


Exploring a Portuguese-English corpus through LancsBox®: possibilities for research in translation studies / *Exploração de corpus bilíngue português brasileiro-inglês através do LancsBox®: possibilidades para uma pesquisa em Tradução*

João Gabriel Carvalho Marcelino *

PhD student in Translation Studies, MsC in Language and Teaching (PPGLE/UFCG). Conducts research in the field of Translation Studies, focusing on Literary Translation, Interlingual Translation in the Portuguese – English pair, Translation and computational tools. He is currently carrying out doctoral research on the translation of elements from the northeastern sertão of Vidas Secas into English.

 <https://orcid.org/0000-0001-6528-0208>

Received in: 28 feb. 2023. **Approved in:** 03 mar. 2023.

How to cite this article:

MARCELINO, João Gabriel Carvalho. Exploring a Portuguese-English corpus through LancsBox®: possibilities for research in translation studies. v. 12, n. 1, *Revista Letras Raras*. Campina Grande, v. 12, n. 1, p. 155-178, apr. 2023.

ABSTRACT

Given the variety of linguistic processing tools, with free or paying access, in this paper presents the exploration of a translation corpus using the LancsBox® tool developed at Lancaster University. Seeking to explore a bilingual parallel corpus of research in translation using LancsBox® tool; the specific objectives are: i) To discuss the use of applications and linguistic processing tools through Corpus Linguistics and Computational Linguistics; and ii) To point out the possibilities of application of LancsBox® in translation research. This experimental study is based on Brezina, McEnery and Wattam (2015), Brezina and McEnery (2021), Berber-Sardinha (2002) and Jurafsky and Martin (2022). Findings show different application possibilities of the tool's functions through the search for terms that cause opacity, using the search for expressions in a simple way or through exploration by RegEx. Results have shown that linguistic processing tools can enrich the analysis of translations, facilitating the localization of the occurrences more efficiently.

KEYWORDS: Corpus Linguistics. Vidas Secas. Barren Lives. Translation Studies.

*

 jogabrielcarvalho@hotmail.com

 [10.5281/zenodo.7909650](https://doi.org/10.5281/zenodo.7909650)

RESUMO

Diante da variedade de ferramentas de processamento linguístico, de acesso livre ou pago, neste artigo observa-se a exploração de um corpus de tradução a partir da ferramenta LancsBox®, desenvolvida na Lancaster University. Buscando explorar um corpus paralelo bilingue de pesquisa em tradução utilizando a ferramenta de processamento de linguagem LancsBox®, caracterizam-se aqui os seguintes objetivos específicos: i) Discutir a utilização de aplicativos e ferramentas de processamento linguístico através da Linguística de Corpus e da Linguística Computacional; e ii) Sugerir possibilidades de aplicação do LancsBox® na pesquisa em tradução. Esse estudo experimental está fundamentado em Brezina, McEnery e Wattam (2015), Brezina e McEnery (2021), Berber Sardinha (2002) e Jurafsky e Martin (2022). Os resultados da exploração evidenciam possibilidades de aplicação das funções da ferramenta através de buscas por termos que causam opacidade, utilizando a busca pelas expressões de maneira simples ou através de exploração por RegEx. Os resultados das buscas mostram que ferramentas de processamento linguístico podem enriquecer as análises de traduções, permitindo localizar de maneira mais eficiente as ocorrências buscadas.

PALAVRAS-CHAVE: Linguística de Corpus; Vidas Secas; Barren Lives; Estudos da Tradução.

1 Introduction

Translation Studies have benefited from the interaction with Computational linguistics and Corpus linguistics, and through this interaction, there is the possibility of using linguistic processing applications, such as AntConc® and Wordsmith tools®, among others. This paper uses LancsBox®, an application developed by Lancaster University, as a model to explore a bilingual corpus for Translation research, to present some possibilities of application for research in Translation Studies. However, these applications were mostly developed for research with monolingual corpora, so it is up to Translation Studies researchers to adapt themselves to use the tools that make up these applications in research on mono and bilingual text corpora. Note that each research has particularities that guide the choice of more appropriate programs, given the variety of tools that the applications offer.

This paper presents an exploration of a bilingual corpus in the Brazilian Portuguese-English pair to expose possibilities of using the LancsBox® tool considering research oriented to Translation Studies. We use as model for the exploration the bilingual corpus elaborated from the novel Vidas Secas, by Graciliano Ramos, and its respective translation. Considering that each research observes language in a different aspect, this article seeks to contribute with possibilities of using tools and search strategies, contemplating translation research corpora and computational tools that optimize linguistic research methodologies.

As a general objective, the paper aims to explore a parallel bilingual corpus of translation research using LancsBox® language processing tool. To this, the following specific objectives were defined: i) Discuss the use of language processing applications and tools through Corpus

Linguistics and Computational Linguistics; and ii) Point out the possibilities of applying LancsBox® in translation research.

This article is based on the research of Brezina, McEnery and Wattam (2015), Brezina and McEnery (2021), Berber Sardinha (2002), among others. Having an experimental nature, it is divided into four sections. The first section presents the theoretical background about Computational linguistics and *Corpus* linguistics, as well as the relation between *Corpus* linguistics and Translation studies. The second section presents the methodology, the *LancsBox*® tool and the *corpus* studied. The third section presents the results of the searches performed in the tool, and finally, the final considerations.

2 Computational and *Corpus* Linguistics

The development of computer technologies such as the microcomputer has made possible the emergence of new approaches to problems in language study. This advent has allowed electronic *corpora* to be built or digitalized, research on linguistic *corpora* to be carried out more rapidly to observe regularities in language, and programs capable of interpreting and generating information about natural language to be developed (VIEIRA, LIMA, 2001). In this sense, computational linguistics has been involved with the development of software tools aimed at natural language processing, applicable to the different fields of linguistic research.

The use of *corpora* as a resource for linguistic research has been recurrent (ALUÍSIO; ALMEIDA, 2021) in view of the different possible observations about language and its functioning, even before the appearance of computational tools. McEnery and Hardie (2012) define *Corpus* Linguistics as the area focused on developing a series of procedures and methods to study language, making it possible for linguists to use such procedures and methods applied to different *corpora* to answer research questions applied to several disciplinary fields aligned to linguistics.

Linguistics has defined *corpora* with different characterizations over time and may classify them as a finite set of texts that can be taken as an object of analysis; or to establish the descriptive grammar of a language; or a set of utterances issued in a language; or even a set of written or spoken texts that can be made available for analysis (ALUÍSIO; ALMEIDA, 2021). Such characterizations converge on the idea that a *corpus* is a set of texts that can be analyzed for

different purposes, considering the different branches of Linguistics, such as Translation in the case of this article.

Corpus linguistics needs a relationship with another field for the analyses to be carried out, which is evident when one observes that the procedures are still considered as 'under development', although some already have greater or lesser degrees of consolidation, such as the concordancing¹ (MCENERY; HARDIE, 2012). Such procedures are developed with the aid of Computational Linguistics and Software Development for applications such as AntConc, *Wordsmith*, *LancsBox*, among others, which perform the processing of texts from digitized *corpora*. The use of *corpora* shows the inexactness of human intuition in understanding language, which can be observed more accurately through electronic tools (BERBER SARDINHA, 2002).

The *corpora* must follow some characteristics in their elaboration for further analysis. Aluísio and Almeida (2021) point out as important issues for the elaboration of a *corpus*: i) Authenticity of the texts; ii) *Corpus* representativeness; iii) Balance (corresponding to the balance of discursive and textual types or genres); iv) Sampling; v) Diversity; and vi) Adequate size for the research. These characteristics are directed by the purpose of this paper, as well as point to the methods of data collection.

The *corpus* characteristics pointed out by Aluísio and Almeida (2021) corroborate the ideas of Knight (2011), who points out four characteristics of parallel *corpora*: i) design and infrastructure, corresponding to the *corpus* structure from construction to presentation; ii) size and purpose, dealing with *corpus* dimensions; iii) nature, dealing with how realistic the *corpus* is for the researched context; and iv) availability and (re)use, considering if the *corpus* is published in full or in parts and if it is available to other researchers for verification or analysis. These characteristics make it possible to structure a *corpus* for linguistic research in different formats and for different purposes, pointing to the possibility of applying different looks to the same *corpus*.

Aluísio and Almeida (2021) consider three stages for the collection of a *corpus*: i) *corpus* design; ii) compilation, manipulation, naming of files and requests for permission to use the data; and iii) structural or linguistic annotation of the working *corpus*. These steps are consolidated in the construction of a linguistic research *corpus* that can be treated and processed through Natural Language Processing (NLP) tools (ALUÍSIO; ALMEIDA, 2021), aligning *Corpus* Linguistics with Computational Linguistics. The following section explores the relationship between Translation

¹ Concordancing refers to the process of checking how words behave in a text, performed by linguistic processing tools such as *LancsBox*, *AntConc*, and others. Concordancers allow you to observe syntactic or word placement patterns for different types of linguistic analysis.

Studies and *Corpus* Linguistics, considering the possibilities of applying computational tools to the field and the proposal presented in this article.

2.1 Translation Studies and *Corpus* Linguistics

Translation Studies, according to Williams and Chesterman (2010), describes the disciplinary field dedicated to the description, analysis and theorization of the processes, contexts, and products of translation, as well as the roles of the agents involved in these observation spaces. This characterization of the disciplinary field allows different theories to be applied to the objects of Translation Studies, allowing the observation of the process, the product, and the agent of translation. Among these theories, *Corpus* Linguistics and Computational Linguistics can be applied to support analysis of the translated text in relation to the source text; in relation to texts produced in the target or source language, as well as in the development of translation tools and the training and education of translators.

Berber Sardinha (2002) stresses that, despite being a very useful tool, the integration between *Corpus* Studies and Translation Studies has been slow. In this sense, the author suggests three hypotheses to justify this slow integration. The first hypothesis being the prejudice of *corpus* linguists in relation to the translated text, for they considered it a deviant text and not representative of the language; the second hypothesis being associated with the negative image that Linguistics had for translators and translation researchers, considering that for a long time Translation was seen as a mere application of different linguistic theories; and the third hypothesis was in the sphere of access to technology.

Berber Sardinha (2002) cites the research directions on the application of *Corpus* Studies in translation in topics such as the elaboration of bi- or multilingual *corpora*; automatic alignment of parallel *corpora*; creation and use of concordancers and other *corpus* processing tools; Machine Translation and Translation Memories; *Corpora* in the training of translators and translation researchers; and the observation of the translation process of professional translators. Olohan (2004) emphasizes that such directions must be well defined for the research, since the elements observed in research on translation must be clearly delimited and the *corpus* must be built aiming at a high representativeness for the aspects studied. In this sense, the definition of the type of Translation research approach conditions the *corpus* collection and annotation method; definition of output format; as well as the tools used, whether for translation or translation analysis.

The possibilities of methods for applying *corpora* in translation research are varied. Berber Sardinha (2002) suggests: i) identifying the formal standardization of L1 and the corresponding functions; ii) identifying *prima facie* translation equivalents for each function; and iii) identifying the formal standardization of L2 and the corresponding functions. Another methodology presented by the author referencing Tognini-Bonelli (2001, 2002)² is the one based on three *corpora*: i) a parallel one, with texts in the source language and their respective translations into the target language; ii) a reference one, with texts produced in the source language; and iii) a reference one, with texts produced in the target language. This methodology suggests mapping the patterns related to the items of interest in the components of the parallel *corpus* as well as in the reference *corpora* to identify whether the usages present in the parallel *corpus* are typical or not.

2.2 Search Possibilities in Literary Translation

The tools developed and explored by Computational linguistics help optimize research in Applied Linguistics, as observed in translation, allowing researchers, through these tools, to find recurring patterns in translation to understand choices in the translation process and theorize about the impacts on the translation product. In this sense, exploring a translation such as the one proposed in this paper requires positioning the text and the elements that guide the observations made.

Graciliano Ramos' *Vidas Secas* (1938) is an episodic narrative about the life of the family of *Fabiano* and *Sinhá Vitória*, retreatants who travel through the *caatinga* with their two children and their dog *Baleia*. Graciliano Ramos' work presents in a brief narrative the setting of the *caatinga*, an exclusively Brazilian biome, which dictates the paths taken by the characters. The narration in third person and the characters' little ability to orally express themselves make the text brief and succinct, written by a "writer who only said what was essential and, for the rest, preferred the silence"³ (CÂNDIDO, 2012, p. 142). Leaving in the brief spaces of writing elements so particular

² The author references Elena Tognini-Bonelli on the book *Corpus Linguistics at Work*, published in 2001 by John Benjamins; and on the book chapter *Functionally complete units of meaning across English and Italian: Towards a corpus-driven approach*, part of the book *Lexis in Contrast: Corpus-Based Approaches*, organized by Bengt Altenberg and Sylviane Granger, and published by John Benjamins in 2002.

³ "escritor que só dizia o essencial e, quanto ao resto, preferia o silêncio" (CÂNDIDO, 2012, p. 142)

to the *Sertão*⁴ that, in translation, the translator needs to deal with this brevity of words to transpose the meanings of the literary text.

The translation explored in this paper was performed in 1964 by Ralph Edward Dimmick and presents elements that make it possible to discuss the strategies adopted in the translation process. Thus, it justifies the use of computational tools to explore the *corpus* in order to identify different translation strategies from Brazilian Portuguese into English.

Oustinoff (2011) exposes that there is no 'neutral' or 'transparent' translation, in which the original text would appear mirrored, identically. In this sense, translation deals with the complexity of meanings in the transposition between languages and different cultures, even more, in literary translation, the translator deals with the different interpretations that the text allows each reader, since the literary text is not fixed in interpretations (BRITTO, 2016).

Even in a text by an author like Graciliano Ramos who sought to say only the essential, the presence of elements that have no correspondence in the target language and culture denote the complexity of the source language and culture, which highlights the creative nature of the work of translation (BRITTO, 2016). To deal with the dryness of the environment, the text, and the biome, the translator needs to make choices and adopt strategies that sometimes erase the foreign, sometimes bring it to the fore.

By observing the translation process in the Brazilian Portuguese-English pair, it is possible to reflect on the recognition of asymmetric relations in translation projects, considering the global hegemony of English (VENUTTI, 2019). This allows translations of Brazilian Portuguese texts that are part of the Brazilian literary system and have been translated into English to be studied aiming to observe how the translator deals with particular elements of the source language and culture in the transposition to the target language and culture. What can be highlighted is the translation tendency that attaches the text to a target culture by submitting it to that language and culture, what Berman (2013) exposes as the idea of translating to give the impression that the author would have written that way, if he wrote for that language.

This allows us to reflect on the processes of Domestication and Foreignization explored by Venuti (2021) which, respectively, reduce the foreign text to the values of the target language by incorporating the author into that environment; or create a pressure on those foreign values to

⁴ Translation note: The *Sertão* is one of the four sub-regions of the Brazilian Northeast Region, being the largest in area. The region covers the states of Alagoas, Bahia, Ceará, Paraíba, Pernambuco, Piauí, Rio Grande do Norte, and Sergipe. This region has four geoclimatic subregions named *Zona da mata* (Atlantic forest); *Agreste* and *Sertão* (Semi-arid regions), and *Meio Norte* (transition zone between Semi-arid and Amazon Rainforest).

register the linguistic and cultural difference of the foreign text, taking the reader outside. Observing the presence or erasure of the Foreign in the text allows denoting which approach predominates in translation, allowing the researcher to identify and demarcate in the *corpus* occurrences that report these tendencies to quantify and analyze them. Therefore, considering these two values in discussion about the literary text does not mean that they are not present at some level in translation (BERMAN, 2013), but that at certain moments distinct strategies can be used to accomplish a functionally equivalent translation.

It is possible to locate in a translation *corpus* elements that occur in the source text, demarcate them with tools that allow *corpora* annotation to identify deforming tendencies, which Berman (2013) exposes as an inevitable play of forces that the translator is in the middle of while transposes the letter (the text), considering that languages in translation operate differently. As well as, to observe what Franco-Aixelá (2013) discusses as Culture-Specific Items, elements that cause ideological or cultural opacity that require different strategies to deal with them in the translation process.

In this sense, by recognizing the type of translation strategy or tendency used, the researcher can recognize initial patterns to perform searches using tools such as Regular Expression search (RegEx) to perform annotations, find frequencies or patterns in the *corpus* from words, constructions or parts of words that may be visibly frequent in the translation process. Considering, for example, the repetition of cultural-specific items through the strategy of conservation or intratextual explanation (FRANCO-AIXELÁ, 2013), or reflecting on translation analytics and deformative tendencies (BERMAN, 2013) in the translation or conservation of culture-specific items such as the elements of the northeastern sertão present in *Vidas Secas* in the *Barren Lives* translation.

In the following section, the methodological aspects of the paper are presented, explaining the tool used, LancsBox®, the *corpus*, the definitions of the searches performed and the discussion of the results.

3 Methodology

This section presents the methodological aspects of the article, so in each subdivision are presented the elements used to perform the searches that are presented in the section of Results and Discussions. Thus, this methodology presents the LancsBox® application, with a description

of the functions and use; and presentation of the interface (BREZINA; MCENERY; WATTAM, 2015; BREZINA; TIMPERLEY; MCENERY, 2018; BREZINA; WEILL-TESSIER; MCENERY, 2021). Then the working *corpus* is presented, describing the steps of elaboration and the statistical data describing the *corpus*; and, finally, the search logic is described by explaining the items searched in the *corpus* using regular expression localization (RegEx).

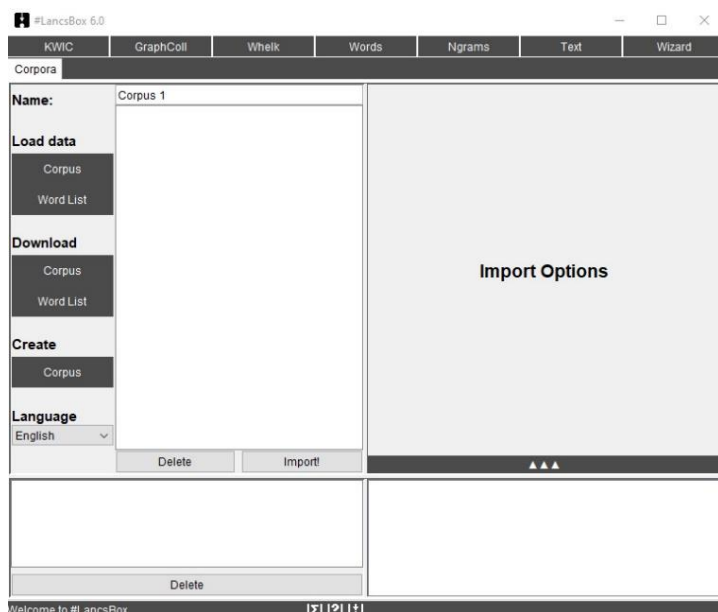
3.1 The LancsBox® app

LancsBox, short for Lancaster University *Corpus* Toolbox, is a *corpus* analysis tool developed by Lancaster University for 64-bit operating systems. The tool can analyze linguistic *corpora* in different formats (.txt, .xml, .doc, .docx, .pdf, .odt, .xls, .xlsx etc.), as well as preparing wordlists and *corpora* extracted from web pages (BREZINA, 2021)⁵.

The LancsBox® interface presents, in the initial screen, the option to add working *corpora*, as well as the tools provided by the application and the option to change the language of the application; however, the Portuguese language is not yet completely available. For this reason, the application has been run in English. On the left you can see the tools for adding a working *corpus*, adding to an existing *corpus*, downloading a *corpus* directly from the application, and creating a *corpus* directly in the application. The *corpora* data entered in the application are saved and listed in the lower left area, allowing you to access them again as many times as needed:

⁵ Available at: <http://corpora.lancs.ac.uk/lancsbox/>

Figure 1 –LancsBox® 6.0 interface



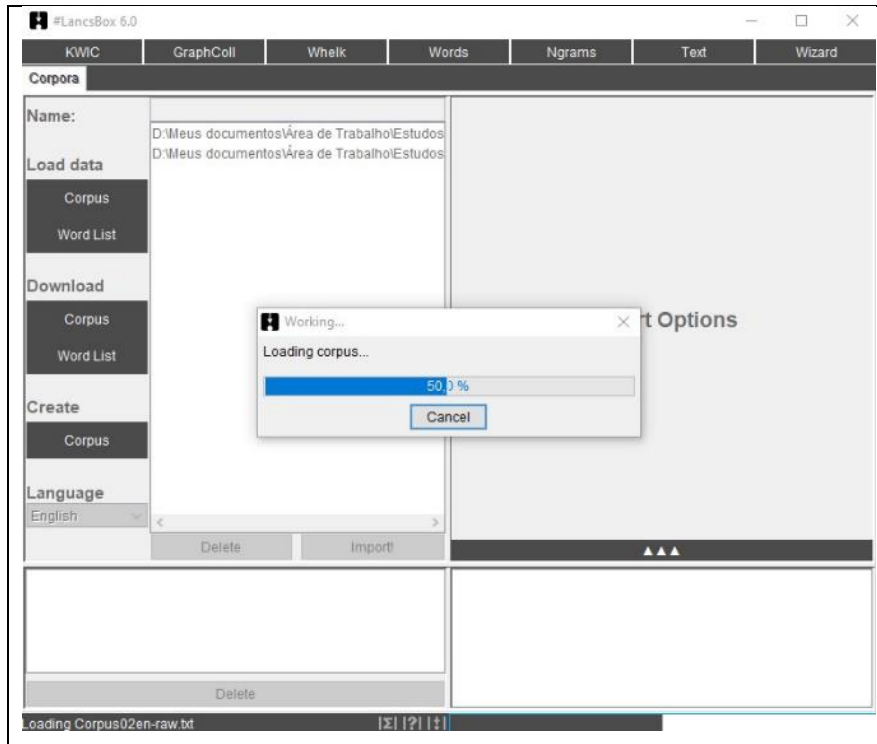
Source: Screenshot by the author

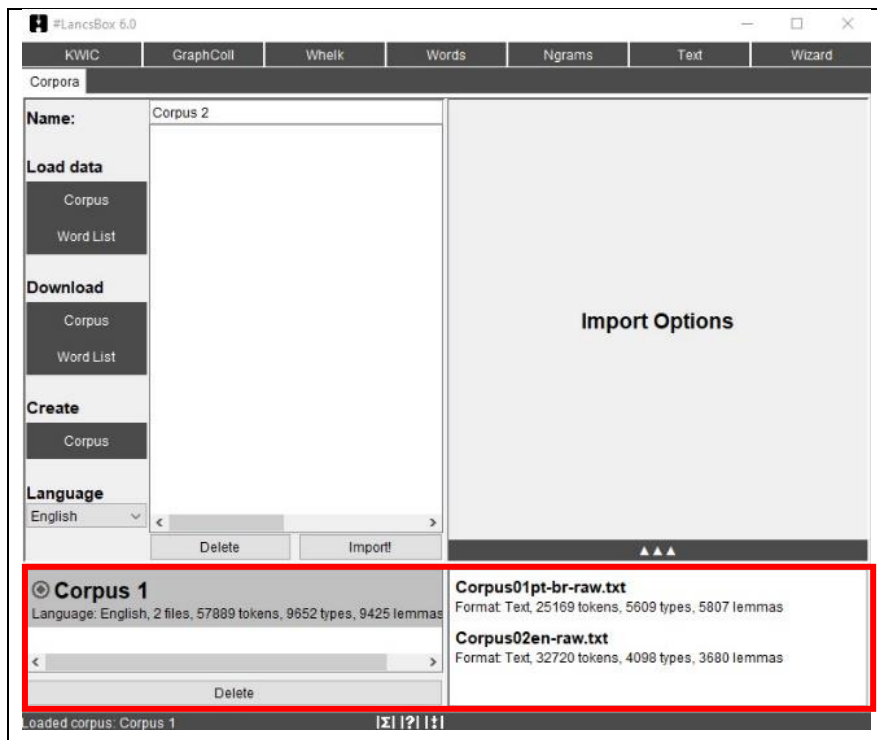
The search tools available in the top bar of the interface shown in Figure 1 direct the user to the following options:

- **KWIC (KeyWord In Context):** keyword search in concordances
- **GraphColl:** Searches for placements and presents the results in graphs or networks.
- **Whelk:** shows the distribution of the search term in the *corpus* files.
- **Words:** allows you to observe frequencies and perform *corpora* comparison in combination with other tools.
- **Ngrams:** Allows you to search and analyze N-gram frequencies.
- **Text:** shows the search terms in sentence contexts.
- **Wizard:** allows you to perform all the previous searches and create an automatic report with the obtained results (BREZINA, 2021)

To use LancsBox®, the working *corpus* was added through the option *Load data >> Corpus* the bilingual *corpus* used, since it is a small *corpus*. The upload to the application was quick and three working *corpora* were created, one with the text in Brazilian Portuguese, one in English, and one with both *corpora*:

Figure 2 – *Corpus* upload process in the app.





Fonte: Screenshot by the author

When *uploading* the files of a *corpus*, *LancsBox* automatically performs token, type and lemma counting⁶, presented in the table highlighted in red. The data presented are used to build the analysis on the *corpus* in different stages of analysis. In the following section, the data from the *corpus* used in the proposed exploration presented here will be presented.

3.2 The Vidas Secas-Barren Lives Corpus

The bilingual *corpus* of *Vidas Secas-Barren Lives* is a *corpus* elaborated for research in the field of Translation Studies, and corresponds to the text of *Vidas Secas* (1938) by Graciliano Ramos, in Portuguese-Brazilian, and, in parallel, the text of *Barren Lives* (1964), translation of *Vidas Secas* by Ralph Edward Dimmick. The *corpus* is elaborated from data that have already passed through the linguistic data collection stage (digitalization, cleaning and pre-processing), *corpus* statistics (quantitative characterization) and labeling (manual or automated), which makes it a *corpus* ready for the linguistic analysis stage (CUNHA, 2020). Table 1 describes the steps and activities performed in the construction of the *corpus* in a simplified way:

⁶ Token corresponds to "words" in a *corpus*; it is used to count words in the *corpus*; *Type* corresponds to particular forms of words in a *corpus*; and *Lemma* to derivations of words in a *corpus* (pebble and quarry are *lemmata* of stone).

Table 1 – Stages of *corpus* elaboration

Language data (scanning, cleaning, and pre-processing)	Conversion of the nato-digital files to .pdf and .txt. Extracting the text and separating the chapters. Text cleanup. Drawing up the <i>corpus</i> for annotation. Preparing the <i>corpus</i> without annotation.
Corpus statistics (quantitative characterization)	Size in kb. Number of words. Number of <i>Lemmas</i> , <i>Types</i> and <i>Tokens</i> . Number of lines
Labeling (manual or automated)	Label annotation.
Linguistic Analysis	Analysis focused on translation studies.

Source: prepared by the author based on Cunha (2020) and the *corpus* data.

To build the text *corpus*, the *Sublime text*⁷ tool was used, which allows writing texts with UTF-8 encoding (universalized format), searching for regular expressions, as well as converting them to different formats, including XML (eXtensible Markup Language), a format used for *tagging*. Given the steps presented by Cunha (2020), the *corpus* prepared has the following data:

Table 2 – Data from the Vidas Secas-Barren *Lives corpus*

Language	Brazilian-Portuguese	English
Name	<i>Corpus01pt-br</i>	<i>Corpus02en</i>
Format	.txt	.txt
Encoding	UTF-8	UTF-8
Size	164kb	193kb
Words	25290	32783
Lines	2077	2115
Tokens	25169	32720
Types	5609	4098
Lemmas	5807	3680

Source: Prepared by the author

⁷ *Sublime text* is a cross-platform source code editor that supports different programming languages and different markup languages such as XML. The application is available for download and purchase of a license from the developer's website <https://www.sublimetext.com/>. The free version of the application does not hamper the usage compared to the paid version.

Given the data presented in Table 2, the searches will be performed considering both *corpora* to observe how translation deals with expressions and words that cause opacity in the translation process between the source language and the target language. The following section explains the terms and justifications for the searches.

3.3 Performed searches

Since LanCSBox® enables searches using *RegEx* (Regular Expressions), the searches performed here are also based on regular expressions. *RegEx*, in short, corresponds to advanced searches performed from any combination of characters presented between slashes (/) used to locate the concatenation of characters sought (BREZINA; WEILL-TESSIER; MCENERY, 2020; JURAFSKY, MARTIN; 2022).

RegEx searches, in the context of *Corpus Linguistics*, allow to specify linguistic elements in the text to extract them (JURAFSKY, MARTIN; 2022), facilitating the location and quantification of occurrences in the *corpus* studied. In this sense, what is determined as regular expressions for this context is based on particles that demarcate decisions in the translation process, such as conservation in translation (located through analogically identified terms, or mechanically located radicals, prefixes, and suffixes) or expressions that complement the meaning of conserved words, establishing a relation of explanation. Thus, the following search categories were established for the exploration presented below:

- i) conservation;
- ii) root words e suffixation;
- iii) placement or frequent occurrences.

The searches presented in the following section were arranged according to the criteria and the regular expression searched for, presenting the images resulting from the searches performed in the *LanCSBox* tool.

4 Results and Discussion

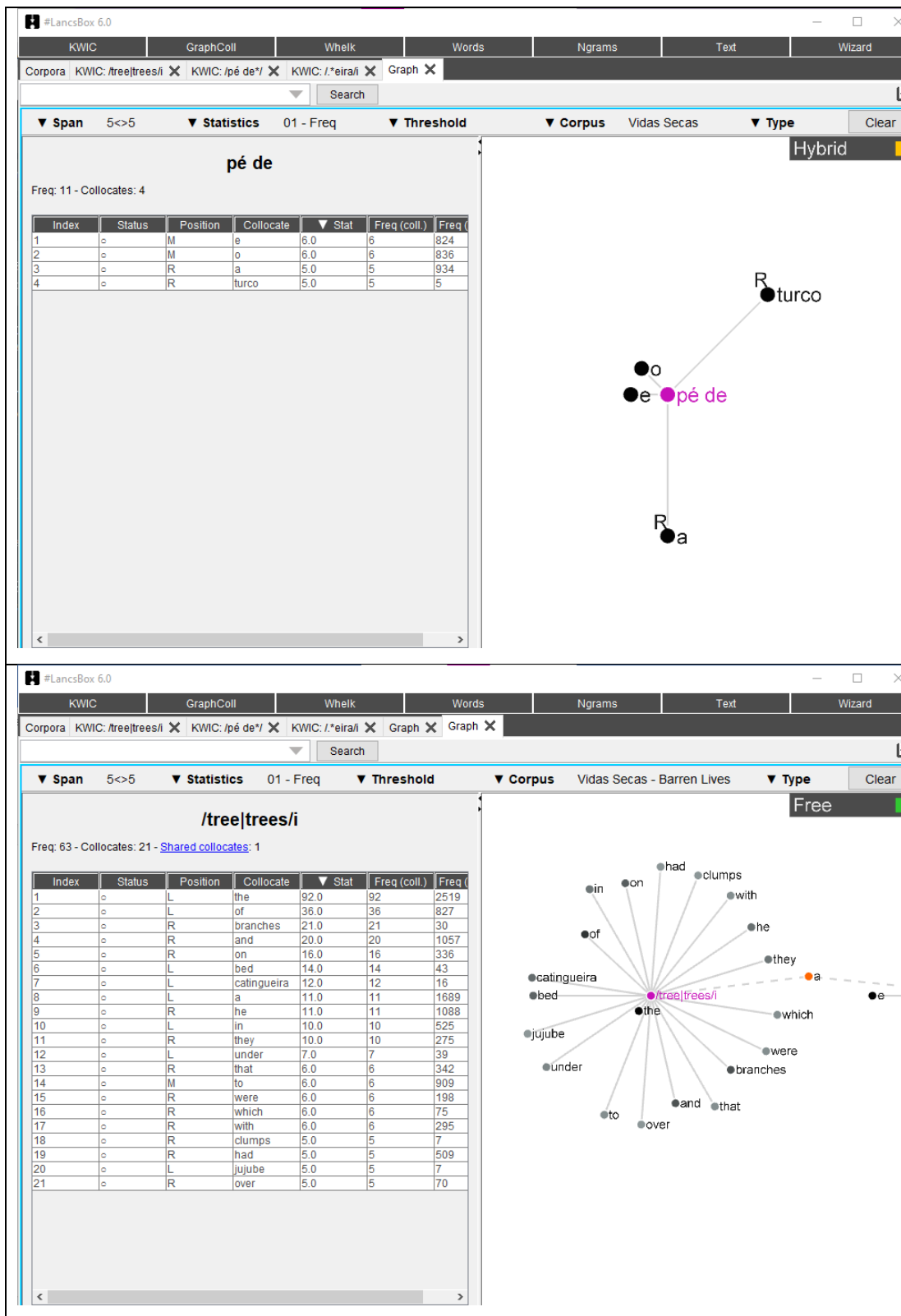
The data are presented in the following order: presentation of the term or regular expression searched; explanation of the search logic; justification for the search; results in the application, and commented results. The tools used were KWIC (Keyword in context) and GraphColl, for demonstration purposes.

4.1 GraphColl

The *GraphColl* searches were performed on the bilingual *corpus* to find recurrent collocations with the words 'pé de' in the Brazilian Portuguese *corpus* and 'tree'/'trees' in the English *corpus*. The regular expressions /pé de/ and /tree|trees/ are formulated with the terms arranged in full, considering the occurrence separated by slash (/) to consider the singular and the plural⁸. The search aimed to locate constructions beginning with 'pé de', such as 'pé de turco' or ending with 'tree' or 'trees' to observe the construction of names of vegetation elements:

⁸ Since the tree expression was sought as a complement to Portuguese terms preserved in the target text, it was not necessary to consider the possibility of case-insensitivity; when necessary, one should use square brackets as in the example: /[tT]ree|[tT]rees/ (JURAFSKY, MARTIN; 2022).

Figure 3 - GraphColl search results /foot of/ and /tree|trees/



Source: Screenshot by the author

The results obtained showed the highest occurrence of constructions ending with the expression *tree* in the translated text to justify the conservation of a term from the target language, performing an intratextual explanation (FRANCO-AIXELÁ, 2013) of the term conserved. The lower occurrence of placements of the expression "pé de" in the target language part of the *corpus* shows that the construction of words by suffixation, ending in -eira naturally points to elements of vegetation, making the presence of the expression not mandatory. In the data obtained from the target language *corpus*, the occurrence of intratextual explanation allows the reader to understand the meaning of the words preserved by context, as in the case of Catingueira, which in the Brazilian Portuguese language context is understood as a tree native to the Caatinga biome, where the narrative of *Vidas Secas* takes place.

4.2 Root words and suffixes

The suffix searches performed on the target text, *Barren Lives*, were done considering names of vegetation elements ending in '-eira'. For this, the regular expression searched was */*eira/* and */*eiro/*, the use of the asterisk (*) to construct the regular expression serves to indicate the occurrence of zero or more occurrences of the character sequence that occurs after the asterisk (JURAFSKY, MARTIN; 2022). The search was conducted to identify the occurrence of conservation of vernacular names of vegetation elements in Brazilian Portuguese present in the English *corpus*:

Figure 4 – KWIC search results for -eira and -eiro in the English corpus



Source: Screenshots by the author

Table 3 – Results extracted from LancsBox®

Corpus: Barren Lives Search Term: /*eira/i Occurrences: 12 (3.67) Texts: 1				
Index	File	Left	Node	Right
1	Corpus02en.txt	empty clay pit, a grove of withered	catingueira	trees, a turk's-head cactus, and the extension
2	Corpus02en.txt	any of them; he was like the	catingueira	and brauna trees. He, Vitória, the boys,
3	Corpus02en.txt	yard, and took refuge under the dry	catingueira	trees beside the empty pond. The dog
4	Corpus02en.txt	the wind shook the branches of the	catingueira	trees, and the roar of the river
5	Corpus02en.txt	of the river flats and reached the	catingueira	trees, which now must be submerged. Surely
6	Corpus02en.txt	had built of small pebbles, under the	catingueira	trees. Now the pond was full and
7	Corpus02en.txt	a few more steps. On reaching the	catingueira	trees, he adjusted his aim and pulled

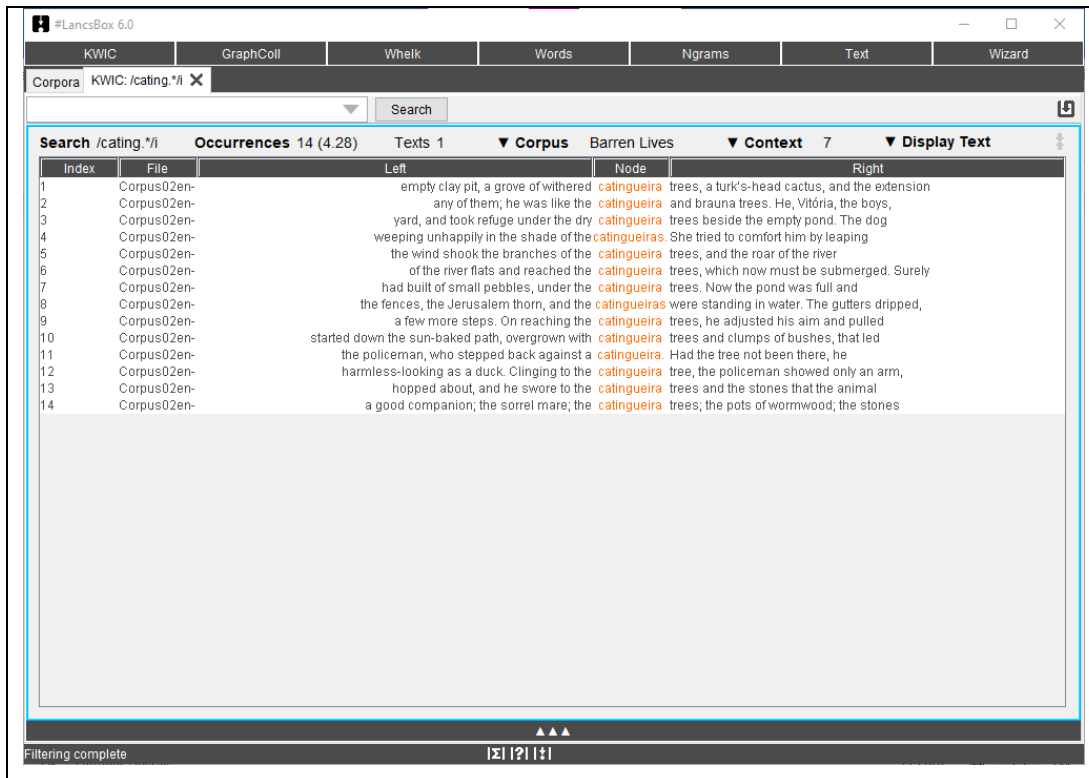
8	Corpus02en.txt	started down the sun-baked path, overgrown with	catingueira	trees and clumps of bushes, that led
9	Corpus02en.txt	the policeman, who stepped back against a	catingueira.	Had the tree not been there, he
10	Corpus02en.txt	harmless-looking as a duck. Clinging to the	catingueira	tree, the policeman showed only an arm,
11	Corpus02en.txt	hopped about, and he swore to the	catingueira	trees and the stones that the animal
12	Corpus02en.txt	a good companion; the sorrel mare; the	catingueira	trees; the pots of wormwood; the stones

Source: Elaborated by the author from the data extracted by LancsBox®

The results obtained, presented in Figure 4 and Table 3, point to the conservation of the word *catingueira*, with the *tree* complementation explained in the previous section. The results obtained, presented in Figure 4 and Table 3, point to the preservation of the word *catingueira*, with the ‘tree’ complementation explained in the previous section, as well as to the translation of vegetation elements ending in *-eiro*, such as *Juazeiro* (translated in the work as *Jujube tree*), which resulted in zero occurrences in the target language *corpus*. Considering that the word *Catingueira* derives from *Caatinga*, spelled *Catinga* in *Vidas Secas*, a root word search was conducted to identify the occurrence of the word *Catinga*, trying to contrast the results of the previous search by suffix, using the regular expression */cating*/* for the root word.

The use of the asterisk at the end of the searched prefix follows similar logic as the suffix search, establishing the search based on zero or more occurrences of the previous sequence (JURAFSKY, MARTIN; 2022), thus resulting in:

Figure 5 - return of searches for the root word cating- in the KWIC function.



Source: Screenshot by the author

Table 4 – Results obtained in LancsBox®

Corpus: Barren Lives Search Term: /cating.* Occurrences: 14 (4.28) Texts: 1				
Index	File	Left	Node	Right
1	Corpus02en.txt	empty clay pit, a grove of withered	catingueira	trees, a turk's-head cactus, and the extension
2	Corpus02en.txt	any of them; he was like the	catingueira	and brauna trees. He, Vitória, the boys,
3	Corpus02en.txt	yard, and took refuge under the dry	catingueira	trees beside the empty pond. The dog
4	Corpus02en.txt	weeping unhappily in the shade of the	catingueiras.	She tried to comfort him by leaping
5	Corpus02en.txt	the wind shook the branches of the	catingueira	trees, and the roar of the river
6	Corpus02en.txt	of the river flats and reached the	catingueira	trees, which now must be submerged. Surely
7	Corpus02en.txt	had built of small pebbles, under the	catingueira	trees. Now the pond was full and
8	Corpus02en.txt	the fences, the Jerusalem thorn, and the	catingueiras	were standing in water. The gutters dripped,
9	Corpus02en.txt	a few more steps. On reaching the	catingueira	trees, he adjusted his aim and pulled
10	Corpus02en.txt	started down the sun-baked path, overgrown with	catingueira	trees and clumps of bushes, that led
11	Corpus02en.txt	the policeman, who stepped back against a	catingueira.	Had the tree not been there, he
12	Corpus02en.txt	harmless-looking as a duck. Clinging to the	catingueira	tree, the policeman showed only an arm,
13	Corpus02en.txt	hopped about, and he swore to the	catingueira	trees and the stones that the animal
14	Corpus02en.txt	a good companion; the sorrel mare; the	catingueira	trees; the pots of wormwood; the stones

Source: Elaborated by the author from the data extracted by LancsBox®

The searches for the radical Cating- returned the same result as the searches for -eira, with two more occurrences due to the plural. With the results obtained it is possible to observe that despite the conservation of the plant name followed by the intratextual explanation (FRANCO-AIXELÁ, 2013) to establish the idea that 'Catingueira' corresponds to a vegetation element, the deletion of the biome name results in the destruction of an adjacent network of meanings (BERMAN, 2013), destroying the relationship between the tree name and the name of the biome from which it derives. When performing the same searches in the target language *corpus*, the KWIC tool obtained the results shown in Figure 6:

Figure 6 - KWIC search returns for -eira and -eiro in the Portuguese *corpus*

Index	File	Left	Node	Right
1	Corpus01-pt-t	verdes. Os infelizes tinham caminhado o dia	inteiro,	estavam cansados e famintos. Ordinariamente andavam pouco,
2	Corpus01-pt-t	culpado, mas dificultava a marcha, e o	vaqueiro	precisava chegar, não sabia onde. Tinham deixado
3	Corpus01-pt-t	mais arrastada, num silencio grande. Ausente do	mpanheir	a cachorra Baleia tomou a frente do
4	Corpus01-pt-t	fazenda sem vida O curral deserto, o	chiqueiro	das cabras arruinado e também deserto, a
5	Corpus01-pt-t	arruinado e também deserto, a casa do	vaqueiro	fechada, tudo anunciava abandono. Certamente o gado
6	Corpus01-pt-t	plantas mortas, rodeou a tapera, alcançou o	terreiro	do fundo, viu um barreiro vazio, um
7	Corpus01-pt-t	alcançou o terreiro do fundo, viu um	barreiro	vazio, um bosque de catingueiras murchas, um
8	Corpus01-pt-t	quis acordá-los. Foi apanhar gravetos, trouxe do	chiqueiro	das cabras uma braçada de madeira meio
9	Corpus01-pt-t	arrabuiu se pedras, arrastou os ventos, cortou	chiqueiro	de novo, fez um minuto localizou se

Index	File	Left	Node	Right
1	Corpus01-pt-t	correia presa ao cinturão, a espingarda de	pedemeirano	ombro. O menino mais velho e
2	Corpus01-pt-t	areia do rio, onde haviam descansado, a	beira	de uma poça: a fome apertara demais
3	Corpus01-pt-t	aligeirou o passo, esqueceu a fome, a	canseira	e os ferimentos. As alpercatas dele estavam
4	Corpus01-pt-t	do rio, acompanharam a cerca, subiram uma	ladeira,	chegaram aos juazeiros. Fazia tempo que não
5	Corpus01-pt-t	do chiqueiro das cabras uma braçada de	madeira	meio roída pelo cupim, arrancou touceiras de
6	Corpus01-pt-t	touceiras de macambira, arrumou tudo para a	fogueira.	Nesse ponto Baleia arrebitou as orelhas, arregaçou
7	Corpus01-pt-t	couro. Fabiano tomou a cuia, desceu a	ladeira,	encaminhou-se ao rio seco, achou no bebedouro
8	Corpus01-pt-t	bem dizer não se diferenciava muito da	bolandeira	de seu Tomás. Agora, deitado, apertava a
9	Corpus01-pt-t	os dentes. Que fim teria levado a	bolandeira	de seu Tomás? Olhou o céu de
10	Corpus01-pt-t	Tomás fugiu também, com a sua	bolandeira	de seu Tomás. Agora, deitado, apertava a

Source: Screenshot by the author

In view of the searches performed, it is observed that the suffixes -eira and -eiro form different words, these suffixes, derived from the Latin -arius, carry several meanings (VIARO, 2008), and their occurrence in Portuguese is used to indicate profession, container, object suitable for something, plant, naturalness, or quality (PRIBERAM, 2008). However, the occurrences that remain in the translation process, as observed in figure 5 and table 4, correspond to names of plants of the *Caatinga*, showing that the context of occurrence of the word tends to define the translation of its meaning, thus the structuring of the meaning of plant elements such as

Catingueira, gains a complementation of meaning through the intratextual description when converting the term to 'Catingueira tree'.

The results obtained from the searches performed in this section denote exploration considering one of different possibilities in a bilingual *corpus*. By observing the choices made to translate elements of vegetation, it is possible to determine traces of foreignization (VENUTI, 2021) in the target text in relation to the source text, as well as to efficiently locate occurrences of translation that erase the foreign and the particular characteristics to a certain space or environment, such as the Caatinga vegetation observed in *Vidas Secas/Barren Lives*.

Final considerations

In view of the above, it is possible to observe that the *LancsBox*® tool, by condensing the different functions in a single application, allows working with *corpora* with security, performing searches for different strategies and obtaining quantitative results which, in conjunction with qualitative analysis, make it possible to obtain satisfactory results in linguistic research. In this sense, the discussion of exploration strategies carried out in a specific area of Linguistics research using a tool such as *LancsBox*® seeks to contribute to the development of strategies by researchers who also use the tool explored in this article, as well as similar tools of more recurrent use.

Considering the use in a bilingual *corpus*, *LancsBox* is a viable option, considering the possibility of performing searches by terms or regular expressions. The searches can be performed in one or both languages of the *corpus* texts, making it possible to obtain dynamic results for linguistic analysis considering the objectives of the research in development. Thus, the tool, with the proper handling, can be an important support for different research carried out in Applied Linguistics.

Given the possibility of searching a bilingual *corpus*, *LancsBox* searches are a viable option for comparisons and analysis in the field of Translation Studies, allowing the tool to be used to search for terminology, suffixes, prefixes, radicals and terms that occur in one or both languages, as well as phrases and constructions in *corpora* for different observations. Therefore, in the experiments presented it is possible to observe both the strategies used to deal with cultural-specific items that cause opacity in the target text, and the consequences of the strategies adopted in terms of the meaning networks related to the language and culture of the source text. Such

results can be used to reflect on translation strategies, translation design, and observable translation trends such as domestication and foreignization and what this represents for the translation product.

CRediT
Acknowledgement: Not applicable.
Financing: Not applicable.
Conflicts of interest: The authors certify that they have no commercial or associative interest that represents a conflict of interest in relation to the manuscript.
Ethical Approval: Not applicable.
Contributor Roles: Conceptualization, Data curation, Formal Analysis, Funding acquisition, Investigation, Methodology, Project administration, Resources, Software, Supervision, Validation, Visualization, Writing – original draft, Writing – review & editing: MARCELINO, João Gabriel Carvalho.

References

- "-eiro", in Dicionário Priberam da Língua Portuguesa [em linha], 2008-2021, <https://dicionario.priberam.org/-eiro> [consultado em 03-02-2023].
- 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. *Calidoscópico*, [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.
- BERBER SARDINHA, Tony. *Corpora* eletrônicos na pesquisa em tradução. *Cadernos de Tradução*. V. 1, No. 9, 2002, p. 15-59. Disponível em: <<https://periodicos.ufsc.br/index.php/traducao/article/view/5980>>. Acesso em: 18 de out. de 2021.
- BERMAN, Antoine. *A tradução e a letra ou o albergue do longínquo*. Tradução de Marie-Hélène C. Torres, Mauri Furlan, Andreia Guerini. 2. ed. Tubarão: Copiart; Florianópolis: PGET/UFSC, 2013.
- BREZINA, V., MCENERY, T. & WATTAM, S. (2015). *Collocations in context: A new perspective on collocation networks*. *International Journal of Corpus Linguistics*, 20(2), 139-173.
- BREZINA, V., TIMPERLEY, M., & MCENERY, A. (2018). *#LancsBox v. 4.x*. [software package].
- BREZINA, V., WEILL-TESSIER, P., & MCENERY, A. (2021). *#LancsBox v. 6.x*. [software package]
- BRITTO, Paulo Henriques. *A tradução literária*. 2. Ed. Rio de Janeiro: Civilização Brasileira, 2016.
- CÂNDIDO, Antônio. *Ficção e Confissão: ensaios sobre Graciliano Ramos*. 4. Ed. Rio de Janeiro: Ouro sobre azul, 2012.

CUNHA, E. L. T. P.. *Contributions to the computational processing of diachronic Linguistic Corpora*. Tese (Doutorado em Linguística / Ciência da Computação) - Universiteit Leiden, Holanda, p. 221. 2020.

FRANCO-AIXELÁ, Javier. Itens Culturais-Específicos em Tradução. Tradução de Mayara Matsu Marinho e Roseni Silva. *In-Traduções*, Florianópolis, v. 5, n. 8, p. 185-218, Jan/jun., 2013. Disponível em: <http://incubadora.periodicos.ufsc.br/index.php/intraducoes/article/viewFile/2119/2996>

JURAFSKY, Daniel; MARTIN, James H. *Speech and Language Processing: An Introduction to Natural Language: Processing, Computational Linguistics, and Speech Recognition*. 3 ed. New Jersey: Prentice Hall, 2022.

KNIGHT, Dawn. The future of multimodal corpora. *Rev. bras. linguist. apl. [online]*. 2011, vol.11, n.2, pp.391-415. ISSN 1984-6398. Disponível em: <http://dx.doi.org/10.1590/S1984-63982011000200006>

McENERY, Tony; HARDIE, Andrew. *Corpus Linguistics*. Cambridge: Cambridge University press, 2012.

OLOHAN, Maeve. *Introducing corpora in translation studies*. New York: Routledge, 2004.

OUSTINOFF, Michaël. *Tradução: história, teorias e métodos*. Tradução de Marcos Marcionilo. São Paulo: Parábola, 2011.

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.

VENUTI, Lawrence. *A invisibilidade do Tradutor: uma história da tradução*. Tradução de Laureano Pellegrin... [et al.]. São Paulo: Editora Unesp, 2021.

VENUTI, Lawrence. *Escândalos da Tradução: por uma ética da Diferença*. Tradução de Laureano Pellegrin, Lucinéia Marcelino Villela, Marileide Dias Esqueda, Valéria Biondo. São Paulo: Editora Unesp, 2019.

VIARO, Mário Eduardo. *A formação do significado agentivo de -eiro*. In: XVI Congreso internacional de la ALFAL, 2011, Alcalá de Henares. *Actas del XVI Congreso Internacional de La Asociación de Lingüística y Filología*. Alcalá de Henares: Universidad de Alcalá, 2011. p. 2671-2679. Disponível em <http://www.usp.br/gmhp/publ/ViaA5.pdf>

VIEIRA, Renata; STRUBE DE LIMA, V. L. . *Lingüística Computacional: princípios e aplicações*. In: Ana Teresa Martins; Díbio leandro Borges. (Org.). *SBC - Jornadas de Atualização em Inteligência Artificial (JAIA)*. Fortaleza - CE, 2001, v. 3, p. 47-86.

WILLIAMS, Jenny; CHESTERMAN, Andrew. *The Map: a beginner's guide to doing research in translation studies*. Manchester, UK & Kinderhook: St. Jerome Publishing, 2010.