Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 849–854
Marseille, 11–16 May 2020

The *BDCamões* Collection of Portuguese Literary Documents: a Research Resource for Language Technology and Digital Humanities

Sara Grilo[1], Márcia Bolrinha[1], João Silva[1], Rui Vaz[2], António Branco[1]
[2]*Instituto Camões I.P.*
Av. da Liberdade 270, 1250-149 Lisboa, Portugal
rvaz@camoes.mne.pt
[1]*University of Lisbon*
NLX—Natural Language and Speech Group, Department of Informatics
Faculdade de Ciências
Campo Grande, 1749-016 Lisboa, Portugal
{srgrilo, msbolrinha, jsilva, antonio.branco}@di.fc.ul.pt

Abstract

This paper presents the BDCamões Collection of Portuguese Literary Documents, a new corpus of literary texts written in Portuguese that in its inaugural version includes close to 4 million words from over 200 complete documents from 83 authors in 14 genres, covering a time span from the 16th to the 21st century, and adhering to different orthographic conventions. Many of the texts in the corpus have also been automatically parsed with state-of-the-art language processing tools, forming the BDCamões Treebank subcorpus. This set of characteristics makes of BDCamões an invaluable resource for research in language technology (e.g. authorship detection, genre classification, etc.) and in language science and digital humanities (e.g. comparative literature, diachronic linguistics, etc.).

Keywords: Portuguese, corpus, language technology, digital humanities, literary studies, history, diachronic linguistics, cultural landmarks, cultural heritage.

1. Introduction

This paper introduces the BDCamões Collection of Portuguese Literary Documents,[1] a new language resource that is suitable for the scientific study and the technological preparation of the Portuguese language. This corpus is a collection of written documents with close to 4 million words from over 200 documents.

In contrast to many existing corpora aimed at primarily supporting the development and testing of natural language processing tools and applications, BDCamões has a number of less common, very interesting set of characteristics that together turn it into an invaluable research resource, complementing other related resources. It is made of complete documents, rather than of fragments; it is made of high quality, carefully edited texts, rather than of content that has been (semi-)automatically scrapped from web pages; it contains texts from a wide time span ranging from the 16th century to the 21st century, rather than only from recent times, or only from old times; it is made mostly of literary texts, rather than from the more usual, easily available domains of news articles, legal documents, etc.; it includes texts of different genres, such as novels, chronicles, poems and short stories, among others; it contains texts from a number of different authors, in different styles, rather than originating from a single author or adhering to a uniform style; its documents have well identified authors, rather than lacking clear authorship; many of its texts are outstanding landmarks of culture expressed in Portuguese language and/or are of high historical relevance (e.g. the first theater plays written in Portuguese) or from the greatest authors (e.g. Luís de Camões, Eça de Queirós, Fernando Pessoa, etc.); and last but not least, its texts adhere to a range of different orthographic traditions or standards that have been used for the Portuguese language. The BDCamões corpus is distributed by PORTULAN CLARIN, the Research Infrastructure for the Science and Technology of Language,[2] under the most permissive licenses possible given their different intellectual property rights.

Given its unique set of features, this corpus turns out to be a highly versatile language resource suitable for numerous research purposes for the science and technology of the Portuguese language. This is strengthened by the circumstance that, alongside the raw text versions, it includes also versions of many of its documents where these were automatically annotated and parsed with a wide range of linguistic information, including on part-of-speech, morphological features, grammatical dependencies and on expressions denoting named entities.

Focusing on language technology, the BDCamões corpus can be used to support the development of computational processing tools for authorship detection, genre classification, grammar checking, orthographic conversion, lexicon construction, etc., on a par of course with the more usual processing tools whose development is also supported by other types of corpora.

Focusing on language science, in turn, this corpus offers a great potential for the research in digital humanities. It will make viable the study of literary works and authors enhanced by computational technology solutions and thus under a new light that previous methods would not support. For instance, it allows for: the rapid development of

[1] The acronym BDCamões relates to the Portuguese designation *Coleção de Documentos Literários em Língua Portuguesa na Biblioteca Digital do Camões I.P.*

[2] `https://portulanclarin.net`

(sub-)vocabularies; accurate indexes of words and their occurrence in the context of specific works or authors; comparative studies on different literary schools, different authors or different creative periods of a given author; diachronic studies concerned with the evolution of the Portuguese language; among many others.

This paper is organized as follows. The next Section 2 provides an overview of some of the existing corpora for Portuguese that may be closer to BDCamões and in what way it differs from them. Section 3 describes in detail the content of the corpus and how it was gathered, while Section 4 informs about its distribution and licenses. Section 5 closes the paper with concluding remarks.

2. Related work

There are a few corpora for Portuguese that are suitable to support language research and the development of language technology. The BDCamões corpus complements them and opens new possibilities for research and innovation that were not so amply available due to its unique set of characteristics. We put it in contrast to some more relevant language resources with which it can be more closely compared with.

CIPM—Corpus Informatizado do Português Medieval (Xavier, 2016) is a corpus of 2,670 texts, totaling 2 million words, from the 12th to the 16th century, comprising several genres, such as historical narratives, religious texts and poetry. It addresses an earlier time span not in BDCamões but lacks coverage from the 16th century onward.

CTA—Corpus de Textos Antigos contains 29 historiographic texts as well as hagiographic, spiritual and novelistic texts originally written or translated into Portuguese until 1525.[3]

Tycho Brahe—Parsed Corpus of Historical Portuguese (Galves, 2018) is a corpus of texts written in Portuguese between the 15th and 19th centuries, with 76 texts from over 50 authors, comprising 3.3 million tokens, which only partly coincide with the texts in BDCamões (6 texts, with about 159,000 words).

LT Corpus—Corpus de Textos Literários (Généreux et al., 2012) is a literary corpus containing 70 documents published between the mid-19th century and the 1940's of the 20th century. While similar in design to and complementing BDCamões, it covers a shorter time span, has less variety of genres, fewer authors, and is smaller in size, with 1.8 million words, which only partly coincide with the texts in BDCamões (23 texts, with about 897,000 words).

CINTIL—Corpus Internacional do Português (Barreto et al., 2006) is a linguistically interpreted corpus with 1 million tokens, mostly from anonymized excerpts of news articles, but also including some works of fiction, and transcriptions of formal and informal speech. It is annotated with a variety of manually verified linguistic information, including morphological information and part-of-speech tags. Its texts are only from a recent period and it lacks some metadata items, such as information on the author, that would be necessary for some types of studies.

3. Corpus description

3.1. Documents gathering methodology

The books to which the documents in the BDCamões corpus relate to were collected and converted to their digital versions in PDF format under appropriate licensing in a dedicated acquisition and processing campaign in Camões I.P., the Portuguese official national organization, which acts under indirect administration of the Portuguese Ministry of Foreign Affairs, that is responsible for proposing and implementing Portuguese language and culture internationalization policies.

These documents, in their digital versions, were deposited in the Digital Library of Camões[4] — which gives the name to the present corpus —, from where they can be freely retrieved and used under the respective licensing conditions.

The documents were also further processed to produce the BDCamões corpus, which was deposited in PORTULAN CLARIN Research Infrastructure for the Science and Technology of Language,[5] belonging to the Portuguese National Roadmap of Research Infrastructures of Strategic Relevance, and part of the international research infrastructure CLARIN ERIC. The corpus can be freely retrieved from its repository and used under the licensing conditions indicated below in Section 4.

The PDF files were either produced from digital scans of the pages of the corresponding physical documents or exported to that format by the editors who allowed them to be published in BDCamões. While these files represent the visual aspect of the original documents, they cannot be processed as text by language processing tools.

The PDF documents were converted in the University of Lisbon into files in plain text format using the command line tool PDFTOTEXT,[6] which extracts any textual content found within a PDF file. This extraction was feasible because, when the PDF files were obtained from the scans of hard books, they underwent a process of optical character recognition (OCR) that secured a textual version of the content within the file.

The texts extracted from PDF files contained many errors, for instance, mistaking `l` (lowercase "L") for `I` (uppercase "i"), or `rn` for `m`, etc., which are indicative of failures in the preceding optical character recognition (OCR) process used to originate the PDF files. Since there is no safe heuristic to automatically detect and fix such cases, we performed an exhaustive manual revision of the converted plain text documents and the errors were corrected by two linguists, the first two authors of this paper, taking into account the input PDF version of the documents. Note that the manual correction only addressed the errors introduced by the conversion. The texts were otherwise transcribed literally, including eventual typographic errors present in the original edition.

The conversion to plain text is necessarily lossy in what

[3] http://beta.clul.ul.pt/teitok/cta/

[4] https://www.instituto-camoes.pt/en/activity-camoes/online-services/service-desk

[5] https://portulanclarin.net

[6] The PDFTOTEXT tool is part of the XPDF toolkit (http://www.xpdfreader.com).

concerns some aspects of formatting (e.g. font style) and layout (e.g. headers and footers). For BDCamões, page headers and page numbering were removed, while the tables of contents (if applicable) and footnote content were preserved. The content of a footnote is placed at the next available paragraph break after its reference so as not to break the sentence where the footnote is referred to.

The construction of the corpus is an ongoing work, and the texts included in this inaugural version are those whose conversion to digital version and subsequent curation has been already concluded.

3.2. Corpus statistics

The BDCamões corpus is composed of 208 documents and has a total of 3,945,943 words.[7]

There are 83 authors whose texts are included in the corpus (see Table 1 for the full list of authors). The number of documents and amount of words from each author vary. While most authors, 59 in all, have only one or two documents in the corpus, others are better represented. For instance, Trindade Coelho (1861–1908) has 18 documents in the corpus, making him the author with the largest number of documents, though not the one with the largest amount of words, as all his works in the corpus are short length tales. Júlio Dinis (1839–1871), in turn, is the author with the largest volume of texts, in terms of word count, with over 13% of the words in the corpus coming from his 5 works (4 novels and 1 tale).

Table 1: Amount of content per author

name	docs.	words
Agustina Bessa-Luís	7	378,522
Alexandre Herculano	8	173,851
Alfredo Margarido	1	9,646
Almeida Garrett	4	123,208
Amadeu Lopes Sabino	1	4,621
Antero de Quental	3	54,211
António Botto	1	2,770
António Feliciano de Castilho	1	5,385
António José da Silva	1	23,877
Aquilino Ribeiro	6	46,295
Armando Silva Carvalho	1	2,131
Augusto Abelaira	1	3,129
Bernardo Gomes Brito	1	8,871
Bernardo Santareno	1	8,247
Brito Camacho	1	4,980
Camilo Castelo Branco	7	177,012
Conde de Ficalho	2	5,521
D. Francisco Manuel de Melo	1	18,591
David Mourão-Ferreira	1	5,623
Eça de Queirós	10	273,011
Fernando Cabral Martins	2	1,798
Fernando Pessoa	1	5,154
Fernando Venâncio	1	2,855
Fernão Lopes	1	36,410
Fernão Mendes Pinto	2	19,004
Ferreira de Castro	1	4,347
Fialho D'Almeida	5	92,185
Francisco Maria Bordalo	1	13,395
Gil Vicente	6	21,068
Gonçalo M. Tavares	3	1,773
Hélia Correia	1	2,567
Jacinto Lucas Pires	1	2,895
Jaime Rocha	1	3,801
Jerónimo Osório de Castro	1	8,319
João Braz de Oliveira	1	5,318
João Vaz	1	8,964
Joaquim Canas Cardim	1	4,443
Joaquim Paço D'Arcos	1	12,521
Joaquim Pedro Celestino Soares	1	10,218
Jorge de Sena	5	37,684
José Cardoso Pires	1	6,447
José de Almada Negreiros	3	14,326
José Luandino Vieira	2	21,089
José Martins Garcia	1	6,946
José Régio	1	10,836
José Rodrigues Miguéis	2	17,934
Júlio Dantas	2	6,774
Júlio Dinis	5	528,249
Lídia Jorge	2	13,942
Luís de Camões	1	146,821
Luísa Costa Gomes	3	16,248
Luísa Dacosta	1	9,798
Manuel de Arriaga	1	21,686
Manuel Maria Barbosa du Bocage	7	19,622
Manuel Teixeira Gomes	5	26,160
Maria Gabriela Llansol	1	2,373
Maria Leonor Buescu	1	32,097
Maria Ondina Braga	1	4,927
Maria Teresa Horta	1	1,498
Maria Velho da Costa	1	1,020
Mário Cláudio	1	578
Mário de Carvalho	5	22,235
Mário de Sá-Carneiro	1	2,218
Mário Henrique Leiria	1	731
Maximiano Lemos Júnior	1	6,263
Nun'Álvares de Mendonça	1	17,568
Nuno Júdice	2	3,850
Oliveira Martins	3	334,693
Padre António Vieira	1	12,038
Pêro Vaz de Caminha	1	9,395
Ramalho Ortigão	6	239,252
Raul Brandão	3	69,207
Ruben A.	1	5,878
Rui de Pina	8	219,031
Sophia de Mello Breyner	1	6,711
Teófilo Braga	5	227,856
Teresa Veiga	1	8,056

(table continues)

[7] Here we consider "word" as any sequence of characters delimited by white space, and the count is obtained by the standard Linux command line tool WC.

Table 1 (continued)

name	docs.	words
Tomaz de Figueiredo	1	4,308
Tomaz Vieira da Cruz	1	4,224
Trindade Coelho	18	127,166
Venceslau de Moraes	2	43,776
Vergílio Ferreira	2	6,247
Vitorino Nemésio	4	41,648
total	208	3,945,943

The corpus covers written texts from several genres, such as tales, novels, chronicles, poems, dramas and essays, among others, as shown in greater detail in Table 2. Similarly to what happens regarding authorship, the proportion of documents and words for each genre varies. Tales are the most common genre in number of documents, providing for more than 44% of the texts in the corpus. As a result of their smaller size, they only account for 17% of the corpus in terms of words. The much longer novels, though being only 12% of the documents, account for over 32% of the words in the corpus.

Table 2: Genre distribution in the corpus

typology	docs.	words
tale	92	656,228
chronicle	26	600,018
novel	25	1,290,327
short story	21	295,724
poem	18	296,296
theater play	11	81,589
essay	8	534,515
travel guide	1	6,016
sermon	1	12,038
other	1	6,507
narrative	1	52,715
memoirs	1	17,568
letter	1	9,395
anthology	1	87,007
total	208	3,945,943

In terms of the time span represented, the corpus contains texts from the 16th century onward. Namely, 7 from the 16th century, 4 from the 17th century 8 from the 18th century, 84 from the 19th century, 82 from the 20th century and 23 from the 21st century. As such, this corpus represents different phases of the Portuguese language, including 13 texts from Middle Portuguese (up to early 16th century) or Classic Portuguese (until mid-18th century). The remaining text are in some form of Modern Portuguese (from the mid-18th century onward; or older but in an edition that has been transcribed into those orthographic norms): 21 are written according to the Portuguese orthographic norm of 1911, and 174 according to the norm of 1945.

The authors, genres and time periods are not equally represented in the collection, as the goal of the effort described here is to gather and transcribe the documents available in the Digital Library of Camões, making them available for various types of studies. Researchers interested in a particular set of authors, genre or time period will then be able to take the BDCamões corpus as a resource in which the relevant documents may be found.

```
<document>
  <header>
    <title> ... </title>
    <author> ... </author>
    <type> ... </type>
    <firstPublicationDate> ... </firstPublicationDate>
    <publisher> ... </publisher>
    <publicationDate> ... </publicationDate>
  </header>
  <text> ... </text>
  <annotation> ... </annotation>
</document>
```

Figure 1: XML structure of a document in BDCamões

3.3. Metadata and linguistic annotation

All the documents written in Modern Portuguese, or which are older but whose edition has been transcribed into that orthographic norm, have been automatically parsed with state-of-the-art language processing tools for Portuguese (Branco and Silva, 2006), and thus annotated with linguistic information that follows from the design of these tools and that can be found in detail in their guidelines and documentation (Branco et al., 2015).

This subcorpus forms the **BDCamões Treebank** with 4,495,379 tokens, of which 3,456,396 belong to the public domain part and 1,038,983 belong to restricted part of the corpus.

The resulting linguistic annotation comprises part-of-speech tags (e.g. PREP, ADV, etc.), morphology (lemmas for words from the open categories; gender and number for words from nominal categories; tense, aspect, person and number for verbs), named entities (in BIO notation), syntactic analysis in terms of graphs of grammatical dependencies (e.g. SJ, OBL, M, etc.), and semantic analysis in terms of semantic roles (e.g. ARG1, ARG2, LOC, etc.). A second version of the dependency graphs was obtained by converting them to the so called Universal Dependencies (de Marneffe et al., 2014).

Each document is stored in a separate file, associated with the metadata record in XML markup shown in Figure 1. The text itself and, when applicable, the corresponding linguistically annotated data appear, respectively, in the two fields `<text>` and `<annotation>`. The remaining fields, in the header, contain the title, author and type (genre) of the work, and information on its publication (the date for the first publication of the work, and the publisher and data of the publication for the edition that was transcribed).

The output of the annotation uses a CoNLL-style format, with one token per line and tab-separated fields. An excerpt of an annotated sentence may be seen in Figure 2. The Universal Dependencies version of the grammatical dependencies appear in a ante-penultimate column, parallel to and after the column with the original analysis.

```
O             O             _             DA    ms     O      SP          2 DET      2  R
nome          nome          NOME          CN    ms     O      SJ-ARG1     5 NSUBJ    5  LR
de            de            _             PREP  _      O      OBL-ARG1    2 CASE     4  LR
Ramalhete     Ramalhete     _             PNM   _      B-LOC  C           3 POBJ     2  LR
provinha      provinha      PROVIR        V     ii-3s  O      ROOT        0 ROOT     0  LR
decerto       decerto       _             ADV   _      O      M-LOC       5 ADVMOD   5  LR
de            de            _             PREP  _      O      C-ARG2      6 CASE     9  LR
um            um            _             UM    ms     O      SP          9 DET      9  LR
revestimento  revestimento  REVESTIMENTO  CN    ms     O      C           7 DEP      6  LR
quadrado      quadrado      QUADRADO      PPA   ms     O      M-PRED      9 AMOD     9  LR
de            de            _             PREP  _      O      OBL-ARG1   10 CASE    12  LR
azulejos      azulejos      AZULEJO       CN    mp     O      C          11 POBJ    10  LR
..............................................rest of the sentence omitted..............................................
```

Figure 2: Excerpt of an annotated sentence from a document in BDCamões. The 11 columns are as follows: (1) word form; (2) normalized word form (e.g. after expanding contracted forms); (3) lemma; (4) part-of-speech; (5) inflection; (6) named entity (BIO notation); (7–8) dependency relation and parent index; (9–10) dependency relation and parent index, in Universal Dependencies; and (11) spacing around the token (e.g. LR indicates the token had spaces to the left and to the right of it in the original sentence).

Table 3: Intellectual rights of the documens

availability	docs.	words
public domain	127	3,121,986
restricted	81	823,957
total	208	3,945,943

4. Corpus distribution

The BDCamões corpus is distributed by PORTULAN CLARIN, the Research Infrastruture for the Science and Technology of Language.[8] The distribution is split into two parts, whose usage is ruled by different licensing conditions, namely Part I includes the documents that are in the public domain and Part II includes the remaining documents (cf. Table 3). The annotated subcorpus BDCamões Treebank is part of the distribution of the BDCamões corpus, and additionally, for the sake of the convenience of its users, it is also distributed separately, split into two parts as well, by PORTULAN CLARIN.

The two parts of the BDCamões corpus are distributed separately, each having a different entry in the PORTULAN CLARIN repository. The two parts of BDCamões Treebank also have their own entries.

The two parts of the corpus are distributed under the most permissive license for each of them. Part I of BDCamões is distributed under the license CC-BY, which requires that when used, the academic authorship of this part of the corpus is acknowledged. Part II has the license CC-BY-NC-ND, which restricts it to research, non commercial usage and does not allow its redistribution.

The two parts of the treebank receive similar licenses.

5. Conclusion

In this paper we presented the inaugural version of the BDCamões corpus, a new language resource to support the science and technology of language. It is a collection of 208 complete documents, mostly literary, from 14 genres, by 83 authors, whose first publication dates cover a time span from the 16th to the 21th century, and totaling close to 4 million words. Its unique set of characteristics, namely being composed of complete documents instead of fragments, the carefully edited text, the variety of genres and authors, and the wide time span and orthographic conventions covered, make of it an invaluable and versatile resource for multiple research purposes, both for language technology and digital humanities.

In future work, we aim at publishing further versions of the BDCamões corpus, extending it with further documents, and from further variants of Portuguese.

Acknowledgements

The research reported here was partially supported by Camões I.P., and by PORTULAN CLARIN—Research Infrastructure for the Science and Technology of Language, funded by Lisboa 2020, Alentejo 2020 and FCT—Fundação para a Ciência e Tecnologia under the grant PINFRA/22117/2016.

Bibliographical References

Barreto, F., Branco, A., Ferreira, E., Mendes, A., Nascimento, M. F., Nunes, F., and Silva, J. (2006). Open resources and tools for the shallow processing of Portuguese. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC)*, pages 1438–1443.

Branco, A. and Silva, J. (2006). A suite of shallow processing tools for Portuguese: LX-Suite. In *Proceedings of the 11th European Chapter of the Association for Computational Linguistics (EACL)*, pages 179–182.

Branco, A., Silva, J., Querido, A., and de Carvalho, R. (2015). CINTIL DependencyBank PREMIUM handbook: Design options for the representation of grammatical dependencies. Technical Report DI-FCUL-TR-2015-05, University of Lisbon.

de Marneffe, M.-C., Dozat, T., Silveira, N., Haverinen, K., Ginter, F., Nivre, J., and Manning, C. (2014). Universal Stanford dependencies: A cross-linguistic typology. In *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC)*, pages 4585–4592.

Galves, C. (2018). The Tycho Brahe corpus of historical Portuguese. *Linguistic Variation*, 18(1):49–73.

[8] http://portulanclarin.net

Généreux, M., Hendrickx, I., and Mendes, A. (2012). A large Portuguese corpus on-line: cleaning and preprocessing. In *International Conference on Computational Processing of the Portuguese Language (PROPOR)*, pages 113–120.

Xavier, M. F. (2016). O CIPM – corpus informatizado do português medieval, fonte de um dicionário exaustivo. In Johannes Kabatek, editor, *Lingüística de corpus y lingüística histórica iberorrománica*, pages 137–156. De Gruyter.