Colonia: Corpus of Historical Portuguese

Marcos Zampieri, Martin Becker
Romance Philology Department
University of Cologne

Abstract

This paper presents a new linguistic resource for Portuguese, Colonia: Corpus of Historical Portuguese. It is a collection of textual material from the 16^{th} to the early 20^{th} century with a user-friendly interface. Colonia is a diachronic corpus available online with over 5,1 million tokens divided into five sub-corpora by century. The corpus was POS tagged using TreeTagger and the framework used to process queries is Corpus WorkBench (CWB) and Corpus Query Processor (CQP). On top of this architecture, CQPWeb provides the interface between query language and users. The methods for corpus compilation as well as the characteristics of the data are presented in detail in this paper.

1 Introduction

Portuguese is currently the 6^{th} most spoken language in the world according to Ethnologue [1]. For geopolitical reasons its importance as a global language has increased significantly over the past decades contributing to raise the interest in Portuguese as a foreign language and in linguistic research on Portuguese. Even though the interest in Portuguese has grown substantially, there is still much to be done in terms of language resources, especially if compared to other European languages such as German, French and English. Language resources such as corpora, taggers, and lexicons are vital for linguistic research and for the computational processing of a language. This paper aims to contribute in this direction and presents a new resource for the study of Portuguese focusing on historical data.

2 Related Work

A number of Portuguese corpora have been developed over the last two decades, and most of them are freely available to the research community. One repository of Portuguese language resources is the Linguateca[1] project [2]. Linguateca hosts a number of written contemporary Portuguese corpora, the most important of them are CETEMPublico [3], a collection of 191 million words retrieved from the Portuguese *Público* newspaper and a 32 million word Brazilian corpus from *Folha de São Paulo*. Most corpora hosted by Linguateca are contemporary, with the exception of Vercial[2], a corpus of old literary Portuguese.

Other Portuguese corpora include the Reference Corpus of Contemporary Portuguese (CRPC) [4], containing written and spoken Portuguese data and the new *Corpus Brasileiro* [5], alleged to be the biggest Portuguese corpus to date with 1 billion words of written Brazilian Portuguese.

When it comes to historical Portuguese texts, there are only a few options available. One of them is Tycho Brahe [6], which contains texts from the 16^{th} to the 19^{th} century, in a total of 2,3 million words and two different layers of annotation: morphosyntactic and syntactic. Tycho Brahe texts are available for download in three versions: raw corpus, POS annotated or syntactically annotated. The corpus can be also accessed through an online interface. Another historical corpus is the one compiled by *Grupo de Morfologia Histórica do Português (GMHP)*[3] at the University of Sao Paulo. This corpus is, however, untagged and has no graphical user in-

[1] http://linguateca.pt/

[2] http://www.linguateca.pt/acesso/corpus.php?corpus=VERCIAL

[3] http://www.usp.br/gmhp/CorpI.html

terface. Finally, a widely used diachronic corpus is *Corpus do Português* [7], a collection of 45 million words spanning the 14^{th} to the 20^{th} century. *Corpus do Português* is available solely through an user interface.

2.1 Another Portuguese Historical Corpus?

Given the high quality of the resources available for historical Portuguese, we believe that the compilation of Colonia should fill a gap between the existing corpora. We present next a table with the basic features of the most important historical Portuguese corpora to date: Tycho Brahe, GMHP[4], Corpus do Português, and the new Colonia.

	Tycho Brahe	**GMHP**	**C. Português**	**Colonia**
Variety	PT - BR	PT - BR	PT - BR	PT - BR
Time Span	14^{th}-19^{th}	15^{th}-20^{th}	13^{th}-19^{th}	16^{th}-20^{th}
Interface	Yes	No	Yes	Yes
Download	Yes	Yes	No	Yes
POS	Yes	No	Yes	Yes
Syntactic	Yes	No	No	No
Size	2,45 mil.	No data	45 mil.	5,1 mil.

Table 1: Comparison Portuguese Corpora

From the four corpora compared here, Colonia is the biggest corpus available both to be downloaded (POS tagged version) as well as to be used through a graphical interface. The other similar option is Tycho Brahe, which has not only POS tags but also syntactic annotation not yet available in Colonia. Tycho Brahe has, however, less than half of Colonia's size and its texts are sampled from a different time span.

[4] There is no available data about the size of this corpus. Based on the material we compiled, we estimate roughly a total of 7 million tokens.

3 Methodology

Corpus compilation was done by collecting material from three main sources: *Dominio Público*[5], a digital library of non-copyrighted media maintained by the Brazilian Ministry of Education, and texts from other two aforementioned Portuguese historical corpora: *GMHP* and Tycho Brahe.

Compilation of texts resulted in a total corpus size of 5,157,982 tokens. The corpus was split into 5 sub-corpora by century. The total number of texts and tokens from each of these sub-corpora is presented in the table number 2 and a complete inventory of the texts used for Colonia is available online[6].

Century	**Texts**	**Tokens**
16^{th} Century	13	399,245
17^{th} Century	18	709,646
18^{th} Century	14	425,624
19^{th} Century	38	2,490,771
20^{th} Century	17	1,132,696
Total	100	**5,157,982**

Table 2: Corpus Size by Century

The difference between Portuguese two main language varieties was taken into account when collecting the texts for Colonia. The corpus contains texts published by Brazilian and European authors in a balanced proportion (52 Brazilian texts and 48 European texts).

3.1 Annotation and Post-Processing

After its compilation, the corpus was POS tagged using the IMS Stuttgart's TreeTagger [8] along with a parameter file for Portuguese [9]. TreeTagger is a language independent probabilistic tagger that arranges annotated data in a three column format (original token, POS tag and lemma). Studies report that TreeTagger can achieve performance higher than 95% accuracy in

[5] http://www.dominiopublico.gov.br/

[6] http://corporavm.uni-koeln.de/colonia/inventory.html

attributing the correct POS tag and lemma of a token [8]. Parsers and POS taggers are designed to be used on contemporary standard data. Given the particularities of historical data, a substantial number of tokens were tagged with an unknown lemma. This occurs because of spelling variation and we tried to address this outcome on a post-processing stage.

The tagset used to annotate the corpus contains not only the classic POS tags (e.g. V, DET, N) but also a couple of compound tags, such as the combination of preposition plus determiner as (PREP+DET) or verb plus pronoun (V+P). The tagset used for the annotation is presented in table 3.

Category	**POS**	**Example**
Adjective	ADJ	bonita
Adverb	ADV	muito
Determinant	DET	os
Cardinal	CARD	primeiro
Noun	NOM	mesa
Pronoun	P	eles
Preposition	PREP	de
Verb	V	fazer
Interjection	I	Oh!
Commas	VIRG	,
Punctuation	SENT	.

Table 3: Tagset

After the annotation, we carried out a post-processing stage. At this stage, our first decision was to not normalize any spelling variation. We processed texts as they were originally collected, hence some texts were already orthographically normalized before being compiled, and this information can be found in the inventory of texts. As the oldest material for Colonia was written in the 16^{th} century, after this point Portuguese orthography had a moderate degree of variation [12] that could be grasped by rules. After the 19^{th} century very little variation has taken place and we therefore focused on post-processing texts ranging from the 16^{th} to the 18^{th} century.

A semi-automatic method was then applied using scripts to address unknown lemma words[7]. We first identified the words that were both attributed an unknown lemma and were not proper nouns. This list of words with unknown lemma was then scrutinized in search for patterns that were coded in scripts to attribute the correct lemma to these words.

3.2 Queries and Interface

After annotation and post-processing, the Corpus WorkBench (CWB) and Corpus Query Processor (CQP) [10] were used to index the corpus and allow search queries in this data. On top of this structure, CQPWeb [11] provides a web interface between data, query processor and the user. CQPWeb is a tool that works together with CQP to allow users to make queries through a user-friendly interface, providing most functions that state-of-the-art corpus processors do. The corpus is in the final stage of processing and is available at the following address: http://corporavm.uni-koeln.de/colonia.

4 Conclusion and Further Perspectives

We presented a new resource for the study of Portuguese, the Colonia Corpus of Historical Portuguese. The corpus contains POS annotation and user-friendly interface which enables researchers to perform shallow syntactic analysis using corpus queries.

We are continuing improving the corpus. At the moment we are carrying out experiments with fine-grained tags that take into account morphological information, similar to the annotation used in the CLAWS[8] tagset [13]. Given the aforementioned challenges of annotating historical data, fine-grained annotation constitutes an additional level of complexity if compared to the coarse-grained tags presented here. We are training RFTagger [14], which is a tool similar to TreeTagger, but developed specially for fine-grained tags. The annotation with fine-grained tags will not

[7] At this stage, we did not handle incorrect POS Tags.

[8] http://ucrel.lancs.ac.uk/claws/

substitute the current annotation but it will be available along with the coarse-grained annotation in the near future.

The work carried out so far is being expanded to medieval Portuguese texts. Moreover, this work is being replicated to other Iberian languages as well, such as Galician and Spanish. We aim to apply the same methodology and annotation described in this paper allowing researchers to make comparative and contrastive analysis across these languages.

Acknowledgments

The authors thank Mario Viaro from *Grupo de Morfologia Histórica do Português (GMHP)* of the University of Sao Paulo and Charlotte Galves from the University of Campinas for part of the material used to compile Colonia. We would also like to thank Sascha Diwersy for the help with CQP and Sanja Štajner for commenting a draft version of this paper.

References

[1] Lewis, P.; Simons, G.; Fennig, C. (2013) *Ethnologue: Languages of the World*. 17th edition. SIL International.

[2] Santos, D. (2000) O Projecto Processamento Computacional do Portugues: Balanço e Perspectivas. *Actas do V PROPOR*. p. 105-113.

[3] Rocha, P.; Santos, D. (2000) CETEMPúblico: Um Corpus de Grandes Dimensões de Linguagem Jornalística Portuguesa. *Actas do V PROPOR (2000)*. p. 131-140.

[4] Bacelar do Nascimento, M. (2000) Corpus de Reference du Portugais Contemporain. *Corpus, Methodologie et Applications Linguistiques*. Presse Universitaires de Perpignan. p. 25-30.

[5] Berber Sardinha, T. (2010) *Corpus Brasileiro*. Available in: http://corpusbrasileiro.pucsp.br/x/

[6] Galves, C.; Faria, P. (2010) *Tycho Brahe Parsed Corpus of Historical Portuguese*. URL: http://www.tycho.iel.unicamp.br/ tycho/corpus/en/index.html

[7] Davies, M.; Ferreira, M. (2006) *Corpus do Português: 45 Million Words, 1300s - 1900s*. URL: http://www.corpusdoportugues.org

[8] Schmid, H. (1994) Probabilistic Part-of-Speech Tagging Using Decision Trees.*Proceedings of International Conference on New Methods in Language Processing*. Manchester, UK.

[9] Gamallo, P. (2005) *TreeTagger Portuguese Parameter File*. URL: http://gramatica.usc.es/ gamallo/tagger.htm

[10] Christ, D. (1994) A modular and flexible architecture for an integrated corpus query system. *Papers in Computational Lexicography (COMPLEX '94)*. p. 22-32.

[11] Evert, S.; Hardie, A. (2012) Twenty-first Century Corpus Workbench: Updating a Query Architecture for the New Millenium. *Proceedings of Corpus Linguistics 2011*. Birmingham, UK.

[12] Castro, I. (1991) *Curso de História da Lingua Portuguesa*. Universidade Aberta, Lisboa.

[13] Garside, R., and Smith, N. (1997) A hybrid grammatical tagger: CLAWS4, in Garside, R., Leech, G., and McEnery, A. (eds.) *Corpus Annotation: Linguistic Information from Computer Text Corpora*. Longman, London, p. 102-121

[14] Schmid, H.; Laws, F. (2008): Estimation of Conditional Probabilities with Decision Trees and an Application to Fine-Grained POS Tagging, *Proceedings of COLING 2008*, Manchester, Great Britain.