Article

CADERNOS de TRADUÇÃO

Normalisation in self-translation: a corpus-based analysis of *An Invincible Memory* by João Ubaldo Ribeiro and a review of the #LancsBox concordancer

Daniele de Sousa Santos
Instituto Federal de Educação, Ciência e Tecnologia do Pará
Belém, Pará, Brazil
daniele.santos@ifpa.edu.br
https://orcid.org/0009-0006-8688-5859

Abstract: A corpus-based study involves working with the available linguistic material. In this paper, I address the tendency of translated texts to have some distinctive features (Baker, 2007), such as simplification and normalisation, which may be found in both standard translations and in self-translations. The focus in this work was on normalisation in the self-translated historical novel *An Invincible Memory* (*Viva o Povo Brasileiro*) by João Ubaldo Ribeiro (1984, 1989) to describe lexical and collocational normalisation. The novel covers 400 years of Brazilian history, numerous common people characters, and entails a non-literate use of language. The objectives were to provide a quantitative analysis of normalisation in self-translation, to indicate how self-translation operates when the direction of translation is from a less frequently spoken source language, such as Portuguese, to a more frequently spoken target language, such as English and, finally, to explore how lexical and collocational normalisation could be represented statistically in this self-translation. The methodology was in #LancsBox and consisted of the compilation of the texts in the form of corpora, the generation of wordlists including the number of items for a statistical analysis, a type/token ratio (TTR) to identify the first signs of normalisation, and the Key Word in Context (KWIC) tool to determine the lengths of the sentences in both texts. Finally, concordance graphics (GraphColls) were generated to summarise the results. The findings provided evidence of the standardisation of the TTR values (0.111 and 0.077, respectively) in the source text and in the self-translation. In addition, the article *the* (raw frequency 13,498) was in the top 10 of the most frequent words, thus denoting the overuse of target language patterns to ensure easy readability via a lower lexical range. The study aimed to provide some insights into a more pragmatic approach to João Ubaldo Ribeiro's work.

Keywords: self-translation; historical novel; normalisation; corpus-based analysis; #LancsBox.

Cadernos de Tradução, 44, 2024, e98752
Graduate Program in Translation Studies
Federal University of Santa Catarina, Brazil. ISSN 2175-7968
DOI https://doi.org/10.5007/2175-7968.2024.e98752

1. Introduction

Although corpus-based translation studies have played a prominent role in translation studies, the available work on the general features of translated texts (Baker, 2007; Baker & Saldanha, 2020) has remained intuitive and qualitative; this indicates that different types of studies are required in this field. The use of a corpus analysis in the present study is an attempt to minimise these intuitive aspects with the aim of providing a more quantitative approach. Focusing on a particular universal law of translation (Toury, 2012), namely normalisation (Baker, 1995), was expected to provide a quantitative bias for the traditionally qualitative analyses of normalisation.

Linguistic fields such as Lexicography focus on statistics to ensure that the candidate words that may potentially be included in dictionaries and glossaries are valid and are in current use. Similarly, the question in Translation Studies, particularly in studies of self-translation, is how word extraction can play a significant role. To answer this question, verification of the use of collocates in translated texts compared to source texts as a reference might reveal how a self-translator makes their choices, as well as whether it is possible to observe a tendency towards a more normalised text in a self-translation (Baker, 1995, 1996, 2000), or if it is even possible to extract different factors that indicate this tendency towards normalisation.

Therefore, the present research comprises a literary study of two single-text corpora (two corpora consisting of one text each), namely the source text and the self-translation of the historical novel *Viva o Povo Brasileiro* (*An Invincible Memory*) by the Brazilian writer João Ubaldo Ribeiro (1984, 1989). The study entailed the compilation of the corpora, the generation of wordlists with the number of items for the statistical analysis, the identification of the type/token ratio (TTR) to observe the first signs of normalisation, and the Key Word in Context (KWIC) tool to determine the lengths of the sentences in both corpora. Finally, the concordance graphs (GraphColls) tool was used to generate graphics to summarise the results. Another research aim was to discuss the inconsistency in the concepts and the political reasons for self-translation that have preserved the hierarchy of language prestige by not considering the bidirectional flow of translation.

2. State-of-the-art

Hegel proposed relationships between the subject, history, and conscience, which gave rise to the literary genre that Lukács (2011) called the historical novel. The French Revolution and the Napoleonic Wars were the backdrop to Hegel's reasoning and gave rise to Lukács' (2011) conceptualisation of the historical novel, which combines history and literature. In effect, narrative fiction entails the quest to reveal the seminal aspects of a people's legacy that official history insists on concealing. Given the Enlightenment concepts on who mankind is for themselves and in History, the historical novel arose as a kind of anti-History account, somehow more reliable, the paradoxical role of which is to provide society with a faithful portrayal of themselves through fictional – and real– characters in their own history.

In this context, historiography, the ancient art of History account, found in the Neopositivism the chance to fulfil the "dream" of mathematical accuracy not only in History but also in science in general that, in fact, came true centuries later (Maestri, 2002) with the advent of the

21st century's technology revolution. However, historiography omitted critical analysis and aimed at a united method that could suppress the state of disarray prevailing in science, including history at that time, which appeared to be superficial, antiscientific, and opposed the very principles of science.

Before starting the novel, the writer assumes the position of a historian in the creation process, and creates a synchronic temporal framing to settle the plot – and History. Using a thorough documentary research approach and the unique characteristics of scientific documents, the author uses their imagination to finally interpret how society was like in a specific period. This imaginative ability "translates" the author to the period in question, and makes the synchronic, temporal framing of the story possible. Groot (2010, p. 28) claimed that "[...] the novel is best when it concentrates on the minor details and the marginalised characters to communicate the 'social and human motives of behaviour'". Therefore, there is not just one central character in the historical novel, but many of them.

In this regard, the development of historical novels provided a suitable scenario for literary self-translations, considering that the best person to translate their people's history faithfully is the author. As Popovič (1976, p. 19) stated, self-translation is conceived of as "[...] the translation of an original work into another language by the author himself". This particular characteristic of this kind of translation stems from the "prevailing mode of thought in the author" (Grutman & Bolderen, 2014, p. 324), i.e., the author/translator has autonomy and authority over the text to be translated. This can be considered to be the very essence of authorship, and is what Jung (2002, p. 29) called the power of "[...] reconstruct[ing] the memory of the original intention, rather than the intention itself", since two different text products are involved.

When advocating for self-translation, Shread (2009) suggested that self-translation consisted to be the creative expansion of a work that is recognised as being authentic. What is at stake is that, irrespective of critic's omission and devaluation, self-translation has proved its value by being used in many published and accepted works over the last decades. This non-standardised type of language transfer has achieved a broader space amongst readers, and requires more attention from Translation Studies. Extending beyond a migration phenomenon, self-translation has become the art of self-identity, self-knowledge, and self-transparency for to paraphrase Shread (2009). Having said this, the description and interpretation of this case study is sustained on the analysis of a self-translation commission asked by a publisher. In other words, João Ubaldo Ribeiro was not exiled, a migrant, or inserted in a foreign culture when he did this work. The present study is also an attempt to provide a different view of the novel.

Depending on each case, self-translation may present a tendency of standardisation that could sound unnatural or even foreignised (Venuti, 1995). This tendency stems from what Baker (2007) described as features of translated texts, or some recurrent traces in translator's choices. One of these features is normalisation, which consists of and entails "[...] a tendency to exaggerate features of the target language and to conform to its typical patterns" (Baker, 1996, p. 182). Accordingly, this work is focused on the traces of normalisation in the self-translated historical novel *Viva o Povo Brasileiro* (*An Invincible Memory*) by João Ubaldo Ribeiro (1984, 1989) so that it describes two types of normalisations, namely lexical and collocational.

The term "normalisation" was firstly coined by Baker (1995), and has been applied to corpus-based translation studies. In this correlated study, the internal chunks of language, such as the lexical

and collocational aspects, which are part of the same bias, the word itself, may help to provide a hybrid approach, that is, the computational statistical-linguistic analysis (Pazienza *et al.*, 2005). One of the most useful functions of applied corpus-based translation studies and concordancers is "providing information missing from dictionaries" (Zanettin, 1998; Frérot, 2016). Corpora are part of the technological tools that have supported translation which help search beyond dictionaries and glossaries, showing their importance in translation practice and research as well.

Normalisation is the predisposition to make changes in translations, such as changing grammatical structures and punctuation patterns, and minimising regional and cultural markers as far as possible to give the text a universal style to increase text intelligibility. Therefore, normalisation is an attempt to naturalise the transfer (Shuttleworth & Cowie, 2014), demonstrating that there is a grammatical tendency towards standardisation. This way, normalisation can also be seen as a predictable and subconscious result of language prestige hierarchy, which Camargo (2006) described as follows: the higher the source language's status, the lower the traces of normalisation, and vice versa.

As demonstrated by several studies (Camargo, 2006; Antunes, 2009; Paiva, 2011, among others), normalisation has been qualitative-oriented, strictly based on Discourse Analysis, while the statistical potential for the other features (Baker, 2007) such as explicitation and simplification has not been considered. However, the under-researched quantitative method of normalisation associated with Corpus Linguistics and Translation Studies might shed light on the development of such a hybrid approach; that is, balancing both qualitative and quantitative perspectives.

Particular attention has been paid to corpus analysis (McEnery, 2006; Brezina *et al.,* 2015; Gablasova *et al.*, 2017), as well as to work on collocation algorithms for term extraction (Evert, 2005; Pecina, 2010; Gries, 2013). These studies have led to the creation, development, and improvement of many concordancers such as WordSmith™, AntConc™ and, more recently, #LancsBox (Brezina *et al*., 2021), which is reviewed in this paper. These concordancers possess tools with the core functional roles of managing statistical cues to recognise terms (Pazienza *et al*., 2005) and extract quantitative results.

Although natural language is not static in terms of use and creativity, corpus analysis has become increasingly reliable by making "steady" calculations, which does not mean that these measures have no range in terms of coefficient of variation and of dispersion, but that they use fixed parameters. At this point, it should be emphasized that these parameters are intended to guarantee the reliability of the method, which has increased in precision by giving rise to many algorithms for association measures in the study of collocations. As has been shown in the literature (Evert, 2005; Pecina, 2010; Gries, 2013), much has been done in all the instances to produce useful studies, which are of crucial importance to understand how corpus-based translation analysis may contribute to the under-researched relationship between literary self-translations and corpus-based analyses in a broader extent.

3. Materials and methods

The methodology was based on the concept of normalisation proposed by Baker (1995, 1996, 2000); the materials consisted of a compilation of the original historical novel *Viva o Povo*

Brasileiro and its self-translation *An Invincible Memory* by João Ubaldo Ribeiro (1984, 1989) in the form of two single-text corpora (each corpus consisted of one text). The concordancer that was used was #LancsBox 6.0 (Brezina et al., 2021), according to which the source text contained approximately 233 K tokens, 25 K types, and 23 K lemmas, while the self-translation presented nearly 258 K tokens, 19 K types, and 17 K lemmas.

The tools used in #LancsBox 6.0 (Brezina et al., 2021) were KWIC, Words, and GraphColl. The KWIC tool was employed for searching for terms, whereby a node (searched term in the concordancer) can be visualised in all lines throughout the novel; that is, the window approach (Gablasova *et al.*, 2017). The window span was L_7 and R_7. The Words tool has a function that generates wordlists and the lexical statistics type/token ratio (TTR), the standardised TTR (STTR), and the moving average TTR (MATTR). GraphColl enabled the observation of collocations using the association measures MI3 and Log-likelihood, as well as the frequency, distance, and exclusivity, with the respective values being summarised in the graphs.

The lexical and the collocational analyses were conducted as described in the following sections.

3.1 Lexical normalisation

With regard to the word distribution in corpora, #LancsBox (Brezina et al., 2021) assists in describing the words in a corpus numerically, as well as indicating how the variety of vocabulary can be understood in terms of normalisation. A lexical analysis begins with the generation of a wordlist, in which every single term in the corpus is separated and has its frequency ranked. At this point, the Words tools made it possible to obtain the lexical statistics that guided this study, namely TTR, STTR, and MATTR.

TTR refers to the range of lexis in a text, and posits that the greater the value, the greater the variety of words throughout a text. Conversely, a lower TTR value indicates a high number of lexical repetitions, which might be a sign of lexical normalisation. Baker (1995) claimed that if the TTR were low, this might indicate that the translation had a lower lexical range, with more repeated words and less varied vocabulary, which can be understood as a means of ensuring that the translation has what Scott (1998, p. 19) called "easy readability".

According to Baker's (1995, 1996) initial studies, the TTR is closely related to the feature simplification. However, she did not mention normalisation, which is the focus of this paper in order to shed light on the use of TTR for the latter. Additionally, if the TTR measures the lexical range in a corpus, it is also possible to consider it to be evidence of lexical normalisation, given that the initial reduction in the TTR values indicates vocabulary repetition. The simple calculation of TTR (Brezina, 2018) consists of dividing the entire number of types (single words) by the number of tokens (running words occurrence), as follows:

$$\text{TTR} = \frac{\text{no. of types in text or corpus}}{\text{no. of tokens in text or corpus}} \quad (1)$$

Although TTR seems to be quite useful, the development of linguistic variables has led to more accurate statistical measures. Equation (1) expresses the raw value of the relationships between types and tokens (Brezina, 2018), based simply on the respective raw frequencies derived from the wordlist. Therefore, they can be considered less accurate, and not as reliable as they could be. To reach more robust values, recent studies have used STTR that was first proposed by Scott (2004), which takes not only the number of types and tokens into account, but also establishes the threshold called standard-size segment, which is a cut-off number to divide the corpus into segments afresh calculated.

Following corpus segmentation, frequency allows for the calculation of the raw TTR for each standardised batch of words. The concordancer then generates a mean value for all these TTRs, resulting in the STTR. Baker (2000) posited that the use of the raw TTR was reliable for texts that had the same length. For diverse ones, she recommended STTR, hence the importance of this statistical measure in this study since both source text and self-translation are of different length. Baker (2000) and Brezina (2018) recommended that for smaller texts or a single-text corpus, the standard-size segment threshold should be lower (e.g.: 100 words).

Over the years, Baker's recommendation for the use of STTR because it is more reliable has been replaced by Brezina's (2018) more robust MATTR, which also segments words, but not into numbered batches. Instead, MATTR uses an overlapping moving window – which is similar to optical character recognition tracking – throughout the corpus to calculate the TTR. As the TTR is calculated for each window position that has text content, the MATTR does not exclude very short final segments, as the STTR does. MATTR then generates a mean value for all the obtained TTRs without any exclusions, which leads to a more accurate coverage of the corpus as a whole. Thus, for a richer discussion, it is interesting to examine these three ratios, which is an easy task because #LancsBox (Brezina *et al.,* 2021) provides their values automatically.

3.2 Collocational normalisation

As shown in Firth (1957, p. 6), "the company words keep" points to the need to reflect on the relationship amongst them. The "company" – the collocates – is the first step in understanding translated texts and how they function in self-translations. Along these lines and according to standard grammar principles, collocations are combinations of two or more words that usually co-occur in texts and corpora (Brezina *et al.*, 2015). They can be analysed in the form of statistical occurrences, which initially stem from the generation of a wordlist and a raw frequency ranking. Wordlists and raw frequencies are thus the first and most elementary step in corpus analysis that can – and should – be followed by a more extensive next step, which includes algorithms, developed over the years by Evert (2008), Gries (2013), Brezina *et al.* (2015), amongst others, in the concordancer, which are known as association measures (AM). Thus, these AMs may also provide compelling evidence for collocational normalisation.

AMs are statistical associations (Evert, 2005) that have been developed to shed light on the automatic identification of collocations (Brezina *et al.*, 2015), and consist of *formulae* (Pecina, 2010) to determine the strength of collocations (Gries, 2013), that is, how close the relationship is between a node and a collocate (a term that matches the node to form a collocation). Hence, the

use of more than one AM to capture as many aspects of a collocation network as possible (Gries, 2013) is of paramount importance. Each AM has the goal of identifying one of the aspects of a collocational relationship according to certain criteria determined by a statistical function. At this point, it should be emphasised that, in translation, the essence of collocation extraction remains the same for both research and professional purposes, that is, choosing a node, running it on a concordancer and identifying its collocates regardless of whether applying advanced statistical measures are applied or not (Fantinuoli, 2016; Brezina, 2018).

The usual criteria (see Evert, 2005; Gries, 2013; Brezina *et al.*, 2015; Gablasova *et al.*, 2017) for analysing collocations are frequency, distance, and exclusivity. Frequency refers to the number of occurrences (tokens) of each word in a corpus. The distance of a collocation depends on the node that appears with it; it refers to the distance between the node's preceding and subsequent running words to the right and to the left, which is called a window span, and which are represented on #LancsBox (Brezina *et al.,* 2021) as R_5 and L_5 by default. The exclusivity of a collocation is linked to the frequency with which two or more terms occur aligned. For example, a term can present a high frequency with a specific preposition, but it can also appear with other collocations. When the opposite occurs, that is, the combination of these two terms is highly frequent or even unique, it is said that such a term is exclusive of the other (Brezina *et al.*, 2015), and they are likely to co-occur.

To assure a more reliable analysis, it is interesting the use of at least one AM. MI3 and log-likelihood were employed in addition to the raw frequency. MI3 calculates the exclusivity of terms; that is, how many and which terms are likely to co-occur. Loglikelihood, in turn, reduces the null hypothesis or even removes it, depending on the context, and on the type of corpus. However, there are some combinations throughout a corpus that do not always indicate a collocation network, which means that there is a certain degree of proximity and dependence between a node and a collocate, and it can be verified in its strength which is, with the AM, translated into a numerical result. What is important here is to avoid potential false collocations as far as possible.

4. Results and discussion

4.1 Lexical normalisation analysis

#LancsBox (Brezina *et al.,* 2021) reads PDF format of texts easily. According to it, the source text under analysis contained approximately 233 K tokens, 25 K types, and 23K lemmas, while its self-translation presented nearly 258 K tokens, 19 K types, and 17 K lemmas. As can be seen, the numbers for tokens (occurrences), types (words) and lemmas were higher in the source text compared to the self-translation. The TTR also differed, as shown in Table 1:

Table 1: Statistics for the source text and for the self-translation

Source Text	Stats	Self-translation
0.111	Raw TTR	0.077
0.748	STTR	0.720
0.749	MATTR	0.719
233k	Text length	258k
19.6	Sentence length	21.7

Source: Author (2024)

The values in Table 1 were estimated. According to Brezina (2018), using the raw TTR is only possible when the corpora have the same text length; for different text lengths, which was the case here, he recommended using STTR and MATTR. Overall, in Table 1, the statistical ratios in the source text show more varied numbers, thus indicating the presence of greater lexical variety, which is typical of source language texts. The opposite is true for the self-translation, where the three ratios possess quite near values and a higher text length, indicating a lower range in vocabulary.

Another interesting point is that the TTR, the STTR, and the MATTR in the self-translation presented values that were relatively close to each other, indicating a trace of normalisation; that is, these numbers demonstrated a limited range of vocabulary that could be confirmed throughout the text. By contrast, the source text presented smaller and more diverse numbers for the same rates, thus reflecting fewer traces of normalisation as a qualitative analysis may denote. With regard to the text and the sentence lengths, the source text had a text length of 233 K words, while the self-translation had a text length of 258 K words, showing that the latter was longer than the former.

The numbers revealed how the author shared his common ground with readers from his own culture and with foreigners. In other words, he was comfortable using a less typical lexicon for local readers, whereas the opposite was true for the self-translation, in which the ratios (0.077, 0.720, and 0.719) and the text and sentence lengths did not denote a lack of language skills, as one might be tempted to infer at first sight, but reflect a deliberate choice to ensure easy cultural readability by normalising the text. As a result, lower values in the self-translation reveal that the author's intention was to make his language use more accessible to non-Portuguese speakers.

It should be noted that the values for the STTR and the MATTR in the source text had higher range, 0.748 and 0.749, respectively. One must understand that the TTR was clearly influenced by the aforementioned values (233 K tokens, 25 K types, and 23 K lemmas). Baker (2000, p. 250) claimed that "[...] a high type/token ratio means that the writer uses a wider range of vocabulary. A low type/token ratio means that a writer draws on a more restricted set of vocabulary items". In other words, this reduction might indicate what Brezina (2018, p. 57) termed "recycled words" for the repeated running words in a corpus, which surely have some impact on the lexical diversity (Brezina, 2018), as well as providing compelling evidence of lexical normalisation in the self-translation.

Baker (1996) also implied that the lower range of lexical diversity in translated texts was because they were generally addressed to a non-native speaking audience, in an attempt to make the text easier to understand. However, this same lower range of lexical diversity could be seen in João Ubaldo Ribeiro's self-translation, which contradicts this statement to some extent. In João Ubaldo Ribeiro's case, the narrower range of vocabulary was a conscious choice to present his novel to the non-native Portuguese-speaking audience as a translated version (Ribeiro, 1990). This background can indicate a concept of lexical diversity usage that differs from the current one and may provide a broader perspective on lexical use. Thus, normalisation in self-translation might denote the author's decision, and not be a sign of poor transfer.

Figure 1 illustrates the top 10 most frequent words in both corpora. Positive keywords (+) represent the source text, and negative keywords (-) represent the self-translation. Normalisation was present in this text transfer, and the lexical similarity was 70%. This list shows that the prominent word classes (part of speech) in both texts were articles (*o, os/the*), prepositions (*para/to;*

de/*of*; *em*/*in*), and conjunctions (*e*/*and*). Therefore, it follows that the proximity between the texts was possibly due to normalisation.

Figure 1: The top 10 most frequent words in both corpora

1/7064	Keywords +
1	que
2	e
3	não
4	o
5	se
6	um
7	de
8	para
9	em
10	os

1/6300	Keywords -
1	the
2	to
3	and
4	of
5	he
6	his
7	in
8	that
9	was
10	it

Source: Author (2024)

There was a tendency towards normalisation regarding the correlation between both of the first top 10, *que* and *the*. The clause *que* (raw frequency 9, 207), which means *that*, is mandatory in Portuguese, while it is optional in English, was used in more times throughout the self-translation. Hence, *que* occupied the first position in Portuguese and the corresponding eighth position in English, with the latter indicating a tendency to being closer to Portuguese usage. Although it ranks as one of the most frequent positions, it can be said to be a trace of normalisation. In native English, *that* would rank in a lower position.

For the article *the* (raw frequency of 13,498 on KWIC tool), the collocations that were derived, included *of the* (raw frequency 1,509), *in the* (raw frequency 1,008), and *on the* (raw frequency 428). The most frequent collocate *of the* reveals an attempt at normalisation to ensure easy readability and to avoid ambiguity by explicitating and simplificating the self-translation. As Scott (1998) suggested, normalisation is part of a broader set of translated text features, which included other aspects, such as explicitation and simplification so that they contribute to normalisation.

If we consider normalisation to be the broadest feature of translation, from which all others stem (explicitation, simplification, levelling out, explanation, etc.), it can be said that (self)translation relies on the key concept of the norm. However, this does not entail establishing patterns of language to be blindly followed by (self)translators and readers, but norm readability, whereby the patterns that are created strictly obey plain feasibility of cultural transfer.

This section aimed to explore lexical normalisation by using word frequency and distribution through the TTR ratios. Normalisation could be observed in the POS distribution, in which the top 10 most frequent words suggest the frequent use of word classes (articles, prepositions, conjunctions) denoting the use of patterns in the self-translation. The same was observed in text and sentence lengths, which presented values that were higher than were those in source text, as well as the STTR and the MATTR, the close values of which provided evidence of normalisation.

4.2 Collocational normalisation analysis

This analysis was conducted using Words and GraphColl on #LancsBox (Brezina *et al*., 2021). The first generated new wordlists with lemmatisation to provide an even distribution of words, which were then saved as plain text to search for the top five high-frequency words. Results pointed to verbs. These verbs, henceforth called nodes, were used to produce the GraphColls.

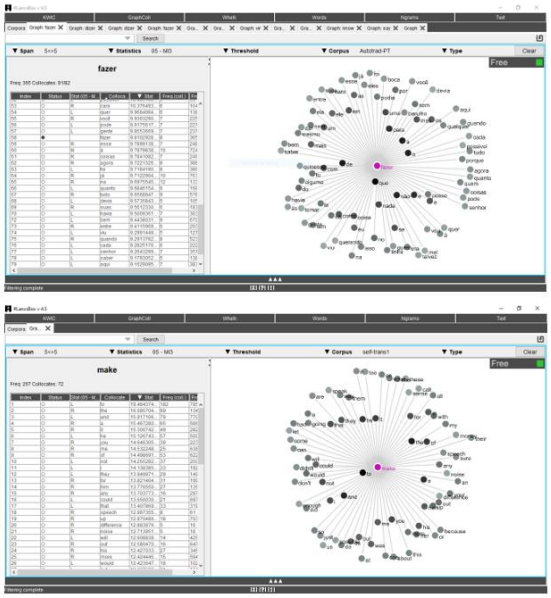
As the corpora were small and the graphs did not present an overpopulation of terms, there was no need to remove function words or to change the threshold, leading to the use of MI3 as the most effective filter; however, loglikelihood was also used for the node *Budiao*. The criteria that were used for the analysis were frequency, distance, and exclusivity.

4.2.1 GraphColls and lexical collocations

In the graphs below, the main AM that was used for data visualisation with the least intrusion in terms of filter parameters such as span size and threshold, was MI3. Function words did not need to be removed due to the small size of the corpora and the low frequency of the top five verbs, as they did not conceal the other most frequent collocates and normalisation.

The graphs show instances of lexical collocations. In other words, the nodes and the collocates illustrate two types of co-occurrences (Xia, 2014). The first is the idioms in both languages (e.g.: *fazer sentido*, *efetuar uma chamada*, *dar um discurso* /*make sense*, *make a call*, *make a speech*, respectively), while the second, lexical collocations, which are nodes that occur with other words but do not form an idiom, showed no exclusivity. See Figure 2.

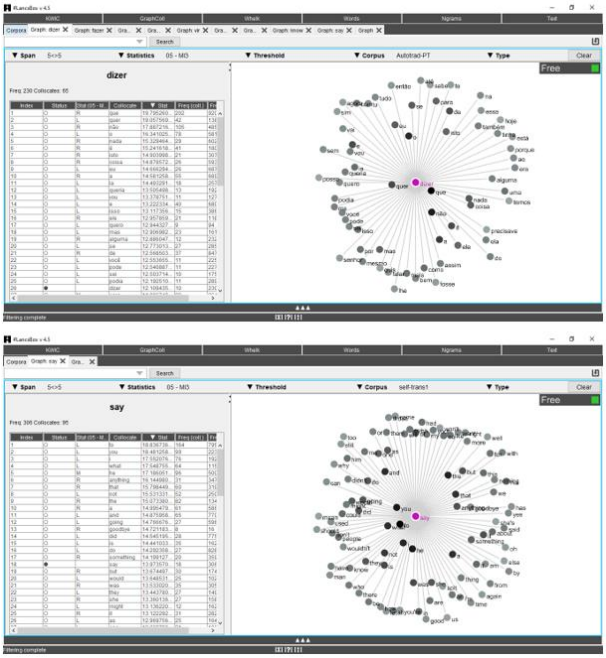
Figure 2: GraphColls for the pair of nodes *fazer* and *make*, the most frequent verbs



Source: Author (2024).

In practical terms, all the GraphColls presented more lexical collocations than idioms. The reason for this was probably to be the small size of the corpora, which consisted of only one long text in each. This particular feature of the corpora also influenced the decision not to remove the function words, since they would not modify the number of expected collocations. See Figure 3.

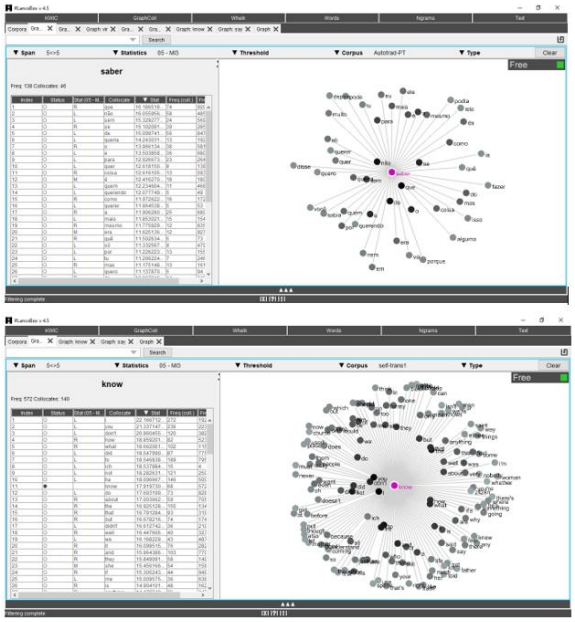
Figure 3: GraphColls for the pair of nodes *dizer* and *say*



Source: Author (2024).

The three criteria for the collocation analysis, namely frequency, distance, and exclusivity, repeatedly revealed traces of normalisation. Broadly speaking, raw frequency in GraphColl was illustrated by more populated graphs (Figure 4) in the self-translation, even when MI3 was applied. Whereas normalisation was previously endorsed via higher numbers for text and sentence length in Section 4.1, populated graphs (for the self-translation) showed not only syntactic differences between both languages but also the self-translator's decision to make his text more explicit for the target culture.

Figure 4: GraphColls for the pair of nodes *saber* and *know* (populated graph for the self-translation)



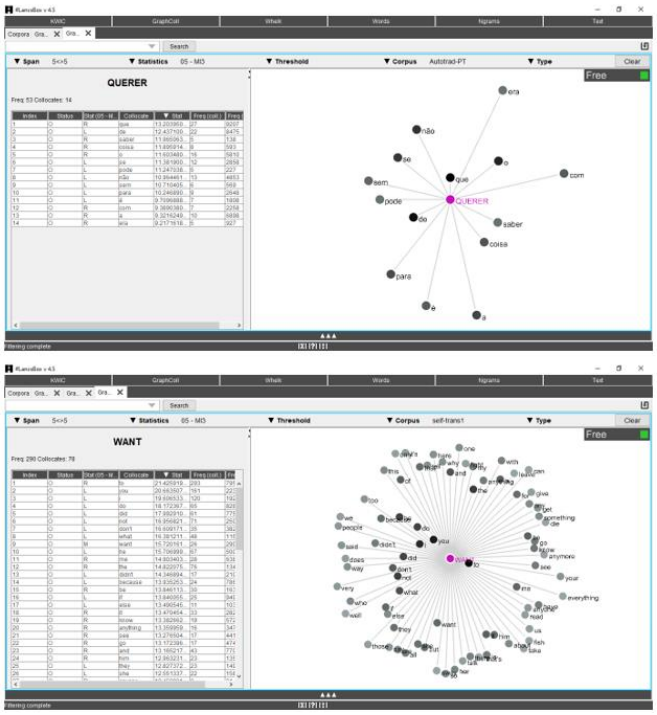
Source: Author (2024).

The distance between the nodes and the collocates was slightly larger in the self-translation. GraphColl showed the collocates in a more peripheral position; that is, the further they were from the node, the larger the number of lexical collocations, which indicates the decision to using a more standardised form to make the text as accessible as possible, even at the expense of producing a longer text in a language that is considered to be more concise than the source one.

This tendency for the collocates to be positioned around the edges of the GraphColl also influences exclusivity, the narrowness of which is not important in small corpora. It can be said that low occurrences of exclusivity indicate a decrease in lexical range, which is additional compelling evidence for normalisation.

The most striking case of normalisation change in GraphColls was the huge decrease in the numbers of collocates for the node *querer* and its self-translated node *want* (Figure 5). When using MI3 as a filter, the node *querer* occurred 53 times and had 14 collocates, all of which were function words, while the node *want* had higher values for frequency (290) and collocates (78). This considerable increase in values in the self-translated corpus indicate evidence of normalisation, since frequency shows a high number of repeated words and traces of explicitation, which is one of the features of normalisation.

Figure 5: GraphColls for the pair of nodes *querer* and *want*



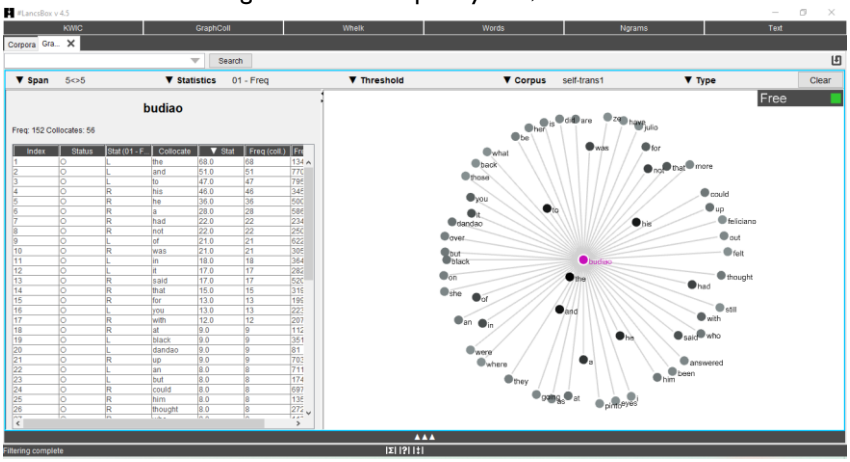
Source: Author (2024).

In brief, the contributions of the small corpora were evaluated critically, and an even distribution of the words in the Portuguese terms compared to the self-translation in English was noted, as shown in the graphs. Simultaneously, this type of self-translation tends to include fewer idioms and more lexical collocations, which can be understood as a trace of normalised text.

4.2.2 The node *Budiao* and the plot analysis

The node *Budiao*, which was the proper name of an important character in the novel, had a raw frequency of 152 and comprised 56 collocates, i.e., it gathered another 56 different terms that constituted what can be said to be a "plot around itself". In other words, to understand the graphs below (Figures 6, 7, and 8) depicting this particular node, it is imperative to consider the concepts of span, collocation network (Brezina *et al.*, 2015), and plot. Plot here is a metaphor that means that each node and its collocates revealed scenes and facts in the plot, since the corpus consisted of a chained account; that is, a sequence of historical facts. Thus, the connections linking terms might not be direct, and exclusivity is difficult to find in a single-text corpus.

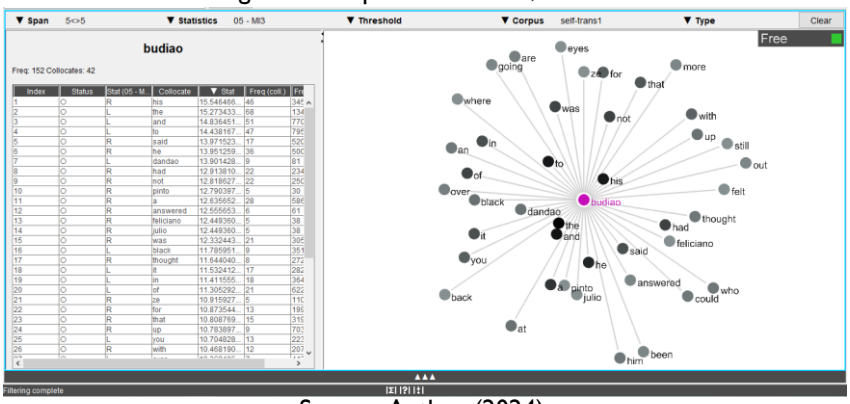
Figure 6: Raw frequency 152, Coll. 56



Source: Author (2024).

At this point, the concepts of span and collocation networks provide valuable insights into plot analysis. Let us consider the main collocate for the node *Budiao* in the self-translation, which was the collocate *the* (Figure 6). The KWIC tool revealed that *the* (as well as the source language equivalents, *o*, *os*) did not appear immediately before the node *Budiao*, as would be expected. The span varied between L_4/ L_5 and R_5. Such a distance between the collocate *the* and node *Budiao* indicated the context of the account, and that their relationship differed from multitext corpora. Moreover, with the span varying between L_4/ L_5 and R_5, the collocate *the* was not exclusive of the node *Budiao*. The same was the case for the single-text corpus in Portuguese, thus showing that the span L_5 and R_5 indicate evidence of normalisation on the author's part in an attempt to decrease the distance between the original and the self-translation for reasons of historical importance.

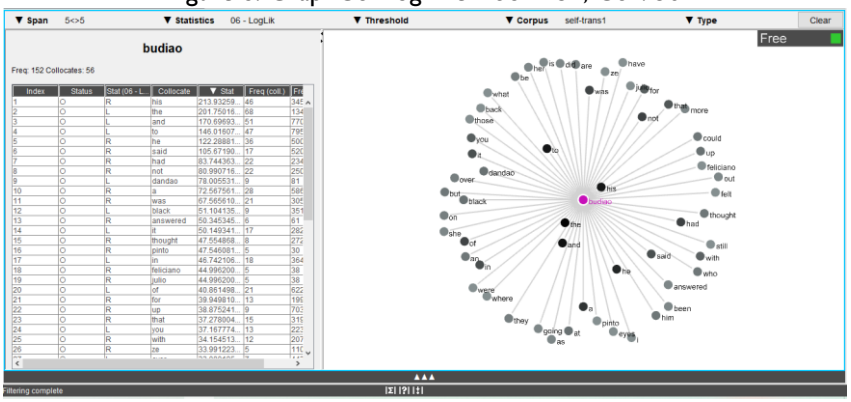
Figure 7: GraphColl MI3 152, Coll. 42



Source: Author (2024).

As said before, the raw frequency is only the first step in an analysis. Employing AMs to produce other results is also possible. For example, MI3 is an AM that is used to reduce the low-frequency terms, thus revealing the most important collocations and their respective networks. With this graphic representation, it is possible to assemble the significant terms to be analysed first in a more filtered display. The graphs illustrated the strongest collocation networks. In Figures 7 and 8, unlike raw the frequency, the top collocate for *Budiao* was *his* using MI3.

Figure 8: GraphColl log-likelihood 152, Coll. 56



Source: Author (2024).

Based on Figures 6, 7, and 8, it is possible to assume that the collocation network in the first level between *his* and *Budiao* filtered and presented all the most important passages that included this character and their relationship of possession. However, the AM loglikelihood, which reduced or even eliminated the null hypothesis, depending on the discourse background, denoted a hybrid collocation network result. It simultaneously presented the same numbers for the frequency and the collocates as the raw frequencies, 152 and 52, respectively, and its top collocate was *his*, which was also the case in MI3. It can be concluded that the AM also shed light on context analysis.

5. Conclusion

Although reflections on how self-translation traditionally contributes to cultural and migration analyses, this work aimed to provide some different insights. While the role of self-

translation in cultural and migration analyses is well established, the goal of this research was to provide a quantitative approach to João Ubaldo Ribeiro's (1984, 1989) historical novel *Viva o Povo Brasileiro* and its self-translation, *An Invincible Memory*. The aim was not to add to the existing literature, but to acknowledge the author's and self-translator's merit in their choices to highlight their value.

This small-scale, exploratory study suggested that all the methods and approaches are interconnected, and form chains that complement each other throughout the translation. This might shed light on another type of view regarding a more practical purpose of self-translation, which somewhere is between bilinguals/polyglots' translations, and standard translations, and suggests a way to resolve the dilemma of normalisation. This is interesting because normalisation may be seen not as a negative outcome, but as a solution found by the author in his attempt to communicate his people's history. João Ubaldo Ribeiro did not consider his work to be a masterpiece of self-translation, but a friendly presentation of his people's fairly complex history.

As for #LancsBox, it presented itself as a quite useful tool for the analysis as the translation was into English language; however, it might require some adjustments for different target languages. Concordances have been found to be useful tools, not only in Corpus Linguistics but also in Translation Studies, and to provide interesting and novel results. This said, #LancsBox can be used in the analysis of both literary and scientific translations, the latter is the goal as future work.

Returning to the question of Neopositivism and the "dream" of mathematical accuracy, not only in history but in science in general, in fact, it seems that this "dream" was realised a century later via what Pazienza *et al.* (2005) called hybrid approach, which refers to a computational statistical-linguistic method for translation studies, as demonstrated throughout this paper. The search for historical accuracy was "translated" into statistical analysis, which demonstrated the real value of historical novels, self-translations, and the combination thereof.

As could be seen in this study, as João Ubaldo Ribeiro was his own translator, a more informal register without many explanations, as in the source text, was expected; however, the traces of normalisation in the text structure revealed quite the opposite. The overall tendency to exaggerate characteristics of the text (Baker, 1996) played a role in the self-translation. At the same time, normalisation could be understood as the broadest amongst the features of translated text, as it contains all the others, which can be considered subsections of normalisation.

It was also noted that the self-translation aimed to reproduce the account's events in the most similar pattern as possible; after all, these were historical facts. Self-translation was used to explain what was culturally inevitable by elucidating it, and included every minor detail within the text itself, without footnotes or endnotes, which again indicated the use of normalisation. Thus, normalisation and self-translation could not be characterised as less creative phenomena. The first could be seen as an effective strategy in the translation process that functioned well in the context of this self-translation. The latter was used in João Ubaldo Ribeiro's work not as an adoption of biculturalism and bilingualism, but as a resource to reveal the characteristics of his own work, which was a historical novel.

In light of these considerations, the historical novel communicated the history of a people without epic resources, but placed value on facts in an attempt to reproduce its true essence, as well as human nature in its most authentic state. Thus, the present work was an attempt to use a

quantitative method to describe the successful combination of normalisation and self-translation as useful resources that are as completely compatible with constructions, processes, and analyses as are any other types of translations.

All the aforementioned conjectures and analyses could be applied to any other work, which proves the equivalent value of self-translation to any other type of translation. Therefore, the main concern in this study was to contribute by providing a more horizontal perspective of non-standard translation processes, such as self-translation, to mitigate the historical hierarchy that has been constructed based on the perspectives of Western translation studies.

References

Antunes, M. G. (2009). Marcas no texto autotraduzido: o caso de João Ubaldo Ribeiro. *Ipotesi*, *13*(1), 57–65.

Baker, M. (1995). Corpora in Translation Studies: An Overview and Some Suggestions for Future Research. *Target*, *7*(2), 223–243. https://doi.org/10.1075/target.7.2.03bak

Baker, M. (1996). Corpus-based Translation Studies: The Challenges that Lie Ahead. In H. Somers (Ed.), *Terminology, LSP and Translation: Studies in Language Engineering in Honour of Juan C. Sager* (pp. 175–186). John Benjamins.

Baker, M. (2000). Towards a Methodology for Investigating the Style of a Literary Translator. *Target, 12*(2), 241–266. https://doi.org/10.1075/target.12.2.04bak

Baker, M. (2007). Patterns of Idiomaticity in Translated vs. Non-translated English. *Belgian Journal of Linguistics*, (21), 11–21. http://dx.doi.org/10.1075/bjl.21.02bak

Baker, M., & Saldanha, G. (Eds.). (2020). *Routledge Encyclopedia of Translation Studies*. Routledge.

Brezina, V. (2018). *Statistics in Corpus Linguistics: A Practical Guide.* Cambridge Press.

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. https://doi.org/10.1075/ijcl.20.2.01bre

Brezina, V., Weill-Tessier, P., & McEnery, A. (2021). Weill-Tessier, P. #LancsBox v. 6.x. [software package].

Camargo, D. C. (2006). Tradução de textos de áreas especializadas e a presença de traços de normalização. *Tradterm, 12*, 55–67.

Evert, S. (2005). The Statistics of Word Cooccurrences: Word Pairs and Collocations [Doctoral thesis]. Universität Stuttgart. https://elib.uni-stuttgart.de/handle/11682/2573

Evert, S. (2008). Corpora and collocations. In A. Lüdeling & M. Kytö (Eds.), *Corpus Linguistics: An International Handbook* (pp. 223–233). Walter de Gruyter.

Fantinuoli, C. (2016). Revisiting Corpus Creation and Analysis Tools for Translation Tasks. *Cadernos de Tradução, 36*(1), 62–87. https://doi.org/10.5007/2175-7968.2016v36nesp1p62

Firth, J. (1957). *Papers in Linguistics.* Oxford University Press.

Frérot, C. (2016). Corpora and Corpus Technology for Translation Purposes in Professional and Academic Environments. Major Achievements and New Perspectives. *Cadernos de Tradução, 36*(1), 36–61. https://doi.org/10.5007/2175-7968.2016v36nesp1p36

Gablasova, D., Brezina, V., & McEnery, T. (2017). Collocations in Corpus-Based Language Learning Research: Identifying, Comparing, and Interpreting the Evidence. *Language Learning, 67*(S1), 155–179. https://doi.org/10.1111/lang.12225

Gries, S. T. (2013). 50-something Years of Work on Collocations: What is or Should be Next... *International Journal of Corpus Linguistics, 18*(1), 137–166. https://doi.org/10.1075/ijcl.18.1.09gri

Groot, J. D. (2010). *The Historical Novel.* Routledge.

Grutman, R., & Bolderen, T. V. (2014). Self-Translation. In S. Bermann & C. Porter (Eds.), *A Companion to Translation Studies* (pp. 323–332). Wiley-Blackwell.

Jung, V. (2002). English–German Self–Translation of Academic Texts and its Relevance for Translation Theory and Practice [Doctoral thesis]. Heinrich–Heine–Universität Düsseldorf.

Lukács, G. (2011). *O Romance Histórico.* (R. Enderle, Trans.) Boitempo.

Maestri, M. (2002). História e romance histórico: fronteiras. *Novos Rumos, 17*(36), 38–44.

McEnery, T. (2006). *Swearing in English: Bad Language, Purity and Power from 1586 to the Present.* Routledge.

Paiva, P. T. (2011). Traços de tradução em artigos de anestesiologia: uma comparação entre os resultados de um corpus paralelo e de um corpus comparável. *Estudos Linguísticos, 40*(2), 1158–1171.

Pazienza, M., Pennacchiott, M., & Zanzotto, F. M. (2005). Terminology Extraction: An Analysis of Linguistic and Statistical Approaches. Berlin Springer-Verlag.

Pecina, P. (2010). Lexical Association Measures and Collocation Extraction. *Language Resources and Evaluation, 44*(1-2), 137–158. https://doi.org/10.1007/s10579-009-9101-4

Popovič, A. (1976). Aspects of Metatext. *Canadian Review of Comparative Literature, 3*, 225–235.

Ribeiro, J. U. (1984). *Viva o povo brasileiro.* Nova Fronteira.

Ribeiro, J. U. (1989). *An Invincible Memory.* Harper & Row Publishers.

Ribeiro, J. U. (1990). Suffering in translation. *P.T.G. Newsletter, Portuguese Translation Group, 3*(3), 3–4.

Scott, M. (2004). *WordSmith Tools Version 4.* Oxford University Press.

Scott, M. N. (1998). Normalisation and Reader's Expectations: A Study of Literary Translation with Reference to Lispector's A Hora da Estrela [Doctoral thesis]. University of Liverpool.

Shread, C. (2009). Redefining Translation through Self-Translation: The Case of Nancy Huston. *French Literature Series, 36*, 51–61.

Shuttleworth, M., & Cowie, M. (2014). *Dictionary of Translation Studies.* Routledge.

Toury, G. (2012). *Descriptive Translation Studies and Beyond*. John Benjamins.

Venuti, L. (1995). *The Translator's Invisibility.* Routledge.

Xia, Y. (2014). *Normalization in Translation: Corpus-based Diachronic Research into Twentieth-century English–Chinese Fictional Translation.* Cambridge Scholars Publishing.

Zanettin, F. (1998). Bilingual Comparable Corpora and the Training of Translators. *Meta*, *43*(4), 613–630. https://doi.org/10.7202/004638ar

Notes

Authorship contribution

Conception and preparation of the manuscript: D. S. Santos
Data collection: D. S. Santos
Data analysis: D. S. Santos
Discussion of results: D. S. Santos
Review and approval: D. S. Santos

Research dataset
Not applicable.

Funding
Not applicable.

Image copyright
Not applicable.

Approval by ethics committee
Not applicable.

Conflict of interests
The authors declare no conflicts of interest.

Data availability statement
The data from this research, which are not included in this work, may be made available by the author upon request.

Publisher
Cadernos de Tradução is a publication of the Graduate Program in Translation Studies at the Federal University of Santa Catarina. The journal *Cadernos* de Tradução is hosted by the Portal de Periódicos UFSC. The ideas expressed in this paper are the responsibility of its authors and do not necessarily represent the views of the editors or the university.

Section editors
Andréia Guerini – Willian Moura

Technical editing
Alice S. Rezende – Ingrid Bignardi – João G. P. Silveira – Kamila Oliveira

Article history
Received: 26-02-2024
Approved: 12-08-2024
Revised: 27-09-2024
Published: 09-2024