Portuguese Corpora of the 18th century: old Medicine texts for teaching and research activities

Maria José Bocorny Finatto*, Paulo Quaresma†, Maria Filomena Gonçalves ‡

* Federal University of Rio Grande do Sul - UFRGS, Linguistics Department, Instituto de Letras, Av. Bento Gonçalves, 9500 - Campus do Vale - Prédio 43221, sala 217 Caixa Postal 15002 - 91501-970 - Porto Alegre –RS - Brasil maria.finatto@gmail.com

† Universidade de Évora, Department of Computer Science, Laboratory of Informatics, Systems and Parallelism, R. Romão Ramalho, 59 - 7000 Évora, Portugal. pq@di.uevora.pt

‡ Universidade de Évora, ECS/Department of Linguistics and Literatures, CIDEHUS-UÉ/FCT (UID/HIS/00057/2013) Largo dos Colegiãos 7002-554 Évora, Portugal. mfg@uevora.pt

Abstract

The aim of this paper is to demonstrate the application of the methodologies of Corpus Linguistics and of the Natural Language Processing (NLP) tools to an 18th century Portuguese medicine book. The general objective of this work is to apply the digital humanities tools to a text that has not yet received this kind of approach, in view of teaching and research activities.

1. Introduction

The aim of this paper is to demonstrate the application of the methodologies of corpus linguistics and of Natural Language Processing (NLP) tools to an 18th century Portuguese medicine book. Therefore, the purpose of this work is to present a preliminary essay with a view to a major project on a historical study of the medical terminology in the Portuguese language. It should be noted that, until now, the Portuguese old terminologies had not been studied with computing tools.

First of all, it is important to draw up the theoretical and methodological framework of the analysis, starting with the concept of Corpus Linguistics. Therefore, the general objective of this work is to apply digital humanities (Berry and Fagerjord, 2017; Marquilhas and Hendrickx, 2016) tools to a text from the 18th century that had not yet received this kind of approach, in view of teaching and research activities.

A historical corpus is a set of documents “intentionally created to represent and investigate past stages of a language and/or to study language change” (Claridge, 2008: 242). Nowadays, as mentioned by Kytö (2011), empirical research in Linguistics has increasingly relied on material drawn from a wide range of electronic corpora. In this regard, the history of various languages has (re)emerged as a research area where electronic resources and various kinds of search tools can represent a new stage in the way research has been carried out to investigate mechanisms involved in language change, as well as the features possibly accounting for different phenomena. This kind of corpora have proved particularly useful in some areas of linguistic research, such as: historical lexicology, terminology and lexicography.

These areas involve problems and procedures that nowadays can be recognized as a new “digital philology” (Driscoll and Pierazzo, 2016; Paixão de Sousa, 2013a, 2013b).

As mentioned by Froehlich (2015), if we have a collection of documents organized as a corpus, it is possible to find patterns of grammatical use, or frequently recurring phrases in it. A researcher may also want to find statistically likely and/or unlikely phrases for a particular author or kind of text, particular kinds of grammatical structures or a lot of examples of a particular concept across a large number of documents in context. Corpus analysis, conducted with the help of different kinds of computational tools, “is especially useful for testing intuitions about texts and/or triangulating results from other digital methods” (Froehlich, 2015).

However, in spite of the progress made by these new digital collections of data, with the support of Natural Language Processing (NLP) and Corpus Linguistics tools, there are many difficulties to overcome when handling old documents in digital format. One of the greatest difficulties remains at the computational processing of written language in ancient texts, whether handwritten or printed. Identifying spelling and even updating them are important challenges for the linguists as well as for the NLP researchers.

Taking this challenge into account, this article presents a set of initial procedures for the design of a corpus consisting of samples of ancient medical texts printed in Portuguese of the 18th century on the subject “diseases and their treatments”. Our starting point was the book Observações medicas doutrináes de cem casos gravissimos (Semedo, 1707). It was printed in Lisbon, Portugal, in 1707, with 635 pages, published by João
Curvo Semedo (1635-1719), a Portuguese physician from Monforte, Alentejo, a region within Portugal.

It is important to emphasize that Semedo produced several medical treatises and handbooks of which the following are examples: Polyanthea medicinal (1697), our selected book Observaçoens medicas doutrinaes de cem casos gravissimos (1707) [Medical and doctrinal Observations of a hundred serious cases as Figure 1 shows] and Atalaya da vida contra as emboscadas da morte (1720) [freely translated as: An Observatory of life against the traps of death]. Thereby, the choice of Curvo Semedo is justified by being one of the “most popular doctors throughout the Portuguese empire in the eighteenth century” (Furtado, 2008: 147) and because the majority of treatments he prescribed (“Curvian secrets”) were made with ingredients from Brazil, Africa and Asia. The works of Semedo confirm the opening of European medicine to products from other regions of the world.

In addition, his work represents, in linguistic terms, the period of the “classical Portuguese” (Castro, 2006: 73, 183-198; Banza and Gonçalves, 2018: 39-47), while illustrating the medical terminology of this period. It should be noted that, although the emergence of Portuguese language terminologies (Verdelho, 1998) represents a true technological metamorphosis of the language, its historical analyses still lacks a systematic study, a situation that also applies to the medical terminology.

In the scenario of the ancient Portuguese lexicography, the terms of Medicine received a specific mark (“medicine term”) as we can see in the Vocabulario Portuguez e Latino (Portuguese and Latin Vocabulary) of Rafael Bluteau (1712-1728). This is a dictionary which is an indispensable work for the study of the different technical and scientific terminologies.

On the other hand, the works of Semedo inspired other treatises, namely works published by Portuguese doctors who practiced Medicine in Brazil. Thus, his book Observaçoens medicas doutrinaes de cem casos gravissimos (hereinafter Observaçoens) and others are relevant to the history of Medicine in that territory and even of the so-called “popular pharmacopoeia”, that is, curative methods based on the empirical knowledge of the properties of nature elements. Semedo himself added to the Medicine jargon some words of these pharmacopoeia, which are not actually terms, but popular names for plants, infusions and other “household remedies”, which could even include blood from different animals, stones, seeds and roots.

At last, it is also important to emphasize that Semedo's proposal intended to present these texts, vocabularies and terminologies in a way to make it accessible to their readers, with special attention for the lower literate “young doctors” of his time, who did not know enough Latin but who could read a text in Portuguese.

For all these reasons, the Observaçoens of Curvo Semedo are a rich source of terminological information to which Digital Humanities research methods need to be applied.

Semedo's Observaçoens deal with 101 cases of a wide range of profiles, offering a historical overview of the most common diseases and intercurrences of the time, affecting different population segments: adults, men, women, pregnant women, newborns, young people, the elderly, children, noblemen, peasants or city people.

Semedo's work was also chosen because it was not registered in any of the great historical corpora, not even by Mark Davies' Corpus1, which has 45 million keywords covering a period between 1200 and 1900.

In file format, this scanned book is available for free at Google Books. In addition to this source, for our work on reading, familiarizing with and transcribing the text, it was important to have another complete digital version made from an original. It was available in the Reservation Sector of the Évora Public Library (BPE) in Portugal. Figure 1 below shows this book frontpage from BPE.

The text of this book, as a corpus-sample, will be part of a website specially dedicated to the study of historical lexicology and terminology topics. It is a corpus with printed texts of the 18th century. These materials are integrated to the didactic initiative “Terminologia Histórica”, within the scope of the TEXTECC Project www.ufrgs.br/textecc at Universidade Federal do Rio do Sul (UFRGS), Brazil. Texts and other data build an e-learning environment, where simple sets of texts and online tools will be offered for exploration to help studies on the historical terminology and, in particular, on the history of medical terminology in Portuguese. The tools planned for this website are: a word list generator, a word-context generator to search expressions in a given corpus text, and a generator of lists of word groups to show blocks of repeated words (clusters) along a given text or several texts. Figure 2 below shows the front page of the didactic environment and some preliminary activities with Semedo's book. Starting from the left menu, the user has an initial sample of the corpus and some guided transcription exercises. It is also possible for

1 Website of the Mark Davies Corpus: http://www.corpusdoportugues.org/interface2016.asp.
2. The tools for text processing tests

TermoStat receives an input text and returns as a main result a list of candidate terms (CT) derived from the text. A term – or a specific word item – can be either simple (a word) or complex (a sequence of words). Each term receives a score based on the frequency of the term in the analyzed corpus, the corpus of analysis (CA), and its frequency in another pre-processed corpus, a corpus of reference (CR). The Portuguese reference corpus has about 10,000,000 occurrences, which corresponds to approximately 542,000 different forms. It is a non-technical corpus. In our study, the input text can be made by an ancient orthography or an adapted one, but it will be compared with the same modern Portuguese corpus, the CR. The CR is a “resident” part of the TermoStat system for its Portuguese module.

On the other hand, AntConc is a freeware corpus analysis toolkit. This tool is useful for searching words in context and helps us to do different kinds of text analysis. AntConc, for example, allowed us to observe the usage of repeated stock phrases throughout much of the text. With AntConc, we can also make a wordlist of a whole text or texts and compare their frequencies. As TermoStat, AntConc receives, as input, a text file that will be processed. This software identifies each set of text characters which is separated by a blank as a “word” (token). Numbers and punctuation marks used in the text are disregarded. Thus, if we have in the ancient corpus three different forms of a Portuguese ancient word (today: PURGAÇÃO [PURGING, using laxatives], as PURGAÇÃO and PURGAÇÃO or PURGAÇAM, the AntConc system will identify them as three different “words”. The same will happen with any flexional forms/variants, as plural and singular for Portuguese nouns, as the word MULHER [WOMAN] or MULHERES [WOMEN].

3. Steps of the pilot study

Some initial results of an experiment, only with the above-mentioned Semedos’ sample processed by AntConc and TermoStat tools, indicate the advantages of dealing with the old orthographic forms (Gonçalves, 2003). More details are described by Finatto (2018). For an initial test, the performance of these tools was compared in processing the old spelling and the updated spelling. Figure 3 shows a complete page of Semedos’ book and illustrates some special examples of problems in handling the orthographic system of this kind of ancient printed material.

As the Figure 3 exemplifies, there is a lot of orthographic challenges to face with our OCR systems and even with the typographical conventions. To support a future large scale better optical character recognition, it will be to use necessary different resources. One option to help us with the tasks of the corpus development with our students is the eDictor system, a tool for philological edition and automatic linguistic annotations (Sousa, Kepler and Faria, 2013). We intend to explore this system in the frame of the above cited e-learning environment “Terminologia Histórica”. The Version beta 1.0 of the eDictor was developed in 2007 (https://www.ime.usp.br/~tycho/participants/psousa/edicto).
r/presentation/edictor_2007.html, and this first version already contained the core functions of the tool: an XML annotation module, the possibility of XSLT transformation, exportation, and a morphosyntactic (Part of Speech) tagging function.

Then, with the TermoStat, tool described above, we have contrasted the frequencies and word distributions used in the old text with the word frequencies of its collection of texts with current Portuguese spelling. With TermoStat, we would argue, in thesis, the major peculiarities of Observação XCII regarding the statistical distribution of a specific vocabulary of the past in relation to a current and broader vocabulary. The test with AntConc was productive. That is, it has met the challenge of recognizing the words in their original (not modernised spelling of our 18th century medical text, even though it was not developed for this purpose. It is worth mentioning that it handled well the diversity and frequency of graphic forms, especially with the measure of the proportional variety of vocabulary (measure known as ‘Type-Token Ratio’) and indication of the proportion of words of single occurrence.

On the other hand, TermoStat worked by identifying and categorizing “words” by morphological classes, then contrasting the vocabulary of the segment from Semedos’ with a large collection of current texts. The results with this tool require further studies on its modes of functioning and performance with ancient texts. It is necessary to consider what this system does, “its statistical functioning and performance with ancient texts. It is this tool require further studies on its modes of functioning and performance with ancient texts. It is necessary to consider what this system does, “its statistical functioning and performance with ancient texts. It is necessary to consider what this system does, “its statistical functioning and performance with ancient texts. It is necessary to consider what this system does, “its statistical functioning and performance with ancient texts.
itself] – along 635 pages, but there are only 10 occurrences for BEXIGA [URINARY BLADDER], word in the singular.

In another contextual frame, designed by the corpus of Gazetas Manuscritas, considered as an ancient journalistic text, we can count 29 occurrences of the word BEXIGAS [in SMALLPOX sense]. Below, we have an example – with the ancient orthography – of an entry of the word BEXIGAS and SANGRIAS/SANGRALO [related to BLEEDINGS]. The emphasis in bold does not exist in the original text:

Com grande susto esteve a corte em húia grande febre do Prinçipe e na contenda dos medicos duvidando húis, e querendo outros sangralo prevaleço a opiniao de que não fizeçe este remedio, e secou a febre de todos os sintomas sahindo húia espeçie de bexigas, tão benigna que senão fosse preçizo á fineza da Prinçeza bem podião chamar-se com outro epiteto, houve preçes, e assistencia dos reys, e de toda a corte foi, como mereçia couza tão justa. A Prinçeza ja se levanta, o Sr. Jrnante D. Carlos melhorou com as sangrias.

Table 1 below shows a comparison of the top-10 nominal expressions in examined Semedo’s book segment Observaçam XCII and in Gazetas Manuscritas.

<table>
<thead>
<tr>
<th>Noun Gazetas Manuscritas</th>
<th>Frequency</th>
<th>Noun Gazetas Manuscritas</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>Parto</td>
<td>15</td>
<td>rey</td>
<td>1004</td>
</tr>
<tr>
<td>Sangria</td>
<td>14</td>
<td>sra</td>
<td>344</td>
</tr>
<tr>
<td>Febre</td>
<td>11</td>
<td>conde</td>
<td>309</td>
</tr>
<tr>
<td>Natureza</td>
<td>9</td>
<td>antonio</td>
<td>188</td>
</tr>
<tr>
<td>Purga</td>
<td>5</td>
<td>duque</td>
<td>183</td>
</tr>
<tr>
<td>Humor</td>
<td>5</td>
<td>infante</td>
<td>142</td>
</tr>
<tr>
<td>Medico</td>
<td>4</td>
<td>eza</td>
<td>137</td>
</tr>
<tr>
<td>Galeno</td>
<td>4</td>
<td>Sr</td>
<td>136</td>
</tr>
<tr>
<td>Purgação</td>
<td>4</td>
<td>annos</td>
<td>134</td>
</tr>
<tr>
<td>Puêrperio</td>
<td>3</td>
<td>diario</td>
<td>101</td>
</tr>
</tbody>
</table>

Table 2 and Table 3 show examples of the most frequent nominal expressions in Semedos’ Observaçam XCII and in Gazetas Manuscritas, respectively.

Table 2: Distribution of nominal expressions in Semedo’s segment book Observaçam XCII

<table>
<thead>
<tr>
<th>Gazetas Manuscritas</th>
<th>%</th>
<th>Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>N</td>
<td>44</td>
<td>rey, conde, filho, dia, srã, cruzado, antonio, duque</td>
</tr>
<tr>
<td>N+N</td>
<td>36</td>
<td>el rey, d. maria, d. antonio, d. anna, s. francisco, d. manuel, d. joão, d. lourenço, campo grande, del rey</td>
</tr>
</tbody>
</table>

As we can see the top-10 nominal expressions are totally distinct, and they reflect the “textual genres” of both texts. In Semedos’segment book the most frequent item is PARTO [childbirth] while in Gazetas the top lexical item is REY [the king]. Indeed, the textual genre not only determines certain terminology characteristics, but the textual genre is also determined by certain factors. As Santos and Costa (2015: 160) point out “texts are the result of social and discursive activities” and “when considered from this perspective, texts are not only linguistics artefacts, but also the product of social, cultural and ideological factors”.
Table 3: Distribution of nominal expressions in *Gazetas Manuscritas*

<table>
<thead>
<tr>
<th>Nominal Expression</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>N+prep+N</td>
<td>15</td>
</tr>
<tr>
<td>N+adj</td>
<td>2</td>
</tr>
<tr>
<td>N+N+N+V</td>
<td>2</td>
</tr>
</tbody>
</table>

4. Initial results: some considerations

As a result of our initial tests with the selected tools, we want to emphasize the importance to have historical corpora - especially in Portuguese - for different kinds of researches in Lexicology, Terminology and related areas as well as indicate the importance of diachronic studies of vocabulary and medical terminologies in ancient documents. However, besides the computational dimension highlighted here, an explicative philological-historical component should be included. This component, of course, is something that needs to be included in the online learning environment in which the corpus and computational tools to explore it will be offered.

Words identified as frequent and as "terminologies" by the computational tools or by a human reader have a source and a history. These ancient terminologies appear in Semedos’ medical handbook as a particular conception of the functions of the human body. Thus, the vocabulary profile of the text manifests an epistemology of the late 17th and early 18th century. It is also concerned to the Semedos' scientific points of view before the Linnaean taxonomy and this scientific revolution to mankind. This prism related to these documental corpora is relevant to understand the language and terminology of the time, besides the automatic and comparative data. This shows a frame of elements that should be considered beyond quantitative evidences.

In addition, Semedo's proposal that intended to present these type of Medicine language, vocabularies and terminologies in a way to make it accessible to their readers serves as a good inspiration for today’s researchers on the topic “plain language” for lower literate audiences.

5. Acknowledgments

We register here our special acknowledgments to the National Council for Scientific and Technological Development (CNPq), trough the financial lines PQ and *Edital Universal*, and to The Coordination for the Improvement of University Personnel (CAPES) in the scope of the Estágio Sênior CAPES, Proc. 88881.119078/2016-01, these are the foremost agencies to promote and support academic research in Brazil. We also thank for special support to the Interdisciplinary Center for History, Culture and Societies, University of Évora (CIDEHUS-UE/FCT), Portugal.

6. References


