# PORTUGUESE CORPORA OF THE 18TH CENTURY: OLD MEDICINE TEXTS FOR TEACHING AND RESEARCH ACTIVITIES

Maria José Finatto

Universidade Federal do Rio Grande do Sul/Linguistics Department

Paulo Quaresma

Universidade de Évora, Department of Computer Science, Laboratory of Informatics, Systems and Parallelism

Maria Filomena Gonçalves

Universidade de Évora/ECS/ Linguistics Department

UNIVERSIDADE DE ÉVORA

FCT
Fundação para a Ciência e a Tecnologia
MINISTÉRIO DA EDUCAÇÃO E CIÊNCIA

CIDEHUS
Centro Interdisciplinar de História, Culturas e Sociedades da Universidade de Évora
UID/HIS/00057/2013

# INTRODUCTION

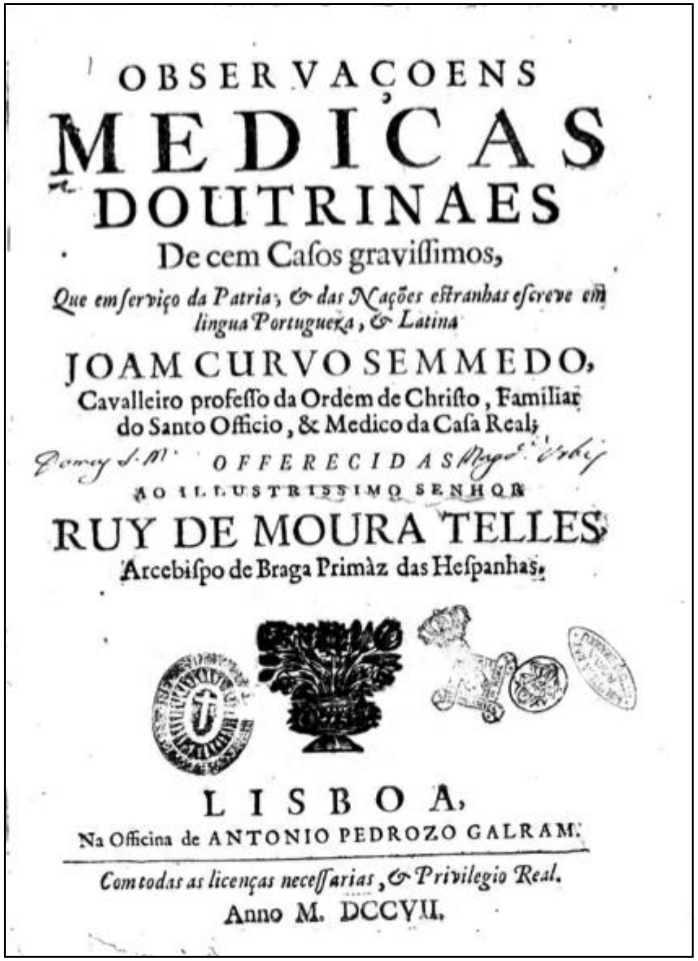
The main goal of this paper is to demonstrate the application of the methodologies of corpus linguistics and of Natural Language Processing (NLP) tools to an 18th century Portuguese medicine book.

The other goal 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.

- This paper 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*** (Curvo Semedo, 1707).
- Curvo Semedo (1635-1719), a Portuguese physician from Monforte, Alentejo, a region within Portugal.



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;

Domus S. M. *OFFERECIDAS* Magd. Urbis

AO ILLUSTRISSIMO SENHOR

RUY DE MOURA TELLES

Arcebiſpo de Braga Primàz das Heſpanhas.

LISBOA,
Na Officina de ANTONIO PEDROZO GALRAM.

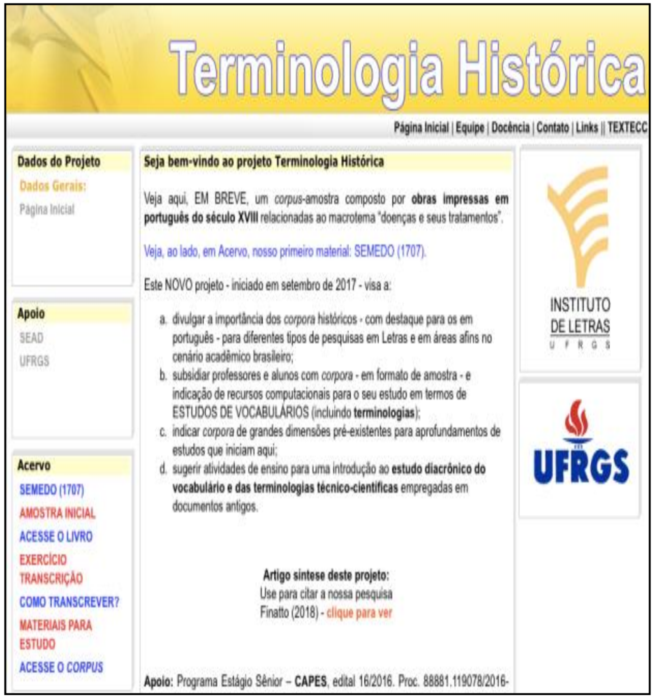
*Com todas as licenças neceſſarias, & Privilegio Real.*
Anno M. DCCVII.

- Semedos’ work was chosen because it was not registered in any of the great historical corpora, not even by Mark Davies’Corpus, which has 45 million keywords covering a period between 1200 and 1900.

- 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.



**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-

- 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).

# 2. THE TOOLS FOR TEXT PROCESSING TESTS

- 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
- 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).

# 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.

# 2. THE TOOLS FOR TEXT PROCESSING TESTS

- 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.
- 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.

# 2. THE TOOLS FOR TEXT PROCESSING TESTS

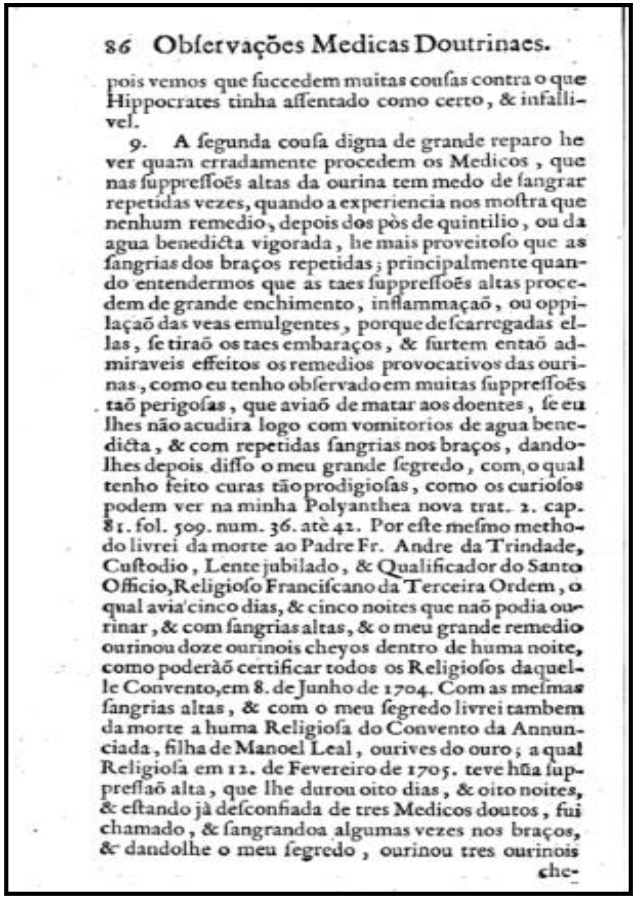
- If we have in the ancient corpus three different forms of a Portuguese ancient word (today: PURGAÇÃO [PURGING, using laxatives]), as PURGAÇAÕ and PURGAÇÃO or PURGAÇAM, the AntConc system will identify them as three different “words.

- 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.

# 3. STEPS OF THE PILOT STUDY



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-

There is a lot of orthographic challenges to face with our OCR systems and even with the typographical conventions.

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.

We intend to explore this system in the frame of the above cited e-learning environment "Terminologia Histórica".

# 3. STEPS OF THE PILOT STUDY

- 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.

- **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*).

# 3. STEPS OF THE PILOT STUDY

- 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.
- 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.
- The results with this tool require further studies on its modes of functioning and performance with ancient texts.

# 3. STEPS OF THE PILOT STUDY

- 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.

# 3. STEPS OF THE PILOT STUDY

- For the modern one, it showed the items PURGAÇÃO, SANGRIA and PURGAÇÃO LOQUIAL [CHILDBIRTH'S PURGING].
- **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.

# 3. STEPS OF THE PILOT STUDY

- 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].

# 3. STEPS OF THE PILOT STUDY

- 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.

- 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%.

# 3. STEPS OF THE PILOT STUDY

- 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.

# 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.

- 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.

# 4. INITIAL RESULTS: SOME CONSIDERATIONS

- 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.

# 4. INITIAL RESULTS: SOME CONSIDERATIONS

- 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.