Automatic Annotation of Enhanced Universal Dependencies for Brazilian Portuguese

Elvis A. de Souza, Magali S. Duran, Maria das Graças V. Nunes, Gustavo Sampaio, Giovanna Belasco, Thiago A. S. Pardo

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

{elvis.desouza99,magali.duran}@gmail.com, gracan@icmc.usp.br,

{gustavo.sampaio,giovannabelasco}@usp.br, taspardo@icmc.usp.br

Abstract. *This paper presents the first attempt to automatically annotate Enhanced Universal Dependencies for Brazilian Portuguese. We use a symbolic annotation system, based on graph rewriting rules, and modify its original rules to better suit the linguistic characteristics of Portuguese using a manually annotated sample from the journalistic portion of Porttinari treebank as ground truth. Our objective is to assess the performance of the automatic annotation for a novel language and to determine the extent of possible improvements through rule modifications. Results demonstrate significant performance enhancements, where linguistic-driven rule adjustments improved the annotation accuracy 11.38 points, achieving 96.05% F1-score.*

1. Introduction

Morphological and syntactic annotation have shown to be relevant for several Natural Language Processing (NLP) initiatives. For instance, tasks of open information extraction (Oliveira et al. 2023) and text simplification (Candido et al. 2009) may directly base their decisions on syntax. Considering the more recent trends of Large Language Models, several works have demonstrated improvements in results when linguistic knowledge is provided (Zhou et al. 2020; Bai et al. 2021; Lin et al. 2021; Bölücü et al. 2023). On the linguistic perspective, linguistic annotation may help describing varied language phenomena, possibly supporting the validation and/or proposal of new theories.

Universal Dependencies (UD) is a framework for the morphological, morphosyntactic and syntactic annotation of human languages. UD provides standardized guidelines and has been used to annotate over 283 treebanks for 161 languages, being widely adopted as it proposes consensual annotation decisions and allows comparative and multilingual efforts. Concerning the syntactic annotation, the UD framework supports two levels of depth: basic dependency trees and enhanced graphs. Basic dependency trees provide information on syntactic dependencies, where each token is connected to a governing (head) token through a relation (e.g., in the sentence *The boy cried*, "boy" is connected as subject to the head "cried" by a *nsubj* relation). Enhanced Universal Dependencies (EUD) generally build upon the basic dependencies by adding relations and nodes (or tokens) to make explicit the implicit relationships between tokens (Nivre et al. 2020) (e.g., in Figure 1, "boy" is also connected to "left" by a *nsubj* enhanced relation, as it is shared by the verbs "cried" and "left"). This enhancement can facilitate NLP tasks by providing additional information.

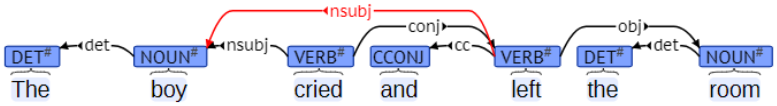

Figure 1. EUD annotation – the red *nsubj* dependency is a new EUD dependency.

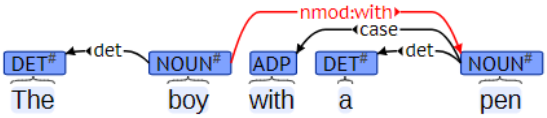

Figure 2. EUD annotation – relation extended with the lexical item "with".

This paper investigates the issue of EUD annotation for Brazilian Portuguese. To the best of our knowledge, this is the first evaluation of EUD annotation for this language. Following two previous shared tasks on EUD annotation (Bouma et al. 2020; Bouma et al. 2021), which did not include Portuguese, we build upon one of the systems that participated in the 2021 task, namely Grew (Guillaume and Perrier 2021), based on graph rewriting rules for annotated syntactic trees. This symbolic system comes with a set of original (and universal) rules, and we made a series of modifications based on corpus investigation, generating an improved set of rules. The two sets of rules were applied to a sample dataset from the journalistic portion of Porttinari (Duran et al. 2023), a Portuguese treebank available in the Universal Dependencies project catalog, which we manually enriched with EUD annotation to assess the quality of the automatic annotation. Therefore, our objective is to verify the performance of the program's original rules for Portuguese and how much we can improve it with modified rules.

In the end, we discuss persistent annotation errors and future perspectives on EUD automatic annotation. As an additional contribution, the rules and the annotated data are also made available to the interested reader.

2. Related Work

EUDs present significant challenges compared to traditional UD annotation. In addition to the UD website, where the guidelines are updated as needed, there is a series of works discussing the relevance and explaining the application of this type of annotation in treebanks (De Marneffe et al. 2014; Nivre et al. 2016; Schuster and Manning 2016; Nivre et al. 2020). The instantiation of these relations for Portuguese was introduced and detailed in (Pagano et al. 2023). Overall, EUDs may include 6 annotation situations:

1. Inclusion of the prepositions, coordinating conjunctions, and subordinating conjunctions lemmas in the label of the relations they introduce (as in Figure 2);
2. Identification of the controlling subject of the null subject in *xcomp* clauses (as in Figure 3);

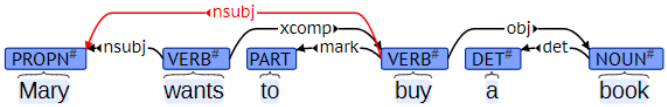

Figure 3. EUD annotation – *nsubj* relation for a verb dependent of *xcomp*.

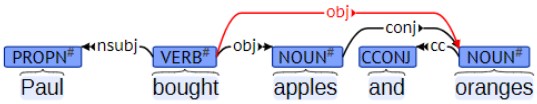

Figure 4. EUD annotation – *obj* relation propagated to the dependent of *conj*.

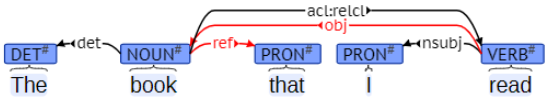

Figure 5. EUD annotation – "book" is the object of "read" and "that" is *ref* of "book".

3. Propagation, to the dependent of *conj*, of the relation that reaches the head of *conj* (as in Figure 4);
4. Propagation, to the dependent of *conj*, of some relations that depart from the head of *conj* (as in Figure 1);
5. Replacement of the relative pronoun in relative clauses with its antecedent, marking the relationship of the relative pronoun with its antecedent with a label exclusive to the EUD: *ref* (as in Figure 5);
6. Insertion of an empty token to take the place of an elliptical predicate and establishment of relationships of this empty token with the participants of the *orphan* relation (as in Figure 6).

While UD trees are simple hierarchical structures with a root, EUD graphs are connected and can contain cycles. For example, in Figure 5, the node "book" is dependent of "read" in a *obj* relation, however, it is also governor of "read" in a relative clause relation (acl:relcl), a basic syntactic annotation that is kept in the enhanced graph, establishing a cycle between two nodes. Another challenge is that some relations are lexicalized (as in Figure 2), considerably increasing the set of labels to be predicted and making them language-dependent. Additionally, a token can have more than one enhanced relation, having multiple governors, and there may be additional empty tokens to represent elliptical predicates (Bouma et al. 2020). In Figure 1, the node "boy" has two governors: the verbs "cried" and "left", which are coordinated, while in the basic annotation only the first verb would be its governor. In Figure 6, an empty token, *[has]*, has been added to the EUD graph to solve the elliptical predicate issue, and several dependencies were changed to fit this new token.

The shared tasks held at IWPT in 2020 (Bouma et al. 2020) and in 2021 (Bouma et al. 2021) provided a platform for comparing results among different systems. To date, there is no treebank annotated with EUD for the Portuguese language, meaning

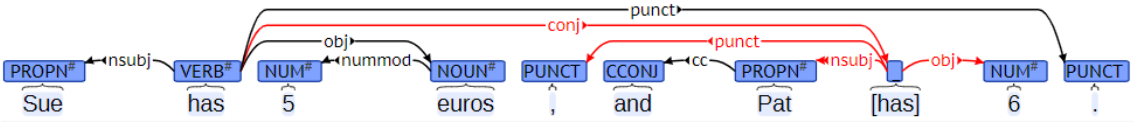

Figure 6. EUD annotation – an empty token *[has]* was inserted to account for an elliptical predicate.

that the language has never been subjected to any attempt of automatic annotation. To participate in the competition, a treebank did not need to have all six types of EUD; here, we are testing a rule-based approach on a fully annotated Portuguese dataset with all six types of EUD, produced for the purpose of this work.

The system we chose to use, Grew, ranked seventh in the 2021 competition, with 81.58% ELAS (a F1-score over EUD relations), being the best ranked symbolic-based system[1]. Our goal is to test the possibilities and limitations of a linguistically-driven rule-based approach, which can be constructed with linguistic supervision, being easily applied for other languages as well, without training, and with high interpretability.

3. Methodology

We use two small gold-standard EUD sets: one for testing (gold-test) and one for development (gold-dev). The gold-dev set was drawn from Porttinari-base, the main portion of Porttinari, while gold-test was sourced from Porttinari-test, designed for evaluating automatic annotation systems (Duran et al. 2023). Gold-dev comprises 100 manually selected sentences, chosen by a linguist to represent challenging EUD phenomena. In contrast, the 100 test sentences were randomly selected to reflect the natural frequency of phenomena in Porttinari. Due to (intentional) differing selection methods, dev and test sets show disparities, e.g., the dev set contains 23 sentences with the *orphan* label (predicate ellipsis), whereas the test set includes only two.

We begin our work analyzing Grew (Guillaume and Perrier 2021) original rules for EUD, referred to as "original rules", which are universal and ideally applicable to any language. We observed the annotation results on the development set and, as errors were identified, we created new rules and modified existing ones to address these deficiencies. Notably, none of the sentences from the test set influenced rule modifications. As a result of the process, we have the rule set named "modified rules."

Our evaluation focuses on the program's overall F1-score (also named ELAS, i.e., labeled-attachment score over enhanced dependencies), as well as F1-score for each of the 6 EUD types. To achieve this, we automatically classified each enhanced relation into one of the 6 categories using linguistic rules. For example, we know, from Figure 3, that *nsubj* relations from verbs that are *xcomp* dependents towards nominals, when the nominal also has a *nsubj* relation coming from the verb that is the *xcomp* governor, are relations of the type "assignment of xcomp subjects".

Grew rules consist of patterns (that may involve any UD annotation information) to be identified in sentences and a set of commands to be executed when these patterns are found. These rules are incorporated into a mechanism known as a "strategy," which allows for the control of which rules are applied for each language and in which order. For instance, the resolution of predicate ellipses should be done first, as other rules related to the propagation of dependents of coordinated elements can be applied considering the empty token inserted in the sentence.

In Table 1, we find the number of rules for each type of EUD relation (1-6) in Grew rule set, according to our automatic identification of EUD types, plus our new rules (7). There are also "unclassified" rules, as they do not produce any visible changes to

[1]The system is 7.66 points below the system that ranked first, TGIF (Shi and Lee 2021).

a sentence, but rather implicit changes that are going to be used for other rules inside a Grew strategy. Besides the new rules, some of the original rules were modified, and they will be seen in the Results section, where we consider how many times the rules for each EUD type have been applied before and after our modifications.

EUD Types	Number of Rules
1 - Addition of prepositions and conjunctions	3
2 - Assignment of xcomp subjects	11
3 - Propagation of conj head	18
4 - Propagation of conj dependents	8
5 - Annotation of relative pronoun referent	13
6 - Inclusion of elliptical predicate	23
7 - New rules	15
Unclassified rules	64

Table 1. Number of rules related to each EUD type

4. Results

Both the test and dev samples were manually annotated for EUD. Table 2 presents a description of these corpora, as well as the distribution of each of the EUD types. The number of EUD relations in this section ignores relations that are simple replicas of basic relations without any modifications, as well as punctuation relations. "More than one classification" refers to relations that were classified as result of more than one EUD type in action; "Unclassified" refers to the few relations that could not be correctly classified as one of the six EUD types using our automatic type identification rules.

	gold-dev	% dev	gold-test	% test
Sentences	100	-	100	-
Tokens	2,213	-	2,012	-
EUD Relations	776	-	587	-
Sentences with elliptical predicates	23	23.0%	2	2.0%
1 - Addition of prepositions and conjunctions	397	51.16%	362	61.67%
2 - Assignment of xcomp subjects	44	5.67%	34	5.79%
3 - Propagation of conj head	67	8.63%	56	9.54%
4 - Propagation of conj dependents	45	5.8%	30	5.11%
5 - Annotation of relative pronoun referent	72	9.28%	55	9.37%
6 - Inclusion of elliptical predicate	113	14.56%	11	1.87%
Relations with more than one classification	37	4.77%	31	5.28%
Unclassified relations	1	0.13%	8	1.36%

Table 2. Distribution of phenomena in the gold-standard samples of EUD

Regarding the distribution of EUD types per sample, we see a reasonably large difference between the two, with the frequency of phenomena always being higher in the

dev sample. Particularly in class 6, the difference (14.56% of relations in gold-dev versus 1.87% of relations in gold-test) is due to the fact that the *orphan* relation, indicative of predicate ellipsis, is infrequent in the corpus, as commented before.

Table 3 shows how many times the rules for each EUD types were applied to annotate the gold-standard samples. The difference in applications from "Original" to "Modif." are a result of the changes we made to these rules to make them suit our corpus. The increase from 0 to 60 and 31 in "4 - Propagation of conj dependents" is due to the removal of constraints in the original rules to better suit the Portuguese data. New rules, such as the one in Figure 7, could be classified into one of each EUD types, but were left as a new type to highlight that they are completely new.

	gold-dev		gold-test	
EUD Types	Original	Modif.	Original	Modif.
1 - Addition of prepositions and conjunctions	415	448	355	374
2 - Assignment of xcomp subjects	36	40	27	29
3 - Propagation of conj head	84	87	57	57
4 - Propagation of conj dependents	0	60	0	31
5 - Annotation of relative pronoun referent	108	117	95	95
6 - Inclusion of elliptical predicate	71	75	6	7
7 - New rules	0	75	0	8

Table 3. Number of rule applications for each EUD type

```
rule iobj_vira_suj_do_depxcomp{
  pattern{
    HEADXCOMP -[1=iobj]-> IOBJ;
    IOBJ [upos=PRON,PronType=Prs,Case=Dat];
    HEADXCOMP -[1=xcomp]-> DEPXCOMP;
  }
  without{
    DEPXCOMP -[1=nsubj]-> IOBJ;
  }
  commands{add_edge f:DEPXCOMP -> IOBJ; f.label = "nsubj"; f.enhanced=yes;}
}
```

Figure 7. A new rule, created to annotate sentences such as "*Essa lei permitiu-lhes ganhar um aumento de salário*" (This law allowed them to earn a salary raise), where "lhes" is a pronominal indirect object (IOBJ) of a governor of *xcomp* relation (HEADXCOMP), "permitiu", thus it should gain a new enhanced relation as *nsubj* of the *xcomp* dependent (DEPXCOMP), "ganhar".

Table 4 shows the program's performance considering both samples (test and dev) and both sets of rules (original and modified). ELAS indicates the overall performance of the program. Items 1 to 6 represent the performance, according to the F1-score metric, for each of the six types of EUD. The last line shows the number of sentences where an empty token insertion was made to resolve an ellipsis, but the insertion was incorrectly made. Considering that sentences with ellipses are more challenging to annotate, as they require the empty token inserted into the sentence to be placed in the correct position, and considering that various relations in the sentence may suffer negative impact due

	gold-dev		gold-test	
	Original	Modif.	Original	Modif.
ELAS	61.36%	78.97%	84.67%	96.05%
ELAS (excluding sentences w/ ellipses)	88.50%	99.07%	88.97%	96.05%
1 - Addition of prepositions and conjunctions	93.35%	98.99%	95.17%	98.90%
2 - Assignment of xcomp subjects	85.39%	89.89%	92.54%	97.06%
3 - Propagation of conj head	72.00%	87.94%	84.13%	96.43%
4 - Propagation of conj dependents	84.11%	92.47%	96.67%	96.67%
5 - Annotation of relative pronoun referent	88.28%	100.0%	94.23%	94.23%
6 - Inclusion of elliptical predicate	9.05%	40.71%	0%	100.0%
Sentences with misplaced empty token	21	8	2	0

Table 4. Overall ELAS and by EUD type

to the incorrect placement of this empty token, we calculated two types of ELAS: one considering the entire sample, and another excluding the sentences with predicate ellipses.

Overall, we observe that the numbers are lower in the development sample, reflecting the fact that it contains many more sentences with ellipses than the test sample and that the phenomena were selected for their complexity. The results are superior using the modified rule set, reaching up to 99.07% ELAS for the development sample, excluding sentences with ellipses. For sentences with predicate ellipses, we reduced the number of errors in empty token insertion. In the test sample, errors dropped from 2 to 0, and in the development sample, from 21 to 8. Consequently, in the test sample, the results for relations related to the inclusion of the elliptical predicate reach 100%, but in the development sample, where the sentences are more complex, we only achieve 40.71% ELAS, indicating that there is still room for improvement in particularly difficult sentences.

Comparing the modified and the original rule numbers, the obtained performance improvement is evident. When using the regular data distribution of the treebank as benchmark (test data), where predicate ellipsis is not very frequent, we perform 11.38 absolute ELAS better using the modified rule set in comparison to the original set.

As noted by the Grew team submission to IWPT 2021 (Guillaume and Perrier 2021), the parser's performance heavily relies on the accuracy of the basic syntactic parser. Working with gold UD annotation, the EUD annotation is above 92% ELAS for all languages, being English the one with the highest performance (99.0% ELAS) and Lithuanian the lowest one (92.1%). Our result for Portuguese, in comparison, would be of 96.05% ELAS using the modified rule set.

We observed that labeling the dependency relations between the empty token and the former participants of the *orphan* relation remains particularly challenging. Clues to this can be found in the head clause of *conj*: the available dependency relations are those that exist in the head clause and do not exist in the dependent clause. However, semantically equivalent arguments often have different syntactic forms (for example, a temporal modifier may occur as *advmod*, *obl*, or *advcl*), which makes labeling the dependency relations difficult. The task is computationally complex, and, since the occurrence of this

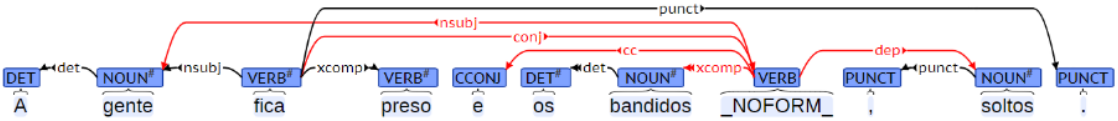

Figure 8. Incorrect EUD annotation of the sentence "We get arrested and the criminals, free" (loose translation).

phenomenon is infrequent, we recommend manually reviewing all relations after insertion of the empty token until we advance in the solutions to improve accuracy.

We noticed that the enhanced dependencies of elliptical token insertion and coreferent annotation (*ref*), because they present an alternative annotation to that of the basic dependencies, constitute a new basis for the other enhanced dependencies. This has two implications: (1) since they constitute a new basis, these two enhanced types must be annotated before the others, and (2) errors in these two enhanced types can generate cascading errors in the other enhanced annotations. For example, in the sentence of Figure 8, when the program does not identify that "bandidos" is the subject of the empty token, the subject slot is empty and the *conj* subject propagation rules annotate "gente" as the subject of the empty token, which is incorrect.[2]

5. Final Remarks

We have addressed the issue of automatic enhanced dependencies annotation for Portuguese, which, to the best of our knowledge, consists in the first attempt for this language. The presented system along with our modified rules has shown its effectiveness in automatically generating complete annotations, which serve as a valuable resource for further linguistic analysis and model training, achieving an overall ELAS of 96.05% over gold basic syntactic annotation.

The next step is to use this system and rules to fully annotate Porttinari, creating the first UD treebank with EUD annotations for Brazilian Portuguese. By leveraging the capabilities of Grew, we aim to provide comprehensive and accurate annotations that include all 6 types of enhanced dependencies, which will be done in batches with human supervision to ensure the dataset quality[3].

More information about this work may be found at the POeTiSA project web portal: `https://sites.google.com/icmc.usp.br/poetisa`

Acknowledgments

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

[2]Further discussion of specific enhancements for Portuguese can be found in the annotation technical report (Duran 2024).

[3]The rules and the data that we used are publicly available at `https://github.com/alvelvis/grew-ed-portuguese`.

References

[Bai et al. 2021] Bai, J., Wang, Y., Chen, Y., Yang, Y., Bai, J., Yu, J., and Tong, Y. (2021). Syntax-BERT: Improving pre-trained transformers with syntax trees. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3011–3020.

[Bölücü et al. 2023] Bölücü, N., Rybinski, M., and Wan, S. (2023). Investigating the impact of syntax-enriched transformers on quantity extraction in scientific texts. In *Proceedings of the Second Workshop on Information Extraction from Scientific Publications*, pages 1–13, Bali, Indonesia.

[Bouma et al. 2020] Bouma, G., Seddah, D., and Zeman, D. (2020). Overview of the iwpt 2020 shared task on parsing into enhanced universal dependencies. In *58th Annual Meeting of the Association for Computational Linguistics*.

[Bouma et al. 2021] Bouma, G., Seddah, D., and Zeman, D. (2021). From raw text to enhanced universal dependencies: The parsing shared task at iwpt 2021. In *Proceedings of the 17th International Conference on Parsing Technologies and the IWPT 2021 Shared Task on Parsing into Enhanced Universal Dependencies (IWPT 2021)*, pages 146–157.

[Candido et al. 2009] Candido, A., Maziero, E., Specia, L., Gasperin, C., Pardo, T., and Aluisio, S. (2009). Supporting the adaptation of texts for poor literacy readers: a text simplification editor for Brazilian Portuguese. In *Proceedings of the Fourth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 34–42, Boulder, Colorado.

[De Marneffe et al. 2014] De Marneffe, M.-C., Dozat, T., Silveira, N., Haverinen, K., Ginter, F., Nivre, J., and Manning, C. D. (2014). Universal stanford dependencies: A cross-linguistic typology. In *LREC*, volume 14, pages 4585–4592.

[Duran et al. 2023] Duran, M., Lopes, L., Nunes, M. G. V., and Pardo, T. (2023). The dawn of the porttinari multigenre treebank: Introducing its journalistic portion. In *Anais do XIV Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana*, pages 115–124, Porto Alegre, RS, Brasil. SBC.

[Duran 2024] Duran, M. S. (2024). Anotação de enhanced dependencies. Disponível em: `https://repositorio.usp.br/item/003209188`. Acesso em: 10 out. 2024.

[Guillaume and Perrier 2021] Guillaume, B. and Perrier, G. (2021). Graph rewriting for enhanced universal dependencies. In *IWPT 2021-17th International Conference on Parsing Technologies*.

[Lin et al. 2021] Lin, Y., Wang, C., Song, H., and Li, Y. (2021). Multi-head self-attention transformation networks for aspect-based sentiment analysis. *IEEE Access*, 9:8762–8770.

[Nivre et al. 2016] Nivre, J., De Marneffe, M.-C., Ginter, F., Goldberg, Y., Hajic, J., Manning, C. D., McDonald, R., Petrov, S., Pyysalo, S., and Silveira, N. (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.

[Nivre et al. 2020] Nivre, J., de Marneffe, M.-C., Ginter, F., Hajic, 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 Twelfth Language Resources and Evaluation Conference*, pages 4034–4043.

[Oliveira et al. 2023] Oliveira, L., Claro, D. B., and Souza, M. (2023). Dptoie: a portuguese open information extraction based on dependency analysis. *Artificial Intelligence Review*, 56(2):7015–7046.

[Pagano et al. 2023] Pagano, A. S., Duran, M. S., and Pardo, T. A. S. (2023). Enhanced dependencies para o português brasileiro. In *Proceedings of the 2nd Edition of the Universal Dependencies Brazilian Festival*, pages 461–470.

[Schuster and Manning 2016] Schuster, S. and 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 (LREC'16)*, pages 2371–2378.

[Shi and Lee 2021] Shi, T. and Lee, L. (2021). TGIF: Tree-graph integrated-format parser for enhanced UD with two-stage generic- to individual-language finetuning. In *Proceedings of the 17th International Conference on Parsing Technologies and the IWPT 2021 Shared Task on Parsing into Enhanced Universal Dependencies (IWPT 2021)*, pages 213–224.

[Zhou et al. 2020] Zhou, J., Zhang, Z., Zhao, H., and Zhang, S. (2020). LIMIT-BERT: Linguistics informed multi-task BERT. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4450–4461.