10.17771/PUCRio.TradRev.30814

Corpora de Tradutores Aprendizes: Construção do CATUERJ e potencial dos CAT para pesquisa e aplicações

Barbara Ramos e Maria Alice Gonçalves Antunes*

1. Introdução

Este trabalho tem como objetivo apresentar uma descrição do processo de construção de um corpus paralelo inglês-português de textos de aprendizes de tradução do Instituto de Letras da Universidade do Estado do Rio de Janeiro (doravante CATUERJ). Para a construção desse corpus, reúnem-se métodos de construção de corpora paralelos tradicionais e corpora de aprendizes de tradução**.**

O CATUERJ é um corpus paralelo de textos traduzidos para o português por formandos em Letras – Inglês/Literaturas, alinhados com os textos originais em inglês. O corpus é de tamanho pequeno e a língua materna dos alunos é o português. O projeto se propõe a criar recurso para informar a pedagogia da tradução, e não foi concebido a partir de um propósito específico de investigação, mas sim para uma agenda de pesquisa em estudos de tradução dentro do Escritório Modelo de Tradução Ana Cristina César (doravante EscrTrad), incluindo

1. explorar variação e escolha na tradução, quando diferentes traduções da mesma fonte são comparadas;

2. comparar os textos traduzidos por aprendizes de tradução aos textos escritos originalmente em português, que pode levar a conclusões sobre o "tradutês", a interlíngua do tradutor-aprendiz e sobre possíveis consequências das restrições que a tradução como atividade comunicativa impõe em oposição à liberdade de expressão;

* Barbara Ramos é mestre em Letras pela UERJ. Maria Alice Gonçalves Antunes é professora adjunta do Instituto de Letras / UERJ.

10.17771/PUCRio.TradRev.30814

3. análise de concordâncias de vários textos traduzidos que podem ajudar a desenvolver e testar hipóteses sobre os itens linguísticos mais propensos a erros ("áreas problemáticas");

4. análise dos erros de tradução mais comuns para descobrir possíveis lacunas na competência tradutória;

5. explorar formas de uso direto do corpus como uma ferramenta de auxílio ao ensino de tradução e como fonte de questões de pesquisa em tradução.

O presente artigo exibe, em primeiro lugar, um breve panorama dos Corpora de Aprendizes de Tradução (doravante CAT) existentes. Em seguida, descreve o projeto do CATUERJ, seu conteúdo e principais características. Finalmente, apresenta algumas perspectivas para novas pesquisas e desenvolvimentos.

2. Projetos recentes na área dos Corpora de Aprendizes de Tradução

O uso de corpora de aprendizes de tradução no ensino e na pesquisa na área dos Estudos da Tradução parece ter sido apresentado pela primeira vez por Robert Spence (1988), da Universidade de Leipzig, Alemanha. Spence compilou um CAT com 49 textos originais – sobre economia, política, turismo, literatura – em alemão (L1), e as respectivas traduções (por alunos) para o inglês (L2). O número de textos traduzidos chega a 1232. Tal número de traduções indica que várias traduções do mesmo original fazem parte do corpus[1].

O PELCRA tem um total de três milhões de palavras que estão divididas em três seções (UZAR, 2008, p. 239-240): a primeira seção conta com três textos originais em inglês, sendo um artigo de jornal, um artigo de revista e um documento sobre contabilidade. Na segunda seção constam três textos originais em polonês com conteúdo equivalente aos textos da primeira seção. A terceira seção é a mais relevante para esse estudo, pois é composta por sessenta traduções dos originais da segunda seção, no par linguístico polonês-inglês, feitas por aprendizes – totalizando 180 traduções. Foi desenvolvido entre 1996, e é uma resposta à necessidade de um grande corpus de referência do polonês para pesquisa e para outras aplicações. O corpus inclui

[1] Disponível em: http://www.spence.saar.de/papers/hongkong1998slides_a.pdf Acesso em 5/04/2015.

10.17771/PUCRio.TradRev.30814

textos escritos (90% do corpus) e textos orais, principalmente contemporâneos (95%), chegando a um total de 3 milhões de palavras. Inclui também um subcorpus com textos acadêmicos escritos em polonês (L1) e suas respectivas traduções (por aprendizes) para o inglês (L2) e uma tipologia de erros construída a partir dos textos incluídos no CAT. Em 2008, o PELCRA tornou-se uma parte do Corpus Nacional do Polonês.

No artigo "Student Translation Archive. Design, Development and Application", Lynne Bowker e Peter Bennison (2003) afirmam que os corpora de aprendizes na área de ensino de língua estrangeira serviram de inspiração para que compilassem "uma coletânea eletrônica de traduções de aprendizes" (p. 104). O *Student Translation Archive* (STA) apresenta textos jurídicos e médicos escritos em francês e/ou espanhol (L2) traduzidos para o inglês (L1). É interessante destacar que a compilação do STA previa o fornecimento de informações (metadados) que permitiriam a identificação das características dos aprendizes bem como das condições em que foram produzidas as traduções. Tal identificação permite a extração de subcorpora do STA através da seleção de textos com características específicas.

Estudos recentes mostram que outros CAT foram compilados por estudiosos do ensino da tradução. O ENTRAD[2] de Celia Florén (2006), o *Russian Learner Translator Corpus* (RusTLC, Kutuzov and Kunilovskaya, 2014) e o MeLLANGE são exemplos desse progresso e do uso dos CAT na pesquisa e na pedagogia da tradução.

O ENTRAD é um corpus paralelo, com textos originais em inglês, a maioria publicados em jornais e revistas, junto a várias traduções para o espanhol. A cada texto em inglês estão relacionadas várias traduções feitas por aprendizes de tradução. Os erros foram classificados visualmente de acordo com uma tipologia criada pelos pesquisadores para essa finalidade. Além dos textos traduzidos pelos aprendizes individualmente, o ENTRAD inclui uma tradução em inglês para cada texto, chamada "de consenso", feita em sala de aula[3].

[2] ENseñanza de la TRADucción.

[3] Disponível em: http://ice.unizar.es/entrad/ Acesso em 05/04/2015.

10.17771/PUCRio.TradRev.30814

O RusTLC[4] é um corpus paralelo com textos alinhados em pares compostos pelos textos originais e os textos traduzidos por aprendizes. Inclui traduções de e para o inglês de alunos cuja língua materna é o russo. Andrey Kutuzov e Maria Kunilovskaya (2014) coordenam o projeto, desenvolvido por uma equipe de professores de tradução e linguistas computacionais, que recolhe traduções de aprendizes provenientes de várias universidades russas, e que "se propõe a criar um recurso representativo e confiável para ser utilizado em estudos da tradução" (p. 315). Em março de 2014, o corpus continha aproximadamente 1.2 milhões de palavras, 258 textos-fonte, e 1795 traduções de aprendizes.

O MeLLANGE resulta de um projeto de larga escala que reúne parceiros de sete países. É um corpus multilíngue, alinhado e anotado com classe gramatical e erros e contém traduções de aprendizes para a língua materna. Por conseguinte, o MeLLANGE torna-se único em termos do número de línguas e de aprendizes de tradução envolvidos. Além disso, o MeLLANGE inclui traduções feitas por profissionais ao lado de traduções feitas por aprendizes, alinhadas no nível do contexto. Sara Castagnoli, Dragos Ciobanu, Kerstin Kunz, Natalie Kübler e Alexandre Volanschi destacam que o MeLLANGE apresenta a oportunidade de identificação de dificuldades de tradução e das possíveis soluções para essas dificuldades (2011, p. 1). Convém esclarecer que o projeto MeLLANGE se encerrou em 30 de setembro de 2007.

Por fim, comentamos sobre o PoNTE (SANTOS, 2014), um corpus de aprendizes bidirecional, português-norueguês e norueguês-português. O PoNTE é um corpus muito abrangente, tanto em relação ao gênero textual (contos, reportagens, blogues, humor, burocracia, promessas eleitorais e texto técnico no que toca aos originais em português, e textos jornalísticos, literários e cartas no que se refere ao norueguês) como em relação à língua dos textos originais (textos brasileiros, moçambicanos, portugueses e angolanos). O PoNTE foi construído de modo a preservar a identidade dos alunos. Dessa forma, não há metadados correspondentes a informações pessoais sobre os aprendizes que fizeram as traduções ou versões. No entanto, o

[4] Russian Learner Translator Corpus.

10.17771/PUCRio.TradRev.30814

pesquisador pode ter acesso a informações como a língua materna do autor de cada tradução e qual variante do português é por ele adotada.

Já existem alguns trabalhos realizados com o PoNTE, comparando pronomes e verbos entre textos originais e suas traduções. Por fazer parte da Linguateca, todo o material foi anotado automaticamente em relação à função sintática e campos semânticos. O PoNTE ainda não recebeu uma anotação de erros, chamada por Santos (2014) de "anotação crítica". No mesmo artigo, a autora questiona qual deve ser o método mais apropriado para a anotação de erros e cita, também, possíveis desdobramentos e pesquisas futuras envolvendo o PoNTE, bem como limitações do trabalho envolvendo corpora de aprendizes.

Os corpora brevemente descritos acima diferem principalmente em termos do número de línguas envolvido, das técnicas / tecnologias utilizadas para a criação do corpus e para a consulta. Os estudos desenvolvidos com base nos corpora descritos acima concentram-se nas traduções para a língua materna dos alunos, a fim de pesquisar fenômenos relacionados à tradução (CASTAGNOLI et al., 2011), com exceção do corpus PELCRA, que foi concebido como um corpus que se presta para estudos sobre a produção dos aprendizes e pode servir também como um instrumento de auxílio ao ensino de línguas estrangeiras (Uzar & Waliński, 2001), e, portanto, inclui a produção dos alunos na língua estrangeira. Ressalte-se que a análise de erro parece ser o tipo mais frequente de análise realizada sobre esses CATs (ESPUNYA, 2013; KUNILOVSKAYA e KUTUZOV, 2014), mesmo que esses corpora também possibilitem pesquisas sobre estratégias de tradução e sobre escolhas tradutórias. A análise de erro é um tipo de anotação que pode ser feita a qualquer momento em um corpus. Vale pontuar que anotações de classe gramatical ou de erro não foram feitas com o CATUERJ até a elaboração deste artigo.

A tabela 1, abaixo, resume as principais características dos corpora a que nos referimos nesta seção.

10.17771/PUCRio.TradRev.30814

Corpus	Língua Fonte	Língua Alvo	Status da língua alvo	Anotação de erro	Disponível online
STA	Francês, Espanhol	Inglês	materna	não	não
PELCRA CAT	Polonês	Inglês	estrangeira	sim	não
RusTLC	Inglês	Russo	materna	sim	não
ENTRAD	Inglês	Espanhol	misto	sim	sim
MeLLANGE	Dinamarquês Inglês Francês Espanhol	Catalão Dinamarquês Inglês Espanhol Italiano Francês	materna	sim	Apenas durante o período no qual o projeto esteve ativo

Tabela 1: Listagem dos CAT descritos neste artigo e suas principais características

No levantamento feito por Ramos (2016), também foram encontrados dois CAT em desenvolvimento no Brasil: o primeiro, criado na UNESP e o segundo, desenvolvido por Joacyr Tupinambás de Oliveira, da USP, como tema de sua dissertação de mestrado, defendida em 2015. Nenhum deles está disponível para consulta on-line.

A respeito do CAT desenvolvido em 2014 pela aluna de graduação e bolsista de iniciação científica Yasmin de Sousa Brito, na Universidade Estadual Paulista Julio de Mesquita Filho, apenas o resumo de seu trabalho no XXV Congresso de Iniciação Científica na própria universidade[5] foi encontrado. No resumo há informações superficiais sobre o corpus. Com base nessas informações, a autora concluiu que o CAT não parece ter um nome; que ele é constituído por versões de textos jornalísticos no par linguístico português-inglês; e que o trabalho realizado com esse CAT tem o objetivo de verificar se colocações que aparecem com maior frequência são utilizadas por falantes nativos de língua inglesa, com o auxílio do COCA, o *Corpus of Contemporary American English*. Não há informações sobre o tamanho do corpus

[5] Disponível em: http://prope.unesp.br/cic/admin/ver_resumo.php?area=100076&subarea=22618&congresso=35&CPF=40960999884. Acesso em 11/01/2016.

10.17771/PUCRio.TradRev.30814

nem sobre o tratamento que a ele foi dado. Tampouco achamos outras publicações ou pesquisas envolvendo esse corpus até a finalização deste artigo. De qualquer forma, ele serve como uma pequena amostra de que o trabalho com CAT está sendo desenvolvido no Brasil, mesmo que esse trabalho pareça ainda um pouco tímido.

O CAT de J. T. Oliveira (2015), que também não tem nome específico, foi compilado com dois objetivos. O primeiro era a organização e alinhamento do corpus, logo após a compilação. O segundo objetivo seria um levantamento de opções feitas pelos tradutores aprendizes e o que os motivou a fazê-las, tanto lexical como morfossintaticamente. O corpus conta com oito textos em inglês que contêm de duzentas a trezentas palavras e suas traduções para o português, englobando aproximadamente cem traduções para cada texto-fonte. Essas traduções foram feitas por tradutores aprendizes nos primeiro e segundo períodos de uma faculdade particular do estado de São Paulo, entre 2010 e 2013. Para documentar o corpus, o pesquisador traçou um perfil dos participantes com informações relacionadas a faixa etária, sexo, conhecimento de inglês, vivência no exterior, idioma nativo, experiência como tradutor e experiência em outras áreas de conhecimento (contato com outro curso superior por pelo menos dois anos). A característica mais relevante desse CAT para minha pesquisa é que o alinhamento foi feito manualmente e utilizando o programa Microsoft Excel, mesmo programa utilizado para verificar o alinhamento do CATUERJ, inicialmente feito pelo *LFAligner*, que será discutido no capítulo 2 desta dissertação.

3. O CATUERJ e sua construção

Inicialmente, a dissertação de mestrado de Ramos (2016) visava à montagem, documentação e anotação[6] de erros de um Corpus de Aprendizes de Tradução (CAT doravante) constituído por traduções de estagiários do EscrTrad, o CATUERJ. A ideia de fazer a anotação de erros no CATUERJ surgiu da expectativa de que tradutores iniciantes e aprendizes cometem muitos erros[7]

[6] "Anotar é delimitar um segmento de texto e atribuir-lhe uma etiqueta (ou conjunto de etiquetas)" que contenha informação linguística associada (Freitas, 2015, p. 36).

[7] Destacamos, nesse ponto, que foi adotada a visão de erro definida por Maria Paula Frota, sendo ele causado por "ignorância - ou seja, o desconhecimento acerca de alguma estrutura lingüística, algum aspecto cultural ou, ainda, alguma informação relativa ao contexto da tradução, como o seu público alvo,

10.17771/PUCRio.TradRev.30814

em suas traduções. No entanto, ao alinhar o corpus, analisando os textos originais em paralelo às traduções, Ramos (2016) constatou que, na verdade, as traduções eram naturais e fluentes, sem uma grande presença de erros que impedissem a comunicação daquele conteúdo da língua-fonte para a língua-alvo – contrariando a expectativa inicial de que tradutores iniciantes e aprendizes em formação cometeriam muitos erros. Desta forma, a autora optou por enfatizar aspectos positivos das traduções e das escolhas tradutórias dos aprendizes, indo de encontro a meu pensamento anterior. Essa escolha seria uma tentativa de colaborar para a construção de uma imagem mais positiva do tradutor aprendiz e do próprio processo de aprendizagem – que pouco realça o acerto –, salientando não apenas aspectos negativos da tradução, mas enfatizando escolhas tradutórias positivas.

O Corpus de Aprendizes de Tradução da UERJ (CATUERJ) é composto por textos originais em inglês e suas respectivas traduções para o português, feitas por estagiários do EscrTrad, entre os anos de 2006 e 2014. Os estagiários cujas traduções foram reunidas trabalham com o par linguístico inglês-português. No EscrTrad, eles executam trabalhos propostos por clientes de fora do Escritório e também tarefas designadas pela coordenadora de inglês do projeto.

Os trabalhos encomendados ao EscrTrad vêm das comunidades interna e externa. Esses trabalhos podem ser traduções, versões e transcrições (de texto oral para escrito, por exemplo). O trabalho chega para o estagiário, que discute com o coordenador a melhor forma de executá-lo. O processo de tradução é iniciado, passando pelas mãos do coordenador quantas vezes quanto forem necessárias, até que estagiário e coordenador cheguem a um produto final. A tradução pronta é, então, entregue ao cliente. Esse tipo de trabalho é encomendado por clientes reais – não se tratando de uma simulação. É importante esclarecer que somente os textos originais e as respectivas traduções realizadas como exercício fazem parte do CATUERJ.

As tarefas pedidas pela coordenadora de inglês, por sua vez, podem englobar pesquisas, produções de texto e traduções. As pesquisas são na área

seu veículo etc. - e um certo fenômeno ou funcionamento psíquico inconsciente que é muito comum e que redunda no lapso de língua (...) [este] consiste em uma manifestação do inconsciente e pode ocorrer em qualquer operação que envolva a linguagem verbal" (Frota, 2006, p. 142).

10.17771/PUCRio.TradRev.30814

de tradução e podem ser de caráter teórico ou prático, como sobre cursos de tradução on-line e ementas dos cursos de tradução oferecidos em universidades, bem como sobre publicações com temas específicos na área de tradução e Estudos da Tradução. As produções textuais também têm a tradução como tema na maior parte das vezes, podendo ser uma resenha de um livro ou artigo sobre tradução, ou mesmo um relatório sobre as pesquisas realizadas. As traduções, por sua vez, são atividades tradutórias propriamente ditas, como traduções de artigos científicos; artigos de páginas virtuais; capítulos de livros sobre tradução. Os gêneros textuais propostos para que o estagiário realize a tradução são diversificados, pois o objetivo principal das tarefas de tradução é melhorar a prática tradutória do tradutor em treinamento.

As traduções e os textos originais compilados para o CATUERJ, como mencionado anteriormente, foram todos provenientes das tarefas tradutórias definidas pela coordenadora de inglês do EscrTrad, entre os anos de 2006 e 2014, por tratarem de um mesmo tema – tradução. O CATUERJ é um corpus paralelo constituído por dois subcorpora. O primeiro, denominado CATUERJ-Ing, é um subcorpus com os textos em inglês (L2), que foram tirados de capítulos de livros sobre tradução. Na Tabela 2 há referências bibliográficas dos capítulos e os respectivos livros que fazem parte deste subcorpus.

O outro subcorpus, o CATUERJ-Tra, é onde estão armazenadas as traduções dos textos que fazem parte do CATUERJ-Ing para o português, língua materna dos estagiários. O CATUERJ, portanto, é um corpus paralelo unidirecional inglês-português.

Como mostra a Tabela 3, os textos em inglês contam com um total de 16.484 *tokens* em sete textos originais diferentes, enquanto os textos traduzidos englobam 18.572 *tokens*, divididos entre oito traduções. Os textos 4 e 5, no CATUERJ-Ing, aparecerem com números diferentes apesar de serem o mesmo texto original – ou seja, esse texto teve duas traduções, cada uma feita por um estagiário em diferentes períodos. Essa duplicação foi feita para que as traduções não se misturassem e pudessem ser relacionadas mais facilmente no corpus alinhado. Ao todo, o corpus é constituído por 35.056 *tokens*.

10.17771/PUCRio.TradRev.30814

Nome do ar-quivo	Referências bibliográficas
TO01	SCHIFFRIN, A. Preface. In: _____. *The business of books:* How International Conglomerates Took Over Publishing and Changed the Way We Read. Londres: Verso, p. 01-14, 2001.
TO02	VENUTI, L. How to read a translation. In: VENUTI, L. *Translation changes everything: theory and practice.* Londres: Routledge, 2013.
TO03	BASSNETT, S. The meek or the mighty. In: ÁLVAREZ, R. e VIDAL, M. (Eds.) *Translation, Power, Subversion*. Clevedon: Multilingual Matters Ltd. p. 10-24, 1996.
TO04 e TO05	BASSNETT, S. Original Sin. In: _____. *Reflections on Translation.* Bristol: Multilingual Matters, p. 12-15, 2011.
TO06	*RAZMJOU, L. To be a good translator. Second International Conference on Critical Discourse Analysis: the Message of the Medium. Iémen, Universidade de Hodeidah, out., 2003. Disponível em:* <http://www.translationdirectory.com/article106.htm.>. *Acesso em: 18/01/2016.*
TO07	BASSNETT, S. Translation or Adaptation? In: _____. *Reflections on Translation*. Bristol: Multilingual Matters, p. 40-43, 2011.
TO08	SHIELDS, K. Gained in Translation: Language, Poetry, and Identity in Twentieth-Century Ireland. Berna: Peter Lang, 2000.

Tabela 2: Referências dos textos originais que compõem o CATUERJ

Fonte: Ramos (2016).

	CATUERJ-Ing	CATUERJ-Tra
Texto 1	1,754	1,823
Texto 2	2,840	2,911
Texto 3	5,727	5,833
Texto 4	1,457	1,325
Texto 5		1,597
Texto 6	1,625	1,909
Texto 7	1,553	1,593
Texto 8	1,528	1,581
Total 35,056	16,484	18,472

Tabela 3: *Tokens* de textos originais e traduzidos do CATUERJ-Tra

Fonte: Ramos (2016).

10.17771/PUCRio.TradRev.30814

Um aspecto que chama a atenção, a partir de um olhar mais detalhado para a Tabela 03, é a pouca diferença do número de *tokens* entre os textos originais e os traduzidos, o que pode indicar a preocupação do tradutor aprendiz em relação à manutenção da quantidade de palavras do original. Além disso, nota-se ainda a quantidade de *tokens* dos TT 04 e 05, que foram traduções de um mesmo original, embora o TT04 tenha quase trezentos *tokens* a mais que o TT05; e o número de *tokens* do TT04 também seja menor que o original, fato considerado incomum em textos traduzidos, que tendem a ser maiores que o original em número de *tokens*.

4. A documentação do CATUERJ

A documentação é uma forma de registrar informações estruturais do corpus, ou seja, dados externos aos textos, como informações bibliográficas, de publicação e catalogação. Segundo Freitas,

> a documentação deve incluir informações quantitativas como o número de palavras; informações relativas à proveniência do material compilado (datas de acesso e ou de criação; sites ou locais de onde os textos foram retirados); critérios de seleção da amostra; tipo de texto; registro; e pode conter ainda informação relacionada à autoria dos textos, ao gênero dos autores, e todas as informações que os criadores do corpus considerarem relevantes [...] A explicitação de tais informações possibilita um melhor entendimento dos resultados, tornando mais fácil a comparação com outros corpora, por exemplo. (Freitas, 2015, p. 34).

Como já mencionado na seção anterior deste artigo, a tipologia de erros é um tratamento valioso dado ao corpus em termos de pesquisa, porém não primordial. Por outro lado, como já ressaltado em publicações da área de linguística de corpus (Baker, 2004; Pravec, 2002; Reppen, 2010), a documentação é uma etapa fundamental (Freitas, 2015, p. 34) no tratamento inicial do corpus. As informações, anexadas ao texto em forma de cabeçalho, facilitam a recuperação do texto, a avaliação da validade daquele corpus para a pesquisa a ser realizada e a geração de outros subcorpora, selecionando textos de um mesmo autor, de certo ano ou de certo gênero. Na documentação de um texto, os metadados da documentação podem anexar ao texto informações externas do próprio texto. Na documentação de um CAT

10.17771/PUCRio.TradRev.30814

pode haver, também, metadados referentes às traduções e aos tradutores aprendizes. As diferentes documentações dos CAT deste artigo demonstram que são comuns informações sobre o tradutor aprendiz, como nome, gênero, idade, grau de experiência e a data na qual a tradução foi realizada.

Segundo Celia Florén e Rosa Lorés (2008, p. 439), a documentação favorece a "busca nas traduções por motivos pedagógicos [...] e torna o material acessível para propósitos de pesquisa [...]"[8]. A documentação de um CAT também permite que o pesquisador esteja ciente de detalhes inerentes ao processo de criação daquele corpus, melhorando a comunicação entre estudante e corpus. Traçando-se um paralelo com a área médica, a documentação pode funcionar como uma espécie de prontuário, no qual se encontram informações referentes ao passado daquele paciente, sendo o paciente o corpus.

A documentação do corpus CATUERJ-Tra foi feita visando à reutilização do corpus para pesquisas futuras, facilitando a interpretação de resultados por outros pesquisadores de forma mais consistente, não somente pelos criadores do corpus. Para definir os itens a serem inseridos nos metadados da documentação do CATUERJ, os critérios utilizados na documentação de alguns dos CAT descritos no início deste artigo foram adaptados. Esse processo de adaptação e seleção resultou em oito itens, sendo eles: (a) Título do TO; (b) Autor do TO; (c) Gênero textual; (d) Ano de publicação; (e) Editora; (f) Idioma do TO; (g) Número de *tokens* do arquivo; (h) Número de segmentos alinhados; e (i) Nome do subcorpus do qual o texto faz parte.

A escolha de não acrescentar informações sobre os tradutores aprendizes foi feita por questões de ordem prática. Seria necessária uma pesquisa mais extensa para recrutar dados sobre os estagiários que realizaram as traduções do subcorpus CATUERJ-Tra. Além disso, essa é uma etapa que pode ser modificada a qualquer momento, podendo o pesquisador adicionar novas informações que lhe sejam úteis para a pesquisa que vai desenvolver.

[8] No original: "the search for translations for pedagogical reasons [...] and also makes material accessible for research purposes […]" (tradução nossa).

10.17771/PUCRio.TradRev.30814

5. O alinhamento do CATUERJ

O alinhamento de textos na linguística de corpus pode ser definido como a busca pela correspondência entre segmentos dos textos-alvo e fonte de um corpus paralelo. Ou seja, ao realizar o alinhamento, o pesquisador busca no texto-alvo – a tradução – o trecho correspondente a determinada passagem do texto-fonte – o original.

O alinhamento pode ocorrer considerando diferentes recortes. É de fundamental importância não considerar que a frase seja a unidade de tradução, pois, em diversas situações, a frase pode perder seu sentido original se tirada do contexto (Gaussier et al., 2000, p. 272). Então, os corpora paralelos podem ser alinhados no nível de palavra, de sentença, de parágrafo, ou por segmentos não específicos, que podem ser trechos de textos de tamanhos diferentes. Os corpora, de forma geral, podem ser alinhados da forma que melhor couber ao objetivo planejado para cada pesquisa.

Os corpora podem ser alinhados manualmente pelo próprio grupo de pesquisa ou automaticamente, com auxílio de um *software*. Tanto o alinhamento manual como o automático têm vantagens e desvantagens. No alinhamento manual, o pesquisador tem a oportunidade de estar próximo de seu corpus em todo o processo, lendo cada segmento de texto a ser alinhado e pode, assim, conhecer melhor seu corpus de pesquisa. A desvantagem está na demora desse processo. O pesquisador precisa ler trecho a trecho tanto no subcorpus de L1 como no subcorpus de L2, consumindo muitas horas hábeis de sua pesquisa.

Esse gasto de tempo não acontece quando o corpus é alinhado automaticamente. O alinhamento automático de corpora paralelos, por ser feito pelo *software*, é um processo quase instantâneo. O pesquisador seleciona os corpora ou subcorpora que deseja alinhar, inserindo-os no *software*. O pesquisador seleciona os corpora ou subcorpora que deseja alinhar, inserindo-os no *software*, podendo essa seleção ser a etapa mais demorada. A etapa seguinte, o alinhamento em si, feito pelo *software*, acontece quase imediatamente, com um gasto de tempo menor que um minuto, independente do tamanho e quantidade de corpora ou subcorpora a serem alinhados. Por ser mais rápido, o alinhamento automático é mais usado quando há grande quantidade de material.

10.17771/PUCRio.TradRev.30814

Andras Farkas[9], idealizador do *LFAligner – software* utilizado para alinhar o CATUERJ – afirma que o alinhamento automático tem em média noventa e cinco por cento de precisão. No entanto, Farkas reconhece que textos que foram autoalinhados devem ser revisados pelo pesquisador para certificar-se de que o resultado foi satisfatório[10]. Há, ainda, a possibilidade de os textos não serem adequados para o autoalinhamento, pois pode haver um alto grau de incompatibilidade de segmentos devido à troca de ordem de frases, por exemplo.

Independente do tipo de alinhamento que seja feito no corpus, o alinhamento de textos paralelos é uma forma de obter informações proveitosas que podem ser aplicadas de diferentes maneiras. No campo da atividade tradutória, ele pode ser transformado em conteúdo para bancos de dados de memórias de tradução, por exemplo. Já na área de terminologia, pode auxiliar a construção de dicionários bilíngues. Por fim, os corpora ou subcorpora alinhados podem, também, servir para a elaboração de material didático para ensino de línguas.

O CATUERJ está armazenado eletronicamente em dois formatos: .doc e .txt. O alinhamento foi feito automaticamente e em nível de sentenças, utilizando o *software LFAligner*. Os textos e suas respectivas traduções foram inseridos no *software* e, em seguida, divididos em segmentos e alinhados pelo próprio programa – o segmento do texto original e seu correspondente traduzido. A Tabela 4 mostra o número de segmentos divididos e alinhados automaticamente em cada texto dos dois subcorpora, o CATUERJ-Ing e o CATUERJ-Tra.

Após obter os dados da Tabela 4, referentes ao alinhamento automático realizado pelo *LFAligner*, modifiquei os pares correspondentes, igualando o número de segmentos de cada subcorpus. Essa etapa foi realizada manualmente com o auxílio do programa Microsoft Excel, tendo como inspiração a metodologia desenvolvida e adotada por Oliveira (2015), que construiu seu CAT nessa mesma ferramenta. No entanto, a revisão manual do

[9] Disponível em http://www.farkastranslations.com/ (último acesso em 12/01/2016).
[10] Disponível em http://www.farkastranslations.com/alignment.php (último acesso em 11/01/2016).

10.17771/PUCRio.TradRev.30814

alinhamento nesta pesquisa só foi viável devido ao tamanho reduzido do corpus, o que possibilitou a conferência de todas as passagens do CAT. Dessa forma, no subcorpus CATUERJ-Tra há segmentos constituídos por mais de uma frase, para que cada segmento esteja equivalente ao conteúdo e sentido do segmento correspondente do CATUERJ-Ing.

	Segmentos no CATUERJ-Ing (L2)	**Segmentos no CATUERJ-Tra (L1)**
TO01-TT01	79	81
TO02-TT02	107	109
TO03-TT03	219	222
TO04-TT04	56	59
TO05-TT05	56	62
TO06-TT06	79	89
TO07-TT07	61	62
TO08-TT08	54	57

Tabela 4: Números de segmentos dos subcorpora CATUERJ-Ing e CATUERJ-Tra

Fonte: Ramos (2016).

É possível notar que, em alguns dos pares, há uma diferença de até dez segmentos a mais do TO para o TT. Isso ocorre devido à quebra que o tradutor em treinamento faz de uma frase longa em inglês ao traduzi-la para o português. A Tabela 5, a seguir, mostra três exemplos dessa ocorrência em diferentes pares dos subcorpora. É importante esclarecer que as traduções utilizadas no CATUERJ foram compiladas antes da revisão da coordenadora do projeto, mantendo a autenticidade das escolhas tradutórias dos estagiários.

A exceção à explicação dada antes da Tabela 05 é o TT05. Ele tem seis segmentos a mais que o TO05, entretanto, ao analisar o par de textos alinhados pelo *LFAligner*, percebemos que o motivo do aumento do número de segmentos foi um deslize do tradutor aprendiz, que deslocou um trecho que seria do final do texto mais para o início, fazendo com que os segmentos 06

a 13 – oito, no total – ficassem sem correspondência e, em seguida, tornou a traduzir os segmentos correspondentes ao TO. A Tabela 6 ilustra essa ocorrência.

Exemplo 1: TO03-TT03
[58] Central to the polysystems approach were certain key assumptions about translation, most crucial of which was the recognition of the role played by translation in shaping the literary polysystem.
[58] Fundamental para a abordagem dos polissistemas eram certas suposições chave sobre a tradução. A mais crucial delas era o reconhecimento do papel desempenhado pela tradução na modelagem do polissistema literário.
Exemplo 2: TO04-TT04
[45] Apart from restricted supplies of manuscripts, scribes to copy them and teachers to disseminate knowledge, not all rules promoted learning, and disease, wars and Viking raiders made consistent study difficult.
[45] À parte de uma oferta restrita de manuscritos, escribas para copiá-los e professores para disseminar o conhecimento, nem todos os governantes promoviam o ensino. Além disso, doenças, guerras e piratas viquingues tornavam difícil o estudo consistente.
Exemplo 3: TO06-TT06
[63] Another important point is that successful translators usually choose one specific kind of texts for translating and continue to work only in that area; for example a translator might translate only literary works, scientific books, or journalistic texts.
[63] Um outro ponto importante a ressaltar é que tradutores consagrados normalmente escolhem um modelo específico de texto para traduzir e tendem a permanecer trabalhando somente nesta área. Por exemplo, um tradutor pode especializar-se em traduzir apenas trabalhos literários, livros científicos ou textos jornalísticos.

Tabela 5: Exemplos de quebras de frases do inglês para o português no CATUERJ

Em seguida, os pares de textos alinhados e revisados foram salvos no formato .xls, ficando disponíveis para consultas futuras através do programa Microsoft Excel. A partir dos arquivos armazenados em formato .xls, o CATUERJ passou a ser um corpus paralelo alinhado sentencialmente e de fácil compreensão para o pesquisador que se debruçar sobre ele, independente de seu objetivo.

10.17771/PUCRio.TradRev.30814

10.17771/PUCRio.TradRev.30814

Número do segmento	TO05	TT05
[04]	Most translators would immediately opt for the second option, being all too aware of the pitfalls of the word-for-word approach.	A maioria dos tradutores, familiarizados com o perigo do uso da palavra-por-palavra, optaria de imediato pela segunda opção.
[05]	After all, a translation that is too literal can be simply unreadable.	Afinal, uma tradução literal demais pode acabar por ilegível.
[06]		A história da tradução literal através da glosa interlinear dos antigos manuscritos não é apenas uma narrativa especializada para acadêmicos, mas a história do nascimento do inglês escrito.
[07]		De modo parecido, a glosa interlinear em outras línguas europeias deu origem a outras formas escritas de vernáculo.
[08]		Isso significa que a literatura oral que tinha circulado por séculos, como os grandes épicos germânicos, as canções, as charadas e as histórias, podia ser registrada em línguas que, no mínimo, equiparavam-se ao latim, ainda que não o suplantassem em estilo, àquela época.
[09]		É interessante refletir sobre o papel da tradução literal no desenvolvimento das competências linguísticas.
[10]		Comecei menosprezando as versões palavra por palavra e ainda defendo que uma boa tradução vai além do literal.
[11]		Entretanto, a transposição exata de um texto serve a um propósito bem definido.
[12]		Os escribas que faziam as anotações anglo-saxãs eram, acredito, muito melhores linguistas do que meu filho.

10.17771/PUCRio.TradRev.30814

[13]		Mas o princípio de alinhamento de palavras, usado para entender como línguas diferentes funcionam, ainda é válido através dos séculos.
[14]	Inexperienced translators seem to go for word-for-word renderings, and it seems to be a universal truth that translation in the tourist industry worldwide is pretty dire.	Já tradutores inexperientes parecem optar pela adoção da palavra-por-palavra, e é quase uma verdade universal que a tradução na indústria de turismo é péssima no mundo inteiro.
[15]	Here are a couple of word-for-word items, one from an Indonesian hotel brochure and the other from a pamphlet produced by the city of Salamanca tourist office:	Encontramos a seguir alguns exemplos do uso da palavra-por-palavra, um retirado da brochura de um hotel na Indonésia, e outro de um panfleto produzido pela secretaria de turismo de Salamanca:

Tabela 6: Ilustração do deslocamento de oito segmentos no par TO05 – TT05

6. CATUERJ: principais características

O CATUERJ pode ser considerado um CAT pequeno, por conter menos de vinte mil palavras (Berber Sardinha, 2000; Koester, 2010), especializado e representativo. Consideramos o CATUERJ representativo de acordo com a noção de representatividade discutida por Almut Koester (2010, p. 70), como sendo a garantia de que o corpus tenha uma variedade de construções linguísticas nele presentes que possam dialogar com o objetivo do pesquisador. Por outro lado, Berber Sardinha (2000, p. 342-345) alega que, quanto maior o corpus, possivelmente maior será sua representatividade, o que não acontece no caso do CATUERJ. Apesar de ser um corpus de tamanho pequeno, sua amostra linguística corresponde à compilação de parte da produção dos estagiários do EscrTrad (tarefas tradutórias). Logo, seu recorte e caráter especializado fazem com que ele seja representativo dentro do contexto onde pode ser mais explorado.

Quanto ao CATUERJ ser considerado um corpus pequeno, Randi Reppen (2010) alega que não há um número de palavras fixo para que um corpus seja compilado. Para ela, o tamanho do corpus deve variar de acordo com o propósito para o qual ele servirá. O pesquisador deve apenas se preocupar com as variáveis de praticidade e representatividade. Para considerar a primeira variável – a praticidade –, a preocupação deve ser em relação ao prazo para que a pesquisa seja realizada e finalizada, entre outros detalhes

10.17771/PUCRio.TradRev.30814

de ordem prática, como datas limites para publicações e emissão de documentos, por exemplo. Já para a segunda variável – a representatividade –, o pesquisador deve determinar se já coletou textos suficientes para representar o item a ser investigado.

A extensão do CATUERJ, ao longo de sua compilação, pareceu ser suficiente e elucidativa para a pesquisa desenvolvida por Ramos (2016). Talvez, porém, para uma pesquisa com outro objetivo, seu tamanho não seja satisfatório. Uma característica positiva do CATUERJ, no entanto, é que sua existência está diretamente ligada ao EscrTrad. Portanto, o pesquisador pode compilar outros textos produzidos por estagiários em anos mais recentes ou com outros assuntos, desde que dentro de seu objetivo de pesquisa, e acrescentá-los ao CATUERJ para que produções futuras de estagiários também possam dialogar entre si, bem como com as já existentes no CATUERJ, aprofundando o aprendizado e o conhecimento sobre o processo de tradução e sobre os alunos interessados no estágio em tradução.

Dentre os motivos pelos quais um corpus pode ser especializado, cito os que Lynne Flowerdew (2004, p. 21) apresenta como sendo os principais: a compilação do corpus com o objetivo de investigar um item lexical ou gramatical, em especial; um único gênero textual; um tipo de texto; um assunto; um registro de linguagem; ou um contexto específico, como os participantes, o propósito comunicativo do texto, ou condições particulares para a criação do texto. O CATUERJ, então, é especializado pois foi compilado visando apenas o contexto de traduções realizadas por tradutores em treinamento que atuam como estagiários em um escritório universitário em particular.

Por último, é fundamental ressaltar que o CATUERJ também é representativo no contexto de estudos com CAT no Brasil. Apesar de não poder ser disponibilizado on-line por motivos legais, já que todos os textos originais estão regidos por direitos autorais, o CATUERJ está entre os primeiros CAT a serem compilados e utilizados para pesquisas de forma concreta no Brasil, ilustrando algumas das possíveis aplicações do conteúdo de um CAT na sala de aula de disciplinas voltadas para formação de tradutores. Até a conclusão deste trabalho, o CATUERJ vinha sendo utilizado como corpus para outra pesquisa de mestrado em andamento, também na UERJ.

10.17771/PUCRio.TradRev.30814

7. Limitações, perspectivas para futuras pesquisas e desenvolvimentos

Enquanto pesquisadoras da área, temos consciência de que o CATUERJ não se trata de uma inovação na área de estudos com corpora, ou mesmo de corpora de aprendizes de tradução. No entanto, julgamos que ele é um CAT que pode contribuir para os estudos com corpora e a formação de tradutores, principalmente dentro da universidade onde foi concebido. Como sua construção foi parte da dissertação de mestrado de Ramos (2016), sabemos também que as limitações de ordem prática influenciam diretamente no cronograma e programação das atividades a serem cumpridas. Nossa expectativa é de que o CATUERJ continue sendo explorado por grupos de pesquisa, seja em nível de graduação ou pós-graduação, assim ampliando suas possibilidades de uso e aplicação e contribuindo, mesmo que em pequena escala, para o avanço dos estudos com corpora no Brasil. Atualmente, há uma dissertação de mestrado em andamento que também fala sobre o CATUERJ.

Por fim, destacamos algumas perspectivas e aplicações futuras para o CATUERJ assim como outros CAT, de maneira geral. No que diz respeito à continuidade de tratamento do próprio CATUERJ, o caminho já traçado por pesquisadores da área sugere a anotação de erros e de classes gramaticais, já explorada em outros CAT. Outro caminho possível, também já explorado na área, seria "a análise da lista de frequências de palavras para que os dados do CATUERJ direcionem uma pesquisa sobre aspectos gramaticais, sintáticos ou semânticos, como já vem sendo feito com corpora em geral (de aprendizes ou não)" (RAMOS, 2016, p. 86). O CATUERJ poderia também ser usado pelos professores de disciplinas sobre tradução em uma tentativa de estudar o processo de aquisição da língua estrangeira pelos tradutores aprendizes. E, a partir desse estudo, auxiliar na categorização de áreas desafiadoras para os alunos, como gramática, a tradução feita palavra por palavra, terminologia específica ou uso de léxico especializado; com o objetivo de conscientizar o tradutor aprendiz sobre a tradução ser feita baseada em escolhas, e não em certo e errado.

Sendo o CATUERJ usado dentro de sala de aula, ele pode ser comparado a corpora paralelos que contenham traduções profissionais. Comparando o processo tradutório dos aprendizes no CATUERJ a corpora com traduções profissionais, os alunos poderiam ter uma formação direcionada para

10.17771/PUCRio.TradRev.30814

a autonomia. Mais uma possibilidade seria o acréscimo de textos, criando um ou mais subcorpora para o CATUERJ, visando sua disponibilização on-line – ainda que parcial. Outra oportunidade concreta para pesquisa decorrente da construção do CATUERJ e da pesquisa de mestrado de Ramos (2016) seria um estudo sobre os procedimentos de tradução mais frequentemente usados pelos estagiários do EscrTrad, e também a elaboração de materiais didáticos que levem os resultados dessa pesquisa à sala de aula.

Referências

BAKER, Mona. "A corpus-based view of similarity and difference in translation". *International Journal of Corpus Linguistics*, v. 9, n. 2, p. 167-193, 2004.

BOWKER, Lynn; BENNISON, Peter. "Student Translation Archive and Student Translation Tracking System. Design, Development and Application". In: Zanettin, Federico; Bernardini, Silvia; Stewart, Dominic (eds.), *Corpora in translator education*. Manchester: St. Jerome Publishing, p. 103-118, 2003.

CASTAGNOLI, Sara; CIOBANU, Dragos; KUNZ, Kerstin; VOLLANSKI, Alexandra. "Designing a learner translator corpus for training purposes". In: Kubler, N. (ed.) *Corpora, Language, Teaching and Resources: From Theory to Practice.* Berna: Peter Lang, p. 221-248, 2011.

FLORÉN SERRANO, Celia; LORÉS SANZ, Rosa. "The application of a parallel corpus English-Spanish to the teaching of translation (ENTRAD project)". In: Muñoz Calvo, M. et al. (eds.), *New Trends in Translation and Cultural Identity*. Cambridge: Cambridge Scholars Publishing, p. 433-444, 2008.

FLOWERDEW, Lynne. "The Argument for Using English Specialized Corpora to Understand Academic and Professional Settings". In: Connor, U.; Upton, T. (eds.), *Discourse in the Professions: Perspectives from Corpus Linguistics*. Amsterdã: John Benjamins, p. 11-33, 2004.

FREITAS, Cláudia. "Corpus, Linguística Computacional e as Humanidades Digitais". In: Leite, Miriam; Teresa Gabriel, Carmen (orgs.), Linguagem, discurso, pesquisa e educação. 1. ed. - Petrópolis: FAPERJ, p. 23-56, 2015.

FROTA, Maria Paula. "Erros e lapsos de tradução: um tema para o ensino". *Cadernos de Tradução*, v. 1, n. 17, p. 141-156, jan/jun., 2006, Santa Catarina. Disponível em:

10.17771/PUCRio.TradRev.30814

https://periodicos.ufsc.br/index.php/traducao/article/view/6859/6411 (acesso em 26 de dezembro de 2015)

GAUSSIER, Eric; HULL, David; AÏT-MOKTHAR, Salah. "Term alignment in use: Machine aided human translation". In: Véronis, J. (ed.). *Parallel text processing: Alignement and Use of translation Corpora*. Kluwer Academic Publishers, p. 253-74, 2000.

KOESTER, Almut. "Building small specialized corpora". In: O'Keeffe, A.; McCarthy, M. (eds.), *The Routledge Handbook of Corpus Linguistics*. London: Routledge, p. 66-79, 2010.

KUTUZOV, Andrey; KUNILOVSKAYA, Maria. "Russian Learner Translator Corpus: Design, Research Potential and Applications". In: Sojka, P. et al. (eds.), *Text, Speech and Dialogue. TSD 2014*. Brno: Springer, p. 315-323, 2014.

OLIVEIRA, Joacyr T. "A linguística de corpus na formação do tradutor: compilação e proposta de análise de um corpus paralelo de aprendizes de tradução". Dissertação (Mestrado em Letras) – Universidade de São Paulo, 2015.

PRAVEC, Norma A. "Survey of learner corpora". *ICAME Journal*, v. 26, p. 81-114, 2002.

RAMOS, Barbara C. M. P. "Corpus de Aprendizes de Tradução: possíveis aplicações na sala de aula de uma disciplina de tradução". Dissertação (Mestrado em Linguística) – Instituto de Letras, Universidade do Estado do Rio de Janeiro, Rio de Janeiro, 2016.

REPPEN, Randi. "Building a Corpus: What are the key considerations?" In: O'Keeffe, A.; McCarthy, M. (eds.), *The Routledge Handbook of Corpus Linguistics.* Londres: Routledge, p. 31-37, 2010.

SANTOS, Diana. "PoNTE: apontando para corpos de aprendizes de tradução avançados". *Linguamática* v. 6, n.1, p. 69-86, 2014.

SARDINHA, Tony Berber. "Linguística de Corpus: Histórico e Problemática". In: *D.E.L.T.A.*, v. 16, n. 2, p. 323-367, 2000.

SPENCE, Robert. "A Corpus of Student L1-L2 Translations". In: Granger, S.; Hung, J. (eds.), *Learner Corpora, Second Language Acquisition and Foreign Language Teaching*. International Symposium on Computer. Hong-Kong: The Chinese University of Hong Kong, 1998.

UZAR, Rafal S. "A corpus methodology for analyzing translation". *Cadernos de Tradução*, v. 1, n. 9, p. 235-263, jan./jun., 2008, Santa Catarina. Disponível em: https://periodicos.ufsc.br/index.php/traducao/article/view/5988 (acesso em 11 de janeiro de 2016).

10.17771/PUCRio.TradRev.30814

Resumo

O presente artigo apresenta um panorama dos Corpora de Aprendizes de Tradução (CAT) e descreve o processo de construção do CATUERJ a partir da metodologia de pesquisa da dissertação de mestrado de Barbara Ramos (2016), exibindo suas principais características e relacionando com outros corpora de aprendizes de tradução internacionais já existentes. Por último, discutimos limitações e o potencial do CATUERJ – e de CAT em geral – para pesquisas e aplicações acadêmicas e pedagógicas.

Abstract

The article gives a brief overview of Learner Translator Corpora and describes the compilation and design of CATUERJ, an LTC developed at the University of the State of Rio de Janeiro, based on Ramos' (2016) research methodology for her masters' research. Such description includes showing its main features and relating it to other LTC which were designed previously. Finally, we discuss the potential of CATUERJ and other LTC for academic research and applications.