CorEGe-PT: Compiling a Large Corpus of Academic Texts in Portuguese

Tanara Zingano Kuhn[1,2,7], **José Matos**[1,3], **Bruno Neves**[1,4], **Daniela Pereira**[5], **Elisabete Cação**[1,6], **Ivo Simões**[1,3], **Jacinto Estima**[1,3], **Delfim Leão**[1,6,7], **Hugo Gonçalo Oliveira**[1,3]

[1]University of Coimbra; [2]CELGA-ILTEC; [3]CISUC/LASI, DEI; [4]BGUC; [5]Independent Researcher; [6]CECH; [7]FLUC

tanarazingano@uc.pt, josematos@student.dei.uc.pt, bneves@uc.pt, danielafsp92@gmail.com, elisabetecacao@gmail.com, ivosimoes@student.dei.uc.pt, estima@dei.uc.pt, leo@fl.uc.pt, hroliv@dei.uc.pt

Abstract

This paper describes the creation of a large-scale corpus of academic texts in Portuguese, dubbed CorEGe-PT, extracted from the institutional repository of a Portuguese university. Its compilation methodology, which combined automatic and manual procedures, is detailed, together with challenges faced and proposed solutions. The process included a thorough analysis of the metadata, which is publicly released together with the documents, extracted in a markdown format. CorEGe-PT covers five areas of knowledge and, with over 34,000 documents and 1B tokens, is the largest corpus of its kind in Portuguese, which will enable in-depth linguistic studies while providing data for adapting Large Language Models to academic Portuguese and related tasks.

Keywords: Corpus Creation, Corpus Linguistics, Academic Documents, Science, Portuguese

1. Introduction

This paper reports on the compilation of the *Corpus do Estudo Geral* – Portuguese (dubbed CorEGe-PT), a corpus of academic texts written in Portuguese and deposited in the repository of a Portuguese university. It comprises more than 34,000 documents and over 1B tokens, due to the great proportion of large documents like theses. Compared to other languages, such as English, Portuguese is less-resourced, and lacks linguistic resources, including specialized corpora. CorEGe-PT fills this gap by providing researchers with a large corpus of Portuguese comprising different academic genres across five areas of knowledge, and covering different levels of academic writing experience. The majority of the documents (87%) is written in European Portuguese, with a small portion of documents in Brazilian Portuguese.

By following the methodology of corpus linguistics (Biber et al., 1998; Meyer, 2002; McEnery et al., 2006; Sinclair, 2003; Wynne, 2006) to compile such a specialized corpus, CorEGe-PT may support in-depth linguistic studies, such as the analysis of lexico-grammatical and discursive features that may contribute to the characterization of academic genres, areas of knowledge, as well as to the development of pedagogical resources like dictionaries, grammars, and teaching materials. Moreover, it will provide domain-specific data for adapting current Large Language Models (LLMs) to academic Portuguese, and support a range of tasks in the scientific domain, from the summarization of scientific literature to question-answering based on this kind of documents. In compliance with FAIR principles and Open Access regulations, CorEGe-PT is publicly available for download from the Huggingface Hub[1] and soon also via a corpora analysis tool[2] .

The structure of this paper is as follows. Section 2 reviews related work. Section 3 presents the repository where the corpus was extracted from. Section 4 describes the methodology of corpus compilation, consisting of metadata acquisition and analysis, metadata revision, text extraction, and postprocessing. In addition, it provides some analyses and discussions of the challenges faced and solutions found. Section 5 describes the resulting corpus, together with a quantitative analysis and some additional figures. Lastly, Section 6 concludes the paper and drafts future plans.

2. Related Work

Academic corpora that are built for linguistic research purposes tend to follow the principles and techniques of Corpus Linguistics closely. Among others, these involve careful text selection and detailed metadata acquisition and annotation. Such corpora are available in several languages, and in different types and sizes, as can be found in repositories such as CLARIN ERIC[3] and Open Language

[1]https://huggingface.co/datasets/NLP-CISUC/CorEGe-PT

[2]https://corpora.celga.iltec.pt/

[3]https://www.clarin.eu/

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 1533–1543
11–16 May 2026.

Archives Community[4], or integrated in corpus management tools such as Sketch Engine (Kilgarriff et al., 2014). Some corpora focus on students' writing, e.g., the British Academic Written English Corpus (BAWE) (Alsop and Nesi, 2009) and the Corpus of Academic Slovene KAS 2.0 (Erjavec et al., 2020), others on specific written genres, like journal articles, e.g., the Corpus of Academic Journal Articles (Kosem, 2010). More rarely, some corpora comprise spoken discourse, with the British Academic Spoken English (BASE) (Thompson and Nesi, 2001) corpus as a renowned example.

Within Natural Language Processing (NLP), and especially with the advent of LLMs, the term "corpus" mostly refers to a simple dataset of texts. This is because these corpora are created for purposes such as pre-training or fine-tuning LLMs to perform multiple tasks in several domains. The compilation process of these corpora usually consists of collecting as much data (i.e., texts) as possible, while other aspects that are crucial for linguistic research, such as detailed metadata and careful selection of sources, are not necessary. For example, a corpus of 1.1M scientific publications from Semantic Scholar has been compiled in the areas of computer science (18% of the papers) and biomedical science (82%) to train SciBERT (Beltagy et al., 2019), a BERT-based model. The training of the decoder-only LLM Galactica (Taylor et al., 2022) was based on a corpus with over 48M scientific papers, textbooks, and lecture notes, covering a broad spectrum of scientific fields from a range of sources, including PubMed, Semantic Scholar and arXiv. Additional examples include a corpus of 1M arXiv papers (Saier and Färber, 2019), and S2ORC (Lo et al., 2020), which originally contained 81.1M academic publications from 20 academic fields, with full text available for 10%.

Another application of academic corpora is in Retrieval-Augmented Generation (RAG) to adapt LLMs to a target domain without adjusting their parameters. Examples of tasks include question answering (Zheng et al., 2024) and identifying relevant papers and summarizing their findings (Asai et al., 2024).

Interestingly, smaller corpora compiled within the corpus linguistics perspective are also useful within NLP. For instance, BAWE was recently used as a reference for assessing academic essays produced by LLMs (Mo and Crosthwaite, 2025), and as part of the data in a RAG chatbot designed for students developing academic writing (Cheung and Crosthwaite, 2025).

The previously cited work refers to English, which is not surprising, given its current role as the international language of science. Nevertheless, science and academic production are made in other languages as well, including in Portuguese. However, publicly available, open access corpora, especially larger ones, are scarce. Examples that stand out are the Corpus of Academic Portuguese (*Corpus de Português Académico* – CPA) (Santos et al., 2020), comprising 768 texts from students at universities in Portugal and Mozambique; and the Corpus of Portuguese from Academic Journals (*Corpus de Português Escrito em Periódicos* – CoPEP) (Kuhn and Ferreira, 2020), a 40M-word corpus, balanced between the varieties from Brazil and Portugal, and covering six great areas of knowledge. Other sources of academic-scientific texts in Portuguese can be found in corpora that are multilingual (Soares et al., 2018) or domain-specific, e.g., on Physics and Chemistry, available within the *Textos Técnicos e Científicos* project[5].

Given the lack of large, openly available corpora of academic texts written in Portuguese, we compiled the CorEGe-PT corpus. Similarly to the methodology applied in the compilation of CoPEP and the Corpus of Scientific texts from the Open Science Slovenia portal OSS 1.0 (Žagar et al., 2023), documents were acquired from a single repository, in this case, Estudo Geral (Section 3), following strict selection criteria (Section 4).

3. Data Source

The contents of CorEGe-PT are collected from Estudo Geral[6], the institutional repository of the University of Coimbra, in Portugal, committed to disseminating and preserving the scientific output of authors affiliated with the institution. Documents are organized into communities and collections, reflecting the organizational structure of the University. At the top level, communities represent the various faculties, which are further divided into subcommunities corresponding to smaller organizational units, such as departments and research centers. Each community or subcommunity is associated with collections that are further organized by document type, such as articles; bachelor, master, and doctoral theses; reviews; and reports. The metadata description for these documents primarily follows the Dublin Core[7] schema. Additional metadata standards are also employed to enable the integration of records into aggregators such as the Open Access Scientific Repositories of Portugal (RCAAP)[8], and OpenAIRE (Miguéis and Neves, 2021). Certain metadata elements use controlled vocabularies recommended by the Guidelines for Literature Repos-

[4] http://www.language-archives.org/

[5] https://www.ufrgs.br/textecc/

[6] https://estudogeral.uc.pt/

[7] https://www.dublincore.org/specifications/dublin-core/

[8] https://www.rcaap.pt/

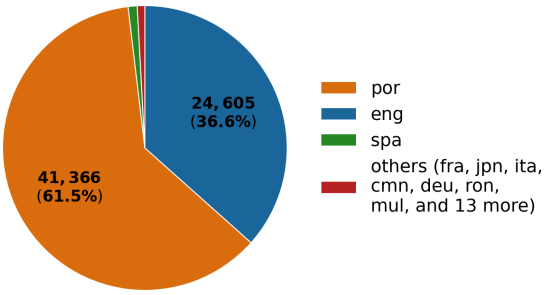

Figure 1: Language (ISO 639-3) distribution in Estudo Geral, as of April 2025.

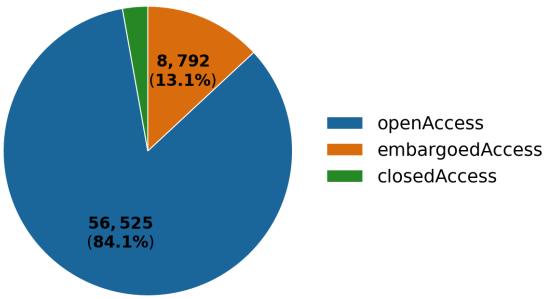

Figure 2: Right of access distribution in Estudo Geral, as of April 2025.

itories v3, particularly for fields such as publication type (`dc.type`), rights (`dc.rights`), and language (`dc.language.iso`), among others.

4. Compilation Process

Three criteria were adopted to select the documents to incorporate in CorEGe-PT. They had to: (1) be in Portuguese; (2) be available in open access; (3) have no copyright restrictions. However, additional steps are required when building a corpus for linguistic research purposes, including metadata acquisition and text processing. The compilation of the CorEGe-PT corpus involved working with metadata of publications from Estudo Geral, combining automatic and manual procedures that were iteratively refined to ensure data quality and consistency, along with processing textual data itself.

4.1. Metadata Acquisition and Analysis

The first step was the acquisition of the metadata for all the 67,227 documents available in Estudo Geral, as of April 2025. These comprised 113 fields for each record, stored in a spreadsheet that was shared among our team. Given the goal of CorEGe-PT, the metadata analysis started by verifying how many documents were written in Portuguese overall. To do this, we counted all records for which the value of the Language field (`dc.language.iso`) was Portuguese, resulting in 41,366 records (≈61%). Figure 1 shows the distribution of languages in the repository, confirming that, as the repository of a Portuguese institution, Portuguese is the most represented language, followed by English (≈37%).

Next, we analyzed the URI of the applied licenses (`dc.rights.uri`) and learned that it was available for just about 40% of the documents. In contrast, the rights of access (`dc.rights`) are available for every record, even if spread across different fields. Its distribution is in Figure 2, showing that a small portion of the documents is in closed access, and a considerable share (13%) is embargoed.

This initial analysis revealed a substantial amount of documents in Portuguese along with a variety of metadata, thus indicating that the corpus could be compiled. We then applied the three abovementioned criteria to define which documents would be downloaded. The first filter was applied to the rights of access and excluded the 1,913 records in closed access, as well as the 8,209 records that explicitly had a *no derivatives* (ND) license. The second filter aimed at keeping only documents written in Portuguese. The third and final filter removed 252 documents that, though not in closed access, were still under embargo in 2025. After applying the three filters, approximately half of the original documents remained, more specifically, 34,482 records matching the three criteria.

4.2. Metadata Revision

When examining the metadata, we noticed that some fields were redundant or complementary, due to successive updates to the repository system. Details such as the DOI, language, or usage rights often appeared across several fields with slight variations. For instance, there were several fields for the license and the usage rights in different languages (e.g., `dc.rights[por]`, `dc.rights[eng]`, `dc.rights[en_US]`), even if, in most cases, only one of them had information.

This kind of issues was resolved by merging the fields, followed by a manual check and minor normalization (e.g., `Open Access` or `open` were normalized to `openAccess`). In fact, Figures 1 and 2 presented earlier already reflect this cleaning. Moreover, to simplify the metadata, we decided to focus on a subset of fields, prioritizing those filled for the majority of the records, namely:

- `Collection`: name of the collection where the publication is deposited.

- `dc.title`: publication title.
- `dc.contributor.author`: name of the author of the publication.
- `dc.date.issued`: publication date (ISO-8601).
- `dc.subject`: publication keywords or subject terms assigned by the repository manager.
- `dc.subject.fos`: classification of the Field of Science and Technology (FOS).
- `dc.identifier.uri`: persistent identifier (repository handle).
- `dc.identifier.doi`: persistent identifier (DOI).
- `dc.type`: publication type (e.g., masterThesis, doctoralThesis, article).
- `dc.rights.uri`: Creative Commons license.
- `dc.description`: general information about the publication or an explanation of its content.

Among the selected fields, the classification of Field of Science and Technoloogy (FOS) was deemed very useful for a corpus of scientific documents. However, the field was not standardized and was populated for just about 5% of the records. Therefore, we developed alternative approaches to ensure that each record had an assigned FOS. As manually filling was impractical, a heuristic was first defined for mapping the field `Collection`, often containing the name of the faculty or research center where the document had been produced, to one of the six high-level OECD's revised FOS classification[9], adopted by the Portuguese Foundation for Science and Technology (FCT): Humanities, Social Sciences, Exact and Natural Sciences, Engineering and Technology Sciences, Medical and Health Sciences, Agricultural Sciences. Since the University of Coimbra does not have a unit dedicated to Agricultural Sciences, Estudo Geral does not contain documents of the latter FOS, thus reducing the number of possible values to five.

The heuristic led to the automatic mapping of 26,348 records (≈76%). Due to inconsistencies in the name of the collection, an additional 920 records had to be mapped manually. Yet, a total of 7,215 records (≈21%) could still not be assigned a FOS. These corresponded to master theses from a generic collection, named *UC – Dissertações e Teses* (UC Thesis and Dissertations). For these documents, the assignment of a FOS was left for automatic classification (see Section 4.4.1).

Overall, the methodology applied for metadata revision resulted in a spreadsheet with coherent and complete fields, despite the challenges due to incompleteness, redundancies, and inconsistent classifications. The metadata of each document was then saved in a textual format, illustrated in Figure 3 for two documents in CorEGe-PT. The considered metadata fields are in the first 11 lines, and the last five lines represent fields added after post-processing, further explained in Section 4.4.

4.3. Text Extraction

After collecting and revising the metadata, we proceeded with automatic extraction of the files from the 34,482 records that matched the three criteria (Section 4.1). It should be noted that, in several cases, more than one file is associated with one single metadata record. For instance, if a thesis contains multiple annexes, it is registered as one single metadata record; however, the main thesis and the annexes are saved in the Estudo Geral repository as separate files. The files were then converted to Markdown, UTF-8 encoded, with the help of Docling (Livathinos et al., 2025), making them readily available for text processing tasks (see Figure 4 for an illustrative excerpt). Markdown[10] is a text-based, structured yet lightweight format, frequently used in modern processing pipelines, including workflows involving LLMs. A useful asset for linguistic research purposes, especially in the area of Academic Discourse Studies, is that different sections of the markdown files are marked with `#`, allowing for their inclusion as search variables in corpus analysis tools.

Of the 34,482 metadata records that met the selection criteria, each potentially associated with more than one file (for example, theses with multiple annexes), 32,111 (93.1%) records had at least one associated file successfully extracted. For 1,952 records, no files could be downloaded at all, due to causes such as internal server errors, records that were withdrawn from the repository between the metadata collection and download phases, and records whose embargo had not yet expired at the time of download, despite being close to expiration when selection criteria were applied, and therefore kept as valid metadata. The remaining 32,530 records resulted in a total of 34,704 downloaded files. During the markdown conversion phase, Docling failed to process 419 of these files (1.2%) due to issues such as non-PDF formats or file corruption, resulting in 34,285 markdown files covering 32,111 unique metadata records (93.1% of the original 34,482).

4.4. Postprocessing

After text extraction, two post-processing steps were still performed: one for classifying the FOS of the records for which this field was not assigned by the heuristic (Section 4.2); another for classifying the variety of Portuguese of the documents (see

[9] https://one.oecd.org/document/DSTI/EAS/STP/NESTI(2006)19/FINAL/en/pdf

[10] https://www.markdownguide.org/

```
Collection:  UC - Dissertações e Teses
dc.title:  Modelo de Inteligência Artificial para inferir automaticamente os estados mentais do paciente
associados ao engagement
dc.contributor.author:  Carvalho, Beatriz Lopes de
dc.date.issued:  2023
dc.subject:  Engagement||Demência||Expressões Faciais||Inteligência Artificial||VGG16
dc.identifier.uri:  https://hdl.handle.net/10316/110650
dc.type:  masterThesis
dc.rights:  openAccess
dc.rights.uri:  http://creativecommons.org/licenses/by/4.0/
dc.subject.fos:  Ciências Exactas e Naturais
dc.description:  Trabalho de Projeto do Mestrado em Engenharia Biomédica apresentado à Faculdade de Ciências
e Tecnologia
----------
fos.assignment:  classifier
pt.auto:  true
pt.mean.confidence.auto:  0.801
pt-pt.auto:  true
pt-pt.mean.confidence.auto:  0.625
```

```
Collection:  FLUC - Secção de Línguas Românicas
dc.title:  Onde está o exemplar de Os Lusíadas de 1572 com oitavas transpostas?
dc.contributor.author:  Marnoto, Rita
dc.date.issued:  2023
dc.subject:  Os Lusíadas||Luís de Camões||Edição princeps de Os Lusíadas||Edição crítica de Os
Lusíadas||British Library, G.11286||DA, Diocese do Algarve
dc.identifier.uri:  https://hdl.handle.net/10316/111175
dc.type:  article
dc.rights:  openAccess
dc.rights.uri:  N/A
dc.subject.fos:  Humanidades
dc.description:  (2023).  Onde está o exemplar de Os Lusíadas de 1572 com oitavas transpostas?, Românica.
Revista de Literatura, 25, 157-184.
-------
fos.assignment:  heuristic
pt.auto:  true
pt.mean.confidence.auto:  0.938
pt.pt.auto:  true
pt.pt.mean.confidence.auto:  0.999
```

Figure 3: Metadata of two documents in CorEGe-PT.

```
## Introdução

## 1.1 Contexto do Projeto

A presente tese foi desenvolvida a partir de uma parceria entre a
Universidade de Coimbra e o Center for Research in Neuropsychology and
Cognitive and Behavioral Intervention (CINEICC).

...

## 1.2 Motivação

Existem no mundo bastantes problemas de saúde que requerem tratamentos
a longo prazo e, por vezes, estes são difíceis de realizar por parte
dos doentes. É geralmente reconhecido que o empenho dos doentes
desempenha um papel fundamental no sucesso desses tratamentos.
Infelizmente, muitas vezes, os doentes perdem a sua motivação, o seu
interesse, o chamado engagement 1, durante o período de tratamento.
Assim, é muito importante detetar as flutuações do nível de empenho
dos pacientes ao longo de toda a duração do tratamento, e produzir
alertas quando este se afasta dos níveis normais, para que os
profissionais possam atuar a tempo de tentar corrigir a situação.
Infelizmente, essas situações não são usualmente comunicadas, ou
porque os doentes não são capazes de as detetar ou porque,
simplesmente, não o querem fazer. A solução poderá passar pela
incorporação de técnicas de Inteligência Artificial (IA) [4] para que
estas situações possam ser detetadas de forma automática e rápida [5].

...

## 1.3 Objetivos e Metodologia

Este estudo tem como principal objetivo detetar engagement em pessoas
com demência, através da análise das suas expressões faciais, durante
a fase de tratamento, uma vez que existe uma grande dificuldade, por
parte dos profissionais de saúde, em conseguir que os seus pacientes
mantenham interesse durante esta etapa.
```

Figure 4: Excerpt from markdown file, with the first metadata in Figure 3.

Section 4.4.2). This resulted in five additional metadata fields, illustrated in the last five lines of each example in Figure 3, and briefly explained as follows:

- `fos.assignment` identifies the method used for assigning a FOS to the document (heuristic or classifier);
- `pt.auto` indicates whether the document was automatically classified as written in Portuguese, with confidence in `pt.mean.confidence.auto`;
- `pt-pt.auto` indicates whether the document was automatically classified as written in European Portuguese, with confidence in `pt-pt.mean.confidence.auto`.

The inclusion of these fields enables additional filtering, depending on the user's goal. Among others, they may: focus on a stricter selection of documents for which both the metadata and the automatic classifier label as Portuguese, possibly considering the confidence; work only on the documents labeled as European Portuguese; or disregard FOS classifications resulting from automatic classification.

4.4.1. Field of Science and Technology Classification

In Section 4.2, we mentioned that it was not possible to map 7,215 records with a FOS. To address this, we took advantage of the records for which this field had been successfully assigned and relied on supervised learning to automatically classify the FOS of the remaining documents.

Initially, the 27,268 records with an assigned FOS were reduced to a subset of 27,019, which excluded the 249 that lacked information in the fields deemed relevant for classification, namely `dc.title`, `dc.subject`, `dc.description`. Then, this subset was randomly split into two stratified subsets with 80% and 20% of the records, respectively used for training and evaluation.

Classification relied on different combinations of the aforementioned fields, as well as on the abstract of the document. The content of these fields was preprocessed to remove double spaces, URIs and bibliographic references.

The training subset was used for fine-tuning different pretrained encoder-based transformers, with their performance assessed in the evaluation subset. All tested models achieved an acceptable performance, with a slight advantage for those pretrained specifically for Portuguese. Table 1 reports the scores for the three best models, namely BERTimbau (Souza et al., 2020) base (110M parameters), Albertina-PTPT (Rodrigues et al., 2023) (100M), and the multilingual BERT (Devlin et al., 2019), including whether the field `dc.description` was used and the number of training epochs.

For explaining the content of the publication, the `dc.description` often mentions the type of document (e.g., PhD/MSc thesis, report), the course, faculty and / or unit where it was presented and, sometimes, the name of the supervisors (see examples in Figure 3). Since our goal was to classify as many documents as accurately as possible, we used the best model (BERTimbau) and the four fields, i.e., including `dc.description`, to classify the FOS of the remaining 7,215 documents. Nevertheless, we also report scores when `dc.description` is not used because, due to its content, it could make the task significantly easier and, unlike the other three fields, may not be found in documents from other repositories. We confirm that not using `dc.description` deteriorates the F1 score by about 1 percentage point.

4.4.2. Language and Variety Identification

As mentioned in Section 3, we can get the language of each document from the metadata of Estudo Geral, as the field is manually assigned when the documents are uploaded. We relied on this field for selecting the documents written in Portuguese. However, we decided that it would be useful to further enrich the metadata by identifying the specific variety of Portuguese used in each document, as it happens in CoPEP (Kuhn and Ferreira, 2020).

For this purpose, we relied on liaad/PtVId (Sousa et al., 2025), a Portuguese BERT model (Souza et al., 2020) fine-tuned for classifying documents as either the European (PT-PT) or the Brazilian (PT-BR) variety of Portuguese, available from the HuggingFace hub[11]. Since liaad/PtVId expects input text to be in Portuguese, we first verified whether the result of an automatic language identification matched the language indicated in the metadata. For this verification, we used the multilingual transformer XLM-RoBERTa (Conneau et al., 2020) base, fine-tuned for language detection, also available from HuggingFace[12].

Briefly, for each document, 31 snippets of up to 800 characters were randomly extracted. In the first stage, i.e., language identification, documents were classified as Portuguese (PT) if the mean confidence score across all snippets was greater or equal to 0.5. In the second stage, i.e., variety identification, a majority voting scheme was applied, i.e., documents with more than half of their snippets classified as PT-PT were labeled accordingly.

Of the 34,285 documents, 32,137 (93.7%) were classified as Portuguese. Given the criteria for selecting just documents in Portuguese, this was not expected. Therefore, in order to identify common causes for this mismatch, two authors of the paper manually inspected a sample of 60 of such documents. The most frequent issues were: problems in text extraction (41%), either with the encoding of some characters or files that had mostly image placeholders, resulting from PDFs that contain many images; actual errors in the metadata of Estudo Geral (28%). Other issues included: documents with multiple languages (13%), e.g., compilations of papers in different languages; false negatives in the automatic classification (10%); PDFs that were incomplete in the repository, e.g., with a significant portion consisting of references (7%). This means that, with the exception of false negatives, language identification was helpful in identifying files that, for most purposes, would not be useful, namely those with extraction issues or incorrect metadata. For consistency purposes, they were left in the corpus, but can easily be removed by considering the added metadata field `pt.auto`.

Of the 32,137 documents classified as Portuguese, 29,766 were classified as PT-PT, which is about 92.6%. Since Estudo Geral is the repository of a Portuguese university, the higher proportion of PT-PT was also expected. The mean confidence scores were 0.91 ± 0.09 for the Portuguese classification and 0.88 ± 0.12 for PT-PT.

5. Corpus Description

CorEGe-PT has 32,111 metadata records and 34,285 markdown files (i.e., documents), comprising 1.1B tokens. The metadata-file mismatch is

[11] https://hf.co/liaad/PtVId

[12] https://hf.co/papluca/xlm-roberta-base-language-detection

Model	dc.description	Epochs	Precision	Recall	F1
mBERT	Yes	5	0.934	**0.936**	0.935
mBERT	No	4	0.926	0.918	0.922
BERTimbau	Yes	2	**0.938**	0.935	**0.937**
BERTimbau	No	3	0.927	0.923	0.925
Albertina-PTPT	Yes	5	0.935	0.933	0.934
Albertina-PTPT	No	4	0.926	0.921	0.923

Table 1: Performance of different models fine-tuned for classifying the scientific area based on their title, keywords, abstract and, optionally, description.

explained by the issues reported in Section 4.3, i.e., documents that could not be downloaded and records with more than one file. This section explains how the corpus files are named and then provides an overview of the contents of the corpus.

5.1. File Naming

Including relevant metadata from corpus files in the filenames provides quick information about the texts without requiring to open them. This is particularly useful for tasks involving the manual selection of a subset for further inspection. Therefore, the naming of both metadata and markdown files follows specific rules, similar for both. Their naming is based on the structure: `FOS_dcType_languageVariety_year_ID`. The only difference is that the metadata file will have a different extension than the markdown files, respectively `.meta` and `.md`. Moreover, a suffix indicating the order of the file is appended, e.g., `_1` for the first and, if there are more files associated with the record, `_2`, `_3` and so on for the following. For illustrative purposes, files with the first and second sets of metadata in Figure 3 have, respectively, the following names: `ExaNat_MAThesis_PTPT_2023_063735_1.md`, `Hum_article_PTPT_2023_064245_1.md`.

5.2. Overview

Table 2 has the distribution of documents and tokens per FOS. Each document has at least one assigned FOS, but 1.3% have two or more (i.e., multiple). Social Sciences is the most represented FOS, with more than a third of the documents, followed by Medical and Health Sciences. Exact and Natural Sciences is the least represented, with just 7% of documents and tokens. Curiously, despite having twice the number of documents in Engineering and Technology, the number of tokens in Medical and Health Sciences in both is comparable.

Figure 5 shows the distribution of document types by FOS. The large proportion of theses in CorEGe-PT stands out, further explaining the high token/document ratio of 32,229. Master thesis is the most frequent type in four FOS and overall. It is the most represented type in Engineering and Technology and in Medical and Health Sciences. Article is the most common type of document in Humanities and the second for Social Sciences. For Engineering and Technology and Medical and Health the presence of articles is residual, probably because articles written in Portuguese are less common in these fields. We also note that Exact and Natural Sciences have the largest proportion of doctoral theses, while Humanities has the largest proportion of book parts.

Figure 6 shows the distribution of documents by publication year. There is an increase in the number of submissions in every five-year period until 2015. Since Estudo Geral was launched in 2008, an increase of documents is expected in those early years. In the 2016–2020 period, the number of documents starts to decrease, mainly in Medical and Health and in Social Sciences, and even more in 2021–2025. However, when inspecting the numbers for the whole repository, i.e., including documents in other languages, we notice that the number of publications by year does not vary much, except for 2024 and, mainly, 2025. So, our hypothesis is that the decrease displayed in Figure 6 is due to external reasons, such as a lower number of publications in Portuguese, especially in certain FOS; the data cut-off in April 2025; and documents being uploaded to the repository months, or even years, after their actual publication.

6. Conclusions

CorEGE-PT is a corpus with more than 34,000 academic documents, comprising 1.1B tokens, written in Portuguese, the largest of its kind for this language, with the majority of documents in European Portuguese. One of the main challenges of creating CorEGE-PT was the attention required to metadata acquisition, revision, and enrichment. The compilation process followed a methodology that combined several manual, semi-automatic and automatic steps, here described in detail, so that others may learn from our work. We also provide several analysis of the resulting corpus, focusing on distribution by document types, publication year, and the five Fields of Science and Technology covered.

FOS	#Docs	%Docs	#Tokens	%Tokens
Humanities	5,618	16.4%	210,983,788	19.1%
Social Sciences	13,003	37.9%	474,378,421	42.9%
Exact and Natural Sciences	2,442	7.1%	73,528,531	6.7%
Engineering and Technology Sciences	4,255	12.4%	154,400,829	14.0%
Medical and Health Sciences	8,506	24.8%	167,624,613	15.2%
Multiple	461	1.3%	24,047,033	2.2%
All	34,285		1,104,963,215	

Table 2: Number of documents and tokens and distribution according to FOS.

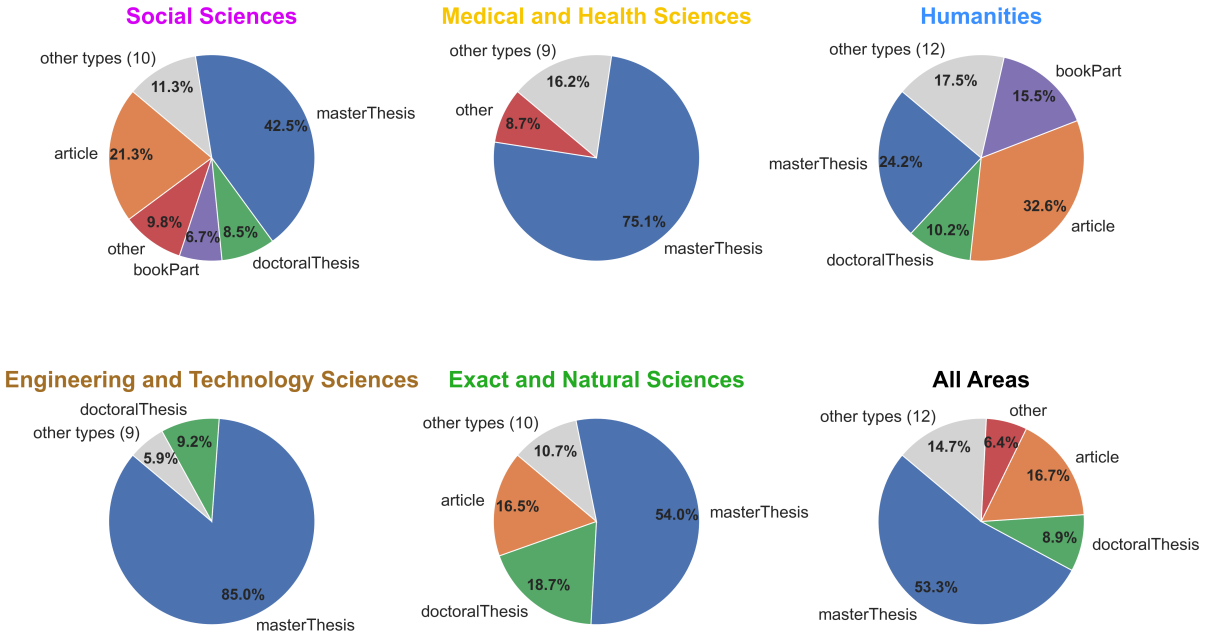

Figure 5: Distribution of document types by FOS. "All Areas" includes documents with a single FOS, as well as those that overlap over multiple FOS.

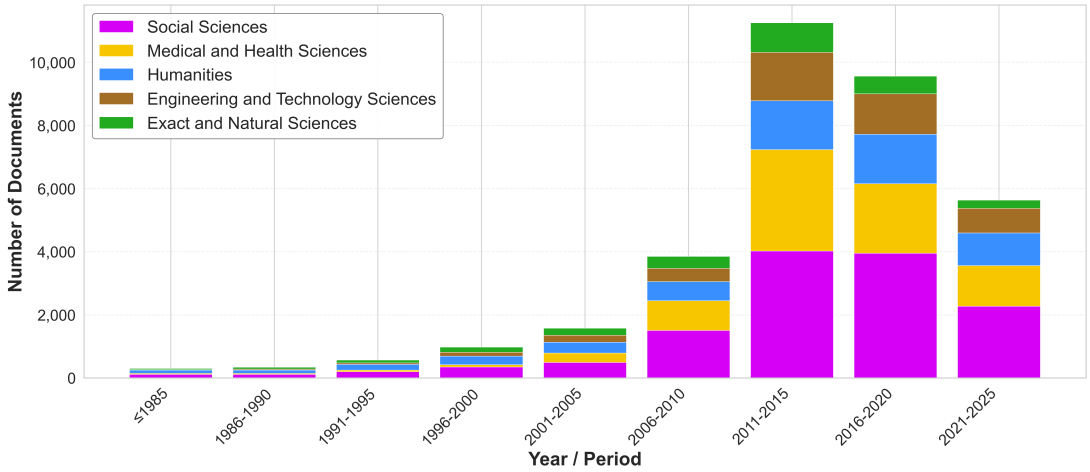

Figure 6: Temporal distribution of documents by FOS.

All the attention given to metadata aimed to ensure that CorEGe-PT can support rigorous studies of academic language. Moreover, the size of the corpus makes it a valuable resource for training LLMs, having in mind their adaptation to the academic domain, either through fine-tuning or RAG.

To make CorEGe-PT available to the research community, the content of documents, together with the revised metadata, are publicly available for download through Huggingface Datasets, in the

Parquet format, where a column exists for every metadata field. Moreover, they will be soon deployed to a corpora analysis tool. This will enable the replication of previous studies and support new research, contributing to advances in linguistics, particularly academic discourse studies, as well as Artificial Intelligence and NLP in the Portuguese language.

Nevertheless, we plan to continue improving the data and the metadata. We will explore alternative tools for the automatic analyses, and inspect the obtained results more systematically, to identify potentially problematic documents. Currently, we are exploring automated approaches for fixing already identified encoding issues. Moreover, we will add headers with metadata to the markdown files and assess the adoption of a customized classification of text types. As new publications are added to Estudo Geral, it will be possible to keep expanding the corpus. Moreover, we aim to apply the methodology described here to other academic repositories, further increasing the availability of data in this domain.

7. Ethical Considerations

All documents included in CorEGe-PT were sourced from a publicly available repository. We ensured that the documents were in open access and, when such information was available, covered by permissive licenses. In the process, we excluded documents in closed access, embargoed, or without the most restrictive licenses (ND). In addition, we acknowledge all the sources in the corpus metadata, thus recognizing the intellectual contributions of authors. To support reproducibility, we provide detailed documentation of the corpus construction process, including data sources, selection criteria, and pre and post-processing steps. The corpus is intended for research use, and we encourage responsible usage in accordance with ethical research practices.

Acknowledgements

This work was partially supported by the AMALIA project, funded by FCT/IP in the context of measure RE-C05-i08 of the Portuguese Recovery and Resilience Program; by the Portuguese Recovery and Resilience Plan (PRR) through project C645008882-00000055, Center for Responsible AI; by international funding, in the framework of the project https://doi.org/10.54499/UID/PRR2/04887/2025; and by national funds through FCT – Foundation for Science and Technology I.P., in the framework of the Project CISUC (UID/00326/2025), and the Project CELGA-ILTEC (UID/04887/2025).

Bibliographical References

Sian Alsop and Hilary Nesi. 2009. Issues in the development of the British Academic Written English (BAWE) corpus. *Corpora*, 4(1):71–83.

Akari Asai, Jacqueline He, Rulin Shao, Weijia Shi, Amanpreet Singh, Joseph Chee Chang, Kyle Lo, Luca Soldaini, Sergey Feldman, Mike D'arcy, et al. 2024. OpenScholar: Synthesizing scientific literature with Retrieval-Augmented LLMs. *arXiv preprint arXiv:2411.14199*.

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. SciBERT: A pretrained language model for scientific text. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3615–3620, Hong Kong, China. Association for Computational Linguistics.

Douglas Biber, Susan Conrad, and Randy Reppen. 1998. *Corpus Linguistics: Investigating Language Structure and Use*. Cambridge University Press, Cambridge.

Lisa Cheung and Peter Crosthwaite. 2025. Corpuschat: integrating corpus linguistics and generative ai for academic writing development. *Computer Assisted Language Learning*, pages 1–27.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Procs. of 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. ACL Press.

Tomaž Erjavec, Darja Fišer, and Nikola Ljubešić. 2020. The KAS corpus of Slovenian academic writing. *Language Resources and Evaluation*, 55(2):551–583.

Adam Kilgarriff, Vít Baisa, Jan Bušta, Miloš Jakubíček, Vojtěch Kovář, Jan Michelfeit, Pavel

Rychlý, and Vít Suchomel. 2014. The Sketch Engine: ten years on. *Lexicography*, 1:7–36.

Iztok Kosem. 2010. *Designing a model for a corpus-driven dictionary of Academic English*. Ph.D. thesis, Aston University.

Tanara Zingano Kuhn and José Pedro Ferreira. 2020. O corpus de português escrito em periódicos-CoPEP. *DELTA: Documentação de Estudos em Lingüística Teórica e Aplicada*, 36(2):1–42.

Nikolaos Livathinos, Christoph Auer, Maksym Lysak, Ahmed Nassar, Michele Dolfi, Panos Vagenas, Cesar Berrospi Ramis, Matteo Omenetti, Kasper Dinkla, Yusik Kim, Shubham Gupta, Rafael Teixeira de Lima, Valery Weber, Lucas Morin, Ingmar Meijer, Viktor Kuropiatnyk, and Peter W. J. Staar. 2025. Docling: An efficient open-source toolkit for ai-driven document conversion.

Kyle Lo, Lucy Lu Wang, Mark Neumann, Rodney Kinney, and Daniel Weld. 2020. S2ORC: The semantic scholar open research corpus. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4969–4983, Online. Association for Computational Linguistics.

Tony McEnery, Richard Xiao, and Yukio Tono. 2006. *Corpus-Based Language Studies: An Advanced Resource Book*. Routledge, Abingdon.

Charles F. Meyer. 2002. *English Corpus Linguistics: An Introduction*. Cambridge University Press, Cambridge.

Ana Eva Miguéis and Bruno Neves. 2021. A visão dos gestores de repositórios. o caso da universidade de coimbra. *Sob a lente da Ciência Aberta: Olhares de Portugal, Espanha e Brasil*, pages 273–294.

Zhishan Mo and Peter Crosthwaite. 2025. Exploring the affordances of generative ai large language models for stance and engagement in academic writing. *Journal of English for Academic Purposes*, 75:101499.

João Rodrigues, Luís Gomes, João Silva, António Branco, Rodrigo Santos, Henrique Lopes Cardoso, and Tomás Freitas Osório. 2023. Advancing neural encoding of Portuguese with transformer AlbertinaPT-*. In *Progress in Artificial Intelligence – 22nd EPIA Conference on Artificial Intelligence, EPIA 2023, Faial Island, Azores, September 5-8, 2023, Proceedings, Part I*, volume 14115 of *LNCS*, pages 441–453. Springer.

Tarek Saier and Michael Färber. 2019. Bibliometric-enhanced arxiv: A data set for paper-based and citation-based tasks. In *BIR@ ECIR*, pages 14–26.

Joana Vieira Santos, Paulo Nunes da Silva, Marta Sitoe, Marteen Janssen, Helga Arnauth, Rute Soares, Carla Ferreira, and Conceição Carapinha. 2020. Corpus de português académicos (cpa). https://corpora.celga.iltec.pt/teitok/cpa/. Coimbra: CELGA-ILTEC.

John Sinclair. 2003. Corpus processing. In Piet van Sterkenburg, editor, *A Practical Guide to Lexicography*. John Benjamins Publishing Company, Amsterdam/Philadelphia.

Felipe Soares, Viviane Moreira, and Karin Becker. 2018. A large parallel corpus of full-text scientific articles. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC-2018)*.

Hugo Sousa, Rúben Almeida, Purificação Silvano, Inês Cantante, Ricardo Campos, and Alipio Jorge. 2025. Enhancing Portuguese variety identification with cross-domain approaches. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39:25192–25200.

Fábio Souza, Rodrigo Nogueira, and Roberto Lotufo. 2020. BERTimbau: Pretrained BERT Models for Brazilian Portuguese. In *Intelligent Systems: 9th Brazilian Conference, BRACIS 2020, Rio Grande, Brazil, October 20–23, 2020, Proceedings, Part I*, pages 403–417, Berlin, Heidelberg. Springer-Verlag.

Ross Taylor, Marcin Kardas, Guillem Cucurull, Thomas Scialom, Anthony Hartshorn, Elvis Saravia, Andrew Poulton, Viktor Kerkez, and Robert Stojnic. 2022. Galactica: A large language model for science. *arXiv preprint arXiv:2211.09085*.

Paul Thompson and Hilary Nesi. 2001. The British Academic Spoken English (BASE) corpus project. *Language Teaching Research*, 5(3):263–264.

Martin Wynne, editor. 2006. *Developing Linguistic Corpora: A Guide to Good Practice*. Oxbow Books, Oxford. Accessed: 22 October 2025.

Kristjan Žagar, Marko Ferme, Milan Ojsteršek, Mateja Jemec Tomazin, and Tomaž Erjavec. 2023. Corpus of scientific texts from the open science slovenia portal OSS 1.0. Slovenian language resource repository CLARIN.SI.

Yuxiang Zheng, Shichao Sun, Lin Qiu, Dongyu Ru, Cheng Jiayang, Xuefeng Li, Jifan Lin, Binjie Wang, Yun Luo, Renjie Pan, Yang Xu, Qingkai

Min, Zizhao Zhang, Yiwen Wang, Wenjie Li, and Pengfei Liu. 2024. OpenResearcher: Unleashing AI for accelerated scientific research. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 209–218, Miami, Florida, USA. Association for Computational Linguistics.