Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 6635–6643
Marseille, 20-25 June 2022

PortiLexicon-UD: a Portuguese Lexical Resource according to Universal Dependencies Model

Lucelene Lopes[1], **Magali S. Duran**[1], **Paulo Fernandes**[2], **Thiago A. S. Pardo**[1]
[1]Interinstitutional Center for Computational Linguistics (NILC)
Institute of Mathematical and Computer Sciences, University of São Paulo, São Carlos–SP, Brazil
[2]Merrimack College, North Andover–MA, USA
{lucelene, magali.duran, lemelle.fernandes}@gmail.com, taspardo@icmc.usp.br

Abstract

This paper presents PortiLexicon-UD, a large and freely available lexicon for Portuguese delivering morphosyntactic information according to the Universal Dependencies model. This lexical resource includes part of speech tags, lemmas, and morphological information for words, with 1,221,218 entries (considering word duplication due to different combination of PoS tag, lemma, and morphological features). We report the lexicon creation process, its computational data structure, and its evaluation over an annotated corpus, showing that it has a high language coverage and good quality data.

Keywords: Lexical Resources, Portuguese, Universal Dependencies

1. Introduction

This paper describes the construction of a robust lexical resource for Brazilian Portuguese – PortiLexicon-UD – with 1,221,218 distinct entries in the form of 4-tuples, composed by information on word, lemma, Part of Speech (PoS) tag and morphological features, following the Universal Dependencies (UD) model and guidelines (Nivre et al., 2020).

Lexical resources are useful for several different tasks in Natural Language Processing (NLP), from supporting PoS tagging and parsing to named entity recognition and information extraction, among many other possibilities. They may also subsidize linguistic studies and corpus linguistics initiatives, helping characterizing specific text genres and domains, and allowing the investigation of patterns of word usage. In particular, the lexicon we propose also contributes to UD-oriented research for Portuguese, in line with what has been done for other languages (Przepiórkowski and Patejuk, 2018; Cecchini et al., 2020; Miletic et al., 2020; Seddah et al., 2020).

The new resource was based on the UNITEX-PB (Muniz, 2004) lexicon as a starting point, first by doing a PoS tagset mapping from more traditional tags to UD tags, and then by incorporating new decisions, which will be commented on later.

To the best of our knowledge, there is no available lexical resource using UD model for Portuguese. In fact, Portuguese resources in UD are limited to annotated corpora, three of which are publicly available at UD site: BOSQUE-UD (Rademaker et al., 2017), a subset of Floresta Sin(c)tática with 210,963 tokens, PUD (Uszkoreit et al., 2017), a small parallel corpus created for CoNLL 2017 shared task, with 21,917 tokens, and GSD (McDonald et al., 2013), a Portuguese translation from Google UD treebank, with 297,057 tokens.

This paper goal is to present a lexical resource, called PortiLexicon-UD, which contains 858,922 unique Portuguese words (homographs are considered only once) and 1,221,218 entries considering word duplication due to different combinations of PoS tag, lemma and morphological features. The resource, available for online search and for download (see the Conclusion), is organized in a computational structure based on multi-decision diagrams and binary search to provide an efficient access to the lexical information.

The next section describes the scope and basic definitions adopted for building the PortiLexicon-UD. The third section describes the mapping of the Portuguese PoS tags and morphological features into UD model. The fourth section describes the computational structure to hold the lexical resource information, as well as its characteristics. The fifth section describes an application of PortiLexicon-UD to a practical case, which also allows assessing the lexicon quality. Finally, the conclusion summarizes this paper contributions and possible future works.

2. Scope and Basic Definitions

The basic UD principles include to represent a sentence in the token level, associating to each token a PoS tag, a lemma, and morphological features (as number and gender for nouns and adjectives; mode, tense and person for verbs; and so on). It is only at the syntactic level that the tokens are associated using dependency relations to compose a tree, but dependency relations are beyond the scope of a lexicon. To build PortiLexicon-UD, we initially did an exploratory study of available lexicons for Portuguese. We analyzed the Portuguese DELAF (Ranchhod et al., 1999), UNITEX-PB (Muniz, 2004), and the MorphoBr (Figueiredo de Alencar et al., 2018) lexicons, and it is noticeable that they share the vast majority of their entries, including automatic word

generation errors, as the past participle "*obteido*" for the verb "*obter*" instead of "*obtido*" ("obtained" in English). They also share a significant portion of lexical information, using the same tagset and the same morphological features. For convenience and familiarity, we adopted UNITEX-PB as a starting point.

To construct a lexicon using UD schema, it is not enough to read UD Guidelines and to know UD tagsets: it is necessary to instantiate the guidelines in the target language, what requires some decisions. This may be observed when we compare the three available UD annotated corpora in Portuguese (BOSQUE-UD, PUD, and GSD), which do not agree upon some decisions concerning assignment of PoS tags. Because of that, we decided to follow a single source of instantiated guidelines for assignment of UD POS tags in Portuguese (Duran, 2021), which has been applied for the construction of a new treebank for Portuguese (Pardo et al., 2021).

Once the design decisions were made, the next step was to map the PoS tags and morphological features of UNITEX-PB to the UD ones. This process was not as simple as first imagined and required a lot of effort to manually review the mapping results and checking data quality, and possibly correcting it. We group these process' tasks into:

- corrections of word forms, lemmas, and morphological features;
- exclusion of entries;
- inclusion of entries; and
- unfolding of entries that admitted more than one lemma, more than one PoS tag, and more than one set of morphological features.

2.1. Corrections

The correction of word forms was made automatically by verifying words from the available corpora that were not found in the lexical resource, and then analysing them manually. This allowed us to correct 594 entries distributed mostly on PoS tag VERB, but also on PoS tags NOUN, ADJ, and ADV. For example, for 42 verbs (e.g., "*faltar*" - "to lack") 10 forms of each verb were incorrect (e.g., for the first person singular of the imperative form it was "*falgue*" instead of "*falte*"). This problem alone motivated 420 word form corrections.

The correction of lemmas was made manually for ADJ and NOUN words annotated in UNITEX-PB with degree feature (augmentative and diminutive words). As there are several ways to produce these words, we consider them generated by derivation and not by an inflection process. This decision also benefits words that seem to be a diminutive or augmentative form but, in fact, are words with their own meaning, as for example "*bandeirinha*" which is the word that designates the soccer side referee (besides meaning "little flag") and "*salão*" ("hall") which is not an augmentative word derived of "*sala*" ("room"). For this reason, we assumed their lemmas are the same word in their original gender and in singular form. For words accepting gender inflection, the lemma is the masculine form. This correction changed the lemma of 3,375 entries.

UNITEX-PB does not systematically include diminutive and augmentative words. MorphoBr attempted to fill this gap, however we realized this kind of inclusion need to be linguistically validated and this is a task that we will address in future work.

Another verification of lemmas was carried out on all NOUNs ending by the letter "*a*" (generally a feminine form), but having a lemma ending with letter "*o*" (generally a masculine form). These words were carefully analyzed, as they might contain errors due to automatic lemmatization. Some feminine words (not a gender inflection) present masculine words as their lemmas (words that really exist, but are not related to the feminine ones). Examples of these mistakes are the NOUNs "*troca*" ("exchange") with lemma "*troco*" ("change") or "*proposta*" ("proposal") with lemma "*proposto*". Once a problem was encountered, we extended our analysis to the plural forms (e.g., "*trocas*" or "*propostas*"). For those cases, the lemma was brought to the feminine singular form, e.g., the lemma of both "*troca*" and "*trocas*" became "*troca*". This analysis resulted in 184 lemma corrections.

For the closed classes, we corrected the lemmas manually for all entries. In total, we made lemma corrections on 5,319 entries.

The corrections of morphological features were made automatically for verbs by checking the existence of forms for all inflections of all verbs in the UNITEX-PB resource. This allowed us to detect several error patterns and allowed us to correct 50,675 VERB entries. The large majority of such corrections was related to errors arising from the automatic generation of verb forms made by UNITEX-PB. For example, for the Present of Indicative, 21 verbs (e.g., "*fugir*" - "to run away" in English) had the third person plural (e.g., "*fogem*") annotated as the third person singular.

The second most frequent correction of morphological features was for adverbs, which affected 2,376 entries. Some corrections were also made on words with PRON, DET, NUM, ADJ, and NOUN tags, resulting in a grand total of 53,946 entries with morphological feature corrections.

2.2. Exclusions

The exclusion of entries from the source resource UNITEX-PB was mostly made with a manual analysis of some targeted groups of words. This is the case, for example, of the NOUN "*jardineirozinha*", an incorrect form to designate a little gardener, probably automatically generated. We adopted a conservative approach for this analysis and a large number of 4,148 entries as such were excluded.

We also promoted the exclusion of masculine nouns found as lemmas for the feminine nouns which were not only incorrect lemmas, but also nonexistent nouns in Portuguese. This was the case, for example, of the NOUN "*tentativo*" that was erroneously stated as lemma of the NOUN "*tentativa*". A total of 228 entries were excluded as such.

Furthermore, we performed the manual exclusion of some nonexistent verb inflections erroneously generated by UNITEX-PB, as the word "*construi*" for the imperative form of the verb "*construir*" ("to build") at the second person singular, which accepts only "*construas*" and "*constrói*". This corresponds to the exclusion of 94 entries.

Finally, we removed automatically all proper nouns, as we consider that proper nouns are usually too specific to be included in a general language lexicon and should be treated in a separate resource. The overall number of excluded entries was 12,743.

2.3. Inclusions

The inclusion of entries contemplated words that were not found in the available UD Portuguese corpora. Most of such words were new loan words, i.e., foreign words recently incorporated into Portuguese, as "*bitcoin*" and "*fintech*". For those words, a morphological feature indicating their foreign origin was added (Foreign=Yes). Some frequent abbreviations of Portuguese verbs were also included, as "*tô*" (short for "*estou*" - in English "I am"). For those words, a morphological feature indicating their nature was added (Abbr=Yes). These cases represented 719 new entries.

Another change that increased the number of entries was related to the adverbial behavior of adjectives. In Portuguese, adjectives are used as adverbs in some situations, such as when there are two coordinated adverbs, thus, the first one does not need to end with "*-mente*" ("-ly" in English). To do so, we duplicated all adjectives that form adverbs with "*-mente*", so that the duplicated ones are labeled as ADV. The included adverbs were noted with the lemma corresponding to the extended form (for example, the ADV "*social*" has "*socialmente*" as lemma - in English, "socially") and the abbreviated nature is also annotated (Abbr=Yes). The total number of such ADV inclusions added up 2,516 new entries.

Another important inclusion was the addition of 103 new verbs that were absent from UNITEX-PB. Some verbs are unusual ones found in the consulted corpora, as "*precificar*" ("to assign a price"), but also common ones, as "*resmungar*" ("to mumble"). Some verbs became very popular since UNITEX-PB was developed and most of them relate to computer science inspired actions, as "*deletar*" ("to delete)" and "*hackear*" ("to hack"). Those verbs included at least 71 conjugations each, adding up a total of 7,325 new entries.

2.4. Unfolding

The unfolding of entries that admit more than one combination of PoS tag, lemma, and morphological features is responsible for the largest increase of entries. It is worthy noticing that the new resource has about 1.2 million entries and the starting point resource has about 850,000.

Most of the extra information was originated from unfolded entries to represent the polycategorization of some words in UD. For example, while in UNITEX-PB verbs are only VERB, in UD some of them can be VERB or AUX. Another example of such unfolding are the possessive pronouns that are PRON in UNITEX-PB, whereas in UD they can be PRON, if used as a nominal (for example, "*meu*", translatable to "mine"), or DET, if used as a modifying pronoun ("*meu*, translatable to "my").

The most abundant source of additional entries is the duplication made for verb past participles, which in PortiLexicon-UD may be classified as VERB and ADJ. In this case, the added ADJs maintain track of their origin by keeping the morphological feature VerbForm=Part. The additional number of entries due to this unfolding process sum up to 347,864 new entries.

3. Mapping Portuguese into UD

The resulting lexicon with all basic definitions described was manually revised in full for the closed classes and by sampling for the other classes. In the next subsections, we present the adopted UD PoS tagset and the morphological features associated to each PoS tag that we use in PortiLexicon-UD.

3.1. Mapping PoS Tags

To map the Portuguese PoS tags used in UNITEX-PB into the UD PoS tags, we considered in our approach the standard 17 UD PoS tags[1] as detailed below.

- ADP, adpositions, a closed class that corresponds to prepositions in Portuguese, such as "*de*" ("of", in English), "*para*" ("to"), and "*com*" ("with");
- ADV, adverbs, an open class that has a closed subset of the primitive adverbs – those not formed after an adjective with the suffix "*-mente*" (as the suffix "-ly" in English). Examples of this closed subset of the adverb class are: "*cedo*" ("early"), "*agora*" ("now"), and "*acima*" ("above"). Examples of adverbs formed by derivation are "*normalmente*" ("normally") and "*insanamente*" ("insanely"). We also include adjectives that may be employed as adverbs, as previously mentioned. Examples of such are expressions as "*social e economicamente*" ("socially and economically");
- CCONJ, coordinating conjunctions, a closed class including, for example, "*e*" ("and"), "*mas*" ("but"), and "*portanto*" ("therefore");

[1] https://universaldependencies.org/u/pos/all.html

- SCONJ, subordinating conjunction, a closed class including, for example, "*conquanto*" ("although"), "*se*" ("if"), and "*segundo*" ("according to");
- DET, determiners, a closed class including, for example, "*cujo*" ("which"), "*diversos*" ("several"), and "*a*" ("the");
- PRON, pronouns, a closed class including, for example, "*eu*" ("I"), "*isso*" ("this"), and "*ambos*" ("both");
- PART, particles, a closed class that is not typically used in Portuguese, but has been assigned for the word "*que*" in specific contexts that can be translatable by "that" or "which" according to the instantiated guidelines we followed (Duran, 2021);
- NUM, the cardinal numbers, an open class that includes all numbers represented as digits (an open subset) or written (a closed subset), for example, "*cinco*" ("five"), and "*um*" ("one");
- AUX, the auxiliary and copula verbs, a closed class in UD, defined by the instantiated guidelines, which encompasses the conjugation of verbs "*ser*" ("to be"), "*estar*" ("to be"), "*haver*" ("to exist"), "*ir*" ("to go") and "*ter*" ("to have") whenever certain context constraints are met;
- VERB, the verbs, an open class in Portuguese, for example, "*canta*" ("(he/she/it) sings"), "*jogar*" ("to play"), and "*chorastes*" ("(you) cried");
- ADJ, the adjectives, an open class in Portuguese, which also includes in UD the ordinal numbers (including the ones represented with digits and ordinal symbols), for example, "*bonito*" ("beautiful"), "*caro*" ("expensive"), "*terceiro*" ("third") and "*50º*" ("50th");
- NOUN, the nouns, an open class in Portuguese, for example, "*presidente*" ("president"), "*quarto*" ("room"), and "*bola*" ("ball");
- INTJ, the interjections, an open class in Portuguese, for example, "*tchau*" ("bye") and "*oi*" ("hi");
- PROPN, the proper nouns, an open class in Portuguese, for example, "*Obama*" (the former American president) and "*Paris*" (the French capital city);
- X, all words that do not naturally belong to the vocabulary, for example, any foreign words, but also onomatopoeias, as "*détente*" ("relaxation") and "*roinc*" (the onomatopoeia for pigs' noise);
- PUNCT, all punctuation marks in Portuguese, for example, periods, parenthesis, and commas;
- SYM, all symbols that are represented or contain non-alphanumeric characters (signs), for example, "*R$*" (the representation of Real, the Brazilian currency) and "*%*" (the percent sign).

Unitex-PB PoS	**UD PoS tag**	
Adverbs	ADV	all adverbs
Prepositions	ADP	all prepositions
Coord. Conj.	CCONJ	all coordinating conjunctions
Sub. Conj.	SCONJ	all subordinating conjunctions
Articles	DET	all articles
Pronouns	DET	all pronouns, except the personal ones
	PRON	all pronouns
	PART	the pronoun "*que*" employed in specific contexts
Nouns	NOUN	all common nouns
Proper Nouns	PROPN	all proper nouns
Adjectives	ADJ	all adjectives
	ADV	all adjectives that have a corresponding adverb ending with "*-mente*"
Interjections	INTJ	all interjections
Numerals	NUM	all cardinals, except the higher forms ("*milhão*", "*bilhão*", etc.)
	ADJ	all ordinals
	NOUN	multiplicatives and the higher forms ("*milhão*", "*bilhão*", etc.)
Verbs	AUX	verbs "*ser*", "*estar*", "*haver*", "*ir*", and "*ter*"
	VERB	all verbs except "*ser*" and "*estar*"
	ADJ	all participle forms
Onomatopoeias	X	all onomatopoeias
	INTJ	if commonly used as interjection in Portuguese
Foreign words	NOUN	if already in common use of Portuguese
	X	if not in common use of Portuguese
Puctuations	PUNCT	all punctuations
Symbols	SYM	all symbols

Table 1: Mapping Unitex-PB PoS tags to the UD PoS tags

To map the Unitex-PB PoS tags into UD PoS classes, we have developed the associations stated in Table 1. However, since PortiLexicon-UD is intended to be a general lexicon of Portuguese, we will not represent PROPN and X classes that are virtually impossible to cover satisfactorily, nor we will represent PUNCT and SYM classes because they include non-alphabetical characters. Similarly, all elements of the

classes NUM, NOUN, and ADJ with non-alphabetical characters (digits or symbols) are left out of our lexical resource. As a result, our lexical resource covers 13 UD PoS tags. It is important to notice that some Unitex-PB PoS classes may have more than one possible PoS tag in UD, depending on the context in which the corresponding words are employed.

3.2. Mapping Unitex-PB Morphological Features to UD Features

The first important decision we took about UD morphological features[2] is to draw a distinction between features that may be associated to most PoS tags (generic features) and features that are naturally present for specific PoS tags (specific features).
The generic features are *Typo*, *Abbr*, and *Foreign*, since they can be associated to words belonging to all PoS classes. The specific features are defined to each considered PoS tag as follows.

For words in the PoS classes ADP, ADV, CCONJ, PART, SCONJ, and INTJ, no specific morphological feature is assigned. Other features employed in available Portuguese UD corpora, such as *Polarity*, *Degree*, *Reflexive*, and *Voice*, have not yet been employed in the lexicon because they require specific studies out of the scope of this paper.

For DET and PRON, we use the features *PronType*, *Definite* (if originally an article), *Gender*, *Number*, *Person*, *Case* (if a personal pronoun), and *Poss* (if a possessive pronoun). The acceptable values for these features are defined in Table 2. For example, the article "*uns*" (indefinite, masculine, and plural in Portuguese - "a" or "an" in English) has the PoS tag DET and is denoted by the following morphological features:

Definite=Ind Gender=Masc Number=Plur PronType=Art

Another example is the personal pronoun "*quantos*" ("how many") that has a PoS tag PRON and is denoted by one of the following morphological features:

Gender=Masc Number=Plur PronType=Ind

Gender=Masc Number=Plur PronType=Int

Gender=Masc Number=Plur PronType=Rel

Finally, a third example is the possessive pronoun "*vossas*" ("yours") that may have either a PoS tag DET or PRON and is denoted by the following morphological features:

Gender=Fem Number=Plur Person=2 PronType=Prs Poss=Yes

For NOUN and ADJ, we adopt the features *Gender* and *Number*, plus the origin of the word with either the features *VerbForm* for participles or *NumType* for ordinal numbers. The acceptable values for these features are defined in Table 3.

[2] https://universaldependencies.org/u/overview/morphology.html

For example, the word "*bola*" ("ball") has a PoS tag NOUN and is denoted by the following morphological features:

Gender=Fem Number=Sing

Another example is the NOUN formerly denoted as augmentative "*portão*" ("gate" and not "large door") which is denoted by the following morphological features:

Gender=Masc Number=Sing

For example, the adjective "*terceiro*" ("third") is originally an ordinal number, but receives the PoS tag ADJ and is denoted by the following morphological features:

Gender=Masc Number=Sing NumType=Ord

Finally, a fourth example is the adjective "*aposentado*" ("retired") that has a PoS tag ADJ and is denoted by the following morphological features:

Gender=Masc Number=Sing VerbForm=Part

For NUM, we adopt the feature *NumType* for cardinal numbers (*Card*), plus optionally the feature *Gender* for forms accepting gender variations, as "*dois*" and "*duas*", or "*duzentos*" and "*duzentas*". It is important to stress that the NUM PoS tag does not accept the feature *Number*. The acceptable values for these features are defined in Table 3. For example, the word "*cinco*" ("five") has a PoS tag NUM and is denoted by the following morphological features:

NumType=Card

Another example is the word "*uma*" ("one" feminine version) that has a PoS tag NUM and is denoted by the following morphological features:

Gender=Fem NumType=Card

For AUX and VERB, the features considered are *VerbForm*, *Mood*, *Tense*, *Gender*, *Number*, and *Person*. The acceptable values for these features are defined in Table 4. For example, the verb "*cantarei*" ("(I) will sing") has a PoS tag VERB and is denoted by the following morphological features:

Mood=Ind Number=Sing Person=1 Tense=Fut VerbForm=Fin

Another example is the auxiliary verb "*seríamos*" ("(we) would be") that receives the PoS tag AUX and is denoted by the following morphological features:

Mood=Cnd Number=Plur Person=1 VerbForm=Fin

Finally, a third example is the verb "*ganha*" ("(you) win (it)", "(he) wins", or "won") that has a PoS tag VERB and is denoted by one of the following morphological features (either being an imperative for the second person singular, a present of indicative for the third person singular, or a participle for feminine singular):

Mood=Imp Number=Sing Person=2 VerbForm=Fin

Mood=Ind Number=Sing Person=3 Tense=Pres VerbForm=Fin

Gender=Fem Number=Sing VerbForm=Part

kind	PoS tag	*PronType*	*Case*	*Definite*	*Gender*	*Number*	*Person*	*Poss*
Articles	DET	Art	-	Def/Ind	Masc/Fem	Sing/Plur	-	-
Demonstrative Pronoun	DET/PRON	Dem	-	-	-/Masc/Fem	Sing/Plur/Inv	-	-
Indefinite Pronoun	DET/PRON	Ind	-	-	-/Masc/Fem	Sing/Plur/Inv	-	-
Relative Pronoun	DET/PRON	Rel	-	-	-/Masc/Fem	Sing/Plur/Inv	-	-
Interrogative Pronoun	DET/PRON	Int	-	-	-/Masc/Fem	Sing/Plur/Inv	-	-
Possessive Pronoun	DET/PRON	Prs	-	-	-/Masc/Fem	Sing/Plur	1/2/3	Yes
Personal Pronoun	PRON	Prs	Acc/Dat/Nom	-	-/Masc/Fem	Sing/Plur	1/2/3	-

Table 2: Specific morphological features for DET and PRON

kind	PoS tag	*Gender*	*Number*	*VerbForm*	*NumType*
Nouns	NOUN	-/Masc/Fem	Sing/Plur/Inv	-	-
Multiplicative Numbers	NOUN	-/Masc/Fem	Sing/Plur/Inv	-	Mult
Adjectives	ADJ	-/Masc/Fem	Sing/Plur/Inv	-	-
Participle Verbs	ADJ	-/Masc/Fem	Sing/Plur/Inv	Part	-
Ordinal Numbers	ADJ	-/Masc/Fem	Sing/Plur/Inv	-	Ord
Cardinal Numbers	NUM	-/Masc/Fem	-	-	Card

Table 3: Specific morphological features for NOUN, ADJ and NUM

kind	PoS tag	*VerbForm*	*Mood*	*Tense*	*Gender*	*Number*	*Person*
Impersonal Infinitive	VERB/AUX	Inf	-	-	-	-	-
Personal Infinitive	VERB/AUX	Inf	-	-	-	Sing/Plur	1/2/3
Gerund	VERB/AUX	Ger	-	-	-	-	-
Participle	VERB/AUX	Part	-	-	-/Masc/Fem	Sing/Plur/Inv	-
Imperative	VERB/AUX	Fin	Imp	-	-	Sing/Plur	1/2/3
Present of Indicative	VERB/AUX	Fin	Ind	Pres	-	Sing/Plur	1/2/3
Imperfect of Indicative	VERB/AUX	Fin	Ind	Past	-	Sing/Plur	1/2/3
Past Perfect of Indicative	VERB/AUX	Fin	Ind	Imp	-	Sing/Plur	1/2/3
Future of Indicative	VERB/AUX	Fin	Ind	Fut	-	Sing/Plur	1/2/3
Pluperfect of Indicative	VERB/AUX	Fin	Ind	Pqp	-	Sing/Plur	1/2/3
Present of Subjuntive	VERB/AUX	Fin	Sub	Pres	-	Sing/Plur	1/2/3
Past of Subjuntive	VERB/AUX	Fin	Sub	Past	-	Sing/Plur	1/2/3
Future of Subjuntive	VERB/AUX	Fin	Sub	Fut	-	Sing/Plur	1/2/3
Conditional Future	VERB/AUX	Fin	Cnd	-	-	Sing/Plur	1/2/3

Table 4: Specific morphological features for AUX and VERB

4. PortiLexicon-UD Data Structure

In order to access the large amount of entries of PortiLexicon-UD, we decided to store the lexical resource as a hierarchical data structure that first delivers all possible PoS tags for each word by an indexing structure. Then, specific entry tables store all words belonging to each of the 13 considered PoS tags. Each of these tables holds all words from the corresponding PoS tag associated to possible lemma and morphological feature pairs. Figure 1 depicts the proposed data structure.

The indexing structure is basically a Decision Diagram structure (Lucchesi and Kowaltowski, 1993) that, to each word, uses its letters to index the possible PoS tags associated to it. Especifically, this data structure is stored using a MTMDD - Multi Terminal Multivalue Decision Diagram (Fernandes et al., 2012), which delivers the more efficient way to determine all possible PoS tags for each known word. All words of the lexical resource are associated to a non-empty set of the 13 considered UD PoS tags. For example, the word "*caso*"

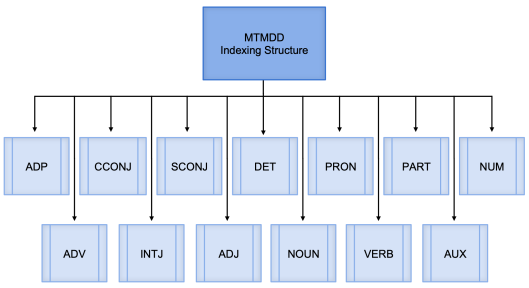

Figure 1: PortiLexicon-UD data structure

can be associated to the following three UD PoS tags:

- NOUN, if translated to the English noun "case" (for example: "*Apresento meu caso*" - "I state my case");
- SCONJ, if translated to the English subordinating conjunction "if" (for example: "*Caso sobreviva, irei*" - "If I survive, I will go"); or

- VERB, if translated to the English verb "(I) marry" (for example: "*Eu caso com ela*" - "I marry her").

This indexing structure is made available as a large alphabetically sorted textual file with a word per line, separated by a comma with all possible PoS tags separated by blank spaces. For example, the line corresponding to the word "*caso*" will be:

caso,NOUN SCONJ VERB

Once we know all possible PoS tags, we access the lemma and morphological features of each word consulting one of the 13 tables that list all possible words of each PoS tag. Given the kind of information stored, these tables are efficiently implemented using a sorted entry table that can be efficiently accessed by a standard binary search, delivering a set of pairs (lemma, morphological features) to each word. That being said, other structures as hash tables or heaps could be adapted with a similar efficiency. Each of these tables is made available in an alphabetically sorted .tsv textual file with one line per possible pair. In this line, the word, the lemma, and the morphological information are separated by a tab ("\t"), being the morphological features separated by the vertical bar ("|").
For example, in the table of PRON, the word "*quem*" has three entries associated to it (three lines in the textual file) - one for when it is employed as indicative, another as interrogative, and another as relative pronoun:

quem quem Number=Inv|PronType=Ind

quem quem Number=Inv|PronType=Int

quem quem Number=Inv|PronType=Rel

Another example is the word "*canta*" in the table of VERB, where it appears as present of the indicative ("(he) sings") or imperative ("sing (you)"), thus corresponds to two lines in the textual file:

canta cantar Mood=Ind|Number=Sing|Person=3|Tense=Pres|VerbForm=Fin

canta cantar Mood=Imp|Number=Sing|Person=2|VerbForm=Fin

Overall, the resulting lexical resource is composed by fourteen files: one master file representing 858,992 different words and the possible PoS tags associated to each of them, plus 13 files with the words of each of the 13 PoS classes with the corresponding possible pairs of lemmas and morphological features. Table 5 presents the number of unique words in each of the files (column *total*), stating how many are ambiguous words, with more than one possible PoS tag (column *amb.*), or not (column *non-amb.*), plus the number of entries in each file (if a word has ambiguity of lemma or morphological features, it generates more than one entry in the file). Finally, the last row of Table 5 summarizes the total number of unique words, splitting the number of words in how many of those are ambiguous or not regarding the PoS tags, plus the total number of entries, simply adding all entries for all 13 PoS tags files.

	number of words			
PoS tag	*total*	*amb.*	*non-amb.*	*entries*
ADP	38	25	13	38
CCONJ	22	14	8	22
SCONJ	14	13	1	14
DET	127	119	8	135
PRON	160	128	32	173
PART	1	1	0	1
NUM	49	38	11	49
ADV	4,936	1,834	3,102	4,942
INTJ	40	9	31	40
ADJ	120,338	77,880	42,458	120,448
NOUN	74,373	25,361	49,012	74,544
VERB	746,621	65,528	681,093	1,020,448
AUX	275	194	81	364
total	858,922	83,072	775,850	1,221,218

Table 5: The number of words in each PoS tag and entries

With PortiLexicon-UD stored in such data structure, the time complexity to access the entries is logarithmic. The search of the words' possible PoS tags takes in the worst case scenario a number of tests that is bounded by the size of the largest word in the lexical resource, which is "*cineangiocoronariográficas*" that has 26 characters, thus taking at the worst case 26 tests to find the possible PoS tags of a word.
The search within the PoS tag specific tables is bounded by a binary search in the largest file (VERB), which is $20 > log_2 1,020,448$. In terms of efficiency, the retrieval of all possible annotations for every word belonging to our lexical resource takes 1.1034 seconds to retrieve 2 million words in a modest portable machine with i7 2.2 GHz processor, 8 Gbytes memory running MacOS 11.6. This experiment was conducted processing 20 random packages of 2 million words and averaging the time took for the 16 median packages, disregarding the two lowest and the two highest times.

5. PortiLexicon-UD Application

To illustrate the use of PortiLexicon-UD, our proposed lexical resource, we employ it to verify the annotation of a Portuguese corpus annotated with UD and manually revised, composed by 168,399 tokens distributed in 8,420 sentences extracted from Folha Kaggle corpus[3]. The 131,710 tokens of this corpus belonging to the 13 UD PoS tags covered were searched in our lexical resource and five possible outcomes were obtained:

- the word was absent from our lexicon (*absent*);

[3] https://www.kaggle.com/marlesson/news-of-the-site-folhauol

- the word was present in our resource and one single combination of PoS tag, lemma and morphological features was found matching the corpus definition (*non-ambiguous OK*);
- the word was present in our resource and multiple combinations of PoS tag, lemma and morphological features were found, and one of them matched the corpus definition (*ambiguous OK*);
- the word was present in our resource and one single combination of PoS tag, lemma and morphological features was found, but not matching the corpus definition (*non-ambiguous NOK*);
- the word was present in our resource and multiple combinations of PoS tag, lemma and morphological features were found, but none of them matched the corpus definition (*ambiguous NOK*).

Table 6 presents the outcome of our experiment. The first observation from these results is that the lexical resource coverage is very good, as less than one percent of the words was absent (0.53%). Some of these absent words are abbreviations that are commonly employed, but were not covered in the lexicon, for example, "*TV*", "*km*", and "*min*", which in Portuguese are also used to abbreviate "television", "kilometer'', and "minute", respectively. Another group of absent words are misspelled words, as wrong forms of infinitive verbs, as "*exercê*" instead of "*exercer*" ("to practice").

outcome	*# tokens*	*%*
absent	692	0.53%
non-ambiguous OK	50,004	37.97%
ambiguous OK	79,339	60.24%
non-ambiguous NOK	358	0.27%
ambiguous NOK	1,317	1.00%

Table 6: Outcome for the experiment of corpus search of 131,710 tokens in PortiLexicon-UD

Another observation from Table 6 is the large number of tokens (37.97%) that could be correctly identified unambiguously for PoS tag, lemma, and morphological features (*non-ambiguous OK*) using a straightforward technique (Lopes et al., 2021). Assuming a disambiguation technique, it is also possible to correctly tag other 60.24% of the words, for the tokens where there was a fully correct option among the PortiLexicon-UD's options (*ambiguous OK*).

Finally, the number of tokens that could not be correctly identified (*non-ambiguous* and *ambiguous NOK*) because no valid annotation alternative was offered by our lexical resource represents about one percent of the tokens (0.27% plus 1.00%).

Individually observing the matching rates for PoS tag, lemma, and morphological features, and considering that either a non-ambiguous correct option or one correct option among ambiguous ones is offered by PortiLexicon-UD, we have the results stated at Table 7, that take into account only the 131,018 (131,710 minus 692) tokens found. This table's results illustrate the quality of our lexical resource to provide the morphosyntactic encoding using UD format.

aspect	*# tokens*	*%*
PoS tag	130,652	99.72%
Lemma	130,742	99.79%
Morph. Feat.	130,158	99.35%

Table 7: Individual correct tokens for PoS tag, lemma, and morphological features for the 131,018 corpus' tokens found in PortiLexicon-UD during the experiment.

6. Conclusion

This paper presented PortiLexicon-UD, a lexical resource for Portuguese delivering morphosyntactic information using UD tags and principles. The novelty of such resource is a benefit in itself given the scarcity of UD resources for Portuguese. Our preliminary experiment shows that PortiLexicon-UD has a good coverage and quality of the delivered information. In particular, our example application showed that almost all searched tokens were found (over 99%). The quality of the delivered information was also good according to the same example application, as for PoS tag, lemma, and morphological features a correct option was offered for nearly all found tokens (nearly 99%).

Future work naturally includes a deeper analysis of words not found in PortiLexicon-UD, with the possible inclusion of additional words. For example, the inclusion of absent words as abbreviations, neologisms, and new words found in new corpora is a good source for improvements.

The analysis of differences of annotation in other UD Portuguese corpora may also bring interesting insights about the annotation decisions taken. From a computational point of view, the addition of new entries to the lexical resource is not an issue, given the efficiency of PortiLexicon-UD's data structure. In fact, the development of lexical resources is never really finished as languages never stop evolving. Nevertheless, the availability of PortiLexicon-UD may already boost the integration of Portuguese with UD initiatives.

The lexicon is publicly available for search and download at `https://portilexicon.icmc.usp.br/`. For the interested reader, more information about this work is available at the web portal of the PO-eTiSA project: `https://sites.google.com/icmc.usp.br/poetisa`.

Acknowledgements

The authors thank the Center for Artificial Intelligence of the University of São Paulo (C4AI[4]), sponsored by IBM and FAPESP (grant #2019/07665-4).

[4] `http://c4ai.inova.usp.br/`

7. Bibliographical References

Cecchini, F. M., Sprugnoli, R., Moretti, G., and Passarotti, M. (2020). Udante: First steps towards the universal dependencies treebank of dante's latin works. In Johanna Monti, et al., editors, *Proceedings of the Seventh Italian Conference on Computational Linguistics, CLiC-it 2020*, volume 2769 of *CEUR Workshop Proceedings*, Bologna, Italy. CEUR-WS.org.

Duran, M. S. (2021). Manual de anotação de pos tags: Orientações para anotação de etiquetas morfossintáticas em língua portuguesa, seguindo as diretrizes da abordagem universal dependencies (UD). Technical Report 0103-2569, ICMC-USP.

Fernandes, P., Lopes, L., Prolo, C. A., Sales, A., and Vieira, R. (2012). A fast, memory efficient, scalable and multilingual dictionary retriever. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2520–2524, Istanbul, Turkey, May. European Language Resources Association (ELRA).

Figueiredo de Alencar, L., Cuconato, B., and Rademaker, A. (2018). MorphoBr: an open source large-coverage full-form lexicon for morphological analysis of portuguese. *Texto Livre*, 11(3):1–25, dez.

Lopes, L., Duran, M. S., and Pardo, T. A. S. (2021). Universal dependencies-based pos tagging refinement through linguistic resources. In *Proceedings of the 10th Brazilian Conference on Intelligent System*, BRACIS'21.

Lucchesi, C. L. and Kowaltowski, T. (1993). Applications of finite automata representing large vocabularies. *Softw. Pract. Exper.*, 23(1):15–30, jan.

McDonald, R., Nivre, J., Quirmbach-Brundage, Y., Goldberg, Y., Das, D., Ganchev, K., Hall, K., Petrov, S., Zhang, H., Täckström, O., Bedini, C., Bertomeu Castelló, N., and Lee, J. (2013). Universal Dependency annotation for multilingual parsing. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 92–97, Sofia, Bulgaria, August. Association for Computational Linguistics.

Miletic, A., Bras, M., Vergez-Couret, M., Esher, L., Poujade, C., and Sibille, J. (2020). Building a Universal Dependencies treebank for Occitan. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 2932–2939, Marseille, France, May. European Language Resources Association.

Muniz, M. C. M. (2004). A construção de recursos linguístico-computacionais para o português do brasil: o projeto Unitex-PB. Master's thesis, Instituto de Ciências Matemáticas e de Computação - Universidade de São Paulo - ICMC/USP.

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.

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

Przepiórkowski, A. and Patejuk, A. (2018). From Lexical Functional Grammar to enhanced Universal Dependencies. In *Proceedings of the Joint Workshop on Linguistic Annotation, Multiword Expressions and Constructions (LAW-MWE-CxG-2018)*, pages 2–4, Santa Fe, New Mexico, USA, August. Association for Computational Linguistics.

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)*, pages 197–206, Pisa, Italy, September.

Ranchhod, E., Mota, C., and Baptista, J. (1999). A computational lexicon of Portuguese for automatic text parsing. In *SIGLEX99: Standardizing Lexical Resources*.

Seddah, D., Essaidi, F., Fethi, A., Futeral, M., Muller, B., Ortiz Suárez, P. J., Sagot, B., and Srivastava, A. (2020). Building a user-generated content North-African Arabizi treebank: Tackling hell. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1139–1150, Online, July. Association for Computational Linguistics.

Uszkoreit, H., Macketanz, V., Burchardt, A., Harris, K., Marheinecke, K., Petrov, S., Kayadelen, T., Attia, M., Elkahky, A., Yu, Z., Pitler, E., Lertpradit, S., Mendonça, G., Rinaldi, L., Popel, M., Zeman, D., de Paiva, V., and Rademaker, A. (2017). Ud portuguese pud. `https://github.com/UniversalDependencies/UD_Portuguese-PUD`. Accessed: 2022-01-11.