UMA BREVE REVISÃO SOBRE SISTEMAS WEB COM BASE EM *CORPUS* NO PAR LINGUÍSTICO INGLÊS-PORTUGUÊS[1] [2]

A BRIEF REVIEW OF CORPUS-BASED WEB SYSTEMS IN THE ENGLISH-PORTUGUESE LANGUAGE PAIR[3]

Rossana DA CUNHA SILVA[4]
Swansea University

Resumo: Com o advento da internet e dos constantes avanços tecnológicos, os *corpora* se tornaram essenciais no crescimento dos Estudos da Tradução Baseados em *Corpus* (ETBC), assim como no desenvolvimento de sistemas de informação e técnicas que fazem uso destes. Este artigo apresenta uma breve revisão de sistemas web baseados em *corpus* no par linguístico inglês-português, a partir de uma perspectiva de aplicação ao ensino, à pesquisa e à prática tradutória. Para tanto, buscamos proporcionar uma significação no âmbito tecnológico por meio de (i) uma breve contextualização teórica sobre o uso de *corpora*, (ii) as suas principais características e (iii) as aplicações mais conhecidas. Posteriormente, apresenta-se uma síntese das ferramentas web gratuitas: COMPARA (2000), CorTrad (2009), COPA-TRAD (2011), OPUS-CORPUS (2012) e VVV (2013). Em seguida, elencamos os usos e benefícios mais comuns de sistemas de compilação, análise, classificação e exploração de *corpora*. Por fim, a análise revela o momento vivenciado nos ETBC por meio de um resumo do aparato tecnológico existente na área. Desta maneira, almejamos que a presente discussão venha a proporcionar o desenvolvimento de pesquisas relacionadas aos sistemas baseados em *corpus*, haja vista a constante evolução tecnológica e a variedade de aplicações que podem se beneficiar do uso de *corpora*, seja no contexto prático ou profissional.
Palavras-chave: Tecnologia de tradução. Estudos da Tradução Baseados em Corpus. Corpora no ensino, pesquisa e prática tradutória.

Abstract: *With the advent of Internet and continuous technological advances, corpora have become essential in the growth of Corpus-Based Translation Studies (CTS), as well as in the development of information systems and techniques that make use of them. This paper presents a brief revision of corpus-based web systems in the English-Portuguese language pair, from a perspective of application in translation teaching, research and practice. To this end, we aim to provide a meaning in the technological scope through (i) a brief theoretical contextualization on the use of corpora, (ii) its key features and (iii) the best-known applications. Afterwards, a summary of the open-source web-based tools is presented: COMPARA (2000), CorTrad (2009), COPA-TRAD (2011), OPUS-CORPUS (2012) and VVV (2013). Next, we list the most common uses and benefits of systems for compiling, analyzing, classifying, and exploiting corpora. Finally, the analysis reveals the moment experienced by CTS through a synthesis of the technological apparatus in the area. To sum up, we aim to encourage the development of corpus-based systems research, due to the constant technological evolution and the variety of applications that can benefit from the use of corpora, either in the practical or professional context.*
Keywords: *Translation Technology. Corpus-based Translation Studies. Corpora in translation teaching, research, and practice.*

1. Introdução

Nos Estudos da Tradução, Palumbo (2009, p. 26) destaca que os primeiros indícios dos estudos baseados em *corpus*[5] de tradução surgiram na década de 1980, quando *corpora* foram utilizados pela primeira vez para descrever padrões encontrados em textos traduzidos em oposição aos textos originais. Laviosa (apud GAMBIER & DOORSLAER, 2010, p. 83) acrescenta que, mais precisamente em 1993, Mona Baker, em seu artigo intitulado "*Corpus linguistics and Translation Studies: Implications and applications*" (Linguística de *corpus* e Estudos da Tradução: implicações e aplicações)[6], percebeu a capacidade da linguística de *corpus*, sendo assim responsável por introduzir *corpora* nos estudos de tradução. Desde então, os *corpora* têm sido utilizados em pesquisas relacionadas aos estudos descritivos de tradução, na formação de tradutores, na avaliação da qualidade de tradução (*Translation Quality Assurance* – TQA) e em conjunto com ferramentas CAT[7] (BAKER & SALDANHA, 2008, p. 59; LAVIOSA apud GAMBIER & DOORSLAER, 2010, p. 83).

Os Estudos da Tradução Baseados em *Corpus* (ETBC) aproveitaram os avanços tecnológicos em recursos computacionais, bem como ferramentas de desenvolvimento e sistemas informatizados. Baker (1995, p. 224) enfatizou como "*[c]orpora* computadorizados foram se tornando cada vez mais populares nas áreas da disciplina, que têm vínculos estreitos com as ciências exatas" (minha tradução)[8]. Ademais, algumas das pesquisas realizadas no campo dos ETBC estão relacionadas com o estilo do tradutor, a ideologia da tradução, recursos de tradução, tradução forense, entre outros (WILLIAMS & CHESTERMAN, 2002).

Devido aos mais diversos enfoques na área dos ETBC, este artigo busca apresentar uma breve versão sobre alguns dos sistemas web gratuitos mais utilizados no par linguístico inglês-português, com o objetivo de proporcionar uma significação no âmbito tecnológico do uso de *corpora*, bem como suas aplicações mais conhecidas. Ao elencarmos características sobre ferramentas da área dos ETBC, almejamos contribuir para o uso e a disseminação de pesquisas relacionadas aos sistemas de compilação e análise de *corpora*.

Por esse motivo, dispõe-se primeiramente uma breve contextualização teórica; em seguida apresentam-se cinco dos sistemas web baseados em *corpus* no par inglês-português: COMPARA (2000), CorTrad (COMET, 2009), COPA-TRAD (2011), OPUS-CORPUS (2012) e VVV (2013); depois, são elencados os usos mais comuns de *corpora* nos Estudos da Tradução; e, por último, serão dispostas as conclusões e encaminhamentos futuros.

2. *Corpora* e suas tipologias: algumas distinções

Diferentes tipologias estão presentes nos ETBC, porém, no escopo da presente pesquisa, considera-se a sugerida por Fernandes (2006), após estudos sobre a tipologia proposta por Baker (1995). Em seu artigo, Fernandes ressalta a necessidade de haver um propósito que norteie a criação de um *corpus*, elencando sete critérios que deverão ser levados em consideração. São eles: (i) o tipo de relação existente entre os textos (comparável ou paralelo); (ii) a área de estudo (linguística ou tradução); (iii) o domínio (geral ou restrito); (iv) o modo (escrito ou falado – atualmente temos também o multimodal); (v) a restrição temporal (diacrônico ou sincrônico); (vi) o número de línguas (monolíngue, bilíngue ou multilíngue); (vii) a direcionalidade (unidirecional, bidirecional ou multidirecional).

Segundo Baker (1995, p. 234), um *corpus* comparável consiste em duas coleções separadas de textos na mesma língua: um *corpus* composto por textos originais na língua A e outro de traduções na mesma língua A (por exemplo, traduções originadas de uma determinada língua B ou C). O *Translational English Corpus* (TEC) e o *British National Corpus* (BNC) são exemplos de *corpora* comparáveis. Ainda segundo a autora (ibid.), o termo *corpus* paralelo se refere a textos-fontes na língua A e suas versões traduzidas na língua B. São exemplos de *corpora* paralelos: EUROPARL (*European Parliament Proceedings Parallel Corpus*), LDC (*Linguistic Data Consortium*), que disponibiliza um grande número de *corpora* paralelos e, por último, ELRA (*European Language Resources Association*), dentre outros.

Com relação à área de estudo, pode-se dividir os *corpora* em projetados para o estudo da língua e desenvolvidos para a investigação de produtos e processos relacionados à tradução. Fernandes acrescenta que, embora os ETBC estejam mais preocupados com o segundo tipo, vários estudiosos utilizam os *corpora* linguísticos na formação de tradutores, como forma de desenvolver a competência linguística de tradutores aprendizes.

O domínio diferencia *corpora* entre geral, que é compilado de maneira equilibrada com amostras da língua a partir de uma grande variedade de registros e gêneros (FERNANDES, 2006), entre eles o BNC e o *Corpus of Contemporary American English* (COCA); e restrito, por exemplo: o *International Corpus of English* (ICE) e o TOEFL11 (*A Corpus of Non-Native English*). Em relação ao modo, temos os *corpora* escritos, falados ou multimodais. Os multimodais possuem imagem, som, escrita (em legendas ou imagens) (FERNANDES, 2006), tais como: *Augmented Multi-party Interaction Corpus* (AIM), *SmartKom Corpus* e *HuComTech Corpus*.

No que se refere à restrição temporal, segundo Atkins et al. (1992 apud FERNANDES, 2006, p. 93), um *corpus* sincrônico tem como característica principal um ponto em particular como objeto de estudo, ou seja, um retrato do uso da língua durante um período de tempo limitado. Já o diacrônico leva em consideração o desenvolvimento histórico relacionado à investigação em andamento (exemplos: *Helsinki Dialect Corpus, Corpus of English Dialogues, Lancaster-Oslo-Bergen corpus* – LOB).

Um *corpus* multilíngue, de acordo com Baker (1995, p. 232), consiste em um "conjunto de dois ou mais *corpora* monolíngues em diferentes línguas, construídos nas mesmas ou em diferentes instituições, e utilizando o mesmo critério de seleção de textos"[9]; como exemplos, temos *Oslo Multilingual Corpus* (dentro do *English-Norwegian Parallel Corpus* – ENPC); e monolíngues: BNC, CorTec (Corpus Técnico Científico).

Finalmente, podemos caracterizar os *corpora* pela sua direcionalidade. Considera-se um *corpus* unidirecional quando temos, por exemplo, textos originalmente escritos em uma língua A e suas respectivas traduções em uma língua B, sendo que isto ocorre apenas em uma direção, de A para B. O *corpus* bidirecional possui textos originais na língua A e suas respectivas traduções na língua B, além de textos escritos originalmente na língua B e suas traduções na língua A. Por último, os *corpora* multidirecionais ocorrem quando existem mais de duas línguas e suas traduções não estão centradas apenas na língua A, mas em todas as línguas presentes no *corpus* (FERNANDES, 2006).

3. Características, usos e benefícios de sistemas para análise de *corpora*

As primeiras ferramentas que surgiram para análise de *corpus* não foram criadas especificamente para a área dos ETBC, e sim destinadas a professores de línguas estrangeiras ou lexicógrafos. Segundo McNery & Hardie (2012), a primeira ferramenta criada para análise de *corpus* foi construída por Roberto Busa em 1951. Os autores (ibid.) acrescentam que apesar de Busa não ter inventado o "concordanciador" (vide abaixo), ele mostrou que a "concordância poderia ser aplicada de forma rápida e efetiva em textos eletrônicos" [10] (MCNERY & HARDIE, 2012, p. 37). No geral, os sistemas de tradução já utilizam *corpus* como recurso integrante, mas os sistemas e ferramentas para análise de *corpora* exploram o uso de *corpus* com maior enfoque e de modo mais analítico e, para isso, necessitam maior interação por parte do tradutor.

Corpas-Pastor (2012, p. 77-76) destaca que o *corpus* é um recurso linguístico que tem sido responsável pelo desenvolvimento de técnicas e ferramentas de tratamento (compilação,

análise, classificação e exploração). À luz do exposto acima, consideramos neste artigo o termo "sistema de tradução baseados em *corpus*" como sinônimo de "sistema de compilação e análise de *corpus*" ou apenas "sistema de análise de *corpus*", visto que se refere ao perfil dos sistemas apresentados na seção 3, pois englobam diversas funcionalidades, não apenas o uso, mas a compilação, análise, classificação e exploração de *corpora*.

A maioria dessas ferramentas permite a manipulação e análise de *corpora*, apresentando informações úteis aos seus usuários. Conforme Kenny (2011, p. 3), uma das ferramentas mais conhecidas para o processamento de *corpus* é o concordanciador, que possibilita aos usuários realizar uma pesquisa por todas as instâncias de uma determinada palavra ou frase dentro de um *corpus*. Alguns sistemas deste tipo possuem recursos auxiliares, como a apresentação de resultados da busca no formato KWIC[11], com a exibição de uma palavra-chave em contexto; ou a criação e manipulação de listas de palavras (i.e., tipos/*types* – palavras únicas que ocorrem em um *corpus*, junto com a indicação de frequência de ocorrência – *token*) (KENNY, 2011; BOWKER & PEARSON, 2002). De acordo com as autoras (2002, p. 120), os concordanciadores bilíngues são usados em *corpora* alinhados e permitem a pesquisa por um termo em determinado idioma, com a apresentação de todas as ocorrências de uma determinada palavra em contexto (com suas colocações), classificadas por ordem alfabética ou frequência.

Concordamos com o fato de o uso das ferramentas de compilação e análise de *corpus* terem se tornado populares aos interessados na área por proporcionarem acesso a informações, tal como a frequência de uma palavra em uma variedade de contextos, ou mesmo oferecendo a possibilidade de observar diferentes variáveis, como o uso de determinado registro, gênero etc., o que não poderia ser imaginado em textos impressos. O uso de *corpora* pode oferecer diversas vantagens, mas isso não significa que deva ser considerado solução para qualquer tipo de problema. De qualquer forma, sabe-se que seu uso pode ser um recurso valioso e/ou complemento útil para outros tipos, como dicionários, textos impressos, peritos em determinados assuntos etc. (BOWKER & PEARSON, 2002).

Uma observação feita por Tiedemann (2011, p. 27) está relacionada ao fato de a internet ter disponibilizado uma quantidade significativa de documentos com suas versões traduzidas, o que tornou possível a recuperação automática de *corpora*. O autor (ibid.) acrescenta que uma das formas conhecidas de se coletar informações da internet é chamada de *web mining* (mineração da web) e tem como exemplo de aplicação a arquitetura denominada STRAND (*Structural Translation Recognition Acquiring Natural Data*)[12], desenvolvida por

Resnik (1999), e que se baseia na observação da padronização de *sites* com suas versões traduzidas. Outros exemplos desse tipo de tecnologia são *Bitextor*[13] e ILSP *Focused Crawler*[14].

Outra característica interessante sobre as ferramentas baseadas em *corpus* está relacionada à sua disponibilidade, pois a maioria delas são conhecidas como *off-the-shelf software*[15], como: *WordSmith Tools, AntConc, Multiconcord, Paraconc.* Outras ferramentas tiveram sua disponibilização na internet, como o *WMatrix*[16], que é uma ferramenta para análise e comparação de *corpus*, e provê uma interface web para as ferramentas de anotação de *corpus* USAS (UCREL *Semantic Analysis System*)[17], CLAWS (*the Constituent Likelihood Automatic Word-tagging System*)[18], e as que serão apresentadas nos próximos subitens. Além das ferramentas que utilizam uma interface web (aplicações *Front-end*), outras fazem uso de *APIs* (*Application Program Interface*)[19], que acessam os *corpora on-line* e podem ser utilizadas por meio de aplicativos em computadores portáteis, ou mesmo em dispositivos móveis (OLOHAN, 2004; BAKER & SALDANHA, 2008). A seguir, serão apresentadas cinco ferramentas de análise de *corpus* no par linguístico inglês-português que estão disponíveis *on-line* e gratuitamente à comunidade em geral.

3.1. COMPARA

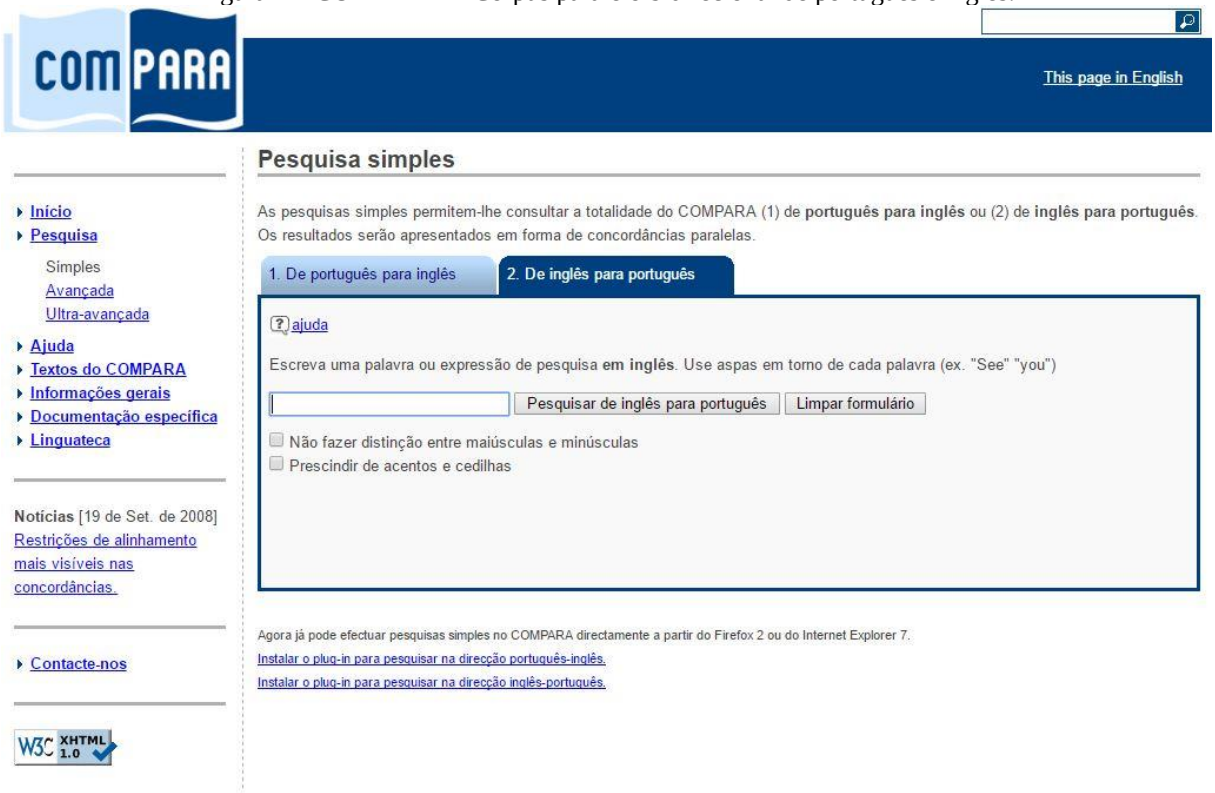
O primeiro programa a ser mencionado é o COMPARA, que é uma "ferramenta que permite estudar a tradução humana e contrastar o português e o inglês através de pesquisas automáticas" (FRANKENBERG-GARCIA; SANTOS, 2003, p. 71). Segundo as autoras, há três características principais:

- aberto, ou seja, que pode crescer em qualquer direção que se mostre importante aos seus usuários, e cujos textos adicionados ao *corpus* possam ser disponibilizados após o término de seus processamentos;
- para pessoas que não são familiarizadas com *corpus*, bem como usuários experientes em *corpus*;
- pesquisável por meio da internet.

O projeto COMPARA teve início em 1999 dentro do Linguateca (financiado pelo governo português desde maio de 2000) e utiliza o sistema DISPARA (DIStribuição de *corpora* PARAlelos na web), desenvolvido como uma ponte para o *IMS Corpus Workbench*.

O *corpus* também está configurado para a possibilidade de um texto fonte ser alinhado com mais de uma tradução. Os textos selecionados para o *corpus* consideraram variações dialetais e linguísticas de português e inglês (por exemplo, português do Brasil e de Portugal; inglês dos Estados Unidos e Reino Unido), bem como textos contemporâneos e não contemporâneos, com a devida autorização no que se refere aos direitos autorais. Há dois tipos de pesquisas disponíveis na interface DISPARA: pesquisa simples, para pessoas que nunca utilizaram ferramentas baseadas em corpus; e pesquisa avançada, com mais opções de filtros e com a possibilidade de realizar consultas mais sofisticadas (FRANKENBERG-GARCIA & SANTOS, 2003, p. 80–81).

Figura 1 – COMPARA – Corpus paralelo bidirecional de português e inglês.



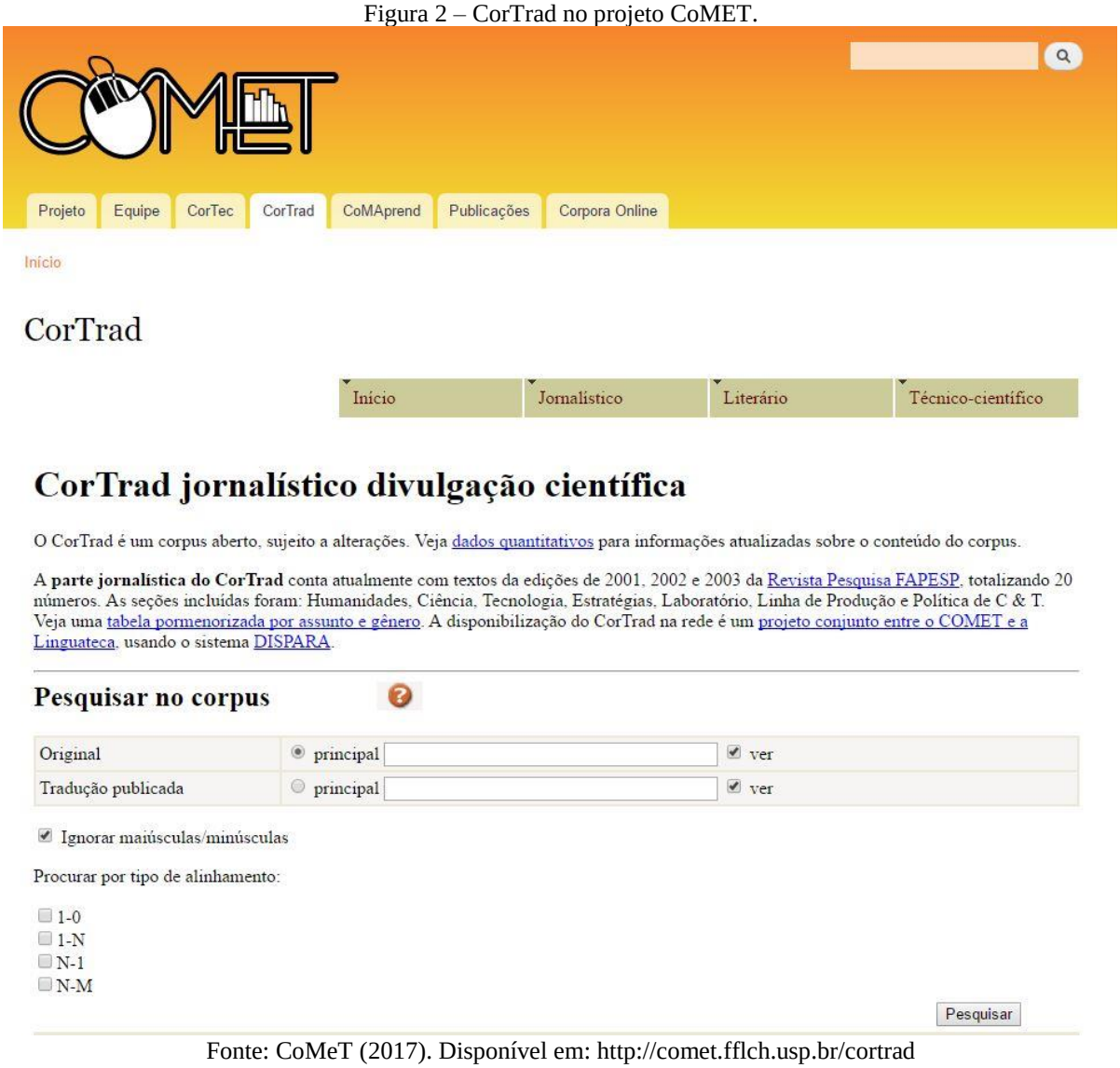
Fonte: COMPARA (2017). Disponível em: http://www.linguateca.pt/COMPARA/psimples.php?language=pt

3.2. CorTrad

O segundo sistema denomina-se CorTrad (*Corpus* Paralelo de Tradução), imerso no projeto CoMET – *Corpus* Multilíngue para Ensino e Tradução (2009), e foi desenvolvido dentro da Universidade de São Paulo (USP). O projeto CoMET é composto de três *corpora* (TAGNIN; TEIXEIRA; SANTOS, 2009):

- CorTrad (*Corpora* de tradução, dividido em três *subcorpora*: CorTrad jornalístico, CorTrad literário e CorTrad técnico científico);
- CoMAprend (*Corpus* de aprendizes)[20];
- CorTec (*Corpus* técnico composto por diversas áreas de conhecimento, como Ciência da Computação, linguística etc.).

Figura 2 – CorTrad no projeto CoMET.



Fonte: CoMeT (2017). Disponível em: http://comet.fflch.usp.br/cortrad

O CorTrad foi resultado da parceria entre o CoMET, a Linguateca e o NILC (Núcleo Interinstitucional de Linguística Computacional, localizado no ICMC – Instituto de Ciências Matemáticas e da Computação da USP de São Carlos), e teve seu início em maio de 2008. As pesquisas são feitas utilizando o sistema DISPARA, semelhante ao COMPARA, permitindo a

utilização de uma interface customizável. Dentre as ferramentas disponíveis, é possível executar pesquisas dentre os *subcorpora* existentes com:

- a possibilidade de se compararem diferentes versões de um mesmo texto (original, versões revisadas e tradução publicada);
- a utilização de mecanismos de busca para cada gênero pesquisado – permitindo, por exemplo, pesquisar seções específicas de diferentes tipos textuais;
- a possibilidade de testar e relacionar hipóteses, através de um refinamento da pesquisa. Isso porque os textos do *corpus* são classificados (etiquetados) morfossintaticamente (COMET, 2009).

3.3. COPA-TRAD (*Corpus* paralelo de tradução)

O terceiro sistema de compilação e análise de *corpus* a ser mencionado é o COPA-TRAD – *Corpus* paralelo de tradução (FERNANDES & SILVA, 2011), desenvolvido no âmbito da Universidade Federal de Santa Catarina (UFSC) a partir de reuniões do grupo de pesquisa TraCor (Tradução e *Corpora*). Segundo os autores (ibid.):

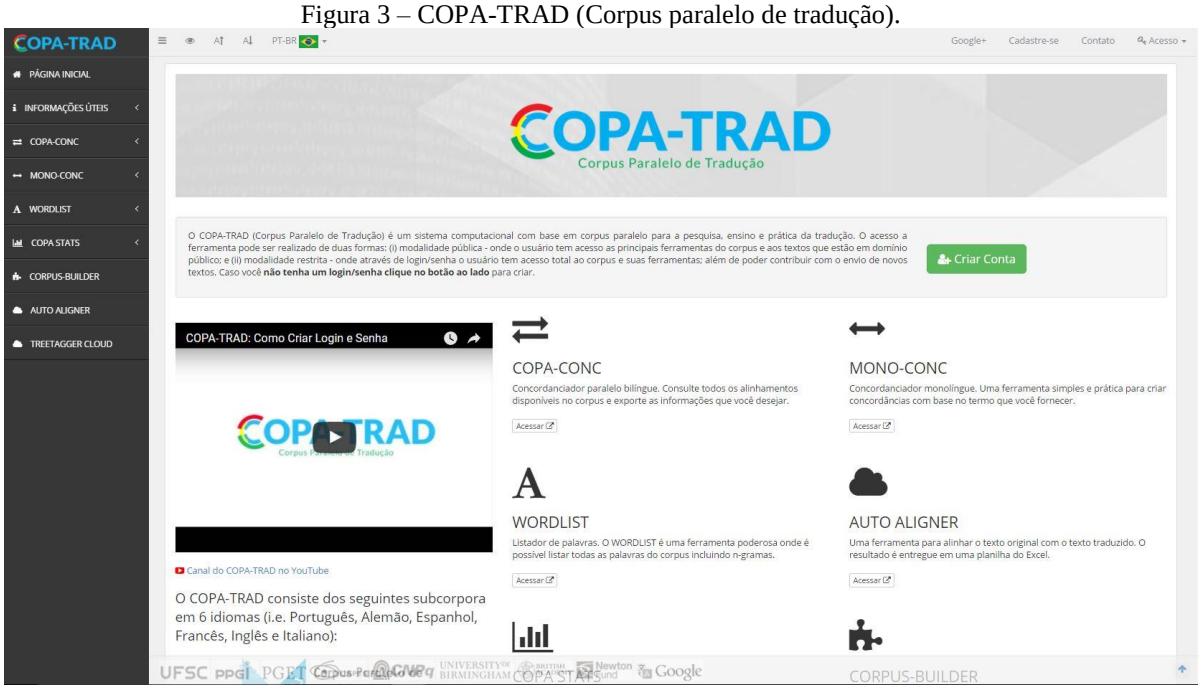
> O COPA-TRAD é um sistema computacional com base em *corpus* paralelo (i.e. um conjunto textos em L1 e sua(s) respectivas traduções em L2). O objetivo principal do COPA-TRAD é oferecer ferramentas computacionais disponíveis *on-line* para a pesquisa, ensino e prática da tradução. Essas ferramentas permitem que o usuário identifique, por exemplo, práticas tradutórias relacionadas a padrões linguísticos específicos dos tipos de texto que constituem o COPA-TRAD. Além disso, dados estatísticos sobre os textos e a possibilidade de criação de *Do-It-Yourself corpora* são outras das ferramentas oferecidas pelo sistema (grifo nosso).

O COPA-TRAD é composto por cinco *subcorpora* (FERNANDES & SILVA, 2011):

- COPA-LIJ, que consiste em textos clássicos de Literatura Infanto-juvenil dos mais variados subgêneros (por exemplo, fantasia, aventura, ficção científica, etc.);
- COPA-TEL, que consiste em textos clássicos da literatura que já se encontram no domínio público (por exemplo, romances, contos, poemas, etc.);
- COPA-RAC, que consiste em resumos de trabalhos acadêmicos das mais variadas áreas de conhecimento;
- COPA-MDT, que consiste em textos teóricos sobre tradução com o objetivo de se investigar questões relacionadas ao meta-discurso da tradução;

- Por último, o COPA-TEJ, que consiste em textos jurídicos como, por exemplo, certidões de nascimento, casamento e óbito; contratos; testamentos; entre outros.

Figura 3 – COPA-TRAD (Corpus paralelo de tradução).

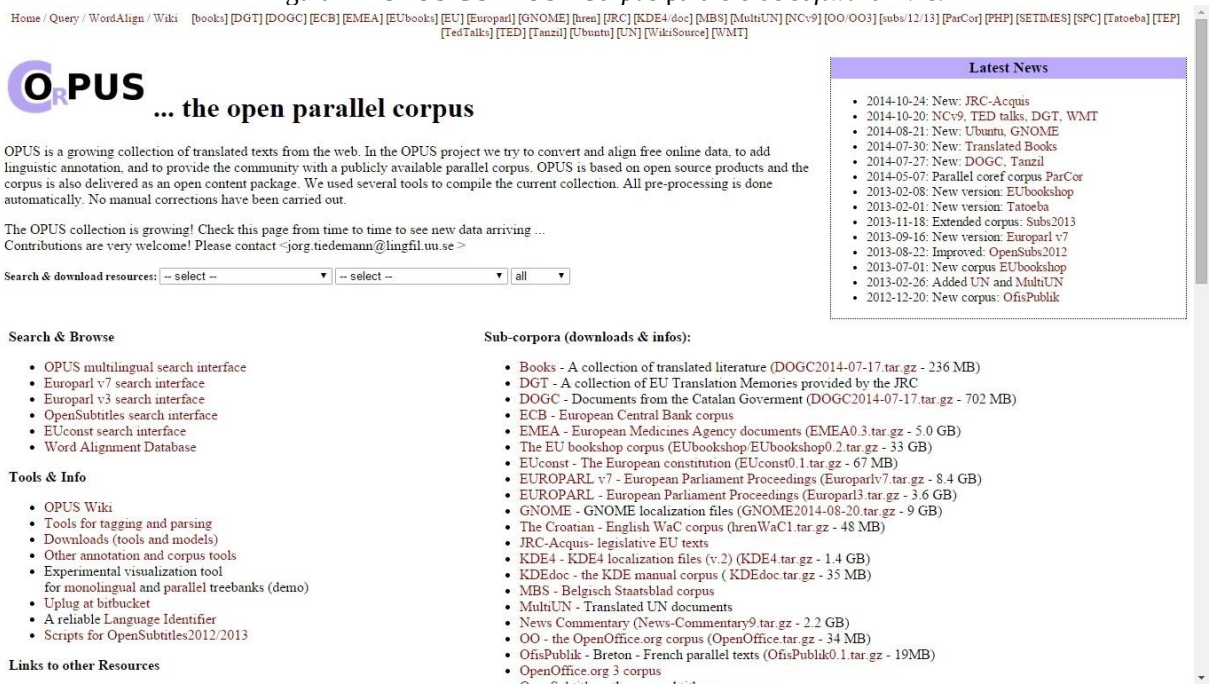


Fonte: COPA-TRAD (2017). Disponível em: https://copa-trad.ufsc.br/

3.4. OPUS-CORPUS

O quarto sistema de análise de *corpus* é o OPUS-CORPUS – *Corpus* paralelo de *software* livre (2012), que é um recurso de *corpora* paralelos e ferramentas relacionadas (por exemplo, ferramentas de busca e concordanciadores). O principal objetivo do OPUS é proporcionar dados que estejam disponíveis gratuitamente pela internet. O *corpus* abrange mais de noventa idiomas e agrupa vários domínios, como textos da área legislativa e administrativa (na sua maioria provenientes da União Europeia e instituições associadas), legendas de filmes traduzidos e localização de projetos de *software* livre (OPUS CORPUS, 2012).

Figura 4 – OPUS CORPUS - *Corpus* paralelo de *software* livre.

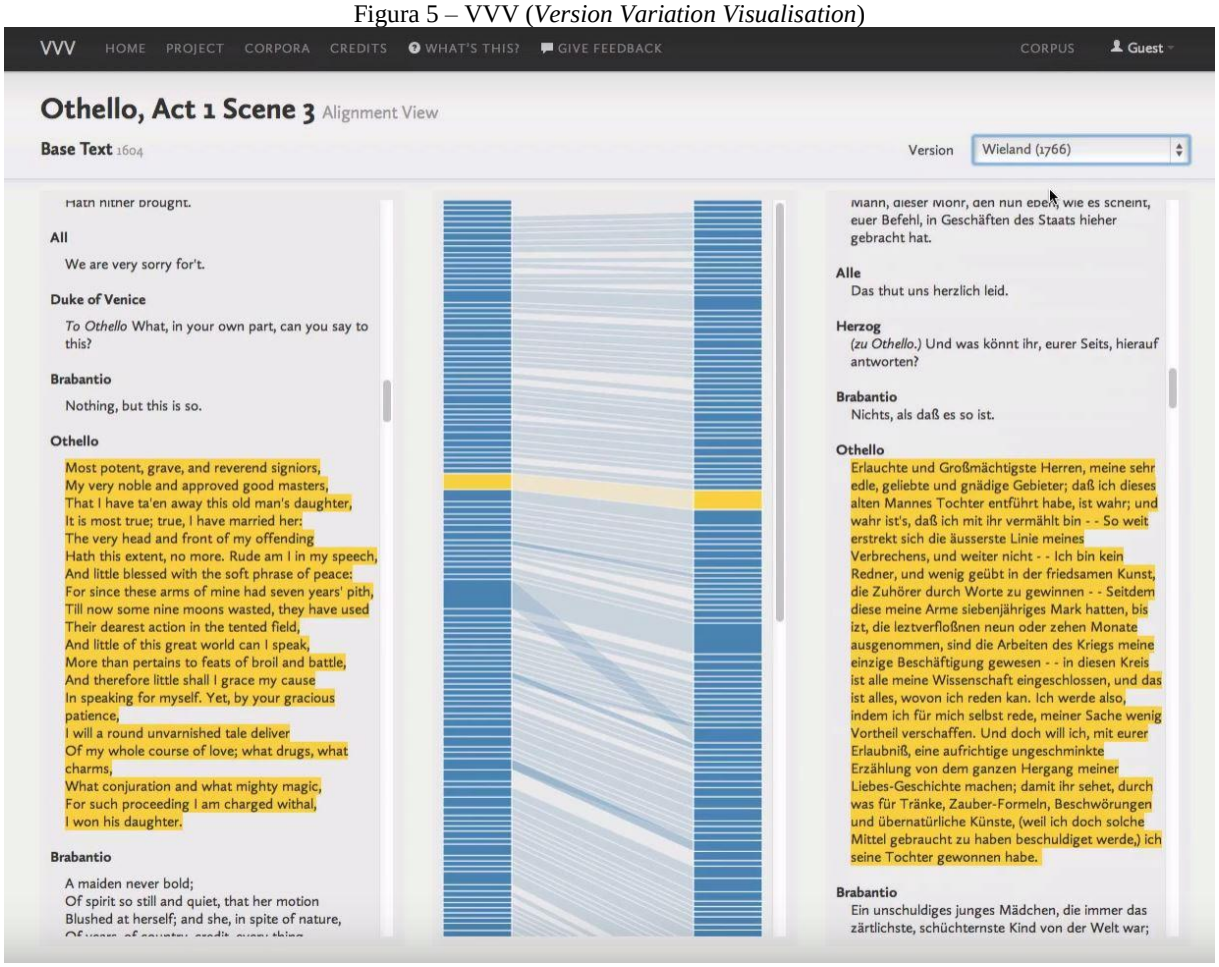


Fonte: OPUS-CORPUS (2017). Disponível: http://opus.lingfil.uu.se/

3.5. *Version Variation Visualisation* (VVV)

O quinto e último sistema baseado em *corpus* denomina-se VVV – *Version Variation Visualisation*[21] (CHEESMAN et al., 2013). Conforme os autores, o VVV é um conjunto de ferramentas digitais desenvolvido com o propósito de auxiliar as pessoas a explorar, comparar e analisar qualquer número de versões múltiplas de qualquer obra (CHEESMAN et al., 2016). O sistema de análise de *corpus* surgiu a partir do projeto *TransVis* desenvolvido no âmbito da Swansea *University*. O VVV permite aos usuários criarem *corpora* paralelos com múltiplas versões alinhadas por segmento de texto. A ferramenta web de análise de *corpus* possibilita analisar múltiplas versões a partir de interfaces visuais, tais como: (i) mapa de alinhamento; (ii) visualização paralela (conforme apresentado na Figura 5); (iii) estilometria; (iv) interface "Eddy & Viv"[22]; e (iii) mapa cronológico (CHEESMAN et al., 2016).

Figura 5 – VVV (*Version Variation Visualisation*)



Fonte: VVV (2017). Disponível em: www.delightedbeauty.org/vvv/

4. A influência e uso de *corpora* no ensino da tradução

Como vimos anteriormente, existem diversas ferramentas que estão disponíveis de forma gratuita e que podem ser utilizadas em diferentes abordagens. Uma delas é a utilização de *corpora* no ensino de tradução e na formação de tradutores, auxiliando alunos a encontrar soluções para os problemas que são pertinentes a uma tradução em específico. Segundo Beeby et al. (2009, p. 10), os *corpora* digitalizados oferecem grande vantagem sobre os físicos, já que podem proporcionar ao tradutor uma vasta quantidade de dados linguísticos sobre uma terminologia específica, apenas com o toque de um botão. Há um número considerável de contribuições exemplificando o uso dos *corpora* paralelos no ensino da tradução e na formação de tradutores. Existem os trabalhos de Meyer et al. (2000) sobre o uso de *corpus* paralelo em terminologia; os estudos de Barlow (2000), sobre colocações e padrões[23]; Bowker & Pearson (2002) fornecem um guia para auxiliar desde o aluno até o

próprio formador de tradutores; Frankenberg-Garcia & Santos (2003) apresentam exercícios por meio da busca no *corpus* paralelo COMPARA, exemplificando estratégias adotadas por tradutores profissionais. Olohan (2004) apresenta também vários trabalhos sobre o assunto; Rodríguez-Inés (2010) explora o uso de *corpora* como parte integrante da aquisição de competência tradutória; e Zanettin & Bernardini (2003) oferece uma coleção de artigos sobre a formação de tradutores.

5. A aplicação de *corpora* na pesquisa de tradução

Os corpora têm sido utilizados em pesquisas sobre equivalências em tradução, terminografia bilíngue e lexicografia, que visam a fornecer dados empíricos aos sistemas de tradução automática (BOWKER & PEARSON, 2002). Outros exemplos de pesquisas são citados por Gambier & Doorslaer (2010), tais como: estilo de traduções; a influência do inglês nas línguas europeias por intermédio da tradução; fornecimento de hipóteses sobre futuras investigações; características linguísticas de dublagem; e artigos sobre características de traduções (explicitação, simplificação, normalização ou interferência)[24], entre outros. Fernandes (2004, 2006, 2009) contribui com diversas formas de pesquisa na área de tradução: além de seus artigos sobre o tema, apresenta o COPA-TRAD como um sistema que resume parte de sua pesquisa. Olohan (2004) também fornece várias formas de pesquisa, tais como estilo e padrões encontrados em tradução.

6. A prática tradutória e o uso de *corpora*

O tradutor técnico ou especializado pode se beneficiar do uso de *corpora* paralelos, em vista da possibilidade de pesquisas terminológicas, ou mesmo encontrar diferentes estratégias de tradução. Seu uso pode também auxiliar na investigação de artifícios e características estilísticas de autores e outros tradutores (OLOHAN, 2004). A utilização de *corpora* também se faz na avaliação da qualidade de tradução (TQA), tomando como base alguns critérios para tentar determinar ou estabelecer até que ponto uma tradução pode ser considerada melhor do que outra (GAMBIER & DOORSLAER, 2010, p. 85). Em seu livro, Bowker & Pearson (2002) apresentam as vantagens de utilização de *corpora* paralelos no auxílio da escrita técnica. Enquanto isso, Bernardini et al. (2003) exploram os benefícios dos *corpora* tanto na formação como na prática tradutória, apresentando vantagens do uso de *corpus* quando em combinação com outros recursos, tais como dicionários, memórias de tradução, ou ferramentas CAT em geral. Varantola (apud ZANETTIN & BERNARDINI, 2003) apresenta o uso do *corpus* "descartável" (*Disposable corpus*)[25], bem como suas preocupações sobre a

competência necessária para a sua utilização, visto que o tradutor deverá saber como compilar um *corpus* e utilizá-lo de forma que o auxilie em suas traduções.

7. Outras considerações e aplicações sobre *corpora*

Cada vez mais, *corpora* se tornam essenciais no crescimento dos ETBC, assim como no desenvolvimento de sistemas e ferramentas que fazem uso destes. Com a crescente utilização de *corpora*, podemos citar outros exemplos de seu uso na área de tradução: *corpora* multimídia de filmes originais e dublados; a lexicografia computacional, que se beneficiou enormemente do uso de *corpora*; a usabilidade de ferramentas de tradução, tais como programa de pós-edição, para citar apenas alguns. A criação de *corpora* também é beneficiada pelo emprego de *crowdsourcing translation* (tradução feita por diversos grupos de pessoas sem limitação geográfica). Segundo Snel et al. (2012, p. 1), *crowdsourcing* vem a caracterizar os *corpora* por meio de identificação de padrões nos textos com diferentes usos, como na tradução automática, na visão computacional e na análise de sentimento[26].

Com a expansão da internet, também foi possível haver um desenvolvimento maior nos sistemas de tradução automática, devido à extração de informações da web, produzindo assim grandes *corpora on-line*. A computação em nuvem permitiu o armazenamento ilimitado de dados, bem como a execução de aplicativos por meio de aplicações web ou móveis, aumentando a gama de ferramentas e seu acesso por diferentes usuários.

8. Conclusões e encaminhamentos

Assim, vimos ao longo do artigo os principais tópicos abordados: Estudos da Tradução Baseados em *Corpus*; *corpora*, suas tipologias e principais distinções; sistemas de tradução baseados em *corpus*; cinco sistemas web gratuitos de análise de *corpus* para o par linguístico inglês-português e, por último, exemplos de como utilizar *corpora* em diferentes segmentos.

Buscou-se estabelecer uma analogia entre os sistemas e tecnologias existentes nos ETBC e os diferentes usos e abordagens na área de ensino, prática e pesquisa de tradução. Desta maneira, representou-se o momento vivenciado por meio de um resumo do aparato tecnológico existente na área.

Embora a análise tenha mostrado algumas das visões atuais da tecnologia de tradução com base em *corpus*, é interessante que outros estudos sejam conduzidos, visto a constante evolução do tema e a variedade de aplicações que podem se beneficiar do uso de *corpora*, seja no contexto prático ou profissional. Como trabalho futuro, pretende-se conduzir um

experimento mais específico com as tecnologias existentes na área e aplicar a pesquisa de forma a levantar informações sob a perspectiva do usuário e como esses (pesquisadores, estudantes, profissionais de tradução) interagem com essas tecnologias.

REFERÊNCIAS BIBLIOGRÁFICAS

BAKER, Mona. **In Other Words**. A coursebook on translation. Londres/Nova Iorque: Routledge, 1992.

_____. Corpus Linguistics and Translation Studies: Implications and Applications. In: **Text and Technology**: In Honour of John Sinclair. Amsterdam/Philadelphia: John Benjamins Publishing Company, 1993, p. 233–250.

_____. Corpora in Translation Studies. An Overview and Suggestions for Future Research. **Target,** Amsterdam, v.7, n.2, p. 223–243, 1995.

BAKER, Mona; SALDANHA, Gabriela. **Routledge Encyclopedia of Translation Studies**. 2ª ed. Londres/Nova Iorque: Routledge, 2008.

BARLOW, Michael. **Parallel texts in language teaching**. Multilingual corpora in teaching and research. Amsterdam: Rodopi, 2000, p. 106–115.

BEEBY, Allison; RODRÍGUEZ-INÉS, Patricia; SÁNCHEZ-GIJÓN, Pilar. **Corpus Use and Translating**: Corpus use for learning to translate and learning corpus use to translate. Amsterdam/Philadelphia: John Benjamins Publishing Company, 2009.

BERNARDINI, Silvia; STEWART, Dominic Stewart; ZANETTIN, Federico. Corpora in translator education: An introduction. In: ZANETTIN, Federico; BERNARDINI, Silvia; STEWART, Dominic. **Corpora in Translator Education**. Manchester: St Jerome, 2003.

BOWKER, Lynne; PEARSON, Jennifer. **Working with specialized language**: a practical guide to using corpora. Londres: Routledge, 2002.

CHEESMAN, Tom; FLANAGAN, Kevin; THIEL, Stephan. VVV. **Version Variation Visualisation**: Project Overview. 2013. Disponível em: <www.delightedbeauty.org/vvv>. Acesso em: 01 mai. 2017.

CHEESMAN, Tom; FLANAGAN, Kevin; THIEL, Stephan; RYBICKI, Jan; LARAMEE, Robert S.; HOPE, Jonathan; ROOS, Avraham. **Multi-retranslation corpora**: visibility, variation, value, and virtue. Digital Scholarship in the Humanities. Swansea, 2016.

COMET. Disponível em: <http://comet.fflch.usp.br/projeto>. Acesso em: 3 abr. 2017.

COMPARA. 2000. Disponível em: <http://www.linguateca.pt/COMPARA/Welcome.html>. Acesso em: 03 abr. 2017.

CORTEC. 2009. Disponível em: <http://comet.fflch.usp.br/O%20que%20e>. Acesso em: 03 abr. 2017.

CORPAS-PASTOR, Gloria. **Corpus, Tecnología y Traducción**. In: XII Jornadas de Lingüística. Cádiz: Servicio de Publicaciones de la Universidad de Cádiz, 2012, p. 75–98.

FERNANDES, Lincoln P. **Brazilian Practices of Translating Names in Children's Fantasy Literature**: A Corpus-based Study. 2004. 189f. Tese (Doutorado em Língua Inglesa e Linguística Aplicada – Tradução) – Curso de Pós-Graduação em Estudos Literários e Inglês, Universidade Federal de Santa Catarina, Florianópolis, 2004.

_____. Corpora in Translation Studies: revisiting Baker's typology. **Fragmentos**, Florianópolis, v.1, n.30, p. 87–95, 2006.

_____. A Portal into the Unknown: Designing, Building, and Processing a Parallel Corpus. **Ctis Occasional Papers**, v. 4. Manchester, Reino Unido, p. 21–43, 2009.

FERNANDES, Lincoln P.; SILVA, Carlos E. **COPA-TRAD** (Corpus Paralelo de Tradução). Universidade Federal de Santa Catarina, Florianópolis, 2011. Disponível em: <http://copa-trad.ufsc.br>. Acesso em: 11 abr. 2017.

FRANKENBERG-GARCIA, Ana; SANTOS, Diana. Introducing COMPARA: The Portuguese-English Parallel Corpus. In: ZANETTIN, Federico; BERNARDINI, Silvia; STEWART, Dominic. **Corpora in Translator Education**. Manchester: St. Jerome, 2003, p. 71–87.

GAMBIER, Yves; DOORSLAER, Luc van. **Handbook of translation studies**. John Benjamins B.V., Amsterdam, 2010.

KENNY, Dorothy. Electronic tools and resources for translators. In: MALMKJÆR, Kirsten; WINDLE, Kevin. **The Oxford handbook of translation studies**. Oxford: Oxford University Press, 2011.

LAN, Lin. Corpus. In: CHAN, Sin-Wai. **Routledge encyclopedia of translation technology**. Nova Iorque: Routledge, 2015, p. 465–479.

MCNERY, Tony; HARDIE, Andrew. **Corpus linguistics**: Method, theory and practice. Nova Iorque: Cambridge University Press, 2012.

MEYER, Renée; OKUROWSKI, Mary Ellen; HAND, Thérèse. **Using authentic corpora and language tools for adult-centered learning.** Multilingual Corpora in Teaching and Research. Amsterdam: Hodopi, 2000, p. 86–91.

OLOHAN, Maeve. **Introducing Corpora in Translation Studies**. Londres/Nova Iorque: Routledge, 2004.

OPUS CORPUS. Disponível em: <http://opus.lingfil.uu.se/>. Acesso em: 11 abr. 2017.

PALUMBO, Giuseppe. **Key Terms in Translation Studies**. Nova Iorque: Continuum Publishing, 2009.

RESNIK, Philip. **Mining the web for bilingual text**. Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics. Maryland: Association for Computational Linguistics, 1999, p. 527–534.

RODRÍGUEZ-INÉS, Patricia. Electronic Corpora and Other Information and Communication Technology Tools: An Integrated Approach to Translation Teaching. **The Interpreter and Translator Trainer**, v.4, n.2, 2010, p. 251–282. Disponível em: <http://dx.doi.org/10.1080/13556509.2010.10798806>. Acesso em: 01 fev. 2017.

SNEL, John; TARASOV, Alexey; CULLEN, Charlie; DELANY, Sarah Jane. **A Crowdsourcing Approach to Labelling a Mood Induced Speech Corpus**. 4th International Workshop on Corpora for Research on Emotion Sentiment & Social Signals (ES³ 2012). Istambul, Turquia, 2012.

TAGNIN, Stella E. O.; TEIXEIRA, E. D.; SANTOS, D. *CorTrad*: a multiversion translation corpus for the Portuguese-English pair. Arena Romanistica. In: **The 28th International Conference on lexis and grammar**. 2009. Bergen, Noruega, 2009, 314–323. Disponível em: <http://www.linguateca.pt/Diana/download/Tagnin-Teixeira-Santos_final.pdf>. Acesso em: 03 fev. 2017.

TIEDEMANN, Jörg. **Bitext alignment**. Synthesis Lectures on Human Language Technologies 4.2. San Rafael: Morgan & Claypool, 2011.

WILLIAMS, Jenny; CHESTERMAN, Andrew. **The Map**: a Beginner's Guide to Doing Research in Translation Studies. Manchester, UK: St. Jerome, 2002.

ZANETTIN, Federico; BERNARDINI, Silvia; STEWART, Dominic. **Corpora in Translator Education**. Manchester: St. Jerome, 2003.

RECEBIDO EM: 22/10/2016

ACEITO EM: 20/05/2017

PUBLICADO EM: Junho de 2017

[1] Este artigo apresenta parte da pesquisa realizada pela autora em seu mestrado, realizado no Programa de Pós-graduação em Estudos da Tradução da Universidade Federal de Santa Catarina (UFSC), sob orientação do prof. Dr. Lincoln P. Fernandes.

[2] A autora agradece ao suporte fornecido pela Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) durante o período em que foi realizada esta pesquisa.

[3] A autora agradece à revisora e tradutora, Marta Oliveira, pelas valiosas sugestões.

[4] Rossana DA CUNHA SILVA – Graduada em Ciência da Computação (2003) pela Universidade Federal do Pará (UFPA). Graduada em Letras – Inglês (2013) pela Universidade Federal de Santa Catarina (UFSC). Mestre em Estudos da Tradução (2016) pela mesma universidade. Atualmente é pesquisadora assistente na Swansea University. Swansea, País de Gales, Reino Unido.
Lattes: http://lattes.cnpq.br/1301817965381234 E-mail: rossanacs@gmail.com

[5] *Corpus* (plural *corpora*): o *corpus* em linguística é uma grande coleção de textos legíveis por computador e que foram compilados com um propósito específico, sendo recuperados com determinado *software* para pesquisa linguística (LAN, 2015, p. 465).
[6] Neste trabalho, todas as traduções são de minha própria autoria.
[7] Ferramentas CAT (*Computer-aided translation*) ou ferramentas de apoio à tradução.
[8] "*Computerised corpora are becoming increasingly popular in those areas of the discipline which have close links with the hard sciences*" (BAKER, 1995, 224).
[9] "*I will use the term 'multilingual corpora' to refer to a set of texts of two or more monolingual corpora in different languages, built up either in the same or different institutions on the basis of similar design criteria*" (BAKER, 1995, 232).
[10] "*Busa showed that, with a little effort, concordancing could be applied rapidly and effectively to electronic texts"* (MCNERY & HARDIE, 2012, p. 37).
[11] KWIC (*Key word in context)*: palavra-chave em um contexto.
[12] STRAND *(Structural Translation Recognition Acquiring Natural Data)* – Reconhecimento da estrutura de tradução para aquisição de dados naturais.
[13] Disponível em: http://bitextor.sourceforge.net/.
[14] Disponível em: http://nlp.ilsp.gr/redmine/projects/ilsp-fc.
[15] *Off-the-shelf software*: tipo de programa que oferece vários recursos genéricos, ou seja, que a maioria de seus usuários poderia utilizar. Exemplos: programa que possui formatação de texto, revisores gramaticais ou ortográficos etc.
[16] Disponível em: http://ucrel.lancs.ac.uk/wmatrix/
[17] USAS (*UCREL Semantic Analysis System*): Sistema de análise semântica da UCREL (*University Centre for Computer Corpus Research on Language*). Disponível em: http://ucrel.lancs.ac.uk/usas/.
[18] CLAWS (*the Constituent Likelihood Automatic Word-tagging System*): Sistema de etiquetagem morfossintático.
[19] API (*Application Program Interface*): Interface do programa da aplicação.
[20] O *Corpus* de Aprendizes é constituído de redações dos alunos da graduação e dos cursos de extensão das áreas do Departamento de Letras Modernas: alemão, espanhol, francês, inglês e italiano (COMET, 2009).
[21] *Version Variation Visualisation* (VVV): Programa que disponibiliza a "visualização de variações de versões".
[22] Eddy & Viv: Algoritmos existentes no sistema de análise de *corpus* VVV. Eddy é um algoritmo que calcula a variação de pequenos segmentos de texto em um *corpus* de traduções. Em seguida, os resultados são agregados e calculados com base nos segmentos do texto fonte com a utilização do algoritmo Viv ("v*ariation in variation*" – variação em variação) (CHEESMAN et al., 2016, p. 9).
[23] Colocações são geralmente consideradas como palavras que "andam juntas" ou são "encontradas em companhia uma da outra". Uma descrição mais técnica, define as "colocações" como palavras que aparecem juntas com uma probabilidade maior do que apareceriam de forma aleatória (BOWKER, 2002, p. 64).
[24] Para maior detalhamento de características e terminologia relacionadas à tradução, consultar Gambier & Doorslaer (2010) e outros livros da área.
[25] *Corpus* "descartável" (*Disposable corpus)* - também conhecido como virtual, *ad hoc* e *DIY web corpora* – pequenos *corpora* especializados criados com a finalidade de traduzir um texto fonte em particular (ZANETTIN & BERNARDINI, 2003).
[26] A análise de sentimento está presente em diversas ferramentas de mídias sociais, como *Instagram, Facebook, Google+*, de forma a identificar, por meio de mensagens dos usuários, como estes se comportam, quais suas necessidades e preferências.