Genipapo – a Multigenre Dependency Parser for Brazilian Portuguese

Ariani Di Felippo[1,2], Norton T. Roman[3], Bryan K. S. Barbosa[1,2], Thiago A. S. Pardo[1]

[1] Núcleo Interinstitucional de Linguística Computacional (NILC)
Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo (USP)
São Carlos/SP – Brazil

[2]Departamento de Letras, Universidade Federal de São Carlos (UFSCar)
São Carlos/SP – Brazil

[3]Escola de Artes, Ciências e Humanidades, Universidade de São Paulo (USP)
São Paulo/SP – Brazil

ariani@ufscar.br, norton@usp.br, bryankhelven@ieee.org
taspardo@icmc.usp.br

Abstract. *In this article, we present a pioneer effort on building a multigenre parsing model for Brazilian Portuguese. Following the Universal Dependencies framework, we trained a current state-of-the-art model in three corpora from different text genres (journalistic, academic and user-generated content – X posts). Our experiments show that our multigenre parsing model achieves better or competitive results in relation to single-genre trained parsers.*

Resumo. *Neste artigo, apresenta-se um esforço pioneiro para o desenvolvimento de um modelo de parsing multigênero para o português brasileiro. Seguindo o projeto Universal Dependencies, treinou-se um dos modelos do estado-da-arte em três corpora gold-standard de diferentes gêneros textuais (jornalístico, acadêmico e conteúdo gerado por usuário – postagens do X). Os experimentos revelam que nosso modelo multigênero de parsing produz resultados melhores ou competitivos em relação aos modelos de gênero único.*

1. Introduction

Syntactic parsing is the task of automatically uncovering the syntactic relations among the words of a sentence, resulting in syntactic trees, which correspond to one of the first analysis levels in Natural Language Processing (NLP) [Jurafsky and Martin 2024]. This task has proved useful for several different applications, such as text simplification, information extraction, automatic summarization, and sentiment analysis, among many others.

As time goes by, parsing takes different importance degrees and attend different desires. In the beginning, it was common to have parsing as a step in NLP applications (*e.g.* grammar checking [Martins et al. 1998] and text simplification [Candido et al. 2009]). Recent advances in deep learning, distributional models, and language modeling have allowed many applications to forgo deeper linguistic analysis, but current research efforts have indicated that the inclusion of linguistic knowledge during model training or in post-processing steps (*e.g.* in neuro-symbolic approaches) may be relevant for improving results [Zhou et al. 2020, Bai et al. 2021, Lin et al. 2021, Bölücü et al. 2023]. Moreover, given the expensive computational requirements for training the above models and the search for explicability and interpretability, linguistic analysis systems have reemerged as relevant alternatives in several research situations.

There are some well known parsers for Portuguese, including those considered classic, such as PALAVRAS [Bick 2000] and *PassPort* [Zilio et al. 2018], and more recent ones aligned to the *Universal Dependencies* (UD) project [de Marneffe et al. 2021], as *UDPipe 2* [Straka 2018] and the current state-of-the-art *Portparser* [Lopes and Pardo 2024] (with accuracy near 95% for news texts).

We propose here to move a step further in parsing for Brazilian Portuguese (BP). Using the different annotated corpora that are available in the UD initiative, and adopting a widely known parsing framework (the *Stanza* pipeline [Qi et al. 2020]), we investigate the issue of multigenre parsing, aiming at producing a parser that works well for different language writing styles, including short and usually syntactically fragmented X posts (formerly known as tweets), "daily language" of news texts and (supposedly) more refined writing of academic texts. The resulting system, named Genipapo[1] (an acronym for "multiGENre PArser for POrtuguese"), achieves better or competitive results in relation to the single-genre trained parsers, consisting in a step to unleash the potential of Portuguese text analysis tools to work on a wide variety of texts.

The rest of this article is organized as follows. Section 2 introduces the UD framework. Section 3 briefly presents the main related work in the area. The adopted resources and methodology are reported in Section 4, whereas Section 5 presents the results of our experiments. We conclude this article in Section 6.

2. The *Universal Dependencies* framework

UD [Nivre et al. 2020] is currently the most used dependency-based framework of morphological and syntactic analysis in NLP [Sanguinetti et al. 2023]. It is an attempt to standardize the annotation of morphology and syntax, proposing a "universal" annotation strategy for all languages, facilitating the development of multilingual taggers and parsers. At the time of this writing, there are already over 240 treebanks available for more than 150 languages, dealing with a variety of textual genres.

In UD, the following morphology information is considered: (i) Part-of-Speech (PoS) tags, (ii) lemmas, and (iii) features. The syntactic annotation consists of typed dependency relations (*deprels*) between words. Currently, the model has 17 PoS tags and 37 *deprels*, plus a non-fixed set of morphological features. Figure 1 shows an example of an annotated post from the *DANTEStocks* corpus [Di-Felippo et al. 2021]. The basic dependency representation is a tree, where exactly one word is the *head* of the utterance (**root**) (*e.g.* "assina" – "sign"), and all the remaining words depend on some other word. The labeled arcs represent the dependency relations, pointing from *heads* to their dependents. PoS tags, lemmas, and morphological features are displayed below the words in Figure 1.

[1]The corresponding fruit, "Jenipapo" (with 'J' instead of 'G'), is a tropical fruit, appreciated in several states of Brazil and used for different purposes, from painting to eating and preparing beverages. By adopting this inspiration for the name of our parser, we sought this symbolic connection with something rooted in the Brazilian culture and language.

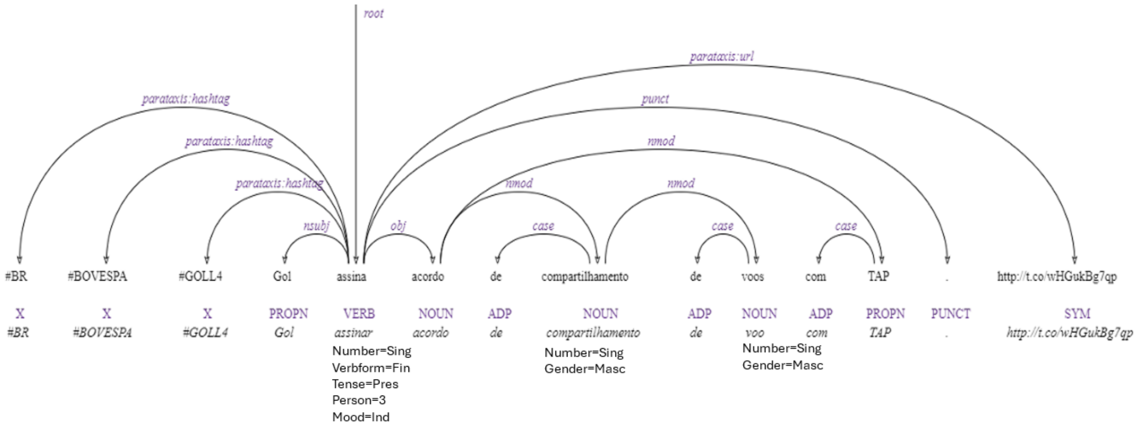

Figure 1. Example of UD morphological and syntactic annotation.

3. Related work

About the linguistic resources for training UD-parsers, there are some available datasets in BP. One of the first corpora with UD annotation for texts in standard (or canonical) Portuguese is the *UD-Portuguese-Bosque* treebank [Rademaker et al. 2017], which contains 210,958 tokens across 9,357 sentences. The Brazilian portion of this corpus consists of 4,213 well-written sentences extracted from journalistic texts. There is also *PetroGold* [Souza et al. 2021], which is a fully revised treebank that consists of academic texts from the oil and gas domain, in a total of 8,946 sentences (and 232,333 tokens). Differently from *UD-Portuguese-Bosque*, *PetroGold* is a specialised or domain-specific corpus. Besides, the UD project makes available the *UD-Portuguese-GSD* corpus [Zeman 2017]. Totaling 12,020 sentences (296,169 tokens) from news texts and blogs, it features two different textual genres, with different degree of canonicalness.

Specifically aiming at growing syntax-based resources for BP, another treebank (with genres beyond newswire texts) has been created. *Porttinari* [Pardo et al. 2021] currently includes two main genres (with others under construction): (i) news texts, representing standard written language, and (ii) user-generated content (UGC), representing informal non-canonical web language (in particular, tweets/X posts).

Concerning parsing models, some dependency UD-parsers are available for BP, specially for news texts. *UDPipe 2* [Straka 2018] is probably the most used model. Using a graph-based biaffine attention architecture, it achieves a *Labelled Attachment Score*[2] (LAS) of 87.04% for news texts. *Stanza* [Qi et al. 2020] is another well-known system, which uses a feature-enriched Bi-LSTM-based deep biaffine neural method. According to the results for the UD version 2.12[3], *Stanza* achieves 87.75% of LAS for news texts. *UDify* [Kondratyuk and Straka 2019] is another important system. It is a semi-supervised multitask self-attention model. There is also the recently released *PortParser* [Lopes and Pardo 2024], which was built by training *UDPipe 2* with *BERTimbau* [Souza et al. 2020] on the *Porttinari-base* corpus [Duran et al. 2023a], which is part of the journalistic portion of the larger *Porttinari*[4] [Pardo et al. 2021] treebank. The model

[2]This score evaluates the output of a parser by considering how many words have been assigned both the correct syntactic *head* and the correct label of the relation [Nivre and Fang 2017].

[3]https://stanfordnlp.github.io/stanza/ performance.html

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

achieved LAS around 95%. This LAS value brings an improvement of around 7% over some well-known existing baselines for standard written Portuguese language.

As a final example, it is important to cite the work of [Zilio et al. 2018]. However delivering lower results than those by more recent works, the authors compared some previous and classical parsing methods for BP. The authors reported that the best model (called *PassPort*) achieved LAS of 85.21% in the UD corpus. In an additional small scale evaluation, the *PassPort* was manually compared to PALAVRAS, using a single corpus of 90 sentences (1,295 tokens), randomly selected from three different genres, to wit, literature, news texts and subtitles. The systems achieved similar results for dependency parsing, with a LAS of 85.02% for *PassPort* against 84.36% for PALAVRAS.

4. Materials and methods

Given the objective of building a multigenre UD parser for BP, three corpora, belonging to three different genres, build our materials.

Our first corpus, *DANTEStocks* [Di-Felippo et al. 2021], comprises 4,048 tweets (with 81,048 tokens) from the stock market domain automatically collected during 2014 (which limits each post to 140 characters). The corpus was built by fetching messages containing a ticker[5] of one of the 73 stocks that composed Ibovespa at that time [da Silva et al. 2020]. *DANTEStocks* presents a combination of standard and non-standard written language, as well as speech marks, domain specific vocabulary and medium (Twitter) features. The dependency relations of the corpus were annotated in two semiautomatic stages [Barbosa 2024]. First, a reference subcorpus of 1,000 tweets was annotated using *UDPipe 2*, which had been trained on *UD-Portuguese-Bosque* and was chosen because it is easily available for use online and offers reliable performance. This subcorpus was then manually revised before being designated as a gold standard. The rest of the corpus was then annotated by customizing Stanza for *DANTEStocks*. We used the combined *Porttinari-base* and reference subcorpus as the initial training set for *Stanza*. The resulting parsing model was used to automatically annotate a new (first) package of data (out of the remaining 3,048 tweets). The first package was manually revised and incorporated to the previous dataset, being used to start a new training run of Stanza. This cycle of training iteration continued incrementally until the last (in a total of 6) package was annotated/revised. Regarding LAS, the final score (6th run) achieved 94.62%, increasing 0.76% from the first run score of 93.86%.

The second corpus, *PetroGold* [de Souza and Freitas 2023], is a gold-standard treebank for the oil and gas (O&G) domain. It integrates the Petrolês corpus, which is a collection of academic and technical documents from public agencies such as Petrobras and "Agência Nacional do Petróleo, Gás Natural e Biocombustíveis" (ANP) [Gomes et al. 2021]. *PetroGold* is composed of 19 academic texts (theses and dissertations), with a total of 9,127 sentences and 253,640 tokens. The syntactic annotation of *PetroGold* also followed a semiautomatic approach. Especifically, four experts were responsable for reviewing the output of a customized version of Stanza, trained on the combination of *UD-Portuguese-Bosque* (v.2.6) and a small collection of data from the O&G domain. Through an intrinsic evaluation using a model created by the UDPipe

[5] A five or six-character alphanumerical string that represents a type of stock from a company, such as "PETR4" for Petrobras' preferred stock.

tool, the corpus achieves 88.53% of LAS. For NLP purposes, the corpus is subdivided into three subsets. The subsets have 7,170, 737 and 1,039 sentences for training (80%), validation (8%) and test (12%).

Our final corpus, *Porttinari-base* [Duran et al. 2023a], is the gold-standard (*i.e.* fully manually annotated and revised) journalistic subcorpus of *Porttinari*, which is composed of 8,418 sentences (168,080 tokens) selected from *Folha de São Paulo* newspaper. The *Porttinari-base* annotation process started with an automatic annotation by *UDPipe 2* using the *UD-Portuguese-Bosque* corpus, which achieved 87% accuracy (in terms of LAS). Next, the dependency relations were manually revised in detail following an annotation manual containing specific guidelines for BP [Duran 2022]. *Porttinari-base* is also subdivided into training, validation and test subsets. The subsets have 5,893, 842 and 1,683 sentences in the train (70%), dev (10%) and test (20%) files, respectively.

For developing our parser, we employed the *Stanza* pipeline, which was trained and evaluated across different corpora. Since both *PetroGold* and *Porttinari-base* corpora already come subdivided in train, validation and test sets, we first set apart their test sets to ensure they would only be used for final evaluation purposes. After this, we unified each corpus' train and validation sets to build a larger training set for each, which was then used in our experiments. Next, we randomly split (from a uniform distribution), *DANTEStocks* in training and test sets, following the same principle of keeping the test set strictly for final testing. Table 1 details each set size across the corpora.

Table 1. Size and proportion of train and test sets across corpora.

Corpus	*Train*		*Test*		*Unit*
	Units	*Proportion*	*Units*	*Proportion*	
DANTEStocks	3,643	90%	405	10%	tweet
UP-Portuguese-PetroGold	7,907	88%	1,039	12%	sentence
Porttinari-base	6,735	80%	1,683	20%	sentence

To assess the model's performance across different genres, we combined the training sets from all three corpora to create a fourth, unified training set, along with a corresponding test set. A grid search was conducted for hyperparameter optimization, focusing on batch size (2000, 3000, 4000, and 5000) and dropout rate (0.2, 0.3, and 0.4), since *Stanza* does not natively support learning rate adjustments. Next, we ran 5-fold cross-validation with grid search (using the above mentioned grid) at each of the four training sets[6], whereby each set was further split in five subsets, with four being used to train the model, and the fifth one being held for validation purposes. This subdivision procedure is repeated five times. We then selected, for each training set, the hyperparameters that produced the highest LAS value across the validation sets during cross-validation.

Having the best set of hyperparameters for each of the four corpora (*DANTEStocks, PetroGold, Porttinari-base* and their union), we retrained the model at each corpus training set, varying its random seed (42, 123, 456, 789 and 101,112), thereby changing the model's innitial configurations. To do so, *PetroGold's* and *Porttinari-base's* training sets were split back into their original training and validation sets, whereas *DANTEStocks'* training set was randomly split into training and validation sets, so that the entire *DANTEStocks* corpus would contain 10% of the data for test, 10% for validation and 80% for

[6]*I.e.* each corpus' individual training set and the largest set built from the union of these training sets.

training purposes. The best performance model, across all seeds, was then selected for each corpus. As a final step, all four models were tested and compared using the previously separated test sets, which had been reserved exclusively for this final evaluation.

5. Results and discussion

Tables 2 and 3 present the results of our model, when trained in each corpus' training set (rows in the tables), and tested at the different test sets of the experiment. Table 2 refers to model results in terms of LAS, whereas Table 3 presents the results in terms of *Unlabelled Attachment Score*[7] (UAS). In the tables, the "Genipapo" lines refer to the model trained in all of the available training sets, *i.e.* our multigenre model, while the "All together" columns refer to the union of all test sets.

Table 2. Model's LAS (%) at each corpus' test set.

Training set	**Test set**			
	Porttinari-base	DANTEStocks	PetroGold	All together
Porttinari-base	94.82	66.10	87.47	88.48
DANTEStocks	87.61	91.95	83.68	86.48
PetroGold	86.74	61.30	**95.33**	87.30
DANTEStocks + Port.-base	94.91	92.67	87.94	91.90
DANTEStocks + PetroGold	87.66	91.85	84.10	86.66
Porttinari-base + PetroGold	94.92	66.75	95.29	91.84
Genipapo	**94.94**	**92.69**	95.11	**94.75**

Table 3. Model's UAS (%) at each corpus' test set.

Training set	**Test set**			
	Porttinari-base	DANTEStocks	PetroGold	All together
Porttinari-base	95.88	75.55	90.38	91.27
DANTEStocks	90.36	93.98	87.51	89.63
PetroGold	89.69	71.45	**95.84**	90.15
DANTEStocks + Port.-base	95.95	94.39	90.97	93.86
DANTEStocks + PetroGold	90.26	93.97	88.04	89.76
Porttinari-base + PetroGold	95.91	76.03	96.00	93.67
Genipapo	**95.97**	**94.42**	95.81	**95.73**

We see that each model trained in isolation produces the best results for its corresponding genre. For instance, considering LAS, training with *Porttinari-base* produced the best results for the test set of *Porttinari-base* (94.82%) and worse results for *DANTEStocks* (66.10%) and *PetroGold* (87.47%). This pattern holds across the genres, where the isolated models consistently perform best when tested on the same genre they were trained on. More interestingly, *Genipapo*, our multigenre parser, outperforms the single-genre trained parsers for 2 of the genres (news texts and X posts), but not for the academic genre. The differences, however, are minimal (less than 1%), suggesting that they could be due to random fluctuation rather than statistically significant differences.

When combining all test sets ("All together" columns in the tables), Genipapo delivers the best results, achieving a 7% improvement in LAS and nearly 5% in UAS

[7]UAS indicates the accuracy of the *head* ignoring the relation's name (deprel) [Nivre and Fang 2017].

compared to the second-best results from single-genre parsers, and a 3% LAS and 2% UAS improvement over parsers trained on pairs. This suggests that Genipapo may be the more suitable choice for processing texts from varied sources, such as diverse web content.

By looking at the results produced by *Genipapo*, when tested on each corpus separately, we see some common mistakes between pairs of *deprels*. One of the most common errors across the three corpora was the confusion between **obl** and **nmod**. This result does not come as a total surprise, since the classification of a nominal as an adverbial adjunct (**obl**) or as a nominal modifier (**nmod**) was already reported in the literature as a challenge for parsing standard Portuguese (and also for humans in some situations), such as in journalistic and academic texts [Duran et al. 2023b, Souza et al. 2021]. Once this phenomenon is also observed in *DANTEStocks'* UGC, this difficulty seems to be unrelated to the degree of "canonicalness" of the corpus. The pairs **acl** (adnominal clause) and **advcl** (adverbial clause) and **obj** (the second argument of a verb) and **nsubj** (a nominal subject) show a relevant confusion only in the standard language corpora. The confusion between **acl** and **advcl** seems to be a case of ambiguity that requires semantic knowledge to be solved, and the confusion between **obj** and **nsubj** occurs when the candidate to the subject is at the right of the verb, since noun phrases at the right can be either object or subject in Portuguese [Duran et al. 2023b].

When comparing the errors of *Genipapo* at each *deprel*, we see the model making a higher number of wrong **root** predictions in *DANTEStocks*, given its error rate of 7.7% against 2.0% in *Porttinari-base* and 0.9% in *PetroGold*. This might be due to the linguistic phenomena of tweets that bring some difficulty to the syntactic annotation of the **root**. Another interesting observation relates to **parataxis**, which is one the the most frequent tag in our UGC corpus (708 cases), but not in the remaining corpora. The relatively low error rate in *DANTEStocks* (9.3%) indicates that this *deprel* has been well learned by *Genipapo* in UGC. Moreover, we could see that the *deprel* tags most wrongfully predicted due to under representation in *Porttinari-base* and *DANTEStocks* are the same: **reparandum, dislocated** and **orphan**. The first two tags do not occur in *PetroGold*, and the only two occurrences of **orphan** in this corpus were wrongly predicted.

As a way to compare *Genipapo*'s performance with that by a state-of-the-art model, we also run *Portparser* in the same testing sets as *Genipapo* (Table 4). We note that the training, validation, and test splits of the *Porttinari-base* used by *Portparser* differ from those publicly available and used in our experiments with *Genipapo*. This discrepancy means that some sentences present in the test sets of *Porttinari-base* and the unified set (All together) may have been included in the training or validation sets of *Portparser*, artificially boosting its LAS and UAS scores. Despite this, *Genipapo* outperformed *Portparser* across all testing sets except for the *Porttinari-base* test set. In terms of LAS, *Genipapo* showed significant improvements over *Portparser* on the *DANTEStocks* test set (92.69% vs. 64.45%), the *PetroGold* test set (95.11% vs. 86.74%), and the combined test set (94.75% vs. 89.51%). However, *Portparser* performed better on the *Porttinari-base* test set (98.06% vs. 94.94%). The same pattern is observed in UAS scores, where *Genipapo* outperformed *Portparser* on *DANTEStocks* (94.42% vs. 75.81%), *PetroGold* (95.81% vs. 90.50%), and the combined test set (95.73% vs. 92.62%). Nevertheless, *Portparser* achieved higher UAS on *Porttinari-base* (98.58% vs. 95.97%).

Table 4. *Portparser*'s LAS and UAS at each corpus' test set.

Test set	LAS (%)	UAS (%)
Porttinari-base	98.06	98.58
DANTEStocks	64.45	75.81
PetroGold	86.74	90.50
All together	89.51	92.62

6. Final remarks

In this paper, we introduced Genipapo, a multigenre UD-parser for Brazilian Portuguese, and showed that it had better or competitive performance in relation to genre specific trained parsers. Future work includes (i) to extend *Genipapo*'s training to other genres and domains, such as audio transcriptions, literary texts, and tweets related to the COVID-19 pandemic, whose corpora are still under construction, and (ii) to explore different parsing strategies and pipelines.

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

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 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, with in the scope of PPI-SOFTEX, coordinated by Softex and published as Residence in TIC13, DOU01245.010222/2022-44.

References

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 EACL*, p. 3011–3020.

Barbosa, B. K. d. S. (2024). Descrição sintático-semântica de nomes predicadores em tweets do mercado financeiro em português. Master's thesis, Programa de Pós-Graduação em Linguísica, Universidade Federal de São Carlos.

Bick, E. (2000). *The Parsing System "Palavras". Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework.* University of Arhus, Århus.

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 2nd Workshop on Information Extraction from Scientific Publications*, p. 1–13, Bali.

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 4th Workshop on Innovative Use of NLP for Building Educational Applications*, p. 34–42, Boulder, Colorado.

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

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

de Souza, E. and Freitas, C. (2023). Explorando variações no tagset e na anotação universal dependencies (ud) para português: Possibilidades e resultados com base no treebank petrogold. In *Anais do XIV Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana*, p. 125–134, Porto Alegre, RS, Brasil. SBC.

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*, p. 335–343, Porto Alegre, RS, Brasil. SBC.

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

Duran, M., Nunes, M. d. G. V., and Pardo, T. A. S. (2023b). Construções sintáticas do português que desafiam a tarefa de parsing: uma análise qualitativa. In *Proceedings of the 2nd Universal Dependencies Brazilian Festival (UDFest-BR)*, p. 424–433, Porto Alegre, RS, Brasil. SBC.

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).

Gomes, D. S. M., Cordeiro, F. C., Consoli, B. S., Santos, N. L., Moreira, V. P., Vieira, R., Moraes, S., and Evsukoff, A. G. (2021). Portuguese word embeddings for the oil and gas industry: Development and evaluation. *Computers in Industry*, 124:103347.

Jurafsky, D. and Martin, J. H. (2024). *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition with Language Models*. 3rd edition. Online manuscript released August 20, 2024.

Kondratyuk, D. and Straka, M. (2019). 75 languages, 1 model: Parsing Universal Dependencies universally. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, p. 2779–2795, Hong Kong, China. ACL.

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.

Lopes, L. and Pardo, T. (2024). Towards portparser - a highly accurate parsing system for Brazilian Portuguese following the Universal Dependencies framework. In *Proceedings of the 16th International Conference on Computational Processing of Portuguese - Vol. 1*, p. 401–410, Santiago de Compostela, Galicia/Spain. ACL.

Martins, R. T., Hasegawa, R., Nunes, M. d. G. V., Montilha, G., and Oliveira, O. N. (1998). Linguistic issues in the development of regra: A grammar checker for brazilian portuguese. *Natural Language Engineering*, 4(4):287–307.

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 Twelfth Language Resources and Evaluation Conference*, p. 4034–4043, Marseille, France. ELRA.

Nivre, J. and Fang, C.-T. (2017). Universal Dependency evaluation. In *Proceedings of the NoDaLiDa 2017 Workshop on Universal Dependencies (UDW 2017)*, p. 86–95, Gothenburg, Sweden. ACL.

Pardo, T., Duran, M., Lopes, L., Felippo, A., Roman, N., and Nunes, M. (2021). Porttinari - a large multi-genre treebank for brazilian portuguese. In *Proceedings of the XIII Brazilian Symposium in Information and Human Language Technology*, p. 1–10, Porto Alegre, RS, Brasil. SBC.

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

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)*, p. 197–206, Pisa, Italy. Linköping University Electronic Press.

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

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

Souza, F., Nogueira, R., and Lotufo, R. (2020). Bertimbau: Pretrained bert models for brazilian portuguese. In *Intelligent Systems*, p. 403–417, Cham. Springer International Publishing.

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*, p. 197–207, Brussels, Belgium. ACL.

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

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*, p. 4450–4461.

Zilio, L., Wilkens, R., and Fairon, C. (2018). Passport: A dependency parsing model for portuguese. In *Computational Processing of the Portuguese Language*, p. 479–489, Cham. Springer International Publishing.