http://dx.doi.org/10.5007/2175-7968.2017v37n3p329

ZANETTIN, Federico. *Translation-Driven Corpora*: Corpus Resources for Descriptive and Applied Translation Studies. New York: Routledge, 2012. 244 p.

Marcia Goretti Carvalho[*]
Universidade Federal do Pará

O uso de textos eletrônicos e de ferramentas tecnológicas tem se tornado constante para quem trabalha com línguas e tradução. Federico Zanettin, em *Translation-Driven Corpora: Corpus Resources for Descriptive and Applied Translation Studies,* oferece um livro sobre a criação de corpora de textos eletrônicos e sua exploração em pesquisa aplicada e descritiva na área da tradução, na perspectiva do uso de tecnologias para analisar textos, aplicá-las na linguística de corpus e nos estudos da tradução e usar corpora em cursos para tradutor e em pesquisas com corpus realizadas por estudiosos da tradução, tradutores, empresas de tradução e linguistas computacionais.

Zanettin é professor Associado de Língua Inglesa e de Tradução na Universidade de Perugia, na Itália. É PhD em Estudos da Tradução pela Universidade de Bolonha. Sua atividade de pesquisa é, principalmente, em estudos da tradução com base em corpus. Publicou artigos em revistas especializadas, em volumes e em enciclopédias.

[*] Professora Adjunta da Universidade Federal do Pará. Doutoranda no Programa de Pós-graduação em Estudos da Tradução, da Universidade Federal de Santa Catarina. Belém, Pará, Brasil. E-mail: marciagoretti@bol.com.br

Também já ministrou cursos, conferências e palestras em universidades italianas e no exterior. Na área de estudos da tradução com base em corpus, publicou *Corpora in Translator Education* (2003), com Silvia Bernardini e Dominic Stewart pela Editora St Jerome.

Segundo Zanettin, as metodologias de corpus têm se tornado quase convencional nos estudos descritivos da tradução com base em corpus. Ao mesmo tempo, aplicativos computacionais com base em corpora, memórias de tradução e sistemas de tradução automática (*machine-assisted translation*), cada vez mais, fazem parte da vida de empresas e de tradutores técnicos ou especializados. Zanettin aborda vários pressupostos teóricos, agregando exemplos e estudos de caso com tarefas práticas e um DVD. Há referências no fim dos capítulos e uma lista de referências no fim do livro.

Além da introdução e da conclusão, o livro é dividido em seis capítulos, bem ilustrados, com tarefas e leituras complementares. Cada capítulo tem uma estrutura autônoma e alguns tópicos são discutidos ou mencionados mais de uma vez com o encaminhamento para as páginas em que esses tópicos foram mencionados. As tarefas estão relacionadas aos exemplos no capítulo e têm a ajuda do DVD. Essas atividades estão impressas e os usuários deverão ter acesso à computação *online* para executá-las.

O Capítulo 2 oferece uma introdução a corpora e a aplicações de metodologias da linguística de corpus aos estudos da tradução. Uma definição de corpus é apresentada no início do capítulo como "a collection of texts put together according to some informed criteria […]" (p. 7). Zanettin comenta a trajetória do uso de corpora e examina, brevemente, alguns corpora como *Brown* e *LOB* corpora, *BNC* e faz referência às listas de palavras e às concordâncias como "the primary data used in corpus-based investigations" (p. 9).

O capítulo discute vários tipos de corpora usados na pesquisa aplicada e descritiva de tradução: corpus comparável monolíngue, cor-

pus bilíngue (paralelo ou comparável) e corpus multilíngue. Ele avalia e discute projetos (como o *TEC*, criado por Mona Baker) e a pesquisa com base em corpus. Em seguida, Zanettin trata das regularidades das traduções com base nos estudos de vários teóricos da área como Baker, Holmes, Toury e outros, e apresenta os universais da tradução que Baker (1993) define como traços que ocorrem tipicamente em textos traduzidos e não são o resultado de interferência de sistemas linguísticos específicos, Chesterman (2004) propõe uma distinção entre *T-Universals* e *S-Universals*, e Zanettin ressalta a distinção entre o processo e o produto da tradução.

Neste capítulo, Zanettin, com base nos teóricos, analisa os universais hipotéticos: a simplificação, a explicitação, a padronização, a tradução de itens únicos, as colocações atípicas e a interferência. O autor aborda também questões sobre as regularidades dos tradutores e das línguas e identifica uma variedade de corpora usados na pesquisa descritiva que geralmente possuem dois ou mais subcorpora, comparados para achar similaridades e diferenças entre textos-fonte e textos-alvo ou entre línguas, para isolar traços distintivos potenciais de textos traduzidos ou de línguas e para estudar estilos de tradução e de gêneros. Alguns estudos investigam variedades da língua traduzida produzidas por tipos específicos de usuários da língua.

Esse capítulo é, na realidade, uma introdução geral da tipologia de corpus referente à tradução e à pesquisa com base em corpus. Zanettin apresenta uma visão geral do uso de corpora no ensino-aprendizagem da tradução e ressalta o uso de corpora na tradução automática e na linguística computacional. Alguns estudos recentes têm apresentado sugestões à pesquisa com base em corpora de interpretação (dublagem de filmes e interpretação em conferência) e aos estudos sobre corpora multimodais.

Três diferentes tarefas (p. 34-38) são propostas, neste capítulo, para praticar as metodologias de pesquisa discutidas. A primeira tarefa

consiste em repetir uma pesquisa feita por Olohan e Baker (2000) sobre o *that*, usando o *Translational English Corpus*. Na segunda tarefa, as mesmas técnicas e procedimentos são empregados para comparar os achados da pesquisa com aqueles obtidos do *COMPARA*. Neste experimento, apenas os componentes em inglês (traduções e não-traduções) do corpus são selecionados e usados. Na terceira tarefa, examina-se a interface *online* do *Learner Translation Corpus* com textos originais e traduzidos em línguas europeias, produzidos por estagiários de tradução e por tradutores profissionais.

O capítulo 3 trata de design de corpus e de aquisição. É um projeto complexo com diferentes estágios e custos identificados por Zanettin, e uma lista para a construção de corpus envolvendo custos e planejamento. Depois das fases principais, Zanettin apresenta o tamanho e a composição de corpora. Em relação ao tamanho, um corpus especializado pode ser restringido para incluir apenas um tipo de texto específico ou de assuntos específicos ou para um público específico. Quanto à composição, sabe-se que um corpus é composto da amostra de uma língua ou de uma variedade específica dela. Surgem, então, critérios para definir a população de textos a serem considerados na amostra e utilizados na seleção de itens textuais a serem incluídos no corpus.

A avaliação da composição interna de um corpus em relação ao seu tamanho é necessária para avaliar a representatividade e a comparabilidade. A representatividade, segundo Zanettin, é garantida pela segmentação, em um número apropriado de categorias, do universo textual abordado pelos compiladores de corpus, e a definição das categorias internas dependerá do tipo de corpus. Isto é especialmente relevante para estudos baseados em corpora referentes à tradução que geralmente envolvem uma comparação de achados derivados de subcorpora nas mesmas línguas e em diferentes línguas.

Para explicar melhor as decisões tomadas no projeto de um corpus, tem-se, nesse capítulo, um estudo de caso detalhado: o design

do corpus *CEXI*, um projeto da Universidade de Bologna, pensado como um recurso para estudos descritivos e um auxiliar para aprendizagem de línguas e treinamento para tradutores. Idealmente, a composição do corpus teria permitido análise de subcorpora paralelos, análise de subcorpora comparáveis monolíngues e análise de corpus comparável bilíngue de textos não traduzidos.

Zanettin ressalta que critérios ideais para design de corpus geralmente precisam ser ajustados a algumas restrições como financiamento de projeto, restrições de direito autorais ou falta de material de corpus apropriado ou de ferramentas. Por isso, é necessário examinar as implicações de se criar corpora da Web, com enormes quantidades de material de textos em formato eletrônico. A Web é uma fonte de dados de corpus em relação ao tamanho e à representatividade. Discutem-se, então, as ferramentas disponíveis para usar a Web como uma língua que, na visão de Zanettin, é uma espécie de corpus quase exaustivo da língua na comunicação escrita eletrônica, entretanto parece que a Web não tem o critério principal de um corpus: a sua construção e montagem para a investigação linguística.

Pode-se criar corpora monolíngues, comparáveis multilíngues e bilíngues, gerais e especializados; fazer download e processar documentos recuperados da Web, usando diretórios e mecanismos de busca da Internet. Tais corpora podem também ser criados pelas rotinas semiautomáticas implementadas por programas e serviços *online*. Em relação à Web e a seus recursos, o autor menciona alguns exemplos de "*subwebs*" especializadas, assim como pode-se limitar uma busca pelo tipo de arquivo como os de PDF, com um conteúdo menos volátil do que outros arquivos. Zanettin apresenta várias figuras de ferramentas computacionais para análise de corpus (p. 59-67).

As tarefas do capítulo 3 (p. 68-72) se referem a um rascunho de um projeto de criação de corpus e ao design de dois corpora DIY (faça-você-mesmo) da Web, decidindo o autor do projeto sobre a

natureza precisa do mesmo. Há também uma grade para esboçar esse projeto de design de corpus como guia para quem for desenvolver o corpus. Essa tarefa continua no capítulo 6 em que o leitor reconsidera algumas questões, com um foco na construção de corpora multilíngues. Os dois corpora DIY serão criados peneirando manualmente resultados de buscas da Web (em Inglês) e compilando semiautomaticamente um corpus comparável bilíngue.

O capítulo 4 trata dos diferentes estágios da compilação de corpus e de seu uso, da codificação de corpus e da anotação para indexação e recuperação de dados, dando atenção aos métodos e aos padrões para a anotação de corpora robustos a serem usados em estudos descritivos da tradução. Em relação à anotação de informação em um corpus, o autor apresenta três diferentes aspectos: informação documental sobre os textos em um corpus como um todo; informação documental sobre cada texto no corpus; e informação estrutural sobre os próprios textos, com informação linguística de cada palavra em cada texto. E cita as vantagens de se usar software de análise de corpus como o *WordSmith Tools* ou o *AntConC*.

Padrões comuns de codificação devem ser adotados pelos pesquisadores e uma abordagem modular pode ser usada para acomodar diferentes camadas de anotação para codificar diferentes traços textuais. Zanetti distingue dois tipos principais de anotação ou marcação: a processual ou de apresentação (formatação visual de um texto) e a descritiva ou estrutural (conteúdo de um documento linguístico e extralinguístico: informação bibliográfica sobre um texto, informação sobre sua estrutura lógica ou sobre as menores unidades textuais). Os textos eletrônicos nativos, por sua vez, precisam ser processados antes de se tornarem parte de um corpus como um arquivo PDF, baixado da Web, precisa ser convertido e padronizado no formato de texto simples, antes que o arquivo de texto seja anotado.

O capítulo apresenta uma pequena introdução a alguns padrões existentes para anotação de corpus como as diretrizes do *Text En-*

coding Initiative e o XML *Corpus Encoding Standard*. Zanettin informa que padrões de codificação se desenvolveram para o uso na codificação de corpora de língua, para a pesquisa e para profissionais da área da tradução assim como procedimentos e padrões de anotação para alinhar e anotar corpora paralelos nos estudos descritivos da tradução e no ensino da tradução. A anotação de um corpus, usando padrões XML, TEI e XCES, certamente requer um conhecimento de como os esquemas de anotação funcionam. Entretanto, procedimentos de anotação podem ser simplificados com a ajuda de protocolos de anotação.

Há, nesse capítulo, um modelo de cabeçalho e uma introdução resumida de como a anotação linguística e estrutural podem ser gravadas em um documento com o formato XML TEI. Diferentes camadas de anotação podem ser armazenadas com a implementação de um modelo no qual a anotação é mantida separada do texto corrente (*stand-off annotation*). Exemplos são apresentados para ilustrar como funciona a anotação *stand-off* (p. 98-100).

As tarefas desse capítulo (p. 101-109) permitem ao usuário criar e pesquisar um corpus de um documento simples. Primeiro, um documento com formato XML TEI é criado de um arquivo PDF, marcando manualmente a informação estrutural e documentária no texto. O documento é, então, linguisticamente anotado, validado e indexado, e finalmente o corpus criado é explorado através de um grupo de buscas/pesquisa de amostras. As diferentes partes do software usadas para processar o texto em vários estágios (conversão de texto, anotação automática e manual, indexação, recuperação do texto) estão livremente disponíveis na Web e incluídas no DVD.

O capítulo 5 apresenta as ferramentas de software usadas para criar, gerenciar e analisar corpora e descreve métodos e técnicas que permitam aos usuários finais compreender os dados do corpus. Discutem-se requisitos de hardware e de software necessários para cumprir os vários estágios da construção de corpus. As ferramentas

e técnicas básicas e avançadas de análise de corpus são ilustradas com exemplos práticos para mostrar como elas podem ser usadas na investigação do padrão lexical. Há, também, tabelas com lemas, listas de palavras e de palavras-chave e figuras de *prints* do "*WordSmith Tools 5"*; do *"The Sketch Engine"*; do *"WordSmith's Concord tool"* e de outras ferramentas (p. 117-139).

Zanettin ressalta que o hardware depende dos sistemas operacionais requeridos pelo software usado. Nem todos os aplicativos especificamente criados para anotação de corpus, gerenciamento e análise trabalharão em todos os ambientes de software. Isso varia de acordo com o objetivo do projeto e, para recursos de corpus estáveis, precisa-se de uma equipe na construção desse corpus, para uso local ou para livre acesso na internet.

Sobre a compilação de corpus, o primeiro estágio é obter versões limpas de textos para serem incluídos em um corpus, ou versões já anotadas de documentos existentes. Isto é feito pela digitalização de textos impressos com um scanner e um programa OCR ou pela conversão de documentos eletrônicos existentes de um formato diferente (PDF, HTML, etc.). Uma vez parcialmente anotados, obtêm-se arquivos de texto que podem ser marcados e anotados usando aplicativos apropriados. Um aplicativo grátis e popular, o *Tree Tagger,* é um etiquetador *POS* e um lematizador, adaptável a diferentes línguas com base em um léxico e em um corpus de treinamento manualmente etiquetado. Uma vez compilado, o corpus precisa ser gerenciado para ser submetido a vários tipos de análises e precisa-se de ferramentas de software para o gerenciamento de corpus e de alguns aplicativos para torná-lo disponível à comunidade científica, dentre outras finalidades.

Ferramentas com base em textos e índices oferecem acesso básico a corpora através de listas de palavras e buscas definidas pelo usuário e essas ferramentas se diferem em como elas permitem ao usuário formular dúvidas e exibir os resultados. Pacotes de softwa-

re também se diferenciam em como eles tratam o *input* (o tipo de buscas que eles permitem) e o *output* (como os resultados de busca são exibidos). Alguns programas têm funções de busca de dados limitadas, enquanto outros oferecem buscas flexíveis e rápidas.

Em relação à análise de dados de corpus, Zanettin menciona as listas de palavras e as concordâncias. Segundo o autor, as listas de palavras são "[…] *an index of word forms, […] sorted according to their alphabetical order or according to their frequency*" (p. 117); e as concordâncias são "[…] *an index of all tokens of a word type, together with their immediate linguistic context*" (p. 124). Diferentes aplicativos, conforme Zanettin, podem usar a informação estatística para computar tipos complexos de relações entre palavras, como colocações, coligações, grupos de palavras e perfis léxicos. As colocações, segundo Barnbrook (1996), são padrões de combinações de palavras em um texto e, para Firth (1957, p. 11), a colocação seria "*the company a word keeps*" (p. 130). Quanto à coligação, ela representa um nível maior de abstração quando relaciona formas de palavras simples ou grupos de palavras com outros grupos de palavras que compartilham a mesma classe de palavra. O autor examina as relações e as ferramentas necessárias para tornar as informações mais fáceis aos usuários identificarem padrões significativos na língua, mas é o analista que interpreta os dados.

Com os programas de análise de corpus, possivelmente, se descobre muito sobre as palavras no corpus e sua frequência. Muitos programas permitem ao usuário clicar em qualquer palavra na lista de palavras para recuperar as suas ocorrências em um corpus. Nesse capítulo, Zanettin apresenta ainda a noção de nuvens, perfis de palavras, preferência e prosódia semânticas.

Nas "Tarefas" (p. 141-146), há exemplos de investigação da colocação e da preferência semântica, com software de análise de corpus. Mostra-se ao leitor como criar, manipular e explorar listas de palavras e de lemas, palavras-chave e concordâncias usando o

corpus comparável bilíngue criado no capítulo 3 e uma versão de texto do *Open American National Corpus,* presente no DVD. Essas tarefas podem ser executadas com o software de análise de textos disponível no DVD ou com software comercial. Utilizam-se, nesse capítulo, ferramentas computacionais mais avançadas para relações léxico-gramaticais.

O capítulo 6 focaliza na criação e no uso de corpora paralelos bilíngues. Os diferentes estágios envolvidos na criação de corpora vão desde o design de corpus e aquisição até a codificação do corpus, anotação e indexação. Várias questões adicionais estão envolvidas nessa criação como a disponibilidade de textos a serem incluídos e o alinhamento de pares de textos em corpora paralelos.

Esse capítulo fornece uma pesquisa de procedimentos e ferramentas para o alinhamento de corpora paralelos no nível de parágrafos, da "sentença" e de palavras. Esse alinhamento é importante no processamento de corpora bilíngues paralelos. As correspondências acontecem entre textos, depois entre segmentos menores nos textos. O alinhamento envolve a segmentação paralela de pares de textos em unidades lógicas menores (parágrafos, sentenças, frases). Em relação aos pares de textos paralelos, eles podem ser alinhados com programas de alinhamento. Para bons resultados, conforme Zanettin, textos paralelos devem ser pré-processados, já que alguns programas de alinhamento presumem que textos-fonte e traduções têm o mesmo número de parágrafos e de frases.

Zanettin analisa fragmentos de mapas de alinhamento e recursos tecnológicos para fazer alinhamento como o *Champollion Tool Kit*, distribuído gratuitamente na linguagem de programação 'Perl'. O *ParaConc Aligner,* com base no Windows, é um concordanciador (ferramenta que permite acesso a informações linguísticas e textuais) paralelo independente, com base no algoritmo de *Church e Gale*. Uma outra ferramenta, dentre tantas outras, é o *Geometric Mapping and Alignment*, software gratuito, na linguagem de pro-

gramação 'Java', no ambiente Unix. Um corpus oferece evidência empírica que permite ao usuário discernir mais facilmente padrões de comportamento linguístico e textual.

Zanettin, com exemplos de jornais e da Web, analisa corpus paralelo e comparável e apresenta a coleção *OPUS* e o corpus *COMPARA*. A coleção *OPUS* é um estudo de caso de ferramentas e de procedimentos usados para construir uma versão alinhada de corpora paralelos. No corpus *COMPARA*, o alinhamento de parágrafo foi seguido por um alinhamento de sentença, realizado automaticamente usando um programa de alinhamento, checado manualmente e revisado com a ajuda de um processador de palavras. Cada frase no texto-fonte foi alinhada com um segmento correspondente na tradução, com uma sentença, mais de uma, ou somente parte de uma sentença.

Em relação a corpus paralelo e a corpora comparáveis na tradução, Zanetti menciona a memória de tradução, uma base de dados com texto-fonte emparelhado e segmentos de tradução, um tipo de corpus paralelo diferente dos outros pelo propósito de sua criação e uso e o formato no qual eles são adequadamente armazenados e recuperados. Uma TM é o principal recurso para tradutores profissionais que trabalham com textos técnicos.

As tarefas (p. 174-179) usam textos do DVD, já parcialmente processados em uma tarefa de capítulo anterior. Dois diferentes programas de alinhamento (no DVD) são usados para alinhar três pares de textos de extensões diferentes e de facilidades de processamento. O alinhamento desses textos envolve diferentes abordagens ao alinhamento automático e diferentes graus de interação entre o usuário e o aplicativo de alinhamento. Pode-se começar o processo com o rascunho, depois selecionar textos paralelos de sua preferência e realizar processamento preparatório básico, ou usar os arquivos no formato de texto sem formatação disponível no DVD, com atividades sobre isso. Outra tarefa consiste em revisar

o projeto de construção do corpus esboçado no capítulo 3 de acordo com as informações adquiridas nos capítulos seguintes e usar uma *checklist* do livro.

No capítulo 7, o autor nos apresenta as ferramentas e técnicas para usar corpora multilíngues com importante papel nos estudos aplicados ou descritivos da tradução. Por exemplo, o *European Comparable and Parallel Corpora*, planejado para dois corpora paralelos – um corpora dos discursos em Inglês no Parlamento Europeu e suas traduções para o Espanhol e um outro na direção oposta da tradução; e dois (sub)corpora monolíngues em duas línguas. O (sub)corpus monolíngue Inglês, com discursos políticos, será usado na combinação com o subcorpus de discursos originais do Parlamento Europeu em Inglês e com o subcorpus de discursos do Parlamento Europeu traduzido do Espanhol para o Inglês. O (sub)corpus monolíngue em Espanhol, com discursos do *Congreso de los Diputados*, pode ser usado em combinação com o subcorpus dos discursos originais do Parlamento Europeu em Espanhol e com o subcorpus dos discursos traduzidos do Parlamento Europeu do Inglês para o Espanhol.

Esse capítulo analisa a exposição e a análise de concordâncias paralelas. Zanettin, sobre linhas de concordância, mostra resultados de uma busca para a palavra 'Alice' em *Alice no País das Maravilhas*, dispostos de acordo com a ordem de triagem para a mesma palavra na tradução Italiana. E dois estudos de caso são apresentados, dependendo do nível de anotação e do software usado para recuperar e exibir concordâncias paralelas. Um desses estudos é o *Rushdie English-Italian Parallel Corpus* (algumas obras de Salman Rushdie e traduções para o Italiano) (ZANETTIN, 2001) (p. 190). Ao todo, o corpus tem traduções um pouco mais longas do que os textos-fonte. São adotadas metodologias para analisar o corpus paralelo dos romances de Rushdie e suas traduções, usando o *ParaConc*. O outro estudo de caso é o *OPUS Word Alignment Database*, com três corpora paralelos, *Europarl 3*, o corpus *OpenSubtitles* e o corpus

European Constitution. A base de dados pode ser consultada buscando por uma palavra em uma língua e recuperando seus 'equivalentes' em uma ou mais línguas, e buscando pelo resultado de alinhamento automático de palavras, classificadas de acordo com a frequência. Um dos exemplos ilustrados no capítulo é a análise contrastiva da palavra 'olho' em inglês e em italiano, com a interface de busca para o *OPUS Multilingual Word Alignment Database*.

Zanettin argumenta que corpora podem ajudar os tradutores a lidar com problemas de tradução de difícil solução. Ferramentas de corpora e de corpus, incluindo a Web, surgiram como ajudas práticas durante o processo de tradução, além dos dicionários e das enciclopédias e fornecem informações sobre diferentes aspectos. Enquanto os dicionários focalizam na palavra, os corpora focalizam acima do nível da palavra, com informações sobre colocação, terminologia e fraseologia.

Nas "Tarefas" (p. 202-205), encontram-se explorações práticas de dois diferentes corpora paralelos: o pequeno corpus paralelo literário Inglês-Italiano, criado no Capítulo 6, e o corpus paralelo multilíngue *Europarl,* com palavras de documentos do Parlamento Europeu em vinte línguas. Para buscar o corpus de textos literários, usa-se a versão demo do *ParaConc*, incluída no DVD, e para o corpus do *Europarl,* usa-se a interface de busca multilíngue *online* do *OPUS online*.

A conclusão ressalta o desenvolvimento das aplicações da tecnologia de computadores para a recuperação de informação linguística e textual em textos eletrônicos importantes para os estudos da tradução. Recomendações feitas nesse livro se referem a possíveis desenvolvimentos futuros de projetos com base em corpus nas pesquisas relacionadas à tradução.

Esse livro permite uma leitura que não exige conhecimento aprofundado de linguística de corpus. É importante destacar que, além

de discussões teóricas, ele oferece tarefas práticas e um DVD que levam os leitores (como tradutores e pesquisadores da tradução) a criarem, usarem e analisarem corpora com o auxílio dos computadores e da Internet.

Recebido em: 11/02/2017
Aceito em: 17/05/2017
Publicado em setembro de 2017