DOI: 10.14393/DL33-v12n1a2018-15

Corpus-amostra português do século XVIII: textos antigos de medicina em atividades de ensino e pesquisa[1]

A sample of a Portuguese *corpus* of the XVIII century: old medicine texts for teaching and research activities

Maria José Bocorny Finatto[*]

RESUMO: De acordo com os princípios da Linguística de *Corpus*, este artigo apresenta um conjunto de procedimentos iniciais para o desenho de um *corpus* composto por uma amostra de textos médicos antigos impressos em português do século XVIII sobre o tema "doenças e seus tratamentos". Este *corpus*-amostra será parte de um ambiente virtual dedicado ao estudo de temas históricos de Lexicologia e de Terminologia. Um estudo piloto foi conduzido para verificar as vantagens e desvantagens do tratamento de um conjunto de textos com a ortografia original e com a ortografia atualizada com o uso de duas ferramentas computacionais para processamento de *corpora*, AntConc e TermoStat. Os resultados iniciais indicam vantagens de se trabalhar com as formas ortográficas antigas. Finalmente, o artigo destaca a importância dos acervos históricos - especialmente em português - para diferentes tipos de pesquisas em Lexicologia e áreas afins, além de indicar a importância dos estudos diacrônicos de vocabulário e terminologias médicas em documentos antigos.

ABSTRACT: According to principles of *Corpus* Linguistics, this article presents a set of initial procedures for the design of a *corpus* consisting of a sample of ancient medical texts printed in Portuguese of the XVIII century on the subject "diseases and their treatments". This *corpus*, as a sample, will be part of a virtual environment dedicated to the study of historical lexicology and terminology topics. A pilot study was conducted to verify the advantages and disadvantages of the treatment of a set of texts with the original spelling and with the updated spelling through the use of two computational tools for corpora processing, AntConc and TermoStat. The initial results indicate the advantages of dealing with the old orthographic forms. Finally, the article highlights the importance of historical corpora - especially in Portuguese - for different kinds of researches in Lexicology and related areas, as well indicates the importance of the diachronic studies of vocabulary and medical terminologies in ancient documents.

[1] O estudo-piloto aqui relatado foi apresentando, como comunicação oral, no XXXIII Encontro Nacional da Associação Portuguesa de Linguística, em setembro de 2017, ocorrido na Universidade de Évora, em Portugal.

[*] Professora do Programa de Pós-Graduação em Letras da UFRGS, em Porto Alegre - RS. Pesquisadora do CNPq. E-mail: mariafinatto@gmail.com.

PALAVRAS-CHAVE: Textos antigos. Textos de Medicina. Lexicologia histórica. Linguística de *Corpus*. Léxico-estatística.

KEYWORDS: Old texts. Medicine Texts. Historical Lexicology. *Corpus* Linguistics. Lexicostatistics.

1. Introdução

Corpora, conforme a Linguística de *Corpus* (doravante **LC)**, são acervos textuais transpostos para formato digital, criteriosamente reunidos, de acordo com objetivos específicos de estudo ou de descrição, passíveis de tratamento computacional, servindo para representar um dado estado de uso de língua (BERBER SARDINHA, 2004). Esses acervos digitais, mesmo para pesquisadores e projetos de investigação que não se alinhem ou se identifiquem especificamente com a **LC**, têm sido associados a muitas investigações da atualidade. Entre essas investigações, destacamos as que tratam do léxico de textos escritos em português, descrevendo a padronização, distribuição e especificidades das "palavras"[2]. Na **LC**, a qual compreendemos como um dos ramos da Linguística Aplicada, a língua é entendida como um sistema probabilístico de combinatórias, de modo que cada palavra do texto será definida pelo conjunto das relações que mantenha com outras palavras, em um dado tipo de uso. Hoje, mesmo que nunca se tenha ouvido falar em **LC**, cada vez mais lidamos com *corpora* digitais ao tratar dos fenômenos da linguagem, afinal, há um acesso facilitado a computadores e a miríades de fontes escritas na internet, sejam essas fontes as dispersas – em uma visão da *web* como *corpus* - ou acervos

[2] O termo "palavra", no âmbito dos Estudos da Linguagem, é bastante controverso, conforme já ensinou Biderman (1999), por isso o uso de aspas ao longo deste texto. O que seja uma "palavra" pode ser entendido de vários modos, daí haver as denominações "lexia", "lexema", "lema" ou "unidade/item lexical", entre outras, para corresponder a diferentes facetas desse conceito. Neste trabalho, dado nos guiarmos pela Linguística de *Corpus* (LC), trataremos "palavra" como sendo uma unidade de forma ligada a um sentido/significado, mormente a palavra escrita, entendida como um conjunto de caracteres impressos separados por um espaço em branco. A noção de palavra em LC também comporta diferentes concepções. Há que se considerar a condição de uso associada a um dado conteúdo para delimitar a dimensão semântica de uma palavra. Não aprofundaremos esta discussão aqui.

documentais temáticos e pontuais, especialmente organizados, para fins determinados.

Uma aproximação entre o investigador da linguagem, textos ou acervos textuais em formato digital e a **LC** pode dar-se: **a)** por sua metodologia, bastante marcada pelo uso de ferramentas informatizadas para verificação de usos padronizados de palavras em textos (*corpora*); ou, **b)** pela adesão do pesquisador a perspectiva teórica diferenciada da **LC** para os fenômenos linguísticos. Nesse caminho de aproximação, a **LC** atual, pelo menos no âmbito brasileiro, tornou-se um interessante ponto de encontro para diferentes estudos, independentemente de filiações teóricas ou áreas de formação dos pesquisadores envolvidos, observando-se ênfase para diferentes tratamentos do texto escrito, com destaque para questões sobre configuração do léxico e da gramática. A diversidade de temas de estudo e de enfoques hoje verificada em **LC** pode ser conferida pelos temas das diferentes edições do "Encontro de Linguística de *Corpus*" (ver http://www.ufrgs.br/elc-ebralc2017). Esse evento ocorre já há mais de 15 anos no Brasil, integrando linguistas – de várias áreas, diferentes estudiosos do Texto e do Discurso, estudiosos de Literatura, da Tradução e do Ensino, além de cientistas da Computação que lidam com o Processamento de Linguagem Natural (especialidade conhecida pela sigla **PLN**).

1.2 *Corpora* históricos do português

A produção de *corpora* digitais históricos, conforme Cucatto (2012), é bem menos abundante se comparada a *corpora* da atualidade. A escassez de materiais relacionados a "textos antigos em formato digital" dá-se, além de diferentes motivos, pela complexidade da tarefa de produzi-los. A isso soma-se o fato de o tratamento computacional das "palavras" em textos antigos, sejam eles impressos ou manuscritos, ser nada trivial.

Textos antigos, via de regra, precisam ser digitalizados, e seus conteúdos precisam ser tratados não somente como imagens. Apenas uma ou várias fotos de um documento-fonte não bastam, pois é preciso poder manipular as palavras, por unidades ou por frases, do texto que elas contenham. Enfim, produzir esses *corpora*, muitas vezes, envolve localizar acervos, identificar obras relevantes e manipular documentos delicados, muitas vezes bastante fragilizados pelo desgaste do papel e das tintas, sem contar os danos físicos pela guarda muitas vezes inadequada. Em resumo, fotografar, ler, interpretar, digitalizar e transcrever documentos antigos requer cuidado, conhecimento e técnicas específicas. Apenas depois de todo o trabalho paleográfico e filológico, é que se iniciam, propriamente, a descrição e a análise da linguagem escrita verificada, o que também é tarefa complexa, especialmente quando o investigador contextualiza todo um espaço-tempo enunciativo remoto.

Nesse tipo de pesquisa, seja realizada como Filologia, Lexicologia, Terminologia ou como Linguística Histórica, para os estudiosos do âmbito do léxico/vocabulário, geralmente será preciso deslocarem-se da dinâmica das grafias do passado à materialidade complexa da escrita como um todo simbólico multinível (DUARTE, 2012, p. 123). Quem lida com materiais escritos impressos, foco específico do nosso interesse neste trabalho e do nosso *corpus*-amostra a seguir caracterizado, enfrentará ortografias e tipos de impressão variáveis, complexidades de leitura e, muitas vezes, a fragilidade das fontes originais. Isso sem nos esquecermos de mencionar o caso de versões fotografadas pouco aproveitáveis, visto terem sido realizadas com tecnologias hoje limitadas.

No acesso do pesquisador ao documento impresso em suporte papel, mesmo que tenha sido corretamente digitalizado, a diversidade de recursos tipográficos – e o contexto de uma tecnologia de imprensa do passado (mais ou menos remoto) – tenderá a representar uma barreira importante para a etapa da composição de

arquivos de texto transcrito a partir da tecnologia de um reconhecimento ótico automático dos caracteres, um sistema atualmente conhecido como "OCR". Considerando esse estado de coisas e a nossa filiação teórico-metodológica à **LC** e aos estudos de Terminologia (KRIEGER; FINATTO, 2004), a partir de um estudo inicial, a seguir relatado, pretendemos contribuir para que:

a) haja mais e variados acervos com *corpora* de textos antigos impressos em português, disponíveis gratuitamente em formato *on-line;*

b) possa-se identificar alternativas para o processamento do texto antigo impresso, a partir da sua forma ou grafia original, no que se refere ao seu tratamento por unidades de palavras gráficas, via uso de ferramentas informatizadas.

A nossa contribuição, em termos mais concretos e imediatos, pelo que exporemos neste artigo, restringir-se-á a algo bastante modesto, envolvendo apenas um *corpus*-amostra inicial, de poucas proporções, planejado apenas para apoiar atividades didáticas no curso de Letras da Universidade Federal do Rio Grande do Sul (**UFRGS**), localizada na cidade de Porto Alegre – RS, sul do Brasil. São relatadas, a seguir, as condições e procedimentos iniciais para a composição do nosso *corpus*-amostra composto por obras impressas em português do século XVIII relacionadas ao macrotema "doenças e seus tratamentos". Nosso *corpus* inicial corresponde ao texto do livro intitulado ***"Observaçoens medicas doutrinaes de cem casos gravíssimos (…)"*** (SEMEDO, 1707) impresso em Lisboa, em 1707, de autoria do médico alentejano João Curvo Semedo (1635-1719). Essa obra conta com 635 páginas.

O desenho desse *corpus*, de propósitos didáticos, foi idealizado durante nosso Estágio Sênior CAPES, de abril a julho de 2017, sob a supervisão da Profa. Dra. Maria Filomena Gonçalves, especialista em Filologia e Gramática do Português, no qual nos ocupamos do tema da Terminologia Histórica, estudando textos antigos

sobre Medicina, publicados em português no século XVIII. Durante esse estágio, junto à Universidade de Évora, tivemos também a parceria do colega Prof. Dr. Paulo Quaresma, especialista no Processamento da Linguagem Natural (PLN), subárea da Ciência da Computação/Informática. O Prof. Quaresma, com seus colegas de grupo de pesquisa em PLN, tem lidado com técnicas automáticas de Recuperação de Informação e de tratamento da linguagem a partir de *corpora* de textos antigos em português, especialmente manuscritos e suas transcrições, seja em grafias originais ou nas adaptadas ao padrão de escrita atual.

Nosso *corpus*-amostra será abrigado em um ambiente virtual de apoio a atividades didáticas para estudantes de Letras hospedado em um servidor da **UFRGS.** Uma versão inicial desse ambiente acessa-se em http://www.ufrgs.br/textecc/ clicando-se na aba "**Terminologia Histórica**". O seu conteúdo está sendo planejado para apoiar atividades de ensino relacionadas com Terminologia diacrônica, Lexicologia e Lexicografia, além de pretender prestar apoio para estudantes e professores da disciplina "Linguística Histórica", a qual integra o currículo do curso de Letras da UFRGS.

O texto deste artigo prosseguirá com os seguintes tópicos: i) uma revisão sobre *corpora* históricos do português já disponíveis; ii) relato de um estudo inicial com um segmento da nossa fonte escrita do século XVIII – o qual foi processado, com duas ferramentas computacionais, em uma vez com a grafia original e, em outra, com a grafia atualizada; iii) discussão sobre o rendimento do processamento do texto com ambas as grafias e perspectivas para disponibilização da obra no ambiente virtual de aprendizagem – associando-se o *corpus* a ferramentas de busca de palavras.

2. Algumas iniciativas importantes

Apesar das dificuldades de produção desse tipo de *corpus*, composto por textos antigos, há vários esforços meritórios para a construção de uma "filologia

digital", na qual já se pode destacar grandes empreendimentos especialmente dedicados ao português (QUARESMA, 2013; PAIXÃO DE SOUSA, 2013).

Nesse caminho, há também hoje *softwares* especiais para apoiar o trabalho com o português antigo, como o E-Dictor (https://humanidadesdigitais.org/edictor/). Essa ferramenta, de acesso gratuito, permite armazenar e visualizar a imagem de um ou de muitos documento-fonte, sejam manuscritos ou impressos. Seu sistema permite que o usuário que lida com textos antigos componha arquivos com transcrições nas grafias original e atualizada, além de oferecer ao utilizador um recurso de buscas pontuais e uma ferramenta de categorização das palavras identificadas nos textos por classes e funções. Entretanto, vale frisar, o E-Dictor, conforme a versão para *download* que consultamos, ainda não gera, sozinho, o reconhecimento de caracteres do documento original e uma transcrição automática, mas é um importante avanço para o trabalho. Seu utilizador deve prover o sistema com a transcrição do documento, mas poderá editá-la em versão original e em versão com grafia atualizada, além de contar com etiquetagem morfossintática de palavras e com um léxico de edições. Com esse léxico, é possível padronizar as apresentações das grafias das palavras à medida que se revisa ou edita um texto sob estudo. Contando-se, assim, com um "dicionário armazenado no sistema" das formas variantes para a grafia de uma mesma palavra-base que ocorram em um mesmo dado documento. Esse recurso pode auxiliar o usuário a lidar com casos de alternância tais como o verificado no nosso *corpus* com as formas ***observaçam*** e ***observaçaó***.

A despeito de dificuldades diversas, especialmente as de obtenção de financiamento para estudos linguísticos históricos no cenário do Brasil atual, podemos destacar importantes acervos que contemplam documentos antigos em português. Sem pretensão de exaustividade, listamos, a seguir, algumas iniciativas de grande porte – conforme buscas e acessos aos *links* indicados. O funcionamento desses recursos foi por nós verificado em novembro de 2017:

a) *Corpus* do Português, Mark Davies, BYU

***Site*:** http://www.*corpus*doportugues.org/interface2016.asp

Descrição: iniciativa desenvolvida pelo Prof. Dr. Mark Davies na *Brigham Young University*, em Provo, Utah, E.U.A. Apresenta-se como uma base de dados com **45** milhões de palavras dos anos de **1200** a **1900**. Visa subsidiar a verificação da história do português. Permite toda uma variada gama de buscas, por períodos e por palavras, inclui dicionários do *corpus*.

Buscas: oferece diferentes tipos de buscas por palavras em www.*corpus*doportugues.org/hist-gen;

b) o *corpus* da pesquisa *Para uma História do Português Brasileiro* (**PHPB**, de 1998).

***Site*:** https://sites.google.com/site/corporaphpb/home

Descrição: o projeto **PHPB** foi organizado em 1998 na Faculdade de Letras da UFRJ – Rio de Janeiro, Brasil, integrando-se à proposta de trabalho coletivo lançada no I Seminário para a História do Português Brasileiro, realizado em abril de 1997 pelo Programa de Pós-Graduação em Filologia e Língua Portuguesa da Universidade de São Paulo (USP). Hoje o **PHPB** tem *sites* com desdobramentos que permitem acesso – mediante solicitação prévia à coordenação - a documentos coletados por regiões do Brasil, como, por exemplo, o projeto "Para uma História do Português Brasileiro do Rio Grande do Sul", fundado pela Profa. Dra. Valéria Neto de Oliveira Monaretto na Universidade Federal do Rio Grande do Sul – UFRGS, em 2013. Ver em: www.ufrgs.br/phpb-rs/sobre-o-phpb-rs/

Sua base principal (mantida na UFRJ – Rio de Janeiro - RJ) oferece acesso imediato a *corpora* manuscritos - séculos XVIII, XIX E XX, e impressos dos séculos XIX e XX.

Buscas: permite acesso aos textos (apenas o segmento RJ), sem ferramentas de busca. Veja-se um exemplo de busca em https://sites.google.com/site/corporaphpb/home/corpora-manuscritos/manuscritos-minas-gerais

c) o *corpus* histórico do *PE Tycho Brahe* (Fase I, 1998-2003, Fase II, 2008-2009)

***Site*:** http://www.tycho.iel.unicamp.br/*corpus*/

Descrição: *corpus* eletrônico anotado, composto de textos em português escritos por autores nascidos entre **1380** e **1881**. Compilado na UNICAMP, Campinas – SP, em projetos de pesquisa sob coordenação da Profa. Dra. Charlotte Galves. Oferece **76** textos (**3.303.196** palavras) para pesquisa livre, com um sistema de anotação linguística em duas etapas: **anotação morfológica** (aplicada em **44** textos, num total de **1.956.460** palavras); e **anotação sintática** (aplicada em **20** textos, num total de **877.247** palavras).

Buscas: permite *download* dos textos do *corpus* e oferece ferramentas de busca. Há necessidade de familiarização do usuário com seu sistema. Ver em: http://www.tycho.iel.unicamp.br/*corpus*/texts/csquery/csquery.html

d) o *corpus BIT-PROHPOR* (Programa para a História da Língua Portuguesa)

***Site*:** http://www.prohpor.org/bit-banco-textos

Descrição: o PROHPOR foi criado em 1990, sob coordenação da Profa. Dra. Rosa Virgínia Mattos e Silva, sua fundadora, na Universidade Federal da Bahia (UFBA), pesquisadora falecida em 2012. Inclui um Banco Informatizado de Textos (BIT) para a história da língua portuguesa. Oferece textos escritos em português, do século XIII ao século XVI (Português Arcaico) e do século XVII ao século XXI (Português do Brasil). Textos de natureza variada. Por meio da diversidade de registros/fontes, possibilita a percepção de variação sociocultural da língua. Junto aos textos, há informações extralinguísticas.

Oferece ligações – com as fontes e recursos de outros acervos – como o *Corpus Informatizado de Textos Portugueses Medievais* - o CIPM. Ver em: <http://cipm.fcsh.unl.pt/gencontent.jsp?id=4>

Buscas: Acesso a textos, sem ferramentas. Ver um exemplo de busca em: <http://www.prohpor.org/bit-banco-pb>

e) o acervo documental do *Dicionário Histórico do Português do Brasil* (**DHPB**)

Site: não localizamos nenhum *site* para acesso direto ao *corpus* reunido.

Descrição: o *corpus* reunido para o **DHPB** – séculos XVI, XVII e XVIII, finalizado em 2013. Não está disponível para acesso público. O DHPB, conforme foi finalizado e entregue ao órgão de fomento que o financiou (CNPq), ainda não está acessível para o público em geral. Conta com aproximadamente 7 milhões e 500 mil ocorrências que permitiram aos redatores do **DHPB** extraírem as unidades lexicais para a redação dos verbetes. Conforme relatou a sua coordenadora, a Profa. Dra. Clodilte de Azevedo Murakawa (MURAKAWA, 2013), o banco de textos do **DHPB** tem o ano de 1500 e a Carta de Pero Vaz de Caminha como ponto de partida para a seleção de documentos, sendo1808, ano da vinda da família real portuguesa para o Brasil, como a sua data-limite.

Buscas: restritas aos pesquisadores diretamente associados ao projeto de produção do **DHPB**. Outros detalhes sobre o dicionário e sobre o acervo/*corpus* encontram-se em:

http://www.academia.edu/16334685/Dicionário_histórico_do_português_do_Brasil_um_modelo_de_dicionário_histórico

3. Da nossa amostra de textos antigos

Nosso *corpus*-amostra de textos impressos do século XVIII está integrado à iniciativa "Terminologia histórica", no âmbito do Projeto TEXTECC www.ufrgs.br/textecc. Serão oferecidos conjuntos de textos e ferramentas *on-line*

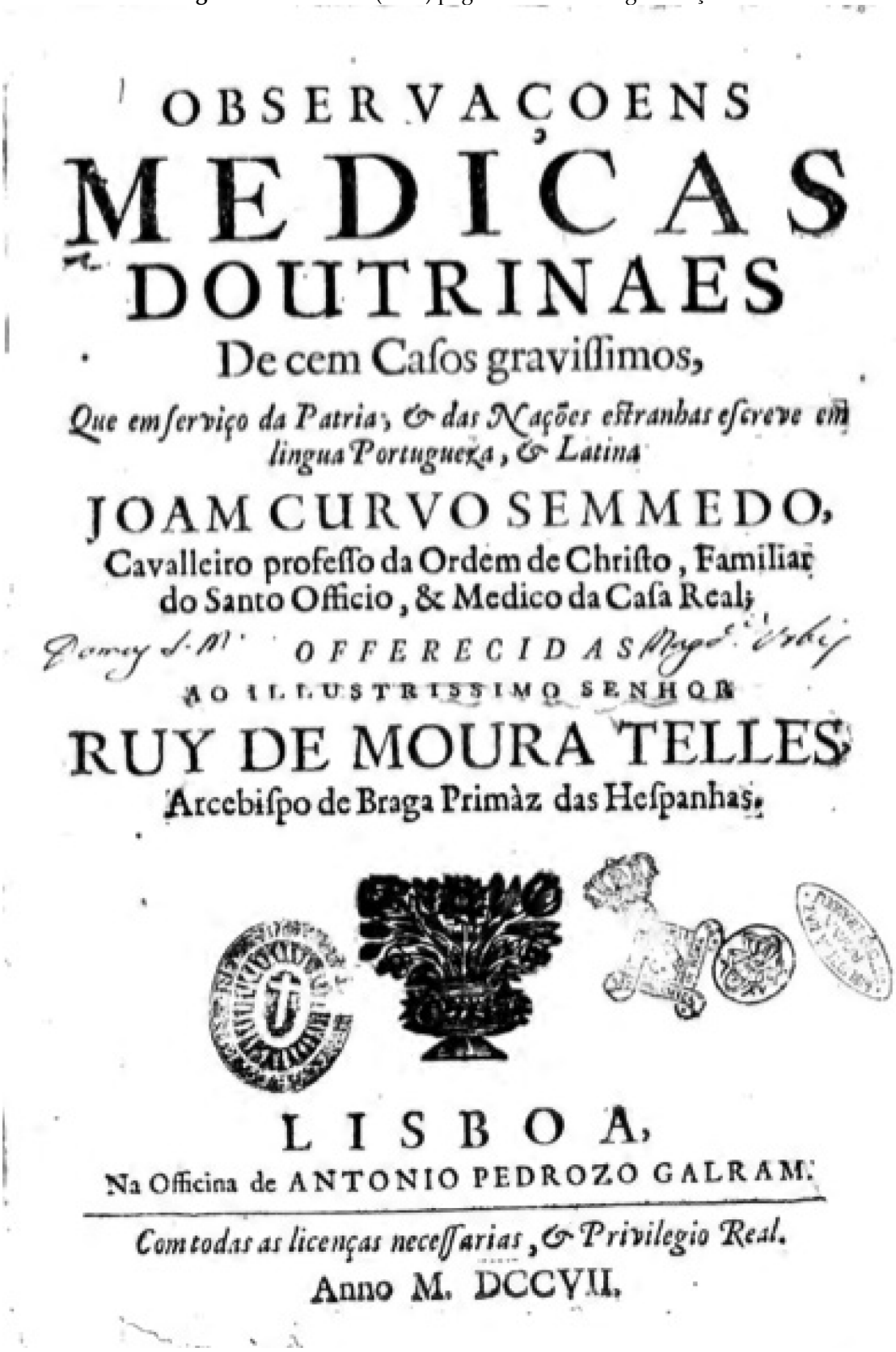
simples para sua exploração em um ambiente de estudos sobre a história da linguagem médica em português. As ferramentas previstas são: um listador de palavras, um gerador de contextos por expressão de busca num dado *corpus*/texto, e um gerador de listas de grupos de palavras que se repetem em blocos ao longo de um dado texto ou de vários textos.

Para projetar o funcionamento dessas ferramentas, partimos do texto do livro ***Observaçoens medicas doutrinaes de cem casos gravíssimos (…)*** impresso em Lisboa, em 1707, com 635 páginas, publicado pelo médico alentejano João Curvo Semedo (1635-1719). Essa obra, entre outras do autor, foi estudada no trabalho de mestrado de Lourenço (2016) "O médico entre a tradição e a inovação: João Curvo Semedo[3]". Semedo foi um dos médicos mais famosos do seu tempo. Médico da Casa Real de Portugal, costumava viajar à Espanha, onde era figura muito admirada, tendo obras citadas e traduzidas ou comentadas, editadas em castelhano. Semedo recebeu o grau de Cavaleiro da Ordem de Cristo e foi designado Familiar do Santo Ofício, tendo sido reconhecido como um inovador da farmacopeia da sua época.

Escolhemos essa obra de Semedo de 1707 justamente por ela não estar contemplada em nenhum dos *corpora* históricos de grande porte antes citados, mesmo no acervo de Mark Davies, que conta com 45 milhões de palavras de textos produzidos entre 1200 e 1900. Em formato arquivo, este livro digitalizado está disponível na íntegra em *Google Books*, com acesso livre e gratuito. Para o nosso trabalho de leitura, familiarização e de transcrição do seu texto, foi importante poder contar com uma outra digitalização completa realizada a partir de um original, fisicamente disponível no *Setor de Reservados* da Biblioteca Pública de Évora (BPE), em Portugal. O material digitalizado pela BPE foi-nos gentilmente cedido, em formato PDF, pela Profa. Dra. Maria Filomena Gonçalves, em abril de 2017. A seguir, a título de exemplo, temos algumas imagens desse livro de Semedo, extraídas do material digitalizado pela BPE.

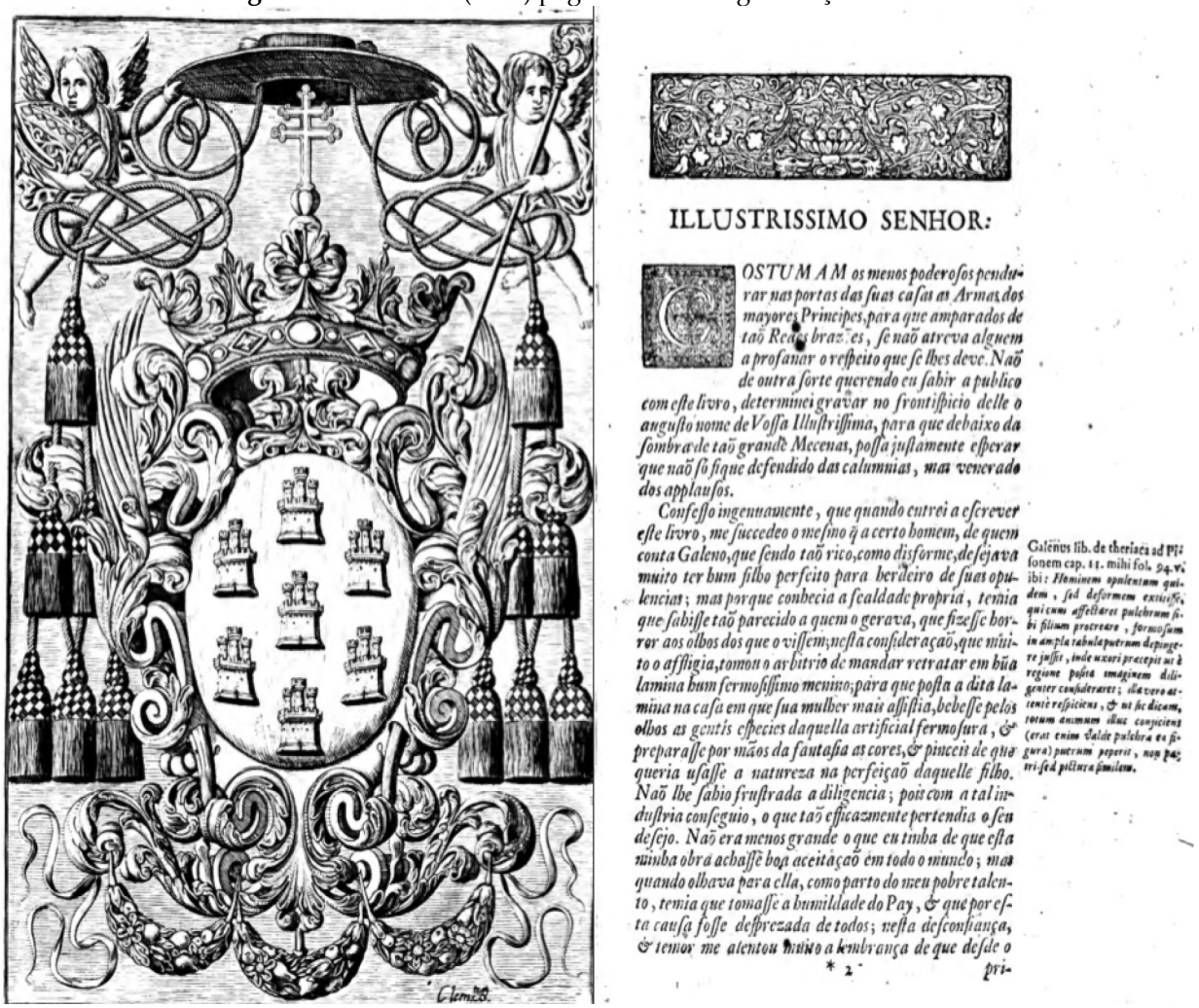
[3] Trabalho disponível em http://www.historia.uff.br/stricto/td/2002.pdf.

Figura 1 – SEMEDO (1707) página 01. Fonte: digitalização da 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.

Figura 2 – SEMEDO (1707) páginas 01-02. Digitalização da BPE.



ILLUSTRISSIMO SENHOR:

COSTUMAM os menos poderosos pendurar nas portas das suas casas as Armas dos mayores Principes, para que amparados de taõ Reaes braz~es, se naõ atreva alguem a profanar o respeito que se lhes deve. Naõ de outra sorte querendo eu fahir a publico com este livro, determinei gravar no frontispicio delle o augusto nome de Vossa Illustrissima, para que debaixo da sombra de taõ grande Mecenas, possa justamente esperar que naõ só fique defendido das calumnias, mas venerado dos applausos.

Confesso ingenuamente, que quando entrei a escrever este livro, me succedeo o mesmo q̃ a certo homem, de quem conta Galeno, que sendo taõ rico, como disforme, desejava muito ter hum filho perfeito para herdeiro de suas opulencias; mas porque conhecia a fealdade propria, temia que sahisse taõ parecido a quem o gerava, que fizesse horror aos olhos dos que o vissem; nesta consideraçaõ, que muito o affligia, tomou o arbitrio de mandar retratar em hũa lamina hum formosissimo menino; para que posta a dita lamina na casa em que sua mulher mais assistia, bebesse pelos olhos as gentis especies daquella artificial formosura, & preparasse por mãos da fantasia as cores, & pinceis de que queria usasse a natureza na perfeiçaõ daquelle filho. Naõ lhe sahio frustrada a diligencia; pois com a tal industria conseguio, o que taõ efficazmente pertendia o seu desejo. Naõ era menos grande o que eu tinha de que esta minha obra achasse boa aceitaçaõ em todo o mundo; mas quando olhava para ella, como parto do meu pobre talento, temia que tomasse a humildade do Pay, & que por esta causa fosse desprezada de todos; nesta desconfiança, & temor me atentou muito a lembrança de que desde o pri-

Galenus lib. de theriaca ad Pisonem cap. 11. mihi fol. 94. v. ibi: *Hominem opulentum quidem, sed deformem existisse, qui cum affectaret pulchrum sibi filium procreare, formosam in ampla tabula puerum depingere jussit, inde uxori præcepit ut è regione posita imaginem diligenter consideraret; illa vero attentè respiciens, & ut sic dicam, totum animum illuc conjiciens (erat enim valde pulchra ea figura) puerum peperit, non patri sed pictura similem.*

Figura 3 – SEMEDO (1707) página 26, trecho do sumário de observações.
Digitalização *Google Books*.

INDICE
DAS OBSERVAÇOENS,
Que ſe contem neſte livro.

Ao longo de mais de 600 páginas, esse livro exibe 101 segmentos de texto. Cada segmento corresponde a uma "Observação" de caso/paciente com comentários diversos do autor-médico sobre os tratamentos por ele desenvolvidos e recomendados para cada caso. Cada segmento relata procedimentos adotados, indicação de medicamentos, conselhos, críticas de procedimentos, como também traz remissões a obras latinas e a conceitos de autores importantes na Medicina da época. A proposta de Semedo era facilitar o acesso, da sua comunidade profissional e dos trabalhadores agregados a ela, a conhecimentos sobre práticas de cura. Isso ele decidiu fazer sob a forma de publicações impressas em português. Pois, como o autor mesmo menciona, as pessoas envolvidas com a Saúde naquela época, em sabendo ler, tinham dificuldades com o acesso e o entendimento de textos em latim.

Assim imbuído, ***Observaçoens Medicas Doutrinaes (...)*** trata de 101 casos de perfis bastante variados, oferecendo um panorama histórico das doenças e intercorrências mais comuns da época, que atingiam diferentes segmentos populacionais: adultos, homens, mulheres, gestantes, recém-paridas, jovens, idosos, crianças, fossem nobres, camponeses ou pessoas comuns, do povo. Combinado a outras obras do autor, este livro serviu para divulgar uma visão doutrinal e uma "escola de pensamento" da Medicina portuguesa do século XVIII associada ao nome de J. C. Semedo. Esse autor também se tornou uma referência de estudo para os curadores e trabalhadores da área no cenário do Brasil Colônia e Brasil Imperial. Seu papel referencial para diferentes "trabalhadores da Saúde" daquela época pode ser conferido na obra "Erário Mineral[4]", um livro também

[4] Uma versão em grafia atualizada do "Erário Mineral", acompanhada de valiosos textos explicativos e glossários, pode ser obtida gratuitamente em http://books.scielo.org/id/ypf34. Esse material também figurará entre as indicações de atividades de estudo do nosso ambiente de ensino na UFRGS. A visão de historiadores sobre o período será de extrema valia para os nossos estudantes.

impresso português, publicado em 1735. Esse livro reúne as experiências e práticas médicas do cirurgião-barbeiro Luís Gomes Ferreira, o qual atuava na capitania de Minas Gerais.

4. Um segmento para exame com ferramentas computacionais

Com partes da obra de Semedo indicadas por Lourenço (2016) como as tematicamente mais relevantes para uma história da Medicina escrita em português, buscamos identificar padrões quantitativos do vocabulário do texto na sua grafia antiga e atualizada, conforme as duas digitalizações do original de que dispúnhamos e contando com o acesso à obra física, em caso de alguma necessidade, junto à BPE. Nessa primeira aproximação, apenas com partes pontuais do livro, quisemos identificar problemas e necessidades para lidar com o seu todo e oferecê-lo sob a forma de *corpus*, passível de exploração com ferramentas identificadoras de palavras gráficas.

Sintetizaremos aqui apenas o trabalho com o segmento ***Observaçam XCII***, um trecho tematicamente importante e que tem 1.317 palavras gráficas considerando-se a sua grafia antiga. Nessa ***Observaçam,*** vemos o relato de problemas enfrentados por Semedo com o atendimento de mulheres recém-paridas, tema que reaparece em outras *Observações* ao longo do livro. Para a geração do texto em formato editável, recorremos ao material do *Google Books* e também à versão da BPE digitalizada em PDF. O texto foi selecionado do arquivo em formato PDF e copiado para um arquivo em formato WORD, para depois ser salvo em formato somente texto (.TXT), tendo sido necessário realizar toda uma série de leituras da imagem original, de releituras das digitalizações e de ajustes para que fosse preservada, o máximo possível, a grafia original do documento-fonte, na edição de 1707. A nossa transcrição do original digitalizado seguiu o

"Léxico de Edições" da obra "Gazetas Manuscritas", material do século XVIII, disponível no *corpus* histórico *Tycho Brahe* antes citado.

Assim, uma vez gerado o nosso arquivo **Observaçam 92**, respeitando-se a sua grafia original, ele foi submetido aos *softwares* AntConc (ANTONY, 2014) e TermoStat (DROUIN, 2003), os quais são recursos computacionais comuns em LC, mas que, frisamos, não foram planejados para processar textos com grafia antiga. Entretanto, cabe frisar, ambas as ferramentas, AntConc e TermoStat, operam com a noção de palavra gráfica, entendida como um conjunto de caracteres separados por um espaço em branco. Essas duas ferramentas para tratamento de *corpora* foram por nós escolhidas por serem grátis, de fácil uso e relativamente familiares a muitos estudantes e professores do curso de Letras da **UFRGS.**

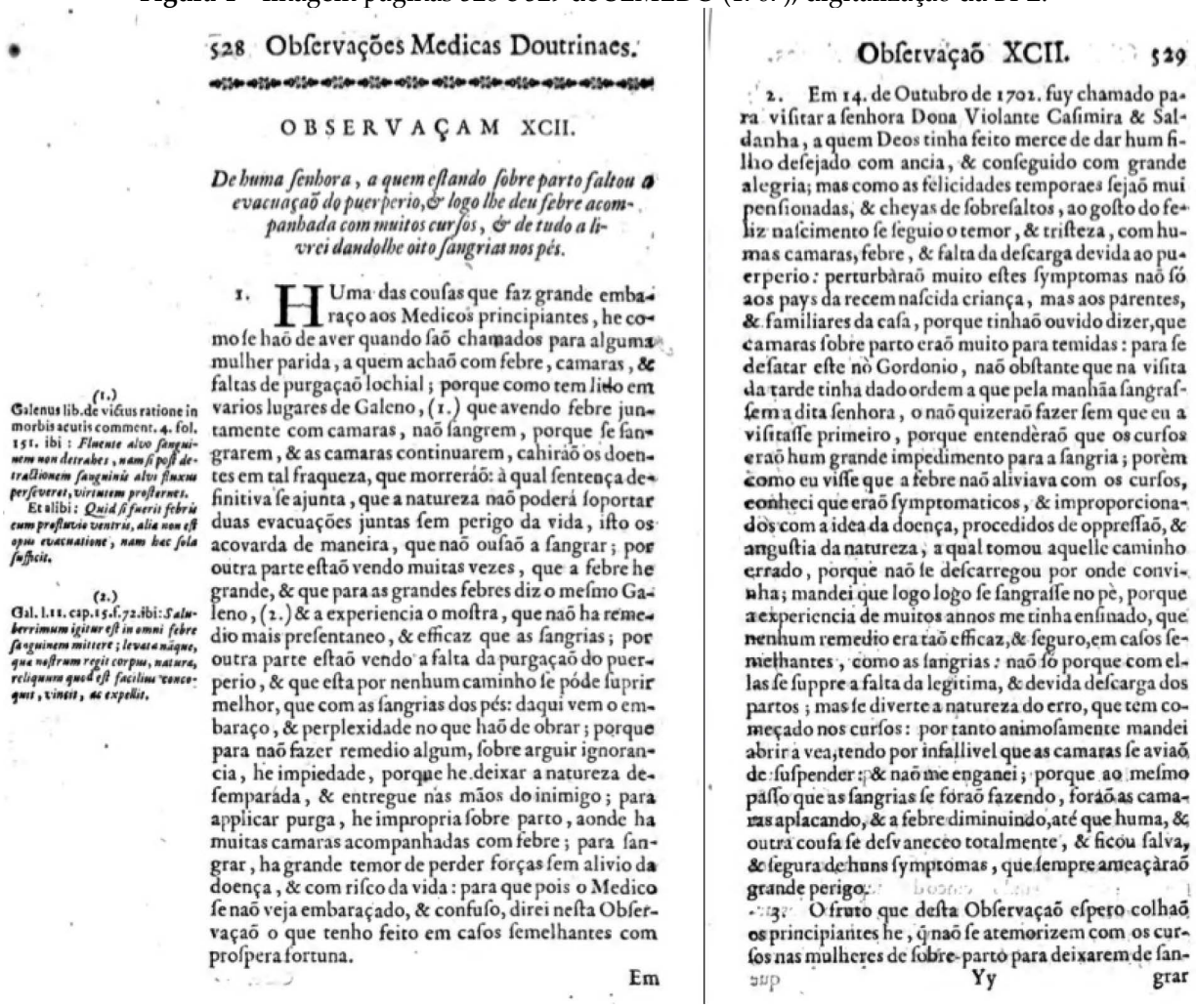
O TermoStat é uma ferramenta bastante reconhecida em estudos de Terminologia por conseguir apontar prováveis "termos técnicos" em um texto de perfil especializado sob exame. Essa ferramenta, ainda pouco utilizada por pesquisadores de **LC**, funciona pela comparação estatística entre as "palavras" presentes num dado texto sob exame e um grande conjunto de textos não especializados (um acervo de textos de jornal) embutido na sua base de dados. As prováveis terminologias, assim, são indicadas pela ferramenta na medida em que correspondam a itens de uso e distribuição peculiares no texto sob exame frente ao acervo geral de textos. A sua lógica de funcionamento, assim, é a de uma comparação entre amostra e a respectiva população.

A seguir, reproduzimos uma parte da nossa **Observaçam 92** e, depois, o correspondente arquivo de imagem do original. Nesse sentido, vale alertar que: a) as remissões laterais de página, em latim, foram desprezadas; b) a correspondência de tipos antigos de impressão com caracteres atuais seguiu o padrão de tratamento do *corpus* Tycho Brahe – conforme antes citado; c) foi

ajustada a hifenação de palavras ao final de cada linha do texto, dado que isso, conforme entendemos, não prejudicaria uma primeira abordagem para o tratamento do vocabulário em teste com dois *softwares* identificadores de palavras selecionados:

> OBSERVAÇAM XCII
> De huma Senhora, a quem estando sobre parto faltou a evacuaçaõ do puerperio, & logo lhe deu febre acompanhada com muitos cursos, & de tudo a livrei dando lhe oito sangrias nos pés.
>
> **1**. Huma das cousas que faz grande embaraço aos Medicos principiantes, he como se hão de aver quando saó chamados para alguma mulher parida, a quem achaõ com febre, camaras, & faltas de purgaçaõ lochial; porque como tem lido em varios lugares de Galeno, (I.) que avendo febre juntamente com camaras, naõ sangrem, porque se sangrarem, & as camaras continuarem, cahiráõ os doentes em tal fraqueza, que morrerão: à qual sentença definitiva se ajunta, que a natureza naõ poderá soportar duas evacuações juntas sem perigo da vida, isto os acovarda de maneira, que naõ ousaõ a sangrar; por outra parte estaõ vendo muitas vezes, que a febre he grande, & que para as grandes febres diz o mesmo Galeno, (2.) & a experiencia o mostra, que naõ ha remedio mais presentaneo, & efficaz que as sangrias; outra parte estaó vendo a falta da purgaçaõ do puerperio, & que esta por nenhum caminho se póde suprir melhor, que com as sangrias dos pés: daqui vem o embaraço, & perplexidade no que haõ de obrar; porque para naõ fazer remedio algum, sobre arguir ignoramcia, he impiedade, porque he deixar a natureza de desemparada, & entregue nas mãos do inimigo; para applicar purga, he impropria sobre parto, aonde ha muitas camaras acompanhadas com febre; para sangrar, ha grande temor de perder forças sem alivio da doença, & com risco da vida: para que pois o Medico se naõ veja embaraçado, & confuso, direi nesta Observaçaó o que tenho feito em casos semelhantes com prospera fortuna.
>
> **Fonte**: Semedo (1707), trecho da página 528, *Observaçam XCII*, transcrição com grafia original.

Figura 4 – Imagem páginas 528 e 529 de SEMEDO (1707), digitalização da BPE.



528 Obſervações Medicas Doutrinaes.

OBSERVAÇAM XCII.

De huma ſenhora, a quem eſtando ſobre parto faltou a evacuaçaõ do puerperio, & logo lhe deu febre acompanhada com muitos curſos, & de tudo a livrei dandolhe oito ſangrias nos pés.

1. HUma das couſas que faz grande embaraço aos Medicos principiantes, he como ſe haõ de aver quando ſaõ chamados para alguma mulher parida, a quem achaõ com febre, camaras, & faltas de purgaçaõ lochial; porque como tem lido em varios lugares de Galeno, (1.) que avendo febre juntamente com camaras, naõ ſangrem, porque ſe ſangrarem, & as camaras continuarem, cahiraõ os doentes em tal fraqueza, que morreráõ: à qual ſentença definitiva ſe ajunta, que a natureza naõ poderá ſoportar duas evacuações juntas ſem perigo da vida, iſto os acovarda de maneira, que naõ ouſaõ a ſangrar; por outra parte eſtaõ vendo muitas vezes, que a febre he grande, & que para as grandes febres diz o meſmo Galeno, (2.) & a experiencia o moſtra, que naõ ha remedio mais preſentaneo, & efficaz que as ſangrias; por outra parte eſtaõ vendo a falta da purgaçaõ do puerperio, & que eſta por nenhum caminho ſe póde ſuprir melhor, que com as ſangrias dos pés: daqui vem o embaraço, & perplexidade no que haõ de obrar; porque para naõ fazer remedio algum, ſobre arguir ignorancia, he impiedade, porque he deixar a natureza deſemparáda, & entregue nas mãos do inimigo; para applicar purga, he impropria ſobre parto, aonde ha muitas camaras acompanhadas com febre; para ſangrar, ha grande temor de perder forças ſem alivio da doença, & com riſco da vida: para que pois o Medico ſe naõ veja embaraçado, & confuſo, direi neſta Obſervaçaõ o que tenho feito em caſos ſemelhantes com proſpera fortuna.

Em

(1.) Galenus lib. de victus ratione in morbis acutis comment. 4. fol. 151. ibi: *Fluente alvo ſanguinem non detrahes, nam ſi poſſ detractionem ſanguinis alvi fluxus perſeveret, virtutem proſternet.* Et alibi: *Quid ſi fuerit febris cum proſtutio ventris, alia non eſt opus evacuatione, nam hæc ſola ſufficit.*

(2.) Gal. l. 11. cap. 15. f. 72. ibi: *Salubberrimum igitur eſt in omni febre ſanguinem mittere; levat enim, quæ noſtrum regit corpus, natura, reliquum quod eſt facilius vincit, vincit, & expellit.*

Obſervaçaõ XCII. 529

2. 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 ancia, & conſeguido com grande alegria; mas como as felicidades temporaes ſejaõ 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: perturbáraõ muito eſtes ſymptomas naõ ſó aos pays da recem naſcida criança, mas aos parentes, & familiares da caſa, porque tinhaõ ouvido dizer, que camaras ſobre parto eraõ muito para temidas: para ſe deſatar eſte nó Gordonio, naõ obſtante que na viſita da tarde tinha dado ordem a que pela manhãa ſangraſſem a dita ſenhora, o naõ quizeraõ fazer ſem que eu a viſitaſſe primeiro, porque entendèraõ que os curſos eraõ hum grande impedimento para a ſangria; porèm como eu viſſe que a febre naõ aliviava com os curſos, conheci que eraõ ſymptomaticos, & improporcionados com a idea da doença, procedidos de oppreſſaõ, & anguſtia da natureza, a qual tomou aquelle caminho errado, porque naõ ſe deſcarregou por onde convinha; mandei que logo logo ſe ſangraſſe no pè, porque a experiencia de muitos annos me tinha enſinado, que nenhum remedio era taõ efficaz, & ſeguro, em caſos ſemelhantes, como as ſangrias: naõ ſó porque com ellas ſe ſuppre a falta da legitima, & devida deſcarga dos partos; mas ſe diverte a natureza do erro, que tem começado nos curſos: por tanto animoſamente mandei abrir a vea, tendo por infallivel que as camaras ſe aviaõ de ſuſpender; & naõ me enganei; porque ao meſmo paſſo que as ſangrias ſe foraõ fazendo, foraõ as camaras aplacando, & a febre diminuindo, até que huma, & outra couſa ſe deſvaneceo totalmente, & ficou ſalva, & ſegura de huns ſymptomas, que ſempre ameaçáraõ grande perigo.

3. O fruto que deſta Obſervaçaõ eſpero colhaõ os principiantes he, q̃ naõ ſe atemorizem com os curſos nas mulheres de ſobre-parto para deixarem de ſan-

Yy grar

4.1 Do experimento com AntConc e TermoStat

Com AntConc, listamos as palavras do texto da **Observaçam 92** conforme a grafia original antiga; obtivemos uma lista com 1.317 palavras (*tokens*), sendo 536 palavras/formas distintas (*types*). Na proporção entre *types tokens*, com a qual se estima a variedade do vocabulário do texto, o segmento mostrou 40% de variedade do vocabulário e um conjunto de 355 palavras de ocorrência única (denominadas *Hapax legomena*). O item CAMARAS (atual **diarreias**) foi o item lexical mais frequente no todo do segmento assim processado.

Depois, com o TermoStat, ferramenta antes descrita, contrastamos as frequências e distribuições de palavras empregadas no texto antigo com as

frequências de palavras do seu acervo de textos com ortografia do português atual. Com o TermoStat, arrolaríamos, em tese, as maiores peculiaridades da **Observaçam 92** quanto à distribuição estatística de um vocabulário específico do passado frente a um vocabulário atual e mais amplo.

As palavras mais "peculiares" empregadas na **Observaçam 92** – conforme estatisticamente percebidas pelo *software* TermoStat - frente ao que se verifica em textos em português atual - foram SANGRIA e MEDICO (sem acento). Reproduzimos, a seguir, na Figura 5, uma tela do processamento de TermoStat com o texto na grafia antiga. Vale observar outros itens apontados como tendo alto escore de especificidade (*Score Spécifcité*): SANGRIA, MEDICO (sem acento) e PURGAÇAÕ (com o til sobre o O).

Figura 5 – Tela de resultados – palavras específicas do segmento **Observação 92**, grafia antiga.

Résultats

Liste des termes | Nuage | Statistiques | Structuration | Bigrammes

Candidat de regroupement	Fréquence	Score (Spécificité)	Variantes orthographiques	Matrice
sangria	14	200.83	*sangria sangrias*	Nom
medico	4	137.78	*medico*	Nom
purgaçaõ	4	137.78	*purgaçaõ*	Nom
galeno	4	137.78	*galeno*	Nom
parto	15	116.41	*parto partos*	Nom
purgação	3	113.64	*purgação*	Nom
puerperio	3	113.64	*puerperio*	Nom
medicos	3	113.64	*medicos*	Nom
purga	5	112.01	*purga*	Nom
medicos principiante	2	83.5	*medicos principiantes*	Nom Adjectif
evacuação junta	2	83.5	*evacuações juntas*	Nom Nom
loquios	2	83.5	*loquios*	Nom
sangria de pé	2	83.5	*sangrias dos pés*	Nom Préposition Nom
humor cacochymicos	2	83.5	*humores cacochymicos*	Nom Adjectif
taes mulher	2	83.5	*taes mulheres*	Nom Nom
foy	2	83.5	*foy*	Nom
falta de purgaçaõ	2	83.5	*falta da purgaçaõ*	Nom Préposition Nom
valerio martins	2	83.5	*valerio martins*	Nom Nom
purgação de parto	2	83.5	*purgação do parto*	Nom Préposition Nom
felicidade temporaes	2	83.5	*felicidades temporaes*	Nom Nom
reynavão soro	2	83.5	*reynavão soros*	Nom Nom

Fonte: do programa TermoStat (DROUIN, 2003).

Com a ferramenta AntConc, listamos todas as palavras gráficas empregadas no texto da **Observaçam 92**. Reproduzimos, a seguir, no **Quadro 1**, uma parte da lista de palavras do todo do segmento, na sua grafia antiga. Depois, segue o respectivo gráfico de frequências de palavras do texto – o Gráfico 1.

O texto exibe um todo de 1.317 palavras, sendo esse universo composto por 536 palavras gráficas diferentes, que são repetidas. A palavra gráfica mais frequente no texto antigo é o QUE, com 61 repetições. Vale observar as duas formas de grafia de NÃO, que essa ferramenta, por sua natureza, trata como duas palavras gráficas diferentes. O **Gráfico 1** mostra um desenho de frequências de palavras bastante normal em relação ao que observa em estudos de lexicoestatística atuais com o português moderno (valendo consultar BIDERMAN, 1998). Isto é, o fato de o texto estar grafado em português antigo não parece repercutir sobre padrões de distribuição lexical genericamente observados, nos quais o mais usual é a alta frequência de palavras gramaticais e um grande conjunto de *Hapax legomena*, desenhando, na curva gerada pelo gráfico, o que se convenciona chamar de "princípio da cauda longa".

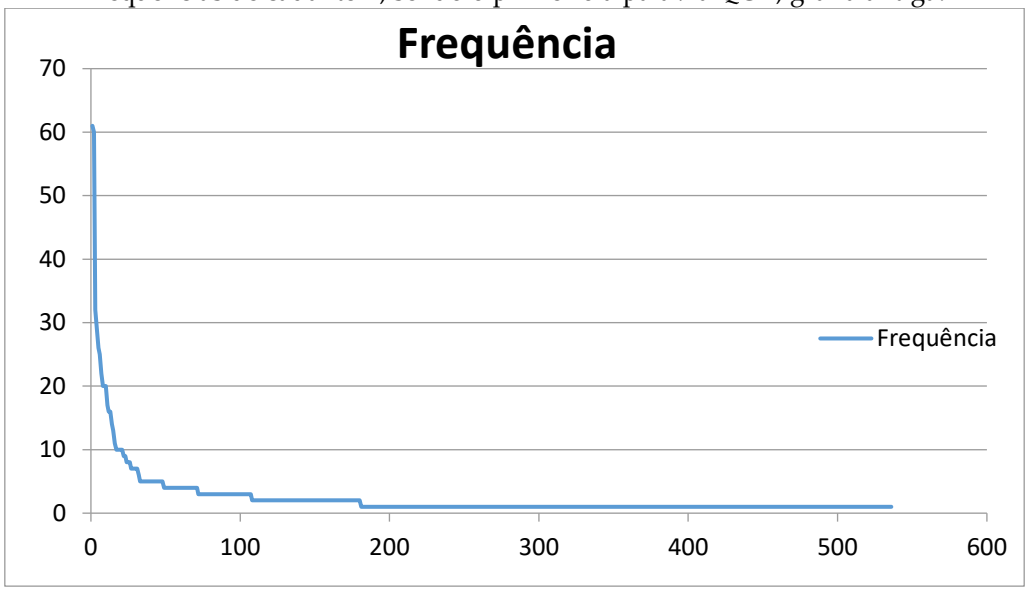
Quadro 1 – Lista das 30 palavras mais frequentes em **Observação 92**, grafia original/antiga, com a ferramenta AntConc.

Ordem	**Frequência/repetições**	**Palavra**
1	61	que
2	60	a
3	32	de
4	29	se
5	26	com
6	25	as
7	22	para
8	20	da
9	20	o
10	20	porque
11	17	camaras
12	16	do
13	16	naõ
14	14	parto
15	13	em
16	11	por
17	10	como

18	10	febre
19	10	os
20	10	ou
21	10	sangrias
22	9	natureza
23	9	paridas
24	8	he
25	8	não
26	8	purgaçaõ
27	7	dos
28	7	falta
29	7	grande
30	7	quem

Fonte: elaborado pela autora.

Gráfico 1 - Distribuição de frequências de palavras. *#Word Types*/ palavras diferentes: 536 – eixo horizontal do gráfico *#Word Tokens*/ total de palavras contadas no segmento: 1.317. No eixo vertical, as frequências de cada item, sendo o primeiro a palavra QUE, grafia antiga.



Fonte: elaborado pela autora.

Depois disso, o mesmo segmento, mas com grafia atualizada, foi reprocessado pelas mesmas duas ferramentas, AntConc e Termostat. Abaixo, um trecho do segmento sob estudo com ortografia por nós atualizada segundo sistemática[5] também do *corpus* Tycho Brahe:

[5] Ver em http://www.tycho.iel.unicamp.br/*corpus*/manual/prep/manual_frameset.html, item **III.1.2.3 Grafia,** acesso em set/2017.

OBSERVAÇÃO XCII

De uma Senhora, a quem estando sobreparto faltou a evacuação do puerpério, e logo lhe deu febre acompanhada com muitos cursos, e de tudo a livrei dando-lhe oito sangrias nos pés.

1. Uma das coisas que faz grande embaraço aos médicos principiantes, é como se hão de haver quando são chamados para alguma mulher parida, a quem acham com febre, câmaras, e faltas de purgação loquial; porque como têm lido em vários lugares de Galeno, (I.) que havendo febre juntamente com câmaras, não sangrem, porque se sangrarem, e as câmaras continuarem, cairão os doentes em tal fraqueza, que morrerão: à qual sentença definitiva se ajunta, que a natureza não poderá suportar duas evacuações juntas sem perigo da vida, isto os acovarda de maneira, que não ousam a sangrar; por outra parte estão vendo muitas vezes, que a febre é grande, e que para as grandes febres diz o mesmo Galeno, (2.) e a experiência o mostra, que não há remédio mais presentâneo, e eficaz que as sangrias; outra parte estão vendo a falta da purgação do puerpério, e que esta por nenhum caminho se pode suprir melhor, que com as sangrias dos pés: daqui vem o embaraço, e perplexidade no que hão de obrar; porque para não fazer remédio algum, sobre arguir ignorância, é impiedade, porque é deixar a natureza de desamparada, e entregue nas mãos do inimigo; para aplicar purga, é imprópria sobreparto, aonde há muitas câmaras acompanhadas com febre; para sangrar, há grande temor de perder forças sem alívio da doença, e com risco da vida: para que pois o médico se não veja embaraçado, e confuso, direi nesta Observação o que tenho feito em casos semelhantes com próspera fortuna.

Com ortografia atualizada, pela ferramenta AntConc, tivemos em suas contagens: 1.363 palavras (*tokens*), 520 palavras diferentes (*types*) e uma proporção de variedade do vocabulário que atinge 38 %. As diferenças entre grafia antiga e original, percebidas pelas duas ferramentas, sintetizamos no Quadro 2, a seguir.

Quadro 2 – Síntese dos achados com as duas ferramentas, ortografia atual e antiga, **Observação 92.**

Resultados	**Grafia original antiga**	**Grafia atualizada**
Número de palavras (AntConc)	1.317	1.363
Palavras diferentes (AntConc)	536	520
% de palavras diferentes (AntConc)	40%	38%
Número de itens de ocorrência única (*hapax*)	355	342
Itens mais específico frente ao português atual TermoStat	SANGRIA MEDICO PURGAÇAÕ	PURGAÇÃO SANGRIA PURGAÇÃO LOQUIAL

Fonte: elaborado pela autora.

Como se vê no Quadro 2, não foram obtidos resultados muito distantes entre o processamento do texto com a grafia antiga e o com a grafia modernizada. Isso, entretanto, não quer dizer que as formas variadas de escrita de um mesmo item tenham deixado de representar um problema a ser enfrentado em termos do processamento desse tipo de *corpus*, para as contagens e para o estudo do vocabulário empregado no texto.

Afinal, cada forma diferente de palavra como, por exemplo, NÃO e NAÕ, nas contagens das duas ferramentas, e nos contrastes que façam com outros *corpora*, sejam de textos antigos, do século XVIII, ou da atualidade, corresponderá a uma palavra gráfica diferente. O número de palavras de ocorrência única, os *Hapax legomena*, que sinaliza a diversidade do vocabulário do texto em termos de variedade, ficou em 355 para o texto na forma antiga e em 342 para a sua forma atualizada.

Quanto à identificação de especificidades do texto do século XVIII frente à linguagem atual, via TermoStat, salientam-se diferentes unidades que muito provavelmente têm valor de terminologia, mas também há muitas coincidências entre o texto antigo e o atualizado. O caso da identificação do sintagma terminológico PURGAÇÃO LOQUIAL, na versão do texto com a grafia atualizada, é bem interessante na medida em que nos aponta um sintagma terminológico, uma unidade poliléxica.

Esses resultados iniciais do estudo-piloto, em tese, sinalizariam, positivamente, para podermos lidar com o todo do livro, com ferramentas computacionais semelhantes a essas, apenas com a versão na ortografia antiga. Assim, pelo menos num primeiro momento de abastecimento do nosso ambiente de estudos e do nosso *corpus*-amostra, o custo de atualizar a ortografia, que é relativamente alto, poderia ser contornado sem grandes prejuízos para o entendimento do texto por parte de um estudante de Letras com pelo menos um ano

de curso. Esse é o perfil do usuário-estudante do nosso ambiente virtual de estudos com textos do século XVIII.

Naturalmente, é preciso ainda ponderar sobre todas as especificidades envolvidas nesse tipo de texto, sobre questões associadas à hifenação do texto na grafia antiga e sobre o próprio perfil de funcionamento particular de cada uma das ferramentas testadas, AntConc e TermoStat. A partir daí, poderemos organizar o *corpus*-amostra da obra no seu todo buscando, justamente, aproveitar essas discrepâncias e contrastes entre linguagem médica antiga e atual, entre a escrita do passado e a escrita do presente, para impulsionar o interesse dos estudantes de Letras da UFRGS para diferentes *corpora* históricos do português. Isto é, poderíamos estimular os nossos alunos a conferirem se as especificidades do nosso acervo-amostra, identificadas por meio de sua experimentação com ele e com essas ferramentas, acessadas no nosso ambiente de estudos, teriam correspondências em outras bases textuais, tais como os grandes *corpora* antes citados. O acesso a dicionários que tratam do português da época e a dicionários atuais de Medicina também é uma atividade interessante, para que se possa verificar o modo de conceituação, antigo e atual, associado a procedimentos e a terminologias mencionados por Semedo.

5. Síntese dos resultados do teste inicial e perspectivas

Inicialmente, utilizar a ferramenta AntConc foi produtivo. Isto é, ela enfrentou bem o desafio de reconhecer as palavras na sua grafia original – não atualizada – do nosso texto médico do século XVIII, sem ter sido desenvolvida para esse fim. Vale destacar seu bom tratamento da diversidade e frequência de formas gráficas, especialmente com a medida da variedade proporcional do vocabulário (medida conhecida como *Type-Token Ratio)* e indicação da proporção de palavras de ocorrência única.

Há que se ponderar, entretanto, que a "percepção" do que seja uma "palavra", nessas ferramentas e nesse modo de observar a linguagem escrita, está balizada apenas pela noção de palavra gráfica. Afinal, há a situação de uma ortografia antiga envolvida nesse tipo de *corpus*, e isso não se deve perder de vista, tampouco devem ser perdidas as limitações dos *softwares* em teste. Um item/palavra como **XVIII**, por exemplo, no AntConc, será considerado uma palavra gráfica, pois é um conjunto de letras, sendo que o seu sistema despreza numerais e sinais de pontuação.

Por sua vez, o TermoStat, como funciona identificando e categorizando as "palavras" por classes morfológicas, para então contrastar o vocabulário do segmento em foco com o de um grande acervo de textos da atualidade, demanda mais estudos sobre seus modos de funcionamento e rendimento em atividades didáticas com textos antigos. É preciso ponderar sobre o que esse sistema faz, na "sua intimidade estatística", com a classificação das grafias não válidas e avaliar em que medida o "erro" – a palavra desconhecida por seu etiquetador morfossintático – ajudaria ou atrapalharia o sistema ao lidar com as nossas ortografias antigas. Ainda que o contraste permitido pelo TermoStat seja entre as palavras do texto antigo *versus* o que há em uma grande massa de textos da atualidade, cremos que se poderia utilizá-lo para algumas finalidades com nossos estudantes, ainda que a comparação antigo-atual seja desigual e problemática em vários pontos.

Assim, a partir deste estudo-piloto, com esse e outros segmentos da obra de Semedo (1707), indica-se viável seguir o trabalho com o todo da obra na sua versão original, *in natura*. Ao que parece, é mais promissor o uso de recursos de processamento semelhantes ao que oferece o AntConc. Afinal, essa ferramenta também aponta especificidades do vocabulário de um texto X sob exame frente ao vocabulário de outros textos – que podem ter o mesmo tipo de grafia do texto X - sem necessidade de se produzir uma cópia de cada segmento com grafia atualizada.

Um teste comparativo interessante, nessa direção, será fazer o contraste do léxico da obra de Semedo com o léxico de textos de temática não especializada, também do século XVIII, tal como as ***Gazetas Manuscritas da Biblioteca Pública de Évora***, encontradas com acesso livre no *corpus* histórico Tycho Brahe (*link* para acesso antes citado). Esse procedimento comparativo, feito com a ferramenta AntConc, é denominado *KEY-WORD LIST* e bastante praticado em estudos de **LC**. O *corpus* de referência antigo, composto pelas ***Gazetas***, na forma de uma *WORD-LIST*, precisará ser inserido pelo usuário na ferramenta para que a comparação do vocabulário possa ser feita. Desse modo, poder-se-ia contrastar a linguagem médica de Semedo com um padrão de linguagem em geral em textos de viés jornalístico produzidos na mesma época, observando-se frequências e distribuições de formas, inclusive das alternâncias de grafia. Ter-se-ia, assim, um *corpus* de estudo e um *corpus* de referência nos moldes da **LC.**

O processo de atualização da grafia do *corpus*, embora gere uma série de facilidades para a leitura do documento, seja a leitura humana ou a "leitura de máquina", faz perder uma série de informações muito importantes para um estudo linguístico-histórico. Ainda assim, a atualização ortográfica do texto é um recurso importante para facilitar a leitura de diferentes pesquisadores, especialmente para quem não seja da área de Letras.

Nessa direção, a produção da atualização do texto/*corpus*-amostra poderia integrar um conjunto de exercícios de aula para os nossos estudantes de Letras, que poderiam, colaborativamente, sob supervisão de seus professores, ajudar a enriquecer e ampliar esse *corpus*. Teríamos, assim, um acervo em diferentes apresentações: imagem digitalizada do documento original, o texto reproduzido com a grafia antiga preservada e sua versão em ortografia atualizada.

Um outro potencial trabalho com nossos estudantes interessados por terminologias médicas da época é o cotejo das palavras ou terminologias nele

contidos com o seu registro em dicionários antigos e em dicionários da atualidade, conforme já citamos. Outro cotejo, importantíssimo, pois nos dá toda uma contextualização histórica, é entre o que lê neste texto de Semedo e o que nos traz a pesquisa de mestrado em História de Lourenço (2016). Essa contextualização pode permitir ponderar em que medida o autor, pela via do texto escrito em português, naquela época, teria conseguido promover uma maior acessibilidade dos seus conhecimentos e práticas de Medicina para os leitores que tinha em mente. Afinal, no Brasil de hoje, esse é ainda um desafio para a comunicação sobre temas de Saúde: promover a acessibilidade textual e terminológica da informação especializada. Precisamos que nossos especialistas e instituições governamentais consigam produzir materiais informativos que possam ser compreendidos tanto pelo cidadão leigo, com diferentes perfis de escolaridade, quanto por públicos profissionais diversos.

Há, enfim, muito a fazer, mas parece viável iniciar, desde já, o *corpus*-amostra, com o texto das 101 "Observações" de Semedo, com a sua grafia original. A tarefa com a produção e o oferecimento de *corpora* de textos antigos, ainda assim, segue com seus custos, visto ser preciso confrontar imagens do original e o texto que se está gerando, seja manualmente, seja com a ajuda de alguma ferramenta que reconheça os caracteres no arquivo-fonte. Os ganhos e as oportunidades de aprendizagem envolvidos, acreditamos, superam os custos.

Agradecimentos

À CAPES, Processo nº 88881.119078/2016-01 do Programa EST-SENIOR, pela bolsa concedida e pela oportunidade de conhecer o acervo da Biblioteca Pública de Évora em Portugal. Ao CNPq, pelo apoio às nossas pesquisas sobre acessibilidade textual e terminológica em temas de Saúde para adultos de escolaridade limitada. Agradeço também aos colegas M. Filomena Gonçalves e Paulo Quaresma, da Universidade de Évora, pelo incentivo para a produção de materiais de apoio ao ensino e à pesquisa – na via linguística e na via informatizada - com textos antigos escritos em português.

Referências bibliográficas

ANTHONY, L. **AntConc** (Version 3.4.3) [Computer Software]. Tokyo, Japan: Waseda University, 2014. Disponível em http://www.laurenceanthony.net/

BERBER SARDINHA, T.. **Linguística de *Corpus***. Barueri-SP, São Paulo: Manole, 2004.

BIDERMAN, M. T. C. Conceito linguístico de palavra. In: BASILIO, M. (Org.). **Palavra**. Rio de Janeiro: Grypho, 1999. v.1, p. 81-97.

BIDERMAN, M. T. C. A face quantitativa da linguagem: um dicionário de frequências do português. **Alfa**, Araraquara, UNESP, v.42, p. 157-78, 1998.

CUCATTO, L. A. Extradev: um sistema de extração semiautomático de deverbais em *corpus* do português histórico e contemporâneo. **Anais do X Encontro do CELSUL** – Círculo de Estudos Linguísticos do Sul UNIOESTE - Universidade Estadual do Oeste do Paraná- Cascavel-PR | 24 a 26 de outubro de 2012 | ISSN 2178-7751, 2012, p. 1-10.

DROUIN, P. 2003. Term extraction using non-technical corpora as a point of leverage. **Terminology**. Amsterdam, The Netherlands, John Benjamins, v. 9 (1), p. 99-117, 2003.

DUARTE, L. G. O correr da pena nas Gazetas Manuscritas. A identidade de formas (1735-1738). **Cadernos de Cultura**. Lisboa: Centro de História da Cultura da Universidade Nova de Lisboa, 2012.

QUARESMA, P. Análise linguística de documentos da Biblioteca Pública de Évora. Uma abordagem informática In: GONÇALVES, M. F; BANZA, A. P. (coord.), **Património textual e humanidades digitais**: da antiga à nova Filologia. Évora: Centro Interdisciplinar de História, Culturas e Sociedades da Universidade de Évora (CIDEHUS)/ Fundação para a Ciência e a Tecnologia (FCT), 2013. Disponível em: http://books.openedition.org/cidehus/1091

KRIEGER, M. da G.; FINATTO, M. J. B. **Introdução à Terminologia**: Teoria e prática. São Paulo: Contexto, 2004.

LOURENÇO, T. S. **O médico entre a tradicão e a inovacão**: João Curvo Semedo. Dissertacão (Mestrado) – Universidade Federal Fluminense, Instituto de Ciências Humanas e Filosofia. Departamento de História, Niterói-RJ, 2016. Disponível em: http://www.historia.uff.br/stricto/td/2002.pdf

MURAKAWA, C. de A. A construção de um dicionário histórico: o caso do Dicionário Histórico do Português do Brasil - séculos XVI, XVII e XVIII. **Estudos de Linguística Galega.** 3 Agosto 2014, vol. 6, no. 0, 2014. [consultado: 21 Novembro 2015]. Disponível em: http://www.usc.es/revistas/index.php/elg/article/view/2084

PAIXÃO DE SOUSA, M. C. A Filologia Digital em Língua Portuguesa: Alguns caminhos. In: In: GONÇALVES, M. F.; BANZA, A. P. (coord.). **Património textual e humanidades digitais**: da antiga à nova Filologia. Évora: Centro Interdisciplinar de História, Culturas e Sociedades da Universidade de Évora (CIDEHUS)/ Fundação para a Ciência e a Tecnologia (FCT), pp. 113-138, 2013. Disponível em http://books.openedition.org/cidehus/1089

SEMEDO, J. C. **Observaçoens medicas doutrinaes de cem casos gravíssimos, que em serviço da pátria, & das nações estranhas escreve em língua portugueza, & latina**. Lisboa: Officina de Antonio Pedrozo Galram, 1707. 616p.

Artigo recebido em: 30.05.2017 Artigo aprovado em: 28.11.2017