PressMint-PT — Compiling a Portuguese Historical Newspaper Corpus

José Aires, Amália Mendes
University of Lisbon, School of Arts and Humanities, Centre of Linguistics
Alameda da Universidade, 1600-214 Lisboa, Portugal
jagc@edu.ulisboa.pt, mendes@edu.ulisboa.pt

Abstract

We present a new European Portuguese corpus of newspapers from the 19th and early 20th centuries, integrated in the recent PressMint project, whose goal is to provide a set of comparable newspaper corpora for European languages in that time frame. We discuss the raw data that was previously available, as well as new data specifically compiled for the project, and the challenges involving OCR, text recognition and different orthographical norms. We describe the pipeline setup for XML encoding and annotation, partially based on work developed for the ParlaMint corpora. The corpus is currently under development and will be made freely available at the end of the project, as part of the PressMint corpora.

Keywords: newspaper corpus, historical corpus, text recognition

1. Introduction

Although corpora composed of European Portuguese are increasingly available for research, finding historical newspaper corpora in Portuguese is still a challenge. The existing initiatives either do not comprise newspaper texts and focus on earlier stages of the Portuguese language, such as the medieval corpus *Corpus Informatizado do Português Medieval*[1], or they only comprise epistolary writing, such as the Post Scriptum corpus[2]. Other corpora (see section 2) frequently do not separate European from Brazilian Portuguese, treat the 20th century as a single period, or follow their own standards.

The situation applies to other European languages, and as a result, the PressMint project aims to compile a multilingual, comparable, annotated, translated, and interoperable set of corpora of European historical newspapers. This initiative follows the work already completed to compile a comparable multilingual corpus of Parliamentary data from several European countries under the ParlaMint project (Erjavec et al., 2025). The PressMint-PT corpus follows the common standards setup in the project for this specific genre and builds upon the expertise gained in Portuguese XML file processing for the ParlaMint-PT project.

The period of the late 19th and first half of the 20th centuries was extremely prolific in terms of newspaper titles. A list, in the form of a dictionary, of the more than three hundred daily newspapers published between 1900 and 2000 in Portugal can be found in Lemos (2020), together with a summary of the history of each newspaper, details of where it can be consulted and a study on the History of the Portuguese Daily Press in the 20th Century.

The evolution of newspapers is directly related to the political and social changes that occur in Portugal (Tengarrinha, 1971). So, a historical newspaper corpus is crucial for historical studies, as it would make available data from a period that covers the Monarchy, then the first Republic, and finally the dictatorship of the Estado Novo. The comparisons that an interoperable set of European corpora offer are valuable for analyzing the political and social events that structured the beginning of the 20th century. It will also be of interest to a historical perspective on Communication Studies, enabling a direct view of the changes that affected the newspaper genre. These are areas of knowledge that are frequently unaware of the available natural language processing methods, or simply underuse them, instead accessing historical newspapers in an unfriendly image format. We are confident that these areas of knowledge that deal with digitized versions of paper-based formats will be greatly enhanced by using the PressMint-PT to access relevant data. Additionally, such a corpus is especially interesting for observing changes that occurred in early contemporary Portuguese from a Historical Linguistics point of view, comparing this period with later stages of the language.

We will present in section 2 other newspaper corpora for Portuguese and in section 3 the data that we have included so far in the PressMint-PT corpus. The actual corpus processing is discussed in section 4, and the XML encoding and annotation pipeline is discussed in section 5, before concluding in section 6.

[1] https://cipm.fcsh.unl.pt
[2] http://teitok.clul.ul.pt/postscriptum/

Proceedings of the First Workshop on Creating Interoperable Corpora of Historical Newspapers (PressMint 2026), pages 16–20

2. Related Work

There are three main corpora of newspaper texts in European Portuguese. One is the CETEMPublico corpus, which comprises 190 million words from the Portuguese newspaper *Público*[3] (Santos and Rocha, 2001). It covers contemporary Portuguese from the 1990's to the 2000's, and falls outside the goals of the PressMint corpus. The *Corpus do Português* (Corpus of Portuguese) contains 45 million words from European and Brazilian Portuguese taken from the 14th to the 20th century[4] (Mark Davies, 2016) (Mark Davies and Michael Ferreira, 2006). The 19th century section covers several genres of European and Brazilian Portuguese and comprises around 10 million words. The 20th century section includes around 3 million words, but most fall into the 1990's and 2000's (an example would be the newspaper Público, which started in the 1990's). The corpus is available for online queries.

The other corpus containing newspaper texts from the 19th and 20th centuries is the *Corpus de Referência do Português Contemporâneo* (Reference Corpus of Contemporary Portuguese). Since we will use part of this corpus for the constitution of the PressMint corpus, the corpus will be presented in detail in section 3.

3. Raw Corpus

For the compilation of a Portuguese newspaper corpus, we relied on the data included in the Reference Corpus of Contemporary Portuguese (CRPC). The corpus focuses mainly on European Portuguese data from the last quarter of the 20th century, but also includes smaller sections on the other varieties of Portuguese spoken in the world, namely data from Brazil, Angola, Cape Verde, Guinea-Bissau, Mozambique, Sao Tome and Principe, Macau, and Timor. As there is very little newspaper historical data in the CRPC, we explored adding new data available in image format to expand our collection. We describe the already existing data in section 3.1 and the new collected data in section 3.2.

3.1. The CRPC corpus and the VARPORT subsection

The CRPC corpus comprises 311 million words of written and spoken texts of different varieties of Portuguese in the world. The written subcorpus includes texts from different genres: newspapers and magazines, fiction, didactic and scientific texts, parliamentary data, law and court rulings, letters, and brochures (Généreux et al., 2012). The written part of this corpus covers 309,812,943 tokens, compiled from 356,208 documents, mostly from 1970 to 2008, although texts from 1800 forward are also included. The corpus was developed at the Center of Linguistics of the University of Lisbon[5] and is available for online queries on CQPweb[6]. The newspapers section of the CRPC, restricted to Portugal, contains 98,579,946 tokens and 160,289 texts. For the PressMint project, we restricted this newspaper section to the period from 1808 to 1945. We provide the number of tokens and the number of files for 30-year periods in Table 1. The table shows that only a small subset of 63,178 tokens fits the time window that was set for the PressMint collection, and highlights the difficulty of finding and processing historical newspaper data.

A large part of the CRPC data mentioned in Table 1, covering the period of the 19th century and early 20th century, was compiled in the framework of the VARPORT project[7]. We will refer to the data that originates from the CRPC corpus and its VARPORT subset as the CRPC/VARPORT corpus.

Table 1: Number of newspaper files and tokens per time period in the European Portuguese CRPC/VARPORT subcorpus

time period	no. of tokens	no. of files
1800-1829	6,239	28
1830-1859	21,245	170
1860-1889	9,682	69
1890-1919	13,344	93
1920-1949	12,668	76
Total	63,178	436

The time period selected for the PressMint corpus is determined by the available material in the CRPC/VARPORT corpus. Each 30-year phase includes three types of data: newspaper articles, editorials, and advertisements/announcements. We decided to keep the latter type in the PressMint corpus for two reasons: first, some announcements are related to information provided by institutions and companies and are not strictly advertising texts; second, even the advertisements can be relevant for future applications, such as the geolocation of companies. The transcriptions of the newspaper texts were performed through digitization, OCR recognition, and manual revision. It should be noted that we kept the original orthography in the

[3] https://www.linguateca.pt/CETEMPublico/

[4] https://www.corpusdoportugues.org

[5] https://www.clul.ulisboa.pt

[6] gamma.clul.ul.pt/CQPweb

[7] VARPORT is a joint venture between the Center of Linguistics of the University of Lisbon (CLUL) and the Federal University of Rio de Janeiro (UFRJ). Project website: https://varport.letras.ufrj.br; supervision: Sílvia Brandão (UFRJ) and Antónia Mota (CLUL)

texts of this set of files.

3.2. Enlarging the corpus

The data in the CRPC/VARPORT corpus is available in digital format and has been previously revised, making it immediately suitable for the next processing step, but the number of tokens is relatively small. To enlarge the corpus, we explored publications from the late 19th and the beginning of the 20th centuries available on the Hemeroteca website[8], in image format. We initially selected three potential newspapers:

- *A Capital*, a daily newspaper;
- *O Ocidente*, a three-times-monthly publication;
- *Revista Universal Lisbonense*, a weekly magazine.

The time period and the number of images for each publication are shown in Table 2.

Table 2: Details from selected Hemeroteca publications

publication	time period	no. of files
A Capital	1910-1938	19,086
O Ocidente	1878-1915	10,845
Rev. Uni. Lisb.	1841-1853	7,042
Total	1841-1938	36,973

The digitized images of these publications raise several challenges to natural language processing methods: not only are some images damaged with tears, wrinkles, or old-age stains, but also the orthography used at the time is different from the current norms, which affects the quality of the OCR process and the accuracy of the linguistic annotation tools. In addition, both the publication in columns and the presence of embedded images make it harder to extract the text in its correct sequence. Such issues will be addressed in the following section 4.

The example in Figure 1 provides the front page of the Lisbon newspaper *A Capital* on its edition of August 1st of 1910, illustrating its compact design and the different orthographical norms where, for instance, *oficiaes* would now be spelled *oficiais*, *monarchico* would be *monárquico*, and *teem* would be *têm*.

4. Corpus Processing

To simplify the production of the required XML documents, we processed the sources mentioned in section 3 above, since even though such text files

[8] https://hemerotecadigital.cm-lisboa.pt


A CAPITAL

DIARIO REPUBLICANO DA NOITE

LISBOA—Segunda-feira, 1 de Agosto de 1910

Campanha eleitoral

CONTRABANDO

Outra assembleia do Credito Predial

O sr. João Albino "despeja o sacco"

Blasco Ibañez

Dr. Bernardino Machado

Ecos do dia

Figure 1: A front page of the newspaper *A Capital*.

were already compiled, they still required additional processing, which will become clear in the following subsections.

4.1. Processing the CRPC/VARPORT Documents

As mentioned before, the CRPC/VARPORT files were already available in text format, but because the purpose for which they were originally created did not account for a clear separation among some data elements relevant to the PressMint project, the available metadata was not quite normalized and had some minor issues, so additional processing was necessary to deal with situations like the following:

- different terms used to identify the same field;
- dates in different formats;
- more than one element in the same line;
- elements spanning more than one line;
- some orthographical errors;
- missing accents;
- extra spaces;

- different orthographies;
- mismatched letter casing.

As an example, we can see below a situation in which the date (Data) and location (local) could be split into separate lines. Also, the date includes the weekday, which is generally unnecessary.

Data e local: Segunda-feira, 18.12.1837 - Lisboa

Some other examples of date normalization include the following:

- domingo, 4 de janeiro de 2026;
- 4 de janeiro de 2026;
- 4-1-2026;
- 4 jan 26 00:00;
- 4.1.2026.

The examples above are all represented by 2026-01-04 after normalization.

The following example shows several fields concatenated and split into several lines in which some fields are even empty.

Jornal: A Capital Número: 8679 Data: Sábado, 13 de Novembro de 1995 Local/Edição: Lisboa Secção: Página: Coluna: Autor: Título: Ficheiro: acordo.txt Introdução: Suporte magnético Revisão:

Such normalization methods required the use of regular expressions, date parsing, and expression substitutions, which greatly simplified the final metadata encoding, described in section 5 below.

4.2. Processing the Additional Image Documents

These documents are the most difficult to process since they are only available as images, requiring an OCR stage to produce the corresponding text. Such stage might even be preceded by a segmentation stage in which the location of the actual text within the image is determined before being presented to the text recognition process.

Taking into account the challenges mentioned in section 3.2, we have conducted a few initial experiments using ImageMagick (ImageMagick Studio LLC, 2024) to modify the images, and tesseract (Ooms, 2026) with its best (most accurate) model to produce the text contained in the corresponding images, with the purpose of finding the best parameters with which to apply the OCR, such as image enhancement, contrast, anti-alias, gamma, and resizing, confirming those have an impact on quality, but so far we have obtained mixed results, gaining quality in some sample areas but losing it in others, so we have still not found a clear winner.

Additionally, we have also prepared a small set of verified examples, with which we fine-tuned the previous tesseract best model, but the images used must also go through some image processing, as well as the number of examples apparently has to be increased before we are able to achieve any relevant positive impact.

Finally, we fear that we might need to tune the process to the different publications since they have some graphical styling differences between them.

To tackle all these difficulties, we plan to start by using a more simplistic programmatic approach, in which we only consider the number of total words and the number of out-of-vocabulary words, hoping to at least be able to make obvious the parameters that will be less likely to produce good results and therefore discard them. Once we have a smaller set of parameter candidates, it will be easier to check their results individually, ideally using a smaller set of random samples to be verified by a human.

5. XML Encoding

Once we normalized the metadata from the CRPC/VARPORT texts, as described in section 4, it became easier to produce their XML and their annotated XML counterpart, as explained in the following subsections.

5.1. XML Documents Generation

The contents of each document took into account not only the actual text, but also its metadata, which became very simple thanks to the processing stage described above. Additionally, in order to apply the naming convention, the files were organized according to their publication dates, followed by their publication source, and finally followed by a number to distinguish files with the same publication date and publication source.

5.2. Annotated XML Documents Generation

The annotation information for each XML file produced above, consisting of lemma, POS, UDR and NER, was obtained using UDPipe 2 (Straka, 2018) and NameTag 3 (Straková and Straka, 2025). However, given the language's old orthographical norm, the quality of the annotation might be suboptimal, so we are considering the implementation of an additional preceding stage in which the old orthography is modernized before being presented to the annotation tools.

5.3. Main XML Documents Generation

The main XML documents include references to the individual XML documents of each text, as well as the sum of several element amounts of each document, such as the number of paragraphs, words, and punctuation marks.

6. Final Remarks

The compilation of the corpus of the additional data in image format to be included in the PressMint-PT set is still underway, which again is the most challenging part of the project. The next step is to evaluate the results of the OCR and text recognition, as well as checking to what extent the annotation tools are capable of (fully or at least partially) automatically dealing with these historical texts, which might require, for instance, an additional orthography modernization stage.

Also, considering the significant progress achieved in the AI field, we intend to explore other alternatives for image text processing, like segmentation with fine-tuned YOLO or Meta's SAM, and models like unsloth/gemma-3-4bit or dots.ocr, to name a few. The segmentation stage will hopefully provide some ideas as to how the articles could be presented, since those are generally non-linear and their borders are not easily determined.

An evaluation of the difficulty of processing individual historical publications will ultimately affect the selection of newspapers to include in the corpus.

7. Acknowledgements

We would like to thank the anonymous reviewers for their helpful comments and suggestions, which have actually provided some ideas we have included and which we intend to explore further. This work was partially supported by CLARIN ERIC PressMint-Interoperable corpora of historical newspapers, and by Fundação para a Ciência e a Tecnologia as part of the project of Centro de Linguística da Universidade de Lisboa (UID/214/2025).

8. Bibliographical References

Tomaž Erjavec, Matyáš Kopp, Nikola Ljubešić, Taja Kuzman, Paul Rayson, Petya Osenova, Maciej Ogrodniczuk, Çağrı Çöltekin, Danijel Koržinek, Katja Meden, Jure Skubic, Peter Rupnik, Tommaso Agnoloni, José Aires, Starkaður Barkarson, Roberto Bartolini, Núria Bel, María Calzada Pérez, Roberts Dargis, Sascha Diwersy, Maria Gavriilidou, Ruben van Heusden, Mikel Iruskieta, Neeme Kahusk, Anna Kryvenko, Noémi Ligeti-Nagy, Carmen Magariños, Martin Mölder, Costanza Navarretta, Kiril Simov, Lars Magne Tungland, Jouni Tuominen, John Vidler, Adina Ioana Vladu, Tanja Wissik, Väinö Yrjänäinen, and Darja Fišer. 2025. *ParlaMint II: advancing comparable parliamentary corpora across Europe*, volume 59. Springer Netherlands.

M. Généreux, I. Hendrickx, and A. Mendes. 2012. Introducing the Reference Corpus of Contemporary Portuguese On-Line. In *LREC'2012 – Eighth International Conference on Language Resources and Evaluation*, pages 2237–2244, Istanbul, Turkey. European Language Resources Association (ELRA).

ImageMagick Studio LLC. 2024. *ImageMagick*.

Mário Lemos. 2020. *Jornais Diários Portugueses do Século XX – um dicionário*.

Mark Davies. 2016. Corpus do português: Web/dialects.

Mark Davies and Michael Ferreira. 2006. Corpus do português: Web/dialects.

Jeroen Ooms. 2026. *tesseract: Open Source OCR Engine*. R package version 5.2.5.

Diana Santos and Paulo Rocha. 2001. Evaluating cetempublico, a free resource for portuguese. In *Proceedings of 39th Annual Meeting of the Association for Computational Linguistics*, pages 450–457, Toulouse, France. Association for Computational Linguistics.

Milan Straka. 2018. UDPipe 2.0 prototype at CoNLL 2018 UD shared task. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 197–207, Brussels, Belgium. Association for Computational Linguistics.

Jana Straková and Milan Straka. 2025. NameTag 3: A tool and a service for multilingual/multitagset NER. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 31–39, Vienna, Austria. Association for Computational Linguistics.

José Tengarrinha. 1971. *História da Imprensa Periódica Portuguesa*. Portugália Editora.