

## A COMPILAÇÃO DE *CORPORA* PARALELOS E COMPARÁVEIS E O USO DE FERRAMENTAS COMPUTACIONAIS COM FINS PEDAGÓGICOS

Este trabalho visa a apresentar iniciativas tomadas por dois pesquisadores em uma universidade pública, no sentido de montar uma base de dados que possibilite discussões teóricas sobre fenômenos linguísticos e questões lexicais, desde a organização de *corpora* textuais (BIBER *et al*, 1998; SINCLAIR, 2001; BERBER SARDINHA, 2004; BERNARDINI, 2011), dos processos de elaboração de dicionários (BIDERMAN, 1984, 2001; BORBA, 2002), treinamento de tradutores (BERBER SARDINHA, 2003; TAGNIN, 2003; ZANETTIN, 2009) e ensino e aprendizagem de línguas estrangeiras (CANALE, 1983; ASTON *et al*, 2004; GABALLO, 2009). Neste sentido, serão apresentados os resultados de sete pesquisas realizadas em nível de graduação e de extensão, nas quais foram utilizados princípios de compilação de *corpora* eletrônicos para que, por meio deles, fossem realizadas análises linguísticas, levantamentos terminológicos e procedimentos tradutórios. Para tanto, foram utilizados programas de memórias de tradução e ferramentas de análise lexical como o *Wordfast* (CHAMPOLLION, 1999), o *WordSmith Tools* (SCOTT, 2005), o *Unitex* (PAUMIER, 2002), além de ferramentas disponibilizadas gratuitamente na Web como o *YouAlign* e o *Google Translator Toolkit*. Os resultados ora apresentados focaram gêneros textuais do discurso especializado, literários, informativos e de divulgação.

No caso dos textos especializados, o primeiro projeto visou a constituição de *corpora* textuais especializados com gêneros do domínio econômico-financeiro em espanhol veiculados em revistas científico-acadêmicas de quatro países de língua espanhola (Espanha, Argentina, Uruguai e Paraguai) para a montagem de uma base de dados por meio da qual seria possível descrever e analisar o uso especializado da língua espanhola (nas variedades peninsular, argentina, uruguaia e paraguaia) a fim de descrever aspectos linguísticos e elaborar dicionários terminológicos. O gênero contemplado na primeira etapa desse projeto foi o “artigo científico”. Selecionaram-se quatro revistas acadêmicas da área da Economia publicadas na Espanha, Argentina, no Uruguai e no Paraguai e coletaram-se artigos para a futura organização do *corpus* que, atualmente, conta com aproximadamente 600 mil itens.

O segundo projeto (PIBIC/CNPq) prevê a organização de um *corpus* textual em língua espanhola com textos de diferentes gêneros da esfera jornalística veiculados em

jornais de três países da América do Sul – Argentina, Paraguai e Uruguai. Esse projeto continuará no segundo semestre de 2012 com a inclusão de jornais de outros países hispanoamericanos bem como da Espanha. O objetivo desse *corpus* é desenhar a macroestrutura de um dicionário bilíngue português-espanhol para a produção de textos em espanhol por aprendizes brasileiros.

No terceiro projeto, três bolsistas realizaram a versão, para a língua inglesa, de textos informativos constantes do site de Química da universidade na qual estudam, utilizando a *Google Translator Toolkit*, uma ferramenta CAT online, que garantiu a centralização dos arquivos, já que, como os textos que seriam vertidos estavam na Internet, não haveria problemas de sincronia dos dados. Além da versão, elaboraram um glossário em forma de memórias de tradução que pôde ser utilizados por todos durante todo o processo tradutório. Este *corpus* paralelo, contendo os textos originais e as traduções, hoje conta com 15.504 itens. Dando continuidade ao projeto de extensão, três novos estagiários iniciarão a versão do site da Faculdade de Farmácia e terão o auxílio de um especialista da área para a consulta terminológica.

Na quarta pesquisa, foram compilados dois *corpora* comparáveis sobre Biotecnologia, nas línguas portuguesa e inglesa, que, por sua vez, foram subdivididos nas subáreas Biotecnologia Ambiental, Celular e Industrial. O *corpus* conta com 20.150 itens, do qual serão levantados termos simples e complexos a partir de palavras-chave para que constem de um glossário que será transformado em memórias de tradução a ser utilizado na ferramenta de tradução *Trados*.

Passamos, agora, a apresentar os trabalhos de cunho literário. O quinto trabalho tratou-se de um estudo comparativo da ocorrência de sintagmas preposicionais na obra *Alice's adventures in Wonderland*, de Lewis Carroll, e da respectiva tradução para o português (CARROLL, 1993, 2002). Para tanto, foi realizada uma busca pela estrutura formada por preposição 1 + (artigo) + substantivo + preposição 2 com o auxílio do programa *WordSmith Tools*, que fez a busca dos sintagmas tendo como palavra de busca as preposições mais frequentes. Em seguida, os textos em inglês e em português foram alinhados pelo programa *YouAlign* que possibilitou a comparação dos sintagmas nas duas línguas. Os resultados mostraram que, na versão portuguesa, o sintagma preposicional foi repetidamente modificado, o que permitiu a elaboração de um glossário bilíngue da obra de Carroll com o contexto de uso dos sintagmas preposicionais nas duas línguas.

No sexto trabalho, uma equipe de cinco alunos de graduação e de pós, sob a orientação de três professoras de língua e literatura inglesa, traduziram a obra *Democratic Vistas*, de Walt Whitman, para a língua portuguesa. Após a utilização das ferramentas de tradução *Wordfast* e *Google Toolkit*, e pelo alinhamento com o *YouAlign* para os que não haviam utilizado as duas primeiras ferramentas, o grupo se reunia para discussão e edição do texto final. Após a conclusão da tradução, a obra, em forma de *corpus* paralelo, foi analisada por meio de levantamento de frequência de palavras, palavras-chave e análise da razão forma/item com o *WordSmith Tools*. Desta forma, foi possível avaliar a variação lexical da obra traduzida e as palavras-chave do domínio ecológico utilizadas pelo autor e pelos tradutores.

Finalmente, no sétimo trabalho, a orientanda traduziu o conto *The Angel at the Grave*, da escritora norte-americana Edith Wharton, que será publicado em uma coletânea de contos de escritores dos séculos XIX e XX. O conto também foi traduzido com o auxílio da ferramenta de edição e de tradução de textos *Google TranslatorToolkit*, que, ao automaticamente alinhar os textos original e a tradução, permitiu a reflexão sobre o processo de tradução e sobre os procedimentos técnicos (BARBOSA, 1990) mais utilizados na tradução de um texto literário. A orientanda pôde adequar a tradução oferecida pelo tradutor automático e discutir as vantagens e desvantagens desta ferramenta ao se traduzir um texto literário, que possui características distintas do texto especializado.

Uma vez concluídos estes trabalhos, o próximo passo a ser tomado será a organização de todos os *corpora* compilados até o momento em uma única base de dados para que seja possível a sua expansão e a realização de outros trabalhos que visem a formação de futuros pesquisadores nas áreas de Linguística Aplicada e Análise Linguística, mais especificamente nas linhas de Lexicologia, Lexicografia (monolinguagem, bilingue, pedagógica), Terminologia, Tradução e Ensino e Aprendizagem de Línguas.

**Tópicos do Trabalho:** Compilação de Corpus; Estudos do Léxico (Lexicologia, Lexicografia e Terminologia/Fraseologia); Tradução Humana e Automática.

### **Referências Bibliográficas**

ASTON, G.; BERNARDINI, S.; STEWART, D. *Corpora and language learners*. Amsterdã: John Benjamins, 2004. 311p.

BARBOSA, H. G. *Procedimentos técnicos da tradução: uma nova proposta*. Campinas: Pontes, 1990.

BERBER SARDINHA, A. P. Uso de corpora na formação de tradutores. *DELTA* 19: Especial, p. 43-70, 2003.

\_\_\_\_\_. *Linguística de Corpus*. São Paulo: Editora Manole, 2004. 410 p.

BERNARDINI, S. Specialised corpus construction for language professionals. In: *Encontro de Linguística de Corpus*, Universidade Federal de Minas Gerais, Belo Horizonte, 2011.

BIBER D., CONRAD, S.; REPPEN, R. *Corpus Linguistics: investigating language structure and use*. Cambridge, Cambridge University Press, 1998.

BIDERMAN, M. T. C. A ciência da lexicografia. *Alfa: Revista de Linguística*, São Paulo, v.28, supl. p.1-26, 1984.

\_\_\_\_\_. *Teoria linguística*. São Paulo: Martins Fontes, 2001.

BORBA, F. S. et al. *Dicionário de usos do português do Brasil*. São Paulo: Ática, 2002.

CANALE, M. From Communicative Competence to Communicative Language Pedagogy. In: J. Richards & Schimidt (eds). *Language and Communication*. London: Longman. 1983.

CARROLL, L. *Alice's Adventures in the Wonderland*, New York: Dover, 1993.

\_\_\_\_\_. *Alice no País das Maravilhas*, Tradução de Maria Luiza Xavier de Almeida Borges, Rio de Janeiro: Jorge Zahar, 2002.

CHAMPOLLION, Y. *Wordfast*. 1999. Disponível em: <<http://www.wordfast.net/index.php>> Acesso em 23 de janeiro de 2007.

GABALLO, V. *English in Translation Studies: methodological perspectives*. Macerata: Eum Edizioni di Macerata, 2009. 150 p.

*Google translator toolkit*. Disponível em <<http://support.google.com/translate/toolkit/bin/answer.py?hl=pt-BR&answer=147809>> Acesso em 16 de agosto de 2010.

PAUMIER, *Unitex*. Disponível em <<http://www-igm.univ-mlv.fr/~unitex/>> Acesso em 21/07/2011.

SCOTT, M. *WordSmith Tools: version 5.0*. Oxford: Oxford University Press, 2007.

SINCLAIR, J. 1991. *Corpus, concordance, collocation*. Oxford: Oxford University Press, 1991.

TAGNIN, S. E. O. Os Corpora: instrumentos de auto-ajuda para o tradutor. *Cadernos de Tradução*, Florianópolis, v. 9, n. 2002/1, 2003. p. 191-213.

*Trados*. Disponível em < <http://www.trados.com/en/>> Acesso em 14 de junho de 2011.

*Youalign*. Disponível em < <http://youalign.com/> Acesso em 13 de julho de 2011.

ZANETTIN, F. Corpus-based Translation Activities for Language Learners. *The Interpreter and Translator Trainer*, (3)2, 209-224, 2009.