Early Experiments on Automatic Annotation of Portuguese Medieval Texts[⋆]

Maria Inês Bico[1,2][0000−0002−6280−9417], Jorge Baptista[3,4][0000−0003−4603−4364], Fernando Batista[4,5][0000−0002−1075−0177], and Esperança Cardeira[1,2][0000−0003−4700−9830]

[1] Univ. Lisboa - Fac. Letras, {mariainesb1,ecardeira}@campus.ul.pt
[2] Centro de Linguística da Universidade de Lisboa
Univ. Algarve - Fac. Ciências Humanas e Sociais,
jbaptis@ualg.pt
[3] INESC-ID Lisboa - Human Language Technology Lab
ISCTE - Instituto Universitário de Lisboa,
fernando.batista@iscte-iul.pt

Abstract. This paper presents the challenges and solutions adopted to the lemmatization and part-of-speech (PoS) tagging of a corpus of Old Portuguese texts (up to 1525), to pave the way to the implementation of an automatic annotation of these Medieval texts. A highly granular tagset, previously devised for Modern Portuguese, was adapted to this end. A large text (∼155 thousand words) was manually annotated for PoS and lemmata and used to train an initial PoS-tagger model. When applied to two other texts, the resulting model attained 91.2% precision with a textual variant of the same text, and 67.4% with a new, unseen text. A second model was then trained with the data provided by the previous three texts and applied to two other unseen texts. The new model achieved a precision of 77.3% and 82.4%, respectively.

Keywords: Automatic Annotation · Lemmatization · Part-of-speech tagging · Old Portuguese

1 Introduction

For a long time, researchers in historical linguistics handpick the traces of the phenomena they choose to study. It is laborious and slow work, and the pressure of time and deadlines usually meant that the scope of the investigation has to be restricted, whether in terms of the phenomena or in the quantity of the data perused. In addition, though the availability of old texts on the web is larger than ever before [1,3,5,11,12,22,19], most of the times they are produced

[⋆] Research for this paper was partially funded by public funds through Fundação para a Ciência e a Tecnologia: J. Baptista and F. Batista (INESC-ID Lisboa, proj.ref UIDB/50021/2020), E. Cardeira (School of Arts and Humanities, Center of Linguistics, University of Lisbon, proj.ref. UIDP/00214/2020) and M.I. Bico by Ph.D grant (proj.ref. UI/BD/152806/2022).

only as a facsimile or, even if transcribed and edited, they are not linguistically annotated, at least for the words' parts-of-speech (PoS), i.e. morphosyntactic categories (noun, verb, adjective, etc.) and their inflection, as well as the words' *lemmata* [8,13]. This presents a challenge to those researchers focused on studying the history of a language. When looking for phenomena that rely on the written word to know how the language was at a particular time, picking up the data manually can both be an valuable asset and a kryptonite. Thus, having texts' words annotated for their lemmas and PoS allows for further linguistic processing, namely automatic syntactic analysis (parsing) and the modelling of former stages of language by way of treebanks [16] and texts' collation [2]. For this reason, Natural Language Processing (NLP) tools and techniques [10] can be very useful to Historical Linguistics [5,6,18]. Not only do they allow new and different kinds of research questions, but they also introduce new research tools and methods regarding the collection of data and speed up its analysis.

This paper is part of a larger project that aims to use NLP methods on the investigation of Old Portuguese, particularly on the texts that make up the *Corpus de Textos Antigos* 'Old Texts Corpus' (CTA)[4], a project started in 2015 by the Center of Linguistics of the University of Lisbon (CLUL)[5]. As a repository of transcribed and edited texts in Old Portuguese, dated up to 1525, this corpus can be a helpful resource to researchers interested in the older stages of the language. Nevertheless, the CTA's texts are not yet annotated neither for their PoS nor for their lemmas. Manual annotation of the entire corpus is a very time-consuming, highly-skilled and costly task, hence a machine-learning approach would better suit these goals. In fact, even if the automatic annotation produced by this language models is not completely accurate, it goes a long way in preparing the textual material for a manual revision and correction, speeding up the human annotation effort. This paper, then, will present some early results of the automatic annotation of a subset of the corpus whose data was prepared for a machine-learning PoS and lemmatization tasks.

2 Corpus of Ancient Texts (CTA)

The *Corpus de Textos Antigos* is a project developed by CLUL's Philology group, which aims to publish all hagiographic, spiritual and didactic texts written in or translated to Portuguese up to 1525 (this is a flexible date, deliberately chosen to allow the inclusion of *incunabula* and also texts that, despite dating from the first quarter of the 16th century, transmit older manuscripts). The main purpose of this project is to offer editions that reproduce the texts with high fidelity to the manuscript (ms.) or incunable [6]. Following this principle, there is little or no editorial intervention when it comes to the correction of errors, the restitution of lacunae or orthographic variation. This is why a simple string search would

[4] http://teitok.clul.ul.pt/teitok/cta/ (last access: July 25, 2022). All the remaining URL in this paper were check on this date.

[5] http://www.clul.ulisboa.pt/grupo/filologia

[6] http://teitok.clul.ul.pt/teitok/cta/index.php?action=criterios

almost never capture all the instances of a word occurring within the corpus, so that lemmatization is an essential previous step towards efficient lexical queries. The corpus uses the web-based framework TEITOK [9,20], an online tool which combines both textually annotated texts with linguistic annotations. With a modular design and the granular customization it allows, TEITOK can be used with very different corpora.

As of April 2022, the corpus consists of 31 editions of 26 different texts. There are three texts with more than one edition: *Horto do Esposo* has two edited manuscripts from the late 14^{th}century; *Vida de Santa Maria Egipcia*, with two mss. from the 15^{th} century; and *Vida e Milagres de Santa Senhorinha de Bastos*, written in the second half of 13^{th} century, has four edited witnesses that date from the early 17^{th} century to the 19^{th} century. The texts differ in extension, from a couple hundreds to more than 150,000 words. As shown, texts also vary both in the date of redaction and in date of production. The oldest text (and manuscript) is the ms. A of *Horto do Esposo*[7] and it dates between 1390–1437. The most recent text (not necessarily the most recent manuscript or edition) is *Memorial da Infanta Santa Joana*[8] and dates between 1513-1525. The most recent manuscript comes from the end of the 18^{th} century, the ms. P of *Vida e Milagres de Santa Senhorinha de Bastos*[9].

3 Text selection, preparation and annotation

In this section, the text selection, preparation and annotation process are described. For the manual annotation, the ms. A of *Horto do Esposo* (henceforward, *HdE-A*), whose both the manuscript and the text date from about the same time (c. 1390-1437), was chosen. For testing, the ms. G1 of *Vida e Milagre de Santa Senhorinha de Basto* (henceforward, *VMSSB-G1*)[10] was chosen. This is a text dated between 1248–1284 and whose manuscript has been dated from 1620-1645. As the corpus has three others, albeit fragmented, witnesses of *Horto do Esposo* (henceforaward, *HdE-DCE*)[11], the testing was also done on this witness. The HdE-DCE ms. was chosen for testing the POS-tagger because of its natural likeness to HdE-A, while the choice of VMSSB-G1 is due to the fact that, being both HdE-A and VMSSB-G1 hagiographic in genre, a greater similarity between their respective lexicons is expected. For another experiment, a second model was trained on these 3 texts, and 2 other texts were selected from the corpus for testing: the *História do mui nobre Vespasiano* (henceforward, *Vespasiano*)[12], printed in Lisbon in 1496, and the text of *Memorial da Infanta Santa Joana* (henceforward, *MISJ*)[13], in a manuscript later than 1525, although it

[7] Biblioteca Nacional (Portugal), Alc.198, fls. 1r–155r.

[8] Biblioteca do Museu de Aveiro, ms. 1 [33/CD], fls. 48a–110b.

[9] Biblioteca Mun. Porto, Safe n. 527 (Cat. n. 683), ff. 196v–208v.

[10] Arq. Mun. Alfredo Pimenta (Guimarães), Ms. da Colegiada 793, fls. 211r–236r.

[11] Arq. Nac. Torre do Tombo. Fragm., Cx. 21, n.26 (Casa Forte). Lorvão, Livro 10, fl. 13r. Fragm., Cx. 21, n.23a (Casa Forte).

[12] Lisboa, Valentim Fernandes, [1496?]. Biblioteca Nacional (Portugal), Inc. 571.

[13] Biblioteca do Museu de Aveiro, ms. 1 [33/CD], fls. 48a-110b.

is thought to have been first written between 1513 and 1525. Table 1 presents the contents of the texts selected from the CTA corpus for the experiments in this paper. The selected texts are indicated by a conventional code with their respective date (see details below). Information on the number of tokens, words, different word forms (case sensitive) and punctuation signs is provided.

Table 1. Texts from Old Portuguese Corpus

Corpus (date)	**tokens**	**words**	**diff.**	**punct**
HdE-A (1390-1437)	154,952	137,710	14,333	17,174
HdE-DCE (1391-1450)	2,694	1,841	722	849
VMSSB-G1 (1248-1284)	13,948	12,403	2,352	1,541
MISJ (1513-1525)	51,680	46,517	7,007	5,162
Vespasiano (1496?)	19,141	17,893	2,759	1,241
Total	242,415	216,364	21,595	25,967

The manual annotation task consisted in attributing to each token the corresponding lemma and the part-of-speech (PoS) tag. A set of guidelines for this task were produced to define the criteria for attributing the *lemmata*, to describe the tagset, and to explicitly guide the PoS-tag attribution, especially in more complex cases. For the lemmatization of the word forms, the modern lemma was adopted whenever possible, in order to ensure an efficient way to query the corpus. The traditional criterion for lemma attribution was generally adopted: the impersonal infinitive for the verbs, the masculine-singular form for the adjectives, the singular form for nouns (masculine or feminine, depending on its gender), and so on. Each PoS-tag consists of a morphosyntactic *category* (v.g. adjective, adverb, conjunction, determiner, interjection, noun, preposition, pronoun or verb) and, if applicable, an *inflection* code indicating the morphological categories relevant to that category (i.e., tense-mood and person-number, for verbs; gender and number for nouns; etc.). We adopt a highly granular tagset, adapting one already developed for Modern Portuguese and presented in [4,14,15]. The formalism here used is generically the same that was originally developed by [7]. Three annotators participated in the task, all linguists familiar with Old Portuguese texts and its grammar. At the end of the process, a set of procedures was put in place to verify and correct eventual inconsistencies.

4 Experiments and Results

Having all words present in *Horto do Esposo* (HdE-A) initially annotated with lemmas and PoS-tag, a thorough revision was made, not only regarding the correctness of the lemmas attribution but also considering the formal consistency of the annotation. Errors and inconsistencies, due to manual annotation, were detected and corrected. A PoS-tagging model was then trained with the TreeTagger [17] and applied to both HdE-DCE and VMSSB-G1. Then, after correcting the annotations produced for these two texts, a new model was trained and applied to both the MISJ and Vespasiano. Table 2 shows the results of the different experiments in automatically PoS-tagging the corpus' texts.

Table 2. Experiments in PoS-tagging. Preliminary results: Precision and error analysis

Corpus	**TP**	**P**	**L-t**	**L-z**	**T-p**	**T-t**	**T-z**	**U-p**	**U-t**	**U-z**	**punct**
HdE-CDE	2,458	91,24%	14	1	6	1	26	70	29	84	5
VMSSB-G1	9,401	67,40%	170	43	24	22	257	1,233	357	1,015	1,426
MISJ	39,956	77,31%	93	1	44	316	442	5,198	1,171	4458	12
Vespasiano	15,768	82,38%	280	30	9	17	254	1,259	652	866	0

A preliminary, manual inspection of the results and the corresponding error analysis was then carried out. Entirely correct matches (lemma, PoS and morphosyntactic tag) are marked as true-positives (**TP**) and precision (**P**) is provided. Then, lemma attribution was considered, either correctly (**L-**), or incorrectly (**T-**) attributed, or, else, not given (unknown, **U-**). Within each of these lemma attributions, the correcteness of the PoS and the morphosyntatic tag were also distinguished: **-p** indicates when both PoS and morphosyntactic tag were correctly given; **-t** indicates that the correctly marked PoS was, but not the morphosyntactic tag; **-z** indicates that neither PoS nor tag were correct. Punctuation (**punct**) marks were often marked by the system as unknown lemmas instead of the conventional notation adopted. Often, these were incorrectly given a PoS and a morphosyntactic tag.

Concerning the first experiment, the better performance of the model on HdE-DCE (precision: 91,24%) could be explained by the fact that it is another witness of *Horto do Esposo*. Both manuscripts (HdE-A and HdE-DCE) thus have the same lexicon and the same syntactical structures. The proximity in the dates of the manuscripts may also have played a role in this results. Several aspects may explain the worst performance of the model on VMSSB-G1 (precision: 67,4%). The ms. VMSSB-G1 dates from the 17^{th} century, which is much later than the date of HdE-A (c.1390-1437). Though some older traces of the language are preserved, VMSSB1-G1 shows some linguistic changes that happened between the two periods. For example, the program did not recognize the form *nao* (adverb 'no') as HdE-A only presents the forms *nõ*, *non*, *nom*, and *nã* This new graphic form *nao* signals the changes in the nasal word endings, converging into the diphthong <ão>. On the other hand, many lemmas could not be ascribed due to graphic differences found in this manuscript, even if the same word appears in both. Also, VMSSB-G1 makes use of the comma 1,426 times, whereas HdE-A only uses the full stop, which explains the punctuation errors signaled in the Table.

The model built upon the data of HdE-A, HdE-DCE and VMSSB-G1 was then applied on the MISJ and Vespasiano. Results show that the new model produced better results on Vespasiano (precision: 82.38%) than in MISJ (precision: 77.31%), though it still fails to recognize a large number of lemmas, especially in the latter. Again, many words show a spelling different from the one used to learn the model.

As for the incorrect annotations (false-positives), there are two types of errors, based on whether the model attributes a lemma to a word (column **T-z**) or not (column **U-z**). A large number of words with unknown lemmas (**U-**) are still adequately tagged as for their PoS (**U-p**) or their morphosyntactic values (**U-t**). This case corresponds to the PoS-tagger being able to correctly guess those values from the surrounding words. Many cases in **T-z** correspond to the typical situation of PoS ambiguity. For example the word *nos* may correspond to different inflections of the personal pronoun (*nós*/*nos*, 'we,us') but also to the contraction of preposition and a definite article (*em*_*os* 'in_the-masc.pl.). As for the cases with unknown lemmas and where the system also fails the PoS and morphosyntactic tags (**U-z**), this may hint at the natural limitations of the machine-learning approach here adopted.

5 Conclusion and future work

This paper presents the preliminary steps taken towards the automatic annotation of the Portuguese *Corpus de Textos Antigos*. This annotation consists in attributing lemmas and PoS-tags to their word forms, both the morphosyntactic categories and their inflection values. An initial annotation task was manually carried out on HdE-A, containing almost 150 thousand tokens. Such data was then used to train a Machine Learning model, and then used to automatically annotate two other smaller documents: another ms. (HdE-CDE) of the same text used for training; and another text, different but of a similar genre (VMSSB-G1). As expected, the second ms. of the HdE text achieved a very high precision (91.24%). With the unrelated text of VMSSB-G1, the model only produced a modest precision (67.4%), mostly because many word forms (8.84%) had not been previously seen by the model, so that their *lemmata* were labelled as *unknown*. Still, the model was able to correctly assign the PoS and the inflection values to most of them.

The preliminary results of the automatic annotation show how the model improves the more data it receives. The performance of the second model on Vespasiano is better than the outcome of the first experiment on VMSSB-G1. Whereas the latter was annotated with a model with the data from only one text, the former had the model trained with three different texts. As for the errors, whenever the tagger inaccurately attributes a lemma, it is often due to the ambiguous nature of the word.

The use of NLP methods on the corpus will allow for new questions to be asked in new approaches to this linguistic data. Based on the lexically annotated corpus, it will now be possible to analyse the irregularity of the forms and linguistic changes. The use of an annotated corpus could also be helpful in determining the affiliation between different witnesses of the same text [2], using automatic collation tools, such as Collatex [21][14].

[14] https://collatex.net/

References

1. Britto, H., Finger, M.: Constructing a parsed corpus of historical Portuguese. In: Proceedings of International Humanities Computing Conference, University of Virginia, Charlottesville. ACH/ALLC (1999)
2. Camps, J.B., Ing, L., Spadini, E.: Collating Medieval Vernacular Texts. Aligning Witnesses, Classifying Variants. In: Digital Humanities Conference (DHC) 2019. Utrecht, Netherlands (Jul 2019), https://hal.archives-ouvertes.fr/hal-02268348
3. Davies, M.: New directions in Spanish and Portuguese corpus linguistics. Studies in Hispanic and Lusophone Linguistics **1**(1), 149–186 (2008)
4. Eleutério, S., Ranchhod, E., Freire, H., Baptista, J.: A system of electronic dictionaries of Portuguese. Linguisticae Investigationes **19**(1), 57–82 (1995)
5. Gamallo, P., Pichel, J.R., Santalha, J.M.M., Neves, M.: Uso de tecnologias linguísticas para estudar a evolução dos sufixos *-çom* e *-vel* no galego-português medieval a partir de corpora históricos. Linguamática **13**(2), 3–17 (2021)
6. Gonçalves, M.F., Banza, A.P.: Da antiga à nova Filologia: o Projecto MEP-BPEDig. In: Actas del XXVI Congreso Internacional de Lingüística y de Filología Románicas. Tome VII. vol. 7, pp. 205–210. Walter de Gruyter (2013)
7. Gross, M.: La construction de dictionnaires électroniques. In: Annales des télécommunications, vol. 44, pp. 4–19. Springer (1989)
8. Hendrickx, I., Marquilhas, R.: From old texts to modern spellings: An experiment in automatic normalisation. J. Lang. Technol. Comput. Linguistics **26**(2), 65–76 (2011)
9. Janssen, M.: TEITOK: Text-faithful annotated corpora. In: Proceedings of the 10^{th} International Conference on Language Resources and Evaluation (LREC'16). pp. 4037–4043. European Language Resources Association (ELRA), Portorož, Slovenia (May 2016), https://aclanthology.org/L16-1637
10. Jurafsky, D., Martin, J.H.: Speech and language processing (draft) (2021), https://web.stanford.edu/~jurafsky/slp3/
11. Lopes, J., Rocio, V., Xaxier, M.F., Vicente, G.: Criação automática de uma colecção de textos de português medieval parcialmente anotados sintacticamente. In: Actas del Segundo Seminário de Escuela Interlatina de Altos Estudios en Lingüística Aplicada. pp. 203–220 (2002)
12. Mendes, A.: Linguística de corpus e outros usos dos corpora em linguística. In: Martins, A.M., Carrilho, E. (eds.) Manual de linguística portuguesa, vol. 16, pp. 224–251. Walter de Gruyter GmbH & Co KG (2016)
13. Parkinson, S.R., Emiliano, A.H.: Encoding Medieval Abbreviations for Computer Analysis (from Latin–Portuguese and Portuguese non-literary sources). Literary and Linguistic Computing **17**(3), 345–360 (2002)
14. Ranchhod, E., Mota, C., Baptista, J.: A computational lexicon of Portuguese for automatic text parsing. In: Standardizing Lexical Resources (SIGLEX'99). pp. 74–80. ACL/SIGLEX, Maryland, USA (1999)
15. Ranchhod, E.M.: O uso de dicionários e de autómatos finitos na representação lexical. In: Ranchhod, E.M. (ed.) Tratamento das Línguas por Computador. Uma introdução à Linguística Computacional e suas aplicações, pp. 13–47. Caminho (2001)
16. Rocio, V., Alves, M.A., Lopes, J.G.P., Xavier, M.F., Vicente, G.: Automated creation of a Medieval Portuguese partial treebank. In: Abeillé, A. (ed.) Treebanks: Building and Using Parsed Corpora, pp. 211–227. Springer (2003)

17. Schmid, H.: Probabilistic part-of-speech tagging using decision trees. In: Proceedings of International Conference on New Methods in Language Processing. pp. 154–163 (1994)
18. Schmid, H.: Deep learning-based morphological taggers and lemmatizers for annotating historical texts. In: Proceedings of the 3^{rd} International Conference onD-digital Access to Textual Cultural Heritage. pp. 133–137 (2019)
19. de Sousa, M.C.P.: O corpus Tycho Brahe: Contribuições para as Humanidades Digitais no Brasil. Filologia e linguística portuguesa **16**(esp.), 53–93 (2014)
20. Vaamonde, G., Janssen, M.: Da edición dixital á análise lingüística. A creación de corpus históricos na plataforma TEITOK, pp. 271–292 (01 2020). https://doi.org/10.17075/cbfc.2020.008
21. van Zundert, J., Haentjens Dekker, R., Van Hulle, D., Neyt, V., Middell, G.: Computer-supported collation of modern manuscripts: Collatex and the Beckett Digital Manuscript Project. Literary and Linguistics Computing **30**(3), 452–470 (Mar 2014). https://doi.org/10.1093/llc/fqu007
22. Xavier, M.F.: O CIPM – Corpus Informatizado do Português Medieval, fonte de um dicionário exaustivo. In: Lingüística de corpus y lingüística histórica iberorrománica, pp. 137–156. De Gruyter (2016)