Polishing the gold – how much revision do we need in treebanks? ⋆

Elvis de Souza and Cláudia Freitas

Pontifical Catholic University of Rio de Janeiro
elvis.desouza99@gmail.com,
claudiafreitas@puc-rio.br

Abstract. We present the second version of PetroGold, a gold-standard treebank for the oil & gas domain in the Portuguese language. The corpus went through a series of revisions guided by three methods tested in the literature: inter-annotator disagreement, inconsistent n-grams and verification rules. We perform an intrinsic evaluation and the model scores 90.92%, 89.09% and 84.07% in the UAS (unlabeled attachment score), LAS (labeled attachment score) and CLAS (content-word labeled attachment score) metrics respectively, CLAS being 1.11% higher than in the first version. We perform an experiment where we verify a negative impact in the intrinsic evaluation when simplifying the annotation related to prepositional verbal arguments and we conclude by discussing the results and future work.

Keywords: Natural Language Processing · Language resources · Corpora reviewing · Treebank

1 Introduction

Annotated corpora are important resources for natural language processing. On the one hand, data-driven NLP approaches use corpora as a learning source for linguistic analysis; on the other hand, approaches based on rules, or oriented by specific knowledge of language, can use it as material to evaluate the results of their analyses. Despite its importance, the number of golden treebanks in Portuguese, with texts from genres other than the journalistic one, still falls short, making it difficult to advance certain NLP tasks for Portuguese, such as information extraction and parsing in diverse domain areas.

⋆ This paper was partially funded by the National Agency for Petroleum, Natural Gas and Biofuels (ANP), Brazil, associated with the investment of resources from the R, D & I Clauses, through a Cooperation Agreement between Petrobras and PUC-Rio. We would like to thank the team at the Applied Computational Intelligence Laboratory (ICA) at PUC-Rio for the generation of morphosyntactic annotation models trained in Stanza, and Elvis de Souza thanks the National Council for Scientific and Technological Development (CNPq) for the Masters scholarship process no. 130495/2021-2.

Some of the main machine-learning-based parsers available to the Portuguese-speaking community (e.g. spaCy [10], UDPipe [24] and Stanza [17]) use the Bosque-UD [18] corpus as training material, achieving a performance of up to 87.81% in the attachment of dependencies, according to the CoNLL 2018 Shared Task [28]. The corpus is composed of morphosyntactically annotated journalistic texts and is part of the Universal Dependencies [14] project, representing a valuable resource for several NLP tasks in general domain texts.

However, tasks that demand the processing of texts from specific domains will face difficulties due to the lack of available training material of diverse domains – fortunately, this scenario is changing with the creation of projects like Porttinari [15]. [25] indicates that, for the English language, a model trained in the Wall Street Journal Treebank sees its performance drop more than 10% when applied in the biomedical domain. Similarly, [6] reports that systems trained with general domain texts do not perform well when applied to academic texts.

In this paper, we present a second version of PetroGold, a gold-standard treebank with texts from the oil & gas domain in the academic genre. The corpus is available at the project webpage[1] and contains 8,949 sentences (250,595 tokens). More than providing an improved version of the material, which includes a systematic treatment of tags related to verbal subcategorization and grammatical multiword expressions, this second round of review aims to (i) evaluate the contribution of different treebank review methods in a robust corpus, offering subsidies for an evaluation of the methodology described in [8], and (ii) evaluate how much the corrections carried out in this second stage impact the performance of language models. In the end, we assess how much differentiating adverbial adjuncts from prepositional verbal arguments impacts the performance of a parser.

2 Treebanks

Syntactically analyzed corpora are called treebanks because syntactic analyses give a hierarchical character to sentences using constituents (in a syntagmatic model) or dependents (in a dependency model). From the point of view of linguistic studies, treebanks are informative about the structure of the language in use, serving as a database for the development of linguistic theories, either as a means of testing them or in carrying out statistical studies. From a NLP point of view, treebanks serve as training and evaluation material, in addition to being the basis for subsequent tasks, such as Open Information Extraction [9].

There are many possible differences between treebanks, some related to the methodology for building the corpus – annotated entirely from scratch or, more commonly in recent times, automatically annotated and revised by linguists – and differences related to the syntactic categories, to the grammar model and others. The pioneering English-language corpora, Penn Treebank [12] and SUSANNE [19], made their syntactic annotation available through constituents;

[1] https://petroles.puc-rio.ai

the Prague Dependency Treebank [4], in turn, uses a syntactic dependencies format.

For the Portuguese language, Linguateca [20] has been dedicated for a long time to the creation of Floresta Sintá(c)tica ([1], [7]), a pioneering project for the construction of treebanks for the Portuguese language. Floresta obtained its morphosyntactic annotation from the automatic analyzer PALAVRAS [3] and it is composed of four parts, which differ in terms of modality (written or spoken) and degree of revision. Bosque is a subset of Floresta, is fully revised by linguists, and it is precisely the revision dimension that made (and makes) Bosque a valuable resource, which is reflected in its conversion to different formats, such as Bosque-UD [18].

3 Building PetroGold v2

PetroGold is composed of 8,949 sentences (250,595 tokens) from 19 theses and dissertations of the oil & gas domain processed in full: only elements such as summary, abstract, appendices and bibliographic references were excluded, as well as figures, graphs, formulas and tables. The corpus was annotated using the Universal Dependencies framework. However, since issues related to the academic genre and the specific technical domain are not covered in the project's annotation guidelines, we needed to discuss how to carry out the analysis of the typical linguistic structures of the corpus. Some of the new problems that have arisen since the release of the first version of the corpus, in addition to the methodology and tools used in the development of this new version will be discussed in this section.

3.1 Annotation challenges

The second version of PetroGold brings at least three major improvements related to the annotation of grammatical multiword expressions, verbal lemmatization and verbal subcategorization.

First, we standardized the annotation of grammatical multiword expressions (MWEs) such as *de acordo com* ("according to"), *por sua vez* ("in turn") and *tendo em vista* ("in view of"), which receive the *fixed* dependency relation. We used as a criterion the recognition of combinations as grammatical phrases (prepositional, conjunctive, adverbial) in Portuguese language grammars and the difficulty of dealing with the combination in a transparent manner, both at the part-of-speech and in the syntactic level. A complete listing of these 227 expressions, which as a whole occur 2,333 times in the corpus, can be found in the documentation accompanying the corpus.

Another improvement is related to the lemmatization of verbs. Since PetroGold was originally annotated by a system trained using a journalistic corpus, many of the verbs specific to the academic genre and the oil & gas domain were not correctly identified by the model, such as *adsorver* ("to adsorb"), lemmatized as *adsorvir* and *absorver* ("to absorb"), lemmatized as *absorvar* – both

very common in the technical domain. In this second phase, we performed a manual verification of all verbal lemmas in the corpus, resulting in 212 corrected lemmas, which occur 621 times in the corpus.

This version also features many corrections regarding adjuncts and verbal arguments, a topic which is thoroughly discussed in [22]. The grammatical guidelines of the Universal Dependencies project for this issue follow the direction of [27]: given the difficulties of distinguishing argument and adjunct already known and reported in the literature ([11], [16], [26], [2]) – and that difficulties are common in corpus annotation at least for most of the languages that make up the project – the project chooses to (partially) shift the discussion to another place: the idea is not to distinguish argument from adjunct, but between the core and the oblique terms.

In short, when related to verbal subcategorization, the core terms are not introduced by preposition and the tags are *obj* and *iobj* – the latter only used with arguments that are oblique pronouns – and oblique terms are preceded by a preposition and the tag is *obl*). However, UD also allows us to annotate a subspecification of the oblique, *obl:arg*, when, in addition to being prepositional, the phrase is also an argument of the verb, if we find the distinction to be important.

While analyzing *obl* and *obl:arg* in PetroGold, we do not seek to characterize arguments based on the transitivity of the verb, but we prioritize the meaning of the prepositional phrase – if it expresses meanings traditionally associated with adverbials (time, place, manner, purpose, causality, conformity etc), we annotate as *obl*, while, in the absence of an adverbial semantics, we analyze as *obl:arg*.

3.2 Methodology

The first version of PetroGold had four annotators working 20 hours a week, for three months, dedicated to reviewing the corpus. In this second version, we had three of the annotators working 20 hours a week for two months. All annotators had previously familiarized themselves with both the UD approach and the type of text that makes up the corpus.

The corpus was originally annotated using a customized Stanza model, which was trained using Bosque-UD plus a small portion of sentences from other texts of the domain with totally revised annotation[2]. The inter-annotator agreement in the human review of the automatic annotation was 95.1% using the κ (*kappa*) metric for the pair of annotators that obtained the highest degree of agreement in the syntactic dependency analysis task, while the worst performing pair obtained 91.9% agreement.

The first version of the corpus used as a review strategy the analysis of confusion matrices, which contrast the analysis of two different parsers, Stanza and UDPipe, in such a way that the divergences between both systems are indicative of possible errors in one of the systems or both, requiring human intervention to choose the correct analysis[3]. This strategy, which we call IAD (Inter-Annotator

[2] In this training material, Bosque-UD represented 93% of the total size.

[3] Since both analysis systems can perform tokenization and sentence segmentation in different ways, we gave UDPipe the corpus already segmented by Stanza.

Disagreement), allows the analysis of errors by clusters of confusion between parsers, making it easier for annotators to detect error patterns and, consequently, to develop different correction rules.

For the second version of PetroGold, we applied the revision strategy schematized in [8], which consists, in addition to the IAD method, in the verification of inconsistent n-grams and the application of general verification rules.

Inconsistent n-grams is a method proposed by [5] and adapted to UD by [13]. The underlying idea is that a pair of dependent lemmas, if repeated in the corpus, must have the same annotation in all occurrences, otherwise it is indicative of annotation inconsistency. For example, in sentences (3) and (4), the same pair of lemmas (**Arai** and **1990**) was analyzed differently in two sentences: in the first, the analysis is of a composite proper name (*flat:name*), and in the second, the analysis is of an adnominal adjunct. Bibliographic references that contain the publication year have the relation between the date and the proper name analyzed as *nmod*, so sentence (3) needed to be corrected to become consistent with the analysis of (4), which is the correct one.

(3) ***flat:name*** – Da mesma forma, **Arai** & Coimbra (**1990**) interpretam que o paleoambiente do Membro Romualdo (...)[4]

(4) ***nmod*** – Desta forma, a característica geral da associação fossilífera encontrada por **Arai** & Coimbra (**1990**) não deixa dúvidas quanto à pertinência dos registros das ingressões marinhas no Andar Alagoas.[5]

Differently from previous authors, we did not require that the words in the context of the pairs should be the same in order to look for divergent pairs annotation because it lowered the method recall.

The other method, a rule-based verification approach, consists of search expressions created to detect errors in the corpus, whether referring to inconsistencies regarding the UD format or the Portuguese annotation.

For example, the comma in (5) after the expression *a seguir* ("next") depended on the verb *seguir* ("to follow"); however, since it is a multiword expression (*fixed*), the comma should depend on the head of the expression, "a", a restriction of the UD model. In (6), the occurrence of a verb in the participle, *denominada* ("denominated"), with a verb *ser* ("to be") depending on it, is typical of passive voice; thus, *ele* ("he"), which is introduced by the preposition *por* ("by"), is a common form of agent of the passive voice in Portuguese, although it had not been analyzed in this way.

(5) **A seguir,** são apresentadas as etapas e a metodologia que foi adotada no trabalho.[6]

[4] Transl. "Similarly, **Arai** & Coimbra (**1990**) interpret that the paleoenvironment of the Romualdo Member (...)"

[5] Transl. "In this way, the general characteristic of the fossiliferous association found by **Arai** & Coimbra (**1990**) leaves no doubt as to the relevance of the records of marine ingressions in the Andar Alagoas."

[6] Transl. "**Next,** the steps and methodology adopted in the work are presented."

(6) Esta zona de falha foi por **ele** denominada "Zona de Transferência de o Funil".[7]

A list with all 61 rules can be found on our GitHub page[8].

In order to evaluate the impact of the revisions, in section 4 we compare three different learning scenarios: (a) the first version against this second revised version, and (b) the second version, which has the *obl:arg* annotation, against the same corpus when the tag is converted to *obl*, simulating what the UD guidelines first suggest.

3.3 Tools

The review was performed using ET, a tool that enables querying, editing and evaluating annotated corpora [21]. ET is divided into Interrogatório, an interface where we search for the most frequent errors and correct their annotation using the correction rules that we developed during the review process, and Julgamento, an interface where we evaluate the linguistic annotation according to the aforementioned methods: inter-annotator disagreement, inconsistent n-grams and verification rules.

Besides seeing the annotation from both a quantitative and a qualitative perspective (for instance, the main tags involved in annotation errors), reading the corpus through the lens of Julgamento provides us with a picture of strengths and weaknesses of the annotation. This picture, in turn, guides us back to Interrogatório: we can search for the same sentences pointed out in the review methods and make corrections manually or in batch, when applicable, to make the review more efficient.

4 Results

PetroGold v2 is slightly smaller than the first version due to some sentences that were suppressed because of incorrect segmentation in the pre-processing stage. Table 1 indicates the differences in the characteristics of both versions of the corpus.

In this second version, the number of tokens corrected since the original annotation from Stanza (summing versions 1 and 2) reached 21,634 – 8.6% of all tokens needed correction –, resulting in 74% of sentences which had at least one token modified by the annotators.

Regarding the review methods described in Section 3.2, Figure 1 illustrates the contribution of each of them in the review process. The figure is an estimation of the relative number of errors found by each method because, since two or more methods can indicate the same token as an annotation mistake, the number of errors found by all methods, when summed, exceed 100% of the corrected tokens.

[7] Transl. "This fault zone was called by **him** 'the Funnel Transfer Zone'".

[8] Available at: https://github.com/alvelvis/ACDC-UD/blob/master/validar_UD.txt. Accessed on 15 Jan. 2022.

	v2	v1
Tokens	250,595	253,640
Corrections	8,802	12,832
Words	221,208	223,707
Sentences	8,949	9,127
Documents	19	19

Table 1. PetroGold features across versions

The most productive method for identifying errors was IAD, which sums up 51.4% of detected errors (11,137 tokens). The general correction rules, in turn, totaled 10.1% (2,202 tokens), while the inconsistent n-grams indicated 9.2% of the corrected errors (2,003 tokens). None of the methods was able to identify 37.8% of the errors (8,188 of the tokens), which were found by the annotators when reading the sentences in the treebank.

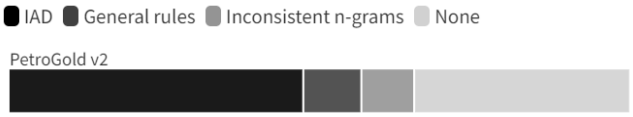

Fig. 1. Methods contribution

From the reviewing process point of view, the IAD method spots the highest number of annotation mistakes. Since we previously showed [8] that this approach also achieves the best F1 among the revision methods (49.7%, 14.4% and 4.8% for IAD, rule-based and inconsistent n-grams, respectively), choosing only IAD is a possibility to be considered when there is not much time or resources to build a revised treebank.

To check the material's consistency, we trained a UDPipe model using the tool's default parameters and PetroGold v2 as the training material. We used the same set of sentences from the experiment performed in [23] in the train and test partitions to allow comparisons between the results from both versions 1 and 2 of the treebank, following the proportion of 95% and 5% of sentences in each partition, respectively. Previously, the performance of the model was compared against the results of a model trained using Bosque-UD, which has a similar size – at that time, PetroGold v1 achieved up to 9% better metrics than the journalistic corpus [23]. This time, we compared the second version of PetroGold, with all the corrections reported, with the first version of the corpus.

As seen in Table 2, the results for PetroGold v2 are not very different from those of the first version with regard to the lemmatization task (LEMMA), part-of-speech assignment (UPOS), syntactic dependency attachment (UAS), classification of dependencies (LAS) and classification of dependencies for content

words (CLAS)[9]. We see a slight improvement in all metrics: LEMMA (0.06%), UPOS (0.21%), UAS (0.27%), LAS (0.56%) and CLAS (1.11%).

Version	LEMMA (%)	UPOS (%)	UAS (%)	LAS (%)	CLAS (%)
v2	98.54	98.40	90.92	89.09	84.07
v1	98.48	98.19	90.65	88.53	82.96

Table 2. Intrinsic evaluation of PetroGold models trained using UDPipe

The results show a modest increase in consistency, suggesting that intrinsic evaluation will not be very sensitive to corpora reviews when they were already internally consistent. The main increase is on CLAS measure, and this can be due to the large review applied to prepositional verbal arguments and adjuncts, which are content words.

Finally, we compared the PetroGold v2 intrinsic evaluation results with the results from the same corpus having converted the *obl:arg* tags to *obl*. In this version, which we call "No *obl:arg*", there is no distinction between adverbial adjuncts and prepositional objects, so that every verb-dependent prepositional phrase receives the tag *obl*, as originally proposed in [27]. It is a modification that affected 1,488 tokens, which are present in 14.8% of sentences.

	LEMMA (%)	UPOS (%)	UAS (%)	LAS (%)	CLAS (%)
No *obl:arg*	98.54	98.40	90.66	88.82	83.48

Table 3. Model evaluation when *obl:arg* is converted to *obl*

The results in Table 3 indicate a drop in all metrics related to dependency analysis (UAS, LAS and CLAS), with emphasis on CLAS, whose performance drop was 0.59%. At first sight, the results seem counter-intuitive, as the *obl:arg* tag represents a semantically oriented analysis, thus a more difficult one – the same phrase introduced by preposition can receive either one tag or another, depending on the meaning of the phrase content words, while the simplified version would be analyzed one way or another based only on the presence of a preposition. However, maintaining this granularity – the distinction between argument and adjunct in prepositional phrases – facilitated the generalization of the system as a whole, indicated by the decrease in all metrics when the distinction is undone.

The results by each dependency relations show that the *obl:arg* is a difficult one (only 62.8% hits) and that the *obl* relation is best learned when we convert all the *obl:arg* relations to *obl* (86.4% against 79.9%). However, many other

[9] Metrics were gathered from [28].

dependency relations were best learned when we had the *obl:arg* relation in the treebank, such as *ccomp* (65.5% with *obl:arg* against 58.6% without it) and *acl:relcl* (95.6% against 93.4%). While there is no direct explanation for the dependency relations hit increase with *obl:arg*, it justifies the overall better F1, which adds up to the many arguments in favor of keeping this annotation in the corpus.

5 Concluding remarks

We presented a small study on treebank revision methods, based on PetroGold corpus. While the most productive review method spots around half of the annotation mistakes, almost 40% of them can not be detected by any method. Besides presenting PetroGold v2, we performed an intrinsic evaluation of annotation consistency, and compared PetroGold v2 against PetroGold v1, which resulted in a 1.11% increase in CLAS. We concluded by confirming that an intrinsic evaluation might not be sensitive to improvements in robust corpora that have previously been reviewed, in spite of the importance of improving some specific annotations for different reasons. Furthermore, we verified that removing the distinction between adverbial adjuncts and prepositional verbal arguments has a slight negative impact in the automatic learning of dependencies.

The first application of PetroGold will be the creation of a morphosyntactic annotation model suitable for texts in the oil & gas domain. The goal is to use this customized model to annotate new texts in the domain in order to proceed with a semantic annotation of named entities, increasingly expanding the coverage of Portuguese NLP directed to specific domains.

References

1. Afonso, S., Bick, E., Haber, R., Santos, D.: Floresta sintá(c)tica: a treebank for Portuguese. In: Rodrigues, M.G., Araujo, C.P.S. (eds.) Proceedings of the Third International Conference on Language Resources and Evaluation (LREC 2002). pp. 1698–1703. ELRA, Paris (29-31 de Maio 2002), http://www.linguateca.pt/documentos/AfonsoetalLREC2002.pdf
2. Bagno, M.: Gramática pedagógica do português brasileiro. Parábola Ed. (2012)
3. Bick, E.: Palavras, a constraint grammar based parsing system for portuguese. Working with Portuguese corpora pp. 279–302 (2014)
4. Böhmová, A., Hajič, J., Hajičová, E., Hladká, B.: The prague dependency treebank. In: Treebanks, pp. 103–127. Springer (2003)
5. Boyd, A., Dickinson, M., Meurers, W.D.: On detecting errors in dependency treebanks. Research on Language and Computation **6**(2), 113–137 (2008)
6. Cohen, K.B., Verspoor, K., Fort, K., Funk, C., Bada, M., Palmer, M., Hunter, L.E.: The colorado richly annotated full text (craft) corpus: Multi-model annotation in the biomedical domain. In: Handbook of Linguistic Annotation, pp. 1379–1394. Springer (2017)
7. Freitas, C., Rocha, P., Bick, E.: Floresta Sintá(c)tica: Bigger, Thicker and Easier. In: Teixeira, A., de Lima, V.L.S., de Oliveira, L.C.,

Quaresma, P. (eds.) Computational Processing of the Portuguese Language, 8th International Conference, Proceedings (PROPOR 2008). vol. Vol. 5190, pp. 216–219. Springer Verlag (8-10 de Setembro 2008), http://www.linguateca.pt/documentos/FreitasRochaBickPROPOR08Poster.pdf
8. Freitas, C., de Souza, E.: A study on methods for revising dependency treebanks: In search of gold (2022), submitted
9. Gamallo, P., Garcia, M., Fernández-Lanza, S.: Dependency-based open information extraction. In: Proceedings of the joint workshop on unsupervised and semi-supervised learning in NLP. pp. 10–18 (2012)
10. Honnibal, M., Johnson, M.: An improved non-monotonic transition system for dependency parsing. In: Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. pp. 1373–1378. Association for Computational Linguistics, Lisbon, Portugal (September 2015), https://aclweb.org/anthology/D/D15/D15-1162
11. Manning, C.D.: Probabilistic syntax. Probabilistic linguistics **289341** (2003)
12. Marcus, M., Santorini, B., Marcinkiewicz, M.A.: Building a large annotated corpus of english: The penn treebank (1993)
13. de Marneffe, M.C., Grioni, M., Kanerva, J., Ginter, F.: Assessing the annotation consistency of the Universal Dependencies corpora. In: Proceedings of the Fourth International Conference on Dependency Linguistics (Depling 2017). pp. 108–115. Linköping University Electronic Press, Pisa,Italy (Sep 2017), https://www.aclweb.org/anthology/W17-6514
14. de Marneffe, M.C., Manning, C.D., Nivre, J., Zeman, D.: Universal dependencies. Computational linguistics **47**(2), 255–308 (2021)
15. Pardo, T., Duran, M., Lopes, L., Felippo, A., Roman, N., Nunes, M.: Porttinari - a large multi-genre treebank for brazilian portuguese. In: Anais do XIII Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana. pp. 1–10. SBC, Porto Alegre, RS, Brasil (2021). https://doi.org/10.5753/stil.2021.17778, https://sol.sbc.org.br/index.php/stil/article/view/17778
16. Przepiórkowski, A., Patejuk, A.: Arguments and adjuncts in universal dependencies. In: Proceedings of the 27th International Conference on Computational Linguistics. pp. 3837–3852 (2018)
17. Qi, P., Zhang, Y., Zhang, Y., Bolton, J., Manning, C.D.: 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 (2020), https://nlp.stanford.edu/pubs/qi2020stanza.pdf
18. Rademaker, A., Chalub, F., Real, L., Freitas, C., Bick, E., de Paiva, V.: Universal dependencies for portuguese. In: Proceedings of the Fourth International Conference on Dependency Linguistics (Depling 2017). pp. 197–206 (2017)
19. Sampson, G.: English for the computer: The susanne corpus and analytic scheme (2002)
20. Santos, D., Simões, A., Frankenberg-Garcia, A., Pinto, A., Barreiro, A., Maia, B., Mota, C., Oliveira, D., Bick, E., Ranchhod, E., et al.: Linguateca: um centro de recursos distribuído para o processamento computacional da língua portuguesa (2004)
21. de Souza, E., Freitas, C.: ET: A workstation for querying, editing and evaluating annotated corpora. In: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations. pp. 35–41. Association for Computational Linguistics, Online and Punta Cana, Dominican Republic (Nov 2021). https://doi.org/10.18653/v1/2021.emnlp-demo.5, https://aclanthology.org/2021.emnlp-demo.5

22. de Souza, E., Freitas, C.: Still on arguments and adjuncts: the status of the indirect object and the adverbial adjunct relations in universal dependencies for portuguese. In: Proceedings of the I Universal Dependencies Brazilian Festival (UDFest-BR) (2022)
23. Souza, E., Silveira, A., Cavalcanti, T., Castro, M., Freitas, C.: Petrogold – corpus padrão ouro para o domínio do petróleo. In: Anais do XIII Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana. pp. 29–38. SBC, Porto Alegre, RS, Brasil (2021). https://doi.org/10.5753/stil.2021.17781, https://sol.sbc.org.br/index.php/stil/article/view/17781
24. Straka, M., Hajic, J., Straková, J.: 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). pp. 4290–4297 (2016)
25. Thompson, P., Ananiadou, S., Tsujii, J.: The genia corpus: Annotation levels and applications. In: Handbook of Linguistic Annotation, pp. 1395–1432. Springer (2017)
26. Vilela, Mário; Koch, I.V.: Gramática da língua portuguesa: Gramática da Palavra, Gramática da Frase, Gramática do Texto/Discurso. Almedina (2001)
27. Zeman, D.: Core arguments in universal dependencies. In: Proceedings of the fourth international conference on dependency linguistics (DepLing 2017). pp. 287–296 (2017)
28. Zeman, D., Hajic, J., Popel, M., Potthast, M., Straka, M., Ginter, F., Nivre, J., Petrov, S.: 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. pp. 1–21 (2018)