A ARQUITETURA DE UM *CORPUS* NA ÁREA DE TEOLOGIA: CONCEPÇÃO E PROCEDIMENTOS DE COMPILAÇÃO*

Solange Aparecida Faria Cardoso[1]

Resumo

O presente trabalho visa descrever a construção de corpus em uma área específica: a Teologia, de acordo com os postulados da Linguística de Corpus. Trata-se de um corpus piloto que não tem a pretensão de ser exaustivo, mas ser apenas uma amostragem da área em foco com o objetivo de exemplificar, de modo concreto, todas as etapas da metodologia usada na organização e compilação de corpora, que servirão à extração das Unidades Terminológicas e à elaboração de proposta de dicionário. A construção do corpus textual levou em consideração a sua representatividade, tanto na área específica quanto junto à comunidade alvo. Observe-se que, a representatividade aqui referida não está relacionada com a extensão, mas com a população envolvida e com a área temática.

Palavras-chave: Linguística de *corpus; corpus* especializado; arquitetura do *corpus;* Teologia

Abstract

This paper aims to describe corpus construction in a specific area: Theology, according to Corpus Linguistics postulates. It is a pilot corpus, which does not claim to be exhaustive, but only be a sample of the focused area, which has as its objective to illustrate, in a concrete way, every step of the methodology used in the organization and compilation of corpora that will serve the Terminological Units extraction and the preparation of a dictionary proposal. The building of the textual corpus has taken into account its representativeness of the specific area as well as the target community. One should note that the representativeness referred to here is not related to the extension, but to the population involved and to the thematic area.

Keywords: Corpus linguistics; specialized corpus; corpus architecture; Theology

1. Introdução

Este artigo é parte de projeto que objetiva a elaboração de proposta de dicionário terminológico da área de Teologia. O recorte que aqui apresentamos integra a metodologia usada na organização e compilação de corpora que servirão à extração das Unidades Terminológicas e à elaboração da proposta de dicionário.

Os corpora textuais estão sendo constituídos de textos escritos digitalizados, utilizados pelos alunos, professores e formadores nas aulas e nos encontros de evangelização, que constituem atividades de formação teológica da Shalom Comunidade e Faculdade de Ensino Superior (FASES) da cidade de Uberlândia/MG.

A estrutura composicional deste trabalho está assim organizada: na seção 1, apresentamos as bases teóricas sobre as quais se alicerçam o trabalho de composição de corpora segundo os postulados da Linguística de Corpus (doravante LC); na seção 2, delineamos o percurso metodológico por nós seguido para a composição dos corpora; na seção 3, fazemos algumas considerações no que se refere às análises preliminares observadas, procurando enfatizar os aspectos mais significativos, tendo em vista os objetivos desse trabalho.

* Gostaríamos de agradecer à Capes (Coordenação de Aperfeiçoamento de Pessoal de Nível Superior) pelo financiamento dos anais da VII Escola Brasileira de Linguística Computacional e do XIII Encontro de Linguística de Corpus, processo nº 3472/2015-87.

[1] Aluna do Programa de Pós-Graduação em Estudos Linguísticos (PPGEL) – curso de Doutorado, do Instituto de Letras e Linguística (ILEEL) da Universidade Federal de Uberlândia (UFU), Uberlândia/MG. solangejac@yahoo.com.br

2. Fundamentação teórica

A Linguística de Corpus é uma área do conhecimento que estuda a linguagem com o auxílio de tecnologias computacionais. Esse estudo é feito por meio da utilização de diferentes quantidades (grandes, médias e pequenas) de dados empíricos relativos ao efetivo uso da linguagem. Ou seja, a principal característica da LC é a observação de dados empíricos de uma ou mais línguas – ou variedades de língua - armazenados em bancos de dados que compõem um corpus, com a utilização de ferramentas eletrônicas. Estas ferramentas são, especialmente, desenvolvidas para auxiliar o pesquisador na análise dos dados, facilitando assim, o trabalho quanto à verificação dos fenômenos da língua em uso. Em outras palavras, podemos afirmar que a LC é uma área que investiga a língua em uso, tanto escrita quanto falada, registrada em formato de arquivo de computador.

Em relação ao conceito de corpus, podemos afirmar que este conceito tem sido definido por distintos posicionamentos, destacando-se semelhanças e diferenças nas indicações dos autores. Como aspecto comum, podemos destacar o fato de que o corpus é apontado como uma coleção de materiais (SINCLAIR, 1991; BERBER SARDINHA, 2004). Em relação a esse conceito, se buscarmos a origem etimológica do termo, podemos verificar que corpus é simplesmente a equivalência em latim da palavra corpo, o que reforça a ideia de conjunto de elementos. Entretanto, McEnery e Wilson (1996) destacam que, quando "corpus" é utilizado em contextos da linguística moderna, o termo adquire conotações mais específicas.

Assim, um corpus é uma coletânea de textos ou transcrições de fala coletados, criteriosamente, para a pesquisa linguística e salvos (textos e ou transcrições) em arquivos com formato eletrônico, como já afirmamos anteriormente. É exatamente o fato de os dados estarem armazenados em formato eletrônico que suscita a característica chave dos estudos desenvolvidos em LC: a utilização de computadores para a análise dos dados.

Considerando-se as definições de corpus consultadas para este estudo (SINCLAIR, 1991; MCENERY e WILSON 1996; BIDERMAN, 2001; BERBER SARDINHA, 2004 etc.), podemos concluir que um corpus linguístico é uma coletânea de textos escritos e/ou orais que ocorrem naturalmente. Esta coletânea é selecionada por meio de critérios específicos, a fim de caracterizar uma variedade da língua ou toda a língua e, para a pesquisa linguística, os textos são processados por computador.

Dentre os tipos de corpora (plural de corpus) utilizados pela LC, podemos destacar o corpus geral e o corpus especializado. O corpus geral, também conhecido como corpus de língua geral, compõe-se de uma coletânea de textos com a finalidade de permitir pesquisa de uma determinada língua, sem levar em conta distinções genéricas, varietais, dialetais, lexicais etc. Sinclair (1991) refere-se ao corpus geral como "um tipo de corpus que armazena uma enorme quantidade de dados e está em constante atualização da língua-alvo".

Entretanto, no caso de interesse de análise de determinada variedade, como o falar de uma determinada região do Brasil (o falar gaúcho, por exemplo); ou domínio específico, como a linguagem jurídica ou da Teologia; esse tipo de corpus não é apropriado e, nesse caso, opta-se pela compilação de um corpus especializado.

O corpus especializado, por sua vez, é desenvolvido para atender às necessidades específicas de um trabalho de pesquisa em particular, de acordo com os objetivos propostos. Uma coletânea de textos da área de Teologia, por exemplo, reunidos com o propósito de verificar a variedade lexical e gramatical em suas diferentes subáreas. Vários corpora especializados têm sido coletados com propósitos e conteúdos distintos, segundo os objetivos da pesquisa e, portanto, diferenciados pelo modo, tempo, seleção, conteúdo, autoria, disposição interna e finalidade (BERBER SARDINHA, 2004).

O objetivo deste trabalho é apresentar o processo de concepção e elaboração do Corpus da Teologia de acordo com os padrões da LC.

3. Metodologia

Segundo Fromm, “um dos pontos básicos para a elaboração de um banco de dados é a criação de uma estrutura para organizar a informação a ser coletada” (FROMM, 2007, p. 38). Assim, a primeira etapa metodológica de nosso trabalho foi a elaboração da árvore do campo pesquisado, por meio da qual nos foi possível o estabelecimento de uma estrutura organizacional para, em seguida, procedermos à constituição do *corpus*.

Nesse sentido, Krieger e Finatto (2004, p. 134) recomendam a utilização de uma árvore de domínio “para que se tenha uma aproximação inicial a uma área de conhecimento.” Para as pesquisadoras, a árvore de domínio seria uma representação, um recurso metodológico na organização de *corpora* à medida que se situa um dado campo de conhecimento, suas denominações e suas inter-relações, ainda que as mais básicas, por ser “uma aproximação inicial” de um campo de conhecimento.

Neste trabalho, consideramos que a elaboração da árvore de domínio cumpre duas funções. A primeira é, como sugerem Krieger e Finatto (2004), a de aproximação inicial à área da Teologia. A segunda função diz respeito ao fato de a árvore de domínio ser elemento facilitador/condutor na criação de uma estrutura para a organização dos corpora a serem coletados. Outra questão a ser considerada é o fato de que, em determinado momento de nossa pesquisa, o próprio procedimento de compilação do corpus poderá ilustrar a configuração da área, determinando subestruturas na árvore. Esclarecemos que a elaboração da Árvore de Domínio foi feita manualmente e com base em estudos bibliográficos e consulta a especialistas.

No que se refere à elaboração da Árvore de Domínio da grande área da Teologia, pesquisamos, inicialmente, o sítio virtual do Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq)[2] e o sítio virtual da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES)[3]. Nessa pesquisa, pudemos constatar que a Teologia está inserida, na Tabela de Áreas do Conhecimento, como uma das áreas das Ciências Humanas, tendo como subáreas História da Teologia, Teologia Moral, Teologia Sistemática e Teologia Pastoral.

Assim temos a tabela:

Tabela 1. Áreas de Conhecimento. Fonte: CAPES.

ÁREA DE AVALIAÇÃO: FILOSOFIA / TEOLOGIA: SUBCOMISSÃO TEOLOGIA TEOLOGIA
71001000 HISTÓRIA DA TEOLOGIA
71002006 TEOLOGIA MORAL
71003002 TEOLOGIA SISTEMÁTICA
71004009 TEOLOGIA PASTORAL

A consulta/pesquisa ao sítio virtual do CNPq e da CAPES resultou na Árvore de domínio da Teologia apresentada na Figura 1.

[2] Disponível em: <http://www.cnpq.br/documents/10157/186158/TabeladeAreasdoConhecimento.pdf>. Acesso em: 22 set. 2014.

[3] Disponível em: <http://www.capes.gov.br/images/stories/download/avaliacao/TabelaAreasConhecimento_042009.pdf>. Acesso em: 22 set. 2014.

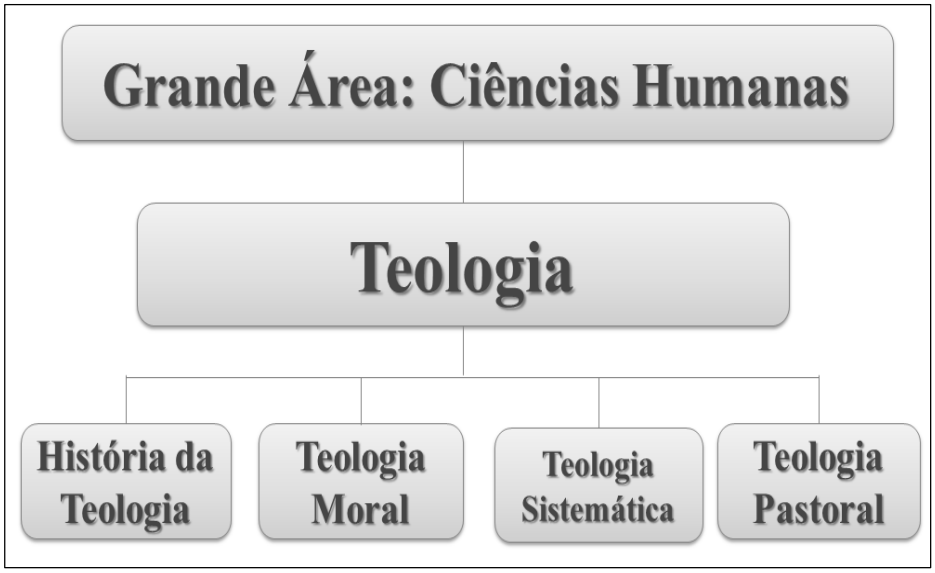

Figura 1. Árvore de domínio da Teologia – CNPq. Fonte: Elaboração própria.

O nosso próximo passo foi a consulta ao documento Matriz Curricular da FASES (ALVES et al., 2015), e por meio das informações nele encontradas, elaboramos a proposta de árvore de domínio desta área de conhecimento. Oportunamente, pudemos apresentar o mapa conceitual (Árvore de Domínio) da FASES a especialistas que atuam nesta área e nesta faculdade. A versão por nós elaborada e apresentada na Figura 2 é, inclusive, o resultado das análises desses especialistas.

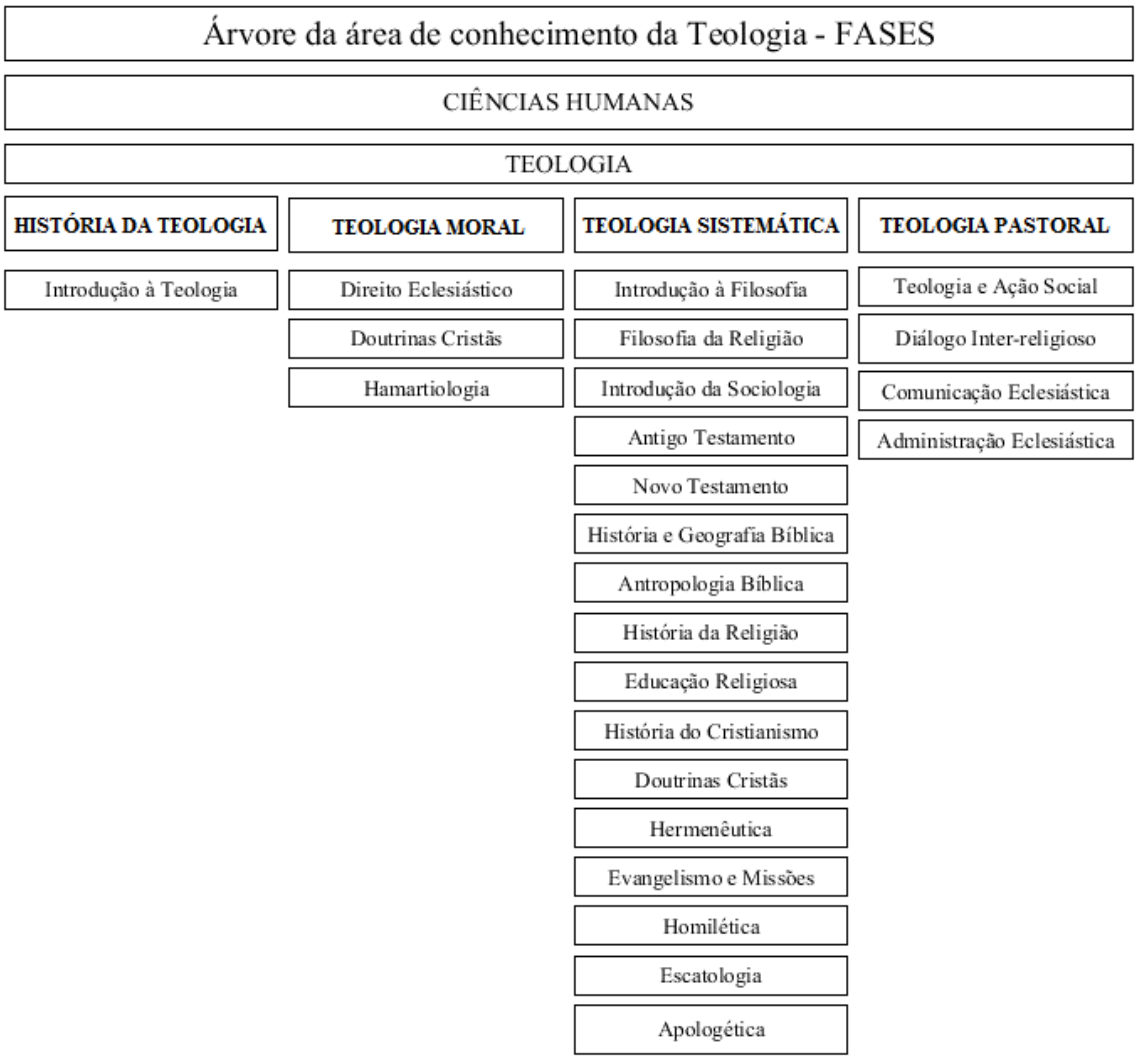

Figura 2. Árvore da área de conhecimento da Teologia – FASES Fonte: Elaboração própria.

A árvore acima (Figura 2) mesmo representando o consenso entre a opinião de alguns especialistas na área, não representa o pensamento unânime entre os estudiosos. Esta proposta consta, portanto, de uma árvore elaborada em três níveis: Teologia > História da Teologia / Teologia Moral / Teologia Sistemática / Teologia Pastoral > subáreas. Esclarecemos que optamos, neste trabalho, pela elaboração da árvore de domínio para a constituição da estrutura organizacional que determinou, até o momento, a compilação dos *corpora*. Esta estrutura possibilitou a criação de um diretório, no computador, com exibição das pastas na mesma estrutura da Árvore de Domínio da área de conhecimento da Teologia – FASES (Figura 3) que constituiu o passo seguinte desta pesquisa.

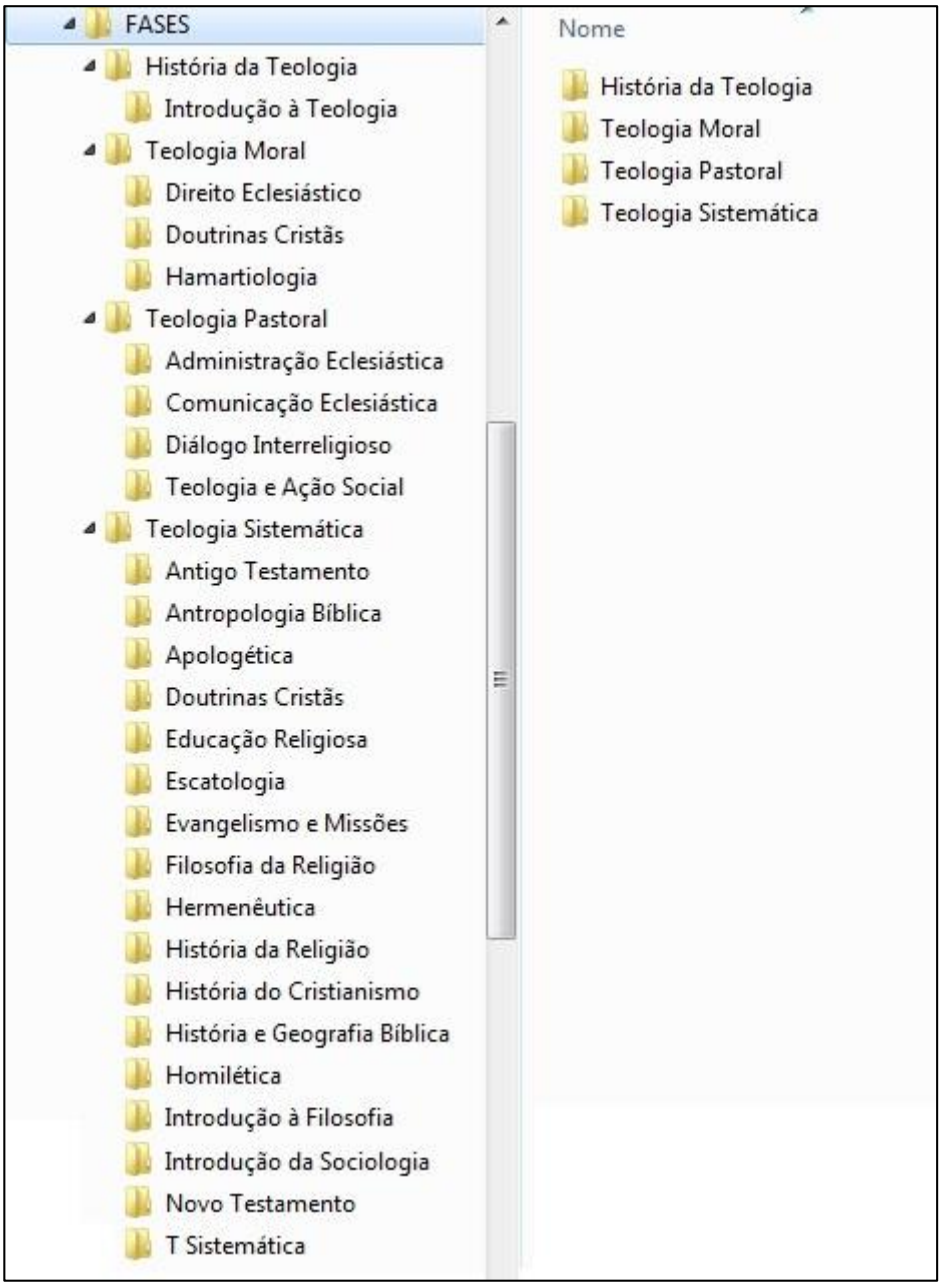

Figura 3. Diretório com pastas na forma da Árvore da área de conhecimento da Teologia – FASES. Fonte: Elaboração própria.

A próxima etapa cumprida diz respeito à compilação dos *corpora*.

Para essa compilação, aproveitamos, inicialmente, a indicação de um total de quinhentos livros sugeridos pelos alunos da FASES. Apresentamos a lista dos quinhentos livros a um pastor e professor dessa faculdade e, após criteriosa avaliação no que se refere ao conteúdo das obras, foram selecionados duzentos e três livros compilados em um CD-R.

No segundo momento, apresentamos a lista com os duzentos e três títulos ao Coordenador e professores do curso de Teologia da Shalom. Solicitamos a esses especialistas a análise dos títulos das obras relacionadas e a consequente indicação daquelas que são ou poderiam ser utilizadas pelos alunos para o ensino/aprendizagem da(s) disciplina(s) ministrada(s) no curso de Bacharelado em Teologia, da Faculdade Shalom de Ensino Superior - FASES. A indicação foi feita por meio de um X assinalado em espaço indicado (ao lado do título da obra correspondente). Consideramos que, procedendo assim, poderemos assegurar que o *corpus* seja constituído de dados autênticos, ou seja, não inventados; legíveis por computador e representativos da língua portuguesa na modalidade escrita. Ressaltamos que, a inautenticidade só seria estabelecida caso os textos fossem redigidos para uma finalidade específica, como a produção de textos para o desenvolvimento desta pesquisa, por exemplo. Assim, as obras não selecionadas permanecem como dados autênticos, não inventados. Apenas não são considerados títulos adequados para o ensino de determinada disciplina.

Fizemos a conversão das extensões dos textos/livros selecionados de *.pdf e *.html para *.txt. e os organizamos no diretório criado no computador (Figura 3).

Apesar de todos os textos, até o presente momento, serem recuperados da compilação em CD-R a nós oferecido por alunos do curso de Teologia da FASES e, para que não haja nenhum tipo de implicações legais relacionadas a direitos autorais, estamos localizando cada texto na internet, salvando-os como arquivos *.txt. e registrando o endereço virtual. Quando não localizamos o texto em publicações virtuais, registramos o CD-R como fonte de resgate do texto.

Para a identificação de cada arquivo, optamos pelo título que aparece no início de cada texto.

Como ainda não estabelecemos critérios para a definição do tamanho dos arquivos de cada uma das áreas de conhecimento, nem a extensão em número de palavras das áreas, organizamos um quadro (exemplificado a seguir) onde registramos: o número de obras e o tamanho do arquivo em *KB*, para facilitar o trabalho futuro de balanceamento do *corpus*.

Quadro 1: Controle número de obras e tamanho dos arquivos em KB. Fonte: Elaboração própria.

ÁREA DE CONHECIMENTO	Nº OBRAS	TAMANHO *KB*
HISTÓRIA DA TEOLOGIA		
Introdução à Teologia	01	375
TEOLOGIA MORAL		
Direito Eclesiástico	00	
Doutrinas Cristãs	04	1 214
Hamartiologia	01	547

4. Análises preliminares

Para a apresentação de algumas considerações no que se refere às análises preliminares observadas, selecionamos o *corpus* já compilado para a disciplina de História e Geografia Bíblica.

Observando a Árvore de Domínio (Figura 2) podemos constatar que História e Geografia Bíblica pertence à subárea da Teologia Sistemática. Para melhor localização, apresentamos o recorte feito para este estudo na Figura 4.

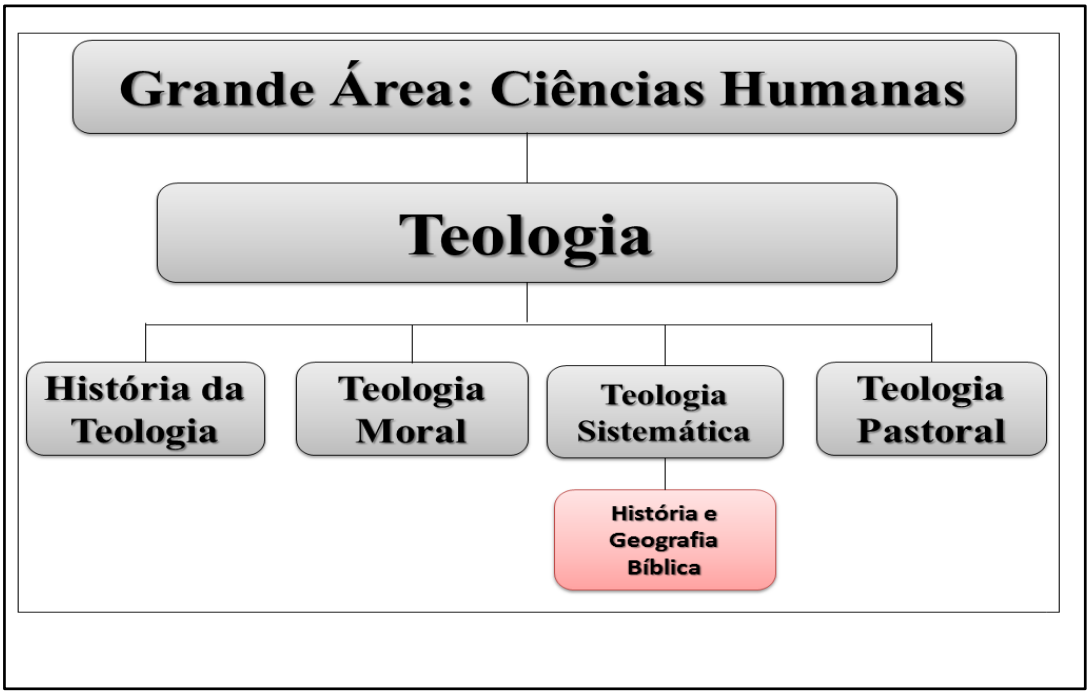

Figura 4. Árvore da área de conhecimento da Teologia – Destaque História e Geografia Bíblica Fonte: Elaboração própria.

Após a conversão das extensões dos textos/livros selecionados de *.pdf e *.html para *.txt. (texto sem formatação) e a organização no diretório criado no computador, procedemos ao tratamento deste *corpus* por meio do uso do programa de análise lexical *WordSmith Tools*, versão 6, de Scott (2012).

Ressaltamos que é considerado *corpus* cru um texto que é tomado do seu habitat natural, (jornal, livro, revista, internet etc.) cujo conteúdo é mantido sem qualquer anotação ou tratamento interno. Essa denominação, *corpus* cru, já é bastante utilizada na LC. Dentre outros estudiosos, Novodvorski (2013) usa a expressão *corpus* cru para definir um *corpus* que não foi etiquetado, ou seja, manipulado quanto a seu conteúdo e quanto à sua disposição espacial, por exemplo. Em oposição ao que Novodvorski (2013) chama de *corpus* cru, "textos em sua versão original" (NOVODVORSKI, 2013, p. 72), o tratamento de *corpus* é qualquer ação de manipulação nos textos, após a coleta. Ou seja, qualquer modificação no arquivo sem que haja alterações no conteúdo lexical dos textos.

Em síntese, sobre o texto cru, original, caso seja feita uma simples conversão a outro formato, o texto já deixaria de ser "cru", ou seja, não original no que se refere, inclusive, à formatação e ao suporte em que, inicialmente, fora publicado.

Inicialmente, usamos a ferramenta *Wordlist* (listagem de palavras) para a produção da lista de palavras com todas as palavras do arquivo selecionado. Esta lista apresenta elencadas em conjunto as frequências absolutas e percentuais de cada palavra. "Também compara listas, criando listas de consistência, onde é informado em quantas listas cada palavra aparece" (BERBE SARDINHA, 2009, p. 11).

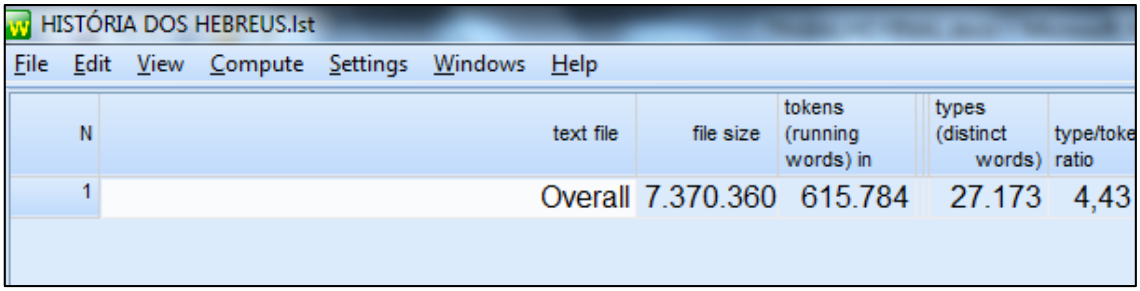

Figura 5. *Wordlist* – História e Geografia Bíblica. Fonte: Elaboração própria.

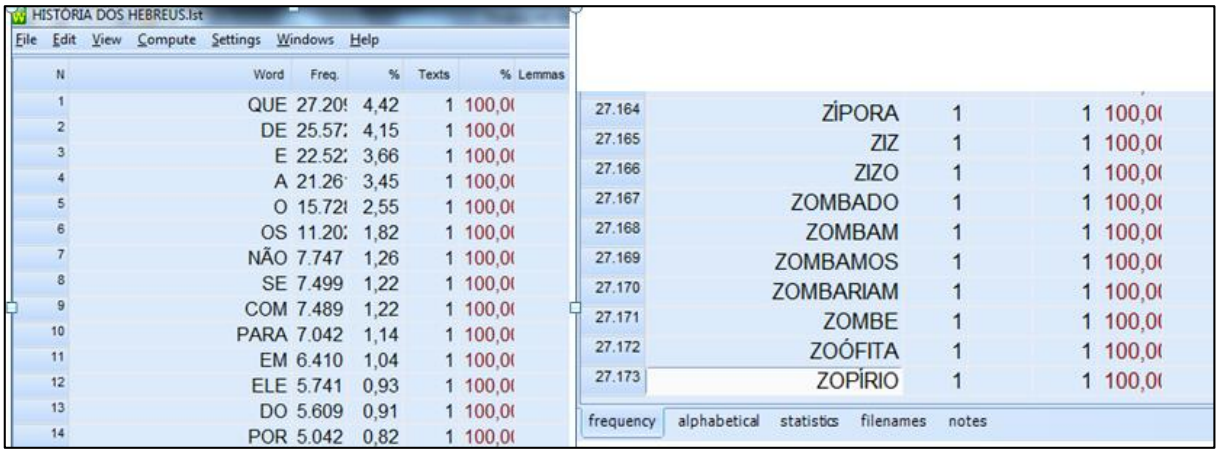

Figura 6. *Wordlist* – História e Geografia Bíblica. Fonte: Elaboração própria.

Como resultado da contagem final das palavras, obtivemos os seguintes resultados:

Quadro 2: Quantidade itens/formas. Fonte: Elaboração própria.

Itens (quantidade total de palavras nos textos)	615.784
Formas (quantidade de palavras diferentes nos textos)	27.173

Em relação a esses resultados, Berber Sardinha (2004) defende que se classifique o *corpus* segundo o número total de palavras, assim:

Quadro 3: Classificação do *corpus* número de palavras. Fonte: Berber Sardinha

Tamanho em palavras	Classificação
Menos de 80 mil	Pequeno
80 a 250 mil	Pequeno-médio
250 mil a 1 milhão	Médio
1 milhão a 10 milhões	Médio-grande
10 milhões ou mais	Grande

Tendo por base o critério de medição desenvolvido por Berber Sardinha (2004), podemos afirmar que o conjunto de material selecionado para esta pesquisa é constituído por um *corpus* de tamanho médio. Isso nos permite afirmar que este é um *corpus* representativo no universo dos textos na subárea da História e Geografia Bíblica. Lembramos que aqui, utilizamos apenas uma parte do *corpus*, o da História e Geografia Bíblica. Se a extensão deste é de 615.784 itens, quando obtivermos material (textos) em todas as outras subáreas, o *corpus* final poderá ser gigantesco.

Como resultado da arquitetura por nós delineada, temos a seguinte tipologia para o *corpus* de História e Geografia Bíblica que se constitui como subárea da Teologia:

Quadro 4 - Tipologia do *Corpus* de História e Geografia Bíblica - subárea da Teologia

Língua	Monolíngue (português)
Modo	Escrito (livro)
Data de publicação	Sincrônico

Seleção	Amostragem, Estático
Conteúdo	Especializado (Teologia)
Autoria	Falantes nativos (português), individual
Disposição Interna	Não comparável
Uso na pesquisa	Estudo (análise terminológica/terminográfica)
Tamanho	Médio (250 mil a 1 milhão de palavras)
Nível de Codificação	Com cabeçalhos, sem etiquetas

5. Comentários finais

O objetivo deste trabalho foi apresentar parte do processo de planejamento de um *corpus* em área específica para análise linguística. Nesse processo, consideramos a elaboração de Árvore de Domínio a metodologia para a aproximação inicial de uma área específica porque, por meio dessa estrutura, é possível delinearmos toda a arquitetura organizacional dos *corpora* da área objeto de pesquisa.

Muito embora esse processo possa resultar em apenas um diagrama composto por termos-chave de uma especialidade, na verdade, a elaboração de Árvore de Domínio só é possível, por meio de muito tempo dedicado a pesquisas bibliográficas, leituras de documentos, consultas e entrevistas com especialistas. O resultado apresentado neste artigo, para além da configuração da Árvore de Domínio e da arquitetura do nosso *corpus* de pesquisa, envolveu também a aproximação e a compreensão de uma área de conhecimento.

Nesta pesquisa, além de nossa aproximação à área de conhecimento da Teologia, a elaboração da Árvore de Domínio possibilitou, também, a organização das informações para a compilação do *corpus* que é, em LC, o ponto de partida para as análises posteriores.

Referências

ALVES, M. E. S.; MOURA, S. S.; SILVA, A. L. et al. *Matriz curricular do curso de teologia*. Faculdade Shalom de Ensino Superior. Uberlândia, 12 fev. 2015. Disponível em: <http://fases.com.br/old/wm/MATRIZ_CURRICULAR_DO_CURSO_DE_TEOLOGIA.pdf> Acesso em: 27 nov. 2015.

BERBER SARDINHA, T. *O que é um corpus representativo?* Direct Paper 44, 2000. Disponível em: <http://www2.lael.pucsp.br/direct/DirectPapers44.pdf>. Acesso em: 5 jul. 2015.

_____. *Linguística de Corpus*. Barueri: Manole, 2004.

_____. Pesquisa em linguística de Corpus com WordSmith Tools. Campinas, SP: Mercado de Letras, 2009.

BIDERMAN, M. T. C. *Teoria Linguística*. São Paulo, Martins Fontes, 2001.

FROMM, G**.** *VoTec: a construção de vocabulários eletrônicos para aprendizes de tradução.* São Paulo, 2007. Tese (Doutorado – Programa de Pós-Graduação em Estudos Linguísticos e Literários em Inglês – Departamento de Letras Modernas). Faculdade de Filosofia, Letras e Ciências Humanas, Universidade de São Paulo.

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

MCENERY, T.; WILSON, A. *Corpus Linguistics.* Edinburgh: Edinburgh University Press, 1996.

NOVODVORSKI, A. *Estilo das traduções de Sergio Molina de obras de Ernesto Sabato:* um estudo de *corpora* paralelos Espanhol/Português. 2013. Tese (Doutorado em Estudos Linguísticos e Tradução) – Universidade Federal de Minas Gerais.

SCOTT, M. *WordSmith Tools version 6.* Liverpool: Lexical Analysis Software, 2012.

SINCLAIR, J. *Corpus, Concordance, Collocation.* Oxford: Oxford University Press, 1991.