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 Colegiais
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, conduced 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 ***Observaçoens medicas doutrinaes 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.

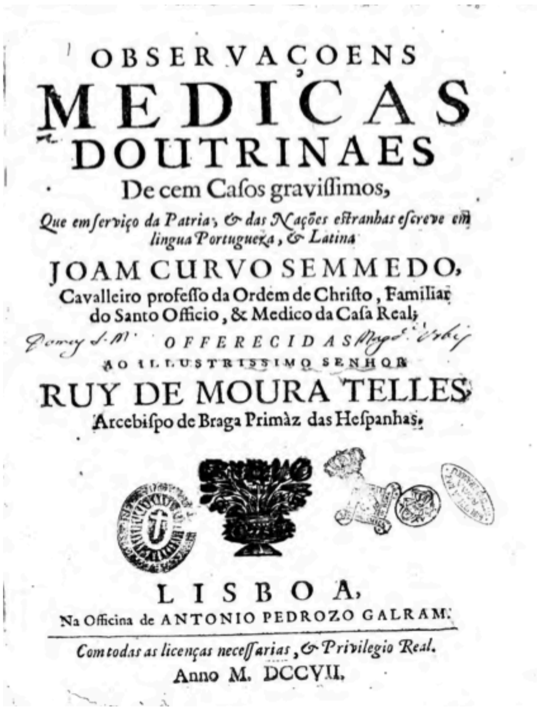
On the other hand, the works of Semedo inspired other treatises, namely works published by Portuguese doctors who practiced Medicine in Brazil. Thus, his book ***Observaçoens medicas doutrinaes de cem casos gravissimos*** (hereinafter ***Observaçoens***) and others are relevant to the history of Medicine 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. Semedos' work was also chosen because it was not registered in any of the great historical corpora, not even by Mark Davies'Corpus[1], 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.



OBSERVAÇOENS
MEDICAS
DOUTRINAES
De cem Caſos graviſſimos,
Que em ſerviço da Patria, & das Nações eſtranhas eſcreve em lingua Portugueza, & Latina
JOAM CURVO SEMMEDO,
Cavalleiro profeſſo da Ordem de Chriſto, Familiar do Santo Officio, & Medico da Caſa Real;
OFFERECIDAS
AO ILLUSTRISSIMO SENHOR
RUY DE MOURA TELLES
Arcebiſpo de Braga Primàz das Heſpanhas.

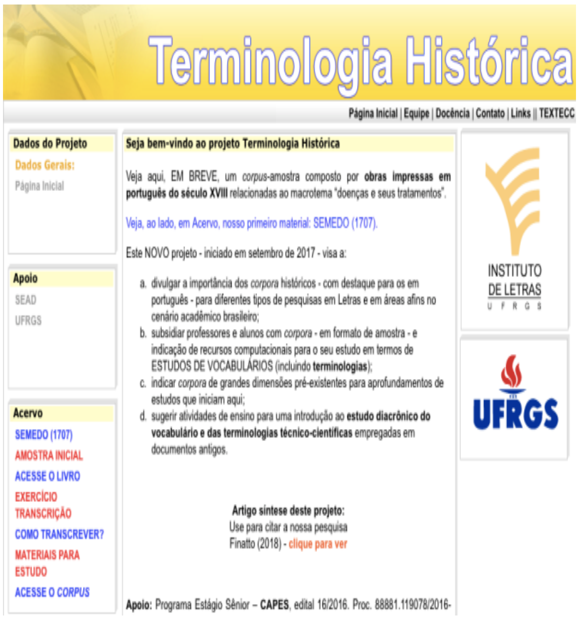
LISBOA,
Na Officina de ANTONIO PEDROZO GALRAM.
Com todas as licenças neceſſarias, & Privilegio Real.
Anno M. DCCVII.

Figure 1: The frontpage of Semedo (1707).
Scanned version by 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 and/or 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.

the user to access Semedo’s book according to the scanned version freely offered by Google Books.



Seja bem-vindo ao projeto Terminologia Histórica

Veja aqui, EM BREVE, um *corpus*-amostra composto por **obras impressas em português do século XVIII** relacionadas ao macrotema "doenças e seus tratamentos".

Veja, ao lado, em Acervo, nosso primeiro material: SEMEDO (1707).

Este NOVO projeto - iniciado em setembro de 2017 - visa a:

a. divulgar a importância dos *corpora* históricos - com destaque para os em português - para diferentes tipos de pesquisas em Letras e em áreas afins no cenário acadêmico brasileiro;
b. subsidiar professores e alunos com *corpora* - em formato de amostra - e indicação de recursos computacionais para o seu estudo em termos de ESTUDOS DE VOCABULÁRIOS (incluindo **terminologias**);
c. indicar *corpora* de grandes dimensões pré-existentes para aprofundamentos de estudos que iniciam aqui;
d. sugerir atividades de ensino para uma introdução ao **estudo diacrônico do vocabulário e das terminologias técnico-científicas** empregadas em documentos antigos.

Artigo síntese deste projeto:
Use para citar a nossa pesquisa
Finatto (2018) - clique para ver

Apoio: Programa Estágio Sênior – **CAPES**, edital 16/2016. Proc. 88881.119078/2016-

Figure 2: The draft version of the website already available at http://www.ufrgs.br/textecc/terminologia/

In order to feed this website and its tools, a pilot study was conducted with Semedos’ books content. The objective was to verify the advantages and disadvantages of the treatment of a set of texts with the original spelling and with the updated spelling. For this purpose, two free access computational tools for corpora processing were tested, AntConc (Antony, 2014) and TermoStat (Drouin, 2003). It is important to emphasize that both tools, developed by Corpus Linguistics researchers, are not built to deal with ancient texts orthography and old print characters. This means that the above-mentioned tools raise a few problems of philological nature, since, in order to comply with the text features, it is necessary to transcribe them and to prepare digital editions (Crane et al., 2008; Paixão de Sousa 2013a, 2013b). The tools will be very briefly described in the next section.

From Semedos’ book only a complete section with 1,317 graphic words considering its spelling was examined. This excerpt, named ***Observaçam XCII*** (pages 528–532), is just one of the 101 that make up the whole book.

In addition, this sample was contrasted with the collection called ***Gazetas Manuscritas*** of the Évora Library (see a part of this in Menezes, 1673), a corpus of ancient journalistic texts (Quaresma, 2016). This is a large set of journalistic texts from the 18th century handwritten in Portuguese. Thus, the ***Gazetas Manuscritas*** [freely translated as “The handwritten News”] was considered as a contrastive reference corpus. In a document with 480,366 characters and 14,832 types (different items), a sample with 85,517 words/tokens was chosen for this contrast. This material is partially available – in a transcribed version – at the Tycho Brahe Corpus (Sousa, 2014): http://www.tycho.iel.unicamp.br/corpus/.

2. The tools for text processing tests

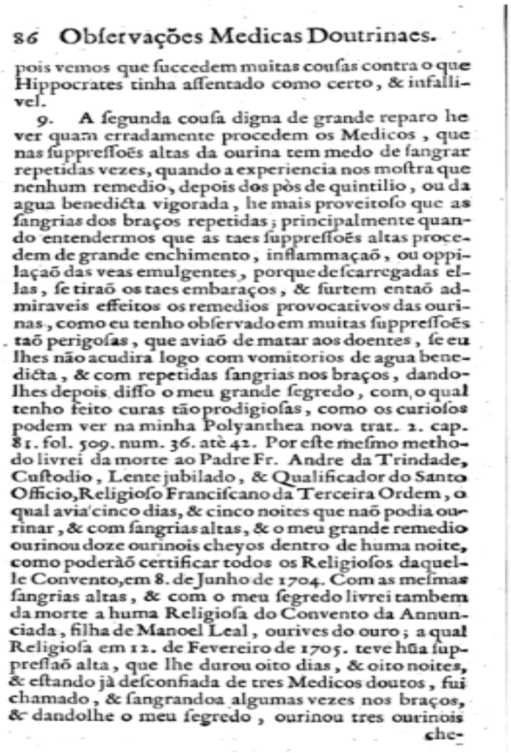
TermoStat receives an input text and returns as a main result a list of candidate terms (CT) derived from the text. A term – or a specific word item – can be either simple (a word) or complex (a sequence of words). Each term receives a score based on the frequency of the term in the analyzed corpus, 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 PURGAÇAÕ 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.



86 Obſervações Medicas Doutrinaes.

pois vemos que ſuccedem muitas couſas contra o que Hippocrates tinha aſſentado como certo, & infallivel.

9. A ſegunda couſa digna de grande reparo he ver quam erradamente procedem os Medicos, que nas ſuppreſſoẽs altas da ourina tem medo de ſangrar repetidas vezes, quando a experiencia nos moſtra que nenhum remedio, depois dos pòs de quintilio, ou da agua benedicta vigorada, he mais proveitoſo que as ſangrias dos braços repetidas; principalmente quando entendermos que as taes ſuppreſſoẽs altas procedem de grande enchimento, inflammaçaõ, ou oppilaçaõ das veas emulgentes, porque deſcarregadas ellas, ſe tiraõ os taes embaraços, & ſurtem entaõ admiraveis effeitos os remedios provocativos das ourinas, como eu tenho obſervado em muitas ſuppreſſoẽs taõ perigoſas, que aviaõ de matar aos doentes, ſe eu lhes não acudira logo com vomitorios de agua benedicta, & com repetidas ſangrias nos braços, dandolhes depois diſſo o meu grande ſegredo, com o qual tenho feito curas tão prodigioſas, como os curioſos podem ver na minha Polyanthea nova trat. 2. cap. 81. fol. 509. num. 36. atè 42. Por eſte meſmo methodo livrei da morte ao Padre Fr. Andre da Trindade, Cuſtodio, Lente jubilado, & Qualificador do Santo Officio, Religioſo Franciſcano da Terceira Ordem, o qual avia cinco dias, & cinco noites que naõ podia ourinar, & com ſangrias altas, & o meu grande remedio ourinou doze ourinois cheyos dentro de huma noite, como poderàõ certificar todos os Religioſos daquelle Convento, em 8. de Junho de 1704. Com as meſmas ſangrias altas, & com o meu ſegredo livrei tambem da morte a huma Religioſa do Convento da Annunciada, filha de Manoel Leal, ourives do ouro; a qual Religioſa em 12. de Fevereiro de 1705. teve hũa ſuppreſſaõ alta, que lhe durou oito dias, & oito noites, & eſtando jà deſconfiada de tres Medicos doutos, fui chamado, & ſangrandoa algumas vezes nos braços, & dandolhe o meu ſegredo, ourinou tres ourinois che-

Figure 3: The page 86 of the Semedo's book ***Observaçoens***

A second round of testing involved the comparison between the ***Gazetas Manuscritas*** sample and ***Observaçam XCII***. These two steps, dealed only with the TermoStat and AntConc systems, are summarized below.

3.1. The first step

With the AntConc tool, a list of all the words from the text of ***Observaçam XCII*** according to the old original spelling was produced. It was a list with 1,317 words (tokens), where 536 were different word forms (types). In the proportion between types-tokens, with which the variety of the vocabulary of the text is estimated, the segment showed 40% of vocabulary variety and a set of 355 words of single occurrence (called *Hapax legomena*).

Below, we have an example of Semedos' book – with the ancient orthography – with entries of the words CAMARAS [today: EPISODES OF DIARRHEA], FEBRE [FEVER] and SANGRIAS/SANGRASSEM [related to BLEEDINGS]. The emphasis in bold does not exist in the original text:

> Em 14 de Outubro de 1702. fuy chamado para visitar a senhora Dona Violante Casimira Saldanha a quem Deos tinha feito merce de dar hum filho desejado com ancja & conseguido com grande alegria; mas como as felicidades temporaes sejaó mui pensionadas, & cheyas de sobresaltos, ao gosto do feliz nascimento se seguio o temor, & tristeza, com humas **camaras**, **febre**, & falta da descarga devida ao puerperio: perturbàraõ muito estes symptomas naó só aos pays da recem nascida criança, mas aos patentes, &familiares da casa, porque tinhaõ ouvido dizer, que **camaras** sobre parto eraõ muito para temidas: para se desatar este no Gordonio, naó obstante que na visita da tarde tinha dado ordem a que pela manháa **sangrassem** a dita senhora, o naó quizeraó fazer sem que eu a visitasse primeiro, porque entendèraó que os cursos era hum grande impedimento para a **sangria**;

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 ***Observaçam 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 guidelines", with the classification of invalid spellings and how it assesses "errors" – the unknown words – that are not recognized by their morphosyntactic parser. Although the contrast allowed by TermoStat is between the words of the unique old text versus a large number of modern texts, we believe that it could be used for some purposes, even if the old-modern comparison can be considered unequal and problematic. As TermoStat pointed out, the words SANGRIA [BLEEDING], MEDICO [DOCTOR] and PURGAÇAÕ [PURGING, using laxatives] are the most typical items with the ancient text. For the modern one, it showed the items PURGAÇÃO, SANGRIA and PURGAÇÃO LOQUIAL [CHILDBIRTH'S PURGING].

3.2. The second step

For a second set of tests we dealt only with texts in the old orthography version and only with the TermoStat tools. As the tool system showed, the main words of Semedos' ***Observaçam XCII*** are SANGRIA, FEBRE [FEVER], PURGA and MEDICO. These words also appear in the ***Gazetas Manuscritas*** text, but not with the highest frequency, as would be expected of a non-specific corpus of Medicine.

On the other hand, if we consider Semedo's entire book (1707), as a medical handbook, there is only 01 occurrence of the item BEXIGAS (plural) [WOUNDS CAUSED BY SMALLPOX or SMALLPOX, the disease 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 hũa grande febre do Prinçipe e na contenda dos medicos duvidando hũns, e querendo outros **sangralo** prevaleço a opinião de que não fizeçe este remedio, e secou a febre de todos os sintomas sahindo hũa 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. Jnfante D. Carllos 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***.

Noun ***Observaçam XCII***	Frequency	Noun ***Gazetas Manuscritas***	Frequency
Parto	15	rey	1004
Sangria	14	sra	344
Febre	11	conde	309
Natureza	9	antonio	188
Purga	5	duque	183
Humor	5	infante	142
Medico	4	caza	137
Galeno	4	Sr	136
Purgaçaõ	4	annos	134
Puerperio	3	diario	101

Table 1: The comparison of the top-10 nominal expressions in Semedo's book ***Observaçam XCII*** and in ***Gazetas Manuscritas***

As we can see the top-10 nominal expressions are totally distinct, and they reflect the "textual genres" of both texts. In Semedos'book segment 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".

It is also interesting to compare the way nominal expressions are created in both textual genres: "noun" is the most frequent word class, but in the ***Gazetas Manuscritas*** there is a high frequency of 'noun + noun' (36.0%) and in Semedo's this represents only 5.0%. Moreover, in Semedo's book the use of multi-word complex nominal expressions including adjectives has a higher frequency than in ***Gazetas Manuscritas*** (25.0% versus 2.0%). This fact suggests the need for more complex nominal structures to describe medical situations in comparison with a general domain text.

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

Semedos' *Observaçam XCII*	%	Examples
Noun (N)	62	parto, sangria, febre, natureza, mulher, falta, caso, perigo, humor, pé
N+prep+N	20	purgação de parto, falta de purgação, sangria de pé, via de purga, sangria de pès, inchação de pé, vizinho de parto, sinaes de crueza, enchimento de sangue, natureza de humor
N+adj	10	caso semelhante, purgação loquial, purgaçã principiante, humor cacochymicos, sentença definitivo, perigo urgente, varão douto, filho desejado, caminho errado
N+N	5	felicidade temporaes, reynavão soro, valerio martins
N+prep+N+adj	5	falta de purgação mensal, embaraço a purgaçã principiante, falta de purgação loquial

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

Gazetas Manuscritas	%	Examples
N	44	rey, conde, filho, dia, sñra, cruzado, antonio, duque
N+N	36	el rey, d. maria, d. antonio, d. anna, s. francisco, d. manoel, d. joão, d. lourenço, campo grande, del rey

N+prep+N	15	filho de conde, duque de cadaval, secretario de estado, conde de assumar, joão de saldanha, rey de frança, cardeal de cunha, duque de aveyro, marques de alegrete, marques de abrantes
N+adj	2	monteiro mor, filho unico, diamante brilhante, sargento mor, camareira mor
N+N+N	2	jnfante d. francisco, jnfante d. carllos, jnfante d. antonio, jnfante d. carlos, el rey catholico, d. anna joaquina, assumar d. pedro, el rey stanislao, jnfante d. manoel, jnfanta d. francisca

Table 3: Distribution of nominal expressions in ***Gazetas Manuscritas***

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

Laurence Anthony. 2014. AntConc (Version 3.4.3) [Computer Software]. Tokyo, Japan, Waseda University. Available from http://www.antlab.sci.waseda.ac.jp/.

Ana Paula Banza and Maria Filomena Gonçalves. 2018. *Roteiro de História da Língua Portuguesa*. Évora, UNESCO Chair in Intangible Heritage and Traditional Know-How: Linking Heritage. http://www.catedra.uevora.pt/unesco/index.php/unesco/Investigacao/Publications-et-al/Books/Roteiro-de-Historia-da-Lingua-Portuguesa.

David Berry and Anders Fagerjord. 2017. *Digital Humanities: knowledge and critique in digital Age*. Cambridge, UK/Malden, Ms, Polity Press.

Rafael Bluteau. 1712-1728. *Vocabulario portuguez e latino* (...). Vol. 1-4 (1712-1713), Coimbra, Colegio das Artes; Vol. 5-8 (171-1721), Lisboa, Pascoal da Sylva; *Supplemento ao Vocabulario Portuguez e Latino*, Vol. 1, Lisboa, Joseph Antonio da Silva; Vol. 2 (1728), Lisboa, Patriarchal Officina da Musica.

Ivo Castro. 2006. *Introdução à história do Português*. 2ª ed. revista e ampliada. Lisboa, Edições Colibri.

Claudia Claridge. 2008. Historical corpora. In: A. Lüdeling and M. Kytö ed., *Corpus linguistics*: *an international handbook*. Berlin/New York: Walter de Gruyter, Handbooks of Linguistics and Communication Science/Handbücher zur Sprach und Kommunikationswissenschaft 29.1-2.

Gregory Crane, David Bamman and Alison Jones. 2008. ePhilology: when the books talk to their readers. In: S. Schreibman and R. Siemens eds., *A Companion to Digital Literary Studies*. Oxford, Blackwell.

Matthew James Driscoll and Elena Pierazzo eds. 2016. *Digital scholarly editing: theories and practices*. Digital Humanities Series, Vol. 4. Open Book Publishers.

Patrick Drouin. 2003. Term extraction using non-technical corpora as a point of leverage. *Terminology*. 9(1):99–117. doi:10.1075/term.9.1.06dro.

Maria José Bocorny Finatto. 2018. Corpus-amostra português do século XVIII: textos antigos de Medicina em atividades de ensino e pesquisa. *Domínios de Linguagem,*12(1):434−464.doi:http://dx.doi.org/10.14393/DL33-v12n1a2018-15.

Heather Froehlich. 2015. *Tutorial. Corpus Analysis with Antconc.*https://programminghistorian.org/lessons/corpus-analysis-with-antconc#introduction.

Júnia Ferreira Furtado. 2008. Tropical empiricism: making medical knowledge in colonial Brazil. In: James Delbourgo and Nicholas Dew ed., *Science and empire in the Atlantic world*, pages 127–152. New York/London, Routledge.

Maria Filomena Gonçalves. 2003. *As ideias ortográficas em Portugal: de Madureira Feijó a Gonçalves Viana (1734-1911)*. Lisboa, Fundação Calouste Gulbenkian/Fundação para a Ciência e Tecnologia.

Hendrick J. Kockaert and Friede Steurs eds. 2015. *Handbook of Terminology*, Vol. 1. Amsterdam/Philadelphia, John Benjamins.

Merja Kytö. 2011. Corpora and historical linguistics. *Revista Brasileira de Linguística Aplicada*, 11(2): 417-457. https://dx.doi.org/10.1590/S1984-63982011000200007.

Rita Marquilhas and Iris Hendrickx. 2016. Avanços nas humanidades digitais. In: A. Maria Martins and Ernestina Carrilho eds., *Manual de Linguística Portuguesa*. MRL 16, pages 252–277. Berlin/Boston, De Gruyter,

Francisco Xavier de Menezes. 1673. *Gazetas manuscritas da Biblioteca de Évora*. Vol. I (1729-1731). http://www.tycho.iel.unicamp.br/corpus/texts/xml/m_008.

Maria Clara Paixão de Sousa. 2013a. A Filologia Digital em Língua Portuguesa: Alguns caminhos. In: Maria Filomena Gonçalves and Ana Paula Banza coord., *Património textual e Humanidades Digitais. Da antiga à nova Filologia*, pages113-138. Évora, Centro Interdisciplinar de História, Culturas e Sociedades da Universidade de Évora (CIDEHUS)/ Fundação para a Ciência e Tecnologia (FCT). http://books.openedition.org/cidehus/1089.

Maria Clara Paixão de Sousa. 2013b. Texto digital: uma perspectiva material. *Revista da ANPOLL*, 35: 17–60.

Maria Clara Paixão de Sousa. 2014. Tycho Brahe: contribuições para as humanidades digitais no Brasil. *Filologia e Linguística Portuguesa*, 16(nº esp. dez.): 53–93.

Maria Clara Paixão de Sousa; Fábio Natanael Kepler, Pablo Picasso Feliciano de Faria. 2013. *e-Dictor* (Version 1.0 Beta 10). Retrieved from https://edictor.net/download.

Paulo Quaresma. 2013. Análise linguística de documentos da Biblioteca Pública de Évora Uma abordagem informática. In: Maria Filomena Gonçalves and Ana Paula Banza coord., *Património Textual e Humanidades Digitais. Da antiga à nova Filologia,* pages 139-155. Évora, CIDEHUS. https://books.openedition.org/cidehus/1091.

Cláudia Santos and Rute Costa. 2015. Domain specificity: semasiological and onomasiological knowledge representation. In: H. J. Kockaert and F. Steurs eds., *Handbook of Terminology*, Vol. 1, pages 153–179. Amsterdam/Philadelphia, John Benjamins.

João Curvo Semedo. 1707. *Observaçoens medicas doutrinaes de cem casos graviissimos, que em serviço da pátria, & das nações estranhas escreve em língua portugueza, & latina*. Lisboa, Officina de Antonio Pedrozo Galram.

Telmo Verdelho. 1998. Terminologias na língua portuguesa (perspectiva histórica). In: Jenny Brumme ed., *La història dels llenguatges iberoromànics d'especialitat (segles XVII-XIX),* pages 98–131. Barcelona, Universitat Pompeu Fabra/Institut Universitari de Lingüistica Aplicada. http://clp.dlc.ua.pt/Publicacoes/Terminologias_lingua_portuguesa.pdf.