Introducing the Reference Corpus of Contemporary Portuguese Online

Michel Généreux, Iris Hendrickx, Amália Mendes

Centro de Linguística da Universidade de Lisboa
Av. Prof. Gama Pinto, 2
1649-003 Lisboa - Portugal

Abstract

We present our work in processing the Reference Corpus of Contemporary Portuguese and its publication online. After discussing how the corpus was built and our choice of meta-data, we turn to the processes and tools involved for the cleaning, preparation and annotation to make the corpus suitable for linguistic inquiries. The Web platform is described, and we show examples of linguistic resources that can be extracted from the platform for use in linguistic studies or in NLP.

Keywords: Corpus management, Preprocessing, Portuguese

1. Introduction

The main purpose of this paper is to introduce a resource for Portuguese that is now available and searchable online: the Reference[1] Corpus of Contemporary Portuguese (CRPC)[2], developed at the *Centro de Linguística da Universidade de Lisboa* (CLUL)[3]. The CRPC is an electronically based linguistic corpus of written and spoken texts, containing 312M tokens. In the next section we present the corpus in more detail, followed by a description of the linguistic and computational uses of the corpus. In the second part of the paper we describe the corpus processing step and explain how the corpus was automatically cleaned and how the linguistic annotation was produced. In the two final sections, we present a brief overview of similar Portuguese resources and conclude.

2. Corpus constitution

The CRPC has been an ongoing project for more than 20 years, and it has recently undergone major changes. We present here the new version of the corpus, which contains now 312M tokens (310M written and 1,6M spoken). The compilation of the CRPC started in 1988 and its main goals then are still valid today: to keep an up-to-date and balanced version of the corpus that can serve as a representative sample for the Portuguese language, both in its written and spoken variety.

With this objective in mind, we have strived during the compilation of the CRPC to sample from several types of written texts (literature, newspapers, magazines, science, economics, law, parliamentary debates, technical and didactic texts, pamphlets). The CRPC represents essentially the European Portuguese language, although it also covers (to a much lesser extent) national and regional varieties of Portuguese, including European, Brazilian, African (Angola, Cape Verde, Guinea-Bissau, Mozambique and São Tomé and Principe) and Asiatic Portuguese (Macao, Goa and East-Timor). From a chronological point of view, our corpus contains texts from the second half of the XIX century up until 2008, albeit mostly after 1970, since the corpus focuses on contemporary Portuguese (Bacelar do Nascimento et al., 2000; Bacelar do Nascimento, 2000). To achieve this broad variety in terms of text types, geography and time period, and considering that a significant part of the corpus was gathered in a period when internet was not yet the important communication channel that it is today, the compilation of texts which were (and still are in many cases) non available in digital format (such as didactic, fiction, pamphlets, documents of non European varieties) required that these had to be scanned with OCR, manually corrected and revised. This time-consuming task has assured a wider representativity of the CRPC, compared to other written corpora of European Portuguese. The same concern was present for the compilation of the spoken subcorpus, which has been enlarged in the scope of several projects, the latest being C-ORAL-ROM[4], that enabled the production of a European spoken corpus comparable to corpora of three other romance languages. The C-ORAL-ROM corpus is available through the ELDA catalogue and on CD-ROM (Bacelar do Nascimento et al., 2005). The two main categories of formal and informal registers are divided into finer-grained types like non-media (e.g. preaching, political debate, teaching), media (news, sports, meteorology), private phone conversations, phone services, conversations, monologues, etc. The transcriptions of the spoken sub-corpus Fundamental Portuguese[5] and the recordings and transcriptions of the sub-corpus Spoken Portuguese[6] are available for download.

[1] The term "reference corpus" is used to convey the idea that the corpus is designed to provide comprehensive information about contemporary Portuguese (and not because it is a reference in the field).

[2] `http://www.clul.ul.pt/en/research-teams/408-crpc-description`

[3] `http://www.clul.ul.pt`

[4] `http://www.clul.ul.pt/en/research-teams/189-c-oral-rom-integrated-reference-corpora-for-spoken-romance-languages`

[5] `http://www.clul.ul.pt/en/research-teams/84-spoken-corpus-qportugues-fundamental-pfq-r`

[6] `http://www.clul.ul.pt/en/research-teams/83-spoken-portuguese-geographical-and-social-varieties-r`

Throughout its history, the CRPC has been constantly enlarged, so this new version (v.2.2) that we present here differs substantially from the previous version as described in (Bacelar do Nascimento et al., 2000). We have increased the size of the corpus considerably as the previous version counted around 92M words, and now 312 M. Most of this additional material was gathered via web crawling. Furthermore, all texts have been automatically cleaned and linguistically annotated with POS-tags and lemmas as will be described in more detail in this paper. The most important improvement however is that currently the full CRPC written sub-part is accessible online using the CQPweb[7] technology. In the older version, only a written subpart of 11,4M tokens was available for online queries, composed of the ELAN corpus[8] with 2,8M tokens and of the RL corpus[9] with 8,6M tokens. This on-line access enabled searches according to different sub-corpora (per text type) and provided concordances and frequencies. In comparison, the CQPweb platform allows powerful queries that will be discussed in section 3.

We believe that, due to its on-line availability, dimension and diversity, the CRPC is a useful resource for all researchers, national and foreign, working on the Portuguese language to whom there is a need for reliable linguistic data. Some *Quick Facts* for the CRPC written sub-part are summarized in table 1 while tables 2 and 3 show text and token distribution of the written part of the corpus. The percentages in table 3 show that the current version is not balanced in terms of different tex types, but if we look at the number of tokens, we observe that we do have substantial amounts of material for all text types, especially if we take into account that types like books and magazines were not retrieved from the internet but manually gathered.

Nb. Types	1.15 M
Nb. Tokens	310 M
Documents	356 K
Annotations	POS, lemmas, NP-chunks
Metadata tags	44
Metadata online	24
Corpus Manager	CQPWeb
URL	http://alfclul.clul.ul.pt/CQPweb/

Table 1: Some Quick Facts regarding the written CRPC.

2.1. Meta-data

Each document in the CRPC is classified in terms of analytic, descriptive and editorial meta-data regarding source, text type (book, review, newspaper, etc.), topic and language variety. In total we have 44 different meta tags in the CRPC. For each major type a particular combination of text-descriptive features is assigned: for example, the set of descriptive meta-data for newspapers includes information on the sections, while for didactic books it covers the course name and the curricular year. Other general descriptive meta-data address a set of bibliographic information like title, editor, country of edition, date of edition and the author's name. Since the corpus covers different time periods and national varieties of Portuguese, a set of descriptive meta-data gives detailed information on the year and country of birth of the author, as well as his first language and the country whose variety he represents. For example, some authors born in Portugal and whose first acquired variety might be European Portuguese have in fact been living in Mozambique and their works are to be classified as pertaining to the Mozambique variety in the corpus. Other descriptive meta-data focus on the file properties: its name, size in tokens and location in the corpus directories. Finally editorial meta-data describe the status of the file in terms of its correction and normalization (e.g. there are two levels of correction for texts that are scanned with OCR: corrected and revised). We only display 24 of the meta data tags in the on-line interface as the other tags are for internal use. For the 24 meta-tags, not all values are consistently annotated and can have a value of zero ("NIL"). In figure 1 we show an example of the 24 meta data tags that have been assigned to a file containing excerpts of a book in the CRPC.

Country	Texts	Tokens
Portugal	93.3%	289,840,619
Angola	5.5%	10,744,627
Cape Verde	0.3%	1,449,269
Macau	0.3%	2,086,763
Mozambique	0.2%	1,126,299
Sao Tome and Principe	0.2%	537,600
Brasil	0.2%	3,539,770
Guinea Bissau	0.04%	364,421
Timor	0.0008%	123,575
Total	100%	309,812,943

Table 2: Text and Token distribution by country

Type	Texts	Tokens
Newspaper	50.8%	110,503,376
Politics	45.9%	163,267,089
Magazine	1.4%	7,581,850
Various	1.2%	4,806,176
Law	0.3%	2,927,953
Book	0.3%	20,557,296
Correspondence	0.03%	88,370
Brochure	0.01%	80,833
Total	100%	309,812,943

Table 3: Text and Token distribution by text type

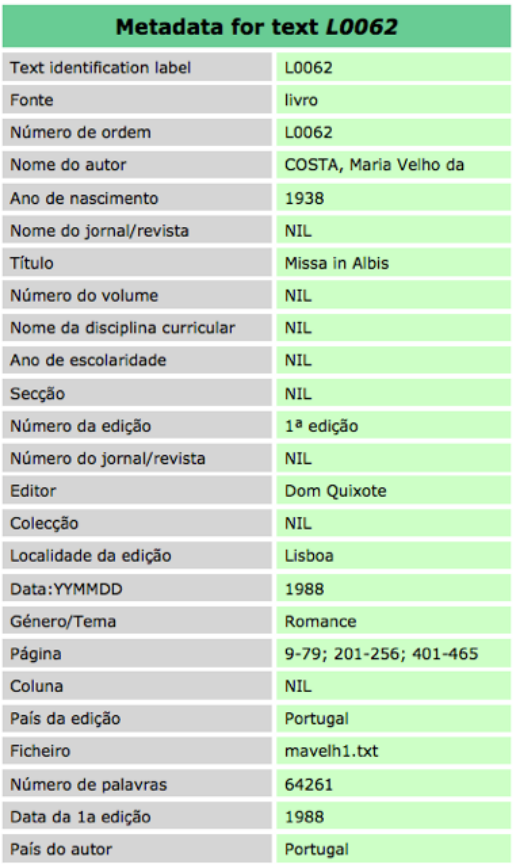
3. Linguistic and Computational Uses of the CRPC

The CRPC and its access through our CQPweb platform provide an important resource for linguistic studies and NLP research on Portuguese especially because it is the first

[7] http://cqpweb.lancs.ac.uk/

[8] http://www.clul.ul.pt/en/research-teams/198-european-language-activity-network-elan

[9] http://www.clul.ul.pt/en/research-teams/192-language-resources-for-portuguese-a-corpus-and-tools-for-query-and-analysis



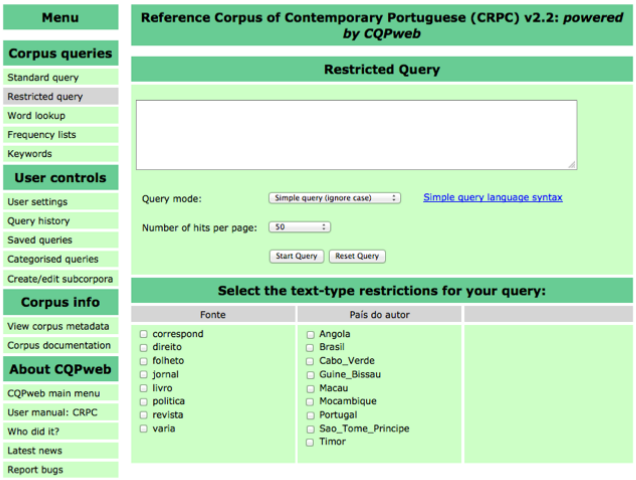
Metadata for text *L0062*	
Text identification label	L0062
Fonte	livro
Número de ordem	L0062
Nome do autor	COSTA, Maria Velho da
Ano de nascimento	1938
Nome do jornal/revista	NIL
Título	Missa in Albis
Número do volume	NIL
Nome da disciplina curricular	NIL
Ano de escolaridade	NIL
Secção	NIL
Número da edição	1ª edição
Número do jornal/revista	NIL
Editor	Dom Quixote
Colecção	NIL
Localidade da edição	Lisboa
Data:YYMMDD	1988
Género/Tema	Romance
Página	9-79; 201-256; 401-465
Coluna	NIL
País da edição	Portugal
Ficheiro	mavelh1.txt
Número de palavras	64261
Data da 1a edição	1988
País do autor	Portugal

Figure 1: Screen shot of meta data tags assigned to a file in containing excerpts of a book entitled "Missa in Albis".

large and diversified corpus of Portuguese to be made available online. The platform provides extensive search options for concordances of word forms, sequences of words and POS categories, and it is already proving extremely useful for ongoing projects. It provides the necessary resources and an extensive collection of data to address or pursue linguistic issues like the status of full predicative verbs vs. light verbs and auxiliary verbs (Duarte et al., 2009), variation in syntactic patterns (Mendes and Estrela, 2008) or modality (Hendrickx et al., 2012).

The search query can be restricted according to country and text type, and also concordances can be further analysed in terms of distribution breakdowns. In figure 2 we show a screenshot of the CQPweb interface for the CRPC. Registered users can create sub-corpora based on metadata, compile and download frequency lists for each sub-corpus. The option *keywords* provides a tool for comparing these frequency lists and it is also possible to automatically identify word forms which occur only in one of these sub-corpora. These options enable contrastive linguistic studies of Portuguese varieties worldwide as well as genre studies.

The result of entering a search query is a list of concordances of the query found in the corpus. The search query itself can be specified using regular expressions including wild cards and one can search for words, phrases, lemmas, POS-tags, NP-chunks or a combination of these. Each concordance can be inspected in more detail and users can enlarge the context surrounding the query; get information on the POS-tags of the concordance or retrieve the meta data of the document from which the concordance was taken. A useful option is the function *thin it* to reduce the number of retrieved collocations to a specific smaller sample size that can be used for inspection or download.



Reference Corpus of Contemporary Portuguese (CRPC) v2.2: *powered by CQPweb*

Restricted Query

Query mode: Simple query (ignore case)

Number of hits per page: 50

Select the text-type restrictions for your query:

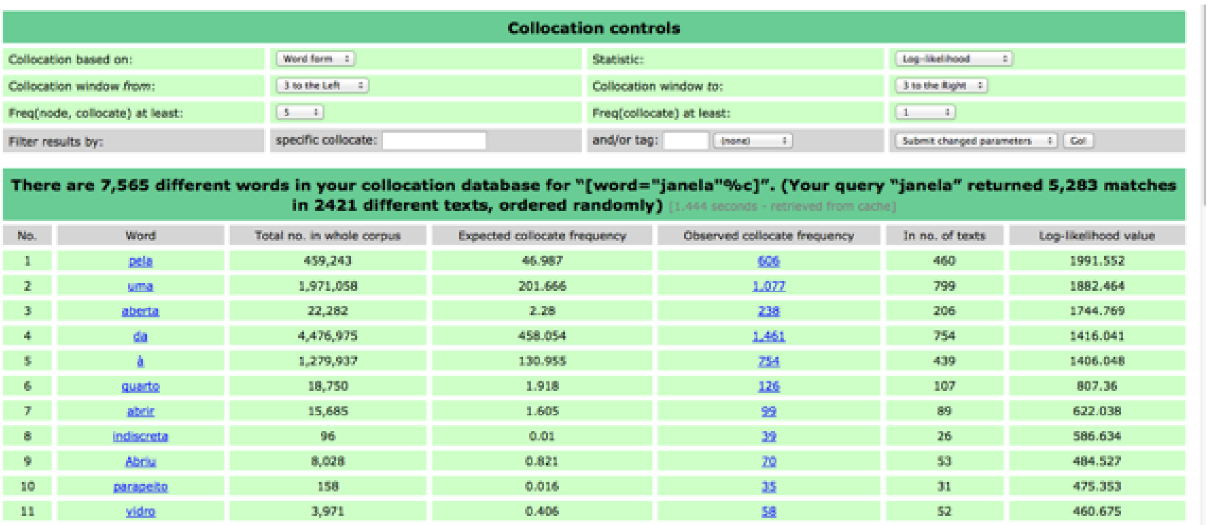
Fonte	País do autor	
□ correspond □ direito □ folheto □ jornal □ livro □ politica □ revista □ varia	□ Angola □ Brasil □ Cabo_Verde □ Guine_Bissau □ Macau □ Mocambique □ Portugal □ Sao_Tome_Principe □ Timor	

Figure 2: Screenshot of the query interface of the CRPC, version 2.2.

The feature *collocation* allows for a full study of the collocational profile of Portuguese words and gives the user the possibility of evaluating results according to different lexical association measures such as mutual information, log-likelihood, or t-score. This feature is used for for a retrieved word or lemma pattern from a standard or restricted query. These results can be used for lexical studies as well as a resource for NLP applications. We show an example in figure 3 for the word *janela* (window) in a search window of 3 words to the left and right using log-likelihood as distance measure, and a frequency threshold of 5. As you can see, contractions of prepositions and determiners (pela (by the), da (of the), à (at the)) are frequently collocated with the word *janela*. The verb *abrir* (to open), the adjective *aberta* (open) also co-occur frequent. The collocation *janela indiscreta* is the portuguese translation of the title of the famous movie "Rear Window" from Alfred Hitchcock, 1954. The word *parapeito* refers to the collocation *parapeito da janela* (window sill).

The possibility of downloading frequency lists provides the NLP researcher with resources to train and develop tools for Portuguese and specific tools targeted at varieties and genres. These frequency lists can be created for word forms, POS-tags, lemmas or NP-chunks and can be filtered on frequency and specific patterns.

The CRPC has already been used in many projects and studies (see webpage of the CRPC), such as a study of comparable CRPC sub-corpora of Portuguese varieties (Bacelar do Nascimento et al., 2008) and a computational study that compares lexicons from pre (1954-74) and post (1974-94) revolution parliamentary discourse in four comparable sub-corpora of the CRPC (Généreux et al., 2010). The application of this diachronic approach to the full CRPC would



Collocation controls

Collocation based on:	Word form	Statistic:	Log-likelihood
Collocation window *from*:	3 to the Left	Collocation window *to*:	3 to the Right
Freq(node, collocate) at least:	5	Freq(collocate) at least:	1
Filter results by:	specific collocate:	and/or tag: (none)	Submit changed parameters Go!

There are 7,565 different words in your collocation database for "[word="janela"%c]". (Your query "janela" returned 5,283 matches in 2421 different texts, ordered randomly) [1.444 seconds - retrieved from cache]

No.	Word	Total no. in whole corpus	Expected collocate frequency	Observed collocate frequency	In no. of texts	Log-likelihood value
1	pela	459,243	46.987	606	460	1991.552
2	uma	1,971,058	201.666	1,077	799	1882.464
3	aberta	22,282	2.28	238	206	1744.769
4	da	4,476,975	458.054	1,461	754	1416.041
5	à	1,279,937	130.955	754	439	1406.048
6	quarto	18,750	1.918	126	107	807.36
7	abrir	15,685	1.605	99	89	622.038
8	indiscreta	96	0.01	39	26	586.634
9	Abriu	8,028	0.821	70	53	484.527
10	parapeito	158	0.016	35	31	475.353
11	vidro	3,971	0.406	58	52	460.675

Figure 3: Screen shot of the collocations for the word janela (window).

provide an insight on lexical changes in Portugal during the last decades.

We are also happy to report that since its introduction online at the end of March 2011, the platform has already responded to an average of more than 1500 queries per month[10] and 41 users from at least eight countries have registered to benefit from extra functionalities. The platform can be accessed without registration, although registered users will be granted user-specific functionalities (e.g. saving queries and creating sub-corpora).

4. Automatic corpus processing

In the next part we present the steps that were taken to transform the raw files into a corpus. First of all, all written material (PDF, OCR scans, word documents, HTML pages) was converted to plain text in UTF-8 character encoding. However, the HTML documents needed an additional cleaning step as is described in the next section. We conclude this section by presenting the automatic annotation of the corpus with POS-tags, lemmas and NP-chunks.

4.1. Cleaning the CRPC

Harvesting a large corpus from heterogeneous sources means that one must be prepared to accept a certain level of noise to be present in the data. In most of the academic contexts, revision and cleaning of such a corpus requires human resources exceeding those normally available. Fortunately, symbolic and statistical-based automatic removal of noisy passages can now be applied with reasonable accuracy to justify no human intervention. This being said, the level of cleaning that is needed remains highly dependent on the sources from which the corpus is drawn. In general, data extracted from the web are by and large the most difficult to clean, because the relevant segments for the corpus are scattered or enclosed among HTML tags, *javascripts* and other meta-data which should not be part of the final corpus. In addition, pages extracted from the web are often littered with adverts and repetitions, let alone cases of outright spamming.

The CRPC is composed of documents from various sources, including internet (88.75% of the documents[11]), 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 HTML markup, 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.

The removal of the markup 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 (Evert, 2008) 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 (Evert, 2008).

NCleaner requires the creation of an annotated corpus to learn to distinguish "relevant" from "not relevant" segments. In (Evert, 2008), 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

[10] As of 14/03/2012.

[11] These include the parliament notes that have a HTML format.

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*.

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 follows the work of Cavnar and Trenkle (1994) and was originally designed to identify the language of a text based on a comparison of the statistical distribution of words and groups of letters (N-grams). The second one is a supervised machine learning approach using Support Vector Machines (SVM) (Joachims, 2002) and deemed successful for text classification tasks[12]. We compare the prediction performance of these three methods on the data set of manually annotated segments that was split in two parts: 75% (4,845 segments) dedicated to learning and 25% (1,615 segments) for testing. We measure the performance in F-score (van Rijsbergen, 1979). The results of this comparison with NCleaner are presented in Table 4. 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[13].

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 4: Comparative evaluation (at the level of the segment) of three approaches for cleaning the corpus

4.2. Linguistic Annotation

All texts in the CRPC were automatically processed to add linguistic information. The texts were tokenized, POS-tagged, lemmatized and chunked at the NP level. For tokenization we applied the LX tokenizer (Branco and Silva, 2003) which removes punctuation marks from words and detects sentence boundaries. This tokenizer was developed specially for Portuguese and can deal with typical Portuguese phenomena such as contracted word forms and verbal clitics (including middle clitics).

We decided to use a slightly adapted version of the CINTIL POS-tagset for POS-tagging the CRPC corpus. This tag set was originally developed for the CINTIL[14] corpus (Barreto et al., 2006), a 1M token sample of the CRPC, annotated with POS and lemma information, manually revised (a joint project of NLX-FCUL[15] and CLUL), based on previous work for the PAROLE corpus and the C-ORAL-ROM corpus.

The main differences between the CINTIL and CRPC corpora are the way word contractions and multi-word units (MWU) are being handled. In CINTIL word contractions were split into the separate word forms, while in the CRPC we kept the contractions to preserve readability. For example the contraction *pela* is split in CINTIL into the preposition *por_* with an underscore to signal the contraction and the determiner *a*. For the CRPC we kept the contracted forms and labeled them with double POS-tags. In the example *pela* is assigned the POS-tag "PREP+DA" indicating that it is both a preposition and a definite article.

In the CINTIL corpus MWU of function words like fixed adverbial or prepositional phrases (for example *por fim* (finally), *de repente* (suddenly)) were tagged with special POS-tags to signal that these tokens form a unit. The written part of CINTIL contains 900 different MWU types and 425 MWU only occur once. When we were preparing an automatic POS-tagger for the CRPC, we noticed that the tagger had many difficulties with these MWU units are as they have a low frequency and are easily confused with other POS tags. Therefor for the CRPC we did not use these MWU POS-tags except for the latin expressions that really have no decompositional meaning otherwise, for example *per capita*.

We decided to use a supervised machine learning approach to train the automatic POS-tagger on an adapted version of the written CINTIL corpus (644K tokens) with contractions and without MWU. In this adapted CINTIL version, we had a set of 80 POS-tag labels which can be considered as a simplified version of the tag set that leaves out the more detailed information about genre, number, time, etc. As automatic tagger we used MBT (Daelemans et al., 1996), a memory-based tagger. To estimate the performance of MBT, we ran some experiments on the adapted CINTIL corpus and compared MBT against another POS-tagger for Portuguese, the LX-tagger (Branco and Silva, 2004). The LX-tagger is a state-of-the-art tagger and has been applied to Portuguese with a reported accuracy of 96.87%. For training and testing, we split the written part of CINTIL in 90% for training and 10% for testing. 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. On the test set of 86K tokens, MBT obtained a F-score of 95.4 against 93.9 F-score for the LX-tagger.

[12] `http://www.crummy.com/software/BeautifulSoup/`

[13] Some documents have been completely emptied of their contents.

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

[15] `http://nlx.di.fc.ul.pt/`

The results on the evaluation set are shown in table 5. As can be expected the performance on known words is much higher than for unknown words. The POS-tagger uses the following information: For known words: information about 3 words and pos-tags to the left and 2 to the right of the focus word that it is trying to tag. For unknown words: the focus word itself was represented with a two-character prefix and a suffix of three characters, whether the word started with a capital letter and whether it contained numbers or hyphens. The local context of unknown words was presented as two words and pos-tags left and right.

Words	Accuracy	# examples
All	95.4	86078
Known	96.0	80414
Unknown	88.2	5664

Table 5: Results of the MBT POS-tagger on the CINTIL evaluation set for known and unknown words.

As we did not encounter a suitable freely available lemmatizer for Portuguese, we decided to convert an existing lemmatizer, MBLEM (Van den Bosch and Daelemans, 1999), that was initially developed for Dutch and English to Portuguese. MBLEM combines a dictionary lookup with a machine learning algorithm to tag words with their lemmas. As dictionary list we used an in-house produced list of lemma and wordform-POS mappings. The dictionary list consists of 102K word forms mapped to 27,860 lemmas with a total of 120,768 wordform-lemma combinations. MBLEM uses the POS information to limit the set of possible lemmas for each word form.

We evaluated the performance of MBLEM on a testing sample of 50K words from the written of the CINTIL corpus. The lemma annotation in CINTIL is limited to content words so only 17K word forms have a gold-standard annotated lemma. As CINTIL has been tagged with another set of POS-tags (80 different tags) than the ones listed in the in-house created dictionary (31 tags), we asked a Portuguese linguist to create a mapping between the two POS-tag sets. In general, this mapping was straightforward as we mapped the fine-grained CINTIL labels to coarse-grained labels for the dictionary. MBLEM achieves a satisfying accuracy of 96.7% on this test set so we could apply MBLEM to lemmatize the full CRPC.

Finally, we chunked the CRPC into noun phrase (NP) constituents. The NP annotations are a slightly modified version of the IOB annotation scheme as proposed by Ramshaw and Marcus (1995). We extended the IOB tags with additional tags to explicitly signal the ends of NPs so that users search on this type of information in the online interface of the CRPC. We used the following tags: *O* (out of NP), *B-NP* (begin a NP), *I-NP* (inside a NP), *E-NP* (end a NP) and *BE-NP* (begin and end a NP). We used the YamCha (Kudo and Matsumoto, 2003) chunker[16] trained on 1,000 random sentences from the CINTIL corpus annotated with complex NPs, which means that NPs may include other constituents (e.g. relative and prepositional clauses, apositives, coordinates, etc.). Despite this challenging endeavour, our chunker obtained a token-level accuracy of 86.5% when cross-evaluated 4-fold on the 1,000 annotated sentences, including 16.5% of (non-critical) errors made solely on specific delimiters and punctuation symbols. In table 6 we show an excerpt from the CRPC that was automatically cleaned, tokenized and annotated with POS-tags, lemmas and NP chunks.

Token	POS	Lemma	NP
Na	PREP+DA	em+a	O
realidade	CN	realidade	BE-NP
,	PNT	,	O
verifica	V	verificar	O
-se	CL	-se	O
que	CJ	que	O
a	DA	a	B-NP
propriedade	CN	propriedade	I-NP
arrendada	PPA	arrenda(r/do)	I-NP

Table 6: Small excerpt from a document of the CRCP corpus that was automatically cleaned, tokenized, POS-tagged, lemmatized and NP-chunked.

5. Related work

Another large corpus of European Portuguese available online is CETEMPúblico (Santos and Rocha, 2001) which contains around 190 million words from the Portuguese newspaper *Público*. It can be accessed and fully downloaded at the Linguateca site, through the AC/DC[17] project (*Acesso a Corpos/Disponibilização de Corpos*). This project aims at having one website where many different corpora (the largest is CETEMPúblico) are available under a practical user interface. The web interface of AC/DC is based on the same architecture underlying the CRPC, the IMS Open Corpus Workbench (CWB).

The Portuguese Corpus[18] 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[19], and the Lácio-Web corpus (see information below). 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.

Several corpora of Brazilian Portuguese have been compiled. The largest is The Bank of Portuguese[20] (Sardinha, 2007) which joined several corpora 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.

The Lácio-Web project[21] (Aluísio et al., 2004) was project aimed at developing a set of corpora for contemporary written Brazilian Portuguese, namely a reference corpus of 8292K tokens, a manually verified portion of the reference

[16] http://chasen.org/~taku/software/yamcha/

[17] http://www.linguateca.pt/ACDC/

[18] http://www.corpusdoportugues.org/

[19] http://www.tycho.iel.unicamp.br/

[20] http://www2.lael.pucsp.br/corpora/bp/

[21] http://www.nilc.icmc.usp.br/lacioweb/

corpus tagged with morpho-syntactic information, a portion of the reference corpus automatically tagged with lemmas, syntactic and POS-tags (Aluísio et al., 2003), 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. There are, of course, many other corpora of smaller dimensions and for other varieties of Portuguese and we refer to (Santos, 2011) for a full overview of the history of corpus development for Portuguese.

6. Conclusion and Future Work

We have presented the preparation and online publication of the Reference Corpus of Contemporary Portuguese, with a focus on available resources for cleaning and preparing such a corpus for queries and navigation as well as on how the platform can be used for developing linguistic resources for NLP. Future work includes a second phase of cleaning that will focus on spotting and discarding excerpts written in a foreign language that currently have not been filtered out. We will also improve segmentation in the cleaning process to consolidate our lexical model and we plan to experiment with the creation of specific lexical models for different text types. In the next version we also plan to add more searchable meta-data besides *fonte* (source) and *país do author* (home country). We also plan to enlarge the CRPC annotation to cover information on nominal and verbal inflection (genre, number, person, tense, etc.) present in the CINTIL annotation schema and to address the issue of MWU. We are gearing up to include the spoken part of the CRPC and respond to requests from members of CLUL wishing to include their own corpora on the platform.

We are currently contacting publishers and authors to acquire authorization for making a part of the CRPC freely available for download. For current online version of the CRPC we chose to make available as much material as possible with the result that some text genres have disproportionally large amounts of material in comparison to other genres. We plan to mitigate this by creating a smaller, balanced version of the material.

Acknowledgement

We thank Luísa Alice Santos Pereira for sharing with us her long experience of the CRPC and Paulo Henriques for his work on annotating the CRPC to produce lexical models for 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, by the FCT Doctoral program Ciência 2007/2008 and by the project METANET4U (CIP-ICT-PSP 270893).

7. References

S. M. Aluísio, J. Marques Pelizzoni, A. R. Marchi, L. de Oliveira, R. Manenti, and V. Marquiafável. 2003. An account of the challenge of tagging a reference corpus for brazilian portuguese. In *Proceedings of PROPOR 2003*, pages 110–117.

S. Aluísio, G. Montilha Pinheiro, A. M. P. Manfrin, L. H. M. de Oliveira, L. C. Genoves Jr., and S. E. O. Tagnin. 2004. The Lacio-Web: Corpora and Tools to Advance Brazilian Portuguese Language Investigations and Computational Linguistic Tool. In *Proceedings of the 4th International Conferenceon Language Resources and Evaluation (LREC2004)*, pages 1779–1782. ELRA.

M. F. Bacelar do Nascimento, L. Pereira, and J. Saramago. 2000. Portuguese Corpora at CLUL. In *Proceedings of the 2nd International Conference on Language Resources and Evaluation (LREC2000)*, volume II, pages 1603–1607.

M. F. Bacelar do Nascimento, J. Bettencourt Gonçalves, R. Veloso, S. Antunes, F. Barreto, and R. Amaro, 2005. *C-ORAL-ROM: Integrated Reference Corpora for Spoken Romance Languages*, chapter The Portuguese Corpus, pages 163–207. Number 15 in Studies in Corpus Linguistics. John Benjamins Publishing Company, Studies in Corpus Linguistics.

M. F. Bacelar do Nascimento, A. Estrela, A. Mendes, and L. Pereira. 2008. On the use of comparable corpora of African varieties of Portuguese for linguistic description and teaching/learning applications. In *Workshop on Building and Using Comparable Corpora, LREC2008*, pages 39–46.

M. F. Bacelar do Nascimento, 2000. *Corpus de Référence du Portugais Contemporain*. H. Champion et Presses Univer. de Perpignan, Paris.

F. Barreto, A. Branco, E. Ferreira, A. Mendes, M. F. Bacelar do Nascimento, F. Nunes, and J. Silva. 2006. 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.

A. Branco and J. Silva, 2003. *Contractions: breaking the tokenization-tagging circularity*, volume 2721, chapter Lecture Notes in Artificial Intelligence, pages 167–170. Spinger.

A. Branco and J. Silva. 2004. Evaluating solutions for the rapid development of state-of-the-art pos taggers for portuguese. In *Proceedings of the 4th International Conferenceon Language Resources and Evaluation (LREC2004)*, pages 507–510. ELRA.

W. B. Cavnar and J. M. Trenkle. 1994. N-gram-based text categorization. In *Third Annual Symposium on Document Analysis and Information Retrieval*, pages 161–175, Las Vegas, April. UNLV Publications/Reprographics.

Crpc.

W. Daelemans, J. Zavrel, P. Berck, and S. Gillis. 1996. Mbt: A memory-based part of speech tagger generator. In *Proceedings of the 4th ACL/SIGDAT Workshop on Very Large Corpora*, pages 14–27.

I. Duarte, M. Colaço, A. Gonçalves, A. Mendes, and M. Miguel. 2009. Lexical and syntactic properties of complex predicates of the type light verb + noun. *Arena Romanistica*, 4:48–57.

S. Evert. 2008. A lightweight and efficient tool for cleaning web pages. In *Proceedings of the 6th Interna-*

tional Conference on Language Resources and Evaluation (LREC2008), Marrakech, Morocco.

M. Généreux, A. Mendes, L. S. Santos Pereira, and M. Fernanda Bacelar do Nascimento. 2010. Lexical analysis of pre and post revolution discourse in Portugal. In *Proceedings of the 3rd workshop on building and using comparable corpora (LREC2010)*, La Valletta, Malta.

I. Hendrickx, A. Mendes, and S. Mencarelli. 2012. Modality in text: a proposal for corpus annotation. In *Proceedings of the eighth International Conferenceon Language Resources and Evaluation (LREC2012)*, Istanbul, Turkey. ELRA.

T. Joachims. 2002. *Learning to Classify Text Using Support Vector Machines*. Ph.D. thesis, Cornell University, USA. Kluwer Academic Publishers / Springer.

T. Kudo and Y. Matsumoto. 2003. Fast methods for kernel-based text analysis. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics*, ACL '03, pages 24–31, Stroudsburg, PA, USA. Association for Computational Linguistics.

A. Mendes and A. Estrela. 2008. Constructions with SE in african varieties of Portuguese. *Phrasis*, 2:83–107.

L.A. Ramshaw and M.P. Marcus. 1995. Text chunking using transformation-based learning. In *Proceedings of the Third Workshop on Very Large Corpora*, pages 82–94.

D. Santos and P. Rocha. 2001. Evaluating CETEMPublico, a Free Resource for Portuguese. In *Proceedings of 39th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 450–457, Toulouse, France, July. Association for Computational Linguistics.

D. Santos. 2011. Linguateca's infrastructure for portuguese and how it allows the detailed study of language varieties. *Oslo Studies in Language*, 3(2).

T. Berber Sardinha. 2007. History and compilation of a large register-diversied corpus of Portuguese at CEPRIL. *The Especialist*, 28(2):211–226.

A. Van den Bosch and W. Daelemans. 1999. Memory-based morphological analysis. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics, ACL'99*, pages 285–292.

C.J. van Rijsbergen. 1979. *Information Retrieval*. Buttersworth, London.