10.17771/PUCRio.TradRev.25346

O USO DE *CORPORA* PARA O ESTUDO DA TRADUÇÃO: OBJETIVOS E PRESSUPOSTOS

Carmen Dayrell

1 – Introdução

Um *corpus* é geralmente definido como uma coleção de textos selecionados e agrupados de acordo com critérios claramente definidos e especificados (Atkins et al., 1992; Baker, 1995; Eagles, 1996; Kenny, 2001:22). Tais critérios são estabelecidos de acordo com os objetivos e finalidades para os quais o *corpus* é compilado. Na lingüística moderna, é natural considerar que esses textos estejam em formato eletrônico, podendo ser analisados de maneiras diversas, automática ou semi-automaticamente (Baker, 1995; Kenny, 2001:22). Baker (1995) esclarece ainda que um *corpus* pode conter tanto linguagem escrita quanto falada, além de oferecer a possibilidade de inclusão de textos das mais diversas fontes, como por exemplo, de autores ou tópicos diferentes.

Corpora representam, portanto, a disponibilidade de um grande volume de dados empíricos, e a incorporação de ferramentas computacionais para análise desses textos revolucionou o estudo da linguagem. Assim sendo, a Lingüística de *Corpus*, ramo da Lingüística que utiliza *corpora* para o estudo da linguagem, abriu novas perspectivas e a possibilidade de explorar e investigar, em grande escala, regularidades e padrões inerentes à linguagem. Esta é, no entanto, uma área extremamente vasta, e uma discussão detalhada sobre o assunto vai além dos objetivos deste artigo. Focalizamos aqui a utilização de ferramentas e metodologias da Lingüística de *Corpus* especificamente para o estudo da tradução. O principal objetivo é, portanto, discutir a importância, benefícios e aplicações teóricas e pedagógicas do uso de *corpora* nos Estudos de Tradução, bem como tratar questões importantes referentes à compilação de *corpora* e à exploração deste valioso potencial disponível para os pesquisadores e teóricos da tradução. Mais especificamente, este artigo visa a abordar os objetivos e propostas da incipiente sub-disciplina Estudos de Tradução com base em *Corpora*.

10.17771/PUCRio.TradRev.25346

2 – O nascimento da disciplina

A incorporação de ferramentas e metodologias da Lingüística de *Corpus* para o estudo da tradução iniciou-se nas áreas da disciplina que utilizam recursos computacionais, tais como a terminologia e a tradução automática. Como explica Baker (1995), o uso de *corpora* teve um impacto favorável no campo da terminologia, onde termos deixaram de ser extraídos de listas pré-estabelecidas e passaram a ser obtidos a partir de textos autênticos. No campo da tradução automática, Baker (1995) destaca o uso de dados empíricos como ponto-chave para o aprimoramento dos sistemas de tradução; *corpora* computadorizados são atualmente usados por lingüistas na elaboração e/ou reformulação de regras lingüísticas e também pelos sistemas de tradução como uma fonte de conhecimento direta.

As metodologias com base em *corpora* também encontraram terreno fértil no ramo pedagógico da disciplina Estudos da Tradução, como ferramenta poderosa para auxiliar no treinamento de tradutores e na prática tradutória. Como destaca Olohan (2004:176), além de extremamente úteis na extração de terminologia, *corpora* eletrônicos podem ser usados para identificar estratégias e soluções adotadas por tradutores profissionais, assim como para avaliar a estrutura textual e discursiva, e ainda para examinar as convenções relacionadas ao tipo de texto ou gênero. Ademais, corpora podem também ser usados para investigar o estilo do autor, ou seja, identificar os artifícios literários e características lexicais, gramaticais e estilísticas que sejam recorrentes e mereçam ser tratados como uma estratégia deliberada por parte do autor (ibidem, p.180).

No entanto, um impacto ainda mais significativo da utilização de *corpora* computadorizados nos Estudos da Tradução deu-se a partir da sugestão original e inovadora da teórica Mona Baker (1993, 1995 e 1996) de utilização das metodologias e ferramentas da Lingüística de *Corpus* para investigar o fenômeno da tradução como um evento comunicativo *per se*, "moldado pelos seus próprios objetivos, pressões e contexto de produção" (Baker, 1996:175, tradução minha). Com a Lingüística de *Corpus*, Baker (1993) explica, pesquisadores e teóricos de tradução teriam em mãos os recursos necessários para explorar e pesquisar a natureza e as características específicas dos textos traduzidos, permitindo assim a redefinição dos principais objetivos, anseios e âmbito de abrangência dos estudos tradutórios. A influência do uso de *corpora* nos Estudos da Tradução como uma área acadêmica de pesquisa é bem ilustrada na afirmação de Baker (1993:235):

10.17771/PUCRio.TradRev.25346

> Grandes *corpora* oferecem aos teóricos de tradução uma oportunidade única para observar seu objeto de estudo e explorar o que o faz diferente de outros objetos de estudo, tais como a linguagem em geral ou mesmo qualquer outra forma de interação cultural. Eles possibilitam explorar também, em uma escala muito maior do que já foi possível até então, os princípios que governam o comportamento tradutório e as limitações sob os quais ele opera. Aí sim estão os objetivos de qualquer investigação teórica: definir e explicar o seu objeto de estudo. (Tradução minha).

Para Baker (1993 e 1996), essa nova abordagem reflete o desenvolvimento de paradigmas nos Estudos da Tradução que prepararam o terreno e contribuíram para uma mudança fundamental no principal foco da disciplina, dos textos-fonte para os textos traduzidos, dando uma atenção especial ao sistema e à cultura de chegada. Estas novas abordagens começaram a questionar a supremacia do texto de origem sobre o texto de chegada, além de reavaliarem a noção de equivalência até então vigente, segundo a qual "traduções deveriam procurar ser o mais equivalente possível aos originais, equivalência esta sendo entendida basicamente em termos de categorias semânticas ou formais" (Baker, 1993:235-6, tradução minha).

Nesse sentido, é importante ressaltar a contribuição significativa dos Estudos Descritivos da Tradução (DTS), em particular os trabalhos de Gideon Toury (1995), ao sugerirem a mudança de uma perspectiva prescritiva para uma orientação descritiva. Como explica Kenny (2001:49), o principal objetivo dos Estudos Descritivos da Tradução é descrever as "traduções como elas realmente ocorrem, e buscar explicar as características observadas nas traduções em relação aos contextos literários, culturais e históricos nos quais elas são produzidas" (tradução minha), contrastando assim com abordagens anteriores cuja principal preocupação era determinar o que uma tradução ideal deveria procurar alcançar. Dentro desta nova perspectiva, o foco de atenção passa a ser direcionado para a cultura de chegada, reservando-se uma ênfase especial aos dados empíricos.

Assim sendo, a Lingüística de *Corpus* e os Estudos da Tradução, considerando-se a perspectiva dos DTS, compartilham interesses comuns. Como explica Olohan (2004:16), ambas disciplinas adotam uma orientação descritiva em relação ao seu objeto de estudo. Ambas insistem na autenticidade dos dados, valorizando a linguagem realmente utilizada ao invés da intuição. Ambas se concentram em regularidades como normas de comportamento, apoiando-se no pressuposto de que ao identificar o típico, freqüente e regular, podemos também investigar o atípico e não-usual. Ambas disciplinas visam descrever a linguagem com base em análises quantitativas e qualitativas dos dados.

10.17771/PUCRio.TradRev.25346

No entanto, a Lingüística de *Corpus* e os Estudos da Tradução também revelam diferenças fundamentais. A primeira e principal diferença refere-se ao foco de interesse de cada uma destas disciplinas. A Lingüística de *Corpus* está interessada no estudo da linguagem em geral e na descrição de suas características. Os teóricos de tradução, por outro lado, estão interessados em tradução, tanto como processo como produto. O objetivo central dos Estudos da Tradução é portanto entender e explicar o processo tradutório e explorar a natureza dos textos traduzidos. Divergências também aparecem em relação à forma como os textos traduzidos são percebidos por cada disciplina. Tradicionalmente, a Lingüística de *Corpus* sempre mostrou uma tendência a menosprezar a linguagem traduzida, considerando-a desviante, distorcida e não representativa da linguagem. Conseqüentemente, os textos traduzidos são geralmente excluídos dos *corpora* de referência. A posição de Teubert (1996:247) ilustra bem esta visão negativa atribuída aos textos traduzidos:

> Traduções, por melhores e quase perfeitas que sejam (mas raramente são), irão sempre dar uma imagem distorcida da língua que elas representam. Os lingüistas nunca devem confiar em traduções para descrever uma língua. É exatamente por esta razão que traduções não são incluídas nos *corpora* de referência. Ao invés de representar a língua nos quais elas são escritas, as traduções são um espelho da suas respectivas línguas de partida. (Tradução minha).

Uma visão totalmente distinta é compartilhada por teóricos da tradução (dentre outros, Even-Zohar, 1990 [1978]; Toury, 1995; Baker, 1993, 1996, 2000 e 2004). Embora reconheçam que a linguagem traduzida seja realmente diferente da linguagem não-traduzida, argumentam e enfatizam que existem diversas razões e justificativas para tais diferenças. As traduções são produzidas em um contexto diferente, sob pressões diferentes, com limitações diferentes, além de refletirem influências e motivações diferentes. A afirmação de Baker (1996:177) reflete mais claramente o pensamento deste grupo:

> Dado que toda linguagem é padronizada e que essa padronização é influenciada pela finalidade para a qual a linguagem é usada e pelo contexto no qual ela é usada, a padronização dos textos traduzidos tem que ser obrigatoriamente diferente daquela dos textos produzidos originalmente em uma língua; a natureza e as pressões do processo tradutório certamente deixam traços na linguagem produzida por tradutores. A tradução é uma atividade lingüística realizada em um contexto único, distinto de uma produção textual normal, inclusive de textos produzidos por estudantes de uma língua estrangeira. (Tradução minha).

10.17771/PUCRio.TradRev.25346

Portanto, na perspectiva dos Estudos da Tradução, a tradução é um evento comunicativo genuíno e as características específicas e próprias dos textos traduzidos merecem ser analisadas, exploradas e explicadas. Estas são, portanto, as principais aspirações e objetivos da incipiente subdisciplina Estudos de Tradução baseados em *Corpora*.

3 – Tipos de c*orpora* para o estudo da tradução

Como mencionado anteriormente, *corpora* são compilados com base em critérios específicos, estabelecidos de acordo com os objetivos e finalidades de cada projeto. No caso dos *corpora* desenvolvidos para o estudo da tradução, os critérios propostos pela Lingüística de *Corpus* necessitam ser ajustados para que possam atender às necessidades dos teóricos de tradução e permitir a investigação de características específicas dos textos traduzidos. Por exemplo, uma atenção especial deverá ser dada ao critério referente às línguas envolvidas e ao contexto de produção, particularmente em relação às características dos tradutores, tais como se são profissionais ou aprendizes, se traduzem para a língua materna ou a partir dela, etc. (Baker, 1995).

Esta seção apresenta os tipos de *corpora* usados nos Estudos da Tradução, enfatizando os benefícios e aplicações destes para os ramos teórico e pedagógico da disciplina. Vale ressaltar que a terminologia empregada para se referir aos tipos de *corpora* usados para o estudo da tradução ainda não se encontra totalmente estabelecida; conseqüentemente, diferentes termos têm sido empregados por diferentes projetos de pesquisa. Adotamos aqui a terminologia utilizada por proeminentes teóricos da área de tradução (Baker, 1995; Kenny, 2001; Olohan, 2004). Três tipos de *corpora* são discutidos, a saber: (1) *corpora* multilíngües, (2) *corpora* paralelos, e (3) *corpora* comparáveis.

3.1 - Corpora multilíngües

Um *corpus multilíngüe* é composto por duas ou mais coleções de textos produzidos originalmente em suas respectivas línguas, ou seja, dois ou mais *corpora* monolíngües de línguas diferentes, compilados de acordo com os mesmos critérios e especificações (Baker, 1995). Como exemplo, Baker (ib.) cita o Projeto de Lexicografia Multilíngüe do Conselho Europeu, cujo objetivo era identificar regularidades no contexto textual de itens lexicais equivalentes em *corpora* de sete línguas européias: inglês, alemão, sueco, italiano, espanhol, húngaro e servo-croata. Para o inglês, por exemplo, foi utilizado o *corpus* Cobuild Bank of English, desenvolvido pela Universidade de Birmingham (Inglaterra). Já para o sueco, o projeto utilizou um *corpus* de sueco contemporâneo com 20 milhões de palavras, compilado pela Universidade de Gotemburgo (Suécia).

10.17771/PUCRio.TradRev.25346

Um exemplo da utilização de um *corpus* multilíngüe com aplicações diretamente relacionadas à área de tradução são os estudos contrastivos de Berber-Sardinha (1999 e 2000) sobre padronização lexical no português e no inglês. Apoiando-se nos trabalhos de Stubbs (1995a, 1995b e 1996) sobre perfil e prosódia semânticos[1] no inglês, Berber-Sardinha (1999 e 2000) examina esses mesmos aspectos para itens correspondentes do português brasileiro, desenvolvendo uma abordagem contrastiva para descrever perfis e prosódia semânticos do inglês e português. Os resultados mostram semelhanças e diferenças importantes entre padrões lexicais do inglês e do português, revelando inconsistências nos atuais dicionários bilíngües. Os dados são extraídos de dois grandes *corpora* de referência do inglês e português, o BNC (British National Corpus) e o Banco de Português respectivamente. O BNC é um *corpus* de inglês contemporâneo em linguagens escrita e falada, contendo aproximadamente 100 milhões de palavras[2] . O Banco de Português, compilado pela PUC/São Paulo, é considerado o maior *corpus* de português brasileiro no momento, com aproximadamente 233 milhões de palavras[3] , também incluindo tanto a linguagem escrita quanto a falada (Berber-Sardinha, 2004).

Além de sua contribuição valiosa para a lingüística contrastiva, especialmente no que se refere à lexicografia bilíngüe, os *corpora* multilíngües oferecem ainda outras aplicações e benefícios para á área de tradução. No campo da tradução automática, por exemplo, este tipo de *corpus* pode ser usado como fonte de conhecimento, contribuindo para um melhor desempenho dos sistemas computadorizados de tradução. Os *corpora* multilíngües servem também como um recurso valioso para o ensino e treinamento de tradutores, por permitirem o acesso a características e padrões lingüísticos em seu contexto natural e disponibilizarem evidências empíricas de itens e estruturas equivalentes em idiomas diferentes (Baker, 1995). Como explica Lindquist (1999), este tipo de *corpus* permite ao tradutor identificar o uso real de um determinado item lexical ou colocação da língua de chegada em um contexto específico. Neste sentido, os *corpora* multilíngües, particularmente aqueles compostos por textos técnicos e especializados, oferecem benefícios práticos para o ensino da tradução, já que podem ser usados como uma ferramenta valiosa para ajudar tradutores aprendizes a se familiarizar com padrões recorrentes da língua de chegada e para a extração de terminologia (Kenny, 1998).

Embora reconheça a importância dos *corpora* multilíngües no campo pedagógico dos Estudos da Tradução, Baker (1995) questiona a utilidade deste tipo de *corpus* para a elucidação de questões teóricas da disciplina. Para Baker (1995:233), o pressuposto básico de que "existe uma forma natural de expres-

10.17771/PUCRio.TradRev.25346

sar qualquer coisa em qualquer língua, e de que tudo que precisamos é encontrar a forma natural de expressar isso na língua A e língua B" (tradução minha), não deixa espaço para que os textos traduzidos sejam tratados como uma atividade lingüística independente e distinta, diferente daquela dos textos produzidos originalmente na mesma língua. Assim sendo, os *corpora* multilíngües não oferecem a possibilidade de investigação da natureza dos textos traduzidos ou do processo tradutório. A proposta de Baker (1993, 1995 e 1996) é, portanto, de uma mudança efetiva do foco dos estudos teóricos de tradução, direcionando-o para o sistema e cultura de chegada e dando ênfase aos textos traduzidos. É neste sentido que os *corpora* paralelos e comparáveis desempenham um papel fundamental para o desenvolvimento da disciplina de Estudos da Tradução.

3.2 - Corpora paralelos

Baker (1995:232) propõe o termo *corpus paralelo* para se referir a dois conjuntos de textos: um conjunto de textos em uma determinada língua de origem e um outro conjunto composto por versões traduzidas destes mesmos textos para um outro idioma. Os *corpora* paralelos são geralmente bilíngües, mas podem também ser multilíngües; ou seja, incluir traduções de um mesmo texto-fonte para diversos idiomas (Kenny, 2001:62; Olohan, 2004:25). Um bom exemplo deste tipo de *corpus* é o projeto COMPARA[4], que é composto por um conjunto de textos originalmente escritos em inglês e de suas respectivas traduções para o português, e por um outro conjunto de textos originalmente escritos em português e suas respectivas traduções para o inglês (Frankenberg-Garcia e Santos, 2002 e 2003). O COMPARA é portanto um *corpus* paralelo bidirecional, ou seja, o português é incluído tanto como língua de origem quanto como língua de chegada. Segundo Frankenberg-Garcia e Santos (ib.), o *corpus* inclui diversas variantes da língua portuguesa (européia, brasileira, asiática e africana) e também traduções de um mesmo texto-fonte para diferentes variantes do português e do inglês. Além disso, não foram impostas restrições quanto à data de publicação, ou seja, o *corpus* possibilita a inclusão de traduções de um mesmo texto-fonte publicadas em épocas diferentes. Em 2004, o COMPARA continha textos do gênero de ficção apenas, compreendendo um total de 2 milhões de palavras.

Um dos principais objetivos de um *corpus* paralelo é possibilitar a identificação de um determinado padrão ou unidade nas línguas de partida e de chegada simultaneamente. Técnicas de alinhamento são utilizadas para que seja possível estabelecer ligações entre os textos de origem e de chegada. Os

10.17771/PUCRio.TradRev.25346

corpora paralelos servem como uma ferramenta preciosa para avaliar o comportamento traducional de um determinado par de idiomas, além de serem extremamente úteis na investigação do relacionamento entre padrões lexicais e sintáticos nas línguas de origem e de chegada, e de ocorrências isoladas de "tradutorês" (Kenny 1998). Como os *corpora* multilíngües, os *corpora* paralelos também desempenham uma função importante no treinamento de tradutores, no desenvolvimento de sistemas de tradução automática e na lexicografia bilíngüe (Baker, 1995; Kenny, 1998). No entanto, para Baker (1995), a mais valiosa contribuição dos *corpora* paralelos para a disciplina Estudos da Tradução é possibilitar a mudança de uma perspectiva prescritiva para uma perspectiva descritiva. Como explica Baker (ib.), os *corpora* paralelos fornecem evidências empíricas de estratégias e alternativas adotadas por tradutores para solucionar dificuldades e obstáculos encontrados na prática tradutória. Tais evidências, além de servirem como um valioso recurso pedagógico para o treinamento de tradutores, podem ser também extremamente úteis na "investigação de normas tradutórias em contextos históricos e sócio-culturais específicos" (Baker, 1995:231, tradução minha).

Um bom exemplo de como um *corpus* paralelo pode ser usado na investigação da influência do processo tradutório no processamento e produção da linguagem é o estudo de Dorothy Kenny (2001) sobre criatividade lexical em tradução. Tendo como principal objetivo abordar o processo de "normalização"[5] lexical em tradução, o estudo examina a tradução de itens lexicais criativos do alemão para o inglês, visando determinar se estes itens foram substituídos por formas mais convencionais na língua de chegada. Para Kenny (2001:31-32), itens lexicais criativos são entendidos como palavras ou colocações não-usuais e atípicas, que revelem criatividade no uso da linguagem. Os dados são extraídos de um *corpus* paralelo bilíngüe, contendo textos experimentais originalmente produzidos em alemão e suas respectivas traduções para o inglês, totalizando aproximadamente um milhão de palavras em cada sub-*corpus*. O ponto de partida é o texto de origem em alemão, e o primeiro passo é selecionar itens ou colocações de acordo com os seguintes critérios: (1) ocorrer apenas uma vez no *corpus*; (2) para os itens ou colocações recorrentes, ter sido usado por apenas um determinado autor (Kenny, 2001:128-129). A convencionalidade, ou não, dos itens ou padrões selecionados é avaliada de acordo com a freqüência dos mesmos em um *corpus* de referência do alemão (*Corpus Mannheimer*) e com a intuição de falantes nativos de alemão. Uma vez considerados criativos, examinaram-se as respectivas traduções desses itens lexicais para o inglês, com o objetivo de avaliar se estes são "normalizados", isto

10.17771/PUCRio.TradRev.25346

é, se o tradutor traduz um item lexical (ou colocação) criativo do texto de origem por um item lexical (ou colocação) igualmente criativo da língua de chegada (p.142-188). A criatividade dos itens traduzidos para o inglês é avaliada com base em um *corpus* de referência do inglês (BNC). Kenny (2001:187) observa que itens lexicais criativos que ocorreram apenas uma vez no sub-*corpus* de textos em alemão tendem a ser normalizados em suas respectivas traduções para o inglês. Por outro lado, itens que apesar de recorrentes são peculiares a um determinado autor não mostraram uma tendência a ser traduzidos por uma forma mais convencional da língua de chegada. A tendência à normalização tampouco é evidente na tradução das colocações (p.207-208). No entanto, Kenny adverte, no caso dos itens recorrentes mas peculiares a um determinado autor, todos os exemplos analisados foram traduzidos por um único tradutor (p.187). No caso das colocações, o estudo examina apenas aquelas de um único item lexical (*Auge*, em alemão, e sua tradução correspondente em inglês, *eye*), sendo que a análise de outros itens talvez possa gerar resultados diferentes (p.207-208). Kenny (2001:210) conclui com a ressalva de que, apesar das evidências empíricas de normalização lexical, o estudo também mostra que a normalização não é uma prática automática na tradução de itens lexicais criativos do texto de origem. Na realidade, os dados revelam a engenhosidade e criatividade de diversos tradutores.

3.3 - Corpora comparáveis

O terceiro tipo de *corpus* proposto por Baker (1995:234) para o estudo da tradução é o *comparável*, que é um *corpus* monolíngüe composto por dois sub-*corpora*: um sub-*corpus* de textos traduzidos para uma determinada língua, a partir de uma ou mais línguas-fonte, e um outro de textos não-traduzidos, ou seja, textos originalmente produzidos na língua em questão. Portanto, na concepção de Baker (ib.), o sub-*corpus* traduzido consiste em textos produzidos por tradutores, e o não-traduzido é composto por textos nessa mesma língua, mas não produzidos via tradução. Baker (1995:234) complementa que esses dois sub-*corpora* "devem cobrir um domínio, variedade de linguagem e período de tempo semelhantes, e ter tamanhos comparáveis" (tradução minha). Em outras palavras, para que seja possível compará-los, é essencial que esses dois sub-*corpora* tenham sido compilados de acordo com os mesmos critérios e especificações, e sejam de tamanho semelhante.

A grande maioria dos estudos de tradução que têm por objetivo comparar textos traduzidos e não traduzidos de uma mesma língua baseia-se no inglês

10.17771/PUCRio.TradRev.25346

(dentre outros, Laviosa-Braithwaite, 1996; Olohan, 2002 e 2003; Mutesayire, 2005). Neste caso, a comparação é geralmente feita entre o Corpus de Inglês Traduzido (TEC - Translational English Corpus)[6] e um sub-*corpus* do BNC (British National Corpus). O TEC foi elaborado e compilado pelo Centre for Translation and Intercultural Studies (CTIS) da Universidade de Manchester e consiste em uma coleção de textos traduzidos para o inglês a partir de diversas línguas fonte. Em 2004, o TEC continha aproximadamente 8 milhões de palavras, incluindo traduções das seguintes línguas-fonte: francês, italiano, espanhol (europeu, sul-americano e centro-americano), português (europeu e brasileiro), alemão, polonês, galês, húngaro, turco, sérvio, sueco, japonês, russo, norueguês, finlandês, árabe, tâmil, tailandês, hebraico e chinês (Olohan, 2004:60; Mutesayire, 2005). O *corpus* é dividido em 4 seções: ficção, biografia, revistas de bordo e artigos de jornais, sendo ficção o gênero predominante, com 80% dos textos. O TEC inclui traduções publicadas a partir de 1983; todos os tradutores são falantes nativos de inglês ou têm o inglês como língua de uso habitual; todos os textos foram incluídos na íntegra (Olohan, ib.; Mutesayire ib.).

Corpora comparáveis estão também disponíveis para o finlandês e para o sueco. O Corpus de Finlandês Traduzido (CFT), desenvolvido pela Escola Savonlinna de Estudos da Tradução (Joensuu, Finlândia), é uma coleção de textos contemporâneos em finlandês traduzido e não-traduzido. O CFT contém aproximadamente 9,6 milhões de palavras: 5,8 milhões de palavras em finlandês traduzido e os restantes 3,8 milhões de palavras em finlandês não-traduzido (Olohan, 2004:60-61). No *corpus* traduzido estão incluídas traduções para o finlandês de textos-fonte em inglês, russo, alemão, francês, espanhol, holandês, norueguês, sueco, húngaro e estoniano. Todos os textos foram publicados entre 1995 e 2000 e, em termos de gênero, o CFT é dividido em quatro seções: ficção, prosa acadêmica, ciência popular e literatura infantil (Olohan, ib.). Já o corpus *comparável de sueco*, compilado pela Universidade de Gotemburgo (Suécia), é composto por 75 romances publicados em sueco no ano de 1976, sendo dividido em duas seções: um sub-*corpus* de textos escritos originalmente em sueco e o outro sub-*corpus* de textos traduzidos para o sueco, a grande maioria traduções do inglês (Kenny, 2001:59).

Um outro exemplo a ser citado é o Corpus Comparável de Português Brasileiro (CCPB), que é composto por textos em português brasileiro traduzido e não-traduzido. Inicialmente elaborado e compilado como parte de uma pesquisa de doutorado (Dayrell, 2005), o CCPB é um projeto a ser expandido com o objetivo de propiciar o desenvolvimento de outros estudos baseados em

10.17771/PUCRio.TradRev.25346

corpora de português brasileiro. O CCPB contém apenas textos literários publicados no Brasil a partir de 1980, sendo que foi dada prioridade às obras publicadas a partir de 1990. Todos os livros incluídos no *corpus* foram considerados *best-sellers* no Brasil durante o período analisado, conforme as listas de *best-sellers* publicadas pela revista *Veja* entre 1991 e 2001. Ademais, considerou-se apenas a literatura adulta, ou seja, o *corpus* não inclui textos classificados como literatura infantil ou infanto-juvenil. Em termos de gênero, o CCPB contém textos de ficção e auto-ajuda. A opção por esses dois gêneros deve-se ao fato de que estes são os gêneros mais populares no Brasil no período analisado (*Veja* 1996, 2001a e 2001b) e, portanto, os que mais têm probabilidade de oferecer um número razoável de textos traduzidos e não traduzidos. Assim sendo, o CCPB é composto por quatro sub-*corpora*: ficção traduzida, ficção não-traduzida, auto-ajuda traduzida e auto-ajuda não-traduzida. O *corpus* contém um total aproximado de 2 milhões de palavras; cada um dos 4 subcorpora contém aproximadamente meio milhão de palavras. Todos os textos foram incluídos na íntegra e tentou-se, na medida do possível, diversificar a seleção de textos em termos de autores, tradutores e editoras. Para a seleção de textos traduzidos, além dos critérios mencionados acima, considerou-se também a língua de origem da tradução, tendo sido selecionadas apenas traduções a partir de textos escritos originalmente em inglês. Traduções indiretas – ou seja, aquelas feitas via outra tradução e não a partir do texto fonte original – não foram incluídas. Todos os tradutores são falantes nativos do português brasileiro, e foi dada prioridade para as traduções cujos textos de origem também tenham sido publicados a partir de 1980.

Uma diferença importante entre os *corpora* paralelos e os comparáveis é que esses últimos não são usados para comparar línguas de partida e de chegada e, portanto, não têm por objetivo identificar normas tradutórias, estratégias adotadas por tradutores nem exemplos de "tradutorês". Como esclarece Baker (1995:235), a principal contribuição de um *corpus* comparável é permitir a investigação de características "que sejam restritas aos textos traduzidos ou que ocorram com uma freqüência consideravelmente mais alta ou mais baixa nos textos traduzidos" (tradução minha) que nos textos não traduzidos. Um *corpus* comparável robusto, composto por uma diversificada gama de autores e tradutores, assim como traduções de diversas línguas fonte, possibilita a identificação de características que são específicas dos textos traduzidos, independentemente da influência da língua de origem ou das preferências estilísticas de tradutores individuais. Como afirma Baker (1996:178), ao compararem textos traduzidos e não-traduzi-

10.17771/PUCRio.TradRev.25346

dos de uma mesma língua, os pesquisadores de tradução podem finalmente "identificar tipos de comportamento lingüístico que são específicos dos textos traduzidos, padrões de comportamento lingüístico que, em outras palavras, são gerados pelo processo de mediação durante a tradução" (tradução minha).

Para ilustrar como os *corpora* comparáveis podem ser usados na investigação das características dos textos traduzidos, vale citar o trabalho pioneiro de Laviosa-Braithwaite (1996), cujo objetivo é investigar o processo de "simplificação"[7] em tradução. Os textos traduzidos e não-traduzidos são analisados sob três aspectos: variedade lexical, carga de informação e tamanho de sentenças (Laviosa, 2002:59-64). Os dados são extraídos de um *corpus* comparável do inglês, consistindo de uma coleção de textos traduzidos para o inglês (TEC) e um outro sub-*corpus* de textos em inglês não-traduzido (extraído do BNC). Cada um destes sub-*corpora* contém aproximadamente um milhão de palavras e inclui dois gêneros: textos literários (ficção e biografia) e textos jornalísticos (Laviosa, ibidem). A variedade lexical é analisada sob três perspectivas: proporção entre palavras de alta e baixa freqüência, proporção de *headwords* (nesse caso, as primeiras 108 palavras mais freqüentes no *corpus*) e quantidade de lemas. Os resultados mostram que proporção entre palavras de alta freqüência e as palavras de baixa freqüência é mais alta no sub-*corpus* traduzido que no sub-*corpus* não-traduzido; a proporção de *headwords* é maior no sub-*corpus* traduzido, ou seja, o nível de repetição de palavras mais freqüentemente usadas é mais elevado no sub-*corpus* traduzido; e a lista de *headwords* do sub-*corpus* traduzido contém um número menor de lemas. Estes resultados são interpretados como indicadores de uma tendência dos textos traduzidos a apresentar menos variedade lexical que os textos não traduzidos. A carga de informação é examinada em termos de densidade lexical, ou seja, a proporção entre itens lexicais e itens gramaticais. Os resultados mostram uma tendência de a densidade lexical ser mais baixa nos textos traduzidos que nos textos não traduzidos. Em relação ao tamanho das sentenças, a média é menor para os textos traduzidos do que para os textos não traduzidos apenas no gênero jornalístico. Nesse estudo, Laviosa-Braithwaite (1996) observa dados que apontam para uma tendência dos textos traduzidos a apresentar uma linguagem mais simplificada que os textos não traduzidos. No entanto, Laviosa (2002:63) adverte, dois fatores talvez possam ter influenciado nos resultados: a análise é baseada em um *corpus* de tamanho limitado e a grande maioria dos textos traduzidos consiste em traduções de textos fonte em línguas românicas.

10.17771/PUCRio.TradRev.25346

4 – Considerações finais

Neste artigo, discutimos os principais objetivos e pressupostos da disciplina Estudos de Tradução com base em *Corpora* (ETC), dando ênfase às aplicações teóricas e pedagógicas do uso de *corpora* para o estudo da tradução. As aplicações são ilimitadas e valiosas; porém, em comparação com outras áreas de pesquisa da linguagem, muito ainda está por ser feito. Esperamos, portanto, que num futuro próximo possamos usufruir do grande potencial dessa nova área de pesquisa.

---

[1] Perfil semântico refere-se "ao teor da colocação, coligação ou prosódia semântica, definido a partir de generalizações a respeito do conteúdo semântico dos itens envolvidos no padrão" (Berber Sardinha, 1999). Prosódia semântica refere-se à conotação – positiva, negativa ou neutra – resultante da associação de itens lexicais. Por exemplo, o verbo 'causar' tende a se associar a itens com uma conotação negativa (problemas, danos, morte, mortes, prejuízos, etc.) (ibidem).

[2] Mais informações sobre o BNC estão disponíveis no site: http://info.ox.ac.uk/bnc (acessado em julho/2005).

[3] Mais informações sobre o Banco de Português estão disponíveis no site: http://lael.pucsp.br/corpora/bp/conc/index.html (acessado em julho/2005).

[4] COMPARA é parte de um projeto para o processamento computacional do português, coordenado pelo Centro de Recursos Português (Linguateca). Mais informações sobre o COMPARA estão disponíveis no site: http://www.linguateca.pt/COMPARA/ (acessado em julho/2005).

[5] O termo *normalização* foi proposto por Baker (1996:176) para indicar uma "tendência [de tradutores] a ajustar-se aos padrões e práticas que são comuns na língua de chegada, chegando até mesmo a exagerá-los" (tradução minha).

[6] Mais informações sobre o TEC estão disponíveis no site: http://www.llc.manchester.ac.uk/Research/Centres/CentreforTranslationandInterculturalStudies/ (acessado em julho/2005).

[7] O termo *simplificação* foi proposto por Baker (1996:176) para indicar "a idéia de que tradutores inconscientemente simplificam a linguagem, mensagem ou ambas" (tradução minha).

10.17771/PUCRio.TradRev.25346

Referências bibliográficas

ATKINS, Sue, CLEAR, Jeremy & OSTLER, Nicholas (1992) "Corpus design criteria". *Literary and Linguistic Computing*, 7 (1): 1-16.

BAKER, Mona (1993) "Corpus linguistics and translation studies: implications and applications". M. Baker, G. Francis & E. Tognini-Bonelli (orgs.) *Text and Technology: in Honour of John Sinclair*, 233-250. Amsterdam/ Philadelphia: John Benjamins.

_____ (1995) "Corpora in translation studies. An overview and suggestions for future research". *Target* 7(2): 223-243.

_____ (1996) "Corpus-based translation studies: the challenge that lie ahead". Harold Somers (org.) *Terminology, LSP and Translation: Studies in Language Engineering in Honour of Juan C. Sager*, 175-187. Amsterdam/ Philadelphia: John Benjamins.

_____ (2000) "The translational English corpus at UMIST (University of Manchester Institute of Science and Technology". W. Kubinski, O. Kubinska and T. Z. Wolanski (orgs.), *Przeladajac Nieprzekladalne: Materialy z I Miedzynarodowej Konferencji Translatorycznej Gdansk-Elblag*, 493-502. Gdansk: Wydawnictwo Uniwersytetu Gdanskiego.

_____ (2004) "A corpus-based view of similarity and difference in translation". *International Journal of Corpus Linguistics* 9:2, 167-193.

BERBER-SARDINHA, Tony (1999) "Estudo baseado em *corpus* da padronização lexical no português brasileiro: colocações e perfis semânticos". *PROPOR"99. IV Encontro para o Processamento Computacional da Língua Portuguesa Escrita e Falada*, Évora, Portugal, 269-287. http://lael.pucsp.br/~Tony. Acesso em julho de 2005.

_____ (2000) "Semantic prosodies in English and Portuguese: a contrastive study". *Cuadernos de Filologia Inglesa*, 9 (1): 93-110, Espanha. http://lael.pucsp. br/~Tony. Acesso em setembro de 2004.

_____ (2004) *Lingüística de corpus*, São Paulo: Manole.

DAYRELL, Carmen (2005) *Investigating lexical patterning in translated Brazilian Portuguese: a corpus-based study*, Tese de doutorado inédita. Manchester: The University of Manchester.

EAGLES (1996) "Preliminary recommendations on corpus typology". J. McH Sinclair (org.), *EAGLES document EAG-TCWG-CTYP/P*, May/1996. http://www.ilc.cnr.it/EAGLES96/home.html. Acesso em dezembro de 2004.

EVEN-ZOHAR, Itamar (1990[1978]) "Polysystem theory". *Poetics Today* (*Special Issue*), 11(1): 9-26.

10.17771/PUCRio.TradRev.25346

FRANKENBERG-GARCIA & SANTOS (2002) "COMPARA, Um *corpus* paralelo de português e inglês na web" in Translation". S. E. O. Tagnin (org.) *Cadernos de Tradução: Corpora e Tradução* 9: 61-79. http://www.cadernos.ufsc.br/ Acesso em dezembro de 2004.

_____ (2003) "Introducing COMPARA, the Portuguese-English parallel corpus". Federico Zanettin, Silvia Bernardini & Dominic Stewart (orgs.) *Corpora in Translation Education*, 71-87. Manchester: St. Jerome Publishing.

KENNY, Dorothy (1998) "Corpora in translation studies". Mona Baker (org.) *Routledge Encyclopedia of Translation Studies*, 50-53. London/New York: Routledge.

_____ (2001) *Lexis and creativity in translation: a corpus-based study*. Manchester: St. Jerome Publishing.

LAVIOSA, Sara (2002) *Corpus-based translation studies: theory, findings, applications*, Amsterdã/Nova York: Rodopi.

LAVIOSA-BRAITHWAITE, Sara (1996) *The English Comparable Corpus (ECC): a resource and a methodology for the empirical study of translation*. Tese de doutorado inédita. Manchester: UMIST.

LINDQUIST, H. (1999) "Electronic corpora as tools for translation". Gunilla Anderman & Margareth Rogers (orgs.) *Word, Text, Translation*, 179-189. Clevedon: Multilingual Matters.

MUTESAYIRE, Martha (2005) *Does translated English favour lexical explicitation as a textual strategy compared to non-translated English: A corpus-based study*. Tese de doutorado inédita. Manchester: The University of Manchester.

OLOHAN, Maeve (2002) "Leave it out! Using a comparable *corpus* to investigate aspects of explicitation in translation". S. E. O. Tagnin (org.) *Cadernos de Tradução: Corpora e Tradução* 9: 153-169. http://www.cadernos.ufsc.br/ . Acesso em dezembro de 2004.

_____ (2003) "How frequent are the contractions? A study of contracted forms in the translational English *corpus*". *Target*, 15(1): 59-89.

_____ (2004) *Introducing corpora in translation studies*, London/New York: Routledge.

STUBBS, Michael (1995a) "Collocations and semantic profiles: on the cause of trouble with quantitative studies". *Functions of Language* 2 (2): 23-55.

_____ (1995b) "Corpus evidence for norms of lexical collocation". Guy Cook & Barbara Seidlhoper (orgs.) *Principle and Practice in Applied Linguistics*, 245-256. London: Oxford U.P.

10.17771/PUCRio.TradRev.25346

_____ (1996) *Text and corpus analysis – computer-assisted studies of language and culture*. Oxford: Blackwell.

TEUBERT, Wolfgang (1996) "Comparable or parallel corpora?". *International Journal of Lexicography*. Special Issue, 9 (3): 238-264.

TOURY, Gideon (1995) *Descriptive translation studies and beyond*. Amsterdam/ Philadelphia: John Benjamins.

Veja (1996) "Ranking renovado: a lista dos mais vendidos terá uma categoria dedicada à auto-ajuda e ao esoterismo". Número 1.474, 11 de dezembro, 130-131.

_____ (2001a) "Os mais vendidos de 2000 – comentário". Número 1.682, 10 de janeiro.

_____ (2001b) "Os mais vendidos de 2001 – comentário". Número 1.731, 19 de dezembro.