O processo colaborativo na compilação de um corpus bilíngue na área de Linguística

Guilherme Fromm
PPGEL/ILEEL/UFU
Linha de Pesquisa: Teoria, Descrição e Análise Linguística
Data de início da pesquisa: 08/2010
Previsão de término: 12/2013

O presente projeto, desenvolvido numa parceria entre o professor pesquisador, alunos de graduação e alunos de pós-graduação, visa a atualização e ampliação de um projeto anterior (FROMM, 2007), o qual previa a elaboração de um *corpus* bilíngue na área de Linguística que servisse de teste para uma ferramenta computacional (baseada em banco de dados) de auxílio aos tradutores nas áreas técnicas. O intuito da atual elaboração de *corpora* é apresentar uma pesquisa de modo que se torne referência na construção de vocabulários (BARBOSA, 2001) no trabalho terminográfico.

A fase inicial consistiu-se na reelaboração da Árvore de Domínio do campo da Linguística apresentada no projeto inicial, tendo em vista o balanceamento do *corpus*. Em virtude da dificuldade de acesso aos especialistas, essa árvore, atualmente com 46 subáreas (27 na área de Linguística e 19 na área de Linguística Aplicada), ainda passa por processo de refinamento. A proposta é a compilação, para cada subárea em cada língua, de quinhentas mil palavras, o que teoricamente geraria um *corpus* de 46 milhões de palavras.

Seguindo a proposta de Berber Sardinha (2004), a tipologia desse corpus é: i. Modo: escrito (textos acadêmicos como resenhas, artigos, dissertações e teses; inicialmente a pesquisa se dá através de procuras por textos em formato PDF nas ferramentas de pesquisa avançada do buscador Google); ii. Tempo: Sincrônico (levantamento realizado entre 2010 e 2013); iii. Seleção: amostragem, estático; iv. Conteúdo: especializado (Linguística); v. Autoria: língua nativa (inglês e português); vi. Disposição Interna: comparável; vii. Finalidade: Estudo (análise terminológica/terminográfica).

A compilação do *corpus* tem-se dado de modo colaborativo. O professor pesquisador ministra uma disciplina na graduação (Língua Inglesa: Estudos Descritivos e Linguística de *Corpus*) e duas na pós-graduação (Lexicologia, Lexicografia e Terminologia; Tópicos em Estudos Analítico-descritivos 2 – Linguística de *Corpus)*. Em todas essas disciplinas são apresentados os conceitos básicos de Linguística de *Corpus* e a avaliação dos alunos, em parte, consiste na compilação, dentro do projeto, de *corpora* de especialidade nas subáreas já estabelecidas da Árvore de Domínio. Os alunos levantam os textos em dupla (cada um trabalhando numa língua) ou individualmente.

No momento, o *corpus* dispõe, de acordo com a ferramenta *WordSmith Tools* 6, de aproximadamente 10,74 milhões de palavras em inglês (540 textos) e 8,23 milhões em português (634 textos), totalizando cerca de 18,61 milhões de palavras Os próximos passos consistem em: (i) finalização da coleta dos textos por parte dos alunos; (ii) balanceamento no número de palavras em cada subárea; (iii) análise e correção do corpus para evitar textos duplicados; (iv) análise da relevância dos textos para cada subárea; (v) construção de um site com ferramentas para análise do *corpus*.

Referências:

BARBOSA, M. A. Dicionário, vocabulário, glossário: concepções. In: ALVES, I. M. (Org.). *A constituição da normalização terminológica no Brasil*. 2 ed. São Paulo: FFLCH/CITRAT, 2001.

BERBER SARDINHA, T. *Linguística de Corpus.* Barueri: Manole, 2004.

FROMM, G. *VoTec:* a construção de vocabulários técnicos eletrônicos para aprendizes de tradução. 2007. 215f. Tese. Departamento de Letras Modernas, FFLCH, Universidade de São Paulo, São Paulo, 2007.