A UD Parser to the Rescue: A Method for Bringing a Classical Annotated Corpus to Life Again

Lucelene Lopes and Magali S. Duran and Thiago A. S. Pardo
Núcleo Interinstitucional de Linguística Computacional, Universidade de São Paulo – Brazil
lucelene@gmail.com magali.duran@gmail.com taspardo@icmc.usp.br

Abstract

This paper reports on an effort to recover the classical morphosyntactically annotated corpus MacMorpho and realign it with the current version of the Universal Dependencies framework. We introduce a knowledge-rich approach grounded in a syntactic parser and on a specially designed tagset compatibility strategy in order to generate a "silver-standard" resource: the MacMorpho-UD-2.17. We evaluate this resource through multiple complementary methods, providing evidence for the quality of both our approach and the resulting annotation.

1 Introduction

The field of Natural Language Processing (NLP) has seen great advances in the last decade, including not only the now widely spread language models, but also new representation languages to codify linguistic information. Naturally, the way NLP researchers tackled many applications changed drastically and, consequently, the computational linguistic tools for preprocessing and preparing the data had to follow the new directions and linguistic resources had to be created or refactored.

The Universal Dependencies (UD) initiative (Nivre et al., 2016; de Marneffe et al., 2021) is one of the achievements of the last ten years. It is a framework for "consistent annotation of grammar", following standard annotation decisions across languages, currently including over 200 treebanks for more than 150 languages in its 2.17 version[1].

There have been several efforts to bring UD to Portuguese, as the pioneering annotation of Bosque (Rademaker et al., 2017), a journalistic corpus, and Porttinari (Pardo et al., 2021; Duran et al., 2023), a recently assembled multigenre corpus (news texts, user-generated content, legal texts, literature, and transcribed speech) that follows more recent annotation directions (Duran et al., 2022) for this language. New tools have also been developed/trained, as state-of-the-art part of speech tagger (Silva et al., 2021) and parser (Lopes and Pardo, 2024).

In this context, this paper reports our endeavor to develop and instantiate a method in order to rescue a valuable resource for Portuguese that got lost along the advances that NLP went through – the MacMorpho corpus (Aluísio et al., 2003) –, aligning it to the current UD framework. MacMorpho has probably been one of the main references for morphosyntactic (part of speech) annotation in Portuguese. It adopted a tagset of its own and, since then, it was used for linguistic studies and development of important part of speech taggers in NLP history for Portuguese. As Bonn et al. (2024) argue, the "long-term preservation [of language resources] is a must, especially for valuable resources such as manually annotated data".

A previous initiative converted the first version of MacMorpho to UD (Freitas et al., 2018), producing the MACMORPHO-UD corpus, but it did not evolve alongside later changes in MacMorpho and UD guidelines. Our contributions address this gap:

- Anchoring on a state of the art UD parser for Portuguese and on a specially developed tagset compatibility strategy, we propose a methodological process to revitalize MacMorpho, using annotation agreement as the main decision parameter, partially inspired by Freitas and de Souza (2023).
- Following the above process, we make available an all-new UD-annotated version of MacMorpho, which can be considered a "silver-standard" resource (in opposition to a gold-standard one, since our method employs a semi-automatic approach and some errors may persist) and that we name **MacMorpho-UD-2.17** (to honor the original conversion initiative and to make explicit its alignment to the current UD version).

[1] https://universaldependencies.org/

Proceedings of the 17th International Conference on Computational Processing of Portuguese (PROPOR 2026)
Vol. 1, pages 240–249
April 13-16, 2026

- Assessing our approach and produced resource in different ways (including comparing the productivity of our method with another one, training an NLP tool over the new data, and evaluating the resource in a data augmentation scenario), we evidence the quality of the method and the data.

This paper is structured as follows. Section 2 briefly synthesizes UD fundamentals and MacMorpho history. Section 3 presents the method that we propose to recover the corpus. Section 4 evaluates the efficacy of our method and what was achieved, while Section 5 concludes this paper.

2 Background

The Universal Dependencies (UD) project aims at establishing consensual tagsets and annotation strategies that work across languages of different families. Basically, it provides morphological features, part of speech tags and dependency relations for corpus annotation and creation of treebanks.

UD data is organized in CoNLL-U files, which are plain text files with information regarding each token of a sentence structured in columns (Table 1). The relevant ones for this paper are LEMMA, UPOS and FEATS. While LEMMA and FEATS are generated through our conversion process (as they were not available in the original MacMorpho corpus), syntactic dependency relations are not included, since MacMorpho does not provide this level of annotation; therefore, the HEAD and DEPREL fields remain empty. The ten fields of the format are left for the interested reader to explore[2], as well as the tagsets of UD[3] and MacMorpho[4].

Column	Field	Description
1	ID	Token index
2	FORM	Word form or punctuation
3	**LEMMA**	Lemma or stem
4	**UPOS**	Universal POS tag
5	XPOS	Language-specific POS tag
6	**FEATS**	Morphological features
7	HEAD	Syntactic head (0 = root)
8	DEPREL	Relation to head
9	DEPS	Enhanced dependencies
10	MISC	Miscellaneous info

Table 1: The ten CoNLL-U columns.

[2] https://universaldependencies.org/format.html

[3] https://universaldependencies.org/u/pos/index.html

[4] https://drive.google.com/file/d/1mOri6rggdh-mOkQyzNLWrSE32p_wPAjo/view

Currently, the UD initiative includes the Portuguese treebanks shown in Table 2. As may be noticed, such corpora represent relevant efforts for this language, positioning the language among the top-5 languages in terms of number of annotated tokens (after Czech, German, Russian and Japanese). The best positioned language - Czech - has almost 3 times more data than Portuguese.

Treebank	Sentences	Tokens
Bosque	9,357	210,958
GSD	12,020	296,169
CINTIL	38,400	441,991
PUD	1,000	21,917
PetroGold	8,949	250,595
Porttinari	8,418	168,080
DANTEStocks	4,042	80,997

Table 2: Portuguese treebanks available in UD version 2.17.

Although having a version that was previously converted to UD, the MacMorpho corpus is not included in the list because it brings only Part Of Speech (POS) tags.

MacMorpho was first published in 2003 (Aluísio et al., 2003) as a corpus with 53,374 sentences (1,221,507 tokens). This version – called v.1 – had a set of 22 POS tags, not assigning tags for punctuation tokens. The basic codification format used by MacMorpho is a simple text file with one sentence per line and each individual token separated from the next one by a blank space, where each token is followed by an underscore symbol and the corresponding POS tag. For example, the sentence "*Eles estão colocando armadilhas nas fazendas onde já ocorreram os ataques.*" is represented in the first row of Table 3.

After a first revision, removing inconsistent sentences and adjusting sentence segmentation and tokenization issues, the version v.2 was published in 2013 (Fonseca and Rosa, 2013). MacMorpho v.2 reduced the number of sentences to 49,990 (1,011,479 tokens, 949,958 words), and created an additional tag for punctuation (**PU**), totalizing 23 POS tags. However, the more notable change was the replacement of token by word representation, which impacts contracted words as "*nas*", which was previously annotated as two tokens "*em as*". For such words, the two corresponding POS tags were annotated after the contracted word, for example: "*nas*_**PREP+ART**". In Table 3, the second row illustrates the same example sentence as encoded in v.2.

The last public revision performed a simi-

v.1	Eles_PROPESS estão_VAUX colocando_V armadilhas_N **em_PREP as_ART** fazendas_N onde_ADV-KS-REL já_ADV ocorreram_V os_ART ataques_N ._.
v.2	Eles_PROPESS **estão_VAUX** colocando_V armadilhas_N **nas_PREP+ART** fazendas_N **onde_ADV-KS-REL** já_ADV ocorreram_V os_ART ataques_N **._PU**
v.3	Eles_PROPESS **estão_V** colocando_V armadilhas_N nas_PREP+ART fazendas_N **onde_ADV-KS** já_ADV ocorreram_V os_ART ataques_N ._PU

Table 3: Encoding of the sentence "*Eles estão colocando armadilhas nas fazendas onde já ocorreram os ataques.*" in the three MacMorpho versions.

lar removal of inconsistencies and segmentation/tokenization adjustments remaining in v.2, thus reducing the number of sentences to 47,935 (977,646 tokens, 911,765 words), as published in 2015 (Fonseca et al., 2015). The more notable changes were the merge of POS tags **ADV-KS-REL** and **ADV-KS**, **PRO-KS-REL** and **PRO-KS**, and **VAUX** and **V**, dropping the number of tags to 20 POS tags. To illustrate the practical change, the example sentence encoded for MacMorpho v.3 is stated in the third row of Table 3.

Later on, Freitas et al. (2018) presented a conversion of the 1st version of MacMorpho (v.1) to UD, directly mapping its POS tags into UPOS tags. All original tags were converted, except for the tag PCP (participle), which was disambiguated into **NOUN**, **VERB**, or **ADJ** following general guidelines and occasional case-specific decisions. The resulting resource, MACMORPHO-UD, represented an important initial effort to align Brazilian Portuguese corpora with UD standards. However, as commented before, it reflects earlier versions of both the MacMorpho corpus and the UD guidelines, which have undergone substantial refinement within the research community.

Next section presents the method that we propose to revitalize MacMorpho.

3 Methodology

Our methodology to produce the silver-standard MacMorpho-UD-2.17 corpus includes three steps: data preparation, automatic corpus annotation, and annotation sifting.

3.1 Data preparation

This initial step aims to adapt the legacy corpus to the goal framework structure – the CoNLL-U format.

From a practical point of view, we started with the last released version of MacMorpho dataset, v.3, which holds 47,935 sentences and 977,646 tokens, preserving its original sentence segmentation and tokenization. Typographic issues such as missing spaces and tag misspellings were corrected, and the corpus was automatically converted into CoNLL-U format through three main transformations:

- **Sentence identification**: Each sentence was assigned a unique ID in the format `# sent_id = MACMORPHO_S000000` for the first sentence and so on until the last one (`MACMORPHO_S047934`);
- **Text reconstruction**: The original sentence text was reconstructed and included in the line `# text =...`, with proper spacing and punctuation restored. Additionally, as required by the file format, we included "*SpaceAfter=No*" in the MISC field of tokens when applicable;
- **Contraction expansion**: Contracted forms (e.g., `do_PREP+ART`) were expanded into two tokens and preceded by the expected fused-form line in the CoNLL-U representation.

For example, the following sentence was translated into the CoNLL-U format shown in Figure 1.

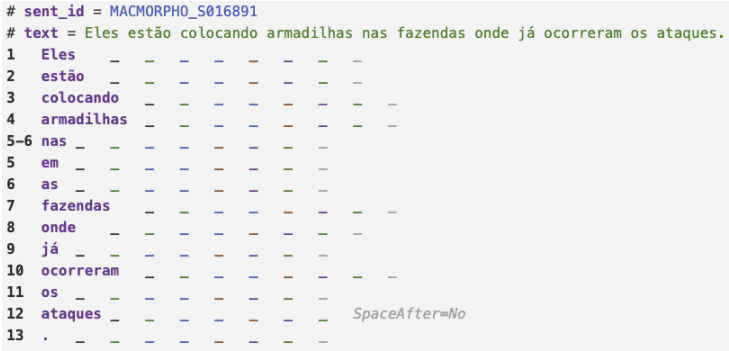
```
Eles_PROPESS estão_V colocando_V
armadilhas_N nas_PREP+ART
fazendas_N onde_ADV-KS já_ADV
ocorreram_V os_ART ataques_N
._PU
```

3.2 Automatic corpus annotation

This step requires a high quality tool to automatically annotate the prepared corpus.

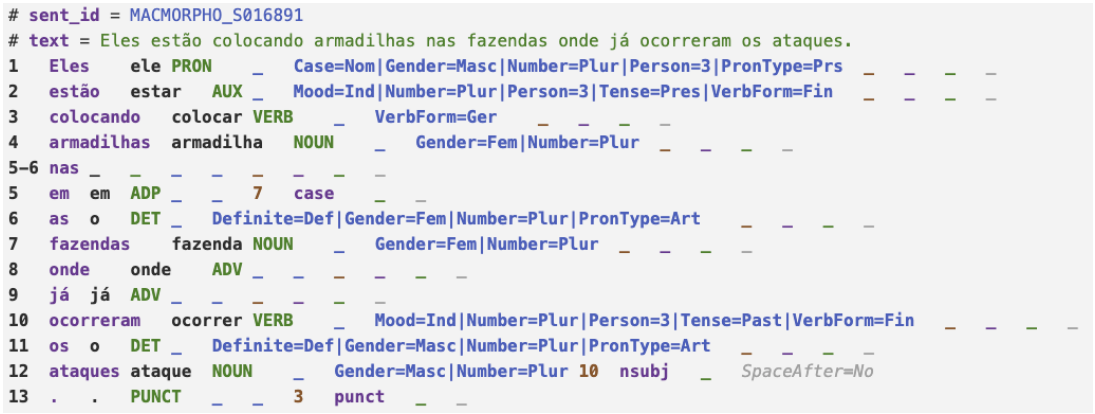
We used the last release (v2) of Portparser[5] (Lopes and Pardo, 2024), a high-precision UD parser for Brazilian Portuguese trained with the Porttinari-base corpus (Duran et al., 2023) (the parser currently achieves over 96% accuracy for news texts). This version of the parser follows the LatinPipe framework (Straka et al., 2024) adapted to Portuguese and performs post-processing adjustments on lemmas and morphological features using PortiLexicon-UD (Lopes et al., 2022), a large lexicon for Portuguese that is aligned to the UD framework. The produced annotation of this pipeline shows about 99% precision for UPOS, LEMMA, and FEATS fields.

[5] https://huggingface.co/spaces/NILC-ICMC-USP/Portparser.v2



```
# sent_id = MACMORPHO_S016891
# text = Eles estão colocando armadilhas nas fazendas onde já ocorreram os ataques.
1   Eles       _  _  _  _  _  _  _  _
2   estão      _  _  _  _  _  _  _  _
3   colocando  _  _  _  _  _  _  _  _
4   armadilhas _  _  _  _  _  _  _  _
5-6 nas        _  _  _  _  _  _  _  _
5   em         _  _  _  _  _  _  _  _
6   as         _  _  _  _  _  _  _  _
7   fazendas   _  _  _  _  _  _  _  _
8   onde       _  _  _  _  _  _  _  _
9   já         _  _  _  _  _  _  _  _
10  ocorreram  _  _  _  _  _  _  _  _
11  os         _  _  _  _  _  _  _  _
12  ataques    _  _  _  _  _  _  _  SpaceAfter=No
13  .          _  _  _  _  _  _  _  _
```

Figure 1: Sentence "Eles_PROPESS estão_V colocando_V armadilhas_N nas_PREP+ART fazendas_N onde_ADV-KS já_ADV ocorreram_V os_ART ataques_N ._PU" from MacMorpho v.3 in the produced CoNLL-U format.



```
# sent_id = MACMORPHO_S016891
# text = Eles estão colocando armadilhas nas fazendas onde já ocorreram os ataques.
1   Eles       ele      PRON  _  Case=Nom|Gender=Masc|Number=Plur|Person=3|PronType=Prs                    _  _  _  _
2   estão      estar    AUX   _  Mood=Ind|Number=Plur|Person=3|Tense=Pres|VerbForm=Fin                    _  _  _  _
3   colocando  colocar  VERB  _  VerbForm=Ger                                                             _  _  _  _
4   armadilhas armadilha NOUN  _  Gender=Fem|Number=Plur                                                  _  _  _  _
5-6 nas        _        _     _  _                                                                        _  _  _  _
5   em         em       ADP   _  _                                  7   case                             _  _
6   as         o        DET   _  Definite=Def|Gender=Fem|Number=Plur|PronType=Art                         _  _  _  _
7   fazendas   fazenda  NOUN  _  Gender=Fem|Number=Plur                                                  _  _  _  _
8   onde       onde     ADV   _  _                                                                        _  _  _  _
9   já         já       ADV   _  _                                                                        _  _  _  _
10  ocorreram  ocorrer  VERB  _  Mood=Ind|Number=Plur|Person=3|Tense=Past|VerbForm=Fin                    _  _  _  _
11  os         o        DET   _  Definite=Def|Gender=Masc|Number=Plur|PronType=Art                        _  _  _  _
12  ataques    ataque   NOUN  _  Gender=Masc|Number=Plur                10  nsubj                        _  SpaceAfter=No
13  .          .        PUNCT _  _                                      3   punct                        _  _
```

Figure 2: Sentence "Eles_PROPESS estão_V colocando_V armadilhas_N nas_PREP+ART fazendas_N onde_ADV-KS já_ADV ocorreram_V os_ART ataques_N ._PU" from MacMorpho v.3 after annotation by Portparser.v2.

The result was an automatically POS tagged corpus of 47,935 sentences. To illustrate the result of this process, the annotation of sentence in Figure 1 is shown in Figure 2. Even though the parser produced the annotation of the fields HEAD and DEPREL, those were discarded and only the relevant columns for this paper were kept.

3.3 Annotation sifting

This is a crucial step responsible for producing the annotated resource with the necessary quality control. We use the figurative sense of "sifting" to refer to the filtering of the annotated sentences.

The sifting strategy consists in (i) checking if the UPOS tags predicted by the parser are compatible with the POS manually annotated in MacMorpho v.3 and (ii) selecting only the sentences in which all tokens show compatible annotation in order to compose the revitalized UD-aligned MacMorpho. Our hypothesis is that it is unlikely that the legacy corpus and the automatic annotation make the same mistakes; therefore, if they show compatible tags, the annotation is probably correct.

First of all, it was necessary to propose the compatibility of MacMorpho v.3 POS tags with UPOS tags. We followed a conservative approach, assuming only safe compatibility options, leveraging the knowledge of previous annotation experiences and the expertise of a specialized linguist. Table 4 synthesizes the proposed compatibility of each of the 20 MacMorpho v.3 POS tags with the UPOS tags.

Some of the assumptions underlying such decisions deserve to be discussed:

- MacMorpho v.3 annotation assigns POS tags **N**, **ADJ**, **ADV** and others regardless of whether the words are foreign or not. In UD, foreign words are tagged as **X**. So, in UD standard, we could not establish direct compatibility of **X** with specific MacMorpho v.3 POS tag. Therefore, except for non-alphanumeric options (symbols and punctuations), we admitted the UPOS **X** as compatible with all MacMorpho v.3 POS tags.

MacMorpho v.3	UPOS	Examples
ADJ	ADJ - PROPN - X	alto_ADJ
ADV	ADV - X	mais_ADV
ADV-KS	SCONJ - ADV - X	onde_ADV-KS
ART	DET - X	o_ART
CUR	SYM	Cr$_CUR
IN	INTJ - X	alô_IN
KC	CCONJ - X	ou_KC
KS	SCONJ - X	caso_KS
N (alpha)	NOUN - NUM - PROPN - X	céu_N
N (symb.)	SYM	%_N
NPROP	PROPN - X	Paulo_NPROP
NUM	NOUN - NUM - X	101_NUM
PCP	AUX - VERB - NOUN - ADJ - X	oculto_PCP
PDEN	ADV - X	também_PDEN
PREP	ADP - X	de_PREP
PROADJ	DET - X	outra_PROADJ
PROPESS	PRON - X	ela_PROPESS
PROSUB	PRON - X	isto_PROSUB
PRO-KS	PRON - DET - SCONJ - X	cujo_PRO-KS
PU	PUNCT	..._PU
V	VERB - AUX - X	foi_V

Table 4: MacMorpho v.3 tagset mapping into UD tagset.

- Proper Nouns are related to named entities and, when a named entity consists of several tokens, how to annotate them is a project decision. The automatic UD annotation we performed uses **PROPN** for each upper case token that integrates a compound proper noun. For example, in "*Agência Nacional de Vigilância Sanitária (ANVISA)*" (National Health Surveillance Agency), all tokens except the preposition "*de*" are tagged as **PROPN**. In MacMorpho v.3, the same named entity is tagged as **N ADJ PREP N ADJ** and only the acronym "*ANVISA*" is tagged as **NPROP**. Therefore, it is expected that **NPROP** and **PROPN** will not be compatible in some cases, and we do not consider this an error. Therefore, we allowed the definition of MacMorpho v.3 POS tags **NPROP**, **N**, and **ADJ** to be compatible with the UPOS **PROPN**.

- The MacMorpho v.3 criteria for choosing the POS tags **N** and **NUM** do not follow a clear definition as in UD standards, so we admitted these MacMorpho v.3 POS tags to be compatible with UPOS **NOUN** and **NUM**.

- The agglutination of MacMorpho v.3 POS tags **V** and **VAUX** into **V** led us to admit those tags to be compatible with either UPOS **VERB** or **AUX**.

- The MacMorpho v.3 POS tag **PRO-KS** refers to relative pronouns; in UD, the relative pronouns are frequently tagged as **PRON**, but the relative pronoun "whose", specifically, is tagged **DET**. There are also pronouns used for topicalization, which are tagged as **SCONJ**. We thus admitted the compatibility of **PRO-KS** with the UPOS **PRON**, **DET**, and **SCONJ**.

- The MacMorpho v.3 POS tag **ADV-KS**, although frequently representing an adverb role, can eventually be employed with subordinative conjunction role. We thus admitted the compatibility of **ADV-KS** with the UPOS **ADV** and **SCONJ**.

- The MacMorpho v.3 POS tag **PCP** is the one with more options, as the UD standard does not contemplate the use of participles as a UPOS, but associates such words with verbs (main and auxiliary), nouns, or adjectives. Therefore, we admitted the compatibility of **PCP** with the UPOS **AUX**, **VERB**, **NOUN**, and **ADJ**.

- The MacMorpho v.3 POS tag **PDEN** is compatible with the UPOS **ADV** when it refers to a single token. However, when **PDEN** is assigned to a fixed expression, MacMorpho v.3 POS tags may correspond to several UPOS, as fixed expressions are not annotated at morphosyntactic level in UD. For example, the expression "*isto é*" (that is), is annotated as **PDEN PDEN** in MacMorpho v.3 and as **PRON VERB** in UD. Therefore, it is not possible to predict the UPOS of tokens annotated with **PDEN** that are part of a fixed expression. For this reason, we disregarded sentences where **PDEN** does not correspond to **ADV**. These cases require another approach and must be addressed in future work.

- All remaining MacMorpho v.3 POS tags have direct compatibility with specific UPOS, except for a very specific situation where the MacMorpho v.3 POS tag **N** is employed for tokens that consist of non-alphanumeric characters, in which case we only admit the UPOS **SYM**.

Considering this tagset compatibility proposal, all tokens in the sentences of MacMorpho v.3 had their predicted UPOS (by the parser) compared to the POS tags originally assigned in MacMorpho v.3. If a sentence had all the tokens with compatible annotation, the sentence annotation (with the predicted UPOS) was considered correct and

included in the new version of the UD-based MacMorpho. If any token did not observe the possible compatibility, the sentence was discarded. As an illustration, the sentence in Figure 2 was kept, as its tokens show compatible UPOS with the original POS tags, as exhibited below:

token	UPOS	compatibility	MacMorpho v.3 POS
Eles	**PRON**	only option for	**PROPESS**
estão	**AUX**	one of the options for	**V**
colocando	**VERB**	one of the options for	**V**
armadilhas	**NOUN**	one of the options for	**N**
em	**ADP**	the only option for	**PREP**
as	**DET**	the only option for	**ART**
fazendas	**NOUN**	one of the options for	**N**
onde	**ADV**	one of the options for	**ADV-KS**
já	**ADV**	the only option for	**ADV**
ocorreram	**VERB**	one of the options for	**V**
os	**DET**	the only option for	**ART**
ataques	**NOUN**	one of the options for	**N**
.	**PUNCT**	the only option for	**PU**

This sifting process resulted in a corpus named **MacMorpho-UD-2.17**, and composed of 30,581 sentences (517,783 tokens - 484,213 words), which corresponds to 63.8% of the sentences from MacMorpho v.3. The reduction in size is due to the filtering criteria applied during the sifting process. Overall, these sentences have an average of 16.93 tokens per sentence, with 65% of the sentences containing between 4 and 21 tokens, indicating a representative coverage of typical sentence lengths in the corpus. The distribution of UPOS tags in these sentences is presented in Table 5.

NOUN	99,773	**DET**	77,725
ADP	76,628	**PUNCT**	69,233
VERB	49,092	**PROPN**	42,573
ADJ	25,308	**ADV**	15,054
PRON	15,052	**AUX**	14,037
NUM	12,322	**CCONJ**	12,038
SCONJ	5,258	**SYM**	2,905
X	703	**INTJ**	82

Table 5: **MacMorpho-UD-2.17** UPOS distribution.

4 Evaluation

To assess our method, we conducted three experiments:

- compatibility validity, evaluating the adequacy of the proposed correspondences between UPOS and MacMorpho v.3 POS tags;
- sifting effectiveness, analyzing the effectiveness of our filtering process in constructing the silver-standard corpus (**MacMorpho-UD-2.17**), in comparison with another more direct strategy to perform the same task;
- usefulness for practical purposes, testing the silver-standard corpus as a resource for training an NLP tool and as additional data for data augmentation.

4.1 Compatibility validity

The first analysis assesses how well the compatibility of UPOS with MacMorpho v.3 POS tags (Table 4) preserves grammatical information. We computed the number of tokens from each original MacMorpho v.3 POS tag that were automatically annotated with each UPOS. The results are shown in Table 6. As observed, the majority of the cases follow the expected compatibility, confirming that our compatibility decisions effectively captured the functional equivalence of the original tags.

One may see that most **ADJ**, **ADV**, and **ART** tokens were consistently annotated as UPOS **ADJ**, **ADV**, and **DET**, respectively. Similarly, **V** tokens were split between UPOS **VERB** and **AUX**, with 41,679 and 13,867 occurrences, respectively. A small proportion of ambiguous or multifunctional tags such as **PRO-KS**, **ADV-KS**, and **PCP** were distributed across multiple UPOS, reflecting genuine syntactic variability rather than compatibility noise.

To help ensuring the correctness of the automatic annotations, all sentences were subsequently automatically checked using the Verifica-UD tool (Lopes et al., 2023), as well as the official UD validate tool[6]. Such tools help guaranteeing that UD annotations follow the directions of the initiative. We could observe that the converted corpus fully complies with UD standards and that no structural or labeling inconsistencies remained.

4.2 Sifting effectiveness

To assess the effectiveness of our compatibility-based sentence-sifting procedure, we compare the silver-standard corpus produced by our method with the corpus produced by a simpler, more direct approach. This alternative "baseline" method compares the automatically parsed MacMorpho v.3 with the MACMORPHO-UD of Freitas et al. (2018) (derived from MacMorpho v.1) and retrieves sentences with the same UPOS annotation, discarding the other ones.

To allow this comparison of methods, we began by identifying the sentences that are strictly identical across the two MacMorpho versions (v.1 and v.3), based on exact token by token form matching. Out of the 53,374 sentences in v.1 and the 47,935 sentences in v.3, a total of 32,776 sentences share the same token sequence. Consequently, all sub-

[6] https://universaldependencies.org/contributing/validation.html

MacMorpho v.3	Portparser automatic annotation - UPOS															
POS Tag	ADJ	ADP	ADV	AUX	CCONJ	DET	INTJ	NOUN	NUM	PROPN	PRON	SCONJ	SYM	VERB	X	PUNCT
ADJ	22,608	–	–	–	–	–	–	–	–	183	–	–	–	–	83	–
PREP	–	76,628	–	–	–	–	–	–	–	–	–	–	–	–	4	–
ADV	–	–	12,236	–	–	–	–	–	–	–	–	–	–	–	15	–
PDEN	–	–	2,343	–	–	–	–	–	–	–	–	–	–	–	–	–
ADV-KS	–	–	475	–	–	–	–	–	–	–	–	38	–	–	–	–
KC	–	–	–	–	12,038	–	–	–	–	–	–	–	–	–	2	–
ART	–	–	–	–	–	69,156	–	–	–	–	–	–	–	–	4	–
PROADJ	–	–	–	–	–	8,473	–	–	–	–	–	–	–	–	–	–
IN	–	–	–	–	–	–	82	–	–	–	–	–	–	–	2	–
N	–	–	–	–	–	–	–	99,660	3,739	2,698	–	–	1,401	–	589	–
NUM	–	–	–	–	–	–	–	2	8,583	–	–	–	–	–	–	–
NPROP	–	–	–	–	–	–	–	–	–	39,692	–	–	–	–	–	–
PROSUB	–	–	–	–	–	–	–	–	–	–	3,215	–	–	–	–	–
PROPESS	–	–	–	–	–	–	–	–	–	–	6,439	–	–	–	2	–
PRO-KS	–	–	–	–	–	96	–	–	–	–	5,398	101	–	–	–	–
KS	–	–	–	–	–	–	–	–	–	–	–	5,119	–	–	–	–
CUR	–	–	–	–	–	–	–	–	–	–	–	–	1,504	–	–	–
V	–	–	–	13,867	–	–	–	–	–	–	–	–	–	41,679	2	–
PCP	2,700	–	–	170	–	–	–	111	–	–	–	–	–	7,413	–	–
PU	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	69,233

Table 6: Occurrences of UPOS with respect to the original MacMorpho v.3 POS tags.

sequent comparisons are restricted to this set of shared sentences, to ensure a fair comparison.

In our sifting method, the silver-standard corpus retains 21,433 of these 32,776 sentences (65.4%) (see Figure 3). In the baseline method, only 7,289 sentences (22.2%) were retained (given that the requirement of direct UPOS matching is a harsh criterion). This demonstrates that the compatibility checking effectively results in more recovered sentences, as expected.

We further observed that the overwhelming majority of the 7,289 sentences with identical UPOS in MACMORPHO-UD (7,192 of them – 98.7%) were also included in the set of sentences retained by our sifting procedure. Only 97 of the 7,289 sentences from MACMORPHO-UD were discarded by our sifting procedure. Most of these instances involve fixed expressions annotated as PDEN in MacMorpho v.3, which do not map to the corresponding UPOS **ADV**. An illustrative example is presented below (under current UD guidelines for Portuguese, "*Ou seja*" would be tagged as **CCONJ** and **AUX** respectively):

```
Ou_PDEN seja_PDEN ,_PU separar_V
aquilo_PROSUB que_PRO-KS é_V
necessário_ADJ do_PREP+ART
inexato_N ._PU
```

4.3 Experiments: usefulness for NLP training and data augmentation

To assess the usefulness of the generated silver-standard corpus for training an NLP tool and for reliable data augmentation, it is important to contextualize the effort required to build high quality manually annotated corpora. Existing gold-standard

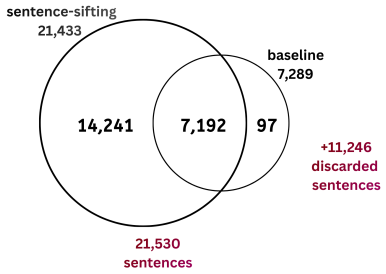


Figure 3: Silver-standard produced by sentence-sifting procedure and baseline comparison over the 32,776 comparable sentences.

resources of similar size require substantial manual work by trained linguists to carefully review and annotate each sentence. The Porttinari-base, for example, contains 8,418 sentences and required a substantial manual effort by a team of linguists to carefully review each sentence (Duran et al., 2023) and produce this gold-standard resource. In contrast, our method generates a corpus nearly four times larger (30,581 sentences) with significantly less manual effort. It is important to emphasize that, in the case of generating the silver-standard from MacMorpho v.3, the automated annotation focuses only on UPOS, LEMMA, and FEATS, without any manual verification, ensuring consistency and quality within this scope.

To illustrate the practical usefulness of this approach, we trained a POS tagger, using the LatinPipe architecture (Straka et al., 2024) with BERTimbau embeddings (Souza et al., 2020) and 80 training epochs, the same setup used to train the newest version of Portparser model over Porttinari-base. The 30,581 sentences of our silver-standard corpus were split into three CoNLL-U files for

training, development, and testing, with 21,407 (70%), 3,058 (10%), and 6,116 (20%) sentences, respectively. The resulting model achieved competitive performance of 99.72%, 99.15%, and 99.45% predicting UPOS, LEMMA, and FEATS, respectively (over the test portion of the silver corpus). This experiment shows that the silver-standard corpus has systematic annotation and can effectively train high-quality models, providing a practical and scalable alternative to fully manual corpus construction while preserving the historical value of legacy resources.

On a complementary view, we also assessed the effectiveness of using the silver-standard corpus for data augmentation purposes. We evaluated three models on three gold-standard corpora of various genres, namely, news texts (Porttinari-base) (Duran et al., 2023), legal texts (PortJur) (Lopes et al., 2025), and academic texts (PetroGold) (De Souza and Freitas, 2022).

The three models considered are:

- Portparser v2 model, trained on the training portion of Porttinari-base;
- a model trained exclusively on the silver-standard corpus; and
- a model trained on the Porttinari-base training data augmented with the silver-standard corpus.

Table 7 reports the precision scores for UPOS, LEMMA and FEATS annotation obtained by each model across the gold-standard corpora.

Porttinari-base test subcorpus - 1,683 sentences			
Model	**UPOS**	**FEATS**	**LEMMA**
Portparser v2	**99.45%**	99.20%	99.36%
Silver-Standard	98.96%	98.26%	99.57%
Porttinari-base+Silver	99.43%	**99.24%**	**99.60%**

PortJur corpus - 2,005 sentences			
Model	**UPOS**	**FEATS**	**LEMMA**
Portparser v2	**98.20%**	**98.03%**	98.61%
Silver-Standard	97.87%	97.62%	**98.62%**
Porttinari-base+Silver	98.15%	97.78%	98.59%

PetroGold dev+test subcorpus - 1,776 sentences			
Model	**UPOS**	**FEATS**	**LEMMA**
Portparser v2	96.82%	86.61%	98.24%
Silver-Standard	**97.17%**	**86.96%**	**98.44%**
Porttinari-base+Silver	96.82%	86.64%	98.23%

Table 7: Experiment results on using the silver-standard corpus for data augmentation.

Overall, we may see that:

- the results of the models trained solely over the silver-standard corpus are sometimes better than the ones produced by training over the gold-standard corpus, which again evidences the quality of the silver-standard corpus and the methodology to produce it;
- the use of the silver-standard corpus as augmented data did not produce significant different results, keeping the quality already achieved by the models without data augmentation, which was somehow expected, as we saw above that the silver-standard corpus has a very good quality and that POS tagging results are already very good, being difficult to improve them.

5 Final remarks

Beyond the revitalization of MacMorpho, as reported in this paper, it is interesting to notice that the proposed method may be adapted for other contexts, e.g., for covering other linguistic levels and for performing knowledge-rich data acquisition.

If the appropriate tools are available, it is not difficult to instantiate our method for syntactic or semantic analysis. Even if there is no previously annotated resource, approaches using other annotation tools may produce annotated data that may serve as surrogate for the manually annotated resource. However, some human expertise may be necessary for the eventual compatibility of tagsets that the method employs.

The recovery of older resources for data acquisition/augmentation may be interesting in situations where there is data scarcity and/or highly specialized linguistic annotation is necessary and current data augmentation strategies (as data synthesis, back-translation and the so-called "easy data augmentation") may be inappropriate or insufficient.

Future work includes improving our method to recover more legacy data (possibly for other linguistic annotation levels as well) and performing in-depth studies to explain the differences between legacy and current annotation directions, which may subsidize divergence analyses for proposing new methodological strategies and/or evidencing linguistic modeling evolution.

For the interested reader, the silver-standard corpus (**MacMorpho-UD-2.17**), as well as other information about this initiative, are publicly available and may be found at the web portal of the POeTiSA project[7].

[7] https://sites.google.com/icmc.usp.br/

Acknowledgments

This work was carried out at the Center for Artificial Intelligence of the University of São Paulo (C4AI - http://c4ai.inova.usp.br/), with support by the São Paulo Research Foundation (FAPESP grant #2019/07665-4) and by IBM Corporation. The project was also supported by the Ministry of Science, Technology and Innovation, with resources of Law n. 8.248, of October 23, 1991, within the scope of PPI-SOFTEX, coordinated by Softex and published as Residence in TIC 13, DOU 01245.010222/2022-44, and by *Instituto Nacional de Ciência e Tecnologia em Inteligência Artificial Responsável para Linguística Computacional, Tratamento e Disseminação de Informação* (INCT-TILD-IAR) (grant #408490/2024-1).

References

Sandra Aluísio, Jorge Pelizzoni, Ana Raquel Marchi, Lucélia de Oliveira, Regiana Manenti, and Vanessa Marquiafável. 2003. An account of the challenge of tagging a reference corpus for brazilian portuguese. In *Computational Processing of the Portuguese Language*, pages 110–117, Berlin, Heidelberg. Springer Berlin Heidelberg.

Julia Bonn, Matthew J. Buchholz, Jayeol Chun, Andrew Cowell, William Croft, Lukas Denk, Sijia Ge, Jan Hajič, Kenneth Lai, James H. Martin, Skatje Myers, Alexis Palmer, Martha Palmer, Claire Benet Post, James Pustejovsky, Kristine Stenzel, Haibo Sun, Zdeňka Urešová, Rosa Vallejos, and 4 others. 2024. Building a broad infrastructure for uniform meaning representations. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 2537–2547, Torino, Italia. ELRA and ICCL.

Marie-Catherine de Marneffe, Christopher D. Manning, Joakim Nivre, and Daniel Zeman. 2021. Universal Dependencies. *Computational Linguistics*, 47(2):255–308.

Elvis De Souza and Cláudia Freitas. 2022. Polishing the gold – how much revision do we need in treebanks? In *Procedings of the Universal Dependencies Brazilian Festival*, pages 1–11.

Magali Duran, Lucelene Lopes, Maria das Graças Nunes, and Thiago Pardo. 2023. The dawn of the Porttinari multigenre treebank: Introducing its journalistic portion. In *Anais do XIV Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana*, pages 115–124, Porto Alegre, RS, Brasil. SBC.

poetisa/resources-and-tools

Magali Duran, Maria das Graças Volpe Nunes, Lucelene Lopes, and Thiago Alexandre Salgueiro Pardo. 2022. Manual de anotação como recurso de processamento de linguagem natural: o modelo universal dependencies em língua portuguesa. *Domínios de Lingu@gem*, 16(4):1608–1643.

Erick R. Fonseca, João Luís G. Rosa, and Sandra Maria Aluísio. 2015. Evaluating word embeddings and a revised corpus for part-of-speech tagging in portuguese. *Journal of the Brazilian Computer Society*, 21(1):2.

Erick Rocha Fonseca and João Luís G. Rosa. 2013. Mac-morpho revisited: Towards robust part-of-speech tagging. In *Proceedings of the 9th Brazilian Symposium in Information and Human Language Technology*.

Cláudia Freitas and Elvis de Souza. 2023. A study on methods for revising dependency treebanks: in search of gold. *Language Resources and Evaluation*, 58(1):111–131.

Cláudia Freitas, Luiza F. Trugo, Fabricio Chalub, Guilherme Paulino-Passos, and Alexandre Rademaker. 2018. Tagsets and datasets: Some experiments based on portuguese language. In *Computational Processing of the Portuguese Language*, pages 459–469, Cham. Springer International Publishing.

Lucelene Lopes, Magali Duran, Paulo Fernandes, and Thiago Pardo. 2022. PortiLexicon-UD: a portuguese lexical resource according to Universal Dependencies model. In *Proceedings of the Language Resources and Evaluation Conference*, pages 6635–6643, Marseille, France. European Language Resources Association.

Lucelene Lopes, Magali Sanches Duran, and Thiago Salgueiro Pardo Pardo. 2023. Verifica-UD: uma ferramenta online para verificação de textos em português anotados no formato CoNLL-U segundo o padrão universal dependencies. Technical Report 445, ICMC-USP.

Lucelene Lopes, Maria Nunes, Magali Duran, and Thiago Pardo. 2025. A sintaxe no tribunal: apresentando e explorando um corpus jurídico em português anotado sintaticamente segundo o modelo universal dependencies. In *Anais do XVI Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana*, pages 220–232, Porto Alegre, RS, Brasil. SBC.

Lucelene Lopes and Thiago Pardo. 2024. Towards Portparser - a highly accurate parsing system for Brazilian Portuguese following the Universal Dependencies framework. In *Proceedings of the 16th International Conference on Computational Processing of Portuguese - Vol. 1*, pages 401–410, Santiago de Compostela, Galicia/Spain. Association for Computational Lingustics.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajič, Christopher D. Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, Reut Tsarfaty, and Daniel Zeman.

2016. Universal Dependencies v1: A multilingual treebank collection. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1659–1666, Portorož, Slovenia. ELRA.

Thiago Pardo, Magali Duran, Lucelene Lopes, Ariani Felippo, Norton Roman, and Maria Nunes. 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.

Alexandre Rademaker, Fabricio Chalub, Livy Real, Cláudia Freitas, Eckhard Bick, and Valeria De Paiva. 2017. Universal dependencies for portuguese. In *Proceedings of the Fourth International Conference on Dependency Linguistics (Depling)*, pages 197–206.

Emanuel Silva, Thiago Pardo, Norton Roman, and Ariani Fellipo. 2021. Universal dependencies for tweets in brazilian portuguese: Tokenization and part of speech tagging. In *Anais do XVIII Encontro Nacional de Inteligência Artificial e Computacional*, pages 434–445, Porto Alegre, RS, Brasil. SBC.

Fábio Souza, Rodrigo Nogueira, and Roberto Lotufo. 2020. Bertimbau: Pretrained bert models for brazilian portuguese. In *Intelligent Systems*, pages 403–417, Cham. Springer International Publishing.

Milan Straka, Jana Straková, and Federica Gamba. 2024. ÚFAL LatinPipe at EvaLatin 2024: Morphosyntactic analysis of Latin. In *Proceedings of the Third Workshop on Language Technologies for Historical and Ancient Languages (LT4HALA) @ LREC-COLING-2024*, pages 207–214, Torino, Italia. ELRA and ICCL.