REVISTA CIENTÍFICA
=TRAMA=
ISSN 1981-4674
unioeste

CONSTRUÇÃO DE *CORPUS* BILÍNGUE PORTUGUÊS-INGLÊS COM MARCAÇÃO XML DE ITENS CULTURAIS-ESPECÍFICOS EM TRADUÇÃO INTERLINGUAL

JOÃO GABRIEL CARVALHO MARCELINO (UFSC)[1]

RESUMO: Este trabalho visa a apresentar a utilização da demarcação em eXtensible Markup Language (XML) para anotação de Itens culturais-específicos na construção de um *corpus* bilíngue no par português-inglês. Para tanto, delimitam-se os seguintes objetivos específicos: i) Apresentar os estágios iniciais da construção de um *corpus* bilíngue com marcação XML; ii) Apresentar uma proposta de lógica de anotação de itens culturais-específicos no *corpus*; e iii) Fazer anotações de itens culturais-específicos no *corpus* com marcação XML. Os textos utilizados para a construção do *corpus* apresentado neste trabalho correspondem aos capítulos *Mudança* e *Fabiano*, da obra *Vidas Secas*, de Graciliano Ramos, e suas respectivas traduções por Ralph Edward Dimmick em *Barren Lives*, intituladas *A New Home* e *Fabiano*, respectivamente. Metodologicamente realiza-se um estudo experimental, orientado aos estudos de *corpora* construindo um *corpus* bilíngue, apontando os critérios de anotação, a construção e a orientação do *corpus* para os interesses de pesquisa no campo dos Estudos da Tradução, a partir de um recorte inicial da obra estudada e sua respectiva tradução. Como fundamentação teórica, parte-se dos estudos de Hardie (2014), Franco-Aixelá (2013), Magalhães (2001), entre outros autores que discutem os Estudos da Tradução, Linguística de *corpus* e Itens culturais-específicos. Com a marcação XML busca-se realizar anotações identificando itens culturais-específicos do sertão nordestino para observar o tratamento dado às especificidades no processo de tradução. Os resultados destacam que a utilização da marcação XML possibilita a identificação de padrões na tradução para lidar com itens culturais-específicos do sertão nordestino, permitindo quantificar traduções, estrangeirizações e domesticações, possibilitando tornar a análise mais eficiente.

PALAVRAS-CHAVE: Vidas Secas. Barren Lives. Tradução Interlingual. *Corpus* Bilíngue.

BUILDING A BILINGUAL PORTUGUESE-ENGLISH *CORPUS* WITH XML TAGGING OF CULTURAL-SPECIFIC ITEMS IN INTERLINGUAL TRANSLATION

ABSTRACT: *This paper aims to present the use of eXtensible Markup Language to annotate cultural-specific items in the building of a bilingual* corpus *in the Portuguese-English pair. To this aim, the following specific objectives are defined: i) Present the initial stages of building a bilingual* corpus *with XML markup; ii) Present a proposal for the annotation logic of specific cultural items in the* corpus*; and iii) Annotate cultural-specific items in the* corpus *with XML markup. The texts used for the construction of the* corpus *presented in this paper correspond to the chapters* Mudança *and* Fabiano*, from the literary work* Vidas Secas*, by Graciliano Ramos, and their respective translations by Ralph Edward Dimmick in* Barren Lives*, entitled* A New Home *and* Fabiano*, respectively. Methodologically, an experimental study is made out, oriented to* corpora *studies, building a bilingual* corpus*, pointing out the annotation criteria, construction, and orientation of the* corpus *for the research interests in the field of Translation Studies, from an initial cut of the work studied and its respective translation.*

[1] Doutorando em Estudos da Tradução (PPGET/UFSC), Mestre em Linguagem e Ensino (PPGLE/UFCG), Licenciado em Letras (FASETE). joaogabrielcarvalho@hotmail.com

As a theoretical framework, we use the studies of Hardie (2014), Franco-Aixelá (2013), Magalhães (2001), among other authors who discuss Translation Studies, Corpus Linguistics and Cultural-Specific Items. With XML markup, we seek to make notes identifying specific cultural items of the Northeastern sertão *to observe the treatment given to specificities in the translation process. The results highlight that the use of XML markup makes it possible to identify patterns in translation to deal with specific cultural items of the Northeastern* sertão*, allowing the quantification of translations, foreignizations and domestications, making the analysis more efficient.*

KEYWORDS: *Vidas Secas. Barren Lives. Interlingual Translation. Bilingual Corpus.*

INTRODUÇÃO

A construção de *corpus* para pesquisas em tradução lida com diferentes problemáticas considerando as diferentes perspectivas de abordagens possibilitadas pelo campo disciplinar. Nesse sentido, a anotação de metadados através do *eXtensible Markup Language* (doravante XML) permite inserir em um *corpus* de pesquisa informações complementares para a construção da análise proposta pelas pesquisas, bem como comentários e notas explicativas sobre o *corpus*. A lógica das etiquetas utilizadas para anotação em XML de natureza simples possibilita que essa linguagem seja aplicada em diferentes campos da Linguística, como nos Estudos da Tradução.

Este trabalho tem como objetivo apresentar a utilização de marcação XML para anotação de Itens culturais-específicos na construção de um *corpus* bilíngue no par português-inglês, utilizando como fonte dos dados a obra Vidas Secas (1938), de Graciliano Ramos, e sua respectiva tradução, intitulada *Barren Lives* (1964), por Ralph Edward Dimmick. A obra de Graciliano Ramos narra a vida de Fabiano, Sinha Vitória, os dois filhos e a cachorra Baleia, grupo de retirantes que atravessa o sertão alagoano e se instala em uma pequena roça enquanto enfrenta o ciclo da seca. A narrativa, inserida no Neorrealismo do romance de 30, apresenta elementos particulares à região nordeste que vão além do ciclo de vida dos retirantes, mas também apresenta elementos sociais, culturais e ambientais da região brasileira.

A tradução para a língua inglesa de *Vidas Secas* (1938) apresentada nesta pesquisa foi publicada pela *University of Texas Press* em 1964 e republicada em 1999. Traduzida por Ralph Edward Dimmick, na obra intitulada *Barren Lives*, chama atenção a presença de elementos particulares ao sertão no processo de tradução, apresentados desde a escolha do título da obra e de capítulos, até o tratamento dado à vegetação da caatinga, bioma particular do Brasil.

Observar a tradução de *Vidas Secas* para a língua inglesa a partir do português brasileiro pode evidenciar estratégias de tradução que domesticam ou estrangeirizam o texto, ao mesmo tempo que possibilita refletir sobre as escolhas tradutórias e seus impactos no texto de chegada. A construção de *corpora* de pesquisas no campo dos Estudos da Tradução permite a discussão de métodos de coleta e limpeza, identificação e marcação que variam de acordo com os interesses de pesquisa. Neste trabalho, portanto, busca-se identificar itens culturais-específicos associados ao sertão nordestino, especificamente no bioma caatinga identificado na obra.

Portanto, para a realização deste trabalho apresentam-se os objetivos específicos: i) Apresentar os estágios iniciais da construção de um *corpus* bilíngue com marcação XML; ii) Apresentar uma proposta de lógica de anotação de Itens culturais-específicos no *corpus*; e iii) Fazer anotações de itens culturais-específicos no *corpus* com marcação XML. Tendo em vista a dimensão deste trabalho, utiliza-se como *corpus* para marcação XML e anotação de itens culturais-específicos os dois primeiros capítulos da obra *Vidas Secas*, intitulados "Mudança" e "Fabiano" e suas respectivas traduções em *Barren Lives*, intituladas "*A new home*" e "Fabiano".

Este estudo se fundamenta nas pesquisas de Hardie (2014), que explora a marcação XML de *corpora* linguísticos de maneira simplificada; Franco-Aixelá (2013), tendo em vista os itens culturais-específicos; e Magalhães (2001), considerando as pesquisas em *corpora* nos Estudos da Tradução, entre outros autores cujas pesquisas são consideradas pertinentes ao estudo.

Este artigo se organiza em quatro seções, a primeira apresenta a fundamentação teórica, discutindo o campo disciplinar dos Estudos da Tradução e sua interface com a Linguística de *Corpus*, pensando a compilação de *corpora* para a análise de traduções; a seção se encerra com a discussão sobre Itens culturais-específicos, considerando a natureza do modelo de anotação proposto. A segunda seção apresenta a metodologia, explicitando a anotação de *corpus* utilizando a linguagem XML para inserir metadados que demarcam a presença de itens culturais-específicos no *corpus* tradutório, descrevendo as etapas e estratégias para a realização das anotações. A terceira seção apresenta os resultados da anotação nos *corpora* observando a distribuição de etiquetas no texto anotado, o formato que se apresenta e refletindo sobre as possibilidades de otimização da lógica de anotação. Por fim, apresentam-se as considerações finais do artigo.

FUNDAMENTAÇÃO TEÓRICA

Tendo em vista que a Tradução pode ser classificada como uma atividade comunicativa que busca transpor uma mensagem entre sistemas linguísticos ou sistemas semióticos distintos, e é realizada pensando na equivalência da mensagem transposta (JAKOBSON, 2004), explorar a tradução enquanto processo ou produto, abre uma gama de possibilidade de observação de como essa transposição é realizada, trazendo à tona questões que orbitam o processo, o produto e até mesmo o próprio tradutor. O campo disciplinar dos Estudos da Tradução, por sua vez, tem se dedicado a observar o processo e o produto tradutório em diferentes contextos. A reflexão sobre a transposição de textos entre diferentes línguas e culturas através do campo permite discutir os mecanismos e estratégias utilizadas por tradutores em diferentes perspectivas.

A ocorrência da tradução em três categorias, como Jakobson (2004)[2] caracteriza em Intralingual – dentro de um sistema linguístico; Interlingual – entre dois sistemas linguísticos distintos; e Intersemiótica – entre sistemas semióticos distintos, permite que a concepção de tradução vá além da tradução interlingual, estendendo-a a diferentes contextos e delineando direções de observação da tradução em meios semióticos e linguísticos distintos. Essa percepção possibilita refletir e problematizar a tradução em diferentes aspectos, a depender de cada objeto traduzido estudado.

As pesquisas no campo dos Estudos da Tradução se dedicam à descrição, análise e teorização dos processos, produtos e contextos do ato tradutório (WILLIAMS; CHESTERMAN, 2010). Diante dessas possibilidades em que o campo tem dedicado suas reflexões, a observação da construção de *corpus* em uma pesquisa de Tradução possibilita compreender a importância da relação entre Linguística de *Corpus* e Estudos da Tradução, ainda considerando as ferramentas de compilação, edição e análise de textos que podem ser aplicadas na construção e análise de *corpora* de pesquisa. Portanto, nesta seção apresenta-se a fundamentação teórica utilizada para embasar a construção de *corpus* bilíngue no par português-inglês com marcação XML para anotação de Itens culturais-específicos,

[2] Apesar de o texto de Jakobson ter sido publicado em 1959, utiliza-se a definição de autor nesse trabalho por ser pertinente à discussão de tradução interlingual.

caracterizando Linguística de *Corpus*, sua relação com os Estudos da Tradução e Itens culturais-específicos.

LINGUÍSTICA DE *CORPUS* E ESTUDOS DA TRADUÇÃO

McEnery e Hardie (2012) descrevem a Linguística de *Corpus* como a área da linguística focada no desenvolvimento de métodos e procedimentos de estudo da linguagem. Para a elaboração desses métodos, linguistas utilizam os *corpora*, conjuntos finitos de textos que podem ser tomados como corpos de análise. Esses conjuntos de textos, ao longo do desenvolvimento da Linguística, têm sido construídos com base em textos em diferentes modalidades, desde os analógicos até os virtuais.

O desenvolvimento de tecnologias computacionais permitiu que pesquisadores da Linguística elaborassem e digitalizassem *corpora* de trabalho para diferentes fins. Essa possibilidade de trabalho com o texto virtual possibilitou que os estudos linguísticos observassem de maneira mais rápida regularidades na língua (VIEIRA; LIMA, 2001). O desenvolvimento tecnológico permitiu então que os *corpora* analógicos fossem digitalizados através de diferentes meios, desde a digitalização manual e a utilização de Inteligência Artificial através de OCR[3] para digitalização de textos analógicos, ou a criação de *corpora* nato-digitais. Esse avanço tornou cada vez mais recorrente a utilização de *corpora* como ferramenta, tendo em vista a variada gama de observações possíveis sobre a linguagem natural através de ferramentas computacionais capazes de interpretar os textos (ALUÍSIO; ALMEIDA, 2021).

Os estudos em Linguística de *Corpus* são aliados aos Estudos da Tradução, permitindo que através dos *corpora* respostas para as questões de pesquisa sejam encontradas (WILLIAMS; CHESTERMAN, 2010). Nesse sentido, a construção e sistematização de *corpora* de pesquisa torna possível que os dados sejam compilados, organizados, alinhados e comparados visando às categorizações propostas em pesquisas no campo do Estudos da Tradução, tendo em vista o volume de informações que podem possuir (PEARSON, 2003). Esse volume de dados pode ser utilizado para comparar Texto de Partida e Texto de Chegada[4] (como neste trabalho) ou textos traduzidos de um mesmo autor, tradutor, sistema ou período, entre outras possibilidades.

Portanto, é importante considerar que há dois tipos de *corpora* textuais que podem ser utilizados nos Estudos da Tradução: i) *corpora* comparáveis – monolíngues e elaborados com textos originais, são utilizados para observar o comportamento linguístico; e ii) *corpora* paralelos – possuem textos e suas traduções, assim como evidências de linguagem produzidas em ambiente mono-, bi- e multilíngue (PEARSON, 2003). Esses *corpora* podem ser utilizados, como observa-se nesta pesquisa, nos estudos descritivos da tradução, assim como na crítica de tradução, ou até mesmo para diferentes fins. A utilização para descrever processos de tradução pode evidenciar escolhas tradutórias, apagamentos e omissões, como o tradutor lidou com termos que podem ser problemáticos para a tradução devido à falta de equivalência, como no caso dos Itens culturais-específicos (FRANCO-AIXELÁ, 2013).

Os *corpora* textuais, sejam comparáveis ou paralelos, podem ser elaborados utilizando diferentes ferramentas de edição de texto, para a posterior análise utilizando processadores de *corpora* (MAGALHÃES, 2001). Diante da variedade de ferramentas de edição de texto, é importante considerar a utilização de ferramentas que utilizem codificação compatível com

[3] *Optical Character Recognition*, tecnologia de reconhecimento de caracteres que consegue extrair textos de imagens digitalizadas.
[4] Utiliza-se a denominação Texto de Partida para o texto na língua fonte (português brasileiro) e Texto de Chegada para o texto na língua alvo (inglês).

diferentes sistemas operacionais, visando ao acesso e à exploração dos *corpora* por outros pesquisadores e estudantes interessados. Bem como para a utilização de ferramentas de processamento de *corpora* que permitem aos pesquisadores encontrar padrões na língua analisada em suas pesquisas, a partir dos *corpora* criados em editores de texto comuns, anotados ou não.

Com o desenvolvimento dos sistemas de computação, as possibilidades de coleta, construção e anotação de *corpora* aumentaram e tornaram-se mais eficientes (CEA; ÁLVAREZ-DE-MON; PAREJA-LORA; PLAZA-ARTECHE, 2002). Desse modo, os sistemas de anotação de *corpora* permitiram que pesquisadores insiram metadados em um *corpus* visando a tornar mais rápida e eficiente a busca e análise de informações sobre os dados do *corpus.* Essas anotações carregam uma variabilidade e adaptabilidade alinhadas aos tipos de pesquisa em que são aplicadas, podendo descrever ou identificar diferentes aspectos da língua.

É necessário, portanto, considerar que os *corpora* possuem natureza mutável, uma vez que sua concepção não é permanente e a atualização é constante (JOHANSSON, 2003). Essa qualidade dos *corpora* também evidencia que as pesquisas não são estáticas, então não é possível esperar que, com o avanço da pesquisa, o *corpus* não sofra alteração, seja pelas hipóteses serem comprovadas ou não, ou em razão da viabilidade ou não de aspectos de pesquisa. Isso também se aplica às anotações realizadas em um *corpus*, sendo necessário revisar e ajustar a lógica de anotação proposta à medida que a pesquisa e a análise se desenvolvem.

ITENS CULTURAIS-ESPECÍFICOS

A tradução, enquanto processo de comunicação, lida com o objetivo de passar um conhecimento de um texto original para um leitor estrangeiro, buscando realizar escolhas que sejam próximas em sentido do original (LEVÝ, 2004). Entretanto, esse processo esbarra em elementos linguísticos que podem não ter correspondência entre as línguas e culturas envolvidas, dada a natureza particular dos contextos envolvidos no processo tradutório. Nesse sentido, as escolhas do tradutor são orientadas pelo texto de partida, e como ele lida com as particularidades da cultura fonte indica as tendências adotadas na tradução.

Tais tendências posicionam o texto traduzido entre estrangeirização, quando o texto conserva, em sua maioria, aspectos da cultura e do texto fonte; ou domesticação, quando o texto é escrito de maneira a se colocar na cultura e língua de chegada como se fosse escrito em tal modelo (VENUTI, 1995). Aspectos que Berman (2013) também aponta ao discutir a tradução Etnocêntrica, que trata como negativo tudo aquilo que não é nativo da cultura e língua de chegada, intencionalmente apagando o estrangeiro para que o texto traduzido seja transparente na língua de chegada.

Nesse caminho, a tradução enquanto processo mediado de comunicação entre línguas (REISS, 2004) tem na figura do tradutor o mediador da comunicação. Desse modo, as escolhas sobre os caminhos percorridos no processo de tradução indicam as tendências adotadas pelo tradutor, sejam por sua escolha direta, ou pela influência do sistema de patronagem (LEFEVRE, 1992) que define 'o que', 'como' e 'por que' traduzir. As tendências adotadas, ao esbarrarem com elementos particulares do contexto e língua de partida, evidenciam que cada comunidade linguística possui particularidades de uso das línguas para diferentes propósitos que podem aproximar ou distanciar as culturas e línguas envolvidas no processo tradutório (FRANCO-AIXELÁ, 2013).

Esse conflito entre as referências para a tradução entre línguas e culturas, para Franco-Aixelá (2013), é o que caracteriza um Item cultural-específico. Um item que, quando traduzido

de uma língua fonte para uma língua alvo, não possui uma equivalência determinada por uso, ideologia, frequência, entre outros, para que seja representado de maneira equivalente (FRANCO-AIXELÁ, 2013). O direcionamento para a observação de Itens culturais-específicos em *Vidas Secas* e sua respectiva tradução, *Barren Lives*, parte do conflito em traduzir elementos particulares ao sertão nordestino, considerando que a narrativa se passa na Caatinga, bioma que ocupa 11% do território brasileiro, tem a vegetação adaptada para períodos de longa estiagem e é exclusivo do Brasil[5]. Essas características da Caatinga posicionam o texto em situação conflituosa para a tradução de elementos naturais e culturais para outros idiomas.

As particularidades da caatinga, como os nomes vernaculares de elementos da vegetação, profissões, títulos e superstições, podem constituir problemas de opacidade para a língua e cultura alvo, o que permite classificar tais elementos como Itens culturais-específicos, pois podem ser apresentados como nomes próprios carregados – que possuem uma margem de indeterminação maior e possuem uma motivação para tal; e os nomes convencionais – que não possuem tal motivação e podem ser repetidos sem problema (FRANCO-AIXELÁ, 2013). Diante do exposto, apresentam-se na seção a seguir as orientações adotadas para a anotação de itens culturais-específicos no *corpus* bilíngue, observando, para a realização deste trabalho, elementos da vegetação, nomes carregados e nomes convencionais.

METODOLOGIA

Esta seção foi elaborada objetivando apresentar uma proposta de lógica para a realização de anotação de itens culturais-específicos em *corpora* traduzidos. Para a construção de *corpus* proposta neste trabalho, utiliza-se como base as orientações de Hardie (2014) acerca da utilização de anotação em *eXtensible Markup Language* (XML). O autor explora a linguagem de anotação, utilizada na programação para inserir metadados em programações com diferentes finalidades, como ferramenta de anotação de metadados de análise para pesquisas em linguística, dada sua natureza simples.

A extensão .xml é utilizada para construção de *Translation Memories* aplicadas em ferramentas de *MachineTranslation* (MOORKENS, 2013), assim como para a marcação e localização de elementos de interesse de pesquisas em tradução, com ferramentas como o AntConc e o LancsBox® (ZANETTIN, 2012; BREZINA; WEILL-TESSIER; MCENERY, 2021), ferramentas de análise linguística gratuitas que possibilitam realizar buscas através de expressões regulares, realizar o Alinhamento e Concordanciamento de diferentes *corpora*, mono-, bi- e plurilíngue.

Para a criação e anotação do *corpus*, utilizam-se como critérios para a elaboração as orientações apontadas por Hardie (2014) de que o *corpus* deve ser simples e de fácil leitura, escrito de maneira objetiva e clara e executável por qualquer linguista. Desse modo, os dados apresentados nesta seção apontam como as *tags*[6] (em português brasileiro, "etiquetas") foram elaboradas, quais valores foram atribuídos, como as *tags* estão alinhadas e os critérios de criação das anotações e identificação do *corpus*.

Hardie (2014) aponta que XML é um formato que permite que as informações de metadados sejam inseridas em *tags* representadas entre '<>'. Tudo o que está no espaço entre

[5] Características da Caatinga. Fundação Joaquim Nabuco. Disponível em: https://www.fundaj.gov.br/index.php/conselho-nacional-da-reserva-da-biosfera-da-caatinga/9193-saiba-quais-sao-as-caracteristicas-da-caatinga. Acesso em: setembro de 2021.

[6] O termo "*tags*" se refere às etiquetas em que estão inseridas as informações sobre os elementos anotados no *corpus*. O vocábulo vem da área da computação e, apesar de possuir uma tradução, a opção por utilizar o termo em língua inglesa considera a divulgação científica do trabalho, o que possibilita a apresentação de resultados que possuem o termo em mecanismos de busca.

<> e </> classifica-se como uma marcação, sendo possível marcar sentenças, períodos, intervenções de personagens e o próprio espaço do texto. Tendo em vista o estilo e a brevidade na escrita de Graciliano Ramos, adota-se no *corpus* apresentado como menor unidade de texto a sentença, que pode ser demarcada por <s> e </s>, ou nativamente após a abertura das *tags* <text> e </text>. Desse modo, as unidades de sentenças são apresentadas da seguinte maneira:

Tabela 1: Unidade de sentença do texto anotado.

<s> Os infelizes tinham caminhado o dia inteiro, estavam cansados e famintos.</s> <s> Ordinariamente andavam pouco, mas como haviam repousado bastante na areia do rio seco, a viagem progredira bem três léguas.</s>
<s> The drought victims had been walking all day; they were tired and hungry.</s> <s> Generally they did not get very far, but after a long rest on the sands of the riverbed they had gone a good three leagues.</s>

Fonte: *Corpus* anotado de *Vidas Secas* (RAMOS, 2018, p. 9) e *Barren Lives* (RAMOS; DIMMICK, 1999, p. 3).

A marcação da menor unidade da sentença possibilita que posteriormente os dados sejam utilizados de maneira mais organizada para a identificação e realização da análise da tradução, orientada pelos Estudos Descritivos da Tradução (LABERT; VON GORP, 2011). No contexto de uma tradução como a de *Vidas Secas*, a anotação pode ser utilizada visando a identificar como a tradução lida com elementos característicos do sertão nordestino e da caatinga no movimento de transposição do português brasileiro para a língua inglesa. Isso permite, através da anotação, obter dados quantitativos organizados para a posterior análise.

Tendo em vista o objetivo de realizar anotações de itens culturais-específicos no *corpus* com marcação XML, e diante do que foi apontado nas seções anteriores, Hardie (2014) aponta que é possível atribuir valores às anotações. Na construção do *corpus* apresentada utiliza-se como elaboração de atributos o seguinte perfil <tag atributo=”valor”> </tag>. No espaço entre aspas são atribuídos os valores que são classificados abaixo como categorias de Itens culturais-específicos, considerando o conceito apontado por Franco-Aixelá (2013), de itens que podem ser problemáticos para a realização da tradução:

i) Nomes vernaculares da Vegetação: Catingueira, Juazeiro etc.
ii) Nomes vernaculares da Fauna: Preá, Rês etc.
iii) Nomes próprios carregados: Tomás da Bolandeira, Baleia etc.
iv) Nomes próprios convencionais: Fabiano, Sinha Vitoria etc.
v) Nomes de produtos e itens: Creolina, Aió etc.
vi) Bioma: Caatinga.[7]

Com essas classificações de itens que podem ser problemáticos para a realização da tradução, as categorias de valores atribuídos para *tags* no *corpus* apresentado nessa pesquisa correspondem a:

i) Vegetação: nomes vernaculares de plantas e o bioma;
ii) Fauna: nomes vernaculares de animais;
iii) Nome próprio carregado.
iv) Nome próprio convencional
v) Nomes de produtos.

[7] Apesar de só haver um bioma mencionado na narrativa, sua categorização vale menção tendo em vista a exclusividade e a derivação no caso do nome Catingueira.

vi) Conservação: o termo foi conservado na LC[8].
vii) Apagamento: o termo foi apagado na LC.
viii) ICE: para item cultural-específico como objetos.

As categorias de valores indicam os seguintes comentários anotados no *corpus* no formato XML:

Tabela 2: Anotações utilizadas no *corpus*

<nota comentario="vegetacao"></nota>	<nota comentario="produto"></nota>
<nota comentário="fauna"></nota>	<nota comentario="apagamento"></nota>
<nota comentário="Nome proprio carregado"></nota>	<nota comentario="conservacao"></nota>
<nota comentario="nome proprio convencional"></nota>	<nota comentario="ICE"></nota>

Fonte: Elaborada pelo autor (2023).

O *corpus* coletado é nato-digital, portanto, a conversão de formato não foi necessária, apenas a limpeza e organização das sentenças de trabalho. A transferência do texto para a ferramenta de edição utilizada, o aplicativo *Sublime text*[9], tornou necessária apenas a limpeza de códigos referentes a imagens não transferidas para o formato do *corpus*. As anotações foram realizadas utilizando o aplicativo, que permitiu a busca e substituição através de expressões regulares para limpeza ou aplicação de etiquetas, como pode ser observado na Figura 1:

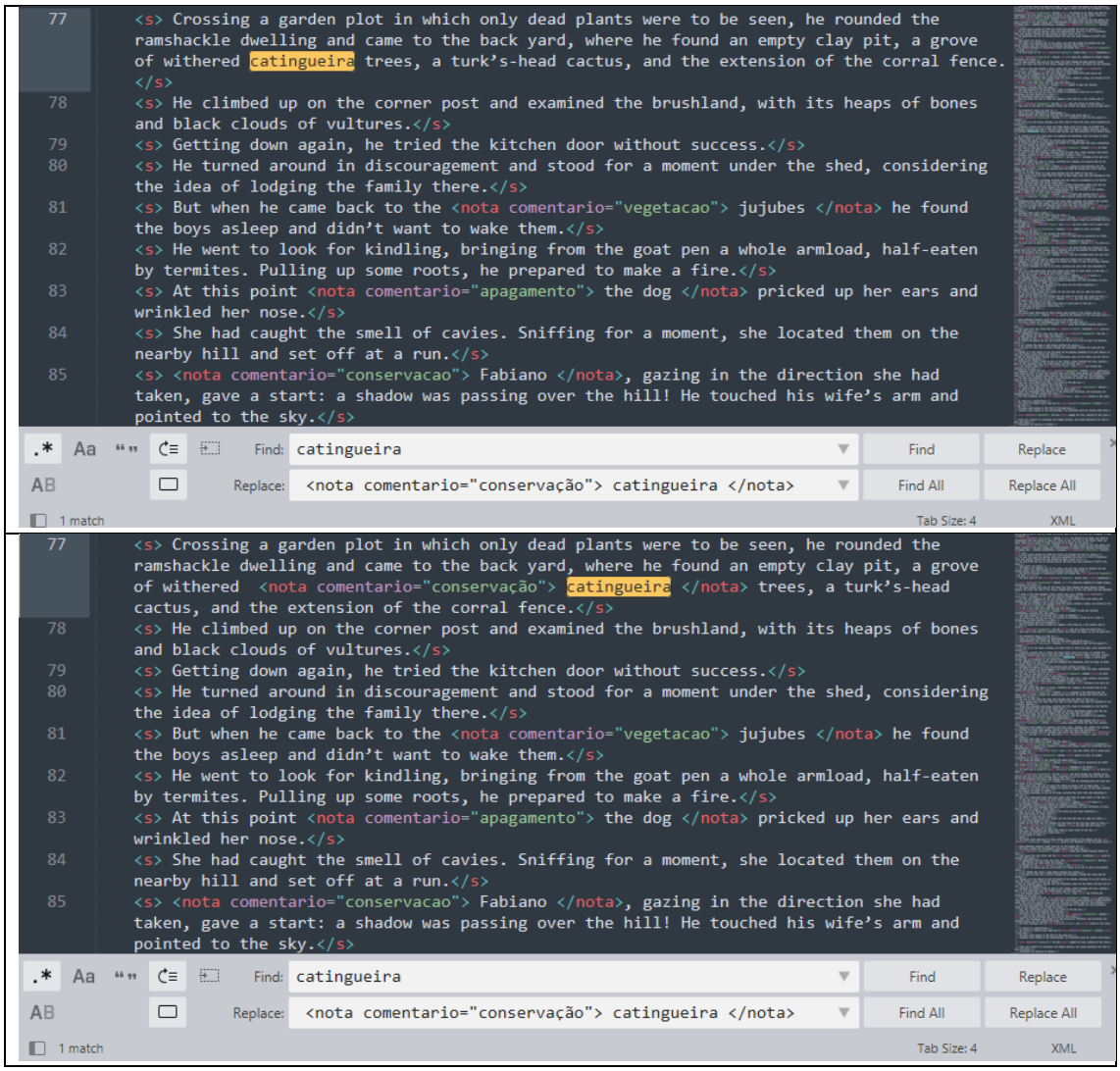
Figura 1: Busca e substituição para limpeza

Fonte: Elaborada pelo autor (2023).

A ferramenta de busca e substituição permite eliminar as ocorrências de códigos como <0x0c>, que corresponde às imagens das ilustrações presentes no arquivo original em PDF que não são legíveis para o formato do *corpus*. Essa utilização permite otimizar a limpeza do *corpus*, porém, é importante ressaltar que a ferramenta de busca e substituição não exclui o olhar humano sobre o texto. A mesma ferramenta pode ser utilizada para a aplicação de etiquetas, como está apresentado na Figura 2:

Figura 2: Busca e substituição para anotação

[8] Língua de Chegada.
[9] Disponível em: https://www.sublimetext.com/. Acesso em: junho de 2021.

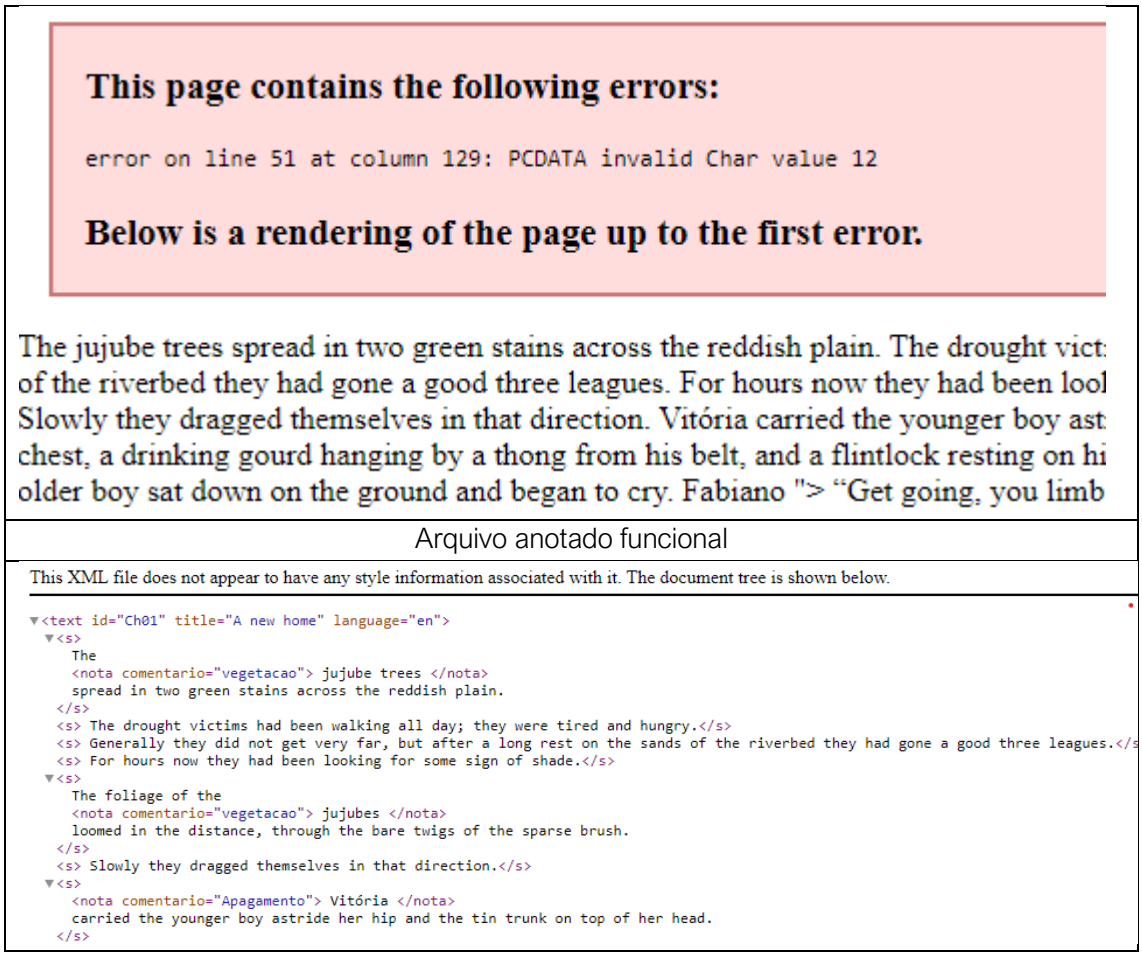


Fonte: Elaborada pelo autor (2023).

A possibilidade de localizar e substituir os termos que já foram reconhecidos como recorrentes no texto possibilita otimizar, por exemplo no texto em inglês, a localização de conservação de termos da língua de partida no texto de chegada, evidenciando a necessidade constante de conferir a etiquetagem para identificar se há erros ou etiquetas sem fechamento. A verificação pode ser realizada abrindo o arquivo no navegador de internet, que mostrará as mensagens a seguir:

Figura 3: Mensagem de erro *versus* confirmação de funcionamento da etiquetagem

Mensagem de erro



Arquivo anotado funcional

Fonte: Elaborada pelo autor (2023).

A mensagem de erro indica a linha e coluna em que há erro, facilitando o processo de correção para verificar novamente o funcionamento. Com a etiquetagem funcional, o *corpus* pode ser explorado considerando a busca pelos metadados ou não.

Na seção a seguir, estão apresentados os resultados da anotação do *corpus*, desde a identificação e separação de seções para anotação, até a marcação de itens culturais-específicos.

RESULTADOS

Apresentando os estágios iniciais da construção de um *corpus* bilíngue com marcação XML, esta seção mostra como as anotações em .xml realizadas utilizando o editor de textos *Sublime text* ficam visíveis no *corpus*. A primeira marcação realizada corresponde à abertura do texto para que os parágrafos fossem alinhados, utilizando a nota <text id="" title="" language="">, para identificar o nome do arquivo com *id* (*ch* correspondendo a *chapter*, e o número do capítulo), *title* para os títulos dos capítulos, *language* para a identificação das línguas de partida e de chegada (*pt-br*, para o português brasileiro; e *en* para inglês):

Tabela 3: Etiquetas de identificação dos textos

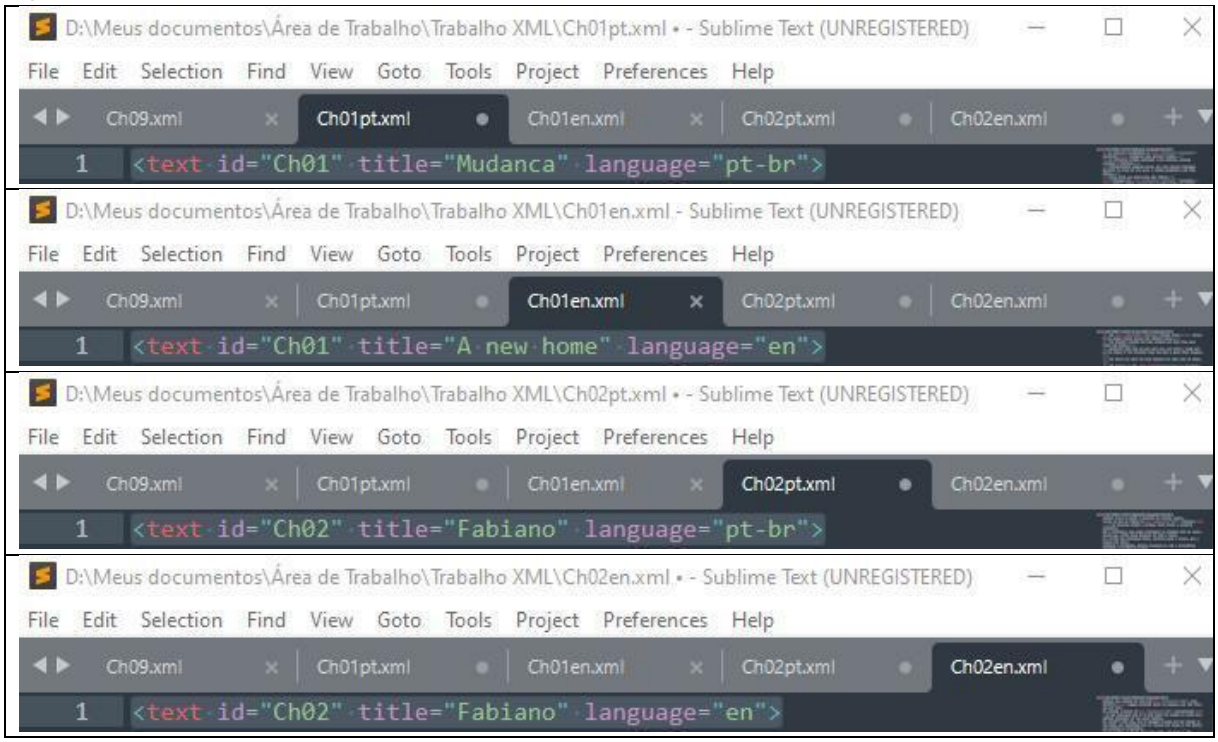
Identificação dos textos	
Português	Inglês
<text id="Ch01" title="Mudanca" language="pt-br">	<text id="Ch01" title="A new home" language="en">

<text id="Ch02" title="Fabiano" language="pt-br">	<text id="Ch02" title="Fabiano" language="en">

Fonte: Elaborado pelo autor (2023).

Desse modo, os títulos dos arquivos em extensão .xml são numerados com a ordem de aparição e tornam mais fácil a localização para o tratamento paralelo dos arquivos capítulo por capítulo. Além disso, a identificação apontada na Tabela 3 também influencia o nome dos arquivos identificados nas imagens apresentadas na Figura 4:

Figura4: Textos identificados no aplicativo *Sublime Text*

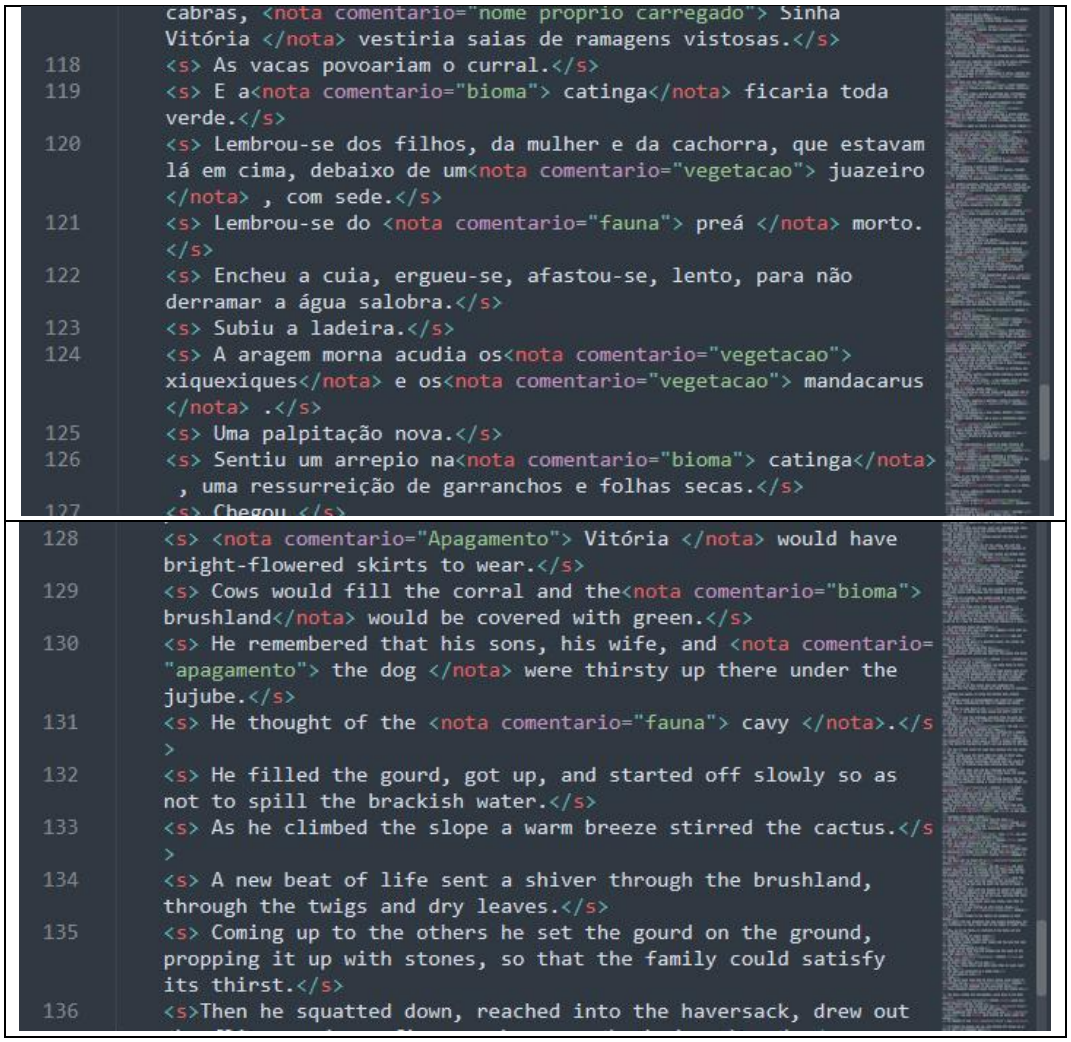


Fonte: Capturas de tela do aplicativo *Sublime Text* realizadas pelo autor (2023).

As abas identificam as unidades do *corpus* como Ch01pt.xml e Ch02pt.xml para os capítulos na língua de partida correspondentes a *Mudança* e *Fabiano*; e Ch01en.xml e Ch02en.xml para as traduções *A new home* e *Fabiano*, permitindo que os arquivos sejam acessados de maneira organizada numericamente e em ordem crescente, o que poderia ser prejudicado pela identificação através de nomes de capítulos. Essa opção se mostrou viável para um *corpus* baseado em poucos capítulos, entretanto, para *corpora* mais extensos, a criação de um único arquivo com a localização do título pode ser mais viável.

A Figura 5 apresenta a ocorrência de anotações no *corpus* em língua portuguesa, utilizando as *tags* apresentadas na seção anterior:

Figura 5: *Corpus* anotado do capítulo 1 – Mudança/*A new home*

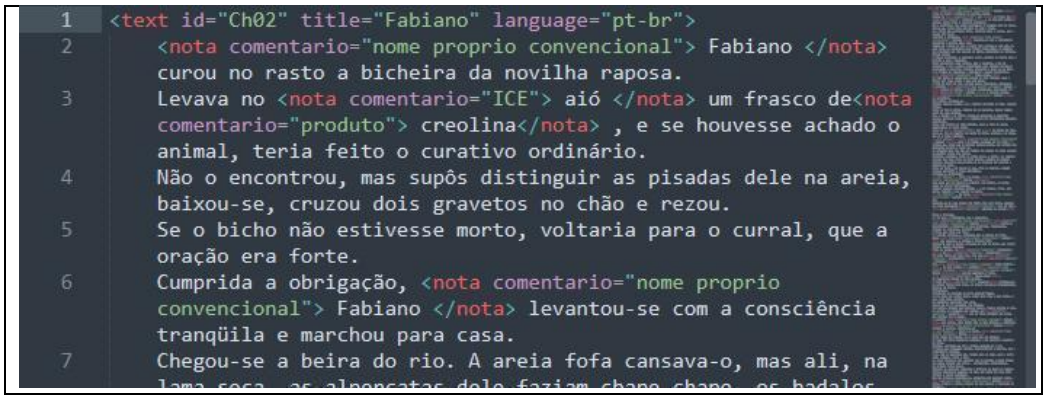


Fonte: Capturas de tela realizadas pelo autor (2023).

A anotação no *corpus* correspondente ao primeiro capítulo foi realizada manualmente, havendo as anotações <s> e </s> para início e fim da sentença e a localização de itens para anotação no texto de partida e as soluções realizadas no texto de chegada. Entretanto, a realização da anotação de sentença manualmente torna o processo mais lento.

Na realização da anotação do capítulo 2 apresentada na Figura 6, a marcação das sentenças foi realizada automaticamente, havendo somente a abertura do texto:

Figura 6: *Corpus* anotado do capítulo 2 – Fabiano



```
<text id="Ch02" title="Fabiano" language="en">
    Fabiano was seeking to cure the<nota comentario="nome"> wily
    heifer's</nota> maggot-infested sores by praying over the trail
    she had left.
    He carried a bottle of<nota comentario="ICE"> disinfectant</nota
    > in his haversack and if he had found the animal he would have
    applied treatment in the usual fashion.
    He couldn't find her, but he thought he made out her tracks in
    the sand, and squatting down he crossed two twigs on the ground
    and uttered an invocation.
    If the animal was still alive she would come back to the
    corral, for his prayer possessed great power.
    His obligation fulfilled, Fabiano straightened up with a clear
    conscience and walked in the direction of the house, taking a
```

Fonte: Capturas de tela realizadas pelo autor (2023).

Com a marcação automática dos parágrafos, há uma otimização no processo de anotação, o texto nativamente assume a posição de cascata, livrando o pesquisador de fazer a entrada de etiquetas para cada linha. Isso permite otimizar a anotação do *corpus* considerando os elementos observados e a lógica de anotação adotada na pesquisa. Essa anotação pode ser realizada manualmente ou utilizando a função Localizar/Substituir, possibilitando a localização e substituição do termo específico pelo mesmo termo anotado ou pela identificação de outros elementos buscados. Entretanto, a opção de localização não substitui a conferência manual para a identificação de apagamentos.

CONSIDERAÇÕES FINAIS

Diante dos dados apresentados, a utilização de anotação de *corpus* utilizando o formato XML para identificação de itens culturais-específicos se mostra promissora para a realização de pesquisas nos Estudos da Tradução. O formato de anotação sugerido por Hardie (2014) possibilita a realização de anotações de metadados alinhadas aos objetivos de pesquisas em diferentes campos dos Estudos Linguísticos. A possibilidade de aplicação nos Estudos da Tradução para anotar itens culturais-específicos permite que a localização de tais itens seja realizada com maior facilidade em pesquisas que observam como termos e expressões são tratadas no processo tradutório.

As anotações realizadas no *corpus* apresentado neste trabalho permitem que a localização das ocorrências de itens culturais-específicos seja realizada de maneira mais eficiente. Isso torna possível, consequentemente, que as estratégias adotadas para lidar com esses itens que podem causar opacidade na tradução sejam quantificados para complementar a análise do produto da tradução e refletir sobre o processo tradutório. Tais dados quantificados podem auxiliar na identificação da orientação tomada na tradução para lidar com o estrangeiro, observando tendências domesticadoras ou estrangeirizantes, permitindo que as pesquisas apresentem esses dados estatisticamente através da contagem dos metadados.

Desse modo, com o desenvolvimento deste trabalho outros questionamentos surgem que podem ser discutidos sobre o processo de anotação. Dentre eles, como anotar tendências deformadoras adotadas na tradução, apagamentos sobre gênero, ou para observações mais específicas como de terminologias, diminutivos, aumentativos, entre outros. Essas perguntas permitem compreender que cada pesquisa, considerando suas particularidades, possibilita o estabelecimento de uma lógica própria de anotação.

Por fim, busca-se, com esse experimento, contribuir com as pesquisas no campo dos Estudos da Tradução alinhadas aos estudos em Linguística de *Corpus,* evidenciando a natureza interdisciplinar do campo.

Recebido em: 25/05/2022
Revisões requeridas em: 15/05/2023
Aceito em: 15/06/2023

REFERÊNCIAS

FRANCO-AIXELÁ, Javier. Itens culturais-específicos Em Tradução. **In-Traduções**, Florianópolis, v. 5, n. 8, p. 185-218, Jan./Jun., 2013.

ALUÍSIO, Sandra Maria; ALMEIDA, Gladis Maria de Barcellos. O que é e como se constrói um *corpus*? Lições aprendidas na compilação de vários *corpora* para pesquisa linguística. **Calidoscópio**, [S. l.], v. 4, n. 3, p. 156–178, 2021. Disponível em: <http://revistas.unisinos.br/index.php/calidoscopio/article/view/6002>. Acesso em: 1 nov. 2021.

BERMAN, Antoine. Tradução etnocêntrica e tradução hipertextual. In: **A tradução e a letra ou o albergue do longínquo.** 2. Ed. Tubarão: Copiart; Florianópolis: Pget/Ufsc, 2013, P. 37-66.

BREZINA, V., WEILL-TESSIER, P., MCENERY, A. **LancsBox**. Versão 5.x. Lancaster, 2020. Disponível em: http://corpora.lancs.ac.uk/lancsbox.

CEA, Guadalupe Aguado de; ÁLVAREZ-DE-MON, Inmaculada; PAREJA-LORA, Antonio; PLAZA-ARTECHE, Rosario. RDF(S)/XML Linguístic annotation of semantic web pages. NLPXML '02: Proceedings of the 2nd workshop on NLP and XML, V. 17 September 2002, p. 1-8. Disponível em: https://doi.org/10.3115/1118808.1118809 Acesso em maio de 2023.

HARDIE, A. Modest XML for *corpora*: not a standard, but a suggestion. **Icame Journal,** Warsaw, v. 38, n. 1, p. 73–103, 2014. DOI: https://Doi.Org/10.2478/Icame-2014-0004

JOHANSSON, Stig. Reflections on *corpora* and their uses in cross-linguistic research. In: ZANETTIN, Federico; BERNARDINI, Silvia; STEWART, Dominic (Eds.). ***Corpora*in translator education**. United Kingdom: St. Jerome Publishing, 2003, p. 135-144.

LEFEVERE, André. **Translation, rewriting and the manipulation of literary fame**. London, New York: Routledge, 1992.

LEVÝ, Jiří. Translation as adecision process. In: VENUTI, Lawrence. **The translation studies reader**. New York: Routledge, 2004, p. 148-159.

MAGALHÃES, Célia M. Pesquisas textuais/discursivas em tradução: o uso de *corpora*. In: PAGANO, Adriana Silvina (Org.). **Metodologias de pesquisa em tradução**. Belo Horizonte: Faculdade De Letras, Ufmg, 2001.

McENERY, T.; HARDIE, A. ***Corpus* linguistics**: method, theory and practice. Cambridge University Press: Cambridge, 2012.

MOORKENS, Joss. The role of metadata in translation memories. In: PELLATT, Valerie. **Text, extratext, metatext and paratext in translation**. Cambridge: Cambridge Scholars Publishing Editors, 2013, p.79-90.

PEARSON, Jennifer. Using parallel texts in the translator training environment. In: ZANETTIN, Federico; BERNARDINI, Silvia; STEWART, Dominic (Eds.). ***Corpora* in translator education**. United Kingdom: St. Jerome Publishing, 2003, p. 15-24.

RAMOS, Graciliano. **Barren Lives**. Tradução De Ralph Edward Dimmick. Usa: University Of Texas Press, 1999.

RAMOS, Graciliano. **Vidas Secas**.139 ed. Rio De Janeiro: Record, 2018.

REISS, Katharina. Type, kind, and individuality of text: decision making in translation. In: VENUTI, Lawrence. **The translation studies reader**. New York: Routledge, 2004, P. 160-171.

ROMAN, Jakobson. On Linguistic Aspects of Translation. In: VENUTI, Lawrence (Ed.). **The translation studies reader**. London/New York: Routledge, 2004, p.113-118.

VIEIRA, Renata; STRUBE DE LIMA, V. L. Lingüística Computacional: princípios e aplicações. In: SBC - Jornadas de Atualização em Inteligência Artificial (JAIA), 2001, Fortaleza. **Anais** […]. Fortaleza, 2001, v. 3, p. 47-86. Disponível em: https://www.inf.pucrs.br/linatural/Recursos/jaia-2001.pdf. Acesso em Maio de 2023.

WILLIAMS, Jenny; CHESTERMAN, Andrew. **The map**: A Beginner's Guide To Doing Research In Translation Studies. Manchester, Uk & Kinderhook: St. Jerome Publishing, 2010.

ZANETTIN, Federico. **Translation-driven *corpora***: *corpus* resources for descriptive and applied translation studies. New York: Routledge, 2012.