NLP for historical Portuguese: Analysing 18th-century medical texts

Leonardo Zilio
FAU Erlangen-Nürnberg, Germany
leonardo.zilio@fau.de

Rafaela R. Lazzari
UFRGS, Brazil
rafaelalazzari@gmail.com

Maria José B. Finatto
UFRGS, Brazil
mariafinatto@gmail.com

Abstract

This paper addresses an important challenge for automatically analysing historical documents: how to overcome the textual barriers imposed by historical writing? The mix of lexical variants, and historical spelling and syntax can be a huge barrier for using NLP tools. This study thus presents a description and lexical analysis of a historical medical corpus, and we propose a pipeline for spelling normalisation that retains alignments with the historical spelling. This allows for the application of NLP tools on the normalised version, while keeping track of the original form. Using this methodology, we observed a gain of more than 4% in part-of-speech tagging precision.

1 Introduction

In this paper we deal with the difficult task of using natural language processing (NLP) tools for analysing historical documents in Portuguese and propose new methods for dealing with the differences in 18th-century spelling. Our focus are samples from three medical books that were published in 1707, 1741 and 1794, covering the span of a century.

When dealing with historical texts published in Portuguese, normalisation is the task of converting the words to some current standard form or norm, so as to standardise the vocabulary through the elimination of historical writing variants. Although it may seem easy at first, modernising the writing of a historical text is a complex and detailed work, which requires specific linguistic, grammatical and historical knowledge from the researcher, and is very time-consuming, as it still cannot be done automatically. In addition, extensive training is necessary to understand the historical writing, typography and printing patterns. As in the 18th century there were no writing norms, in the same text, by a single author, several forms of the same word can be found, such as “agoa” and “agua” for the current form “água” [water], in addition to old characters like the long S (ſ), the joining of words that are now separate (*e.g.*, “em quanto” instead of “enquanto” [while]) and *vice versa*. In the normalisation process, the original word form is usually replaced, but it can facilitate reading and computational processing, increasing accessibility to the content of historical materials, especially for those who are not specialists in Linguistics, History of Portuguese or Philology.

Faced with these challenges, our work involved finding a method to help to computationally process the text of three medical works using a normalised version, while keeping links to the original, historical form. The medical documents under scrutiny are the following (the original spelling was preserved in the Portuguese titles): *Observaçoens Medicas Doutrinaes de Cem Casos Gravissimos* [Medical and Doctrinal Observations of a Hundred Severe Cases] (Semedo, 1707), *Postilla Religiosa, e Arte de Enfermeiros* [Religious Postil, and Art of Nurses] (de Sant-Iago, 1741) and *Aviso a' Gente do Mar sobre a sua Saude* [Advice to Sea People about their Health] (Mauran, 1794). We started by manually normalising (*i.e.*, modernising) the spelling of some chapters of each work, building a sample of original and modernised texts. With the aid of a computer-assisted translation tool, we were then able to keep the modern and historical version of sentences paired. Using these alignments between normalised and historical spellings, we applied NLP tools to the normalised corpus and were able to use their results for the original texts.

The main aim of this paper is to highlight a new methodology for working with historical texts that allow for the processing of historical writing by using a normalised spelling version as proxy. We also present a description of the content of three works published in the 18th century in Portuguese, focusing on spelling variants, and create new lexical resources based on these texts. These new lexical resources are available for the future development of tools that can properly process 18th-century Portuguese texts[1].

Our main contributions to the study of historical Portuguese texts using NLP tools are:

- A novel methodology for normalising historical texts, keeping the alignments between original text and its modernised version.
- An aligned corpus with original transcription and modernised spelling of samples from three historical specialised texts. The corpus is aligned at sentence and word level, and it is annotated with part-of-speech (POS) and dependency tags.
- A keyword analysis of each subcorpus using Corpus Linguistics tools.
- A lexicon of variants with lexical units from 18th-century medical texts, and an analysis of spelling variants.
- An evaluation of the improvement that spelling normalisation can provide in using NLP tools with historical texts.

The remainder of the paper is organised as follows: Section 2 discusses other work dealing with historical texts; Section 3 describes tools and resources used for processing our historical corpus; Section 4 displays our NLP pipeline for working with historical documents; in Section 5, we present our corpus and go over a keyword analysis; Section 6 describes the spelling normalisation process; Section 7 discusses word-level alignment; Section 8 contains a lexical analysis of spelling variants; Section 9 presents an experiment showing improvements that spelling normalisation can bring; finally, Section 10 briefly discusses our main achievements and hints at future work.

2 Related work

Several studies have been developed in relation to historical Portuguese. In this section, we present papers that describe work with historical Portuguese and that discuss challenges of working with historical documents.

Cambraia (2023) presents an interesting summary of decisions with which text critics (*i.e.*, those who work with the recovery of textual content from historical sources) are faced when transcribing a historical text. Although in this study we used already transcribed versions of historical texts, we can relate to these issues, as, during our manual spelling normalisation process, we sometimes had to check whether the source text (*i.e.*, the original transcription) was actually following the genuine form (*i.e.*, the one presented in the original historical document).

Regarding lexical variants, Cameron et al. (2020) describe historical variants of Portuguese, and Cameron et al. (2023) propose a categorisation of variants, which can support automatic standardisation of historical texts.

Several papers also discuss the complexity and evaluate the use of NLP tools in historical texts for achieving different tasks, especially information extraction (Quaresma and Finatto, 2020), named-entity recognition (Vieira et al., 2021; Cameron et al., 2022; Zilio et al., 2022), and textual complexity (Zilio et al., 2023).

Finally, we highlight the work of Gonçalves (2020) in describing the *Postilla Religiosa, e Arte de Enfermeiros* (de Sant-Iago, 1741), which we use as part of our corpus. The author goes from chapter to chapter, focusing on historical treatments and providing historical context for textual extracts.

3 Tools and resources

We processed our corpus in several ways, starting by manually normalising historical spellings, then aligning sentences and tokens, and finally compiling lists of keywords, variant spellings, and parsing the aligned texts to add *lemmata*, POS- and dependency-tag information. In this section, we briefly go over tools and resources used in this process.

An important point here is that none of the tools used in this study were originally devel-

[1] The resources are freely available on Github under a GPL 3.0 licence: https://github.com/uebelsetzer/NLP_for_18th-century_Portuguese_medical_texts.

oped for processing historical texts, and this in itself brings innovation in terms of their newfound applications. Also, all resources and language models were developed and trained based on modern-day language, so they bring their own challenges to the adaptation for working with historical documents.

3.1 AntConc and lexicon of variants

Before doing any type of processing, we used AntConc (Anthony, 2004) to check word lists and keyword lists based on the original historical texts. AntConc is a light-weight tool used in corpus analysis that can provide several types of information: besides the aforementioned lists, it can display concordances, calculate collocations, show phrase-distribution patterns, and present word clusters and n-grams.

To generate keyword lists, a reference corpus or reference word list is needed, so we used the list of variant spellings that was compiled by Giusti et al. (2007) based on the historical corpus of Brazilian Portuguese (Murakawa and Gonçalves, 2015). The list contains variants organised under an entry word, and each variant has a frequency register. This list of variants and frequencies was then matched against the word lists from our historical corpus to generate keyword lists.

It is important to bear in mind that our corpus contains texts that were originally written in European Portuguese. By using a list extracted from a historical corpus written in Brazilian Portuguese, we are assuming that the differences between both variants in the 18th century were negligible. If this assumption is wrong, we can then expect an impact on the results of the keyword analysis and in the evaluation of variants that we present, respectively, in Sections 5.4 and 8. Unfortunately, we could not test the correctness of our assumption or precise how big this impact is, because we could not find any similar, computationally processable list for the European variant.

3.2 OmegaT

Our working pipeline starts with spelling normalisation, by converting the original writing into a modern spelling. Here we opted for using OmegaT[2], a tool that was originally designed for computer-assisted translation (CAT). The advantage of a CAT tool is that it displays the historical text along with the new text. This helps in reviewing and avoids issues such as jumping over parts of the original text, which can easily happen, for instance, in a normal text editor or annotation tool. It also has the advantage of splitting the text in sentences and keeping the original and the modern segments aligned at all times.

In addition, CAT tools store the original and normalised text in a TMX file[3], which is an aligned version of the text, and have integrated automatic aligners. In this study, we used OmegaT's automatic aligners for organising aligned sentences. Finally, CAT tools provide access to glossaries and translation memories, which can improve modernisation consistency, and they offer the option of integrating machine translation systems, which can help improve the speed of modernisation.

3.3 Tokeniser and word aligner

After having a sentence-level alignment provided by the CAT tool, we moved on to align the texts at the word level. However, before this word-level alignment, we tokenised the text using NLTK's[4] tokeniser with its default language settings (*i.e.*, without setting its language parameters to Portuguese). This may seem counter-intuitive at first, but the idea behind this decision is that we tried to ensure that words were only split at spaces and punctuation, avoiding any other type of language-specific tokenisation. This decision was made to facilitate the word-level alignment.

We then applied SimAlign (Sabet et al., 2020) on the tokenised sentences to align them at word level. SimAlign requires a pre-trained language model for using language-specific embeddings, so we selected the recently released Albertina model (PT-PT) (Rodrigues et al., 2023).

[2] OmegaT is an open-source tool that is available at: https://omegat.org/.

[3] TMX stands for translation memory exchange file. This file format uses an XML structure for storing aligned sentences and preserving translation metadata.

[4] NLTK's website: https://www.nltk.org/.

3.4 POS tagging and parsing

We tested two parsers to annotate the texts with normalised and original spelling: spaCy[5] and Stanza (Qi et al., 2020). Both are robust parsers that have support for Portuguese, and both allow for using a custom tokenisation and sentence segmentation process, which was important in our case because of the previously mentioned alignment process.

After checking the output from both parsers, both from a fully automated pipeline and from a customised one, we ended up opting for Stanza, as it was more straightforward to set up for maintaining the tokenisation and sentence splitting that we provided.

4 NLP pipeline for historical texts

One of the main contributions of this paper is a new methodology for working with historical texts. Figure 1 represents this methodology. The original, transcribed text is normalised using a CAT tool, and then its sentence-aligned version is used as input for a word-level aligner. The word-aligned output is then used as basis for the application of NLP tools.

By analysing the original transcriptions via the normalised text, new resources (for instance, glossaries or translation models) can be created, which can then be fed back into the CAT tool for facilitating the normalisation process.

5 Corpus description

Our corpus consisted of chapters selected from three medical works from the 18th century. All are written in Portuguese, but, as a reflection of their time period, they do not present a normalised spelling. These three books span almost the full century, starting in 1707 with João Curvo Semedo's *Observaçoens Medicas Doutrinaes de Cem Casos Gravissimos*, then moving on to the middle of the century, 1741, with Fr. Diogo de Sant-Iago's *Postilla Religiosa e Arte de Enfermeiros*, and ending in 1794 with G. Mauran's *Aviso a' Gente do Mar sobre a sua Saude*. In this section we briefly describe each of them.

5.1 *Observaçoens Medicas Doutrinaes de Cem Casos Gravissimos* (Semedo, 1707)

João Curvo Semedo's work was one of the first medical treatises to be published in Portuguese language (Gonçalves, 2020). It was printed in Lisbon, in 1707, and the author was a physician from Monforte, Alentejo, a region in Portugal, who also wrote other medical treatises and handbooks, such as the *Polyanthea medicinal* (1697) and the *Atalaya da vida contra as hostilidades da morte* [An observatory of life against the hostilities of death] (1720). These books, among others from Semedo, have more than 600 pages. This extensive bibliography made Semedo one of the "most popular doctors throughout the Portuguese empire in the eighteenth century" (Furtado, 2008, p.147).

In addition to some well-known and manipulated chemical substances at that time, some innovative treatments prescribed by Semedo, called "the Curvian secrets", were made with ingredients from Brazil, Africa, and Asia. Semedo's new authorial treatments – some very bizarre by today's standards – are always highlighted in his books. They indicate that European medicine was open to using products from other regions of the world.

For this study, we selected three observations (*i.e.* chapters): *Observaçam XLII*, *Observaçam LXXXVIII*, and *Observaçam XC*. As a criterion for the text selection, which was also applied, to a certain extent, to the samples from the other two books, we used the subject of "fever", so all these observations deal with some sort of fever. The three selected observations contain a total of 5,472 tokens and 1,642 types in their non-standardised spelling, according to Antconc (Anthony, 2004).

5.2 *Postilla Religiosa, e Arte de Enfermeiros* (de Sant-Iago, 1741)

Similar to Semedo's *Observaçoens*, Sant-Iago's *Postilla* was a pioneer work in Portuguese in addressing how nurses should provide health care (Gonçalves, 2020). In the 18th century, nurses were commonly part of religious institutions, so the book contains information for the treatment of both the body and the spirit.

The book is split in three main treatises: in the first treatise, each chapter is an advice to

[5] spaCy's website: `https://spacy.io/`.

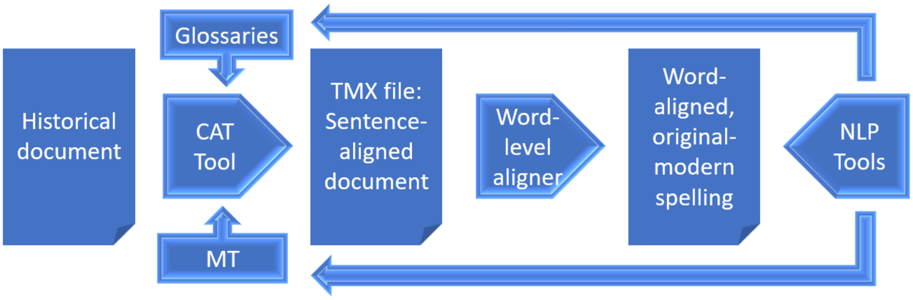
Glossaries
Historical document
CAT Tool
MT
TMX file: Sentence-aligned document
Word-level aligner
Word-aligned, original-modern spelling
NLP Tools

Figure 1: Pipeline for working with historical texts and NLP tools

people in religious positions, such as novices, priests, and bishops, and the text has little to do with health care; the second treatise, which comprises the bulk of the 300-page book, offers advice and instructions on how to prepare and administer medications and treatments to patients; and the third treatise contains information about how to help prepare someone for their impending death, palliative care, and general treatments for the spirit, including instructions on how to perform a "very effective exorcism" in Chapter VI.

The chapters in the second treatise were mostly very short, with brief instructions on how to perform certain treatments or how to concoct certain medications. In total, there are 59 chapters in the second treatise, and, to compose a reasonable corpus, which would be comparable in size to the samples extracted from the other two books, we selected a total of 16 chapters from the *Postilla*: Chapters 17, 22, 29, 30, 32, 33, 34, 40, 41, 42, 43, 44, 46, 47, 48, and 58. Considering their original, non-standardised spelling, these 16 chapters together contain a total of 5,889 tokens and 1,257 types, as seen in Antconc.

5.3 *Aviso a' Gente do Mar sobre a sua Saude* (Mauran, 1794)

The work of G. Mauran was originally published in Marseilles, France, in 1786, with the original title of *Avis aux gens de mer sur leur santé*. It was translated and adapted to Portuguese by Bernardo José de Carvalho, High Surgeon of the Royal Armada, and published in 1794. So, besides being from the end of the century, it is also a book that was not originally written in Portuguese, but was deemed important enough to be translated. The book is a medical treatise and contains information regarding several diseases, including their treatment, so it has chapters dedicated, for instance, to fever, scurvy, and the pest. This book has not received much attention so far, but it offers several points of criticism against bad medical practices that were common at the time.

The chapters in Mauran's *Aviso* are fairly long, so we selected three chapters to be part of this investigation: Chapters 4, 8 and 13. Again, the subject of "fever" was used as a criterion for the selection of these chapters. This subcorpus has a total of 8,724 tokens and 1,803 types, as observed in Antconc, considering their non-standardised spelling.

5.4 Keywords

We generated lists of keywords by matching word lists generated by Antconc (Anthony, 2004) based on the original texts of our corpus against a word list from the historical corpus of Brazilian Portuguese (Giusti et al., 2007). Table 1 shows the **top 15 nouns** for the whole corpus and for each subcorpus, along with their ranks (based on keyness[6]) and frequency.

As expected, the top three keywords in the corpus are content words related to the medical area: "doentes" [sick / sick people], "febre" [fever], "enfermo" [sick / sick person]. The appearance of "fever" is also not surprising, as it is a direct reflection of our methodology for selecting our corpus. As for the rest, the medical theme is prominent, and there are some similarities between the subcorpora, but, most importantly, differences. So, for

[6] We used the keyness metric as set up by default on Antconc: 4-term log-likelihood, considering p <0.05 (with Bonferroni) as threshold.

instance, the *Postilla* does not use the word "doente" [sick person], preferring instead the word "enfermo", which is a synonym. This shows that, in the 18th century, there already is a vocabulary specialisation, and the book that is devoted to nurses [in PT: *enfermeiros*], exclusively uses a word more closely related to the profession, while both physicians' handbooks use "doente". Interesting is also the nonexistence of the word "paciente" [patient], which is more common in nowadays medical works (Scheeren et al., 2008; Zilio, 2009).

Some further elements of notice are: "bezoartico" [type of medicine], in *Observaçoens*, as it is one of the medicines that Semedo himself developed and sold, so it is only natural for him to promote his own "bezoartico", often associating it to seemingly miraculous cures (for instance, in *Observaçam XLII*); the spelling variants "cordeal" and "cordial", which appear as keywords in *Observaçoens* and in *Postilla*; the reference to seemingly common words, such as "camas" [beds] and "camizas" [shirts] in *Postilla*, as these were important items in the work of nurses; and, finally, the reference to "pombos" [pigeons], whose use is actively promoted in *Observaçoens*, and completely rejected in *Aviso*, for the treatment of patients as a way of extracting "evil humours" by eviscerating the animal and deposing its dead body, along with the exposed organs and blood, on the head of the patient.

6 Normalising the corpus

So far, we discussed the corpus in its original spelling. However, a huge part of this study was dedicated to the normalisation of spelling forms. This normalisation ensures that, for instance, "cordeal" and "cordial" can both be associated to the current word "cordial".

As a way of streamlining the standardisation of spelling variants and for the reasons already described in Subsection 3.2, we employed a computer-assisted translation tool. The whole normalisation process was done manually, by going through each segment of the original text and converting words from their original spelling into a modern spelling. In this way, we modernised **only the spelling**, so there was **no change** in word order **nor any adaptation** to make the texts sound modern.

The spelling normalisation of the 22 chapters in the corpus was carried out by an undergraduate student of Translation and a linguist. Table 2 shows differences in number of tokens and types: as expected, the number of tokens remained similar[7], while the number of types was reduced in all subcorpora.

The result of this normalisation process was a corpus of aligned sentences portraying original and modernised spellings. Each normalised chapter was saved, along with its original version, as a TMX file. This sentence-level aligned corpus is the first of our main contributions with this paper.

7 Word-level alignments

Having TMX files as basis, we used SimAlign (Sabet et al., 2020) to automatically align the whole corpus at the word level. Although the amount of change introduced by the modern spelling is not really huge, and most of words are actually aligned one-to-one at the index level, the word-level alignment still presented some issues. For instance, simple words such as "um" [a] and "água" [water], which were commonly spelt, respectively, as "hum" and "agua/agoa" were consistently misaligned, even when their modern counterpart was at the exact same position in the sentence (*i.e.*, where a simple index-based alignment would have worked).

The size of our corpus is relatively small, so we did not want to leave such errors in the alignment get in the way of further processing the documents. To mitigate such issues caused by the historical spelling messing up with the automatic alignments, after the automatic word-level alignment was done, the aligned documents were semi-automatically scrutinised. Tokens that had not been automatically aligned were then manually aligned, and tokens that were aligned with two or more words could have their alignment corrected, if necessary. This semi-automatic alignment was an important step to ensure that the align-

[7] In the *Observaçoens*, the difference in tokens was much larger, but this was probably an issue with how Antconc counts tokens – in this case, for instance, it was set to ignore punctuation –, and not with the actual number of tokens. For comparison, in the tokenised and parsed text, which we will discuss later in the paper, the difference is not 273 tokens, but mere 11 tokens.

Corpus			**Observaçoens**			**Postilla**			**Aviso**		
Rank	*Freq*	*Keyword*	*Rank*	*Freq*	*Keyword*	*Rank*	*Freq*	*Keyword*	*Rank*	*Freq*	*Keyword*
1	76	doentes	1	32	febre	1	112	enfermo	1	63	doentes
2	70	febre	4	20	bezoartico	2	30	enfermeiro	2	38	doença
3	118	enfermo	5	17	quinaquina	3	47	agoa	3	33	febre
5	46	doença	6	16	pombos	5	32	medico	4	34	febres
8	54	febres	8	13	doentes	6	17	purga	5	27	pulso
15	37	estomago	9	31	doente	8	14	banho	9	18	peripneumonia
16	29	sangrias	11	12	humores	10	28	enfermos	13	16	ventre
18	30	enfermeiro	12	18	estomago	12	13	untura	17	18	sangrias
20	59	agoa	14	10	sezaõ	14	12	cordial	19	13	pleuriz
21	41	medico	16	8	cordeal	16	11	unturas	20	13	escarros
22	30	pulso	18	13	febres	17	10	cozimento	22	23	dôr
25	21	humores	19	7	doença	18	13	sangria	27	10	bebida
26	58	doente	25	7	virtude	19	9	cama	32	13	symptomas
27	20	bezoartico	26	7	vitriolo	20	9	camiza	34	9	lado
32	18	peripneumonia	28	9	pès	21	9	unguento	35	9	pontada

Table 1: Main noun keywords in the corpus and in each subcorpora ranked by keyness.

		Tokens	**Types**
Observaçoens	***O***	5472	1642
	M	5745	1596
Postilla	***O***	5889	1257
	M	5892	1249
Aviso	***O***	8724	1803
	M	8716	1738
Corpus	***O***	20085	3633
	M	20353	3433

Table 2: Differences in tokens and types in the original and the standardised spelling of the corpus and each subcorpus. [O = original spelling; M = modern spelling.]

ments were as correct as possible for the analysis of spelling variants and for parsing.

8 Lexicon of variants

The word-level alignments generated in the previous step allowed us to automatically generate a lexicon of variants. With this lexicon, we could check how much variation there was in the original spelling of the texts, and how much this spelling varies from our current spelling standards. We also compared the variants in our texts with the variants in the historical corpus of Brazilian Portuguese.

Our historical corpus has a total of 3,902 types, while the version with modernised spelling has 3,635 types[8]. This results in 1.07 type in the original for each type in the normalised corpus. We can thus notice that the variation in specialised, and, most importantly, printed texts is smaller than, for instance, in handwritten texts (compare, for instance, Cameron et al., 2023). Still, there were some interesting variants to be found, such as "hum" and "hũ" for "um" [a / one], "sezaõ" and "cezaõ" for "sezão" [type of fever / malaria], "terçans" and "terçã" for "terçã" [type of fever / malaria], "damno" and "dano" for "dano" [damage], "sima" and "cima" for "cima" [up], "couza" and "cousa" for "coisa" [thing], and "agoa" and "agua" for "água" [water].

In total, 1,228 types in the original texts had different spelling than their normalised counterparts. This means that almost a third (31.46%) of the types needed to be normalised. This is why resources like ours, which present alignments between original and modern spelling, are important for the long-term objective of automatising the normalisation process.

We also compared the vocabulary that is present in our corpus with the lexicon of variants that was extracted from the historical corpus of Brazilian Portuguese by Giusti et al. (2007). In this comparison, we noticed that, from the 3,703 different word types (*i.e.*, disregarding punctuation and numbers), 1,547 are not present in the lexicon of variants of that larger corpus. Although there are some less relevant entries, such as roman numbers, and verbs with clitics, the main bulk of these new variants are words that belong to the specialised domain of historical medicine.

[8] This number is different from the one in Section 6, because here we are using an NLTK-tokenised version.

Items such as “bezoartico” [type of medicine], “peripneumonia” [old word for pneumonia], “quinaquina” [type of medicine], “sezaõ” [type of fever / malaria], “vitriolo” [vitriol], and “unturas” [ointments] reflect a specialised vocabulary that was not present in other domains and that deserve to be analysed in more details on their own, as they could help improve existing resources based on historical Portuguese, potentially expanding their scope.

9 POS precision and parsing of historical texts

Parsing can give us important information about the lexicon, morphology, and syntax of a text, but modern tools were not trained on historical writing, and usually have news as training corpus, so any tagging on a historical medical corpus will probably not work very well. In this study, we already have an aligned corpus, so we can use the normalised, modern spelling for tagging the text, and then use the alignments to apply the information to the original, non-normalised text. However, even if we normalised the spelling, we are still leaving the original sentence structure untouched, which can have impact on both POS tagging and parsing. So here we devised an experiment to evaluate if there is an actual gain in using normalised spelling for POS tagging.

Stanza (Qi et al., 2020) was selected as main tagger and parser, but sentence splitting came from TMX files, and we used NLTK’s tokeniser. The parser was thus applied on the same tokenised corpus that was used in the alignments, and we parsed each chapter of the corpus using both its original and its normalised version. We then collected 50 random sentences for analysis, which amount to a total of 2,652 tokens in the original corpus (*i.e.*, more than 13% of the corpus). The same 50 sentences were collected from the original and the modernised version, so that the results of the analysis were comparable across the two types of spelling. The same two annotators who normalised the texts also analysed the POS tagging (each analysed 30 sentences, where 10 sentences were in common) in both normalised and original versions. The analysis was done in terms of precision, as the annotators evaluated whether the POS tag attributed to each token was correct or not. Inter-anotator agreement based on 295 tokens (10 sentences) was overall good, with $k = 0.79$ for the original spelling (agreement on 95.93% of the tokens), and $k = 0.57$ for the normalised spelling (agreement on 94.92% of the tokens).

	Measure	**Original**	**Normalised**
Inter-annotator agreement	**Cohen’s kappa**	0.79	0.57
	Tokens %	95.93	94.92
POS precision	**Tokens %**	91.26	95.55
POS precision, no punctuation	**Tokens %**	89.83	94.83

Table 3: Inter-annotator agreement and variation in POS precision in both original and normalised versions of the texts.

Results are shown in Table 3. As we can see, POS tagging on the normalised texts performed 4.29% better, even without making any changes to word order and without using modern-day writing patterns. This difference rises to 5% when ignoring punctuation (which is usually 100% correct). As such, by using a modernised spelling, together with token alignments, we were able to provide a more precise tagging for historical medical texts.

An important caveat is that, on both normalised and original versions, the tagger was partially hindered not only because the texts are from a specific domain – and use historical terminology –, but also because the tokeniser was set to split between words and punctuation, without caring for separating agglutinations (*e.g.*, “do” [of the], “na” [in the], “pelas” [by the]) or clitics that are attached to verbs (*e.g.*, “apartando-se” [moving away from each other], “tirar-lhes” [to take from them], “dar-se-há” [will be given / will give to oneself]).

10 Final remarks

In this paper we presented a series of new resources for historical medical texts. By using texts from three different time periods in the 18th-century (beginning, middle, and end) we covered historical spelling, and also were able to account for some interesting facts related to the 18th-century medicine. The normalisation and later alignment of original and normalised versions of the texts gave rise to a new method for applying modern NLP tools to historical texts.

The use of computer-assisted translation tools, as far as we know, is a novel idea to ensure that the texts are aligned at the sentence level during normalisation. They also allow for the use of glossaries to ensure consistency with normalisation guidelines (for instance, for storing complicated normalisation cases), and for consultation of translation memories (TMX files) with past normalisation decisions. Finally, it also ensures that each sentence is worked on, without any risk of sentences being left without normalisation by mistake.

Our word-level aligned corpus is the first of its kind dedicated to 18th-century medical handbooks. It is an important resource in the future development of automatic normalisation tools. And it is also part of the result of a ground-breaking methodology for the work with historical texts, as we showed, through the case of POS tagging, that NLP tools' performance can be greatly improved by spelling normalisation.

As future work, we intend to investigate methods for automatic or semi-automatic spelling normalisation (such as neural machine translation), so that we can quickly increase the size of the corpus available for analysis. This could then provide the basis for a full-fledged work on historical terminology, leading to the recovery of even more knowledge about medical practices of the past and furthering the studies of their relation with modern medicine within the scope of Digital Humanities and other related disciplines.

Acknowledgements

We would like to thank the Chair of Computational Corpus Linguistics at FAU Erlangen-Nürnberg and the PPG-LETRAS at UFRGS for support, and also the following Brazilian Institutions: PROPESQ - UFRGS – CNPq, for a PIBIC undergraduate research grant; CNPq, for funding a Productivity Research Grant (06/2019, 308926/2019-6) and a research project (26/2021 – PDE – 200051/2023-7); and FAPERGS-CAPES, for funding a research project (06/2018 – INTERNAC., 19/2551-0000718-3).

References

Laurence Anthony. 2004. Antconc: A learner and classroom friendly, multi-platform corpus analysis toolkit. In *IWLeL 2004: An Interactive Workshop on Language e-Learning*, pages 7–13.

César Nardelli Cambraia. 2023. O estilo na crítica textual: domínios de aplicação e a questão da variação linguística/style in textual criticism: application domains and the issue of linguistic variation. *Caligrama: Revista de Estudos Românicos*, 28(1):6–25.

Helena Cameron, Maria Filomena Gonçalves, and Paulo Quaresma. 2020. Linguistic and orthographical classic portuguese variants. challenges for nlp. In *Proceedings of the Workshop on Digital Humanities and Natural Language Processing, co-located with International Conference on the Computational Processing of Portuguese, DHandNLP@PROPOR, Evora, Portugal, March 2, 2020.*, pages 43–48. CEUR.

Helena Cameron, Fernanda Olival, Renata Vieira, and Joaquim Santos. 2022. Named entity annotation of an 18th-century transcribed corpus: problems and challenges. In *Proceedings of the Second Workshop on Digital Humanities and Natural Language Processing (2nd DHandNLP 2022) co-located with International Conference on the Computational Processing of Portuguese (PROPOR 2022), Fortaleza, Brazil, 21st March, 2022*, pages 18–25. CEUR.

Helena Freire Cameron, Fernanda Olival, and Renata Vieira. 2023. Planear a normalização automática: tipologia de variação gráfica do corpus das memórias paroquiais (1758). *LaborHistórico, Rio de Janeiro, ISSN*, pages 2359–6910.

Fr. Diogo de Sant-Iago. 1741. *Postilla religiosa, e arte de enfermeiros: guarnecida com eruditos conceitos de diversos authores, facundos, moraes, e escriturarios*. Officina de Miguel Manescal da Costa, Lisboa, Portugal.

Júnia Ferreira Furtado. 2008. Tropical empiricism: making medical knowledge in colonial Brazil. In *Science and empire in the Atlantic world*, pages 127–151. Routledge.

Rafael Giusti, Arnaldo Candido Jr, Marcelo Muniz, Lívia Cucatto, and Sandra Maria Aluísio. 2007. Automatic detection of spelling variation in historical corpus: An application to build a brazilian portuguese spelling variants dictionary. In *Proceedings of the Corpus Linguistics Conference*, pages 1–20.

Maria Filomena Gonçalves. 2020. A arte de enfermeiros (1741): aspetos do léxico relativo a doenças e remédios no século XVIII. *Panace@*, XXI(52):68–85.

G. Mauran. 1794. *Aviso a' Gente do Mar sobre a sua Saude*. R. Typ. de João Antonio da Silva, Lisboa, Portugal. Translated from the French original edition and extended with some notes by Bernardo José de Carvalho.

Clotilde Murakawa and Maria Filomena Gonçalves. 2015. The corpus of the Dicionário Histórico do Português do Brasil (DHPB). *Planning nonexistent dictionaries*, page 19.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A Python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*.

Paulo Quaresma and Maria José Bocorny Finatto. 2020. Information extraction from historical texts: a case study. In *Proceedings of the Workshop on Digital Humanities and Natural Language Processing, co-located with International Conference on the Computational Processing of Portuguese, DHandNLP@PROPOR, Evora, Portugal, March 2, 2020.*, pages 49–56. CEUR.

João Rodrigues, Luís Gomes, João Silva, António Branco, Rodrigo Santos, Henrique Lopes Cardoso, and Tomás Osório. 2023. Advancing neural encoding of Portuguese with transformer Albertina PT. *arXiv preprint arXiv:2305.06721*.

Masoud Jalili Sabet, Philipp Dufter, François Yvon, and Hinrich Schütze. 2020. Simalign: High quality word alignments without parallel training data using static and contextualized embeddings. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1627–1643.

Fernanda Scheeren, Elisandro Migotto, and Leonardo Zilio. 2008. Estudo exploratório sobre artigos de cardiologia em alemão e português: macroestruturas e usos dos termos Herzinsuffizienz-insuficiência cardíaca. *Salão de Iniciação Científica. Livro de resumos*.

João Curvo Semedo. 1707. *Observaçoens Medicas e Doutrinaes de Cem Casos Gravissimos*. Officina de Antonio Pedrozo Galram, Lisboa, Portugal.

Renata Vieira, Fernanda Olival, Helena Cameron, Joaquim Santos, Ofélia Sequeira, and Ivo Santos. 2021. Enriching the 1758 portuguese parish memories (Alentejo) with named entities. *Journal of Open Humanities Data*, 7:20.

Leonardo Zilio. 2009. Colocações especializadas e 'Komposita': um estudo constrastivo alemão-português na área de cardiologia. Master's thesis, Federal University of Rio Grande do Sul.

Leonardo Zilio, Maria José Bocorny Finatto, and Renata Vieira. 2022. Named entity recognition applied to Portuguese texts from the XVIII century. In *Proceedings of the Second Workshop on Digital Humanities and Natural Language Processing (2nd DHandNLP 2022) co-located with International Conference on the Computational Processing of Portuguese (PROPOR 2022), Fortaleza, Brazil, 21st March, 2022*, pages 1–10. CEUR.

Leonardo Zilio, Maria José Bocorny Finatto, Renata Vieira, and Paulo Quaresma. 2023. A natural language processing approach to complexity assessment of 18th-century health literature. *Domínios de Lingu@gem*, 17.