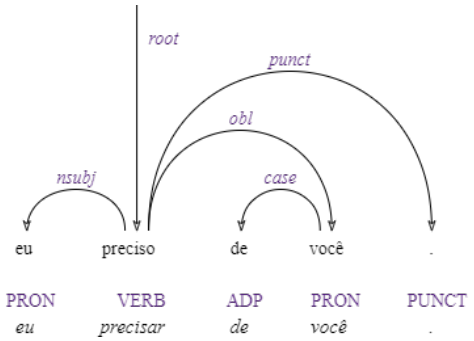


Figure 5. Sentence with the verb ”precisar” classified as an indirect transitive verb (deprel obl)

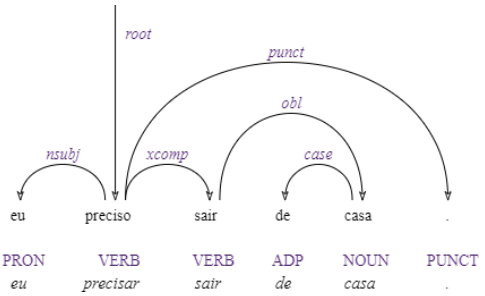


Figure 6. Sentence with the verb ”precisar” without the mandatory use of a preposition (deprel xcomp)

In Brazilian Portuguese, the verb ”precisar” usually requires a preposition in constructions like that in transcription (1), which makes it an indirect transitive verb and leads to its annotation with the deprel *obl* (oblique nominal). In the illustrated example, the refugees use ”precisar” without the preposition, which resulted in the segment being annotated as *obj*, contrary to the pattern observed in Brazilian Portuguese for similar sentences.

This phenomenon seems to have its origin in the transposition of the argument structure of the Spanish verb *necesitar* (to need) into Brazilian Portuguese. Therefore, it is a contact phenomenon. As we can observe in Figure 7 below (translation: ”I need help to buy chicken”), automatically annotated using the AnCora-UD 2.6 model for Spanish [Taulé et al. 2008], the verb *necesito* does not require a preposition in the construction *necesito ayuda*. Due to the proximity between the Spanish language and Brazilian Portuguese, what appears to occur is that speakers transpose the argument structure of the Spanish verb *necesito* onto the Brazilian Portuguese verb *preciso*.

Another phenomenon observed quite productively in the signs was typical phonological phenomena of spoken Brazilian Portuguese, especially in informal speech. As we can see in Figure 8 below, the orthographically proper way in Brazilian Portuguese would be ”preciso de ajuda” (*I need help*) and ”pedimos dinheiro” (*we ask for money*) for

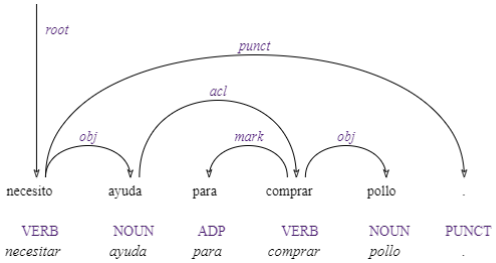


Figure 7. Sentence annotated with the AnCora v. 2.6 model in Spanish

the content in the third line of the sentence. In contrast, we observe the occurrence of "diajuda", where the absence of tonic accent on the preposition forms a single phonetic group [di.a.ˈʒu.dɐ], and it was equally represented as a single word orthographically. Furthermore, there is vowel harmony in "p[i]dimo ∼ p[e]dimo[s]", with the elision of the [s] sound. The elision of the [s], representing the plural mark, also occurs in the fifth line of the sign in the segment "02 menino 03 minina" (*two boy three girl*), which also features vowel harmony in "m[e]nina ∼ m[i]nina". Additionally, we observe diphthongization in the verb "ser" (to be) conjugated in the first person plural in the present indicative tense in the first line of the sign (somos ∼ so[u]mos).

The presence of these occurrences, typical of the oral modality of Brazilian Portuguese, represents an absorption of the language by refugees and especially reveals the specific Portuguese modality to which they have access. They likely do not have contact with the orthographically appropriate written modality or formal instruction in the Portuguese language, which is why they transpose speech phenomena to writing. This transposition also occurs in various textual genres among Portuguese speakers, especially in informal contexts such as daily speech or tweets. These phenomena have proven to be particularly challenging for automatic annotation, and we believe that this difficulty is likely to be encountered when annotating texts exclusively in Brazilian Portuguese, but from text genres that are predominantly informal, as mentioned above.

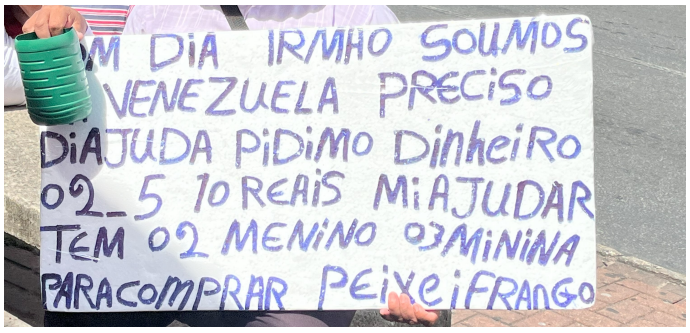


Figure 8. Example of a Sign created by refugees

Furthermore, certain aspects of the morphosyntax of the Warao language, an isolated agglutinative language, appear to be of paramount importance in explaining certain recurring phenomena that we observe in the data. In all the examples presented throughout this article, we notice the absence of conjunctions, for instance. This is possibly due to the lack of this grammatical class in the Warao language, as described by [Romero-Figueroa 1997]. Conversely, according to the author, the connection between

clauses in the language is achieved through parataxis. As a result, in the transcriptions and annotations of the texts, we observe that the majority of connections between clauses in the corpus do not occur through structural marking.

Other significant structural aspects of the Warao language that possibly influence the described contact include the constituent order, which, as per [Romero-Figueroa 1985], is an OSV language; the alternation between facultative use and absence of a copula verb in sentences; the absence of prepositions in the language, instead employing postpositions; the lack of marking for personal pronouns (zero morpheme) for various cases of the second and third person in singular and plural. Finally, the usage conditions and the organization of the pronominal and adposition systems of the Warao language appear consistently divergent from Romance languages. For instance, in Warao, the equivalents of the prepositions ”to” and ”for” for transitive verbal actions can occur through the dative case markers -ma and -to, or through the postposition ”saba”. The equivalents of the preposition ”of” in Warao are the postpositions ”a” and ”abitu”, solely expressing possession.

4. Final remarks

The objective of the current article was to report on the ongoing documentation of linguistic contact among the Warao language, Venezuelan Spanish, and Brazilian Portuguese resulting from Warao migration to Brazil, utilizing the UD project. Additionally, the article discussed language annotation and contact phenomena through UD, along with theoretical and methodological insights made up to the present moment during the annotation and transcription of the collected data.

The upcoming steps involve obtaining additional photographs to enhance the volume of data and facilitate more refined analyses. This will be followed by a stage of reviewing the annotations carried out by experienced researchers in UD. Furthermore, the plan is to also incorporate photographs and recorded interviews available on the web to increase the quantitative data in the treebank. This approach will not only expand the dataset but also offer indications of contact in other contexts and communicative modalities.

Lastly, numerous other theoretical considerations arise from the annotations, such as: Is the emerging language a pidgin or a jargon? Taking into account the reflection by [Holm 2000] on linguistic relatedness, does the linguistic proximity between Venezuelan Spanish and Brazilian Portuguese, even in contact with the Warao language, hinder the emergence of a pidgin? What is the influence of the Warao language's morphosyntax on the emergence of this language, considering linguistic attitudes and the linguistic landscape in which the refugees are situated?

References

Bhat, I., Bhat, R. A., Shrivastava, M., and Sharma, D. (2018). Universal dependency parsing for hindi-english code-switching. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, volume 1, page 987–998.

Braggaar, A. and van der Goot, R. (2021). Challenges in annotating and parsing spoken, code-switched, Frisian-Dutch data. In *Proceedings of the Second Workshop on Do-*

main Adaptation for NLP, pages 50–58, Kyiv, Ukraine. Association for Computational Linguistics.

Buzato, D. and Vital, Á. (2023). O contato linguístico em placas de refugiados venezuelanos em belo horizonte e região metropolitana: observações preliminares. In *Anais do Congresso Nacional Universidade, EAD e Software Livre*, volume 1.

Caron, B., Courtin, M., Gerdes, K., and Kahane, S. (2019). A surface-syntactic ud treebank for naija. In *Proceedings of the 18th International Workshop on Treebanks and Linguistic Theories (TLT, SyntaxFest 2019)*, pages 13–24. Association for Computational Linguistics.

Çetinoğlu, Ö. and Çöltekin, Ç. (2019). Challenges of annotating a code-switching treebank. In *Proceedings of the 18th International Workshop on Treebanks and Linguistic Theories (TLT, SyntaxFest 2019)*, pages 82–90, Paris, France. Association for Computational Linguistics.

Crystal, D. (1987). The cambridge encyclopedia of language. *UK: Cambridge University*.

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 ICMC 435, Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos-SP.

GarcÍa-Castro, Á. (2006). Migración de indígenas warao para formar barrios marginales en la periferia de ciudades de guayana, venezuela. *De Quito a Burgos: migraciones y ciudadanía. Burgos: Gran Vía*.

Holm, J. (2000). *An introduction to pidgins and creoles*. Cambridge University Press.

Luft, C. P. (2010). *Dicionário Prático de Regência Verbal: Nova Ortografia*. Ática, São Paulo, 9 edition.

Mesquita, R. (2020). Diaria o fixo: fotografias sociolinguísticas de boa vista–roraima e as novas perspectivas para as pesquisas do contato linguístico na fronteira. In Cruz, A. and Aleixo, F., editors, *Roraima entre línguas: contatos linguísticos no universo da tríplice fronteira do extremo-norte brasileiro*. Editora da UFRR.

Nivre, J., De Marneffe, M.-C., Ginter, F., Goldberg, Y., Hajic, J., Manning, C. D., McDonald, R., Petrov, S., Pyysalo, S., Silveira, N., et al. (2016). Universal dependencies v1: A multilingual treebank collection. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1659–1666.

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.

Romero-Figueroa, A. (1985). Osv as the basic order in warao. *Lingua*, 66:115–134.

Romero-Figueroa, A. (1997). *A Reference Grammar of Warao*. Lincom Europa, München.

Romero-Figueroa, A. (2020). *El contacto warao-espanol: Consideraciones sobre el proceso de aculturacion lexica de la lengua nativa del delta del Orinoco*. Editorial Académica Española.

Seddah, D., Essaidi, F., Fethi, A., Futeral, M., Muller, B., Suarez, P. O., Sagot, B., and Srivastava, A. (2020). Building a user-generated content north-african arabizi treebank: Tackling hell. In *Proceedings of the 58th annual meeting of the Association for Computational Linguistics*, pages 1139–1150.

Straka, M., Hajic, J., and Straková, J. (2016). Udpipe: trainable pipeline for processing conll-u files performing tokenization, morphological analysis, pos tagging and parsing. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 4290–4297.

Taulé, M., Martí, M. A., and Recasens, M. (2008). Ancora: Multilevel annotated corpora for catalan and spanish. In *Lrec*, volume 2008, pages 96–101.

UNHCR (2021a). Os warao no brasil - contribuições da antropologia para a proteção de indígenas refugiados e migrantes. Technical report, Brasília.

UNHCR (2021b). Perfil socioeconômico da população indígena refugiada e migrante abrigada em roraima. Technical report, Brasília.