A large Portuguese corpus on-line: cleaning and preprocessing

Michel Généreux, Iris Hendrickx, and Amália Mendes

Centro de Linguística da Universidade de Lisboa,
Av. Prof. Gama Pinto, 2,
1649-003 Lisboa - Portugal
{genereux,iris,amalia.mendes}@clul.ul.pt,
WWW home page: http://www.clul.ul.pt

Abstract. We present a newly available on-line resource for Portuguese, a corpus of 310 million words, a new version of the Reference Corpus of Contemporary Portuguese, now searchable via a user-friendly web interface. Here we report on work carried out on the corpus previous to its publication on-line. We focus on the processes and tools involved for the cleaning, preparation and annotation to make the corpus suitable for linguistic inquiries.

Keywords: Corpus, Cleaning, Linguistic Preprocessing

1 Introduction

The aim of this paper is to present our work in preparing a large Portuguese corpus, the Reference Corpus of Contemporary Portuguese (CRPC[1]), into a suitable format for on-line publication and the enrichment of the corpus with automatically assigned pos-tags and lemmas. We hope that sharing our experience in preparing a Portuguese corpus for on-line querying can be of genuine general interest, given the predominance of platforms designed and developed mostly for English. The approaches, practices and techniques described are not novel, although we present them in such a way as to underline key points and potential pitfalls. Language technologists engaged in preparing and publishing corpora on-line may find here some useful insights. The CRPC [3] has been developed at the *Centro de Linguística da Universidade de Lisboa* (CLUL[2]) for more than two decades. This is an electronically based linguistic corpus of written and spoken materials, with a total of 311 million tokens. The written part of this corpus covers 309,812,943 tokens, 1,146,189 types, compiled from 356,208 documents and it is now available online. The corpus covers essentially the chronological period between 1970 to 2008, although texts from 1850 forward are also included (mainly fiction books and parliamentary debates). Our main focus is European

[1] A full description can be found here: http://www.clul.ul.pt/en/research-teams/408-crpc-description

[2] http://www.clul.ul.pt/

Portuguese (see table 1), but other varieties of Portuguese are represented. These sub-parts are not comparable in size since they depend on the availability of data (we try to assure that only texts from native speakers of these varieties with no external linguistic influences are included). The corpus materials are taken by sampling from several types of written texts, chosen to assure as much text diversity as possible, but also according to the availability of the materials. Texts were obtained from different sources and this will reflect strongly on the cleaning procedure presented in the following section. Most recent newspapers, books and magazines were downloaded from the internet, others were obtained directly in digital format from their owners, like the parliamentary debates. But to assure diversity in terms of time period, text type, technical and didactic texts, and Portuguese varieties, we needed to use original texts in paper format. These were prepared in a time-consuming three-step process: digitalization with OCR, manual correction and final revision by a different team member. Our objective when compiling the corpus was closer to the notion of a monitor corpus and, for this online version, although we excluded some of the data, we decided to make as much of this material available as possible. Text diversity and corpus balance are yet aspects to improve in future versions.

Country	Texts	Tokens	**Type**	Texts	Tokens
Portugal	93.3%	289,840,619	Newspaper	50.8%	110,503,376
Angola	5.5%	10,744,627	Politics	45.9%	163,267,089
Cape Verde	0.3%	1,449,269	Magazine	1.4%	7,581,850
Macau	0.3%	2,086,763	Various	1.2%	4,806,176
Mozambique	0.2%	1,126,299	Law	0.3%	2,927,953
Sao Tome and Principe	0.2%	537,600	Book	0.3%	20,557,296
Brasil	0.2%	3,539,770	Correspondence	0.03%	88,370
Guinea Bissau	0.04%	364,421	Brochure	0.01%	80,833
Timor	0.0008%	123,575	–	–	–
Total	100%	309,812,943	Total	100%	309,812,943

Table 1. Text and Token distribution of the CRPC

The corpus and its access through its web-interface[3] provide an important resource for linguistic studies and NLP research on Portuguese especially because it is the largest and diversified corpus of European Portuguese to be made available on-line. The platform provides extensive search options for concordances of word forms, sequences of words and POS categories. It allows for restricted query per variety and text type (and other meta-data if one uses the CQP query syntax), and provides collocations using different statistical measures. The full set of options is described in the CRPC manual on the platform. This new platform is already proving extremely useful for ongoing projects.

[3] `http://alfclul.clul.ul.pt/CQPweb/`

2 Related Work

We refer to [15] for a full overview of the history of corpus development for Portuguese. Here we only discuss corpora which are available online and which share similar features and purpose with CRPC.

The Lácio-Web project[4] [1] was a 2.5 year project aimed at developing a set of corpora for contemporary written Brazilian Portuguese, namely a reference corpus of size 8,291,818 tokens, a manually verified portion of the reference corpus tagged with morpho-syntactic information, a portion of the reference corpus automatically tagged with lemmas, syntactic and POS-tags [2], two parallel and comparable corpora of English-Portuguese and a corpus of non-revised texts. In total, the Lácio-Web corpora together comprise around 10 million words. These corpora can be accessed online and are a follow-up of the NILC Corpus, a corpus of 32M tokens, developed at NILC and available at the Linguateca site, in the scope of the AC/DC project.

The Portuguese Corpus[5] contains 45 million words from Brazilian and European Portuguese taken from the 14th to the 20th century. It includes texts from other corpora, such as the Tycho Brahe corpus[6] and the above mentioned Lácio-Web reference corpus. The corpus is available online via a web interface that allows users to search for word lemmas, pos-tags, frequencies, collocations and restrict their queries for registers, countries or time periods.

The AC/DC[7] project (Acesso a Corpos/Disponibilização de Corpos) aims at having one website where many different corpora are available under a practical user interface. The web interface is based on the same architecture underlying the CRPC, the IMS Open Corpus Workbench (CWB). CETEMPúblico [16] is the largest of the available corpora and contains around 190 million words from the Portuguese newspaper *Público*.

The Bank of Portuguese[8] [17] is a result of joining several corpora together to form one large corpus of nearly 230 million words. A small part of the corpus, 1.1 milion words, is available for online search of concordances.

3 Cleaning

The CRPC is composed of documents from various sources, including internet (88.75% of the documents), which makes it challenging to clean automatically. It seemed therefore appropriate for cleaning the corpus to focus our efforts on a two-step approach, the first designed to get rid of metatags, and the second addressing directly lexical content. This two-step approach allows specialized algorithms to work more efficiently, as it proves much more difficult to process data coming from diverse sources in one single pass.

[4] http://www.nilc.icmc.usp.br/lacioweb/
[5] http://www.corpusdoportugues.org/
[6] http://www.tycho.iel.unicamp.br/
[7] http://www.linguateca.pt/ACDC/
[8] http://www2.lael.pucsp.br/corpora/bp/

The removal of meta-tags does not require extensive processing, as these labels usually follow a specific structure easily modelled by simple rules. In contrast, the cleaning of the remaining lexical content requires a more sophisticated approach, including methods based on learning lexical models from annotated content according to whether it is relevant or not (such as advertising or spam). In this context, the tool NCleaner [11] appears well suited for cleaning the corpus. This tool has proven very successful on a task aimed at cleaning web page content (*CLEANVAL* 2007). In addition, NCleaner automatically segments the text into short textual units, mainly paragraphs. To our knowledge, NCleaner has not been evaluated for a language other than English, so we provide a comparative evaluation of its application to Portuguese. For details of the approaches used in NCleaner, the reader is referred to [11].

NCleaner requires the creation of an annotated corpus to learn to distinguish "relevant" from "not relevant" segments. In [11], 158 documents (about 300,000 words and 2 million characters) were used to create a model of English vocabulary. For our Portuguese model, we have annotated 200 documents (about 200,000 words and 1.7 million characters) randomly selected among all the 359k documents included in the corpus. These 200 documents were first stripped of meta-tags and segmented by NCleaner. These documents were then handed over to an annotator. The task of our annotator, who was already familiar with the corpus and work in corpus linguistics in general, was to identify typical irrelevant segments that should be removed from the final corpus. This work has produced 1,474 irrelevant segments among the 6,460 segments included in the 200 documents. The most frequent classes of irrelevant segments we found were *titles*, *web navigation controls*, *copyrights* and *dates*. Some examples of irrelevant segments:

- *OUTROS TÍTULOS EM SOCIEDADE* [Title]
- *Regresso à página anterior* [Web navigation control]
- *Copyright 1998 Sojornal. Todos os direitos reservados.* [Copyright]
- *TERÇA-FEIRA, 30 DE JULHO 1996* [Date]

Regardless of the category to which they belong, these segments share a common characteristic: they do not represent a typical use of language within a collection of texts of a specific genre and on a defined subject, and distort the analysis of language that human experts, but especially NLP tools, could produce. However, we recognize that this definition of *noise* in the corpus is rather schematic and may be advantageously complemented by a more comprehensive list of general categories.

We also wanted to compare the lexical cleaning phase of NCleaner with two other approaches. The first approach of [8] originally designed to identify the language of a text is based on a comparison of the statistical distribution of words and groups of letters (N-grams). The second approach is that of SVM (*Support Vector Machine*) [13] and deemed successful for text classification tasks[9]. The results of this comparison with NCleaner are presented in Table 2.

[9] See also BeautifulSoup: `http://www.crummy.com/software/BeautifulSoup/`.

Approach	Parameters setting	F-score
N-GRAMS	Sequences of 5 letters or less	82%
SVM	500 Most frequent words	89%
NCLEANER	We keep accented letters	90%

Table 2. Comparative evaluation (at the level of the segment) of three approaches for cleaning the corpus

All of the 6,460 annotated segments were used for the evaluation, 75% (4,845) dedicated to learning and 25% (1,615) for testing. We see that NCleaner performs best with an F-score comparable to the results obtained for English during *CLEANVAL* 2007 (91.6% at the word level). Applied to the entire corpus corpus, NCleaner reduced the number of tokens from 433 to 310 millions, a reduction of about 28%. The number of documents decreased from 359k to 356K[10].

4 Linguistic Preprocessing

We conducted the following types of automatic linguistic preprocessing: tokenization, POS-tagging and lemmatization. For tokenization we applied the LX-tokenizer [6] which splits punctuation marks from words and detects sentence boundaries. This tokenizer is developed specially for Portuguese and can handle typical Portuguese phenomena such as contracted word forms and clitics (including middle clitics).

For POS-tagging we compared two POS-taggers against each other and used the most performant to tag the full corpus. We compared MBT [10], a memory-based tagger to the LX-tagger [7]. The LX-tagger is a state-of-the-art tagger and has been applied to Portuguese with a reported accuracy of 96.87%.

We used the written CINTIL[11] corpus for training and evaluating the taggers, this corpus consists of a mixture of newspaper and fictional texts and is annotated with POS and lemma information and manually verified [4]. The tagset used in the CINTIL corpus is the result of extensive testing of different annotation options in previous projects and seems to best fit requirements for linguistic analysis. For the evaluation experiment we split the data in a 90% training set and a 10% test set. As MBT has features and parameters to be set, we ran ten-fold cross-validation experiments on the training set for finding a suitable setting. The LX-tagger was used without any modification. We measured the performance on the test set of 86,078 tokens in accuracy and F-score as shown in Table 3. We can observe that MBT outperforms the LX-tagger and therefore we used MBT to tag the full corpus. MBT splits the data in two categories: known and unknown words. Known words are the words present in the training data. For the known words only a limited set of POS labels is available, and the

[10] Some documents having been completely emptied of their contents.

[11] `http://cintil.ul.pt/cintilfeatures.html`

tagger can pick a label from this set. However, for unknown words, all labels need to be considered. Here the context of the unknown word as well as prefix and suffix information are useful features to predict the POS-label. For the known words, MBT achieves an accuracy of 96% on the test set while for the small subset of unknown words (5,664 tokens), it has an accuracy of 88.2%. The most frequent errors that are made by MBT match the type of decisions that are also difficult for humans to make such as distinguishing if the word "que" is a relative or a conjunction, and if a word like "italiano" is a common noun or an adjective. Also, proper names are often a source of errors as they are the same as normal dictionary words.

For the labelling of the full corpus, we trained the POS-tagger on a slightly adapted version of the CINTIL data. In CINTIL, contracted word forms are all split, while for our purpose we want to keep the contraction in the corpus. For example the contraction *das* (*de* "from" and *as* "the") receives a double tag 'PREP+DA' indicating that it is both a preposition and a definite article. After studying the CINTIL annotation we noticed that the multi-word units (MWU) as labelled in CINTIL were problematic for the tagger as they have a low frequency and are easily confused with other POS tags, let alone the fact that they are not always consistently annotated within the CINTIL corpus. CINTIL contains 900 different MWU types of which 425 occur once and many are genuine idiomatic expressions. We decided to decompose the MWU. The POS-tagger was trained on this adapted version of the CINTIL corpus containing 643,697 tokens and 30,344 sentences. We used the main POS-tag labels in CINTIL[12], which can be considered as a "simple" version of the tags that leaves out the more detailed information about genre, number, time, etc.

Evaluation	MBT	LX-tagger
Accurary	95.48	94.06
F_micro	95.42	93.92
F_macro	70.10	66.40

Table 3. Results on the test set of the two POS-taggers MBT and LX-tagger.

As freely available Portuguese lemmatizers are scarce, we decided to convert an existing lemmatizer to the Portuguese language. We chose MBLEM [5], a lemmatizer developed by the ILK research group[13] with a very good performance for Dutch and English. MBLEM combines a dictionary lookup with a machine learning algorithm to produce lemmas. The classifier is a memory-based learning algorithm [9] which is very well suited for lemmatization as all previous seen cases are stored in memory. In practice, this means that the full dictionary is stored and all exceptional cases (e.g. irregular verb forms) are kept. The learning algorithm

[12] We refer to the CINTIL annotation manual for details on the tag set.
[13] `http://ilk.uvt.nl`

learns to associate transformation rules that map word forms to their lemmas. As basis for the dictionary we used a list of wordform - POS-tag combinations mapped to lemmas. This list was produced in-house. The dictionary used in MBLEM contains 102,196 word forms combined with 27,860 lemmas, leading to 120,768 wordform-lemma combinations.

To evaluate the performance of MBLEM we used a sample of 50,000 words from the written CINTIL as test set. As CINTIL has been tagged with a different set of POS-tags (80 different main tags) than the set of tags listed in the dictionary (31 tags), we asked a Portuguese linguist to create a mapping between the two POS-tag sets. The mapping was quite straight-forward, almost all CINTIL tags could be mapped against a suitable coarse-grained dictionary tag, although in some cases (e.g. numeral adjectives) categories were treated differently in the two tag sets. Note that MBLEM predicts a lemma for each token in the file, but in CINTIL not all tokens have a lemma, function words such as prepositions and adverbs do not. In our test set, 17,117 word forms have a gold-standard annotated lemma. MBLEM achieves an accuracy of 96.7% on this test set.

5 Conclusion and Perspectives

We have presented the on-line publication of the written sub-part of the CRPC, a large and diverse Portuguese corpus. We have discussed its internal constitution, available resources for cleaning and preprocessing such a corpus, and the new platform used for online queries. The current version of the corpus can be used for lexical studies as well as a resource for NLP applications. The corpus has already been used in many projects and studies, the most recent being a study of comparable subcorpora of Portuguese varieties [14] and a computational study that compares lexicons from different time period [12]. Future work includes a second phase of cleaning that will focus on improving segmentation and consolidating our lexical model, as well as adding more searchable meta-data tags and introducing a language spotter for the few remaining pockets of foreign languages present in the corpus. We are planning the development of a constituent chunker for Portuguese so that the corpus can be enriched with syntactic annotations. We also plan to enlarge the corpus annotation to cover information on nominal and verbal inflection (genre, number, person, tense, etc.) based on the CINTIL annotation schema and to address the issue of MWU. Since its publication online, the platform has been visited and used extensively by users from all over the world.

Acknowledgement

We thank Luísa Alice Santos Pereira for sharing with us her long experience of the CRPC and Paulo Henriques for his work on the cleaning task. This work is financed by the Fundação para a Ciência e a Tecnologia (FCT) for the project PEst-OE/LIN/UI0214/2012, by Fundação Calouste Gulbenkian and by the FCT Doctoral program Ciência 2007/2008.

References

1. Aluísio, S., Pinheiro, G.M., Manfrin, A.M.P., de Oliveira, L.H.M., Jr., L.C.G., Tagnin, S.E.O.: The lacio-web: Corpora and tools to advance brazilian portuguese language investigations and computational linguistic tools. In: Proceedings of 4th conference on International LREC. pp. 1779–1782 (2004)
2. Aluísio, S.M., Pelizzoni, J.M., Marchi, A.R., de Oliveira, L., Manenti, R., Marquiafável, V.: An account of the challenge of tagging a reference corpus for brazilian portuguese. In: Proceedings of PROPOR 2003. pp. 110–117 (2003)
3. Bacelar do Nascimento, M.F., Pereira, L., Saramago, J.: Portuguese Corpora at CLUL. In: Second International Conference on Language Resources and Evaluation (LREC2000. vol. II, pp. 1603–1607. Athens (2000)
4. Barreto, F., Branco, A., Ferreira, E., Mendes, A., Bacelar do Nascimento, M.F.P., Nunes, F., Silva, J.: Open resources and tools for the shallow processing of portuguese. In: Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC2006). Genoa, Italy (2006)
5. van den Bosch, A., Daelemans, W.: Memory-based morphological analysis. In: for Computational Linguistics, A. (ed.) Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics, ACL'99. pp. 285–292 (1999)
6. Branco, A., Silva, J.: Contractions: breaking the tokenization-tagging circularity, vol. 2721, chap. Lecture Notes in Artificial Intelligence, pp. 167–170. Spinger (2003)
7. Branco, A., Silva, J.: Evaluating solutions for the rapid development of state-of-the-art pos taggers for portuguese. In: Proc. of LREC 2004. pp. 507–510 (2004)
8. Cavnar, B., Trenkle, J.M.: N-gram based text categorization. In: Proceedings of the Third Annual Symposium on Document Analysis and Information Retrieval. pp. 161–175 (1994), uNLV Publications/Reprographics
9. Daelemans, W., Van den Bosch, A.: Memory-Based Language Processing. Cambridge University Press, Cambridge, UK (2005)
10. Daelemans, W., Zavrel, J., Berck, P., Gillis, S.: Mbt: A memory-based part of speech tagger generator. In: Proceedings of the 4th ACL/SIGDAT Workshop on Very Large Corpora. pp. 14–27 (1996)
11. Evert, S.: A lightweight and efficient tool for cleaning web pages. In: 6th International Conference on Language Resources and Evaluation (LREC 2008). Marrakech, Morocco (2008)
12. Généreux, M., Mendes, A., Pereira, L.A.S., do Nascimento, M.F.B.: Lexical analysis of pre and post revolution discourse in portugal. In: Proceedings of the 3rd workshop on building and using comparable corpora (LREC 2010 (2010)
13. Joachims, T.: Learning to Classify Text Using Support Vector Machines. Ph.D. thesis, Cornell University, USA (2002), kluwer Academic Publishers / Springer
14. do Nascimento, M.F.B., Estrela, A., Mendes, A., Pereira, L.: On the use of comparable corpora of african varieties of portuguese for linguistic description and teaching/learning applications. In: Proceedings of the Workshop on Building and Using Comparable Corpora (LREC 2008) (2008)
15. Santos, D.: Linguateca's infrastructure for portuguese and how it allows the detailed study of language varieties. Oslo Studies in Language 3(2) (2011)
16. Santos, D., Rocha, P.: Evaluating CETEMPublico, a Free Resource for Portuguese. In: Proceedings of 39th Annual Meeting of the Association for Computational Linguistics (ACL). pp. 450–457. Association for Computational Linguistics, Toulouse, France (July 2001)
17. Sardinha, T.B.: History and compilation of a large register-diversied corpus of Portuguese at CEPRIL. The Especialist 28(2), 211–226 (2007)