CorTrad: um novo corpus paralelo multiversão para o par de línguas português-inglês

Elisa D. Teixeira (Projeto CoMET), *Diana Santos* (Linguateca / SINTEF ICT), *Stella E. O. Tagnin* (USP)

A tradução de textos de diferentes gêneros e tipologias textuais exige estratégias tradutórias distintas, que respeitem as características micro- e macrotextuais que tipificam esses gêneros e tipologias em cada língua e cultura. Uma maneira de identificar essas características é observando textos de grande circulação nas comunidades técnicas e/ou científicas e suas respectivas traduções. Um corpus paralelo, ao confrontar original e tradução lado a lado, constitui uma rica fonte de exemplos e informações, tanto para tradutores profissionais e aprendizes quanto para pesquisadores.

Uma maneira ainda mais refinada de observar essas características é comparando diversas versões de um mesmo texto, i.e., o original, a tradução publicada e uma ou mais versões intermediárias da tradução. Um corpus assim configurado fornece uma visão sobre como tradutores, revisores e editores trabalham para obter um texto fluente, que corresponda às expectativas da comunidade lingüística a que se destina. Também destaca as dificuldades que o par de línguas oferece para aquela tipologia textual específica, tanto nos casos em que o tradutor é falante nativo da língua de partida quanto nos casos em que não é.

Diversos pesquisadores têm demonstrado a relevância do uso de corpora paralelos para o estudo, o ensino e a prática da tradução. McEnery & Wilson (1993) e Sommers (1993) discutem o potencial de corpora paralelos alinhados para a tradução automática. Malmkjaer (1998) e Peters *et al.* (2000), entre tantos outros, apontam para a utilidade de corpora paralelos alinhados como fonte de equivalentes na prática da tradução. Apesar da disponibilidade de alguns corpora paralelos para o par de línguas português-inglês, como o COMPARA (Frankenberg-Garcia e Santos, 2003) e o EuroParl (Koehn, 2005), há uma escassez de recursos no que diz respeito à variante brasileira do português e a corpora não-literários. O objetivo deste artigo é apresentar o CorTrad (www.fflch.usp.br/dlm/comet/consulta_cortrad.html), um novo corpus paralelo multiversão para o par de línguas português-inglês, e demonstrar sua utilidade com alguns exemplos de pesquisa possíveis. A parceria entre o Projeto CoMET, a Linguateca e o NILC para disponibilização gratuita do CorTrad na rede teve início em maio de 2008 e encontra-se em andamento.

O CorTrad está subdividido em três subcorpora. O acesso ao corpus é feito por meio do DISPARA (Santos, 2002), um sistema de disponibilização de corpora na rede com interface customizável que tem por base o sistema de processamento de corpus IMS-CWB (Christ *et al*., 1999). Uma das suas vantagens é a possibilidade de se criar mecanismos de busca específicos para cada subcorpus, tal como no AC/DC (Santos & Bick, 2000), permitindo pesquisas refinadas em diferentes gêneros e, até mesmo, em partes distintas de textos de uma mesma tipologia - p. ex., apenas na lista de ingredientes das receitas culinárias.

O subcorpus de divulgação científica do CorTrad Jornalístico conta atualmente com 1076 textos provenientes de vários números (entre 2001 e 2003) da Revista Fapesp, totalizando cerca de 47 mil palavras em português original e 37 mil em inglês traduzido. O corpus foi etiquetado usando o PALAVRAS (Bick, 2000). O usuário pode, p. ex., estudar a diferença entre o uso do verbo “pesquisar” em áreas científicas diferentes, dado que a cada texto está associada a classificação temática do LácioWeb (Aluísio *et al.*, 2003).

O subcorpus de contos do CorTrad Literário contém por ora 27 contos australianos, mas será logo acrescido de contos canadenses, ainda em fase de tratamento. No caso dos contos australianos, além do original e da tradução publicada, o CorTrad permite a visualização da primeira tradução e de uma segunda versão revisada, anterior à publicada. Por se tratar de traduções feitas por aprendizes, o original e a primeira versão configuram um corpus de aprendizes, que pode ser investigado para detectar os problemas mais freqüentes desses alunos. O usuário pode, entre outras coisas, estudar a maneira como foram traduzidos os termos de cunho cultural, assim como as diferenças mais freqüentes entre uma primeira tradução do inglês e as suas sucessivas revisões. Por exemplo: nas 5 ocorrências de *swag* do corpus, a palavra foi traduzida uma vez por “trouxa”, outra por “manta” e nas outras 3 vezes foi mantida na sua forma original.

O subcorpus de culinária do CorTrad Técnico-científico conta no momento com quatro versões de um livro de culinária de 315 páginas em português brasileiro: o original (cerca de 130.000 palavras), publicado em 2005; a primeira versão da tradução em inglês, feita por duas brasileiras; a revisão, feita por uma falante nativa do inglês; e a versão final, publicada em 2008 (ainda por ser incluída no corpus). Os textos foram alinhados por sentença e as diversas seções do livro e seus subtextos foram marcados. Por exemplo: os textos introdutórios foram identificados para poderem ser separados das receitas nas pesquisas. Dentro de cada receita, a lista de ingredientes, o título e o rendimento também foram marcados para permitir buscas específicas.

Vejamos um exemplo de busca neste corpus: a palavra “iogurte” e seu candidato a equivalente, *yogurt*. Se o usuário digitar simplesmente “iogurte" na janela de busca correspondente ao original e escolher como resultado uma concordância, verá que a tradução proposta inicialmente para “iogurte natural”, *whole milk plain yogurt* foi alterada pela revisora para *plain whole milk yogurt*. Perceberá também que “iogurte”, ao ser usado como sabor de sorvete, teve a tradução alterada para *frozen yogurt* (em vez de apenas *yogurt*). Se o usuário, por outro lado, digitar *yogurt* em uma das janelas correspondentes às versões da tradução verá, p. ex., que a palavra está presente em expressões como *Greek yogurt*, cujo equivalente em português (= "coalhada seca") não possui a palavra "iogurte", e que o plural sugerido como tradução para “iogurtes” foi deixado no singular depois da revisão. Por fim, se o usuário digitar a expressão de busca - “whole” @[] - numa dessas janelas e pedir como resultado a distribuição das formas, verá que o adjetivo modifica vários ingredientes em inglês, tais como: *milk*, *wheat*, *chicken*, *garlic* e *almonds*.

Em suma, há um universo de potencialidades para explorar no CorTrad, que gostaríamos de partilhar com a comunidade que pesquisa, pratica, ensina ou aprende tradução.

Palavras-chave: corpus paralelo; revisão de tradução; tradução técnica; tradução literária.