João Silva*, Sara Grilo, Márcia Bolrinha, Rodrigo Santos, Luís Gomes, António Branco, and Rui Vaz

Where do I Belong in Six Centuries of Literature?

Datasets and AI-based tools for Portuguese literary documents made possible and available by PORTULAN CLARIN

Abstract: Enhancing the availability of corpora and processing tools for language research is a central endeavour of the CLARIN research infrastructure. In this chapter we report on how PORTULAN CLARIN, with the support of the national institute for the promotion of the Portuguese Language, Camões I.P., has contributed to this effort through the development of BDCamões. This is a collection of Portuguese literary documents suited to a variety of research purposes in the science and technology of the Portuguese language. This collection complements existing corpora by virtue of being composed of complete documents, from various genres and prominent authors, covering a wide time span, and offers an important potential for language science and for the development of language technology tools. This chapter also presents and discusses an exemplar case of the exploration of that potential where an automatic authorial style attribution system was developed by resorting to BDCamões.

Keywords: language resources, literary corpora, AI-based language processing tools, language technology, authorial style attribution, Portuguese language

1 Introduction

The oldest document known to have been written in Portuguese (Castro 2015) is "Notícia de Fiadores", a legal text dating back to the 12th century, or 1175 to be precise (Martins 1999). This is also the oldest document written in Portuguese of which a copy is distributed by PORTULAN CLARIN (Branco et al. 2020), and

***Corresponding author: João Silva, Sara Grilo, Márcia Bolrinha, Rodrigo Santos, Luís Gomes, António Branco,** PORTULAN CLARIN and University of Lisbon, Departamento de Informática, Faculdade de Ciências de Lisboa, Campo Grande, 1749-016 Lisboa, Portugal, jsilva@di.fc.ul.pt, srgrilo@fc.ul.pt, msbolrinha@fc.ul.pt, rsdsantos@fc.ul.pt, luis.gomes@di.fc.ul.pt, antonio.branco@di.fc.ul.pt
Rui Vaz, Camões I.P., Rua Rodrigues Sampaio, 113, 1150-279 Lisboa, Portugal, rvaz@camoes.mne.pt

which is included in the CIPM – Corpus Informatizado do Português Medieval (Xavier 2016), a corpus containing texts covering a period from the 12th to the 16th century. The oldest literary document written in Portuguese, in turn, of which a copy is also distributed by PORTULAN CLARIN, and which is included in the same corpus referred to above, is the compilation of love poems "Cantigas d'Amigo", dating back to the 13th century (Cohen 2003).

Legal rules and the rules of love. These are perennial themes for humankind and the two domains that the historical contingency capriciously happened to select for the first documents written in Portuguese that survived until the present day; and that PORTULAN CLARIN is now ensuring that can be read, enjoyed, studied, and preserved, under appropriate and necessary conditions, for future generations.

This aim definitely lies at the heart of the mission of the CLARIN research infrastructure. But CLARIN has been doing far more than just providing help – which is of the utmost importance – to the colleagues who have authored and collected the CIPM corpus, specifically by distributing this corpus through the CLARIN repository so that it can reach the largest possible number of users, readers, and researchers. By means of the Portuguese node PORTULAN CLARIN, the availability of Portuguese literary texts for research has been advanced in two other important directions.

On the one hand, PORTULAN CLARIN complemented the efforts already being made by the authors of the CIPM corpus. With the crucial support of Camões I.P., the national institute for the promotion of the Portuguese language, a new digital collection of literary documents, named BDCamões (Grilo et al. 2020), was developed, covering a historical period starting in the 16th century, precisely where the period covered by the CIPM corpus ended. In its inaugural version, BDCamões includes close to 4 million words from over 200 complete documents by 83 authors in 14 genres, covering a period from the 16th to the 21st century, and adhering to different orthographic conventions. Importantly, many of the texts in this corpus have also been automatically parsed with state-of-the-art language processing tools. This set of characteristics makes of the new BDCamões corpus an invaluable resource for research in language technology (e.g., authorship attribution, genre classification, etc.) and in language science and digital humanities (e.g., comparative literature, diachronic linguistics, etc.), which is now also being distributed by PORTULAN CLARIN.

On the other hand, on the basis of these corpora, and resorting to Artificial Intelligence techniques based on machine learning with artificial neural networks, PORTULAN CLARIN developed an innovative research instrument for the literary studies of Portuguese. This is an automatic authorial style classification tool that takes as its input an excerpt of text and delivers as its output the indication of the most probable literary writers, from among those represented in BDCamões, who

could have authored the input excerpt as a literary text. These achievements by PORTULAN are examples of how CLARIN can accomplish its mission and serve its users in advanced, unheard-of ways, and are examples of initiatives that can be replicated in other languages and literary corpora.

Our goal in this chapter for the volume celebrating the 10th anniversary of CLARIN is thus to expand on the initiatives and results referred to above by describing them in detail, to report on how PORTULAN CLARIN has been undertaking its mission, and to contribute to further spread and improve what CLARIN can do for its users and the advancement of research in the science and technology of language and in Digital Humanities.

The remainder of this chapter is structured as follows: Section 2 describes the BDCamões collection in more detail; Section 3 presents the experiment on authorial style attribution; and Section 4 concludes the chapter.

2 The BDCamões Collection

With close to 4 million words in its 208 documents by 83 authors, the BDCamões Collection of Portuguese Literary Documents possesses a number of characteristics that set it apart from the majority of existing corpora that are primarily aimed at supporting the development of natural language processing tools and applications, typically as training and testing data sets. These characteristics, which make BDCamões an invaluable research resource that complements other related resources (to be more extensively referred to below in Section 2.1), are the following:

- it is composed of complete documents, rather than of fragments or excerpts;
- the texts that form it are of high quality and have been edited carefully, rather than being content that has been automatically or semi-automatically scrapped from web pages;
- it covers a wide time span of six centuries, from the 16th century to the 21st century, rather than being circumscribed by a particular time period;
- it is composed mostly of literary texts, rather than from the more usual, more easily sourced domains of news articles, official documents, social media, legal documents, etc.;
- it includes texts from different genres, such as novels, chronicles, poems, and short stories, among others;
- it contains texts by a number of different authors, in different styles, rather than originating from a single author or adhering to a uniform style;
- its documents have positively identified authors, rather than lacking clear authorship;

- many of its texts are outstanding landmarks of culture expressed in the Portuguese language and/or are of particular historical significance (e.g., the first theatre plays written in Portuguese) or are written by great authors (e.g., Luís de Camões, Eça de Queirós, Fernando Pessoa, Agustina Bessa-Luís, etc.);
- and last but not least, its texts adhere to a range of different orthographic traditions or standards used in Portuguese, *de jure* or *de facto*.

The unique set of characteristics outlined above makes BDCamões a versatile and flexible language resource that is well-suited for a variety of research purposes in the science and technology of the Portuguese language. This is further strengthened by the fact that, alongside the raw text versions of the documents, BDCamões also includes linguistically annotated versions of many of the documents in the collection, with a wide range of linguistic information (cf. Section 2.4), including part-of-speech categories, morphological features, grammatical dependencies, and expressions denoting named entities.

Focusing on the language science applications, this corpus offers a great potential for research in the Digital Humanities and related fields. It makes viable the study of literary works and authors enhanced by computational technology solutions, and thus shows them in a new light that previous methods would not support. For instance, it allows for: the rapid development of (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 within the career of a given author; diachronic studies concerned with the evolution of the Portuguese language; and many other applications.

Focusing, in turn, on the language technology applications, BDCamões can be used to support the development of computational processing tools for authorship analysis, 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. A concrete example is presented in Section 3, using a case study in which an authorial style attribution system was quickly developed by utilising BDCamões.

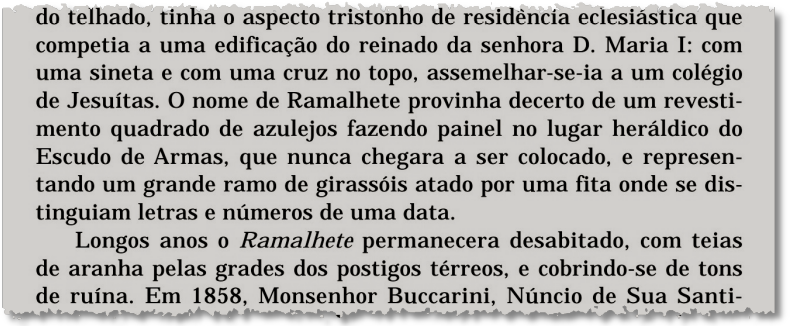
2.1 Related corpora

There already exist a few corpora for Portuguese that can be used to support language research and the development of language technology. In the remainder of this section, we contrast BDCamões with some of the more relevant language resources with which it can be closely compared.

- CIPM – Corpus Informatizado do Português Medieval (Xavier 2016) is a corpus of 2,670 texts, totalling 2 million words, from the 12th to the 16th century, comprising several genres, including historical narratives, religious texts, and poetry. It addresses an earlier time span not covered by 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 up to 1525.[1]
- Tycho Brahe – Parsed Corpus of Historical Portuguese (Galves 2018) is a corpus of texts written in Portuguese between the 14th and 19th centuries, with 76 texts from over 50 authors, comprising 3.3 million tokens, which only partly coincide with the texts in BDCamões (an overlap of 6 texts, totalling about 159,000 words). Subsets of Tycho Brahe have been annotated with part-of-speech tags (44 texts) and parsed (27 texts).
- LT Corpus – Corpus de Textos Literários (Généreux, Hendrickx, and Mendes 2012) is a literary corpus containing 70 documents published between the mid-19th century and the 1940s. While similar in design to and complementing BDCamões, it covers a shorter time span, has a smaller variety of genres, fewer authors, and is smaller in size, at about 1.8 million words, which only partly coincide with the texts in BDCamões (an overlap of 23 texts, totalling about 897,000 words).
- CINTIL – Corpus Internacional do Português (Barreto et al. 2006) is a linguistically interpreted corpus containing 1 million tokens, mostly from anonymised 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 all 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.

BDCamões, due to its unique characteristics already outlined above, complements these other corpora and opens up new possibilities for research and innovation that were not so amply available before.

1 http://teitok.clul.ul.pt/cta/


do telhado, tinha o aspecto tristonho de residência eclesiástica que competia a uma edificação do reinado da senhora D. Maria I: com uma sineta e com uma cruz no topo, assemelhar-se-ia a um colégio de Jesuítas. O nome de Ramalhete provinha decerto de um revestimento quadrado de azulejos fazendo painel no lugar heráldico do Escudo de Armas, que nunca chegara a ser colocado, e representando um grande ramo de girassóis atado por uma fita onde se distinguiam letras e números de uma data.

Longos anos o *Ramalhete* permanecera desabitado, com teias de aranha pelas grades dos postigos térreos, e cobrindo-se de tons de ruína. Em 1858, Monsenhor Buccarini, Núncio de Sua Santi-

Figure 1: Snippet of a PDF page from "Os Maias" (background darkened for contrast)

2.2 Document gathering

The digital documents that form the BDCamões collection evolved from a set of works collected by Camões I.P., the official national organisation, acting under the indirect administration of the Portuguese Ministry of Foreign Affairs, responsible for promoting the Portuguese language abroad.

The collection campaign undertaken by Camões I.P. covered the conversion of the works into their digital versions in PDF format under appropriate licensing. These documents were deposited in the Digital Library of Camões I.P.[2] – which gives the name to the collection – from where they can be freely retrieved and used under their respective licensing conditions.

The PDF files were either provided in that format already by the editors of the works or produced from digital scans of the pages of the corresponding physical documents. In either case, while the files represent the visual aspect of the original documents (see Figure 1), they cannot be processed as text by language processing tools.

To allow the document to be processed by language processing tools, they were converted by PORTULAN CLARIN into files in plain text format using the command line tool PDFTOTEXT,[3] which extracts any textual content found within a PDF file. This extraction was feasible in the case of the PDF files that were obtained from scanning physical documents because these underwent a process of optical character recognition (OCR) that secured a textual version of the content within the PDF file.

2 https://www.instituto-camoes.pt/en/activity-camoes/online-services/service-desk

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

As is to be expected, the OCR process introduces some errors in the transcription, especially for those documents that use uncommon fonts, adhere to old typographic norms, or whose digital scan is of poor quality to begin with. Examples of typical OCR failures are mistaking `l` (lowercase "L") for `I` (uppercase "i"), mistaking `rn` for `m`, and the transcription of typographic ligatures.

There is no safe heuristic to automatically detect and fix such cases. As such, we performed an exhaustive manual revision of the converted plain text documents and the errors were manually corrected by linguists, taking into account the source PDF version of the documents. Note that the manual correction only addressed the errors introduced by the OCR process. The texts were otherwise transcribed literally, including eventual orthographic errors present in the original edition.

The conversion to plain text is necessarily lossy with regard to some aspects of formatting (e.g., font style, such as italics), hyphenation, and page layout (e.g., headers and footers). For BDCamões, hyphenation was reverted, page headers and page numbering were removed, while the tables of contents (if applicable) and footnote content were preserved. For footnotes, their content is placed at the next available paragraph break after the cue so as not to break the sentence where the footnote is introduced.

2.3 Corpus composition

The construction of the corpus is an ongoing work, and the texts included in the collection are those whose conversion to digital version and subsequent curation has already been concluded. In its current version, the BDCamões corpus is composed of 208 documents and has a total of 3,945,943 words.[4]

There are 83 authors represented in the corpus, with a varying number of documents and amount of words from each (see Table 1 for a summary). While a majority of authors – 59 in all – only have one or two documents in the corpus, others are represented more prominently. For instance, Trindade Coelho (1861–1908) has 18 documents in the corpus, making of 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 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).

The corpus covers written texts from several genres, such as tales, novels, chronicles, poems, dramas, and essays, among others, as shown in greater detail

4 Here we consider "word" to be any sequence of characters delimited by white space, and the count is obtained by the standard Linux command line tool `wc`.

Table 1: Amount of content per author

Name	Docs	Words	Name	Docs	Words
Agustina Bessa-Luís	7	378,522	José Luandino Vieira	2	21,089
Alexandre Herculano	8	173,851	José Martins Garcia	1	6,946
Alfredo Margarido	1	9,646	José Régio	1	10,836
Almeida Garrett	4	123,208	José Rodrigues Miguéis	2	17,934
Amadeu Lopes Sabino	1	4,621	Júlio Dantas	2	6,774
Antero de Quental	3	54,211	Júlio Dinis	5	528,249
António Botto	1	2,770	Lídia Jorge	2	13,942
A. Feliciano de Castilho	1	5,385	Luís de Camões	1	146,821
António José da Silva	1	23,877	Luísa Costa Gomes	3	16,248
Aquilino Ribeiro	6	46,295	Luísa Dacosta	1	9,798
Armando Silva Carvalho	1	2,131	Manuel de Arriaga	1	21,686
Augusto Abelaira	1	3,129	M.M. Barbosa du Bocage	7	19,622
Bernardo Gomes Brito	1	8,871	Manuel Teixeira Gomes	5	26,160
Bernardo Santareno	1	8,247	Maria Gabriela Llansol	1	2,373
Brito Camacho	1	4,980	Maria Leonor Buescu	1	32,097
Camilo Castelo Branco	7	177,012	Maria Ondina Braga	1	4,927
Conde de Ficalho	2	5,521	Maria Teresa Horta	1	1,498
Dom F. Manuel de Melo	1	18,591	Maria Velho da Costa	1	1,020
David Mourão-Ferreira	1	5,623	Mário Cláudio	1	578
Eça de Queirós	10	273,011	Mário de Carvalho	5	22,235
Fernando Cabral Martins	2	1,798	Mário de Sá-Carneiro	1	2,218
Fernando Pessoa	1	5,154	Mário Henrique Leiria	1	731
Fernando Venâncio	1	2,855	M. Lemos Júnior	1	6,263
Fernão Lopes	1	36,410	Nun'Álvares Mendonça	1	17,568
Fernão Mendes Pinto	2	19,004	Nuno Júdice	2	3,850
Ferreira de Castro	1	4,347	Oliveira Martins	3	334,693
Fialho D'Almeida	5	92,185	Padre António Vieira	1	12,038
Francisco Maria Bordalo	1	13,395	Pêro Vaz de Caminha	1	9,395
Gil Vicente	6	21,068	Ramalho Ortigão	6	239,252
Gonçalo M. Tavares	3	1,773	Raul Brandão	3	69,207
Hélia Correia	1	2,567	Ruben A.	1	5,878
Jacinto Lucas Pires	1	2,895	Rui de Pina	8	219,031
Jaime Rocha	1	3,801	Sophia de Mello Breyner	1	6,711
J. Osório de Castro	1	8,319	Teófilo Braga	5	227,856
João Braz de Oliveira	1	5,318	Teresa Veiga	1	8,056
João Vaz	1	8,964	Tomaz de Figueiredo	1	4,308
Joaquim Canas Cardim	1	4,443	Tomaz Vieira da Cruz	1	4,224
Joaquim Paço D'Arcos	1	12,521	Trindade Coelho	18	127,166
J.P. Celestino Soares	1	10,218	Venceslau de Moraes	2	43,776
Jorge de Sena	5	37,684	Vergílio Ferreira	2	6,247
José Cardoso Pires	1	6,447	Vitorino Nemésio	4	41,648
José Almada Negreiros	3	14,326			

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 Table 2. Much as we saw regarding authorship, the proportion of documents and words for each genre varies. Tales are the most common genre in terms of the number of documents, accounting for more than 44% of the texts in the corpus, though they only account for 17% of the corpus in terms of words, due to their small size. The much longer novels, though making up only 12% of the documents, account for over 32% of the words in the corpus.

In terms of the time span represented, the corpus contains texts from the 16th century to the present day, 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 the early 16th century) or Classical Portuguese (until the mid-18th century). The remaining texts 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 norm): 21 are written according to the Portuguese orthographic norm of 1911, and 174 according to the norm of 1945.

The various authors, genres, and time periods are not equally represented in the collection, as the goal of BDCamões is to gather and transcribe the documents available in the Digital Library of Camões I.P., 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.

2.4 Linguistic annotation and metadata

To broaden the possible uses of BDCamões, a linguistically annotated version of the documents is made available separate from the plain text version. The annotation was automatically obtained using state-of-the-art language processing tools for Portuguese (Branco and Silva 2006). These tools have been developed with Modern Portuguese in mind and, accordingly, the annotation was done only for the subset of documents that were originally written in Modern Portuguese, or which are older but whose edition has been transcribed into that orthographic norm. This annotated subcorpus, BDCamões DependencyBank, contains 195 documents and a total of 4,495,379 tokens.[5]

The resulting linguistic annotation comprises part-of-speech tags (e.g., PREP, ADV, etc.), morphological and inflectional information (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, and annotated with their type), 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.). The dependency annotation follows the linguistic principles presented in (Branco et al. 2015). Additionally, given the popularity of the so-called Universal Dependencies (de Marneffe et al. 2014) format, BDCamões also provides a second version of the dependency graphs obtained by converting them from their original scheme to Universal Dependencies.

The annotation follows a CoNLL-style format, with one token per line and its linguistic annotation over several tab-separated fields. An excerpt of an annotated sentence may be seen in Figure 2. The 11 columns represent, as follows: (1) raw word form; (2) normalised word form (e.g., after expanding contracted forms); (3) lemma; (4) part-of-speech; (5) morphology and inflection; (6) named entity (BIO notation, with type); (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).

Each document is stored in a separate file, associated with the metadata record in XML markup shown in Figure 3. The text itself and, when applicable, the corresponding linguistically annotated data appear in the fields `<text>` and `<annotation>`, respectively. The remaining fields in the header contain the title, author, and type (genre) of the work, and information on its publication (the date

5 The token count is done after tokenisation, a process that expands contracted forms into multiple tokens and detaches punctuation symbols. As such, the number of tokens far exceeds the number of words.

```
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 BDCamões document

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

Figure 3: XML structure of a document in BDCamões

for the first publication of the work, and the publisher and date of publication for the edition that was transcribed).

2.5 Licensing and distribution

The BDCamões corpus is distributed by the PORTULAN CLARIN Research Infrastructure for the Science and Technology of Language.[6] Due to differences in the licensing conditions regarding document usage, the distribution is split into two parts (cf. Table 3), namely Part I, which includes the documents that are in the public domain; and Part II, which includes the remaining documents. The annotated sub-corpus BDCamões DependencyBank is part of the distribution of the BDCamões corpus, and additionally, for the convenience of its users, it is also distributed separately, again split into two parts, by PORTULAN CLARIN. The URL handles for these various parts are listed in Table 4.

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,

6 http://portulanclarin.net

Table 3: Availability of the documents in BDCamões

Availability	Docs	Words
public domain (Part I)	127	3,121,986
restricted (Part II)	81	823,957
total	208	3,945,943

Table 4: URL handles for the parts of BDCamões

BDCamões sub-corpus	Location
plain text - Part I	https://hdl.handle.net/21.11129/0000-000D-F89B-D
plain text - Part II	https://hdl.handle.net/21.11129/0000-000D-F8AB-B
DependencyBank - Part I	https://hdl.handle.net/21.11129/0000-000D-F8AA-C
DependencyBank - Part II	https://hdl.handle.net/21.11129/0000-000D-F8A8-E

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 is restricted to research, non-commercial usage, and does not allow the material to be redistributed. The corresponding two parts of the annotated corpus have similar licenses.

3 An experiment on automatic authorial style attribution

BDCamões can be used to support a wide range of research in language technology applications. In this section, we present an experiment where we developed systems for automatic authorial style attribution,[7] with implementations at different levels of complexity and performance, by resorting to BDCamões and off-the-shelf software. All the systems run over plain text and do not require any kind of linguistic annotation.

Note that the classifiers learn to assign authors to texts given by users, but it would be inaccurate to say that they are performing authorship attribution, as the authors in BDCamões have not, strictly speaking, authored the user-provided texts,

7 We have also experimented with assigning an historical period (century) to texts. Apart from the class that is to be learned, nothing else changes in the setup of the experiment, so the system descriptions that follow refer only to assigning authorial style. Results for both tasks are given in Section 3.3.

unless the user inputs a text from one of those authors. We thus frame the task as authorial style attribution, that is, classifying the given text as being written in the style of a certain author.

3.1 Baseline classifier

For the baseline classifier, we aimed for a system that should be simple to implement, presenting a low barrier to entry for people who may not be well versed in natural language processing or programming, but still achieving competitive performance.

The implementation was done by using scikit-learn (Pedregosa et al. 2011), a Python package for machine learning that strives for accessibility and ease of use. The package comes with functionality for text processing, which makes it straightforward to apply to text-based tasks.

The features used to represent a document are extremely simple and rely solely on the raw text. A document is represented by a bag of character n-grams for all n in the 2–5 range. That is, a vector with the number of occurrences of every sequence of characters of length 2–5. Such counts would, of course, be large for n-grams that are very frequent throughout the corpus and thus not very helpful in terms of discriminating between authors. As such, the values are normalised by the commonly used tf-idf weighting technique, which downplays n-grams that occur over many documents and gives greater importance to n-grams that are distinctive to a few documents. All these functionalities, i.e., the feature extraction and tf-idf weighting, are provided by scikit-learn.

The classification algorithm is a support vector machine (SVM) with a linear kernel. These are effective even in high dimensional spaces, as in this case,[8] and when there are comparatively few samples. The SVM classifier is provided by scikit-learn. It handles the fact that the task is one with multiple classes, despite the SVM being a binary classifier, by automatically recasting the task as multiple one-vs-rest binary classifications.[9]

8 The feature space is the set of character n-grams for n in the 2–5 range which, for the training set being used (described in Section 3.3), amounts to over 356,000 features.

9 In one-vs-rest, for each author A there will be a binary classifier that only outputs whether a given text has been authored by A, with some certainty score. The assigned author is that of the classifier whose prediction has the greatest certainty.

The feature extraction, training, and evaluation required about a dozen lines of Python code, very similar to those from the "Working with text data" scikit-learn tutorial.[10]

3.2 Neural classifier

Deep neural models have come to the fore as their performance steadily advanced the state of the art in a variety of machine learning tasks and applications. In NLP in particular, the Transformer encoder-decoder architecture of Vaswani et al. (2017) has become a dominant paradigm and the basis for whole families of system architectures.

An encoder-decoder architecture is composed of two parts: the encoder, which maps the input into a compact representation, and the decoder, which takes that compact representation and produces the output. A typical example is found in machine translation, where the encoder maps text in the source language into a compact representation of its meaning and the decoder produces the text in the target language from that representation.

The Transformer makes extensive use of the so-called attention mechanism, which circumvents the requirement to pack the whole input into a single representation – a major bottleneck for previous systems – by allowing the decoder to access (or "pay attention to") the representations of the individual tokens being processed by the encoder.

The descriptions given above of the encoder-decoder architecture and the Transformer are overly simplistic and leave out several details, since an in-depth explanation would be outside the scope of this chapter. We direct the interested reader to (Vaswani et al. 2017).

We have experimented with two neural models, each from a different family, though both are ultimately based on the Transformer. One model is from the BERT (Devlin et al. 2019) family of architectures that take only the encoder part of the Transformer, and the other from the GPT (Radford et al. 2018) family of architectures that take only the decoder part. Both have in common that they are first pre-trained over a large amount of raw text, building up a task-agnostic language model which is then extended with an additional layer, the classification head, and fine-tuned on labelled data for the task at hand.

10 https://scikit-learn.org/stable/tutorial/text_analytics/working_with_text_data.html

3.2.1 BERT-style

The BERT-style architectures use only an encoder and are pre-trained using some sort of input reconstruction task. For instance, the encoder is given an input sentence where a random token has been masked, and the encoder has to predict what that token is.

The BERT neural model we experiment with is RoBERTa (Liu et al. 2019), a BERT architecture with small adjustments that make it more robust. The model has a vocabulary size of 32,000 subwords,[11] 6 layers and 12 attention heads, for a total of 67.5 million parameters.

The RoBERTa model was pre-trained on a data set of 20 million tokens, 10 million in Portuguese and 10 million in English, from the Oscar corpus (Ortiz Suárez, Sagot, and Romary 2019), an automatically filtered and cleaned subset of the huge (multiple terabytes) Common Crawl corpus. The fact that English text is included in the pre-training of the model may be surprising, given that the classifier is to be used for Portuguese texts only, but similar choices are found in the literature, since the additional pre-training data, even if in a different language, can lead to better performance.[12] For this experiment, we found that adding English pre-training data does indeed help.

After the pre-training phase is finished, the model is fine-tuned on the authorial style attribution task. For this, an extra layer, the classification head, is added to the model. This is a fully connected layer that takes the output of the RoBERTa language model and outputs the author. The weights of this layer, and of the underlying RoBERTa language model, are adjusted during fine-tuning.

3.2.2 GPT-style

The GPT-style architectures use only a decoder and are pre-trained using a language modelling task which typically consists of, given a span of tokens, predicting the token that is most likely to follow.

11 The vocabulary of modern neural architectures is not strictly composed by words. It is instead formed by subwords, which are strings from which words are formed. In this work, a method called byte-pair encoding (Sennrich, Haddow, and Birch 2016) is used.

12 This is likely to hold only if there is a large enough amount of data in the language used to extend the pre-training corpus (English thus being the common choice) and if the languages are not very different from each other.

The GPT model we experiment with is GPorTuguese-2.[13] As before, both English and Portuguese texts have been included in the pre-training of the model. The authors took the GPT-2 small model[14] and performed additional pre-training on 1 GB of the Portuguese Wikipedia. It has a vocabulary of 50,257 subwords,[15] 12 layers, and 12 attention heads, for a total of 124.4 million parameters.

Fine-tuning on the authorial style attribution task is done in a similar way to that used on the previous model. The model is extended with an extra layer, the classification head, which is a fully connected layer that takes the output of the language model and outputs the author. The weights of this layer, and of the underlying GPorTuguese-2 language model, are adjusted during fine-tuning.

3.3 Experimental results

The training and testing data set splits are formed by taking, from each document, a randomly chosen 90% of the lines for training and the remaining 10% for testing. Thus, all documents are represented in the training set and in the testing set, in a proportion roughly matching their proportion in the full corpus.[16]

Assigning authorial style and assigning time period (century) are run as separate experiments. For authorial style classification each document is associated with its author (83 classes), and for time period classification each document is associated with the century of its publication (6 classes).

The baseline classifier works at the document level. Each training instance is composed by 90% of the lines of the original document and each testing instance by the remaining 10%. The architectures of the neural classifiers limit the length of the input to 250 words for RoBERTa and 500 words for GPorTuguese-2. As such, instead of working at the document level, the neural classifiers work at the level of the lines in the document. Each line in the training data set, up to the cutoff length, is a training instance. For testing, only the first words in the test instance, up to the cutoff length, are used for classification.

13 https://huggingface.co/pierreguillou/gpt2-small-portuguese

14 GPT-2 (Radford et al. 2019) is the successor to GPT. Much larger than its predecessor, it has 1.5 billion parameters and was pre-trained on 8 million web pages. In this work a much reduced version of it, called GPT-2 small, is used.

15 Like with RoBERTa, byte-pair encoding is used for the vocabulary.

16 Splitting by lines should approximate splitting by words, is easier, and ensures that sentences are not cut short. We have not experimented with balancing the data set as it would require either greatly under-sampling the common classes or greatly over-sampling the rare classes.

Table 5: Experimental results

(a) assigning authorial style

System	Accuracy	Macro-$\mathbf{F_1}$
baseline	0.7500	0.5367
RoBERTa	0.8448	0.7346
GPorTuguese-2	0.9036	0.8505

(b) assigning century

System	Accuracy	Macro-$\mathbf{F_1}$
baseline	0.7644	0.6827
RoBERTa	0.8803	0.8525
GPorTuguese-2	0.8883	0.8370

The neural models involve a certain amount of randomness in the process, such as in the initialisation of the weights in the network. To smooth out the variations caused by this, the results for RoBERTa and GPorTuguese-2 are the average of three runs.

Table 5 summarises the results, showing the accuracy and macro-F_1 score[17] of each system on the two tasks. Both neural models outperform the SVM baseline by a large margin and GPorTuguese-2, the larger model, outperforms RoBERTa. This is in line with what has been commonly reported in the literature for various tasks, where deep neural approaches outperform other techniques and where, as long as there is enough data, larger models perform better than smaller models.

Note that GPorTuguese-2 falls behind RoBERTa in terms of macro-F_1 but not in accuracy, for the task of assigning century. A plausible explanation is that GPorTuguese-2 is over-fitting the data, tending more heavily towards the most common classes (the 19th and 20th centuries), a choice that can lead to an inflated accuracy but is penalised by the F_1 metric.

A remark on computation time.

The neural models clearly outperform the baseline. It is worth noting, though, that the amount of compute they employ is orders of magnitude higher. Each fine-tuning run of RoBERTa takes around 3 hours, while each fine-tuning run of GPorTuguese-2 takes close to 6 hours, and working with these architectures is only feasible with GPU hardware support.[18] A training run of the baseline takes only 2 minutes when assigning authorial style, and 30 seconds when assigning century, and does not require a GPU.

17 The F_1 is the harmonic mean of precision and recall. Macro-F_1 means that F_1 is calculated for each class and the results averaged, giving equal importance to each class.
18 We used a single NVidia GeForce RTX 2080 with 12 GB.

4 Conclusion

This chapter presented how PORTULAN CLARIN, with the support of Camões I.P., has contributed to enhancing the availability of Portuguese literary corpora for research by developing BDCamões, a novel corpus of complete literary texts, from various genres and authors, covering a wide time span.

As mentioned in Section 2.3, the construction of the corpus is an ongoing work and the collection will keep growing as Camões I.P. gathers more texts and converts them into their digital versions.

To showcase an application of BDCamões in the development of language technology tools, we also presented an experiment in authorial style attribution where several systems, at different levels of complexity and performance, were quickly built by using this corpus and off-the-shelf software. This experiment in authorial style attribution was partly intended as an inspiring example of an application of BDCamões in the development of language technology tools.

The GPT-based version of this tool is being integrated into the PORTULAN CLARIN Workbench[19] as the LX-AuthorialStyle online service[20] – see Figure 4 for a screenshot of the current in-development interface. This tool joins a range of other language processing services that PORTULAN CLARIN makes available to its users, as detailed in Chapter 13 (Gomes et al. 2022) of this volume.

We plan to extend the experiments presented here, on the task of authorial style attribution, to the different task of authorship verification, by means of which two given texts are checked to ascertain whether they have been authored by the same person, and is not restricted to a pre-defined set of authors. While the task of authorship verification has an important application for Linguistic Forensics, we expect that the current functionality of authorial style attribution now presented, based on a set of well-known, prominent historic Portuguese literary authors, may also be interesting for the general public (e.g., "write a text and find which author you are more similar to", "complete a given text according to the style of an given author", etc.) and contribute to demonstrating the role of the research infrastructure for the advancement of the science and technology of language.

19 The PORTULAN CLARIN workbench consists of a number of language processing services based on a large body of research work contributed by different authors and teams, which continues to grow and is acknowledged here: Barreto et al. (2006); Branco et al. (2010); Cruz, Rocha, and Cardoso (2018); Veiga, Candeias, and Perdigão (2011); Branco and Henriques (2003); Branco et al. (2011); Branco and Nunes (2012); Silva et al. (2009); Branco et al. (2014); Rodrigues et al. (2016); Branco and Silva (2006); Rodrigues et al. (2020); Costa and Branco (2012); Santos et al. (2019); Miranda et al. (2011).

20 https://portulanclarin.net/workbench/lx-authorialstyle/

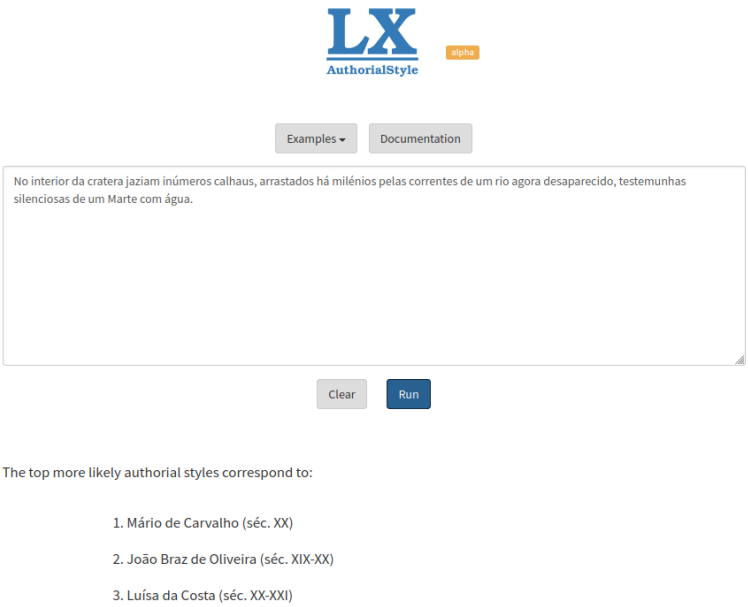


Figure 4: Screenshot of LX-AuthorialStyle service in the PORTULAN CLARIN Workbench

Acknowledgment: This work was done in collaboration with Camões I.P.

Funding: This work was partially supported 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.

Bibliography

Barreto, Florbela, António Branco, Eduardo Ferreira, Amália Mendes, Maria Fernanda Nascimento, Filipe Nunes & João Silva. 2006. Open resources and tools for the shallow processing of Portuguese: The TagShare project. *In Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC)*, 1438–1443.

Branco, António, Sérgio Castro, João Silva & Francisco Costa. 2011. CINTIL DepBank handbook: Design options for the representation of grammatical dependencies. Technical Report DI-FCUL-TR-2011-03, University of Lisbon.

Branco, António, Francisco Costa, João Silva, Sara Silveira, Sérgio Castro, Mariana Avelãs, Clara Pinto & João Graça. 2010. Developing a deep linguistic databank supporting a collection of treebanks: the CINTIL DeepGramBank. *In Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC)*, 1810–1815.

Branco, António & Filipe Nunes. 2012. Verb analysis in a highly inflective language with an MFF algorithm. *In Proceedings of the 11th International Conference on the Computational Processing of Portuguese (PROPOR)*, Lecture Notes in Artificial Intelligence no. 7243, 1–11. Springer.

Branco, António, João Rodrigues, João Silva, Francisco Costa & Rui Vaz. 2014. Assessing automatic text classification for interactive language learning. *In Proceedings of the IEEE International Conference on Information Society (iSociety)*, 72–80.

Branco, António & Tiago Henriques. 2003. Aspects of verbal inflection and lemmatization: Generalizations and algorithms. *In Proceedings of XVIII Annual Meeting of the Portuguese Association of Linguistics (APL)*, 201–210.

Branco, António, Amália Mendes, Paulo Quaresma, Luís Gomes, João Silva & Andrea Teixeira. 2020. Infrastructure for the science and technology of language PORTULAN CLARIN. *In Proceedings of the 1st International Workshop on Language Technology Platforms (IWLTP 2020)*, 1–7.

Branco, António & João Silva. 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)*, 179–182.

Branco, António, João Silva, Andreia Querido & Rita de Carvalho. 2015. CINTIL DependencyBank PREMIUM handbook: Design options for the representation of grammatical dependencies. Technical Report DI-FCUL-TR-2015-05, University of Lisbon.

Castro, Ivo. 2015. Formação da língua portuguesa. In Eduardo Raposo, Fernanda Bacelar, Antónia Mota, Luísa Segura & Amália Mendes (eds.), *Gramática do português*, 7–13. Fundação Calouste Gulbenkian.

Cohen, Rip. 2003. *500 cantigas d'amigo: Edição crítica / critical edition*. Campo das Letras.

Costa, Francisco & António Branco. 2012. Aspectual type and temporal relation classification. *In Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, 266–275.

Cruz, André Ferreira, Gil Rocha & Henrique Lopes Cardoso. 2018. Exploring Spanish corpora for Portuguese coreference resolution. *Fifth International Conference on Social Networks Analysis, Management and Security (SNAMS)*, 290–295.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee & Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. *In Proceedings*

of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL), 4171–4186.

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

Gomes, Luís, Ruben Branco, João Silva & António Branco. 2022. Open and inclusive language processing: Language processing services by PORTULAN to meet the widest needs of CLARIN users. In Darja Fišer & Andreas Witt (eds.), *CLARIN. The Infrastructure for Language Resources*. Berlin: deGruyter.

Grilo, Sara, Márcia Bolrinha, João Silva, Rui Vaz & António Branco. 2020. The BDCamões collection of Portuguese literary documents: a research resource for Digital Humanities and Language Technology. *In Proceedings of the 12th International Conference on Language Resources and Evaluation (LREC)*, 849–854.

Généreux, Michel, Iris Hendrickx & Amália Mendes. 2012. A large Portuguese corpus online: cleaning and preprocessing. *In Proceedings of the International Conference on Computational Processing of the Portuguese Language (PROPOR)*, 113–120.

Liu, Yinhan, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer & Veselin Stoyanov. 2019. RoBERTa: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692*.

de Marneffe, Marie-Catherine, Timothy Dozat, Natalia Silveira, Katri Haverinen, Filip Ginter, Joakim Nivre & Christopher Manning. 2014. Universal Stanford dependencies: A cross-linguistic typology. *In Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC)*, 4585–4592.

Martins, Ana Maria. 1999. Ainda "os mais antigos textos escritos em português": Documentos de 1175 a 1252. In Isabel Hub Faria (ed.), *Lindley Cintra: Homenagem ao Homem, ao Mestre e ao Cidadão*, 491–534. Cosmos.

Miranda, Nuno, Ricardo Raminhos, Pedro Seabra, Joao Sequeira, Teresa Gonçalves & Paulo Quaresma. 2011. Named entity recognition using machine learning techniques. *In Proceedings of the 15th Portuguese Conference on Artificial Intelligence (EPIA)*, 818–831.

Ortiz Suárez, Pedro Javier, Benoî Sagot & Laurent Romary. 2019. Asynchronous pipeline for processing huge corpora on medium to low resource infrastructures. *In Proceedings of the 7th Workshop on the Challenges in the Management of Large Corpora (CMLC-7)*.

Pedregosa, Fabian, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot & Édouard Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research* 12: 2825–2830.

Radford, Alec, Karthik Narasimhan, Tim Salimans & Ilya Sutskever. 2018. Improving language understanding by generative pre-training. OpenAI blog. https://openai.com/blog/language-unsupervised/.

Radford, Alec, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei & Ilya Sutskever. 2019. Language models are unsupervised multitask learners. OpenAI blog. https://openai.com/blog/better-language-models/.

Rodrigues, João, Francisco Costa, João Silva & António Branco. 2020. Automatic syllabification of Portuguese. *Revista da Associação Portuguesa de Linguística*, vol. 1.

Rodrigues, João, António Branco, Steven Neale & João Silva. 2016. LX-DSemVectors: Distributional semantics models for the Portuguese language. *In Proceedings of the*

12th International Conference on the Computational Processing of Portuguese (PROPOR'16), 259–270.

Santos, Rodrigo, João Silva, António Branco & Deyi Xiong. 2019. The direct path may not be the best: Portuguese-Chinese neural machine translation. *In Proceedings of the 19th Portuguese Conference on Artificial Intelligence (EPIA)*, 757–768.

Sennrich, Rico, Barry Haddow & Alexandra Birch. 2016. Neural machine translation of rare words with subword units. *In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL)*, 1715–1725.

Silva, João, António Branco, Sérgio Castro & Ruben Reis. 2009. Out-of-the-box robust parsing of Portuguese. *In Proceedings of the International Conference on the Computational Processing of Portuguese (PROPOR)*, 75–85.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aida Gomez, Łukasz Kaiser & Illia Polosukhin. 2017. Attention is all you need. *Advances in Neural Information Processing Systems*, 5998–6008.

Veiga, Arlindo, Sara Candeias & Fernando Perdigão. 2011. Generating a pronunciation dictionary for European Portuguese using a joint-sequence model with embedded stress assignment. *In Proceedings of the 8th Brazilian Symposium in Information and Human Language Technology*.

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