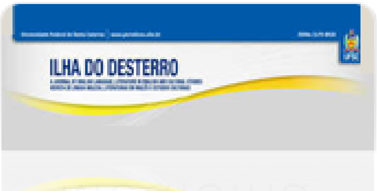


Ilha do Desterro
ISSN: 2175-8026
Universidade Federal de Santa Catarina

Durão, Adja Balbino de Amorim Barbieri; Kloeppel, Paulo Roberto
Children's Literature Parallel Corpora: a Hybrid Experimental Model to Evaluate Transfers of Language Complexity Via Linguistic Transcoding
Ilha do Desterro, vol. 71, no. 1, 2018, January-April, pp. 27-51
Universidade Federal de Santa Catarina

DOI: 10.5007/2175-8026.2018v71n1p27

DOI: 10.5007/2175-8026.2018v71n1p27

CHILDREN'S LITERATURE PARALLEL CORPORA: A HYBRID EXPERIMENTAL MODEL TO EVALUATE TRANSFERS OF LANGUAGE COMPLEXITY VIA LINGUISTIC TRANSCODING

Adja Balbino de Amorim Barbieri Durão[*]
Universidade Federal de Santa Catariana

Paulo Roberto Kloeppel[**]
Universidade Federal de Santa Catariana

Abstract

This article aims at proposing a hybrid model to evaluate language complexity of source and target texts written both in English and Portuguese so that one can analyse what extent language complexity has been transferred from a text to its translation. Here, *hybrid model* points to paralleled approaches to lexical repetition, lexical diversity and lexical density, readability and word unusualness with the help of some Corpus Linguistics tools. The article also stands for developing adjustments to Paul Nation's word family lists and Gunning's GFI formula so that they can be applied to the Portuguese language. Aiming at checking the reliability of the model, the article also presents a case study based on contrastive investigations on *The Secret Garden* by Burnett and its translation into Portuguese *O Jardim Secreto* by Perota and Carvalho.

Keywords: Linguistic transcoding; Language Complexity; Children's Literature; Parallel Corpora.

[*] Professora doutora de pós-graduação em Linguística e Estudos da Tradução e coordenadora da Pós-graduação em Estudos da Tradução da UFSC. Seu e-mail é: adjabalbino@gmail.com.

[**] Mestre em Estudos da Tradução na área de Lexicografia, tradução e ensino de línguas, pela Universidade Federal de Santa Catarina (2015), possui graduação em letras-inglês pela Universidade Federal do Paraná (2001). Seu e-mail é: kpersonal2012@gmail.com.

Introduction

When one wonders about the age range appropriateness of some books of classical Children's Literature , she or he will probably endure a time-consuming and hardly conclusive task. In effect, those who have ever tried to check the age appropriateness of *The Adventures of Tom Sawyer* by Mark Twain or *The Wonderful Wizard of Oz* by L Frank Baum, or any other, have probably become amazed by the various results found in the WEB. For instance, in the website The Scholastic Store, one finds Baum's book is suitable to children 8-11 years old, but the Common Sense Media says it is appropriate to children over 9 years old. Similarly, while this website ranks Twain's book as fitting to children over 11 years old, The Scholastic Store classifies it as appropriate for those over 12. And most distinctively, the website Kidzworld classifies this book as appropriate for those over 9 years old. Even Mark Twain himself has made the task more complicated by stating, in the preface for his book, that "although his book was intended mainly for the entertainment of boys and girls, he hoped it would not be shunned by men and women on that account" (Twain, 1874), which suggests that the content and language level of his book are not restricted to children and juveniles' receptions. Also, this seems to state that his writing style may be refined enough to fulfil the reading expectations of adults as well as simple enough to be read by young girls and boys.

In addition, in parallel with these discrepancies concerning age appropriateness of children's books, one scarcely is provided with clear-cut criteria, whether based on language or content, or both, adopted by the sites to evaluate Children's Literature. As the appropriateness of books' content according to ages seems to be bounded by socio-cultural beliefs and expectations, which can vary from country to country, from one region to another of a country, and even vary from people to people, evaluating language complexity seems to be a safer starting point to classifying children's books, since it may be less controversial than otherwise.

So being, as far as we are concerned, it becomes even tougher to judge age appropriateness of translations of Children's Literature, considering that, besides the uncertainties previously raised, the effect of source languages on translations is strong enough to make the languages of translated texts different from the target languages (MCENERY and XIAO, 2007, p. 6). Hence, the language of translations of children's books may be less naturally processed by their target readers, and by consequence the latter may be exposed to language that is more complex than that which target audiences are used to, sometimes even more complex than that of the source text. To sum up, due to translation processes, unlikely collocations, weak semantic prosodies and not frequently used lexicon and colligations may come out in translated texts to such an extent that the language of translated texts may end up being "translationese". Concerning this, Baker advises,

> …an awareness of aspects of information flow and potential ways of resolving tension between syntactic and communicative functions is important in translation. The fact that certain strategies which can be shown to be useful in translation have not been made use of so far suggests that translators are simply not aware of them, rather than that they are familiar with them but consciously or subconsciously choose not to use them. (Baker, 2004, p.172)

Focusing on "consciousness and sub-consciousness" and the strength of the effect of source languages on translations from the above citations, reasonable reasons for the language of translated texts being translationese may stem from the fact that, according to Schleiermacher's model of translation, "in the domain of cultural capital […] translation can most clearly be seen as to construct cultures" (Bassnet and Lefevere, 2001, p. 7) and on the fact that translation "is vital to the interaction between cultures" (ibid. p. 6). In effect, as the latter presupposes transfers from one culture to others and the former points to the importance of such transfers in the building of cultural capitals of target cultures, it is acceptable that translators consciously or subconsciously aim at transferring as much as possible traces of the writings of source texts to target texts, in order to make authors' writing styles and other inter-related textual features of their books available to target cultures. Nevertheless, such tentativeness may culminate in translations whose language complexity may be beyond their target readers' language skills. This seems to be especially true when we concern Children's Literature, since children and juveniles' language skills are still under construction, mostly according to the formal educational level in which they are enrolled.

However, this is also true in the case of Literature in general, considering that, due to differences among writing styles, even texts written aiming at highly formally educated readers are hardly expected to fit language skills of all audiences. For instance, reading Joyce's innovative language found in *Ulysses* requires different language skills from those needed for reading the poetic language of *The Picture of Dorian Gray* by Oscar Wilde. Therefore, transferring Joyce's innovative language style and Wilde's poetic language through translational processes is crucial to those translations that aim at audiences, within target cultures, that align to those of the source text culture. Nonetheless, such procedure may yield translated texts accessible only to small audiences. Conversely, supported by the thoughts of the Skopos Theory, translators may choose not to transfer all marks of Joyce's writing style and normalize their translations in order to reach different audiences. Yet, this is risk-taking since, for example, instances of Joyce's innovative language make up chains of isotopies in *Ulysses*, and these chains in their turn build bigger chains of isotopies that constitute what Halliday and Hasan call "texture" (1976, p. 2), which cannot be "broken", since it is made of inter-related cohesion devices, such as reference, nominal and verbal substitutions, ellipsis and lexical cohesion (Halliday and Matthiessen, 2004). Recontextualizing Barthes' allegory of the text (1987, p.82), it is like a spider web, if you break some threads of a spider's web you do not smash the whole web, but that snatched part fails to fulfil its function as an insect trap.

The concept of isotopies is a key one in Corpus Linguistics due to its relationship with word frequency, which is the basis of empirical investigations of texts that are electronically carried out. Roughly, the term "isotopies" is designated to refer to reoccurrences of semantic, phonological and morphosyntactic features (ECO, 1984; GREIMAS, 1966) that create semiotic networks that often characterize texts as unique among others within a field.

That said, by empirically investigating the isotopies of source texts and the target texts, one can evaluate to what extent features of the former have been transferred to the latter, which is one of the purposes of this article. Thus, assuming that the previous paragraphs have been effective enough to highlight the role of writing styles and their inter-related text features in the ranking of age appropriateness of Children's Literature, this article proposes the application of some electronic tools of Corpus Linguistics to a small parallel corpus, made of *The Secret Garden* by Frances Hodgson Burnett and its translation to Portuguese by José Luiz Perota and Bianca Carvalho, in order to investigate transfers of language complexity to the translated text.

Aiming at reaching its goals, this article (i) presents some Corpus Linguistics (CL) concepts, tools and practices that are applicable to bidirectional contrastive studies concerning levels of language complexity of source text (ST) and target texts (TT); (ii) presents an experimental corpus-assisted method for evaluating language complexity, crossculturally; (iii) for the sake of being more pedagogical, it presents a case study in which the experimental model is tested via contrastive investigations; (iv) and finally, some conclusions drawn upon the outcomes from the case study are presented, although – and this must be clear – this paper focuses rather on presenting the model itself than on the outcomes of the case study, due to limitations concerning corpus magnitude, and by consequence, statistical representativeness.

I - Parallel Corpora and their applicability in bilingual contrastive studies

When one thinks about levels of language complexity, naturally vocabulary comes into mind. And, as the latter is roughly measured by both lexical diversity and lexical repetition, measuring the frequency of occurrence of words in a text is the primary step to follow, since the frequency of occurrences of distinct words (*tokens*), of text is the bases for CL tools to calculate mathematic averages, frequency ratios and means deviation, which are key-concepts to Statistics. In this regard, CL software, such as WordSmith Tools (Scott, 2010) and AntConc (Anthony, 2014), provide their users with word lists and the frequency of occurrence of words, from which other CL tools calculate Type-Token Ratios (TTR) and Standard Type-Token Ratios (STTR), which respectively measure the proportions of lexical repetition and lexical diversity of texts, in relation to the total number of words (*tokens*). Nevertheless, these ratios have to be carefully read since, as the developer of WordSmith software, Mike Scott says, "TTR and

STTR are both pretty crude measures even if they are often assumed to imply something about 'lexical density'" (SCOTT, 2015, p. 318). What is behind this quotation is the fact that TTRs are highly affected by differences concerning text length, as well as STTRs are affected by repetitions of content-driven vocabulary (Ibid.). Nonetheless, when single texts are contrasted to their translation, TTRs and STTRs may be helpful to acquire overviews of the vocabulary of source and target texts, once the latter tend to be similar to the former in terms of both text length and are not supposed to differ in terms content.

Another way of analyzing the vocabulary level of texts is measuring their lexical density, which differs from lexical diversity owing to the fact that it takes word complexity into account, while lexical diversity ratios take into account only the number of distinct words of texts, regardless their complexity. That is to say, texts may be lexically complex but not lexically diverse, or the reverse. To measure the lexical density of texts AntWordProfiler 1.4.0w tool (Anthony, 2013) ranks portions of vocabulary of texts according to levels of lexical complexity, being the last ones based on lists of the most frequent words of corpora that are taken as statistically representatives of languages. The core of this measure is: the higher the frequency of a word, the lower is its complexity, once such word is highly primed by language speakers (Hoey, 2005). For instance, based on the research of Paul Nation (Nation, 2004; Nation & Webb, 2011), the AntWordProfiler 1.4.0w tool contrasts the vocabulary of a text with Nation's word family lists, which contain the most frequent words of BNC and COCA, in order to measure which amount of this or that word is among the ones classified as of levels 1, 2 and 3 of Nation's word families, which stand for three different levels of lexical complexity.

Analyzing language level of texts, nevertheless, comprises more than simply investigating their vocabulary; it also involves analyzing sentence length, since the longer sentences are, the more difficulty to be read and processed they become, given that their intricate texture demands higher levels of reading and comprehension skills than do textures of short ones. In this regard, many word counters provide their users with numbers of sentences containing, say, 2, 3 4, 5 words per sentence and calculate average length of sentences, in a way that allows users to overview levels of complexity of texts. What is implicit in this logic is the readability of texts, which can be quantitatively measured by calculating the Gunning Fog Index (Gunning, 1968), which is based on mathematic relations involving average sentence length and number of complex words. Unfortunately, this index cannot be strictly applied to Portuguese texts, since English words containing three syllables or more are taken as complex ones by the Gunning Fog Index (GFI), but in Portuguese three-syllable words are very common. Besides, the Gunning Fog Index expresses levels of reading skills based on the American Educational System. Nevertheless, it is our assumption that, with some adjustments, the GFI can be applied to Portuguese as well. This will be set forth in the case study.

That said, despite some limitations of the previously discussed ratios and CL tools, the gathering of their outcomes can give researchers overviews of language complexity of texts that can be contrasted with overviews of their translation, so that one can empirically compare the levels of language of texts.

Moving forward, levels of language of texts can be investigated in the domains of language systems and textual genres. That is to say, they can also be measured by contrasting frequencies of occurrences of lexical items and/or morphosyntactic structures in a text with their frequencies of occurrences in bigger corpora that are taken as statistically representative of a language or a textual genre. In this regard, both the WordSmith tools and the AntConc provide their readers with lists of keywords , which are the words of a text whose frequencies are unusually high or low in relation to a reference corpus. This tool provides "a useful way to characterise a text or a genre. Potential applications include: language teaching, forensic linguistics, stylistics, content analysis, text retrieval" (Scott, 2015, p. 229). In practice, for instance, by contrasting a range of Children's Literature texts, written by the same author – a small corpus – to a bigger corpus – the reference corpus – made of texts of the same genre written by other authors, it is possible to identify marks of that author's writing style in KeyWords lists (BERBER SARDINHA, 2004). On the other hand, by contrasting the same small corpus to a bigger reference one made of literary texts, preferably one that does not contain Children's Literature, it is possible to identify registers and discourse marks that are characteristic of this sort of literature in given culture (Ibid.). Broadening this, when analogous contrasts are based upon corpora made of source texts and their translations, the KeyWords tool is a powerful tool to evaluate to what extent registers, marks of discourse and authors' writing style have been transferred through translational processes.

Still in the realm of unusuality of frequencies of words, the Log-likelihood Coefficient, which is the basis for generating KeyWords lists, allows CL researchers to analyze to what extent one finds a word or a group of words (collocations) whose frequencies of occurrences in a corpus are unusually high or low in relation to a corpus that is taken as representative of a language or textual genre.

Another way of empirically investigating language level of texts comes from the key concept "collocation", which, according to Hoey, "is, crudely, the property of language whereby two or more words seem to appear frequently in each other's company (e.g. *inevitable* + *consequence*)" (HOEY, 2005, p.2). Or, as it is defined in *English Grammar Today: The Cambridge A-Z Grammar of English*, "collocation refers to how words go together or form fixed relationships" (Carter, McCarthy, Mark and O'Keeffe, 2016). From this concept comes the concept of "Semantic Prosody", which stands for the senses, bad or good, that some collocations create (Sinclair, 2003). For example, Sinclair has studied the collocates that "set in" establishes and ended up concluding that "set in" tends to be related to unpleasant events since it often collocates with words such "rot", "despair", "decadence", "prejudice", "disillusion", to mention a few (Ibid). According to him, the term comes from "semantic because it deals with meaning, and prosody because it typically ranges over combinations of words in an utterance rather than being attached just to one" (Ibid. p.116). In this regard, Hoey's Priming Theory states that in natural language processing frequent collocations are primed by the brain, "and ambiguity (or humour) will always result from our use of a word in ways not in accordance to with this priming" (Hoey, 2005, p. 81). In other words, unusual collocations are harder to be processed than primed ones. And this is the point where this concept becomes a key one to this article, since, pushed by the force that the languages of source texts have over the languages of translated texts (McEnery e Xiao, 2007), translators may "coin" unusual or non-primed collocations, which certainly make the language level of such translated texts higher than that of their source texts. Indeed, by being semantic prosody rather a psychological feature of language in use (Hoey, 2005), it is also culturally bounded; therefore, semantic prosodies that are primed within a culture may not be primed in others. On the other hand, using unusual or non-primed collocations may be a writing strategy to build up chains of isotopies of texts, which also take their role in the building up of writing styles. Consequently, if translators replace such collocations by interlinguistic correspondent ones that are more usual or primed within target cultures, the language level of translated texts will be simplified, as well as writers' style will be affected.

Finally, as all the concepts so far discussed can be approached by tools of Corpus Linguistics, the former, except for collocations and semantic prosodies, have been taken as a basis for developing an experimental model for evaluating language levels of source and target texts. The reason for not including investigations concerning collocations and semantic prosodies in the experimental model is purely practical, since investigating them would only demand far more pages than it is expected for academic articles. In reality, given the complexity that involves these concepts, investigating them via LC deserves specific articles.

II - Evaluating language level: an experimental corpus-assisted model

Basically, the experimental model has been drawn upon contrastive empirical investigations that approach two texts, herein *The Secret Garden* (1905) and *O Jardin Secreto* (2012), and two reference corpora made of British and Brazilian Children's Literature books, being the English ones published from 1898 to 1937 and the Portuguese ones from 2011 to 2014. The core of the model, as has been discussed so far, is to contrast the overviews of the language level of the investigated texts.

In constructing the Children's Literature reference corpora, special attention has been devoted to collecting texts that are contemporary to the source and target texts under investigation as well as to building two corpora quite similar in size, which, according to Berber Sardinha (2004), have to be at least five times bigger than the investigated texts. These precautions have tried to get as much corpus representativeness and corpus balance (Sinclair, 2004) as possible, aiming at minimizing the chances of happening distortions in the outcomes, such as those mentioned when the TTR and STTR measures were approached.

The overviews of the language level were obtained on the basis of the following criteria:

1. Applying the WordSmith 6.0 WordList tool, the TTR and STTR of the ST and TT were obtained and contrasted so that one could identify possible superiority, equality or inferiority concerning lexical repetition and lexical diversity of the ST and TT.

2. The obtained TTR and STTR of the texts contrasted to their counterparts of the British and Brazilian Children's Literature reference corpora so that one could evaluate whether these measures align to those of the reference corpora or not.

3. With the help of AntWordProfiler 1.4.0w, the lexical density of the source and target texts was measured and contrasted.

4. From the Statistics of the WordList interface of WordSmith, the average lengths of sentences and the numbers of complex words of the source and target texts were copied and inserted in the formula of the Gunning Fog Indexes. The formula results were contrasted in order to identify lexical density superiority, equality or inferiority between the texts. This is going to be better explained in the case study.

5. On the basis of the British and Brazilian Children's Literature reference corpora, the KeyWords lists of the source and the target texts were obtained in order to look for interlinguistic correspondence likeness that could point to transfers of Baum's style to the translated text.

6. Using the British National Corpus (BNC, 2010) and *Corpus Brasileiro* (Berber Sardinha, Alambet and Moreira Filho, 2013), as reference corpora, both taken as representative of their respective languages, the Log-likelihood coefficient of some selected keywords has been obtained as a means of checking to what extent the frequencies of those words in the target and source texts are unusually high or low in relation to the reference corpora.

III - The case study

In order to be as pedagogical as possible, the case study follows the above presented sequence of steps, in which some readings of tables and graphs are placed, and below the readings some words are devoted to draw previous conclusions.

Before moving to such steps, let us spend some time discussing the treatment given to the corpora and sub-corpora approached by the case study. Following the simplest concept of a corpus, as being made of a single text or a collection of texts, the source text *The Secret Garden* and target text *O Jardim Secreto* were considered as being two sub-corpora of a parallel corpus. And, as the electronic tools of CL recognize strings of characters between two blank-spaces, not words attached to natural languages, sometimes the English and Portuguese texts were paralleled when applying one of the tools of WordSmith 6.0. Nevertheless, when the CL tools worked upon reference corpora, these texts had to be treated separately. Concerning the reference corpora, the case study was developed on the basis of two Children's Literature ones, whose collecting has already been featured, as well as on the basis of the BNC, since the source text is a British one, and on the basis of the *Corpus Brasileiro* and *Corpus do Português* (Davies and Michael, 2006), all taken as statistically representative of their respective languages. In the building of the Children's reference corpora, the collected texts were cleaned up; that is to say, all parts of the books that do not stand for the stories themselves, such as authors' prefaces and acknowledgements, were excluded from the corpora. The architecture of the Children's Literature reference corpora is as following.

Texts	**Tokens**
The Railway Children by Edith Nesbit (1906)	60,255
The Dark Frigate by Charles Boardman Hawes (1923)	70,995
Moonfleet by J. Meade Falkner (1898)	83,771
The Story of Doctor Dolittle by Hugh Lofting (1920)	26,371
The Lost World by Sir Arthur Conan Doyle (1912)	79,061
Peter and Wendy by James Matthew Barrie (1911)	47,438
The Hobbit by J. R. R. Tolkien (1937)	96,951
Corpus Size	**464,842**

Table 1 - British Children's Literature Reference Corpus

Portuguese texts	**Tokens**
Reinações de Narizinho by Monteiro Lobato (2011)	79,177
Conexão Magia by Helena Gomes (2011)	125,787
A Profecia de Samsara by Leticia Vilela (2014)	47,092
Luna Clara & Apolo Onze by Adriana Falcão (2013)	38,930
O fazedor de velhos by Rodrigo Lacerda (2013)	35,934
O mestre dos games by Afonso Machado (2011)	32,342
Amanda e os Nanorobôs by Eliú Quintiliano (2014)	77,178
Pena dourada by José Luiz da Luz (2013)	20,265
Corpus Size	**456,705**

Table 2 -Brazilian Children's Literature Reference Corpus

In the running of WordSmith Tools 6.0, Stop Words Lists were applied, being them made of the most frequent grammar words, those that rather take part on the structuring of texts than on their meaning, and those words that are known by computer science as noisy words, *i.e.*, words or symbols that often appear in books, but do not belong to their texts themselves, such as number of chapters and pages, the words "chapter" and "page" themselves, and alike. The purpose of using Stop Words Lists is to exclude such words from analysis because they are not relevant to the investigation. For instance, "you might want to make a word list or analyse key words excluding common function words like *the, of, was, is, it*" (Scott, 2010).

Let us now move forward and present the first step of the case study.

III - 1 Contrasting the TTRs and STTRs of the source text to the target text

Lexical repetitions and lexical diversities of the ST and TT were analyzed by WordSmith Word List tool.

Texts	**Tokens**	**Types**[1]	**TTR (%)**	**STTR (%)**
The secret garden	81,314	4,944	6.08	38.93
O jardim secreto	78,727	7,287	9.26	44.29
ST / TT Discrepancies[2]			-3.18	-5.36

[1] Distinct words; [2] Discrepancies are expressed in relation to Source Text data.

Table 3 - TTR and STTR of ST and TT

Keeping in mind the aforementioned crudeness of these measures, initially we can say the lexical diversity of the ST is 5.36 percentage points lower than that of the TT, but its level of lexical repetition is higher than it is in the TT, since the lower the TTR, the higher is the lexical repetition; and the higher the STTR, the higher is the lexical diversity (BERBER SARDINHA, 2004). The behaviors of these ratios are somehow expected since, according to Kloeppel's research , lexical repetitions are more frequent in English literary texts than they are in Portuguese ones (Kloeppel, 2015). Therefore, in terms of language complexity, one could draw any conclusion only if these ratios were very higher or very lower than the ratios found in Kloeppel's research.

III - 2 Contrasting the TTRs and STTRs of source and target texts to the analogous of Children's Literature reference corpora

Similarly to what has been done with the ST and TT, the TTR and STTR of English and Brazilian Children's Literature corpora were obtained.

Corpora	**Tokens**	**Types**	**TTR (%)**	**STTR (%)**
The secret garden	81,314	4,944	6.08	38.93
British Literature Ref. Corpus	463,609	17,307	3.73	40.92
O jardim secreto	78,727	7,287	9.26	44.29
Portuguese Lit. Ref. Corpus	455,723	26,672	5.85	47,52

Table 4 - TTR and STTR of Children's Literature reference corpora

Observing Table 4, we notice that lexical repetition is approximately 60% (3.73/6.08) **lower** in *The Secret Garden* in relation to it in the British Literature Reference Corpus, as well as its lexical diversity is 4.86 (40.92/38.93) percentage points lower. Similar behaviors are noticed in the contrast between the *O Jardim Secreto* and Portuguese Literature Reference Corpus. However, when the two reference corpora are contrasted, we notice that lexical repetition is *higher* in the British one and lexical diversity is lower in it. Therefore, quite accurately we can assume that Burnett's writing style is featured by avoidance of lexical repetition, and this has been transferred to its translation. In other words, in terms of lexical repetition the language complexity of the ST and TT seems to be alike.

III - 3 Contrasting the Lexical Density with AntWordProfiler 1.4.0w

As has already been discussed (see section II), the AntWordProfiler 1.4.0w measures the lexical density of English texts on the basis of three levels of complexity as they have been proposed by Nation (Nation, 2011). Basically, this CL tool compares the wordlist of a text with three wordlists of word families, which were generated according to frequency and range data words, in order to locate portions, expressed by percentages of frequencies, within each of the three Nation's levels. For example, if words 'a', 'b' and 'c' make up a 'text', and 'a' and 'c' are among those of level 1, their frequencies of occurrences are summed and are counted as the portion of level-1 words of such text, and if word 'b' is among those of level 2, it is counted as the portion of level-2 words of the same text. Thus the lexical density of this text would be said to be made of 66.66% of level-1 words and 33.33% of level-2 words. On the other hand, if a text contains words 'w', 'x' and 'y' and all of them are among those of level 2, the lexical density of such text would be taken as being 100% of level-2 words. As a result the latter, lexical density differs from that of the former, and by extension, the levels of language complexity of these texts differ as well.

The problem is that Nation's levels are based on the morphosyntactic, semantics and grammar of the English language; hence they do not apply to other languages· Nevertheless, as the AntWordProfiler provides its users with the possibility of setting other level lists, using lists of 1,000, 2,000 and 3,000 most frequent words (MFW) of a Portuguese reference corpus, aligning with Longman Communication 3000 list, which, based on the 390-millionword Longman Corpus Network, states that the "3,000 most frequent words [MFW] in spoken and written English account for 86% of the language" (LONGMAN, 2007), would help us solve the problem. This is so, especially, considering that Longman's percentage aligns to Davies' claim that, by citing Nation (2001), says that the 1,000 MFW account for approximately 80%, the 2,000 MFW for 85%, and the 3,000 account for approximately 89% of this language (Davies, 2005 apud EDDINGTON, 2005, p.106). However, one could raise questions concerning the applicability of the 1,000-level parameter to the Portuguese language or to any other language but the English Language.

On the other hand, Randall's study, concerning the German language, shows that the 3,000 MFW of "a subset of approximately 10% of the BYU/Leipzig Corpus of Contemporary German" (RANDALL, 2006 p. 117) account for 89% of spoken German and account for 80.8 % of literature (Ibid. p. 119), figures that are close to those of Longman's list and Davies' article, despite the fact that "German has a much more complex morphology than English and uses far more compounding of nouns, adjectives and verbs" (Ibid. p. 115). Therefore, supported by Randall's study and the fact that Portuguese morphology, quite similarly to that of German, is more complex than English morphology, it seems reasonable to assume that, at a first glance, the 1,000-word parameter for levels of Longman Dictionary can be set as an alike to distinguish lexical densities of Portuguese texts as well as to measure the transfer of lexicon-based linguistic levels from English source texts to Portuguese target texts.

Even so, based on the frequencies of occurrences of the 776,072,384 words of *Corpus Brasileiro*, obtained from the Linguateca WEB page, with the help of Microsoft Excel, we noticed that the 3,000 top words of this corpus account for approximately 78% of the corpus, about 8-10 percentage lower than those of Longman Dictionary and Davies' article. Although expected, knowing that Portuguese verbs are much more inflected to express verbal aspects than English verbs, this is a high difference considering the magnitude of *Corpus Brasileiro*. Thus, probably the 1,000-word parameter should be enlarged to minimize this percentage difference. And, considering that "the most important aspect of the difference of frequencies between linguistic traits is that they are not random" (BERBER SARDINHA, 2004, p.31, our translation), it seemed reasonable to assume that proportionally adjusting the 1,000-word parameter might work well for Portuguese language. Hence, taking the 390-million-word Longman Corpus Network as reference, such adjustment should be done via a basic linear equation, the so-called rule of three, as follows:

X ———— 776,072,384
3,000 ——— 390,000,000

This rule of three gave us back the figure of 5.97 as the proportional adjustment, meaning that instead of three levels of 1,000 top words each, we should work with 6,000 top words for Portuguese texts, or three levels of 2,000 top words each. This seems to be a reasonable ratio to cope with the superior complexity of Portuguese morphology in relation to English morphology, even because in the WordList of *Corpus Brasileiro*, in the 6,000th word-position the cumulative percentage reaches approximately 86.25%.

In this respect, in order to test the applicability of contrasting 1,000 top English words to 2,000 top Portuguese words, with the help of AntWordProfiler, taken the ST and TT as basis, a simple contrastive test for the verbs 'be', 'have' and 'go' and their respective correspondents *ser*, *estar*, *ter* e *ir* has been carried out, based on the verb forms 'am', "'m', 'is', "'s', 'are', "'re' , 'was', 'were' and 'will be', 'has', 'have' and 'had', and 'go', 'goes', 'went' and 'gone', and the 48 conjugations of the verbs *ser* or *estar,* 24 of the verb *ter* and 24 of the verb *ir*, all in the *presente, pretéritos perfeito* and *imperfeito,* and *futuro* of the Portuguese indicative mood, the last ones due to their role in compound verb forms typical of Brazilian colloquial communication. Before analyzing Table 5, it must be said that when running the above English verb forms in AntWordProfiler, in which the three-word family lists as proposed by Paul Nation are included by default, all verb forms appear in the first 1,000-word base.

Verb forms	Level 1,000	Level 2,000	Level 3,000	**Total %**	**Others[1]**	**Others %**
be	**9**	**0**	**0**	**100%**		
Verb forms	Level 2,000	Level 4,000	Level 6,000			
ser* and *estar	**15**	**15**	**6**	**75%**	**12**	**25%**
Verb forms	Level 1,000	Level 2,000	Level 3,000			
have	**3**	**0**	**0**	**100%**		
Verb forms	Level 2,000	Level 4,000	Level 6,000			
ter	**9**	**5**	**2**	**67%**	**8**	**23%**
Verb forms	Level 1,000	Level 2,000	sLevel 3,000			
go	**2**	**0**	**0**	**100%**		
Verb forms	Level 2,000	Level 4,000	Level 6,000			
Ir	**7**	**3**	**2**	**50%**	**12**	**50%**

[1] The percentage of words, whose frequencies are lower than those of the words in levels 1, 2 and 3.

Table 5 – Testing the proposed 2,000- top-word levels.

Firstly, one can notice that, at least for the verbs 'be', 'have' and 'go', the 1,000-word parameter has not produced results that differ from those obtained by AntWordProfiler using Nation's word-family levels. In relation to Portuguese verb forms, we see that, although they are distributed in a certain decreasing regularity, the verb *ir* is one that the 6,000 MFW cope for only 50%. This may be explained by misuses of some verb forms in colloquial language. For instance, instead of saying, *irei jogar bola amanhã* (I'll play football tomorrow) many Brazilians say *vou jogar bola amanhã*, but *vou* is in the present tense, which is among those 7 in the 2,000-top-word level that is the least complex one. Therefore, considering the rule of thumb of Corpus Linguistics that states that the higher the frequency of use of words the lower is their complexity, we can assume that the verb forms of *ser*, *estar*, *ter* and *ir,* that are among the three levels of Table 5, are the least complex ones for Brazilian speakers to communicate well with their peers.

Drawing our attention the percentages of the Portuguese verb forms in the column "others %" of Table 5, which seem to be very high, they may be easily explained by two linguistic features of spoken Brazilian Portuguese: (i) verb forms of future are more frequent in written texts, since Brazilian people tend to use more some colloquial compound forms using the verb forms of the *Presente do Indicativo* of the verb *ir*, such as instead of saying, *Estudarei matemática amanh*ã (I will study math tomorrow), they say *Vou estudar amanhã*; (ii) a great deal of Brazilians either avoid using inflected verbs for the second person singular and plural or misuse them in such a way that some verb forms as *vais*, *ides*, *estivestes*, *estais,* and some others, have been quite rarely found in spoken Portuguese nowadays. In effect, *vais*, for instance, figures in the 219,336th position in the Word List of the Corpus of Portuguese, as well as *ides* in the 63,014th.

That said, it has been assumed that applying the 2,000-top-word parameter for the three levels of lexical complexity can help us get more accurate quantitative views of the lexical density of Portuguese texts, based on three levels of required linguistic competence to understand them. Nonetheless, when these views are contrasted to the ones of English texts, we must be careful not to draw false, straight conclusions based on contrasts involving any of the three 1,000-top-word levels to its counterpart based on the 2,000-top-word parameter, since the latter rather tell us something about the Portuguese lexicon than about the complexity of acquiring it. On the other hand, as either the 3,000 MFW of the English reference corpus and the 6,000 MFW of the Portuguese reference corpus account for approximately 86% of these corpora, if we take as contrastive references only the amounts of words of source and target texts that are among the remaining 14%, we can have a more reliable view of the transfer of lexical density from a ST to its TT.

On account of this logic, which has been assumed as a reliable one for the purposes of this case study, the lexical density levels here presented were based on lists of 1,000, 2,000 and 3,000 MFW of the British National Corpus (BNC) and 2,000, 4,000 and 6,000 ones of *Corpus Brasileiro* of Linguateca. And, these levels were set in AntWordProfiler 1.4.0w, in order to measure how much of each text is made of words whose frequencies of occurrences are lower than those among each language level attributed for the BNC and the *Corpus Brasileiro*, making it possible to quantitatively analyze the transfers of the attributed lexicon levels to the translated text. However, it must be clear that no claims to relate these three levels of required linguistic competence to age ranges can be made, once such correlation would have to follow levels of linguistic knowledge made available by a range of educational levels, which, in their turn, vary from country to country.

The figures in table 6 display the outcomes of this procedure.

Corpora	Word types	Level 1	Level 2	Level 3	Others[1]
The secret garden	4,808	76.62%	5.75%	3.88%	13.74%
	Nº of listed types	718	470	369	3,251
O jardim secreto	7,288	40.11%	20.93%	15.08%	23.87%
	Nº of listed words	651	589	509	5,539

[1] The percentage of words, whose level of complexity is higher than those of levels 1, 2 and 3

Table 6 – Lexical Density of ST and TT

What calls our attention in Table 6 is that, although the three levels of the Portuguese language contain twice as many words as the ones of English language, 5,539 words of the TT are among less frequent ones in the *Corpus Brasileiro*, while 3,251 words of the ST are among less frequent ones in the BNC, which suggests that lexical density is higher in the TT. This is reinforced by the percentages in the column "others", once 13.74% of the ST is made of such less frequent words, but 23.87% of the TT are made of them. Certainly, the superiority of the lexical diversity of the TT (see table 3) has its effect over this 23.87, yet such effect is not proportional, since if we divide the two percentages (23.87/13.74), we find a ratio of 1.73, while when we divide the number of word types (7,288/4,808), we find a ratio of 1.51. Therefore, the higher lexical density of the TT cannot be accounted just to its lexical diversity superiority.

Let us now analyze the readability of the ST and the TT.

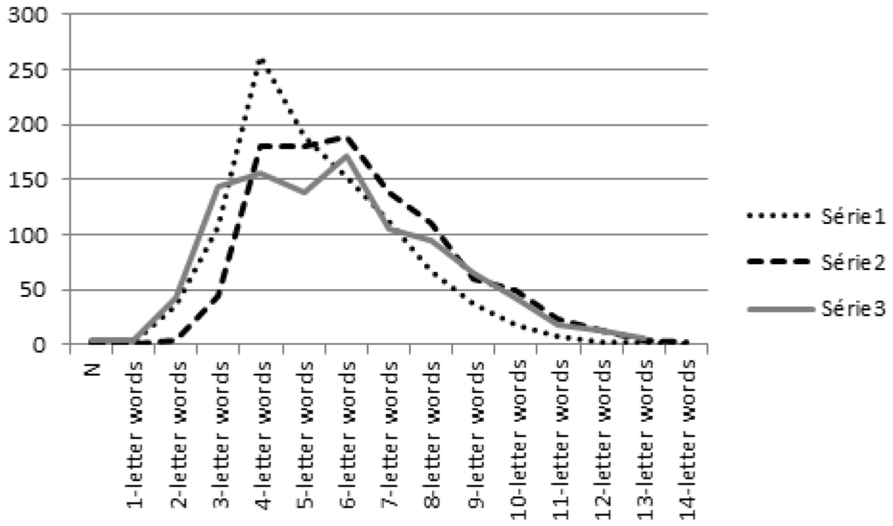
III - 4 Contrasting the readability of the ST and the TT

As the Gunning Fog Index is expressed by the formula 0.4 [(words / sentences) + 100 (complex words / words)], in which the first parenthesis calculates the "average sentences length" in words of a text, and the second one the "average of complex words" of it, in which all words with three or more syllables are complex ones, applying this formula to Portuguese texts would not be a problem, except for setting appropriate numbers of syllables to define Portuguese complex words, since three-syllable words are very common in this language. Besides, classifying words into complex or not complex and counting them, via WordSmith Tools, in which word length is measured by the number of letters not by the number of syllables, demanded a different criterion.

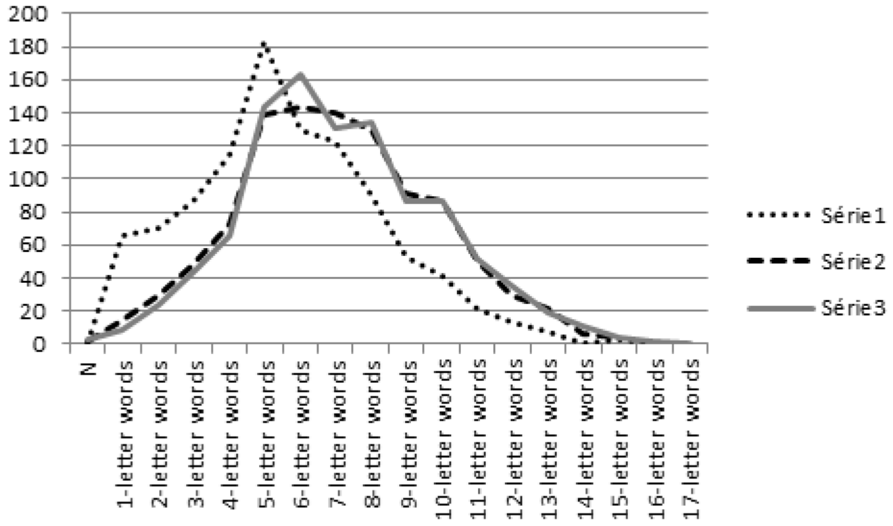
That said, we propose a mixed criterion that aligns Gunning's criterion and Nation's MFW family-words concept, on the basis of the 1,000, 2,000 and 3,000 MFW of BNC and of *Corpus Brasileiro*, as follows:

1. Each MFW list of the reference corpora was run in the WordList tool that has counted and grouped the numbers of words containing the same numbers of letters;

2. The obtained data of each list were inserted in Microsoft Excel and this software created the following graphs, in which series 1 indicates the 1,000, series 2 indicates 2,000 and series 3 indicates 3,000 MFW of each reference corpus.



Graph 1 – English words by numbers of letters.



Graph 2 – Portuguese words by numbers of letters.

3. By observing the graphs, one notices that all graph curves are very similar, which seems to support the assumption that word complexity can be measured by number of letters; otherwise, they would show some drastic distinctions, as, for example, if the highest point of one curve were positioned right over, say, the 12-letter word mark of the x-axis. Shorter words seem to be more common in English, since the three highest points of their curves refer to 260, 170 and 160 MFW within the interval from 4-letter to 6-letter words, while the highest points of curves related to Portuguese refer to 180, 160 and

> 140 MFW within the interval from 5-letter to 7-letter words. Therefore, it is likely that complex English words have fewer letters than those of Portuguese. As series 2 and 3 curves overlap in both graphs, and they are roughly parallel to series 1 curves, which refer to the 1,000 MFW, we can assume that the last ones better reflect the relationship between word complexity and number of letters of words in both languages. Looking closer to English series curve, one notices that the convex part of its circular shape is accentuated in a point that horizontally aligns with the mark of 20 of the Y axis and vertically aligns with the mark 9-letter words of the X axis. This change in the shape of the curve means that, at this point, it accentuates its fall toward the X axis, indicating fewer occurrences of words, and moves away from the Y axis, indicating words having more letters. This X-Y axis relation points to increasing degrees of word complexity. Thus, English words with nine or more letters can be assumed as the most complex ones. And, as previously mentioned, all the curves of the two graphs are very close in shape; the same point, horizontally aligned with the mark 20 in the Y axis, of curve of the Portuguese words for being vertically aligned with 11-letter words indicates that Portuguese words with 11 or more letters can be taken as the most complex ones. Nevertheless, being an experimental model for bilingually contrasting text readability it has its limitations and a few words of the two languages may be misclassified, as such limitations are reflected in both languages, mathematically, they will have little interference in the contrast results, since limitations in one language compensate limitations in the other.

Having defined the parameters for word complexity to be used in this case study, by running the ST and TT in the WordSmith wordlist tool, the average sentence length and the average of occurrences of complex words were obtained, so that it was possible to measure and contrast the "readability" of both the ST and TT.

Texts	Average Sentence	Average of Complex words	Readability Index
The secret garden	13.84	0.003089	6.78
O Jardin secreto	12.56	0.002091	5.86

Table 7 – Readability of ST and TT – a Gunning/Nation model

Considering that in the Gunning Fog Index each unit (1.0) refers to one level of the USA school educational system, based on Table 7, it seems *The Secret Garden* is more difficult to be read by English native speakers than is its translation by Portuguese native speakers, being both in homologous school levels. This conclusion may sound a little speedy since expected reading competencies of students enrolled in each school level may obviously vary from country to country, but it is not based on school levels themselves. The key to draw this conclusion is the well-known understanding that the higher are school levels, the higher are the expected reading skills of students. Also, this conclusion is supported by the reliability of the Gunning Fog Index, which as far as we have already studied is accepted by many scholars. Besides, it is also sustained by the reliability of Davies', Longmans' and Nations' studies concerning the 3,000 most frequent English words and their role in linguistic competencies. Moreover, the conclusion is drawn from reliable arithmetic, once all the variables of both languages have been treated in the same ways.

Let us now check the KeyWords of ST and TT.

III - 5 Crossing KeyWords List

In this step of the experimental model, WordSmtih KeyWords lists of the ST and TT were contrastively investigated, having primarily the lists of the previously described British and Brazilian Children's Literature corpora as reference ones, in order to identify likenesses between unusuality of occurrences of words in the ST and TT. Also, some other investigations were carried out with other reference corpora, namely, the corpora of Brazilian and English literature that were built for Kloeppel's research and the historical sub-corpus of the *Corpus do Português* and the BNC.

As already said the Key-Word lists, in their simplest reading, have just displayed keywords whose frequencies are very high or very low in the ST and TT in relation to their frequency in a reference corpus, often one bigger enough to be taken as such. However, deeper investigations based on keywords helped us to identify some textual features that might have active roles in the fabric of text complexity. For example, in *The Secret Garden* there are many words that belong to the semantic field of Botanic, such as name of flowers, plants and vegetables and so on, as one sees in this passage of the text:

> . "Martha," she said, "what are those white roots that look like onions?" "They're bulbs," answered Martha. "Lots o' spring flowers grow from 'em. Th' very little ones are snowdrops an' crocuses an' th' big ones are narcissuses an' jonquils and daffydowndillys. Th' biggest of all is lilies an' purple flags. Eh! they are nice. Dickon's got a whole lot of 'em planted in our bit o' garden."

Considering that 'daffydowndillys', according to Collins Dictionary, is an archaic or dialect form of 'daffodil', which is the common name for "any other plant of the genus *Narcissus*" (HARPERCOLLINS, 2017), one could assume that this word is a complex one and it would have acted as element of the complexity of the texture of the ST. This perception could be quite useful were it not for the fact that the three occurrences of the word in the text seem not be unusual enough for WordSmith Key-Word tool to read 'daffydowndillys' as a key-word; hence, we could not measure to what extent this word accounts for the complexity of the ST, unless we carried out an in-depth manual investigation on the whole text. Conversely, the nine occurrences of "snowdrops" in the text have been included in the Key-Word list in relation to the British Children's Literature, meaning that this word has an important role in the ST texture, and it may be a more complex word among the words of the ST, since its use seems to be less common in British Children's Literature. Moreover, investigating the role of this word in the ST complexity might be carried out electronically and be supported by statistical theories. Yet, it must be clear that simply assuming that KeyWords are the most complex ones of texts is inaccurate, if not silly, since very simple words, such as personal pronouns, may figure in KeyWords lists, as happened with the pronoun 'she' that is the first one displayed in the Key-Word list of the ST.

In order to investigate the trueness of these perceptions, after analyzing the KeyWords list of *O Jardim Secreto*, seven keywords have been selected for further analysis: three of them, *i.e.*, (i) *urbes*, *anêmonas* and *bulbos* due to their link with the semantic filed of Botanic and their relationships with the theme of the ST; (ii) one, *fuligem*, for being a word that we have assumed as being out of children's vocabulary; (iii) three others, *novamente*, *repentinamente*, *vagarosamente*, for being derived adverbs that are not often used in spoken Brazilian Portuguese. Moreover, these adverbs have been chosen because all of them are not among the 6,000 MFW of the *Corpus Brasileiro*. Besides, because three of their English correspondents of the ST, *i.e.*, 'snowdrops', 'soot' and 'bulbs', figured in its Key-Word List.

These seven words have been tested for log-likelihood test, which statistically measures the probability of a word of a text or of a corpus occuring in another corpus, often a bigger one, taken as a statistically representative sample of a linguistic system, a text genre, a language variety, and alike. The log-likelihood coefficients have been calculated in Microsoft Excel using the formula of the calculator of UCREL , in relation to the above mentioned reference corpora. The purpose of this procedure has been to check whether or not high unusuality of the seven words highlighted by WordSmith KeyWords tool would remain as such in relation to the other corpora. Basically, when the log-likelihood coefficient is equal or higher than a critical value , here set as 15.13, it indicates that the significance of the difference between the frequencies of occurrences of a word/type in two corpora is considerable, pointing to probabilities of the frequency of occurrence of such a word/type in one of the corpora being random.

KeyWords	**Frequency in the target text with 78,727 tokens**	**Log-Likelihood coefficients**		
		Brazilian Children's literature with 455,723 tokens	**Brazilian Literature with 820,829 tokens**	**Historical *Corpus do Português* with 45,000,000 tokens**
Urze(s)	14	53.63	61.04	88.06
Anêmona(s)	9	34.47	43.85	68.11
Bulbo(s)	10	31.92	48.72	81.01
Fuligem	20	63.84	89.58	172.49
Novamente	91	99.95	158.40	425.66
Repentinamente	23	70.46	112,05	128.02
Vagarosamente	12	26.96	**7.21**	44.10

Table 8 - Log-likelihood coefficients of some keywords of TT

Before analyzing the data of Table 8, let us see how ST English correspondents for these Portuguese keywords behave when tested for the Log-Likelihood coefficient.

KeyWords	**Frequency in the target text with 81,314 tokens**	**Log-Likelihood coefficients**		
		British Children's literature with 463,609 tokens	**English Literature with 1,358,814 tokens**	**BNC with 100,000,000 tokens**
Heather(s)	14	32.98	61.87	52.06
Snowdrop(s)	9	34.24	51.74	64.17
Bulb(s)	8	30.44	31.18	22.20
Soot	20	76.09	114.97	154.46
Again	119	**1.63**	**0.30**	79.44
Suddenly	44	**0.46**	**9.04**	70.49
Slowly	49	22.07	60.25	120.07

Table 9 - Log-likelihood coefficients of the English correspondents found in the ST

Knowing that the higher the log-likelihood coefficient, the more significant is the difference between two frequency scores (UCREL), when this coefficient is lower than a critical value, it indicates that the null hypothesis (H0) fails. In statistics, if H0 is confirmed it means there is no relationship between two measured phenomena. In Corpus Linguistics, this means that no linguistic and/ or literary feature has driven the occurrences of words in two corpora. That said, by observing Table 9, we notice that the highlighted adverbs 'again' and 'suddenly' are the only ones for which the log-likelihood coefficients are lower than the critical value 15.13. Therefore, it is assumed that there may be some linguistic and/or literary features that have pushed their occurrences in *The Secret Garden*. Overtly, as in relation to BNC their log-likelihood coefficients are higher than the critical value, it seems these words were overused in this text in relation to their frequency of occurrences in natural uses of English. In other words, the uses of 'again' and 'suddenly' seem to be common in British Literature.

Nevertheless, when we contrast the log-likelihood coefficients for the Portuguese correspondents for 'again' and 'suddenly', we perceive that the adverbs *novamente* and *repentinamente* seem to be overused in *O Jardim Secreto* in relation to all Portuguese reference corpora. This suggests that no linguistic and/ or literary feature has pushed their occurrences in the TT. Therefore, one cannot assume that these adverbs are not commonly used in Brazilian Literature, and their overuses in the TT have probably been pushed by the force of the ST and its language over the TT. And thus, Portuguese words that are more uncommon than their counterpart of the ST were used instead of the commoner nominal groups *de novo* ou *de repente* that are respectively correspondents of 'again' and 'suddenly' as well. That is to say, concerning these Portuguese adverbs and their English correspondents, the language of the target text is more complex than that of the source text. In order to check the validity of this outcome, the frequencies of occurrences of the words *novamente* and *repentinamente* were searched in the Corpus COMPARA (Frankenberg-Garcia and Santos, 2002), which is a bidirectional parallel corpus containing 2,978,688 words of ST and TT in English and in Portuguese. The searches have showed that while there are 502 occurrences of 'suddenly' in the corpus, there are only 20 of *repentinamente*, but there are 214 occurrences of *de repente*. And, there are 1,520 occurrences of 'again' against 133 of *novamente*, in contrast to the 477 occurrences of "de novo". Therefore, it is possible to trust that, in this respect, the language of the *O Jardim Secreto* is more complex than that of *The Secret Garden*. Note that, although it is inaccurate to link the log-likelihood coefficient with complexity of words, under the circumstances of the on-going investigations, especially concerning its interlinguistic dimension, this seems acceptable, since overuses of uncommon words tend to culminate in more complex texts. The reverse is also true. Let us draw a few lines concerning the other keywords that are under investigation.

When we observe the log-likelihood coefficients for the third adverb of Table 8, *vagarosamente*, in relation to the Brazilian Literature corpus, we perceive that the coefficient is lower than the critical value, which suggests that the use of this adverb may be common in literature, differently from its English correspondent of the ST, since the log-likelihood coefficients for 'slowly' are higher than the critical value in relation to all English reference corpora.

When we contrast the log-likelihood coefficients for *anemone, fuligem,* 'snowdrop' and 'soot' we notice that the coefficients increase according to the increase of the sizes of the reference corpora, keeping a certain similarity between the two languages, despite the fact that the increases in the sizes of the reference corpora of each language do not keep any similarity. It certainly points to a similar unusuality in both languages, which points to a balance between the complexities of these words in their linguistic systems.

Concerning the words 'heather' and 'bulb' nothing can be safely stated, considering the polysemy of the words 'bulb' and *bulbo* as well as taking into account that heather can refer to the flower and any women whose name is Heather as well, since these facts interfere in the calculation of the coefficients.

Some other brief comments about the KeyWords lists of the ST and TT are necessary. Word Smith has detected over uses of th', o' and em' in *The Secret Garden*, which refer to agglutinations of words that commonly happen in spoken English. Also overused was the word 'nowt', a word of a Northern England dialect (HARPERCOLLINS, 2017). These outcomes of WordSmtih point to efforts to bring marks of oral registers typical of colloquialism to the written text, which can be inferred as being a mark of Burnett's writing style. Otherwise, they would not be selected by the software.

Similarly, the Key-Word list of *O Jardim Secreto* showed overuses of the words *dizê, levá, pegá*, *pensá*, *perdê* and *trazê*, in which the acute and circumflex graphic accents highlight a certain over stressing on the last syllable of verbs in the infinitive form that happens when, in spoken Brazilian Portuguese, the last letter 'r', typical of this verb form, is suppressed. In other words, the translators have attempted to transfer Burnett's writing style to their translation.

To finish this article, let us now group the outcomes of each step of the case study.

IV Conclusion

Besides any possible conclusion from the case study, after applying the proposed experimental model, we are inclined to trust that the model can make room for future investigations concerning transfers of language level via translation processes. In effect, despite the already mentioned limitations, the outcomes of the proposed criteria to investigate levels of language complexity of both English and Portuguese texts seem to give grounds for our conclusion, as the following review of the case study seems to show.

In the first step of the case study, the analyses of TTR and STTR made by WordSmith Word list tool have showed that, when compared to the Children's Literature reference corpora, no discrepancies between the ST and TT could be observed. In the second step, after developing a criterion for adjusting the three 1,000-MFW-family lists proposed by Nation to the Portuguese language, AntWordProfiler 1.4.0w has showed that the lexical density of TT is higher than that of the ST. Therefore the language of the TT seems to be more complex. In the third step, based on the proposed model, which interchanges Gunning's model for measuring the readability of text and Nation's three MFW families, we got to know that the readability of the ST is suitable to higher level of reading skill when compared to the measured readability of the TT. Finally, the contrast between the log-likelihood coefficients of each of the seven words of two sets of English and Portuguese words selected from the Key-Word lists of the ST and TT evidenced that the Portuguese adverbs *novamente* e *repentinamente* were clearly overused in TT in relation to all Portuguese reference corpora. And, as these adverbs have been proved to be much less frequently used in Portuguese, they can be taken as complex ones. This outcome seems to align to that raised in the investigation concerning the lexical density of the ST and TT. Notice that *repentinamente* has 14 letters, a word of complex readability, according to the criterion adopted in the case study. Gathering these outcomes we can draw a conclusion: the target text seems to be made up of a somewhat more complex language than that of the source text, despite the lower readability of the latter.

Above all, by reading the interrelationships among the first three outcomes, we can infer that the proposed methodology can be a prototype to develop a more sophisticated methodology for further investigations that aim at contrasting the complexity of English and Portuguese source and target texts. Certainly such methodology would include contrastive investigations on collocations, colligations and semantic prosodies. We can argue that developing more in-depth studies like the one detailed in this article would generate outcomes that could help Brazilian teachers and teachers from other target cultures to choose translated English Children's Literature to be approached in their classes.

Notes

1. In this article Children Literature comprises children and juvenile books.
2. Normalize in the sense of the Universal of Normalization proposed by Baker (Baker, 1996).
3. In CL the concept keywords differ from the well-kwon notion, which often stands for the most important words, concerning the text content.
4. According to Berber Sardinha (2004) reference corpora have to be at least five times bigger than the corpus that is being under investigation.
5. KeyWords written with two capital letters is the name of the tool in WordSmith Tool software.
6. This year was set to open the period of time due to the Portuguese Language Orthographic Agreement that began being implemented in 2009, on the assumption that digital texts released or written after this year tend to follow the New Portuguese Orthography.
7. This research, based on a bi-directional parallel corpus, has showed that a great deal of morphosyntactic relations of the English language culminates in the emergence of lexical repetitions in English literary texts. And, when the last ones are compared to their translations, the frequency of lexical repetition tends to be approximately 4% superior than such frequency of Portuguese texts (KLOEPPEL, 2015, p. 59).
8. As WordSmith Tools processes the contracted verbal form together with their syntactic subjects, as in 'I'm', these verb forms were investigated as follows: am, I'm, is, it's, she's, he's, are, you're, they're, we're, was, were and will (be).
9. For further information the formula, see http://ucrel.lancs.ac.uk/llwizard.html
10. For further information about how this critical value have been set, see http://ucrel.lancs.ac.uk/llwizard.html.

References

Anthony, L. (2013). *AntWordProfiler 1.4.0w*. [Computer Software]Tokyo: Waseda University. Retrieved from http://www.antlab.sci.waseda.ac.jp/

_______. (2014). *AntConc (Version 3.4.3)*. [Computer Software] Tokyo: Waseda University. Retrieved from http://www.laurenceanthony.net/

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

Barthes, R. (1973). *O prazer do texto*. Translation from French by J. Guinsburg (1987), São Paulo: Perspectiva.

Bassnett, S. & Lefevere, A. (2001). *Constructing cultures: essays on literary translation.* Shanghai: Shanghai Foreign Language Education Press.

Berber Sardinha, T., Alambert, E. and Moreira Filho, J. L. (2013). *Corpus Brasileiro.* São Paulo: Linguateca. Retrieved from http://corpusbrasileiro.pucsp.br/cb/Acesso.html.

Berber Sardinha, T. (2004). *Linguística de corpus*. São Paulo: Manole.

BNC (2010) *The British National Corpus*, version 3 (BNC XML Edition). Oxford University Computing Services on behalf of the BNC Consortium, 2007. Retrieved from http://www.natcorp.ox.ac.uk/.

Carter, R., McCarthy, M., Mark, G. and O'Keeffe, A. (2016). *English grammar today: The Cambridge A-Z grammar of English.* Cambridge: Cambridge University Press.

Davies, M. and Michael, F. (2006). *Corpus do Português: 45 million words, 1300s-1900s*. Retrieved from http://www.corpusdoportugues.org.

Davies, M. (2005). *Vocabulary range and text coverage: insights from the forthcoming Routledge frequency dictionary of Spanish.* In *Selected Proceedings of the 7th Hispanic Linguistics Symposium* (2005), ed. D. Eddington, Somerville, MA: Cascadilla Proceedings Project, p 106-115.

Eco, Umberto. (1984). *Semiotics and philosophy of languages.* Blumington: Indianna University Press.

Frankenberg-Garcia, A. and Santos, D. *COMPARA, um corpus paralelo de português e de inglês na Web***.** Cadernos de Tradução IX. 1, 2002, pp. 61-79. Universidade de Santa Catarina. ISSN: 1676-7047. Retrieved from http://www.linguateca.pt/COMPARA/.

Greimas, A. J. (1966). *Semantique structurale – Recherche De Méthode*. Translated from French by Hakira Osakabe, (1966), São Paulo: Cultrix.

Gunning, R. (1968). *The technique of clear writing.* Michigan: McGraw-Hill.

Halliday, M. A. K and Hasan, R. (1976). *Cohesion in English.* London: Longman.

Halliday, M. A. K. and Matthiessen, C. (2004). *An Introduction to functional grammar*. 3.ed. Revised by Christian Matthiessen. New York: Oxford University Press Inc.

Harpercollins (2017). *Collins Online English Dictionary*. Retrieved from http://www.collinsdictionary.com/.

Hoey, M. (2005). *Lexical Priming: a new theory of words and language.* Oxon: Routledge.

Kloeppel P. R. (2015). *Repetições de palavras: estudo contrastivo em textos literários em português e inglês e suas traduções via linguística de corpus.* (Master's thesis, Federal University of Santa Catarina, Florianópolis, Brazil). Retrieved from http://www.pget.ufsc.br/curso/teses_e_dissertacoes.php.

Longman (2007). *Longman Communication 3000.* Pearson Longman, Harlow.

MCenery, A. M. e Xiao, R. Z. (2007). Parallel and comparable corpora: What are they up to? In *Incorporating corpora: translation and the linguist.* Clevedon, UK: Multilingual Matters, 2007, p. 18-31.

Nation, I. S. P. (2004). A study of the most frequent word families in the British National Corpus. In P. Bogaards & B. Laufer (Eds.), *Vocabulary in a Second Language: Selection, Acquisition, and Testing* (pp. 3-13). Amsterdam: John Benjamins.

NATION, I. S. P., & WEBB, S. (2011). *Researching and Analyzing Vocabulary.* Boston: Heinle Cengage Learning.

Randall, J. L. (2006). An analysis of lexical text coverage in contemporary German. In Corpus Linguistics around the world, ed. by A. Wilson, Dawn Archer & Paul Rayson, New York, NY, Editions Rodopi B. V., 2006, p. 115 – 119.

Scott, M. (2010). *WordSmith Tools 6*. Oxford: Lexical Analysis Software Ltd. & Oxford University Press.

Scott, M. (2015). *WordSmith Tools.* Stroud – UK: Lexical Analysis Software Ltd.

Sinclair, J. (2003). *Reading concordance: an introduction.* London: Pearson Education limited.

Sinclair, J. (2004). *Corpus and Text: Basic Principles, in Developing linguistic corpora: a guide to good practice*. Ed. M. Wynne. Oxford: Oxbow Books: 1-16, 2005.

Twain, M. (1874). *Adventures of Tom Sawyer* 1st Ed., Chicago: The American Publishing Company.

Recebido em: 28/03/2017
Aceito em: 30/05/2017