Language Resources and Evaluation
https://doi.org/10.1007/s10579-024-09757-5

SURVEY

Historical Portuguese corpora: a survey

Tomás Freitas Osório[1] · **Henrique Lopes Cardoso**[1]

Accepted: 10 June 2024

Abstract
This survey aims to thoroughly examine and evaluate the current landscape of electronic corpora in historical Portuguese. This is achieved through a comprehensive analysis of existing resources. The article makes two main contributions. The first is an exhaustive cataloguing of existing Portuguese historical corpora, where each corpus is meticulously detailed regarding linguistic periods, geographic origins, and thematic contents. The second contribution focuses on the digital accessibility of these corpora for researchers. These contributions are crucial in enhancing and progressing the study of historical corpora in the Portuguese language, laying a critical groundwork for future linguistic research in this field. Our survey identified 20 freely accessible corpora, comprising approximately 63.9 million tokens, and two private corpora, totalling 59.9 million tokens.

Keywords Electronic historical corpora · Historical Portuguese · Electronic resources

1 Introduction

Portuguese is the 8th most spoken language in the world Eberhard et al. (2023). Due to geopolitical reasons, the global interest in Portuguese as a foreign language and for linguistic research has increased in the last decades Zampieri and Becker (2013). Similarly, interest in computational linguistic research has also increased, giving rise to contemporary Portuguese corpora with billions of words/tokens, such as the Oscar Corpus Abadji et al. (2022) Portuguese subset with 15 billion words or the brWaC Corpus Wagner Filho et al. (2018) with 2.7 billion

✉ Tomás Freitas Osório
tomas.s.osorio@gmail.com

Henrique Lopes Cardoso
hlc@fe.up.pt

[1] Faculdade de Engenharia da Universidade do Porto (FEUP), Laboratório de Inteligência Artificial e Ciência de Computadores (LIACC), Rua Dr. Roberto Frias, Porto 4200-465, Portugal

Published online: 18 July 2024

tokens. However, historical Portuguese corpora are smaller and more dispersed; hence, this work aims to survey electronically available Portuguese historical corpora and their annotated data.

A historical corpus is a collection of documents that have been produced in a handwritten or mechanical way Philips and Tabrizi (2020). Often, these documents contain essential historical textual information about persons, places, events, and laws, but can also hold mundane details Kissos and Dershowitz (2016). Historical documents are usually stored in archives, which are physical facilities containing large collections of documents. These collections typically contain records spreading an individual or organisation's lifetime, allowing a better understanding of that entity.

To enable their processing through computational means, these documents must first be digitised, which consists of converting a physical record to a machine-readable form by scanning or photographing a document and then submitting it through special-purpose software to extract text (Philips and Tabrizi, 2020; Nguyen et al., 2022). Even though extracting text from images has been an active research field for over 30 years, results are still imperfect, especially for historical documents. In fact, historical documents with challenging layouts or deterioration signs, such as newspapers, still resist modern Optical Character Recognition (OCR) and Handwritten Text Recognition (HTR) (Rigaud et al., 2019; Rijhwani et al., 2020; Manjavacas and Fonteyn, 2022; Scheible et al., 2011). For example, historical Portuguese text can contain graphical variations that are no longer in use.

Another issue with historical texts is that they often have non-consolidated spelling Manjavacas Arevalo and Fonteyn (2022) and contain a high degree of orthographic variation, particularly in Western European languages that only established their modern spelling norms in the 18th or 19th centuries. Coupled with OCR errors exhibiting nearly random distributions, reducing word errors becomes a challenging task Manjavacas and Fonteyn (2022). Thus, post-processing OCRed text is commonly applied to help reduce text noise, increasing the informational value. However, for less-resourced languages, such as historical Portuguese, only a limited number of automated OCR text post-processing tools are available.

Having historical documents in a digital format opens an opportunity to use data mining and information retrieval techniques that help historians, archivists, and researchers explore vast amounts of documents automatically. It also makes documents readily available and easily accessible, regardless of their fragility, avoiding physical use damage, and also ensures their virtual persistence even if the original physical support is destroyed Philips and Tabrizi (2020). Furthermore, document digitisation enables better research practices, opening the possibility for virtual archives to be hyperlinked to a research project. It allows a new form of citation to help the reader access a historian's evidence of their claims Ogilvie (2016). Thus, historical corpora are of great interest to historians, literary scholars, and computational linguists (Curzan, 2000; Zampieri, 2017).

A historical corpus focuses on periods when language differed from the present day's, thus, at least one generation away. Present-day language pertains to the current state of language, encompassing vocabulary, grammar, and communication

norms in the contemporary era – i.e. language currently spoken, written, and understood by individuals Claridge (2008).

A historical corpus can be either synchronic or diachronic; the former contains documents from a specific period, such as texts from a particular century, while the latter includes documents spread in broader time frames, for example, spanning several centuries. A diachronic corpus can also be used in a synchronic way if a section of it is selected. Diachronic corpora provide valuable insights into understanding the historical evolution of languages Sánchez-Martínez et al. (2013). Just as with contemporary text corpora, historical corpora can be composed of documents from a single genre or domain (useful for more focused research) or from multiple genres or domains (enabling a wider range of linguistic research) (Claridge, 2008; Kytö, 2010).

Besides making available the raw text, historical corpora sometimes also include annotations for lemmas, parts-of-speech (PoS), or named-entities (NE), among others. However, most annotations come from automatic tagging systems designed for contemporary language. When these systems are applied to historical texts, their effectiveness is compromised due to unexpected language structures and spelling variations, resulting in lower annotation quality. For example, according to Claridge (2008), while PoS tagging accuracy for contemporary English can reach 97%, for Early Modern English material, we can observe a drop to 80%, depending on the text's date.

The two principal contributions of this survey are as follows. Firstly, the meticulous cataloguing of existing Portuguese historical corpora, which encompasses detailed descriptions of each corpus, including their linguistic period, geographic origins, and thematic content. Secondly, we highlight the digital accessibility of these corpora, detailing the platforms and repositories where researchers can readily access them. Both aspects are essential in facilitating and advancing historical Portuguese language studies research.

This paper is organised as follows: Sect. 2 discusses the importance of historical corpora, detailing their applications, potential future uses, and collections of corpora for different languages. Section 3 provides a brief overview of the evolution of the Portuguese language from its inception to its current usage. Section 4 outlines the methodology we adopted for discovering and processing historical Portuguese corpora. Section 5 explores the collected corpora and is divided in two: Sect. 5.1 focuses on diachronic corpora, while Sect. 5.2 concentrates on synchronic corpora. Section 6, analyses and compares the collected corpora. Finally, Section 7 offers concluding observations and remarks.

2 Digital historical Corpora

Researchers working with historical textual material are well aware analysing and interpreting them cannot be approached with present-day intuitions Tahmasebi and Risse (2017) due to the constant evolution of languages on all levels of linguistic structure. Language changes over time and exhibits variations across regions and social groups; pronunciation evolves, new words are borrowed or invented, the

meaning of old words drifts, and morphology can grow or decay. These changes can be attributed to many factors, such as contact between different communities, during language learning, social differentiations, or other natural processes of language usage (Zampieri and Becker, 2013; Alatrash et al., 2020; Schieffelin and Ochs, 1986; Fishman, 1964; Kerswill, 2006; Blank, 1999; Hamilton et al., 2016).

In the last two decades, interest in diachronic language change has emerged, driven by technological advancements such as the digitisation of historical texts, enhanced computational capabilities, and the availability of historical corpora tailored for diachronic studies (Tahmasebi et al., 2019; Tang, 2018; Bowern, 2019). This field aims to detect and substantiate general trends in language development with changes manifesting in various aspects like lexicon, grammatical structures, and textual stylistics (Alatrash et al., 2020; Ciobanu et al., 2013; Popescu and Strapparava, 2015; Kutuzov et al., 2018).

Historical corpora in digital format ease statistical analysis of relationships between linguistic phenomena and linguistic or extra-linguistic factors in language change. Additionally, historical corpora serve as a bridge between past and present language studies, offering a platform to test modern sociolinguistic theories. This approach allows for a systematic and empirical examination of recent and ongoing language changes, avoiding dependence on anecdotal observations (Kytö, 2010; Zampieri, 2017). However, for these studies, access to large amounts of corpora from different registers and levels of language is necessary Finatto et al. (2018), but is not always the case.

Extensive historical corpora also enable the training and utilisation of language models (LMs) for predicting the likelihood of specific word or letter sequences within a particular language or sublanguage and providing valuable insights into the language's evolution over time when applied to texts from different periods. With their ability to assign accurate probabilities to text sequences, LMs can also be effectively employed in developing spelling and normalisation tools. By addressing the challenges posed by varied spellings in historical texts, we can, for instance, make search processes within these texts more efficient Pettersson and Megyesi (2018).

Open access to large amounts of historical corpora is essential for historical, linguistic, and computational research. These collections, spanning extensive time frames, provide valuable linguistic data, facilitating empirical studies on language evolution, syntactic constructions, and semantic transformations. Through quantitative analysis, researchers can identify patterns and changes, contributing to a scientifically robust understanding of each language's historical progression. Furthermore, the online availability of these corpora encourages collaborative research and interdisciplinary inquiry, promoting transparency and scholarly exchange within linguistic research. However, these extensive collections exist for highly-resourced languages like English, but for less-resourced languages (such as Portuguese), they are scarce.

In the realm of English historical corpora, there are currently thirty to forty English historical corpora available or in progress, summing more than 630 million words Kytö (2010), including the 400 million word Corpus of Historical American English (COHA) Davies (2012), the 100 million word Time Corpus Time Corpus (2024) and 52 million word Old Bailey Corpus Old Bailey Corpus (2024). The

biggest, COHA, is a widely used resource for investigating lexical, syntactic, and semantic changes. COHA was developed by Brigham Young University and comprises carefully selected historical English texts from newspapers, magazines, and fiction and nonfiction books published between 1810 and 2009. Penn Corpora of Historical English Kroch (2020) contains British English prose texts from the earliest Middle English to the First World War, totalling 4,4 million tokens. Some other available corpora include the Early English Books Online corpus (2024) (1473–1700), the Evans Early American Imprints Collection (2024) (1639–1800), Eighteenth Century Collections Online Eighteenth Century Collections Online (2024) (1701–1800), the Corpus of Late Modern English Texts (2006) (1710–1920), Early English Correspondence Corpus (2024) (1400–1800), Early English Medical Writing (2024) (1375–1800), Corpus of English Dialogues Culpeper and Kytö (1997) (1560–1760), the Lampeter Corpus of Early Modern English Tracts Lampeter Corpus of Early Modern English Tracts (2024) (1640–1740) and Zurich English Newspaper Corpus (2024) (1661–1791).

Spanish also has several historical corpora available, such as the *Oralia diacrónica del español* (ODE) corpus, encompassing manuscripts written from 1492 to the late 19th century from the old Kingdom of Granada and the northern half of Spain Calderón Campos and Díaz-Bravo (2021). ODE belongs to the international network *Corpus Hispánico y Americano en la Red: Textos Antiguos* (CHARTA), which aims to create a diachronic corpus that includes Spanish texts from Spain and Hispanic America—the texts in CHARTA span from 1200 to 1800, amounting to approximately 1,3 million words Calderón Campos and Díaz-Bravo (2021). There is also the *Corpus Léxico de Inventarios* (CorLexIn) (2024) that contains texts from the 17th century, totalling around 738 thousand words from almost all the Spanish provinces and some American regions. The *Corpus diacróico y diatópico del español de América* (CORDIAM) (2024) have texts that span from 1494 till 1905 from Latin America and contains around 9.5 million words. The Post Scriptum Vaamonde et al. (2014) corpus is a collection of private letters from Portugal and Spain from the 16th to 19th century, where the Spanish version has around 871 thousand tokens. The IMPACT-es Sánchez-Martínez et al. (2013) corpus encompasses over one hundred books between 1481 and 1748 of a diverse range of authors and genres such as prose, theatre, and verse, totalling around 8 million words. And the *Compania de Documentos Espaoles Anteriores* corpus (CODEA) [51] has more than 1,500 documents from the 12th to 17th centuries, mainly from archives such as letters and administrations, legal or ecclesial records.

As for historical Dutch, there are several corpora. For instance, Delpher (2024), a database managed by the Koninklijke Bibliotheek, comprises over 130 million scanned and digitised pages of historical newspapers and books from 1618 to the end of the 20th century. The Digital Library of Dutch Literature (DBNL) (2024), a collaborative effort of Dutch and Flemish libraries with over 5 million digitised pages, includes literature dating from the Middle Ages to present times. While Delpher contains OCR text of varying quality, the DBNL, due to meticulous digitisation, generally offers high-quality transcriptions Manjavacas Arevalo and Fonteyn (2022). There is also the Digital Compilation Corpus of Historical Dutch (DCCHD) Coussé (2011) that contains two subcorpora, one from the Middle Ages, covering

Flanders, Brabant, and Holland from 1250 until 1800, and another with texts from Holland from the late sixteenth century to present times.

Concerning historical French corpora, there is the FREnch Early Modern (FREEM) corpus Gabay et al. (2022) that has 185 million tokens, as well as annotations for lemmas, PoS tagging, linguistic normalisation, and NEs. FREEM contains more than 22 corpora, including the FRANTEXT intégral, the most extensive collection of French texts from 1500 to 1800, though only a limited portion is openly accessible through FRANTEXT Démonstration. The *ANalyse auTOmatique et NumérisatiOn des MAZarinades*[1] (Antonomaz) includes more than six hundred texts written between 1648 and 1653. The Paris Speech in the Past corpus Marot (2001) has documents from 1296 till 1790. There is also the *Corpus Électroniques de la Première Modernité*[2] (CEPM; 17th century), *Bibliothèques virtuelles humanistes*[3] (BVH, 16th century), Corpus Descartes[4] (CD, works of René Descartes), *Mercure Galant Project*[5] (MGP, 1672-1710), among many other corpora.

Regarding historical German corpora, there is the GerManC corpus Scheible et al. (2011), focusing on Early Modern German from 1650–1800, which encompasses nine genres, subdivided into three 50-year periods, covering five major dialectal regions of the German Empire at that time, comprising 900,000 words. The *Deutsches Textarchiv* (DTA) Geyken et al. (2010) corpus aims to digitise a comprehensive selection of printed works in modern New High German Language from approximately 1600 to 1900, with 2,422 online texts, totalling more than 650,000 digitised pages and roughly 1,1 billion characters. The *Referenzkorpus Mittelhochdeutsch* (ReM) corpus Klein and Dipper (2016) contains diplomatically transcribed and annotated texts of Middle High German (1050-1350) with approximately 2 million words.

It should be noted that the corpora identified for these languages do not represent an exhaustive list. Additionally, this section does not cover corpora for other languages, as this article's primary focus is Portuguese.

3 The Portuguese language & orthographic agreements

Portuguese is a Western Romance language originating in the Iberian Peninsula and has approximately 250 million native speakers and 24 million second-language speakers Eberhard et al. (2023). Portuguese is an official language in Angola, Brazil, Cabo Verde, Mozambique, Guinea-Bissau, Portugal, and São Tomé and Príncipe Comunidade dos Países de Língua Portuguesa (2022), and a co-official language in East Timor, Equatorial Guinea and Macau. Several orthographic agreements have

[1] www.cahier.hypotheses.org/antonomaz
[2] www.cepm.paris-sorbonne.fr
[3] www.bvh.univ-tours.fr
[4] www.unicaen.fr/puc/sources/prodescartes/
[5] www.obvil.sorbonne-universite.fr/corpus/mercure-galant/

been proposed to unify orthography over the years, a process towards language regularisation Carvalho and Cabecinhas (2013).

The first unification of the Portuguese language was made in 1671 by João Franco Barreto, who published *Ortografia da Língua Portugueza* Barreto (1671) (Orthography of the Portuguese Language) after the monarchy restoration so Portuguese would become an autonomous language Teyssier (2001). In 1739, João Moraes Madureira Feijó published *Orthographia ou Arte de escrever e pronunciar com acerto a língua portugueza* Feijó (1739) (Orthography or the Art of writing and pronouncing the Portuguese language correctly), revising orthography and also instructing professors on how Portuguese should be written.

In 1900, Aniceto dos Reis Gonçalves Viana conducted a questionnaire called *Ortografia Nacional. Simplificação e uniformização sistemática das ortografias portuguesas* Reis Gonçalves Viana (1904) (National Spelling. Simplification and systematic standardisation of Portuguese spellings), collecting the Portuguese orthography rules. Ten years later, after the implementation of the Republic, a Commission for Spelling Reform was appointed to establish a simplified orthography to be used in official publications and education based on that questionnaire. This reform was profound and completely changed the aspect of the written language, bringing it very close to nowadays. Nevertheless, Brazil did not accept this reform and remained using the old pseudo-etymologic orthography since Portugal made this reform alone. In 1915, the Brazilian Academy of Letters decided to harmonise the spelling with Portugal by establishing the first unofficial spelling agreement; however, it was revoked in 1919. Similarly, in 1945, a new spelling agreement was created but was only adopted in Portugal.

More recently, in 1990, a new Orthographic Agreement of the Portuguese Language was established, consisting of an international treaty aiming to unify Portuguese orthography used by all Portuguese-speaking countries, being signed by official representatives from Angola, Brazil, Cabo Verde, Guinea-Bissau, Mozambique, Portugal, and São Tomé and Principe, in Lisbon. In 2009, a new reform was made to unify the vocabulary used in Brazil and Portugal, making its usage mandatory from 2013 onward, which affected about 0,5% of Brazil's vocabulary and 2% of Portugal's. Yet, the impact is much higher in practice since the words that changed orthography were among the most frequently used ones Ricardo (2009).

4 Methodology

In our pursuit to comprehensively catalogue Portuguese historical corpora, we extended beyond the traditional academic sources such as Scopus[6] and Web of Science[7] since many corpora lack associated published articles. Therefore, the search

[6] www.scopus.com
[7] www.webofscience.com

was broadened to include the PORTULAN CLARIN[8] repository, Linguateca,[9] and research groups working on historical text. Additionally, general search engines like Google were employed.

PORTULAN CLARIN is a Research Infrastructure for the Science and Technology of Language, belonging to the Portuguese National Roadmap of Research Infrastructures of Strategic Relevance and part of the international research infrastructure CLARIN ERIC.[10] Linguateca is a distributed language resource centre for Portuguese that was launched due to the Computational Processing of Portuguese project, an initiative by the Portuguese Ministry of Science and Technology to boost the computational processing of the Portuguese language.

In our comprehensive search, we focused on identifying linguistic resources, including corpora, documents, books, and collections containing texts predominantly from before the twenty-first century.

Many of the collected corpora contain meta-information such as the number of documents per year/decade/century, total number of words/tokens, document type, and genre, among others. However, we observed that the methodologies for calculating the word or token counts are often not explicitly detailed. To enable accurate and meaningful comparisons across various corpora, we clean the text in all the datasets, ensuring the token count is comparable. Therefore, we mainly remove text that does not belong to the original documents. This includes eliminating HTML tags, meta-information about the document or annotations, and similar non-essential elements. We also remove punctuation from the text. Finally, we employ the NLTK[11] word tokenizer, a well-established natural language processing tool, to count the number of words per document.

This survey additionally analyses the accessibility of corpora, examining aspects such as licensing conditions and ease of access.

5 Historical Portuguese corpora

Conducting this survey, we identified twenty-two corpora, as detailed in Table 1. We gained access to twenty of these corpora, containing historical text spanning from the 13th to the 21st century. The texts originated from authors with diverse backgrounds, including medics, clerics, soldiers, writers, and others, therefore varying literacy levels, and encompass a spectrum of genres.

Figure 1 illustrates the distribution of tokens and documents per century for all corpora presented in Table 1, to which we had access, with the exceptions of the LT Corpus and CEDOHS. The former comprises a single document that combines various texts, and the latter is excluded due to the absence of date information for individual documents. Figure 1 also serves the purpose of distinguishing between

[8] www.portulanclarin.net
[9] www.linguateca.pt
[10] www.clarin.eu/resource-families/historical-corpora
[11] www.nltk.org

Table 1 Comprehensive repository of historical Portuguese datasets.

Dataset	N. Tokens	Time scope	Source	Open access
C. Português Davies (2006)	45 M	13th–19th	ad hoc	✗
Gutenberg Project Gutenberg (2023)	19,4 M	14th–20th	ad hoc	✓
VERCIAL (2022)	14,9 M	16th–20th	L	✗
DiaPT Pichel Campos et al. (2018)	10.74M	12th–21th	journal	✓
GMHP (2022)	7,03 M	15th–20th	ad hoc	✓
ELTeC Santos (2021)	6,55 M	1840–1920	CA	✓
Colonia Zampieri and Becker (2013)	4,58 M	16th–20th	journal	✓
BDCamões Grilo et al. (2020)	3,89 M	15th–21th	journal	✓
Tycho Brahe Galves (2018)	3,37 M	14th–19th	journal	✓
CIPM Xavier (2016)	3,32 M	13th–16th	journal	✓
LT Corpus Généreux et al. (2012)	1,48 M	19th–20th	journal	✓
Pessoa ARQUIVO PESSOA (2023)	1,20 M	20th	ad hoc	✓
PS Corpus Vaamonde et al. (2014)	763k	16th–19th	journal	✓
WochWel Pereira (2015)	465k	14th–16th	journal	✓
PPM As Memórias Paroquiais de 1758 (2023)	272k	18th	ad hoc	✓
CEDOHS Corpus Eletrônico de Documentos Históricos do Sertão (2023)	229k	16th–20th	ad hoc	✓
CHLMP Corpus Histórico da Linguagem da Medicina em Português (2023)	192k	18th	ad hoc	✓
AdA Arquivo dos Açores (2023)	182k	19th–21th	PC	✓
Fly Corpus Gomes et al. (2012)	131k	20th	journal	✓
EPISA Falcão et al. (2022)	69k	20th	ad hoc	✓
CTA CTACorpus (2022)	29,5k	13th–16th	ad hoc	✓
CDAIII Chancelaria de D. Afonso III (2023)	18k	13th	PC	✓

These datasets originate from diverse sources, ranging from ad hoc collections to curated archives, including journals and reputable platforms such as PORTULAN CLARIN (PC), Cost Actions (CA), and Linguateca (L). For a more detailed overview, see Table 22

synchronic corpora—represented with a single data point—and diachronic corpora—where the data points are connected by dashes or lines, providing a visual representation of their respective time frames.

This section is divided into two parts: one focusing on diachronic corpora and the other on synchronic corpora. This analysis explores the linguistic characteristics and features within each corpus. Together, these sub-sections provide a comprehensive overview of the linguistic landscape, contributing to a nuanced understanding of the temporal and linguistic dimensions of the historical Portuguese language corpora.

5.1 Diachronic corpora

Diachronic corpora comprise documents that cover wide time periods. The following ones have been identified: *Corpus do Português* (CdP) Davies (2006), *Project Gutenberg* (PG) Project Gutenberg (2023), *Vercial* VERCIAL (2022), *DiaPT*

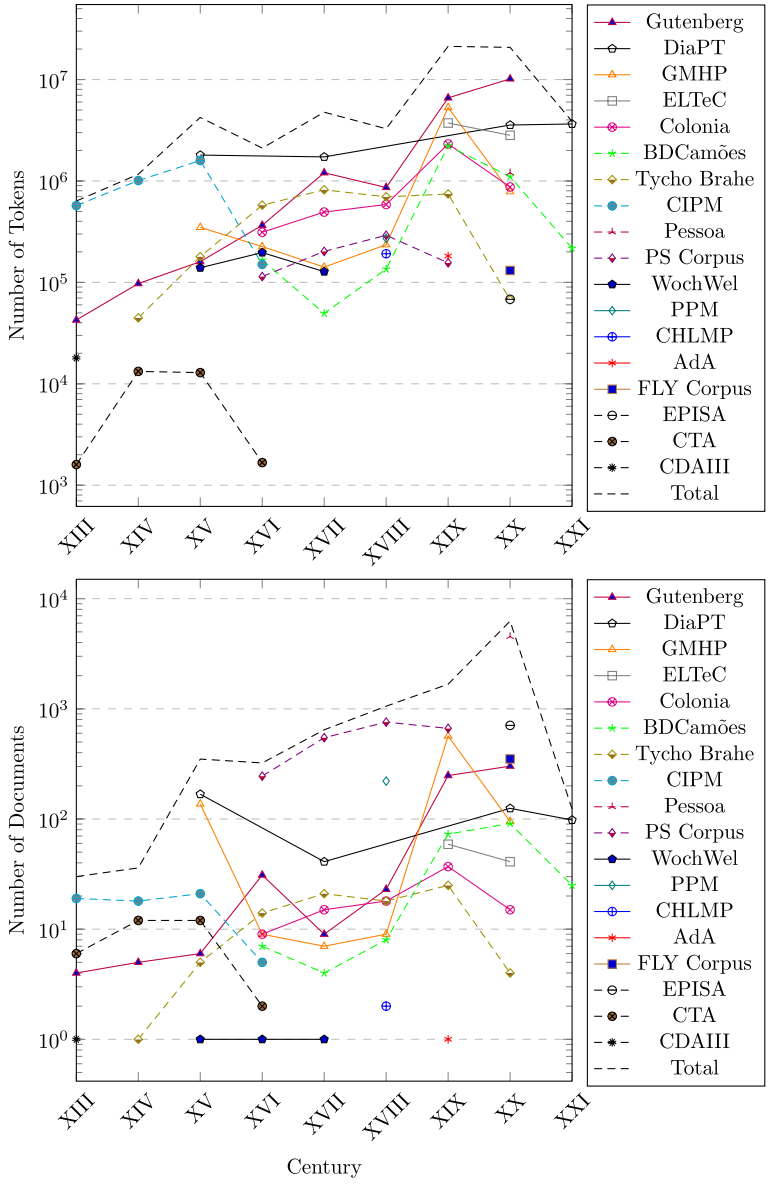

Fig. 1 Distribution of tokens and documents across centuries in each corpus

Pichel Campos et al. (2018), *Grupo de Morfologia Histórica do Português* GMHP (2022) (GMHP), *European Literary Text Collection* Santos (2021) (ELTeC), Colonia (Col) Zampieri and Becker (2013), *BDCamões* (BDC) Grilo et al. (2020), *Tycho Brahe* (TB) Galves (2018), *Corpus Informatizado do Português Medieval* (CIPM) Xavier (2016), *Literary Text Corpus* (LT) Généreux et al. (2012), *Post-Scriptum* (PS) Vaamonde et al. (2014), *Word Order and Word Order Change in Western*

Table 2 Summary of available metadata for each diachronic corpus

Corpus	PG	DiaPT	GMHP	ELTeC	Col	BDC	TB	CIPM	LT	PS	WochWel	CEDOHS	CTA
Title	✓	✗	✗	✓	✓	✓	✓	✓	✗	✗	✓	✗	✓
Author	✓	✗	✓	✓	✓	✓	✓	✗	✗	✓	✓	✗	✓
Author gender	✗	✗	✗	✓	✗	✗	✗	✗	✗	✓	✗	✗	✗
Author birth	✗	✗	✗	✓	✗	✗	✓	✗	✗	✗	✗	✗	✗
Author death	✗	✗	✗	✓	✗	✗	✗	✗	✗	✗	✗	✗	✗
Author residence	✗	✗	✗	✗	✗	✗	✗	✗	✗	✓	✗	✗	✗
Author Job	✗	✗	✗	✗	✗	✗	✗	✗	✗	✓	✗	✗	✗
Author social stat	✗	✗	✗	✗	✗	✗	✗	✗	✗	✓	✗	✗	✗
Pub. date	✓	✗	✗	✓	✓	✓	✗	✗	✗	✓	✓	✗	✓
Doc. genre	✓	✗	✓	✓	✗	✓	✓	✗	✗	✓	✓	✓	✓

Table 3 URLs where the diachronic corpus data were sourced

Corpus	Obtained	URL
PG	Scrape	https://www.gutenberg.org/browse/languages/pt
DiaPT	Download	https://github.com/gamallo/Perplexity
GMHP	Download	http://www.usp.br/gmhp/CorpI.html
ELTeC	Download	https://zenodo.org/records/4288235
Col	Download	https://www.kaggle.com/datasets/rtatman/colonia-corpus-of-historical-portuguese
BDC	Download	https://hdl.handle.net/21.11129/0000-000D-F89B-D https://hdl.handle.net/21.11129/0000-000D-F8AB-B https://hdl.handle.net/21.11129/0000-000D-F8AA-C https://hdl.handle.net/21.11129/0000-000D-F8A8-E
TB	Download	https://www.tycho.iel.unicamp.br/corpus/en/
CIPM	Download	https://hdl.handle.net/21.11129/0000-000D-F934-0 https://hdl.handle.net/21.11129/0000-000D-F91D-B
LT	Download	https://hdl.handle.net/21.11129/0000-000B-D33D-3
PS	Download	https://hdl.handle.net/21.11129/0000-000D-F91F-9 https://hdl.handle.net/21.11129/0000-000D-F924-2
WochWel	Download	http://alfclul.clul.ul.pt/wochwel/Parsedannotation.html
CEDOHS	Download	http://www5.uefs.br/cedohs/view/coletaneas.html
CTA	Scrape	www.teitok.clul.ul.pt/teitok/cta/index.php?action=textos

The **Obtained** column categorises the method used for data collection: **Download** indicates direct data retrieval. In contrast, **Scrape** uses a custom script to extract information from the specified URLs

European Languages (WochWel) Pereira (2015), *Corpus Eletrônico de Documentos Históricos do Sertão* (CEDOHS) Corpus Eletrônico de Documentos Históricos do Sertão (2023), and *Corpus de Textos Antigos* (CTA) CTACorpus (2022). Table 2 presents the metadata associated with these corpora, while Table 3 lists the URLs from which diachronic corpus data were obtained.

5.1.1 Gutenberg corpus

Project Gutenberg Project Gutenberg (2023), the oldest digital library, was founded in 1971 by American writer Michael S. Hart. This archive distributes eBooks that have been digitised and have copyright clearance under United States copyright law; thus, Project Gutenberg does not claim new copyright on titles it publishes and encourages their free reproduction and distribution.

Most documents available on Project Gutenberg consist of books or individual stories within the public domain, offering free accessibility in various formats, including plain text, HTML, PDF, EPUB, MOBI, and Plucker, whenever possible. While most documents are in English, other languages, such as French, German, Finnish, Dutch, Italian, and Portuguese, are present.

Concerning Portuguese literature, Project Gutenberg, at the time of writing, contains 628 pieces from the 13th to the 20th century. Yet, most pieces fall between the 19th and 20th centuries, as illustrated in Fig. 1.

Table 4 CTA metadata example (partially translated to English for better understandability)

Title	*Do Dia do Juízo*
Autor	Unknown
Translation/editing	Unknown, probably translated from Latin
Date of redaction/translation	1450
Date of testimony	1431–1446
genre	Education and spirituality
Stored	*Biblioteca Nacional de Lisboa*

Table 5 CTA documents and tokens genre distribution

Genre	N. Docs	N. Tokens
Hagiography	21	17k
Edu and spirituality	7	5k
Spirituality	2	6k
Novel	1	0.4k
Unknown	1	0.7k
total	32	29.5k

Project Gutenberg also provides some meta-information for each text piece, such as the author, title, publication date, and genre.

5.1.2 Corpus de Textos Antigos

Corpus de Textos Antigos CTACorpus (2022) was created by the Center of Linguistics of the University of Lisbon (CLUL) and contains 32 semi-diplomatic texts written originally or translated to Portuguese between the 13th and 16th centuries. Additionally, meta-information is provided for each document regarding the author, redaction date, title and archive where it is stored, among others, as exemplified in Table 4. Table 5 presents the distribution of the number of documents and tokens per genre.

As can be observed in Fig. 1, CTA exhibits longer documents from the 14th and 15th centuries.

5.1.3 Post-scriptum Portuguese corpus

The *Post-Scriptum* corpus Vaamonde et al. (2014) was compiled by CLUL and contains Spanish and Portuguese private letters from the 16th to the 19th century. The Portuguese corpus partition comprises 2,215 letters, with the majority archived in the *Arquivo Nacional da Torre do Tombo*[12] (Lisbon, Portugal); however, documents from the archives of Évora, Braga, Porto (Portugal), and Goa (India) are also

[12] www.digitarq.arquivos.pt

Table 6 Socioeconomic backgrounds of authors and recipients in the PS corpus letter collection

Social	Count
Ordinary	1089
Unknown	894
Ecclesiastical	509
Military	170
Nobility	155
Inquisitorial	132
Universitary	27
Knightly orders	14
Ordinary	1
Slave	1
Nobility	1
Total	2993

Table 7 Gender of authors and recipients in the PS corpus letter collection

Gender	Count
Male	2467
Female	502
Unknown	24
Total	2,993

Table 8 Examples of PS Corpus Annotations: Original Form, Normalised Version, and Lemmatised Representation

Original	Minha	querida	**thia**	Recebi	a	sua	carta	
Normalised	Minha	querida	**Tia**	Recebi	a	sua	carta	
Lemma	meu	querido	tia	receber	a	seu	carta	
Original	mas	eu	estou	**tam**	faminto	**q**	nem	**cõ**
Normalized	Mas	eu	estou	**tão**	faminto	**que**	nem	**com**
Lemma	mas	eu	estar	tão	faminto	que	nem	com

included. As shown in Fig. 1, the Portuguese partition of the PS corpus exhibits an increase of documents towards the later centuries, similarly to token-wise, with the exception of the final century.

Letters within this collection are from authors and recipients from different social backgrounds, including masters and servants (described in Table 6), adults and children, men and women (shown in Table 7), as well as individuals such as thieves, soldiers, artisans, priests, political activists, and other social agents. The subjects addressed in these letters typically revolve around everyday issues of past centuries and often exhibit an (almost) oral rhetoric. Consequently, many letters feature numerous orthographic errors and abbreviations.

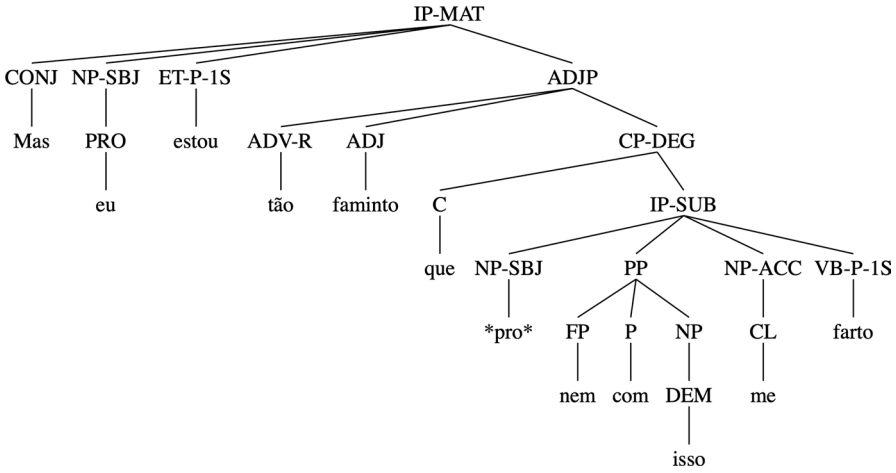

Fig. 2 Example of parse tree with syntactic annotation from the PS Corpus

Table 9 BDCamões documents and tokens genre distribution

Genre	N. Docs	N. Tokens
Tale	93	652 k
Chronicle	26	598 k
Novel	25	1,27 M
Short story	20	282 k
Poetry	18	293 k
Drama	11	79 k
Essay	8	532 k
Narrative	1	52 k
Anthology	1	87 k
Allegory	1	12 k
Letter	1	9 k
Scripts	1	6 k
Memoirs	1	17 k
Other	1	7 k
total	208	3,89 M

The PS corpus provides access to the original transcriptions of letters, PoS annotations (represented in Fig. 2), and a standardised version that eliminates abbreviations and adheres to contemporary Portuguese norms, as exemplified in Table 8. The syntactic annotations were obtained using the Penn-Helsinki annotation system, thus following an annotation standard shared by other Portuguese diachronic syntax projects, such as WochWel Pereira (2015) and Tycho Brahe Galves (2018).

These letters can be a working tool for different humanistic studies, particularly those closely tied to modern history, cultural history, textual criticism, diachronic

Table 10 BDCamões annotation example in CoNLL Hajič et al. (2009) format

Form	organismo	que	sente	musica	mysteriosa
Raw	organismo	que	sente	musica	mysteriosa
Lemma	ORGANISMO	–	SENTIR	MUSICO	MYSTERIOSO
PoS	CN	REL	V	CN	ADJ
Infl	ms	–	pi-3 s	fs	fs
NE	O	O	O	O	O
DepRel	M-PRED	SJ-ARG1	M-PRED	DO-ARG2	M-PRED
Parent	7	12	10	12	14
UDepRel	MOD	NSUBJ	AMOD	DOBJ	AMOD
UParent	7	12	10	12	14
Space	LR	LR	LR	LR	LR

linguistics and corpus linguistics. Additionally, comprehensive biographical data about the authors and recipients, encompassing details such as their lifestyles, social interactions, gender, birthplace, and more, is also accessible.

5.1.4 BDCamões corpus

The *BDCamões* corpus Grilo et al. (2020), compiled by the Natural Language and Speech Group of the University of Lisbon and the *Instituto Camões I.P.*, contains 208 lary documents from 83 authors in 14 genres, such as novels, chronicles, poems, short stories, among others, as shown in Table 9. Spanning from the 15th to the 21st centuries, the corpus features texts that adhere to diverse orthographic traditions and standards. As Fig. 1 illustrates, the BDCamões corpus exhibits a similar number of documents and tokens among the 16th, 17th, and 18th centuries, with a notable increase in the subsequent periods, namely the 19th and 20th, and decreases in the 21st century.

BDCamões exclusively holds complete documents rather than fragmented ones to increase its text quality. It integrates six texts from Tycho Brahe, totalling 159,000 words, and 23 texts from the LT Corpus, amounting to 897,000 words. The remaining documents were obtained from the Digital Library of Camões,[13] with transcriptions obtained through OCR and corrected manually by two linguists. The texts were transcribed literally, thereby preserving any typographic errors present in the original versions.

BDCamões also provides automatically generated linguistic annotations using LX-Suite Branco and Silva (2006), including PoS tagging, morphology, NEs, syntactic analysis in the form of grammatical dependency graphs, and semantic roles, as shown in Table 10, in CoNLL Hajič et al. (2009) format.

[13] www.instituto-camoes.pt/en/activity-camoes/online-services/service-desk

Table 11 Colonia annotation example

Original	Da	vida	o	indício	a	superfície	emerge
Lemma	de	vida	o	indício	a	superfície	emergir
PoS	PRP+DET	NOM	DET	NOM	PRP	NOM	V

Table 12 Tycho Brahe documents genre distribution

Genre	N. Docs
Theatre	26
Letters	26
Narrative	22
Journals	9
Dissertations	7
Grammatical	3
Acts	2
total	95

5.1.5 Colonia

The *Colonia* corpus Zampieri and Becker (2013), developed at the University of Cologne, comprises textual pieces from the 16th to early 20th century. The material was collected from three main sources: *Dominio Público*,[14] a digital library of non-copyrighted media maintained by the Brazilian Ministry of Education, and texts from two additional Portuguese historical corpora, GMHP (2022) and Tycho Brahe Galves (2018).

The Colonia corpus contains 100 documents, organised into five sub-corpora by century, and maintains an equitable distribution between Portuguese and Brazilian documents. Unfortunately, we only had access to 94, totalling 4,58 million tokens. Similarly to other diachronic corpora, as seen in Fig. 1, the number of textual pieces and tokens increases towards the later centuries, except for the 20th century, which decreases compared to the previous century.

The corpus also includes PoS tagging and lemma annotations, which have been annotated using the IMS Stuttgart's TreeTagger Schmid (1994) along with a parameter file designed for Portuguese. TreeTagger is a language-independent probabilistic tagger that arranges annotated data in a three-column format (original token, PoS tag and lemma), as shown in Table 11. Additionally, some texts are orthographically normalised, meaning they adhere to contemporary Portuguese norms, which were already compiled before Colonia.

Colonia has been employed in diverse research applications, including temporal text classification (Niculae et al., 2014; Zampieri et al., 2016), diachronic

[14] www.dominiopublico.gov.br

Table 13 Tycho Brahe contemporary normalisation (CN) example

CN	não	se acharia	o	hospital	com que	satisfazer	o	sustento
Original	não	seachria	o	Hospital	comque	satisfazer	o	sustento

Table 14 GMHP documents and tokens genre distribution

Genre	N. Docs	N. Tokens
Poetry	284	209 k
short story-chronicle	201	449 k
Songs	136	305 k
theatre	90	968 k
Novel	81	4, 6 M
Prose	31	388 k
Medieval	1	44 k
total	824	7 M

morphology Nevins et al. (2015), lexical semantics Santos and Mota (2015) and lexicographical evaluation Bick and Zampieri (2016).

5.1.6 Tycho Brahe parsed corpus of historical Portuguese

The *Tycho Brahe* Corpus Galves (2018), compiled at Campinas University, includes 95 Portuguese texts from over 50 authors born between the 14th and 19th centuries. This collection spans various literary forms, including letters, journals, acts, books, and dissertations, among other types. Tycho Brahe has been processed using the Penn-Helsinki annotation system, having part-of-speech tagging annotations on 60 documents and syntactic annotations on 31, which are represented in the same way as in Fig. 2 from PS Corpus. As illustrated in Fig. 1, Tycho Brahe predominantly focuses on documents from the 16th to the 19th centuries.

The Tycho Brahe corpus, comprising 3,37 million tokens, features diverse genres, as shown in Table 12, including twenty-six documents tagged as letters. Among these, three exhibit numerous orthographic deviance and abbreviations, amounting to 131 thousand tokens and being labelled as written by various authors between the 1800 s and 1900 s. Additionally, two documents tagged as acts and labelled as written by multiple authors in the same period follow the same issue, summing up to 62 thousand tokens. On the other hand, documents categorised as journals and labelled as written by various authors do not exhibit such issues. Finally, documents have a contemporary normalisation of the original text, as exemplified in Table 13.

The Tycho Brahe corpus continually adds new textual pieces, so the number of pieces might differ as of the time of writing.

Table 15 ELTeC annotation example

Original	Duque	de	Bragança	fizera	construir	no
Lemma	Duque	de	Bragança	fazer	construir	em+o
PoS	PROPN	PROPN	PROPN	VERB	VERB	ADP+DET
NE	Person	Person	Person	-	-	-

Table 16 ELTeC metadata example

Author	Castro e Almeida, Virgínia de
Title	Inocente
Birth	1874
Death	1945
Gender	Female
First-edition	*Bibliotrónica Portuguesa*

Table 17 ELTeC gender distribution

Gender	Count
Male	83
Female	17
Total	100

5.1.7 Literary text corpus

The *Literary Text* Corpus Généreux et al. (2012), compiled by CLUL, contains 70 copyright-free classics (61 from Portugal and nine from Brazil) spanning the mid-19th (40 pieces) to the mid-20th centuries (30 pieces).

With 1.48 million tokens, the LT Corpus is a subset of the Reference Corpus of Contemporary Portuguese (CRPC) Généreux et al. (2012), a project developed over two decades. CRPC covers the chronological period between 1970 and 2008, although some texts from 1850 onward are also included.

The LT Corpus consolidates all textual information into a single text file, making it impractical to segregate by epoch. However, details regarding the selected pieces and their respective publication dates are provided. Moreover, the LT Corpus includes automatically annotated PoS. The texts were cleaned using NCleaner Evert (2008), reducing the initial corpus size by around 28%.

5.1.8 GMHP corpus

Grupo de Morfologia Histórica do Português is an interdisciplinary group from the University of São Paulo, established in 2005, focusing on diachronic studies of flexion, derivation and composition of the Portuguese language. As such, they have created the GMHP GMHP (2022) corpus containing 824 documents, divided into six

Table 18 DiaPT annotation example. **PN** line represents the spelling-normalised version based on phonetics

Original	militar	de	31	de	janeiro	de	1891	caracterisou-se
PN	militar	de		de	janeiro	de		karacterisou se

epochs from the 15th to the 20th centuries. The textual pieces are categorised across seven genres; their distribution is shown in Table 14.

As shown in Fig. 1, the 15th, 19th, and 20th centuries exhibit the highest document count. However, in terms of token count, there is a comparable number from the 15th to the 18th centuries, followed by a substantial increase in the subsequent centuries.

5.1.9 European literary text collection

The *European Literary Text Collection* Santos (2021), funded by the COST Action: *Distant Reading for European Literary History*, comprises texts in 10 languages, including Portuguese. Each language has 100 original novels written between 1840 and 1920 with comparable internal structures. The corpus also contains annotation for morphological tagging and NEs, as shown in Table 15, using PALAVRAS (Bick, 2006, 2014).

ELTeC includes a range of meta-information for each document, as outlined in Table 16. A particular aspect, detailed in Table 17, is the author's gender distribution, which reveals a significant skew towards male authors.

5.1.10 Corpus informatizado do Português medieval

The initial development of *Corpus Informatizado do Português Medieval* Xavier (2016) dates back to 1993, overseen by the Linguistics Research Centre of NOVA University Lisbon, contains 63 texts from the 13th to 16th century. The textual pieces can be religious, notarial, or literary texts in prose or verse, written in medieval Portuguese. In addition to transcriptions, CIPM also provides PoS annotations for six documents.

As illustrated in Fig. 1, CIPM predominantly has documents and tokens between the 13th and 15th, reducing in the 16th century. This corpus exhibits a lack of systematic punctuation and features a significant degree of orthographic and morphological variation, along with editorial annotations.

5.1.11 DiaPT

The *DiaPT* Pichel Campos et al. (2018) corpus, a collaborative effort between the University of Santiago, the University of the Basque Country, and Imaxin Software, offers a comprehensive collection of literature and nonfiction spanning from the 12th to the 20th century. This corpus is divided into six historical periods, encompassing 432 documents and 10.74 million tokens.

For DiaPT compilation, resources from CIPM, Project Gutenberg, Wikisource,[15] OpenLibrary,[16] Tycho Brahe corpus, Domínio Público,[17] Arquivo Pessoa, Linguateca, CTA and Colonia were used, only being considered documents which had the original spelling.

DiaPT also provides a spelling-normalised version based on phonetics, as exemplified in Table 18; for this, all Portuguese historical periods were transliterated into Latin script and then normalised using a generic orthography approximation to phonological issues. This version is encoded using 34 symbols, representing ten vowels and 24 consonants, designed to cover the most commonly occurring sounds, including several consonant palatalisations and various vowel articulations.

5.1.12 Corpus Eletrônico de documentos Históricos do Sertão

The *Corpus Eletrônico de Documentos Históricos do Sertão* (2023) was created by the Portuguese Language Studies Centre from the State University of Feira de Santana. It comprises two sets: one containing letters produced between 1820 and 2000 by Brazilians born after 1720 and voice records from the 90 s in Bahia; another containing manuscripts produced between 1640 and 1820 by Brazilians born after 1590, along with documentation generated by the Portuguese during the initial 150 years of colonisation.

Unfortunately, the initial set of CEDOHS was the only one accessible to us, comprising 927 documents with a total of 229 thousand tokens. Furthermore, information about the dates of the textual pieces is not uniformly provided.

5.1.13 Word order and word order change in Western European languages

The *Word Order and Word Order Change in Western European Languages* (WochWel) project, compiled by CLUL, aimed to increase the availability of syntactic data for linguistic resources from Old Portuguese. To achieve this, CLUL used the Kepler tagger Kepler (2005) for annotating four documents. These include the *Livro de José de Arimateia*, a medieval text preserved to modern times in a sixteenth-century copy of an original thirteenth-century manuscript; the *Demanda do Santo Graal*, another medieval manuscript dating back to the 15th century; *A Crónica Geral de Espanha*, a chronicle compiled in 1344 by Pedro Afonso; and a collection of notarial texts. The annotation method employed in these documents is consistent with the approach illustrated in Fig. 2.

[15] https://en.wikisource.org/wiki/Category:Portuguese_authors
[16] https://openlibrary.org/
[17] http://www.dominiopublico.gov.br/pesquisa/DetalheObraForm.do?select_action= &co_obra=16090

Table 19 Overview of metadata available for each synchronic corpus

Corpus	PA	PPM	CHLMP	AdA	FLY	EPISA	CDAIII
Title	✓	✗	✓	✓	✓	✗	✓
Author	✓	✗	✓	✓	✗	✗	✓
Author gender	✓	✓	✗	✗	✓	✗	✗
Author birth	✓	✗	✗	✗	✓	✗	✗
Author residence	✗	✓	✗	✓	✓	✗	✓
Author Job	✓	✓	✗	✗	✓	✗	✓
Author social stat	✓	✓	✗	✗	✓	✗	✗
Pub. date	✓	✓	✓	✓	✓	✗	✗
Doc. genre	✗	✓	✓	✓	✓	✓	✓

Table 20 URLs where the synchronic corpus data were sourced

Corpus	Obtained	URL
PA	Scrape	http://arquivopessoa.net/textos/
PPM	Download	https://hdl.handle.net/21.11129/0000-000D-F8CE-4
CHLMP	Download	https://sites.google.com/view/projeto38597
AdA	Download	https://hdl.handle.net/21.11129/0000-000D-F8 C0-2
FLY	Download	https://hdl.handle.net/21.11129/0000-000D-FE7C-B https://hdl.handle.net/21.11129/0000-000D-F93E-6
EPISA	Download	https://rdm.inesctec.pt/dataset/cs-2022-005
CDAIII	Download	https://hdl.handle.net/21.11129/0000-000D-FE7C-B

The **Obtained** column categorises the method used for data collection: **Download** indicates direct data retrieval. In contrast, **Scrape** uses a custom script to extract information from the specified URLs

5.1.14 Vercial & corpus do Português

The *Vercial* (2022) corpus, compiled by the ALGORITMI research group of the University of Minho, contains 309 digitised literary pieces of 55 Portuguese authors from 1500 to 1933, totalling 14,9 million tokens.

The *Corpus do Português* Davies (2006) (CdP), compiled by Mark Davies, comprises approximately 45 million words with nearly 57 thousand Portuguese texts across different genres from the 1300 s to the 1900 s, with around 20 million words specifically from the 1900s.

Unfortunately, these corpora are private, and we could not access them at the time of writing.

5.2 Synchronic corpora

Synchronic corpora contain texts from a particular period. From the collected corpora, the subsequent ones have been identified as synchronic: *Pessoa Archive* ARQUIVO PESSOA (2023) (PA), *Portuguese Parish Memories* (PPM) corpus As Memórias

Table 21 EPISA corpus distribution of document and tokens per document type

typology	N. Docs	N. Tokens
Structured report	328	32 k
Letter	162	24 k
Unstructured report	98	6 k
Process cover	60	2 k
Theatre cover	60	4 k
Total	708	69 k

Paroquiais de 1758 (2023), *Corpus Histórico da Linguagem da Medicina em Português* (CHLMP) (Corpus Histórico da Linguagem da Medicina em Português, 2023; Finatto et al., 2018), *Arquivo dos Açores* (AdA) (2023), *Forgotten Letters Years* (FLY) corpus Gomes et al. (2012), *Entity and Property Inference for Semantic Archives* (EPISA) dataset Falcão et al. (2022), and *Chancelaria de D. Afonso III* corpus (CDAIII) Chancelaria de D. Afonso III (2023). Metadata related to these corpora is detailed in Table 19, and Table 20 provides the URLs where the data for the synchronic corpus was sourced.

5.2.1 Pessoa archive

The *Pessoa Archive* ARQUIVO PESSOA (2023) consists of a database with 4,528 textual pieces from Fernando Pessoa, one of the most significant Portuguese literary figures of the 20th century—poet, writer, literary critic, translator, publisher, and philosopher. These textual pieces are mainly poems, proses, and letters, but other literary genres are also included. The Pessoa Archive contains 1,20 million tokens.

This corpus contains most of Pessoa's published texts, written between 1888 and 1935. The selection methodology prioritised first editions when multiple editions exist for a given piece. Notably, the corpus exclusively features the original versions of the texts, with exceptions made for specific pieces translated into Portuguese.

As we can observe by the number of pieces in this archive, Pessoa was a prolific writer who used approximately seventy-five heteronyms; some that stand out are Alberto Caeiro, Álvaro de Campos and Ricardo Reis. Pessoa called them heteronyms instead of pseudonyms to capture their true independent intellectual life better since some of these imaginary figures sometimes held unpopular or extreme views.

5.2.2 Entity and property inference for semantic archives corpus

The *Entity and Property Inference for Semantic Archives* dataset Falcão et al. (2022) was produced by the Institute for Systems and Computer Engineering, Technology and Science (INESC TEC) to evaluate automatic text recognition tools. The dataset contains 708 manual transcriptions and the corresponding scanned documents from 1910 till 1974 of letters, structured reports, non-structured reports, processes covers, and theatre plays covers, as shown in Table 21, gathered from two collections of the *Arquivo Nacional Torre do Tombo*: the General Administration of National Treasury (DGFP) and the National Secretariat of Information (SNI).

5.2.3 Corpus Histórico da Linguagem da Medicina em Português

The *Corpus Histórico da Linguagem da Medicina em Português* (Corpus Histórico da Linguagem da Medicina em Português, 2023; Finatto et al., 2018) was a collaboration between the Federal University of Rio Grande do Sul and the University of Évora, where two medicine books, *Observaçoens medicas doutrinaes de cem casos gravíssimos*, written in 1707 by João Curvo Semedo and comprises 101 chapters spanning 634 pages, and *Postilla Religiosa, e Arte de Enfermeiros*, written by Diogo de Santiago in 1741 with 59 chapters across 340 pages, were transcribed. The first book describes 101 cases of illness, allowing a historical view of the healthcare and medical science of the 18th century; the latter addresses religion and the infirmary, with a particular focus on the technical part and assistance to death. It is noteworthy to highlight that digital versions of both books are readily accessible.

5.2.4 Arquivo dos Açores

Arquivo dos Açores (2023) is established as a reference work for historical research on the Azores archipelago, having two series and 20 volumes. The first series, comprising 15 volumes, spans from 1878 to 1959. Volumes 13, 14, and 15 were only published in 1920, 1927, and 1959, respectively. The second series, consisting of 5 volumes and coordinated by Mário Viana, commenced 40 years after the conclusion of the first series in 1999 and continued until 2013. However, the second series is currently halted due to insufficient funding.

The first twelve volumes have been digitised by the Internet Archive[18] with funding from the University of Toronto. However, extracted textual information exhibits numerous OCR errors. To address this issue, the University of Azores curated the first volume, 594 pages. This volume covers an extensive historical period from 1439 to 1877 and was published in 1878.

AdA has readily available digital versions of the first twelve volumes.

5.2.5 Chancelaria de D. Afonso III

The *Chancelaria de D. Afonso III* corpus Chancelaria de D. Afonso III (2023) was produced by the Interdisciplinary Centre for History, Culture and Societies of the University of Évora. CDAIII contains 34 documents written between 1255 and 1279.

These texts originated from the chancellery of King Afonso III of Portugal, tasked with drafting laws, letters, and various documents on the King's behalf. Consequently, this corpus holds indispensable insights into the historical events, linguistic practices, and the language employed within the Royal court during that period.

[18] www.archive.org

5.2.6 FLY corpus

The *Forgotten Letters Years* corpus Gomes et al. (2012) was curated by CLUL. Comprising 2,000 informal letters from 572 authors of different social classes in 351 documents, the FLY corpus captures the sentiments expressed in the context of war, migration, imprisonment, and exile, from 1915 to 1974. The authors wrote during a period when literacy and writing skills were gradually becoming more widespread. Thus, these private letters allow the study of ordinary people within society regarding their linguistic knowledge, behaviour and social identity.

The corpus also contains automatically annotated data featuring morphosyntactic details such as PoS, lemmas, and inflexions, similar to Fig. 2 from PS Corpus. Each letter also includes meta-information provided by the annotator, encompassing details such as the type and characteristics of the original document, dimensions, and a concise summary of its content.

5.2.7 Portuguese parish memories

The *Portuguese Parish Memories* As Memórias Paroquiais de 1758 (2023) corpus was compiled by the Research Centre for Development in Human and Social Sciences of the University of Évora and consists of survey responses gathered between 1758 and 1761 from parish priests overseeing dioceses regarding their local communities. This inquiry was limited to the Portuguese Kingdom, not including dioceses from the Atlantic islands. Given the priests' high literacy at the time, the corpus has high textual quality.

The PPM survey includes 60 questions, 26 related to the locality, 13 about close mountains and 20 others regarding nearby rivers. As expected, the parish priests only responded to what suited their territory. The questions concerning the locality inquired about administrative and jurisdictional (ecclesiastical and secular) issues, demographic data, fairs, local fruits, the impact of the 1755 earthquake, and seaports or ramparts. The ones related to mountains asked about their size, water sources, medicinal herbs, mines, lagoons, villages, monasteries, and churches, which are essential to studying their natural landscape and use of resources. Regarding the rivers, it questioned their size and flow, navigability, current direction, fish and fishery-related activities, bridges, mills and cultivation on the margins, among others. Each of these parts ended with an open question: *And all that is worthy of memory and was not included in the survey?*, to invite the priests to describe other relevant pieces of information about each place.

PPM consists of 44 volumes archived in the *Arquivo Nacional Torre do Tombo*, being accessible online.

Table 22 Overview of corpora

Corpus	N. Tokens	Time scope	Doc.Aggr.	Prov.	Genre	Annot.	License	Acc.	Data Format	Source
CP Davies (2006)	45,0 M	13th-19th	U	U	U	U	Private	✗	-	ad hoc
Gutenberg Project Gutenberg (2023)	19,4 M	14th-20th	Cont	NA	P, Pr, N, T, L	–	Open	✓	TXT	ad hoc
Vercial (2022)	14,9 M	16th-20th	U	U	U	Lem	Private	✗	–	Ling
DiaPT Pichel Campos et al. (2018)	10,74 M	12th-21th	TB	Por, Bra	L, M, D, G, T, P, Pr, Nl, N	PF	GPL	✓	TXT	journal
GMHP (2022)	7,03 M	15th-20th	TB	NA	P, Pr, Nl, T, N	–	CC BY-NC-ND	✓	TXT	ad hoc
ELTeC Santos (2021)	6,55 M	1840-1920	Cont	Por	Nl	NE, PoS, Lem	CC BY 4.0	✓	XML	CA
Colonia Zampieri and Becker (2013)	4,58 M	16th-20th	Cont	Por, Bra	L, M, D, G, T, P, Pr, Nl, N	PoS, Lem, CN	CC BY-NC-ND	✓	TXT	journal
BDCamões Grilo et al. (2020)	3,89 M	15th-21th	Cont	Por, Bra	Nl, P, N, L, T, Pr	PoS, Lem, SA	MS-NC-NoReD-ND	✓	TSV, XML	journal
Tycho Brahe Galves (2018)	3,37 M	14th-19th	Cont	Por, Bra	L, N, D, G, N, T	PoS, SA, CN	CC BY-NC-ND	✓	TXT	journal
CIPM Xavier (2016)	3,32 M	13th-16th	TB	Por	Pr, P	PoS	CC BY-NC-ND	✓	TXT	journal
LT Corpus Généreux et al. (2012)	1,48 M	1850-1940	GB	Por, Bra	Pr, N	PoS	ELRA end-user NC	✓	cqpweb, TXT	journal
ARQUIVO PESSOA (2023)	1,20 M	20th	Cont	Por	L, P, Pr	–	No License	✓	HTML	ad hoc
PS Corpus Vaamonde et al. (2014)	763k	16th-19th	TB	Por, As	L	PoS, CN	CC BY-NC-ND	✓	TXT, PSD	journal
WochWel Pereira (2015)	465k	14th-16th	Cont	Por	M, N	PoS, SA	CC BY-NC-ND 4.0	✓	PSD	journal
PPM As Memórias Paroquiais de 1758 (2023)	272k	18th	GB	Por	R	–	CC BY	✓	DOCx	ad hoc

Table 22 (continued)

Corpus	N. Tokens	Time scope	Doc.Aggr.	Prov.	Genre	Annot.	License	Acc.	Data Format	Source
CEDOHS Corpus Eletrônico de Documentos Históricos do Sertão (2023)	229k	16th-20th	GB	Bra	ATr, R	–	No License	✓	PDF	ad hoc
CHLMP Corpus Histórico da Linguagem da Medicina em Português (2023)	192k	18th	Cont	Por	R, N	–	No License	✓	TXT	ad hoc
AdA Arquivo dos Açores (2023)	182k	19th-21th	Cont	Por	R	–	MS-NC-NoReD-ND	✓	TXT, PDF	PC
Fly Corpus Gomes et al. (2012)	131k	20th	Cont	Por, Afr, As,NAm, SAm	L	PoS, SA	CC BY-NC-ND 3.0	✓	XML	journal
EPISA Falcão et al. (2022)	69k	20th	GB	Por	L, R, T, M	–	Open Def. 2.1	✓	TXT	ad hoc
CTA CTACorpus (2022)	29,5k	13th-16th	Cont	–	N, Pr	Lem	CC BY-NC 4.0	✓	HTML	ad hoc
CDAIII Chancelaria de D. Afonso III (2023)	18k	13th	GB	Por	R	–	CC BY-NC-SA	✓	DOCx	PC

The **Doc. Aggr.** (Document Aggregation) column categorises documents based on three methods: **Cont** (Continuous), where each document is identified by its publication or writing date; **TB** (Time Bucket), documents are grouped into time frames, often by centuries; and **GB** (genre Bucket), documents are organised by genre. **Prov.** (Provenance) indicates the document's origin, with abbreviations such as **Por** (Portugal), **Bra** (Brazil), **Afr** (Africa), **SAm** (South America), **NAm** (North America) and **As** (Asia). The **Genre** column describes the types of genres present in each corpus, including **L** (Letter), **M** (Minutes), **N** (Narrative), **Nl** (Novel), **D** (Dissertation), **G** (Grammars), **N** (News), **T** (Theatre), **ATr** (Audio Transcriptions), **P** (Poetry), **Pr** (Prose) and **R** (Report). **Annot.** (Annotations) outlines the shared annotations in each corpus, which may include **NE** (Named-entity), **PoS** (Part-of-Speech), **PF** (Phonetic Normalisation), **CN** (Contemporary Normalisation), **SA** (Syntactic Annotations) and **Lem** (Lemma). The **Acc.** (Access) column indicates whether the corpus is accessible, while the **Source** column specifies the origin of these datasets, which can be **CA** (Cost Actions), **PC** (PORTULAN CLARIN), **Ling** (Linguateca), **journal**, or **ad hoc**

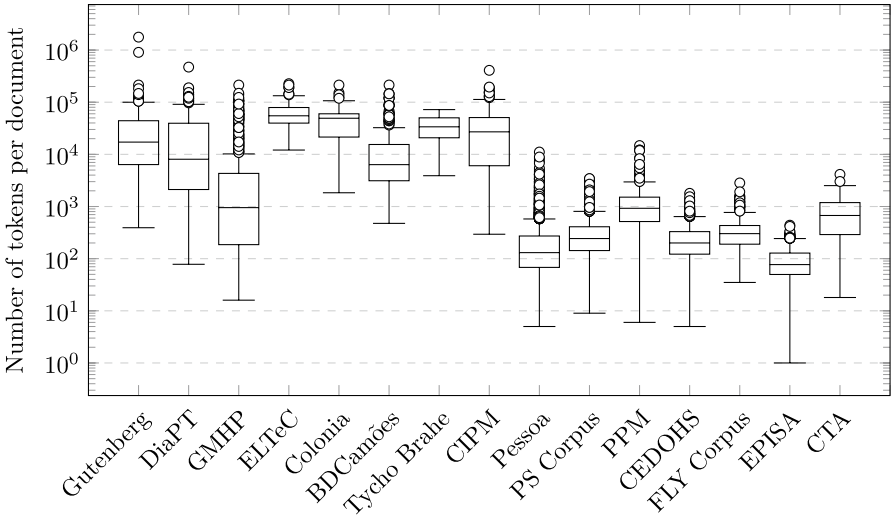

Fig. 3 Distribution of tokens per document in each corpus

6 Discussion

The process of discovering and collecting corpora was not always straightforward. While certain corpora, like those accessible through the PORTULAN CLARIN repository (such as PS corpus, BDcamões, LT Corpus, CIPM, AdA, CDAIII, FLY, and PPM), provided straightforward downloadable links, others demanded more intricate procedures like web scraping. For example, in the case of the Pessoa corpus, CTA, and Gutenberg, the data collection process involved scraping information from websites, adding a layer of complexity. The remaining ones, namely Tycho Brahe, ELTEC, EPISA CHLMP, CEDOHS, DiaPT, WochWel and Colonia, are hosted on dedicated websites, requiring the download of zip files or PDF documents.

Regarding the overall document and content length distribution, as illustrated in Fig. 1, there is a discernible upward trend in both document availability and content as we approach more recent periods. This surge can be attributed to several factors, including higher literacy rates and advancements in documentation technologies. For instance, before the printing press, documents were reproduced by hand copying, a labour-intensive process that took a copyist in the ninth or tenth centuries one day to copy four pages. In the fifteenth century, Gutenberg invented the movable-type printing press, which launched an information revolution, as the printed word was used to disseminate ideas and concepts throughout the world. Similarly, at the end of the nineteenth century, the invention of the Remington typewriter revolutionised paper usage in industry, enabling cost-effective low-volume replication Liu (2004).

Examining the content length per document, Fig. 3 provides a visual summary of the distribution of tokens per document across several openly accessible corpora, excluding LT Corpus, CHLMP, WochWel and AdA. The LT Corpus aggregates

multiple documents in a single file, while AdA, CHLMP, and WochWel consist of one, two, and four documents, respectively. Upon analysis, variations in content density become evident across different corpora. For instance, as expected, EPISA exhibits the lowest median (77 tokens) since it mainly comprises reports and letters. In contrast, the one with the highest median (54,902 tokens) was ELTeC, a corpus composed of novels, closely followed by Colonia. Examining the highest interquartile range, CIPM takes the lead (44,711 tokens), closely followed by ELTeC, Colonia, and Gutenberg. EPISA also records the lowest interquartile range (78 tokens).

Table 22 presents a comprehensive overview of the identified corpora, including information such as time scope, document aggregation, document provenance, genre, annotations, license, access, and data format. Notably, the corpora exhibit variations in data format, with TXT being the predominant format and the most straightforward for processing plain transcriptions, compared to PDF or DOCx. XML and HTML are the most user-friendly when annotations are available due to their structured data organisation.

In terms of document origin, despite the majority being from Portugal and Brazil, certain corpora, namely the FLY Corpus and the PS Corpus, include textual pieces from diverse regions such as Asia or Africa. However, these corpora consist of letters, which likely reflects the Portuguese's extensive exploration and colonisation.

On the other hand, examining the distribution of genres/document types, the most common are letters and prose, followed by narratives and poems. Regarding annotations, the most common are part-of-speech tags and lemmas. It is worth noting that the majority of such annotations was obtained via automatic means; these systems exhibit lower tagging accuracies when applied to historical texts, compared to their performance on modern texts Claridge (2008).

Several corpora also offer valuable metadata, extending their utility across various tasks. This metadata encompasses essential information about the documents, including details about the authors, genres, and other contextual aspects. This enhances the versatility of corpora, enabling research on different tasks like authorship detection, genre classification, orthographic conversion, and lexicon construction.

Regarding licensing, Project Gutenberg offers the least restrictive one, as it does not claim copyright over its published titles and actively promotes their free reproduction and distribution. Similarly, DiaPT operates under the GNU license, which grants the freedom to share and modify all versions. EPISA permits use, modification, and sharing, with the primary stipulation being to maintain provenance and openness.

In contrast, PPM and ELTeC authorise their material's reuse, distribution, remixing, adaptation, and enhancement if proper attribution is provided. They also allow both academic and commercial applications. CDAIII and LT Corpus allow academic usage while restricting commercial applications but permit modifications and distributions with proper attribution.

The Tycho Brahe, PS Corpus, GMHP, CIPM, WochWel, and FLY corpus have more restrictive licenses, prohibiting commercial use and the creation of derivative works. Their use is limited to academic purposes and requires proper credit.

Colonia, lacking an explicit license, is assumed to be the most restrictive among the corpora used; therefore, it has the same restrictions as the previous ones.

CTA, BDCamões, and AdA have similar constraints, allowing its use solely for academic purposes and forbidding redistribution or derivative works. Pessoa, CHLMP, and CEDOHS do not have explicit licenses. Finally, Vercial and CdP are private corpora.

There has been an increase in the digitisation of historical texts; however, in comparison to modern corpora, their availability remains limited. Despite these constraints, researchers continue generating historical embedding models, yet their quality is not always satisfying Sahlgren and Lenci (2016). Therefore, assessing these embeddings becomes essential for advancing distributional semantics studies in historical languages. Addressing this need, the Benchmark of Assessing word embeddings in Historical Portuguese (BAHP) Tian et al. (2021) has been developed, which evaluates embedding models through four tests – analogy, similarity, outlier detection, and coherence – employing sources like the CIPM and Colonia corpora.

7 Conclusion

Historical corpora are essential for historical, linguistic and computational research. They provide valuable insights into the evolution of language, enabling researchers to trace linguistic changes over time and serve as a repository of historical documents. Additionally, historical corpora play a crucial role in advancing computational linguistics and natural language processing of historical texts, allowing the training and use of language models for numerous tasks.

In terms of historical corpora for Portuguese, as shown, there are already some available collections that span wide periods and contain different document types, offering valuable insights into the evolution of the Portuguese language. As previously mentioned, some researchers (Pichel Campos et al., 2018; Niculae et al., 2014; Zampieri et al., 2016; Nevins et al., 2015; Santos and Mota, 2015; Bick and Zampieri, 2016) have used them to trace linguistic changes, study syntactic structures, and explore vocabulary shifts across centuries. However, compared to other languages, available resources for Portuguese remain relatively limited. Consequently, there is an ongoing need to continue to contribute and expand existing collections or create new ones to address this gap and foster a more comprehensive understanding of the linguistic history of Portuguese.

Given the number of available corpora, another possible direction with the resources found is that it is now feasible to consider adapting pre-trained large language models specifically for historical Portuguese. This approach could significantly enhance tasks such as post-OCR correction, NER, and PoS tagging for historical texts. Such targeted improvements in language modelling would refine the accuracy and efficiency of processing historical documents and expand the range of computational tools available to researchers working with historical Portuguese corpora.

Acknowledgements Not applicable.

Author Contributions TFO researched, processed, analysed, and wrote the article with the supervision and review of HLC.

Funding Open access funding provided by FCT|FCCN (b-on). This work was financially supported by Base Funding (UIDB/00027/2020) and Programmatic Funding (UIDP/00027/2020) of the Artificial Intelligence and Computer Science Laboratory (LIACC) funded by national funds through FCT/MCTES (PIDDAC).

Data availability Regarding data availability, links are provided; however, the actual datasets are not shared due to restrictions imposed on their distribution by some corpora. Additionally, scripts are not provided to avoid potential loading problems on certain websites since they can inadvertently cause Distributed Denial of Service (DDoS) issues if improperly configured.

Declarations

Conflict of interest The authors declare that they have no Conflict of interest.

Ethical approval Not applicable.

Consent to participate Not applicable.

Consent to publication Not applicable.

References

Abadji, J., Ortiz Suarez, P., Romary, L., & Sagot, B. (2022). Towards a cleaner document-oriented multilingual crawled corpus. arXiv e-prints, 2201–06642 arXiv:2201.06642 [cs.CL]

Alatrash, R., Schlechtweg, D., Kuhn, J., & Walde, S. (2020). CCOHA: Clean Corpus of Historical American English. In: Proceedings of the Twelfth Language Resources and Evaluation Conference, pp. 6958–6966. European Language Resources Association, Marseille, France. https://aclanthology.org/2020.lrec-1.859 Accessed 2023-09-23

Arquivo dos Açores. https://hdl.handle.net/21.11129/0000-000D-F8C0-2. Accessed: 16-5-2023

ARQUIVO PESSOA. http://arquivopessoa.net/. Accessed: 15-05-2023

As Memórias Paroquiais de 1758. http://www.cidehusdigital.uevora.pt/portugal1758. Accessed: 15-05-2023

Barreto, J. F. (1671). Ortografia da Língua Portugueza. Biblioteca Nacional, [L-323-V] purl pt, biblioteca nacional digital, Portugal

Bick, E. (2006). Functional aspects in portuguese ner. In: Vieira, R., Quaresma, P., Nunes, M.d.G.V., Mamede, N.J., Oliveira, C., Dias, M.C. (eds.) Computational Processing of the Portuguese Language, pp. 80–89. Springer, Berlin, Heidelberg

Bick, E. (2014). In: Sardinha, T., Ferreira, T. (eds.) PALAVRAS - A Constraint Grammar-Based Parsing System for Portuguese, pp. 279–302. Bloomsbury Academic, New York.

Bick, E., & Zampieri, M. (2016). Grammatical annotation of historical portuguese: Generating a corpus-based diachronic dictionary. In P. Sojka, A. Horák, I. Kopeček, & K. Pala (Eds.), *Text, Speech, and Dialogue* (pp. 3–11). Cham: Springer.

Blank, A. (1999). In: Blank, A., Koch, P. (eds.) Why do new meanings occur? A cognitive typology of the motivations for lexical semantic change, pp. 61–90. De Gruyter Mouton, Berlin, Boston. https://doi.org/10.1515/9783110804195.61

Bowern, C. (2019). Semantic change and semantic stability: Variation is key. In: Proceedings of the 1st International Workshop on Computational Approaches to Historical Language Change, pp. 48–55. Association for Computational Linguistics, Florence, Italy. https://doi.org/10.18653/v1/W19-4706. https://aclanthology.org/W19-4706

Branco, A., Silva, J. R. (2006). A suite of shallow processing tools for Portuguese: LX-suite. In: Demonstrations, pp. 179–182. https://aclanthology.org/E06-2024

Calderón Campos, M., & Díaz-Bravo, R. (2021). An online corpus for the study of historical dialectology: Oralia diacrónica del español. Digital Scholarship in the Humanities **36**(Supplement_2), 30–48. https://doi.org/10.1093/llc/fqaa066. https://academic.oup.com/dsh/article-pdf/36/Supplement_2/ii30/41091229/fqaa066.pdf

Carvalho, M. S. d., & Cabecinhas, R. (2013). The orthographic (dis)agreement and the portuguese identity threat. Lusofonia and Its Futures, 82–95

Chancelaria de D. Afonso III: documentos em português. https://hdl.handle.net/21.11129/0000-000D-FE7C-B. Accessed: 16-5-2023

Ciobanu, A. M., Dinu, L. P., Şulea, O.-M., Dinu, A., & Niculae, V. (2013). Temporal text classification for Romanian novels set in the past. In: Proceedings of the International Conference Recent Advances in Natural Language Processing RANLP 2013, pp. 136–140. INCOMA Ltd. Shoumen, BULGARIA, Hissar, Bulgaria. https://aclanthology.org/R13-1018

Claridge, C. (2008). Historical corpora. In: Lüdeling, A., Kytö, M. (eds.) Corpus Linguistics : an International Handbook ; Volume 1

Comunidade dos Países de Língua Portuguesa (CPLP). https://www.cplp.org/id-2597.aspx. Accessed: 2022-10-26

Corpus diacróico y diatópico del español de América. https://www.cordiam.org/. Accessed: 23-04-2024

Corpus Eletrônico de Documentos Históricos do Sertão. http://www5.uefs.br/cedohs/view/home.html. Accessed: 16-5-2023

Corpus Histórico da Linguagem da Medicina em Português (Século XVIII): Terminologia Diacrônica e Humanidades Digitais. https://sites.google.com/view/projeto38597. Accessed: 16-5-2023

Corpus Léxico de Inventarios. https://corlexin.unileon.es/el-corpus/. Accessed: 23-04-2024

Coussé, E. (2011). Een digitaal compilatiecorpus historisch nederlands. Lexikos **20**, https://doi.org/10.5788/20-0-136

CTACorpus. http://teitok.clul.ul.pt/cta/. Accessed: 13-12-2022

Culpeper, J., & Kytö, M. (1997). Towards a corpus of dialogues, 1550-1750. In: Language in Time and Space : Studies in Honour of Wolfgang Viereck on the Occasion of His 60th Birthday. Zeitschrift für Dialektologie und Linguistik. Beihefte, vol. 97, pp. 60–73. Franz Steiner. Stuttgart., ???

Curzan, A. (2000). English historical corpora in the classroom: The intersection of teaching and research. *Journal of English Linguistics, 28*(1), 77–89. https://doi.org/10.1177/007542400022004884

Davies, M. (2006). Corpus do Português: 45 Million Words, 1300s - 1900s. http://www.corpusdoportugues.org

Davies, M. (2012). Expanding horizons in historical linguistics with the 400-million word corpus of historical american english. *Corpora, 7*, 121–157. https://doi.org/10.3366/cor.2012.0024

Delpher. https://www.delpher.nl/over-delpher/delpher-open-krantenarchief/wat-zit-er-in-het-delpher-open-krantenarchief#e6bce. Accessed: 23-04-2024

Digital Library of Dutch Literature. https://www.kb.nl/en/research-find/datasets/dbnl-dataset. Accessed: 23-04-2024

Early English Books Online Corpus. https://www.english-corpora.org/eebo/. Accessed: 23-04-2024

Early English Correspondence Corpus. https://www.helsinki.fi/en/researchgroups/variation-contacts-and-change-in-english/research/corpus-of-early-english-correspondence. Accessed: 23-04-2024

Early English Medical Writing. https://varieng.helsinki.fi/series/volumes/14/taavitsainen_pahta/. Accessed: 23-04-2024

Eberhard, D. M., Simons, G. F., & Fennig, C. D. (2023). Ethnologue: Languages of the World, 26 edn. SIL International, Dallas. http://www.ethnologue.com

Eighteenth Century Collections Online. https://www.gale.com/primary-sources/eighteenth-century-collections-online. Accessed: 23-04-2024

Evans Early American Imprints Collection. https://textcreationpartnership.org/tcp-texts/evans-tcp-evans-early-american-imprints/. Accessed: 23-04-2024

Evert, S. (2008). A lightweight and efficient tool for cleaning web pages. In: Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08). European Language Resources Association (ELRA), Marrakech, Morocco. http://www.lrec-conf.org/proceedings/lrec2008/pdf/885_paper.pdf

Falcão, M., Dias, M., & Lopes, C. T. (2022). Manual Transcriptions of Typewritten Digital Representations of Portuguese Cultural Heritage Documents from the 20th Century. INESC TEC. https://doi.org/10.25747/WPNA-JE39

Feijó, J. M. M. (1739). Orthographia Ou Arte de Escrever e Pronunciar Com Acerto a Língua Portugueza. Biblioteca Nacional, L-5049-A] purl.pt biblioteca nacional digital, Portugal.

Finatto, M. J., Quaresma, P., & Gonçalves, M. F. (2018). Portuguese corpora of the 18th century: old medicine texts for teaching and research. Proceedings of the Conference on Language Techonologies & Digital Humanities

Finatto, M. J., Quaresma, P., & Gonçalves, M. F. (2018). Portuguese corpora of the 18th century: old medicine texts for teaching and research. Proceedings of the Conference on Language Techonologies & Digital Humanities, 114–120. 10174/23606

Fishman, J. A. (1964). Language maintenance and language shift as a field of inquiry. a definition of the field and suggestions for its further development. *Linguistics, 2*(9), 32–70. https://doi.org/10.1515/ling.1964.2.9.32

Gabay, S., Ortiz Suarez, P., Bartz, A., Chagué, A., Bawden, R., Gambette, P., Sagot, B. (2022). From FreEM to d'AlemBERT: a large corpus and a language model for early Modern French. In: Proceedings of the Thirteenth Language Resources and Evaluation Conference, pp. 3367–3374. European Language Resources Association, Marseille, France. https://aclanthology.org/2022.lrec-1.359

Galves, C. (2018). The tycho brahe corpus of historical portuguese: Methodology and results. *Linguistic Variation, 18*, 49–73. https://doi.org/10.1075/lv.00004.gal

Généreux, M., Hendrickx, I., & Mendes, A. (2012). A large portuguese corpus on-line: Cleaning and preprocessing. Computational Processing of the Portuguese Language. PROPOR, 113–120

Geyken, A., Haaf, S., Jurish, B., Schulz, M., Steinmann, J., Thomas, C., & Wiegand, F. (2010). Das deutsche textarchiv: Vom historischen korpus zum aktiven archiv. In: Digitale Wissenschaft. Stand und Entwicklung Digital Vernetzter Forschung in Deutschland, pp. 157–161

GMHP. http://www.usp.br/gmhp/CorpI.html. Accessed: 14-06-2022

Gomes, M., Guilherme, A., Tavares, L., & Marquilhas, R. (2012). Project FLY: a multidisciplinary project within linguistics. In: Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12), pp. 2833–2837. European Language Resources Association (ELRA), Istanbul, Turkey. http://www.lrec-conf.org/proceedings/lrec2012/pdf/1031_Paper.pdf

Grilo, S., Bolrinha, M., Silva, J., Vaz, R., & Branco, A. (2020). The bdcamões collection of portuguese literary documents: a research resource for language technology and digital humanities. Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), 849–854.

Hajič, J., Ciaramita, M., Johansson, R., Kawahara, D., Martí, M.A., Màrquez, L., Meyers, A., Nivre, J., Padó, S., Štěpánek, J., Straňák, P., Surdeanu, M., Xue, N., & Zhang, Y. (2009). The CoNLL-2009 shared task: Syntactic and semantic dependencies in multiple languages. In: Hajič, J. (ed.) Proceedings of the Thirteenth Conference on Computational Natural Language Learning (CoNLL 2009): Shared Task, pp. 1–18. Association for Computational Linguistics, Boulder, Colorado. https://aclanthology.org/W09-1201

Hamilton, W. L., Leskovec, J., Jurafsky, D. (2016). Diachronic word embeddings reveal statistical laws of semantic change. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 1489–1501. Association for Computational Linguistics, Berlin, Germany. https://doi.org/10.18653/v1/P16-1141. https://aclanthology.org/P16-1141

Kepler, F. (2005). Um etiquetador morfo-sintático baseado em cadeias de Markov de tamanho variável. https://doi.org/10.11606/D.45.2005.tde-20210729-141428

Kerswill, P. (2006). Migration and Language vol. Volume 3, pp. 2271–2285. De Gruyter Mouton, Berlin • New York. https://doi.org/10.1515/9783110184181.3.10.2271

Kissos, I., & Dershowitz, N. (2016). Ocr error correction using character correction and feature-based word classification. 2th IAPR Workshop on Document Analysis Systems (DAS), 198–203, https://doi.org/10.1109/DAS.2016.44

Klein, T., & Dipper, S. (2016). Handbuch zum referenzkorpus mittelhochdeutsch. In: Bochumer Linguistische Arbeitsberichte, vol. 19

Kroch, A. (2020). *Penn Parsed Corpora of Historical English LDC2020T16*. Web download. Philadelphia: Linguistic Data Consortium.

Kutuzov, A., Øvrelid, L., Szymanski, T., & Velldal, E. (2018). Diachronic word embeddings and semantic shifts: a survey. In: Proceedings of the 27th International Conference on Computational Linguistics, pp. 1384–1397. Association for Computational Linguistics, Santa Fe, New Mexico, USA. https://aclanthology.org/C18-1117

Kytö, M. (2010). Corpora and historical linguistics. Revista Brasileira de Linguística Aplicada **11**, 417–457. https://doi.org/10.1590/S1984-63982011000200007

Lampeter Corpus of Early Modern English Tracts. http://korpus.uib.no/icame/manuals/LAMPETER/LAMPHOME.HTM. Accessed: 23-04-2024

Liu, Z. (2004). The evolution of documents and its impacts. *Journal of Documentation, 60*(3), 279–288. https://doi.org/10.1108/00220410410534185. Accessed 2023-12-08.

Manjavacas Arevalo, E., & Fonteyn, L. (2022). Non-parametric word sense disambiguation for historical languages. In: Proceedings of the 2nd International Workshop on Natural Language Processing for Digital Humanities, pp. 123–134. Association for Computational Linguistics, Taipei, Taiwan. https://aclanthology.org/2022.nlp4dh-1.16.

Manjavacas, E., & Fonteyn, L. (2022). Adapting vs. Pre-training Language Models for Historical Languages. Journal of Data Mining & Digital Humanities **NLP4DH**, https://doi.org/10.46298/jdmdh.9152

Marot, .-. Clément, Bergerac, .-., Bergerac, .-., Molière, .-., Brécourt, .-. Guillaume Marcoureau de, Jouin, .-. Nicolas, Coustelier, d.. Antoine Urbain: Paris speech in the past. Oxford Text Archive (2001). http://hdl.handle.net/20.500.12024/2423

Nevins, A., Rodrigues, C., & Tang, K. (2015). The rise and fall of the l-shaped morphome: diachronic and experimental studies. *Probus, 27*(1), 101–155. https://doi.org/10.1515/probus-2015-0002

Nguyen, T. T. H., Jatowt, A., Coustaty, M., & Doucet, A. (2022). Survey of post-ocr processing approaches. *ACM Computing Surveys, 54*, 1–37. https://doi.org/10.1145/3453476

Niculae, V., Zampieri, M., Dinu, L., & Ciobanu, A. M. (2014). Temporal text ranking and automatic dating of texts. In: Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, Volume 2: Short Papers, pp. 17–21. Association for Computational Linguistics, Gothenburg, Sweden. https://doi.org/10.3115/v1/E14-4004. https://aclanthology.org/E14-4004

Ogilvie, B. (2016). Scientific archives in the age of digitization. *ISIS, 107*, 77–85. https://doi.org/10.1086/686075

Old Bailey Corpus. https://www.oldbaileyonline.org/. Accessed: 23-04-2024

Pereira, S. (2015). A anotação sintática de textos medievais portugueses. In: Scriptum Digital, vol. 4, pp. 125–142

Pettersson, E., & Megyesi, B. (2018). The histcorp collection of historical corpora and resources. In: Digital Humanities in the Nordic Countries Conference. https://api.semanticscholar.org/CorpusID:19243754

Philips, J., & Tabrizi, N. (2020). Historical document processing: A survey of techniques, tools, and trends. Proceedings of the 12th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management **1**, 341–349, https://doi.org/10.5220/0010177403410349

Pichel Campos, J. R., Gamallo, P., & Alegria, I. (2018). Measuring language distance among historical varieties using perplexity. application to European Portuguese. In: Zampieri, M., Nakov, P., Ljubešić, N., Tiedemann, J., Malmasi, S., Ali, A. (eds.) Proceedings of the Fifth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2018), pp. 145–155. Association for Computational Linguistics, Santa Fe, New Mexico, USA. https://aclanthology.org/W18-3916

Popescu, O., Strapparava, C. (2015). SemEval 2015, task 7: Diachronic text evaluation. In: Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015), pp. 870–878. Association for Computational Linguistics, Denver, Colorado. https://doi.org/10.18653/v1/S15-2147. https://aclanthology.org/S15-2147

Project Gutenberg. https://www.gutenberg.org/browse/languages/pt. Accessed: 15-05-2023

Reis Gonçalves Viana, A. (1904). Ortografia Nacional. Simplificação e Uniformização Sistemática Das Ortografias Portuguesas. Lisboa Viuva Tavares Cardoso, Portugal.

Ricardo, M. M. C. (2009). Breve história do acordo ortográfico. Revista Lusófona de Educação **13**

Rigaud, C., Doucet, A., Coustaty, M., & Moreux, J.-P. (2019). Icdar 2019 competition on post-ocr text correction. International Conference on Document Analysis and Recognition (ICDAR), 1588–1593 https://doi.org/10.1109/ICDAR.2019.00255

Rijhwani, S., Anastasopoulos, A., & Neubig, G. (2020). Ocr post correction for endangered language texts. Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), 5931–5942 https://doi.org/10.18653/v1/2020.emnlp-main.478

Sahlgren, M., & Lenci, A. (2016). The effects of data size and frequency range on distributional semantic models. In: Su, J., Duh, K., Carreras, X. (eds.) Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, pp. 975–980. Association for Computational Linguistics, Austin, Texas. https://doi.org/10.18653/v1/D16-1099. https://aclanthology.org/D16-1099

Sánchez-Martínez, F., Martínez-Sempere, I., Ivars-Ribes, X., & Carrasco, R. C. (2013). An open diachronic corpus of historical spanish. *Language Resources and Evaluation, 47*(4), 1327–1342.

Sánchez-Prieto Borja, P. Desarrollo y explotación del "corpus de documentos españoles anteriores a 1700" (codea). Scriptum digital. Revista de corpus diacrònics i edició digital en Llengües iberoromàniques (1), 5–35 (1)

Santos, D. (2021). Portuguese novel collection (eltec-por). https://doi.org/10.5281/zenodo.4288235

Santos, D., & Mota, C. (2015). A admiração à luz dos corpos. *Oslo Studies in Language, 7*, 57–77. https://doi.org/10.5617/osla.1462

Scheible, S., Whitt, R. J., Durrell, M., & Bennett, P. (2011). A gold standard corpus of early Modern German. In: Proceedings of the 5th Linguistic Annotation Workshop, pp. 124–128. Association for Computational Linguistics, Portland, Oregon, USA. https://aclanthology.org/W11-0415

Schieffelin, B. B., & Ochs, E. (1986). Language socialization. *Annual Review of Anthropology, 15*(1), 163–191. https://doi.org/10.1146/annurev.an.15.100186.001115

Schmid, H. (1994). Probabilistic part-of-speech tagging using decision trees. In: Proceedings of the International Conference on New Methods in Language Processing, Manchester, UK.

Tahmasebi, N., & Risse, T. (2017). On the uses of word sense change for research in the digital humanities. In J. Kamps, G. Tsakonas, Y. Manolopoulos, L. Iliadis, & I. Karydis (Eds.), *Research and Advanced Technology for Digital Libraries* (pp. 246–257). Cham: Springer.

Tahmasebi, N., Borin, L., & Jatowt, A. (2019). Survey of computational approaches to lexical semantic change. arXiv preprint arXiv:1811.06278

Tang, X. (2018). A state-of-the-art of semantic change computation. *Natural Language Engineering, 24*(5), 649–676. https://doi.org/10.1017/S1351324918000220

Teyssier, P. (2001). História da Língua Portuguesa, pp. 31–35. Sá da Costa, Portugal.

The Corpus of Late Modern English Texts (Extended Version). 2006. Compiled by Hendrik De Smet. Department of Linguistics, University of Leuven. https://varieng.helsinki.fi/CoRD/corpora/CLMETEV/. Accessed: 23-04-2024

Tian, Z., Jarrett, D., Escalona Torres, J., & Amaral, P. (2021). BAHP: Benchmark of assessing word embeddings in historical Portuguese. In: Proceedings of the 5th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature, pp. 113–119. Association for Computational Linguistics, Punta Cana, Dominican Republic (online). https://doi.org/10.18653/v1/2021.latechclfl-1.13. https://aclanthology.org/2021.latechclfl-1.13

Time Corpus. https://www.english-corpora.org/time/. Accessed: 23-04-2024

Vaamonde, G., Costa, A. L., Marquilhas, R., Pinto, C., & Pratas, F. (2014). Post scriptum: Archivo digital de escritura cotidiana. Humanidades Digitales: desafíos, logros y perspectivas de futuro, 473–482.

VERCIAL. https://www.linguateca.pt/acesso/corpus.php?corpus=VERCIAL. Accessed: 14-06-2022

Wagner Filho, J. A., Wilkens, R., Idiart, M., & Villavicencio, A. (2018). The brWaC corpus: A new open resource for Brazilian Portuguese. In: Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018). European Language Resources Association (ELRA), Miyazaki, Japan. https://aclanthology.org/L18-1686

Xavier, M. F. (2016). In: Kabatek, J. (ed.) O CIPM - Corpus Informatizado do Português Medieval, fonte de um Dicionário exaustivo, pp. 137–156. De Gruyter, Berlin, Boston. https://doi.org/10.1515/9783110462357-007.

Zampieri, M. (2017). Compiling and processing historical and contemporary portuguese corpora. CoRR arXiv:1710.00803

Zampieri, M., & Becker, M. (2013). Colonia: Corpus of historical portuguese. Non-Standard Data Sources in Corpus-based Research. ZSM-Studien Series - Vol. 5.

Zampieri, M., Malmasi, S., & Dras, M. (2016). Modeling language change in historical corpora: The case of Portuguese. In: Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16), pp. 4098–4104. European Language Resources Association (ELRA), Portorož, Slovenia. https://aclanthology.org/L16-1647

Zurich English Newspaper Corpus. https://www.es.uzh.ch/en/Subsites/Projects/zencorpus.html. Accessed: 23-04-2024