Universal Grammatical Dependencies for Portuguese with CINTIL Data, LX Processing and CLARIN support

António Branco, João Ricardo Silva, Luís Gomes and João Rodrigues
University of Lisbon
NLX – Natural Language and Speech Group, Department of Informatics
Faculdade de Ciências, Campo Grande, 1749-016 Lisboa, Portugal
{ambranco,jrsilva,lmdgomes,jarodrigues}@fc.ul.pt

Abstract

The grammatical framework for the mapping between linguistic form and meaning representation known as Universal Dependencies relies on a non-constituency syntactic analysis that is centered on the notion of grammatical relation (e.g. Subject, Object, etc.). Given its core goal of providing a common set of analysis primitives suitable to every natural language, and its practical objective of fostering their computational grammatical processing, it keeps being an active domain of research in science and technology of language. This paper presents a new collection of quality language resources for the computational processing of the Portuguese language under the Universal Dependencies framework (UD). This is an all-encompassing, publicly available open collection or mutually consistent and inter-operable scientific resources that includes reliably annotated corpora, top-performing processing tools and expert support services: a new UPOS-annotated corpus with 675K tokens and a new UD treebank with nearly 38K sentences; an UPOS tagger and an UD parser trained on these corpora, available both as local stand-alone tools and as remote web-based services; and helpdesk support ensured by the Knowledge Center for the Science and Technology of Portuguese of the CLARIN research infrastructure.

Keywords: dependency treebank, Universal Dependencies, Portuguese, dependency parsing, UD parser, UD tagger, PORTULAN CLARIN

1. Introduction

Grammatical analysis provides for the mapping between the surface form of natural language expressions and some representation of their syntax and semantics. It has been at the core of Natural Language Processing both as an essential support or just as an instrumental procedure for many language processing tasks and applications.

Among the grammatical representations delivered by different possible approaches to grammatical parsing, of differing depth in the mapping between form and meaning, the graph-based representation of grammatical functions or dependencies (e.g. Subject, Object, etc.) among words in a given expression has become mainstream, most likely because of its success in capturing core relations in the interface between syntax and compositional semantics.

And among approaches to grammatical dependencies, of different theoretical persuasions and empirical adequacy, the so called Universal Dependencies format has been attracting increasing attention since its inception as, among other factors, it was adopted and promoted by a big tech company (de Marneffe et al., 2021). Given the specific grammars of the different languages and thus the language specific aspects of their grammatical analysis, there being an annotation format for grammatical analysis that is applied to an increasing number of languages is a major asset for Natural Language Processing as this permits the cross-language interoperability and reuse of many language processing tools and applications — and this is certainly another major factor for the success of Universal Dependencies framework and associated endeavour in terms of developing treebanks, parsers and other processing tools for different languages.

The goal of this paper is to present an open collection of mutually consistent language resources, processing tools and support services concerning Universal Dependencies for the Portuguese language with a unique set of combined characteristics and volume, including a treebank whose text is over three times larger than the treebanked text previously available in the literature. This collection encompasses:

- CINTIL-UPos, an UPOS annotated corpus that contains close to 675K annotated tokens;
- LX-UTagger, a stand-alone, top-performing UPOS tagger;
- web-based services for remote and browser-based usage of this UPOS tagger;
- CINTIL-UDep, an UD treebank that contains close to 38K annotated sentences;
- LX-UDParser, a stand-alone, top-performing UD parser;
- web-based services for remote and browser-based usage of this UD parser;
- expert support ensured by the Knowledge Center for the Science and Technology of Portuguese of the CLARIN international research infrastructure.

These scientific resources are publicly available from the repository and workbench of PORTULAN

CLARIN[1] and from a dedicated site in the GitHub platform.[2]

This paper is organized as follows. The next Section 2. presents the POS tagger and describes the respective annotated corpus upon which it was trained and evaluated. In Section 3., the development of the treebank and its companion parser is introduced. Section 4., in turn, is concerned with the web-based processing services. The support services are described in Section 5.. Finally, Section 6. overviews related work, while Section 7. wraps this paper up with concluding remarks.

2. Part of speech

2.1. Annotated corpus

To support research, including the training and evaluation of language processing tools concerned with UD for Portuguese, we developed a corpus annotated with information on part of speech (POS) and on morphological features and lemmas, the CINTIL-UPos corpus.

The collection of data on which this was developed is, to the best of our knowledge, the largest corpus for Portuguese publicly available that was manually annotated, the CINTIL corpus (Barreto et al., 2006). With 1 Million tokens, this corpus is composed of written texts from news (34%), novels (17%) and written speech transcriptions (42%).

This is a quality linguistically interpreted corpus that was manually annotated by experts in Linguistics whose labeling decisions were harmonized by annotation guidelines specifically designed to take into account the Portuguese language (Barreto et al., 2005).

The CINTIL-UPos corpus was built from the CINTIL corpus by adding to a subset of the later an extra annotation layer that is compliant with the UPOS tagset. This subset used excluded the portion of CINTIL concerned with speech transcriptions, and comprises close to 675K tokens.

The UPOS layer was obtained by the mapping between the CINTIL tagset and the UPOS tagset presented in Annex 7.. Given that the size of the later is smaller than the former, and that every category in the CINTIL tagset, except the tag UM, maps univocally to a UPOS tag, it was not necessary to supplement the automatic mapping with an exhaustive process of manual validation. The only exception concerned the ambivalent CINTIL tag UM used only for the Portuguese word *um* (masc.) / *uma* (fem.), ambivalent as indefinite article and cardinal number, which was mapped to the UPos tag DET.[3]

The frequency of the different tags on the CINTIL-UPos is presented in the table in Annex 7..

2.2. Tagger

LX-UTagger is the POS tagger that is the companion to the CINTIL-UPos corpus.[4]

It is based on the pre-trained language model for Portuguese BERTimbau (Souza et al., 2020), specifically the `bert-base-portuguese-cased` model[5]. More specifically, it is an instance of the `BERTForTokenClassification` model that is part of the Hugging Face(Wolf et al., 2020) Transformers library and it was fine-tuned and evaluated on the CINTIL-UPos corpora.

Training and evaluation were performed under a 10-fold cross validation procedure, using 90% of the corpus for training and 10% for evaluation on each one of ten folds. The corpus was randomly shuffled, once, prior to making the 10-fold partitions, in an attempt to make each fold containing as much linguistic phenomena variety as the others. The accuracy scores for each of the 10 folds, are presented in Table 1.

The performance of the LX-UTagger shows a state of the art performance, attaining an accuracy score of 99.01%.

Though not comparable because of having been trained on different data sets, it is interesting to note that 98.04% is the published best score for UPos tagging of Portuguese (Table 2), with UDPipe (Straka, 2018), and 99.18% is the best score for UPos tagging the 75 languages addressed in (Kondratyuk and Straka, 2019), for UDPipe and UDify, namely when they were trained with the 68.5 Ktoken Czech PDT corpus.

The model supporting the publicly available distribution of the LX-UTagger is the one trained on fold 3, which achieved the highest accuracy score among the ten folds (99.06%).

Each fold was trained for 5 epochs, the model parameters being saved at the end of each epoch and the best performing set of parameters being taken at the end.

The last layer, used for classification, has 20 units, as many as the number of distinct tags in the corpus, taking into consideration that some of these tags are compound, resulting from contracted expressions such as *do*/ADP+DET, the contraction of the two words, *de*/ADP *o*/DET (Eng. *of the*). During training and evaluation, compound tags were left as single tags, but when tagging text for any other practical purpose, the contractions are expanded, both the tokens and their respective tags.

[1] `https://portulanclarin.net`

[2] `https://https://github.com/nlx-group/ud-portuguese`

[3] Given that all occurrences of *um*/*uma* are tagged with DET, this will permit to retrieve all and only the relevant instances in case one might be interested to study this ambivalence or re-annotate them under possibly different tagging criteria.

[4] LX-UTagger is the alias of convenience for the LX-Tagger (Branco and Silva, 2004) set for UPos.

[5] `https://huggingface.co/neuralmind/bert-base-portuguese-cased`

fold #	1	2	3	4	5
accuracy	98.97	99.03	99.06	99.05	99.01
fold #	6	7	8	9	10
accuracy	98.99	98.98	99.01	99.00	99.03
		average:	99.01		

Table 1: Accuracy scores for LX-UTagger in each of the 10 folds used during cross-validation.

3. Grammatical dependencies

3.1. Treebank

To support research, including the training and evaluation of UD parsers for Portuguese, we developed a treebank annotated with information on universal grammatical dependencies, the CINTIL-UDep treebank.

The collection of data on which this annotated corpus was developed is, to the best of our knowledge, the largest set of treebanks for Portuguese publicly available that were manually validated, namely the collection of CINTIL treebanks (de Carvalho et al., 2016; Branco et al., 2014; Branco et al., 2012; Branco et al., 2010).

With 37,780 sentences (473,929 tokens), the CINTIL-UDep treebank (93%) includes tokens from the CINTIL DependencyBank— whose text is a subset of the text in the CINTIL-UPos corpus described above in Section 2.1.—, and (7%) includes tokens also from the CINTIL DependencyBank Premium.

For the portion of CINTIL-UDep treebank coming from the CINTIL DependencyBank, the manual annotation was supported by a deep computational grammar. The respective sentences were input to LXGram (Costa and Branco, 2010), a computational grammar for the deep processing of Portuguese, developed in the HPSG grammatical framework (Pollard and Sag, 1994). For each sentence, its forest parse with all possible grammatical analysis by the grammar was pruned manually until a parse was left, with which that sentence got annotated.

This was performed with the support of the [incr tsdb()] grammar profiling and treebanking tool (Oepen and Flickinger, 1998) by experts in Linguistics, under a double blind annotation followed by adjudication treebanking procedure.

The usage of the grammar ensured both theoretical consistency across all sentences in the treebank and the well-formedness of the grammatical representation of each sentence. The double-blind procedure, in turn, ensured a very high standard of reliability for human annotated data sets.

The ensuing CINTIL-DeepBank contains fully fledged, deep grammatical representations, encompassing syntactic and semantic information. This is the basis of a few other treebanks where only part of that information was retained to annotated each sentence (Branco et al., 2010). For instance, besides morphological information including part of speech and inflection they include, the CINTIL Treebank is a streamlined version that retains only information on syntactic constituency, the CINTIL PropBank only information on semantic roles, the CINTIL DependencyBank only information on grammatical dependencies, and the CINTIL LogicalFormBank only on the deep semantics using a Minimal Recursion Semantics representation (Copestake et al., 2005).

For the portion of CINTIL-UDep treebank coming from CINTIL DependencyBank Premium, in turn, for the treebanking of each sentence, the dependency graph was drawn by experts in Linguistics and associated to that sentence, with the support of the Webanno tool (Eckart de Castilho et al., 2016). The annotation guidelines (Branco et al., 2015) followed by the annotators ensured full consistency of this treebank with the dependency treebank annotated with the support of the grammar.

This treebank contains thus sentences that instantiate types of linguistic phenomena that may not be represented in the grammar-based CINTIL DependencyBank, given that, like other computational grammars, LXGram has suboptimal text coverage.

The CINTIL-UDep, of interest in the present paper, was thus developed by joining together the CINTIL-DependencyBank and the CINTIL-DependencyBank Premium, and by annotating the resulting data set with a layer of structured information compliant with the UD guidelines.[6]

This was obtained by converting the previous dependency layer and replacing it with the outcome of a conversion tool developed for that purpose. This tool relies on the tregex package (Levy and Andrew, 2006) to implement hand-crafted rules that match against the dependency tree and perform deterministic actions involving relabeling relation names and rerouting arcs. As its refinement proceeds, it is reaching a range of LAS values typical of automatic parsers, with manual validation against samples of corpus sentences.

3.2. Parser

LX-UDParser is the parser that is the companion to the CINTIL-UDep corpus.[7]

LX-UDParser is based on the nlp4j framework, a transition-based, non-projective parsing algorithm with linear-time performance. It is reported to achieve 92.26 UAS and 91.93 LAS for English on WSJ, and to be efficient, taking 9 miliseconds per sentence (Choi and McCallum, 2013).

This parsing framework allows to specify flexible feature templates to be used in learning. We used the example configuration as a basis, but with word forms

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

[7] LX-UDParser is the alias of convenience for the LX-DepParser (Silva et al., 2010) set for UD.

instead of lemmas as features, and AdaGrad as the optimizer.

The LX-UDParser was trained and evaluated under a 10-fold cross-validation methodology, and its performance scores, taking the average over the 10 folds, is 90.87 for UAS and 88.01 for LAS.[8]

Though not comparable because of having been trained on different data sets (of different volume and syntactic diversity), it is interesting to note that 92.54 and 91.15 are the two published best scores for UD parsing of Portuguese (Table 2), for UDify (Kondratyuk and Straka, 2019) and UDPipe (Straka, 2018) respectively, and 93.87 is the best LAS score for UD parsing for the 75 languages addressed in (Kondratyuk and Straka, 2019), for UDify, namely when it was trained with the 8.5K sentences Slovak SNK corpus.

We will continue experimenting with other parsing frameworks, of diverse efficiency and accuracy, and LX-UDParser variants will be made available accordingly.

4. Language processing services

The two UD processing tools for Portuguese, LX-UTagger and LX-UDParser, just described in the sections above are two stand-alone pieces of software, apt to be installed and run locally, publicly available for download and reuse from a sign-in free distribution platform of language resources, the PORTULAN CLARIN repository.[9]

In order to further promote their availability as well as the ease and dissemination of their use, they were also embedded in and made to support remote language processing services that can be run for free from a number of convenient web-based interfaces associated to each of these two tools, available from the PORTULAN CLARIN workbench.[10] The present section is aimed at briefly introducing those services.

4.1. Online service

What we termed as *online service* is the central web-based interface for each tool, which is accessible via a web browser.[11] It allows users: to experiment with a tool by changing its input and its possible options for the output format and immediately see the respective effect in the output window; to run one-click usage examples that help users start experimenting with the least amount of effort; to have access to several forms of documentation; and to provide an entry point to the other remote services for the tool, namely as a *web service*, a *file processing* or a *notebook service* interface; to provide pointers to download the respective tool and annotated corpus.

As an example, Figure 1 presents the front page of the online service interface for the parser.

4.2. Web service

The *web service* is a remote procedure call (RPC) type of interface, through which it is possible to interact remotely with the tools by means of computer programs.[12]

To start using the web service, a user will click the “Web Service” button in the tool’s online service interface, which will bring up a dialog window that contains detailed information about the technical requirements that have to be met for this service to be used. In addition, a very simple and self-contained Python program is displayed, which can be copied to a local version and used as a starting point for users with little programming experience to develop their own programs that call the remote web service.

4.3. File processing service

The *file processing* interface is a multi-step workflow, supported by a sequence of dialog windows, that is launched by clicking on the “File Processing” button at the top of the online service interface.

If the file is small enough so that it can be processed under two minutes, then processing will start immediately after the file is uploaded and as soon as the processing is complete, the user will be able to download the processed output files by clicking on a “Download” button.

If the file being uploaded, in turn, is large enough such that its processing time is estimated to be longer than two minutes, then the processing will take place in the background, without requiring the user to suspend other activities waiting for its completion. Instead, in this type of job, when the processing is complete, the user will receive an email message with an URL for downloading the output file, sent to an email address asked on-the-fly to the user.

4.4. Notebook service

The *notebook interface* is launched by clicking on the “Notebook” button at the top of the online service interface. This interface provides an easy path for users to explore the combinatorial affordances of web services

[8] For the definition the UAS (unlabeled attachment score) and LAS (labeled attachment score) metrics, as well as the evaluation script we used, check the site of the CoNLL 2018 shared task “Multilingual Parsing from Raw Text to Universal Dependencies” `https://universaldependencies.org/conll18/evaluation.html`

[9] `https://portulanclarin.net/repository/search`

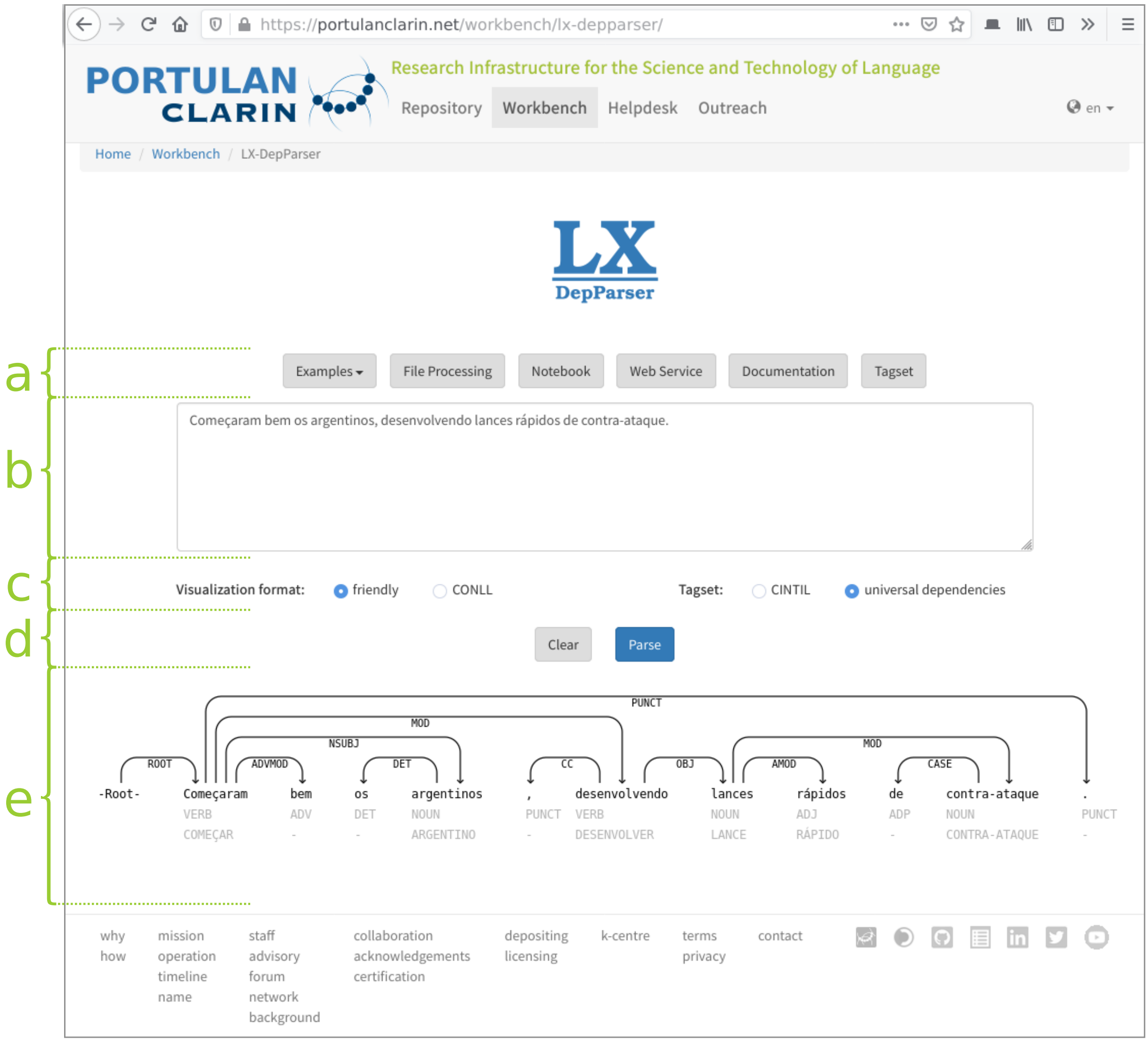
[10] `https://portulanclarin.net/workbench`

[11] The browser should be directed at the following addresses:

`https://portulanclarin.net/workbench/lx-depparser/` for the parser

`https://portulanclarin.net/workbench/lx-tagger/` for the tagger

[12] We chose to implement this service using JSON-RPC, which is a light-weight and programming language-agnostic protocol for which implementations are readily available in many programming languages.



https://portulanclarin.net/workbench/lx-depparser/

Começaram bem os argentinos, desenvolvendo lances rápidos de contra-ataque.

Visualization format: ● friendly ○ CONLL

Tagset: ○ CINTIL ● universal dependencies

Figure 1: Example of the online service interface. In area (a), a row of buttons that give access to examples, other types of interfaces (file processing, notebook and web services), and documentation; in (b), the user input; in (c), options for controlling the tool's behavior and output; in (d), buttons to start processing or clear input; and finally, in area (e), the results are shown.

in a non-local fashion, by resorting to a browser for coding and to non-local servers to run their respective code.

At a fundamental level, notebooks are documents that are both human-readable and machine-executable: they contain textual and graphical elements such as headings, paragraphs, lists, equations and figures, as well as executable code (e.g. Python code). Notebooks also incorporate visualizations of the results produced by executing their code sections. Furthermore, notebooks can be edited interactively, and the code sections re-run at will, making them ideal for experimentation.

The web-based notebook interface for our UD tools is based on Jupyter notebooks (Project Jupyter et al., 2018) and offers its users different, local and non-local, options to execute notebooks: launch notebooks on the Binder platform,[13] and make the code run remotely in the respective Jupyter Project's servers; launch notebooks on Google's Colab platform[14] and make the code run remotely in Google's servers; download notebooks and execute them locally on the users' own computer.

5. Support

In order to support users in handling and exploiting the scientific resources presented in the present paper, in all their different access modalities, expert support is provided by the Knowledge Center for the Science and Technology of Portuguese of the CLARIN research infrastructure.[15]

[13] https://jupyter.org/binder

[14] https://colab.research.google.com/

[15] https://portulanclarin.net/helpdesk

This support is freely available for research purposes and it is targeting all types of users, from AI to Digital Humanities from researchers to companies. It is provided by researchers in the science and technology of language who belong to the centers in the scientific advisory network of PORTULAN CLARIN and who are experts on the Portuguese language.

6. Related work

There is a range of language resources, both data and software, concerning UD for different languages that have been distributed and are accessible from diverse types of channels, including from distribution platforms dedicated to language science and technology (e.g. ELRA, CLARIN, etc.) to specific web sites for particular resources and/or languages. Given the scope of the present paper, this section is not apt to deliver an overarching overview of UD touching on different languages, on theoretical foundations of UD, major initiatives and projects, etc. — which can be found elsewhere (de Marneffe et al., 2021) —, and we will rather focus on related work concerning UD for Portuguese.

6.1. Annotated corpora

Comprehensive information on annotated corpora available concerning UD for different languages has been gathered in the site https://universaldependencies.org/.

To the best of our knowledge, and also from what one can find in that site, there are two other major UD treebanks for Portuguese whose source result from a manually annotated and/or validated endeavour: Bosque (approx. 10K sentences, 210K tokens) (Freitas et al., 2008), and GSD (approx. 12K sentences, 300K tokens), where the annotation layer of the latter was converted from the annotation layer of the former, and thus with an almost complete overlap of text between the two (Rademaker et al., 2017).[16]

In terms of volume, with 37,780 sentences and 473,929 tokens, the CINTIL-UDep presented here is three times larger than the largest one of them, thus representing a most relevant contribution to improve on the volume and variety of UD treebanked data that was manually validated and is publicly available for Portuguese.

If in turn, one considers only UD POS-annotated data, with its 26,779 sentences and 675,530 tokens, the CINTIL-UPos is over 15 times larger than the largest one of previously available corpora.

Like these other treebanks, the CINTIL-UDep was annotated manually in non-UD style and automatically converted to UD. It is however worth noting that the non-UD treebank from which CINTIL-UDep is obtained was treebanked with a double blind annotation followed by adjudication procedure that ensures the highest reliability level for data annotation.

For the sake of completeness, it is also worth mentioning BDCamões Dependency Bank (208 documents, 180K sentences, 4.5M tokens), also developed by our team (Grilo et al., 2020). Differently from the manually annotated corpora above, this is a treebank annotated in an automatic fashion only (in non-UD style and then automatically converted to UD). Likely with more noise in their grammatical representations, its large volume however makes of it a potentially useful resource when size may be a subordinating factor.

There is also another small corpus, automatically annotated, PUD (1K sentences, 22K tokens), part of a parallel treebank created for the CoNLL 2017 shared task on "Multilingual Parsing: from Raw Text to Universal Dependencies" (Zeman et al., 2017), where the sentences were randomly picked from on-line newswire and Wikipedia (750 sentences in English, and 250 in German, French, Italian or Spanish) and translated by professional translators.

6.2. Processing tools

There are a number of publicly available UD parsers for Portuguese with published performance, namely PassPort[17] (Zilio et al., 2018), UDPipe[18] (Straka, 2018), COMBO[19] (Klimaszewski and Wróblewska, 2021b), UDify[20] (Kondratyuk and Straka, 2019), and the models from Spacy framework.[21]

These processing tools are supported by an array of different parsing approaches. PassPort runs the Stanford parser (Chen and Manning, 2014) on Portuguese data. Spacy is based on the non-monotonic arc-eager transition-system described in (Honnibal and Johnson, 2015). UDPipe resorts to Stanford's biaffine attention parser (Dozat et al., 2017). COMBO parser is presented in (Klimaszewski and Wróblewska, 2021a). And UDify is a multilingual multi-task neural network model able to perform part-of-speech tagging, morphological analysis, lemmatization and dependency parsing simultaneously by leveraging a multi-lingual BERT (Devlin et al., 2019) self-attention model pretrained on 104 languages.[22]

The training and evaluation of these processing tools has resorted to the GSD treebank, derived from Bosque, except UDPipe, which resorts to Bosque, as these two treebanks have been in the Universal Dependencies repository distributions (Nivre et al., 2020). The respective performance scores are in Table 2

[16] In https://universaldependencies.org/, there is a mention to other two corpora, namely DHBB, which is not available and seems to have been deprecated.

[17] https://cental.uclouvain.be/resources/smallasmille/passport/

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

[19] https://gitlab.clarin-pl.eu/syntactic-tools/combo

[20] https://github.com/Hyperparticle/udify

[21] https://spacy.io/models/pt

[22] https://github.com/google-research/bert/blob/master/multilingual.md

	UPOS	UAS	LAS
PassPort	n.a.	87.55	85.21
Spacy	97	90	86
UDPipe	96.37	89.48	87.04
COMBO	98.06	92.72	91.15
UDify	98.04	94.22	92.54

Table 2: Publicly available UD parsers for Portuguese and their published performance in terms of accuracy for POS tagging (UPOS), unlabeled (UAS) and labeled attachment score (LAS).

7. Conclusion

In this paper we presented a collection of mutually consistent and inter-operable scientific resources for the computational processing of the Portuguese language under the Universal Dependencies framework. This is an all-encompassing, publicly available open collection that includes reliably annotated corpora, top-performing processing tools and expert support services.

This represents a major extension for the universe of language resources concerning UD for Portuguese previously publicly available and reported in the literature, in terms of either quantity, quality, availability or breadth.

With a new UD treebank with nearly 38K sentences, in terms of quantity, this contribution is 3 times the volume of treebanked text previously available; and with a new UD POS-annotated corpus with 675K tokens, it extends over 15 times the volume of POS annotated data previously available.

With the procedure adopted for the manually annotation of the data, double blind annotation followed by adjudication, in terms of quality, these annotated corpora are the only ones for UD in Portuguese having adopted this expensive and demanding procedure that ensures the highest reliability level for human annotated data.

With the UD POS-tagger and the UD parser trained on these corpora, and made freely available both as local stand-alone tools and as remote web-based services of various sorts (online service, web service, file processing service, notebook service), in terms of availability, the computational processing of Portuguese under the UD framework can be said to have become available in as many ways as it is possible one to get it with currently affordable technology.

With the helpdesk support—targeting all types of users, from researchers to language professionals, from AI to Digital Humanities—ensured by the Knowledge Center for the Science and Technology of Portuguese of the PORTULAN CLARIN research infrastructure, in terms of breadth, the UD open and diverse ecosystem for Portuguese has its profile enhanced to a superior level, which goes beyond only more data and processing and encompasses also dedicated expert support.

The scientific resources and support presented in this paper are available from the PORTULAN CLARIN Research Infrastructure for the Science and Technology of Language.[23]

Acknowledgments

The research reported here was supported by PORTULAN CLARIN—Research Infrastructure for the Science and Technology of Language `https://portulanclarin.net`, funded by Lisboa 2020, Alentejo 2020 and FCT—Fundação para a Ciência e Tecnologia under the grant PINFRA/22117/2016.

Appendix: POS mappings

The following table shows the CINTIL-UPos tag distribution as well as the POS mapping between the original CINTIL tagset and the UD POS tagset.

UD		CINTIL
ADJ	5.00%	ADJ, ORD
ADP	14.86%	PREP
ADV	4.50%	ADV
AUX	0.24%	VAUX, VAUXGER, VAUXINF
CCONJ	2.70%	CJ†
DET	14.90%	DA, DEM, IA, QNT, UM
INTJ	0.08%	DM, ITJ
NOUN	17.26%	CN, EOE, MGT, MTH, PADR, STT, WD
NUM	1.50%	CARD, DFR, DGT, DGTR
PRON	5.18%	CL, IND, INT, POSS, PRS, REL
PROPN	5.95%	PNM
PUNCT	13.77%	PNT
SCONJ	1.71%	CJ†
SYM	0.10%	LTR, SYB, TERMN
VERB	12.26%	GER, INF, PPA, PPT, V

†mapping depends on the word form

Table 3: CINTIL-UPos tag distribution and mapping from CINTIL

The CINTIL POS tagset also includes tags with the form L(POS)n for labeling the tokens in multi-word expressions (MWE), where POS is the part-of-speech of the MWE and n the index of the token in the MWE, for instance “LPREP1” and ”LPREP2” would be assigned to the first and second tokens in a prepositional MWE. When converting to UD POS, the tokens in MWE are annotated individually.

References

Barreto, F., Branco, A., Mendes, A., Bacelar do Nascimento, M. F., and Silva, J. R. (2005). CINTIL corpus internacional do Português: Convenções de etiquetação.

[23] `https://portulanclarin.net`

Barreto, F., Branco, A., Ferreira, E., Mendes, A., Bacelar do Nascimento, M. F., Nunes, F., and Silva, J. R. (2006). Open resources and tools for the shallow processing of Portuguese: the TagShare project. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC 2006)*, pages 1438–1443.

Branco, A. and Silva, J. (2004). Evaluating solutions for the rapid development of state-of-the-art POS taggers for Portuguese. In *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC 2004)*, pages 507–510.

Branco, A., Costa, F., Silva, J., Silveira, S., Castro, S., Avelãs, M., Pinto, C., and Graça, J. (2010). Developing a deep linguistic databank supporting a collection of treebanks: the CINTIL DeepGramBank. In *Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC 2010)*.

Branco, A., Carvalheiro, C., Pereira, S., Avelãs, M., Pinto, C., Silveira, S., Costa, F., Silva, J., Castro, S., and Graça, J. (2012). A PropBank for Portuguese: the CINTIL-PropBank. In *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC 2012)*, pages 1516–1521.

Branco, A., Silva, J., Gonçalves, P., Costa, F., Silveira, S., Gaudio, R. D., Rodrigues, J., Castro, S., Rodrigues, L., Martins, P., Nunes, F., Ferreira, E., Alves, J., de Carvalho, R., Querido, A., Campos, M., and Rendeiro, N. (2014). The CINTIL and LX companion collections of language resources and tools for Portuguese. In *Proceedings, ToRPorEsp - Workshop on Tools and Resources for Automatically Processing Portuguese and Spanish, 11th International Conference on the Computational Processing of Portuguese (PROPOR2014)*, pages 1516–1521.

Branco, A., Silva, J., Querido, A., and de Carvalho, R. (2015). CINTIL DependencyBank PREMIUM handbook: Design options for the representation of grammatical dependencies. Technical report, University of Lisbon, Faculty of Sciences, Department of Informatics. TR-2015-05, DOI:10451/20226.

Chen, D. and Manning, C. (2014). A fast and accurate dependency parser using neural networks. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 740–750, Doha, Qatar, October. Association for Computational Linguistics.

Choi, J. D. and McCallum, A. (2013). Transition-based dependency parsing with selectional branching. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1052–1062.

Copestake, A., Flickinger, D., Pollard, C., and Sag, I. A. (2005). Minimal recursion semantics: An introduction. *Research on language and computation*, 3(2):281–332.

Costa, F. and Branco, A. (2010). Lxgram: A deep linguistic processing grammar for Portuguese. In *International Conference on Computational Processing of the Portuguese Language*, pages 86–89. Springer.

de Carvalho, R., Querido, A., Campos, M., Pereira, R., Silva, J., and Branco, A. (2016). CINTIL DependencyBank PREMIUM: A corpus of grammatical dependencies for Portuguese. In *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC 2016)*, pages 1552–1557.

de Marneffe, M.-C., Manning, C. D., Nivre, J., and Zeman, D. (2021). Universal Dependencies. *Computational Linguistics*, 47(2):255–308, 07.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding.

Dozat, T., Qi, P., and Manning, C. D. (2017). Stanford's graph-based neural dependency parser at the CoNLL 2017 shared task. In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 20–30, Vancouver, Canada, August. Association for Computational Linguistics.

Eckart de Castilho, R., Mújdricza-Maydt, É., Yimam, S. M., Hartmann, S., Gurevych, I., Frank, A., and Biemann, C. (2016). A web-based tool for the integrated annotation of semantic and syntactic structures. In *Proceedings of the Workshop on Language Technology Resources and Tools for Digital Humanities (LT4DH)*, pages 76–84.

Freitas, C., Rocha, P., and Bick, E. (2008). Floresta sintá(c)tica: bigger, thicker and easier. In *Proceedings of International Conference on Computational Processing of the Portuguese Language (PROPOR 2008)*, pages 216–219. Springer.

Grilo, S., Bolrinha, M., Silva, J., Vaz, R., and Branco, A. (2020). The BDCamões collection of Portuguese literary documents: a research resource for digital humanities and language technology. In *Proceedings of the 12th Language Resources and Evaluation Conference (LREC 2020)*, pages 849–854.

Honnibal, M. and Johnson, M. (2015). An improved non-monotonic transition system for dependency parsing. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1373–1378, Lisbon, Portugal, September. Association for Computational Linguistics.

Klimaszewski, M. and Wróblewska, A. (2021a). COMBO: A new module for EUD parsing. 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 158–166, Online, August. Association for Computational Linguistics.

Klimaszewski, M. and Wróblewska, A. (2021b). COMBO: State-of-the-art morphosyntactic analysis. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 50–62, Online and Punta

Cana, Dominican Republic, November. Association for Computational Linguistics.

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)*, pages 2779–2795, Hong Kong, China, November. Association for Computational Linguistics.

Levy, R. and Andrew, G. (2006). Tregex and Tsurgeon: Tools for querying and manipulating tree data structures. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC 2006)*, pages 2231–2234.

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, May. European Language Resources Association.

Oepen, S. and Flickinger, D. (1998). Towards systematic grammar profiling. test suite technology 10 years after. *Journal of Computer Speech & Language*, 12(4):411–436.

Pollard, C. and Sag, I. A. (1994). *Head-Driven Phrase Structure Grammar*. The University of Chicago Press, Chicago.

Project Jupyter, Matthias Bussonnier, Jessica Forde, Jeremy Freeman, Brian Granger, Tim Head, Chris Holdgraf, Kyle Kelley, Gladys Nalvarte, Andrew Osheroff, Pacer, M., Yuvi Panda, Fernando Perez, Benjamin Ragan Kelley, and Carol Willing. (2018). Binder 2.0 - Reproducible, interactive, sharable environments for science at scale. In Fatih Akici, et al., editors, *Proceedings of the 17th Python in Science Conference*, pages 113 – 120.

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, Pisa,Italy, September. Linköping University Electronic Press.

Silva, J., Branco, A., Castro, S., and Reis, R. (2010). Out-of-the-box robust parsing of Portuguese. In *Lecture Notes in Artificial Intelligence, 6001*, pages 75–85.

Souza, F., Nogueira, R., and Lotufo, R. (2020). BERTimbau: pretrained BERT models for Brazilian Portuguese. In *9th Brazilian Conference on Intelligent Systems, BRACIS, Rio Grande do Sul, Brazil, October 20-23 (to appear)*.

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, Brussels, Belgium, October. Association for Computational Linguistics.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Scao, T. L., Gugger, S., Drame, M., Lhoest, Q., and Rush, A. M. (2020). Huggingface's transformers: State-of-the-art natural language processing.

Zeman, D., Popel, M., Straka, M., Hajič, J., Nivre, J., Ginter, F., Luotolahti, J., Pyysalo, S., Petrov, S., Potthast, M., Tyers, F., Badmaeva, E., Gokirmak, M., Nedoluzhko, A., Cinková, S., Hajič jr., J., Hlaváčová, J., Kettnerová, V., Urešová, Z., Kanerva, J., Ojala, S., Missilä, A., Manning, C. D., Schuster, S., Reddy, S., Taji, D., Habash, N., Leung, H., de Marneffe, M.-C., Sanguinetti, M., Simi, M., Kanayama, H., de Paiva, V., Droganova, K., Martínez Alonso, H., Çöltekin, Ç., Sulubacak, U., Uszkoreit, H., Macketanz, V., Burchardt, A., Harris, K., Marheinecke, K., Rehm, G., Kayadelen, T., Attia, M., Elkahky, A., Yu, Z., Pitler, E., Lertpradit, S., Mandl, M., Kirchner, J., Alcalde, H. F., Strnadová, J., Banerjee, E., Manurung, R., Stella, A., Shimada, A., Kwak, S., Mendonça, G., Lando, T., Nitisaroj, R., and Li, J. (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*, pages 1–19, Vancouver, Canada, August. Association for Computational Linguistics.

Zilio, L., Wilkens, R., and Fairon, C. (2018). PassPort: A dependency parsing model for Portuguese. In *Proceedings of the International Conference on Computational Processing of the Portuguese Language (PROPOR)*, pages 479–489.