DOI: 10.14393/LL63-v32n1a2016-16

Assessing complexity and difficulty levels of machine-translated texts

Verificando níveis de complexidade e dificuldade de textos traduzidos automaticamente

Norma Fonseca[*]
Fabio Alves[**]

ABSTRACT: This paper addresses a proposal for assessing complexity and difficulty levels of machine-translated texts in Portuguese to be further post-edited without the support of the source text (monolingual post-editing) in an experimental setting. By using two objective standard parameters, namely readability indexes and word frequency, and by proposing post-editors' perception of difficulty to comprehend and to post-edit machine-translated texts as a new parameter, we sought to select texts with similar textual complexity or difficulty levels. This selection was necessary to carry out an experiment with four monolingual post-editing tasks in Portuguese involving machine-translated texts from three different source languages (English, Spanish, and Chinese). The application of readability indexes in conjunction with word frequency based on a corpus to analyze machine-translated texts into Portuguese to be used in experiments showed to be consistent and adequate. This method can also be applied to select texts to be used in Portuguese language classrooms and to select Portuguese texts to be included in Portuguese language textbooks. The findings can also be applied to the translation classroom, in which teachers can use the same methodology to select texts to be translated or post-edited or encourage students to analyze the texts themselves before performing a task, so students can become aware of the potential effort to be invested on a task or the real effort invested on the task

RESUMO: Este artigo aborda uma proposta para verificar o nível de complexidade e dificuldade de textos traduzidos automaticamente para o português a fim de serem pós-editados sem acesso ao texto-fonte (pós-edição monolíngue) em um estudo experimental. Com o uso de dois parâmetros padrão objetivos, quais sejam índices de legibilidade e frequência de palavras, e a percepção de pós-editores sobre a dificuldade para compreender e pós-editar os textos traduzidos pela máquina como um novo parâmetro, procurou-se selecionar textos traduzidos automaticamente que guardassem entre si níveis de complexidade e dificuldade textuais semelhantes. Essa seleção foi necessária para a realização de um experimento com quatro tarefas de pós-edição monolíngue em português envolvendo textos traduzidos automaticamente a partir de três línguas-fonte diferentes (inglês, espanhol e chinês). A aplicação de índices de legibilidade e da frequência das palavras com base em um corpus para analisar textos traduzidos automaticamente para o português para fins de uso em experimentos mostrou ser consistente e adequada. Esse método também pode ser aplicado para selecionar textos para serem trabalhados em aulas de português e para selecionar textos em português para serem incluídos em livros didáticos. Os resultados também podem ser aplicados para aulas de tradução. Professores podem usar a mesma metodologia para selecionar textos para serem traduzidos ou pós-editados em sala de aula e

[*]Norma Fonseca holds a PhD in Applied Linguistics (UFMG).
[**]Fabio Alves is Full Professor of Translation Studies at Universidade Federal de Minas Gerais (UFMG).

after performing it. Finally, post-editors' perception proved to be a sound parameter to validate text selection.

incentivar alunos a analisar os textos antes da execução das tarefas para que esses alunos possam se conscientizar do esforço a ser investido em uma tarefa ou do esforço real após executá-la. Além disso, a percepção dos pós-editores mostrou ser um parâmetro válido para a seleção dos textos.

KEYWORDS: Text complexity and difficulty. Readability indexes. Experimental texts. Machine translation. Monolingual post-editing.

PALAVRAS-CHAVE: Complexidade e dificuldade textual. Índices de legibilidade. Textos experimentais. Tradução automática. pós-edição monolíngue.

1. Introduction

Within Translation Process Research (henceforth TPR) researchers need to take into consideration many aspects of the research design before proceeding to collect data. One of them is to select suitable texts to be translated or post-edited by participants in experiments.

The importance of this research stage has been pointed out by several authors, including Saldanha and O'Brien (2013) and Krings (2001). Saldanha and O'Brien (2013, p. 116) state, "Having identified participants and data elicitation techniques, we might next turn our attention to selecting appropriate texts for the task." Krings (2001), in turn, highlights that a meaningful selection of experimental texts has to be made prior to an experiment.

At this stage, researchers cannot simply select a text by opening the first page of a magazine or a book or by choosing a text from a random website. In other words, selecting texts for experiments is not as easy as some people might think. Among other features, it is necessary to consider their length, textual genre and, above all, levels of difficulty or complexity.

TPR literature features experimental texts of different lengths. Läubli et al. (2013) had their participants post-edit three texts containing 50, 64 and 118 words, respectively, while O'Brien (2006) had participants post-edit a 1,777 word-long text. Nevertheless, in recent eye-tracking research within TPR, short texts have been proved to be of essence given some experimental constraints, namely: large size of eye-tracking data, i.e., data that has been collected with an eye tracker connected to a computer, large font size, double line spacing, and text to fit the screen without scrolling (SALDANHA; O'BRIEN, 2013).

Rodrigues (2002) includes textual genres as one of the variables in many TPR studies. According to the author, researchers usually had their participants translate predominantly newspaper or journal articles, or alternatively tourist, informative or scientific texts. He also highlights that few studies provided a grounded description of the text selection process and

justified the choice of a particular textual genre (e.g., KÖNIGS, 1986, 1987, 1989; ALVES, 1995; LAUKKANEN, 1996 as cited in RODRIGUES, 2002).

Rodrigues (2002) suggests that an appropriate theoretical framework is necessary for text selection so that findings can be comparable across TPR studies. This concern is apparent in a statement by Sun and Shreve (2014, p. 98), "the use of different translation strategies in terms of type and frequency might vary depending on the translation difficulty level of the texts."

Since comparability may be an issue in studies comparing performance in the translation of two or more texts, the predominance of studies involving only one text, as pointed out by Krings (2001), could be a potential explanation for the little attention heeded to text selection by that time.[1] However, a lack of grounded description of text selection holds true in present days, even in studies involving two or more texts (ALVES et al., 2016, MOORKENS et al., 2015). Consequently, after reading a TPR study, readers usually wonder how the researchers chose the text(s) to carry out their experiments and how they assessed the level of difficulty or complexity for translation.[2]

O'Brien (2013, p. 117) also notes this lack of adequate operationalization of text selection procedures and claims the use of readability indexes to measure text translatability is "still highly questionable". Other authors have similar criticisms, including Mishra, Bhattacharyya and Carl (2013), Campbell and Hale (2002), Sun and Shreve (2014). However, authors such as Jensen (2009) and Sun and Shreve (2014), for instance, have shown that the use of such indexes in conjunction with other measures can be useful for assessing translation difficulty or complexity and thus for selecting texts for TPR.

The present study aims to provide a grounded description of text selection for a research design approaching monolingual post-editing. Post-editing can be generally defined as the human process of correcting a raw output provided by a machine translation system in order to meet some quality standards, which could be either for the sake of understanding (light post-editing) or for publication purposes (full post-editing). When the source text is available during

[1] Krings (2001, p. 74) mentions that only two studies used more than two texts. However, in a table provided on page 75, the author lists three studies that used more than two texts.

[2] We agree with Jensen's (2009) observation that there is a clear distinction between text complexity, i.e., an objective notion, and text difficulty, i.e., a subjective notion. However, some authors do not make such a distinction, and it seems that some of them refer to translation difficulty and complexity as the same concept. Throughout this paper, we present the way they refer to these concepts in their publications, rather than the way we believe they understand those concepts. Moreover, we refer to "difficulty or complexity" as a general term to encompass both concepts.

the post-editing task, we call it bilingual post-editing. In monolingual post-editing, the post-editor performs the task without the aid of the source text, and does not necessarily have any linguistic knowledge of the source language.

This paper consists of five sections, including this Introduction. Section 2 introduces some concepts, models and theories that have been used to assess text complexity and select experiment-suitable texts. Section 3 provides details of the materials and methods applied to select four machine-translated texts, which are reproduced in the Appendix. Section 4 presents the results and statistics obtained after applying the method proposed to select texts. Section 5 draws some conclusions from the study and points out some contributions to future research.

2. Theoretical framework

Aiming at studies that deal with text difficulty, Campbell (1999) proposed a methodology called Choice Network Analysis (CNA). This method is applied to translation by Campbell and Hale (1999, 2002) with the purpose of investigating translators' mental processes. This methodology entailed looking at what has been produced as target texts and comparing "the renditions of a single string of translation by multiple translators in order to propose a network of choices that theoretically represents the cognitive model available to any translator for translating that string" (CAMPBELL, 2000, p. 215).

Campbell and Hale (1999) used CNA to probe into source text difficulty in translations from English. They distinguished comprehension difficulty from production difficulty and investigated the possibility of comparing production difficulty in different target languages (Arabic, Spanish, and Vietnamese) as a means to find what they refer to as universal translation difficulties. To that end, Campbell and Hale (1999, para. 11) analyzed "the translations of the chunk *The patient will be closely monitored*" and drew up "the most parsimonious networks that [would] account for all the data pertaining to grammatical choices".

Aiming at assessing students' and examinees' learning development or proficiency in different languages, Campbell and Hale (1999) showed how to create or select texts of a given level of difficulty by weighting from zero (prediction of no difficulty) to *n* all items in a translated text. Items that were harder to translate would result in higher weights, with the weights determined according to item types, such as problematic words vs. grammatical structures. Then all weights would be added up to eventually measure the total text difficulty for any participant.

Campbell and Hale (2002) expanded the use of CNA to assess source text difficulty for selecting texts for translator training and examination. By carrying out an empirical study to assess difficulty of English source texts to be translated into Arabic and Spanish, they tried to determine which kinds of lexical items and syntactic structures could lead to the highest difficulties in the translation process based on the alternative renditions produced by the participants. The authors also assessed the potential correlation between the number of choices available to the translator and the level of translation accuracy, with a view to finding implications for our understanding of the notion of translation difficulty. Their findings showed that quantitatively all item types entailed similar difficulty levels to the participants, but the causes of difficulty were qualitatively different across both item types and participants.

Jensen (2009) took a different approach and provided a detailed description of the methodology applied to assess translation complexity of potential experimental texts. Facing the challenge of selecting three English-language source texts with different levels of complexity, the author adopted three objective criteria at a preliminary stage: readability indexes, word frequency, and non-literalness.

Readability indexes have been used to measure the level of textual difficulty for comprehension purposes since the 1920s. Jensen (2009) applied seven of them (ARI - Automated Readability Index, Flesch-Kincaid Index, Coleman-Liau Index, Gunning Fog Index, SMOG Index, Flesch Reading Ease Score, and LIX) because they can be used to "(...) assess the relative amount of both production effort and comprehension effort needed during a translation process" (Jensen, 2009, p. 61-2). The indexes that were used to assess text complexity and their formulas are reproduced in Table 1:

Table 1. Reading index formulas.

Index	Formula
ARI	4.71*characters/words+0.5*words/sentences-21.43
Flesch-Kincaid	11.8*syllables/words+0.39*words/sentences-15.59
Coleman-Liau	5.89*characters/words-0.3*sentences/(100*words)-15.8
Gunning Fog	0.4*(words/sentences+100*((words >= 3 syllables)/words))
SMOG	square root of (((words >= 3 syllables)/sentence)*30) + 3
Flesch Reading Ease	206.835-84.6*syllables/words-1.015*words/sentences
LIX	words/sentences+100*(words >= 6 characters)/words

Source: Jensen (2009, p. 64).

The first five indexes return a value that refers to years of schooling that a reader should have to fully understand what is written. In turn, the Flesch Reading Ease index returns a score from 0 to 100, ranging from 0-30 for very difficult texts, which are best understood by college graduates, to 90-100 for very easy texts, which can be easily understood by an average 11-year old student. In contrast, the LIX formula, which is the only one used for comprehension of texts in a language other than English (namely, Swedish), classifies scores of 0-24 as indicative of very easy texts and scores of 55 or above as indicative of very difficult texts.

After applying these different readability indexes as a proxy to assess text translation complexity, Jensen (2009) found a progression of levels of textual complexity in three experimental texts (A, B and C). While the author assumed textual complexity to be similar for both readers and translators, he also acknowledged the limitations of using readability indexes: they do not provide conclusive evidence of how a translator would perceive the effort and how much effort would be really needed to translate the texts (thus, producing new texts), since these indexes were created to measure text comprehension, rather than text production. Mishra, Bhattacharyya and Carl (2013, p. 346), for instance, question not only the application of readability indexes to translation, but also other approaches that focus on reading because "translation is not merely a reading activity".

The absence of conclusive evidence for the use of readability indexes may have led Jensen (2009) to determine a second objective indicator of text complexity, i.e., word frequency. Based on the assumption that "the less frequently a word appears in a corpus, the more effort is generally needed to translate it", Jensen (2009, p. 62) applied two frequency bands (K1 and K2-K10) using the *British National Corpus* (BNC) to classify the words in the

three texts used in his experiment. As suggested by Cobb (2008)[3] and Heatley and Nation (1994)[4], as cited in Jensen (2009, p. 162), K1 frequency band encompasses the first one thousand most common words in English (1-1000), and K2-K10 frequency band refers to less common words (1001-10000).

Jensen (2009) supported his previous findings with this second indicator of text complexity. However, he also provided several examples which prove that this indicator has its own limitations, such as the translators' different levels of familiarity with a word, the lack of correlation between the occurrence of a word in a K2 band and the translators' difficulty to understand or translate it, and the lack of correlation between the occurrence of a word in a K1 band and the participants' easiness to translate it.

Jensen (2009) also proposed a third objective indicator of text complexity, i.e., non-literalness, which encompasses the density of idioms, metonyms and metaphors within texts A, B and C. Assuming a relationship between non-literalness and processing effort, the author suggested that idioms, as well as metonyms and metaphors, can increase the amount of processing effort to comprehend and to translate because they are non-literal, i.e., they use "figurative language". The results for idioms confirmed the previous findings of a progression in text complexity. However, Jensen (2009) warned that idiom density, i.e., "[n]umber of idiom units per 1,000 running words" (GOTTLIEB, 1997, p. 331), has also to be considered with caution, because it provides rough measures of complexity.

Sun and Shreve (2014) stated that readability formulas, as used by Jensen (2009), to assess translation complexity need to be empirically challenged. Thus, they conducted an empirical study to develop a method to measure a text's translation difficulty level. They had undergraduate and graduate students assess the level of translation difficulty of fifteen text excerpts, with an average length of 125 words, before and after the translation task itself. The authors intended to find a way to assess (1) a text's level of translation difficulty without having it translated in the first place and (2) the sources of translation difficulty on the basis of translation errors.

In their method, Sun and Shreve (2014) also investigated the reliability of NASA-TLX (Task Load Index) to measure translation difficulty, the possibility of using translation quality

[3] COBB, T. Web Vocabprofile [accessed 19 March 2009 from http://www.lextutor.ca/vp/], an adaptation of Heatley & Nation's (1994) *Range*. 2008.
[4] HEATLEY, A.; NATION, I. S. P. Range: A program for the analysis of vocabulary in texts [software]. Retrieved from http://www.victoria.ac.nz/lals/staff/paul-nation/nation.aspx. 1994.

scores (i.e., accuracy) and the time a task takes to measure or represent translation difficulty, and the possibility of using Flesch Reading Ease formula (or readability formulas in general) to predict a text's level of translation difficulty. NASA-TLX is commonly used to assess workload:

> NASA-TLX is a multi-dimensional rating procedure that derives an overall workload score based on a weighted average of ratings on six subscales. These subscales include Mental Demands, Physical Demands, Temporal Demands, Own Performance, Effort and Frustration. It can be used to assess workload in various human-machine environments such as aircraft cockpits, command, control, and communication (C3) workstations; supervisory and process control environments; simulations and laboratory tests. (retrievable from: http://humansystems.arc.nasa.gov/groups/tlx/. Access on: 25 Jul., 2014).

Sun and Shreve (2014) indicated that NASA-TLX is a reliable indicator to assess translation difficulty. They also found a weak correlation between the time a task takes to be translated and the level of translation difficulty. Another correlation was found between the Flesch Reading Ease formula and translation difficulty as measured by NASA-TLX.

Mishra, Bhattacharyya and Carl (2013) provided an explanation for the unreliability of time to perform a task as an indicator of text difficulty. These authors argued that the time taken to translate may not be strongly related to translation difficulty due to the fact that "it is difficult to know what fraction of the total translation time is actually spent on the translation-related thinking" and "the translation time is sensitive to distractions from the environment" (MISHRA; BHATTACHARYYA; CARL, 2013, p. 347).

Alves, Pagano and da Silva (2010) also provided information about the selection of experimental texts and a new approach to profiling texts for TPR. The authors used Rhetorical Structure Theory (RST), as developed by Mann and Thompson (1987) to profile two source texts (one in English to be translated into Portuguese and one in Portuguese to be translated into English). As defined by Mann and Thompson (1987, p. 1), RST is "a descriptive theory of a major aspect of the organization of natural text. It is a linguistically useful method for describing natural texts, characterizing their structure primarily in terms of relations that hold between parts of the texts." Alves, Pagano and da Silva (2010) selected two source texts based on their similar rhetorical complexity using RST, number of words and domain (medicine: sickle cell disease). When O'Brien (2013) posits there is a need for new, more reliable measures to profile texts, she cites Alves, Pagano and da Silva's (2010) study as an example of a new proposal.

However, as she states, their proposal "has not yet seen much uptake in cognitive translatology research, but has scope for further investigation and testing" (O'BRIEN, 2013, p. 7).

Mishra, Bhattacharyya and Carl (2013) also introduced a new approach to predicting sentence translation difficulty, which can be used to select texts for experimental studies. They created the Translation Difficulty Index (TDI) based on cognitive evidence from eye-tracking data. In their method, the sum of fixation and saccade counts are used to quantify the translation difficulty of a sentence considering lexical and structural properties. TDI, according to the authors, was then correlated with length, polysemy level, and structural complexity. After calculating TDIs for sentences from the Translation Process Research Database (TPR 1.0) − a publicly available database of recorded translation sessions (CARL; SCHÄFFER; BANGALORE, 2016) −, Mishra, Bhattacharyya and Carl (2013) trained a Support Vector Regression (SVR) system to predict TDIs for new sentences. Those sentences were classified as "easy", "medium" and "hard" to translate. By applying TDI, the authors claim, a researcher can select texts based on their translation difficulty and even compare the translation of texts with different categories of translation difficulty.

As shown above, Jensen (2009), Mishra, Bhattacharyya and Carl (2013), Campbell and Hale (2002), Sun and Shreve (2014) and Alves, Pagano and da Silva (2010) describe methodologies related to text difficulty or complexity for translation. Krings (2001, p. 186) can be added to this list because one of his concerns was to choose a text type that would have a simple structure, "which would be expected to yield fairly usable machine translation results". Consequently, this selected experimental text type, i.e., instructive technical text would not pose difficulty for the participants to post-edit it (both with and without the aid of the source text) or to translate it. Krings (2001) also considered aspects such as language pairs involved, number of texts, text topic and text length when selecting the experimental texts.

The following section describes the reflections and procedures taken into account to assess difficulty of text machine-translated into Portuguese to be further post-edited in an experiment setting using eye-tracking, keylogging, think-aloud protocols (TAPs) and retrospective protocols. Results of the experiment itself, which included four different monolingual post-editing tasks, are available in Fonseca (2016).

3. Methodology

3.1. Research questions and hypotheses

Jensen (2009) poses some questions to be examined in further investigations. The author wonders, for example, whether the application of readability indexes can reveal the level of difficulty a translator would experience when translating a text, to what extent the readability indexes predict the easiness to comprehend and/or to translate texts, and how the frequency of a word can be compared to the translator's familiarity with it.

For the present purposes, one may also wonder how researchers could best use the objective criteria Jensen (2009) proposes to select texts in order to carry out monolingual post-editing experiments involving different source languages. This can be translated into the following questions:

1) Can we use readability indexes to analyze machine-translated texts into Portuguese using Google Translate?

2) How can we compare word frequency to word familiarity as proposed by Jensen (2009)?

3) How can we assess text difficulty and complexity without assessing the texts ourselves?

4) How can we assess the relationship between the easiness to comprehend and the easiness to post-edit texts?

The following hypotheses emerged from those questions:

1) Machine-translated texts into Portuguese using Google Translate can be analyzed building on readability indexes usually applied to English texts, since they take into account objective measures, such as average sentence length, average number of syllables and complex words (i.e., words with three or more syllables), which can be used to infer text complexity across different languages.

2) A large Brazilian Portuguese language corpus can be used as a parameter to word frequency.

3) We can assess text difficulty and complexity by applying readability indexes, word frequency and post-editors' perception to analyze texts.

4) Post-editors' perception can be used to assess both text difficulty and the relationship between easiness to comprehend and easiness to post-edit.

3.2. Participants

Nineteen participants were asked to assess the dificulty to comprehend and post-edit machine-translated texts into Portuguese without the aid of the source texts. They were undergraduate students at a university in Brazil, where they intended to receive their BA in languages with a focus on translation. All students were native speakers of Brazilian Portuguese and had been studying at the university for a period ranging from two to five semesters: 13 for two semesters, three for three semesters, two for four semesters, and one for five semesters. Their education level ranged from 13 to 38 years of schooling (*M*=19.26 years, *SD*=6.34 years, *Mdn*=18 years).

3.3. Materials and Methods

Following a recent trend within TPR, the research design first involved selecting short texts of approximately 100 words. Then, four short texts in three different source languages (English, Spanish, and Chinese) were selected. The texts were supposed to be comparable not only in length, but also in terms of textual genre and domain knowledge. The choice was for popular science news texts on sustainability because such a textual genre usually has a standard structure and is considered a suitable candidate for post-editing. Moreover, sustainability is a current issue that concerns everyone. However, one question remained: How should text difficulty or complexity level be best assessed so that the four experimental texts could be comparable regarding the potential effort invested to perform the tasks? This will be approached in Section 4.

The English, Chinese and Spanish texts were then machine-translated into Portuguese on April, 7th, 2014 using Google Translate (available at https://translate.google.com/), a free machine translation system, before proceeding with the analysis described in this paper. As the texts had been selected for monolingual post-editing tasks, it was more reasonable to analyze the machine-translated texts into Portuguese than the source texts.

The first idea of approach to profiling the experimental texts was using the methodology proposed by Alves, Pagano and da Silva (2010), i.e., RST, which was also suggested by O'Brien (2013) for further investigation and testing. However, RST seemed to be inadequate for the present purposes because the presence of the same relations within the texts does not necessarily seem to guarantee that the experimental texts would have the same level of text difficulty or complexity for translation. Moreover, some relations in the source text can be changed when the raw output is generated by the machine translation system, making it more difficult to

annotate these relations in machine-translated texts. For instance, the first sentence of the Spanish source text (T3), "La continua evolución del mercado tecnológico provoca la obsolescencia de miles de teléfonos móviles" was machine-translated into Portuguese as "A evolução contínua da obsolescência mercado de tecnologia faz com que milhares de telefones celulares".

3.4. Readability indexes and word frequency

Readability indexes and word frequency, as used by Jensen (2009), seemed to suit the purpose of selecting machine-translated texts for monolingual post-editing experiments. Although the use of readability indexes has been criticized when applied to measuring translation difficulty level, they remained as indicators applied in conjunction with others, as detailed below, to select our experimental texts. Following Jensen (2009) and Sun and Shreve (2014), the underlying assumption was that such indexes would help to predict the level of difficulty participants would have to post-edit the texts.

As for word frequency, one can wonder whether it is an actual indicator itself or a component of the readability indexes, as most of them include word frequency in their formulas: "The text properties usually are average sentence length, normally based on samples of 100 words, and an estimate of word difficulty, typically based on syllable length or occurrence on a list of high-frequency words" (DAVISON; GREEN, 1988, p. 2). In addition, one can wonder whether such indicators are applicable across languages as readability formulas are based on a list of high frequency words in English.

Bearing this in mind, it was hypothesized that the readability indexes used by Jensen (2009) could be applied to Brazilian Portuguese machine-translated texts. This was because their formulas are based on quantitative graphological aspects such as number of sentences, number of syllables and complex words, which seem to be language independent.

Moreover, a readability index that is already adapted to Portuguese could be used to confirm findings. Among the few references to using readability indexes to test difficulty of Brazilian Portuguese texts is Martins et al. (1996), who developed a software tool to apply Readability Flesch Score to assess excerpts from Brazilian Portuguese textbooks of different educational levels. The authors adapted the Flesch formula by adding 42 points to all Flesch scores of the Portuguese-language texts: after comparing the Flesch score of English-language source texts (Introductory Physics textbooks) and their translations into Portuguese, they found

that the translated texts scored on average 42 points lower than the source texts. This difference, as they explained, accounts for the fact that words usually have more syllables in Portuguese than in English (on average, 1.8 syllables per English word vs. 2.2 syllables per Portuguese word).

Although Martins et al. (1996) also employed other readability indexes (i.e., ARI, Kincaid and Coleman-Liau) in subsidiary tests, obtaining the same results as those in the Readability Flesch Score, they concentrated on Flesch because it follows a scale from 0 to 100, instead of years of schooling as the other indexes. The adapted Flesch formula provides four levels of readability: first four years of schooling (very easy texts – scores from 75 to 100), fifth to eighth grade (easy texts – scores from 50 to 75), high school and college (fairly difficult texts – scores from 25 to 50), and academic texts (very difficult texts – scores below 25). Moreover, several tests with different text types and educational levels showed that the adapted Flesch formula had a good performance, demonstrating that the readability scores are meaningful for Brazilian Portuguese texts. Eventually, as Martins et al. (1996) pointed out, the adapted Flesch formula might contribute to screening Portuguese texts to a target audience.

A reference to the application of a computer tool based on readability indexes to analyze text characteristics in Brazilian Portuguese is Scarton, Almeida and Aluísio (2009). This study uses Coh-Metrix-Port, an adaptation of Coh-Metrix (developed at University of Memphis, USA) produced by researchers of USP/UFSCar/UNESP, Brazil: Coh-Metrix-Port (available at http://141.225.42.101/CohMetrixHome/documentation_indices.html) "is a computational tool that produces indexes of the linguistic and discourse representations of a text." This computer tool uses built-in readability indexes to analyze the texts, i.e., Flesch Reading Ease, Flesch-Kincaid, and Coh-Metrix L2.

Scarton, Almeida and Aluísio (2009) described the creation of Coh-Metrix-Port and its application to analyzing original texts and their simplified versions for children. This tool (available at http://www.nilc.icmc.usp.br:3000) "can help assess whether texts in Portuguese, for instance those available on government websites, are suitable for functional illiterates and people with other cognitive impairments, such as aphasia and dyslexia, and also for children and adults in a literacy phase…"[5] (ALMEIDA; ALUÍSIO, 2009, p. 1).

[5] Our translation from: "O Coh-Metrix-Port pode ajudar a avaliar se textos em português, por exemplo os que estão disponíveis em sites do governo, são adequados para analfabetos funcionais e pessoas com outras deficiências cognitivas, como afasia e dislexia, e também para crianças e adultos em fase de letramento..."

In addition to indexes typical of Coh-Metrix, Coh-Metrix-Port also includes Flesch Reading Ease, adapted to Portuguese by Martins et al. (1996), and Flesch-Kincaid, from which it incorporates years of schooling. Thus, Coh-Metrix-Port does not only analyze the text characteristics, but it also predicts text difficulty.

As for word frequency, references of the Brazilian Portuguese language in use can be found, among others, in two large corpora, namely *Corpus Brasileiro* (available at http://corpusbrasileiro.pucsp.br/cb/Inicial.html) and *Corpus do Português* (available at http://www.corpusdoportugues.org). The first one, created by Tony Berber Sardinha (PUC-SP, Brazil), encompasses more than 1 billion words. The second one, developed by Mark Davies (Brigham Young University) and Michael Ferreira (Georgetown University), currently contains 45 million words from texts of different genres in Brazilian Portuguese and European Portuguese from the 12th through the 20th century.

Furthermore, based on 20 million words from contemporary Portuguese (20th century) − 10 million words from Brazilian Portuguese texts, and 10 million from European Portuguese texts−, Davies and Preto-Bay (2008) compiled the *Frequency Dictionary of Portuguese*, which contains the top 5,000 most frequently used Portuguese words. The underlying assumption for this number of frequent words is: "the 4,000-5,000 most frequent words account for up to 95 percent of a written text" (Nation, 1990, as cited by Davies and Preto-Bay's, 2008, p. vii). Moreover, authors have agreed that for adequate reading comprehension, the vocabulary coverage rate, i.e., the percentage of the vocabulary a reader knows, should be 95% (LIU NA; NATION, 1985; LAUFER, 1989).

3.5. Predicting text difficulty

To complement readability indexes and word frequency as indicators of text complexity as proposed by Jensen (2009), this study proposes the use of a subjective measure, namely, translation and/or language students, translators' or post-editors' perception of the difficulty level of the texts to be used in translation and post-editing experiments. These participants should have a profile at least similar to those participants who will take part in the experiments, such as language or translation students and professional translators and post-editors. In this case, nineteen translation language undergraduate students with focus on translation from a public university were asked to evaluate the difficulty to comprehend and to post-edit four texts which had been previously selected based on readability indexes and word frequency. Their

assessment was carried out as homework assigned by one of their professors within an undergraduate translation course. The same student's (post-editor's) perceptions would be assessed at two different times: first after reading each of the four experimental texts, then in sequence after performing the monolingual post-editing of these texts.

4. Results

This study uses readability indexes and word frequency—as used in Jensen (2009)— as well as post-editors' perception of text difficulty in order to select experimental texts. Results of each of these indicators are presented in the subsections below. These results are followed by a statistical analysis.

4.1. Applying English readability indexes to machine-translated texts into Portuguese

Seven readability indexes were first applied as suggested by Jensen (2009), followed by Flesch Reading Ease-Coh-Metrix-Port as adapted to Portuguese by Martins et al. (1996), to select four potential machine-translated texts. However, one formula, SMOG, was changed: while Jensen (2009) uses the formula "square root of (((words >= 3 syllables)/sentence)*30) + 3", this study sticks to the original formula created by McLaughlin (1969): "1.0430*square root of (((words >= 3 syllables)/sentence)*30) + 3.1291", not rounding up the last number, "3.1291", for more accuracy.

Figure 1 shows the results of five readability indexes that return the years of schooling necessary to comprehend texts, according to the US system. The experimental texts were identified through numbers: Texts 1 and 2 were machine-translated from English into Portuguese, Text 3 was machine-translated from Spanish into Portuguese, and Text 4 was machine-translated from Chinese into Portuguese. Two texts were machine-translated from English into Portuguese to investigate the effect of think-aloud protocol (TAP). The texts are all popular science pieces of news extracted from websites that address sustainability.

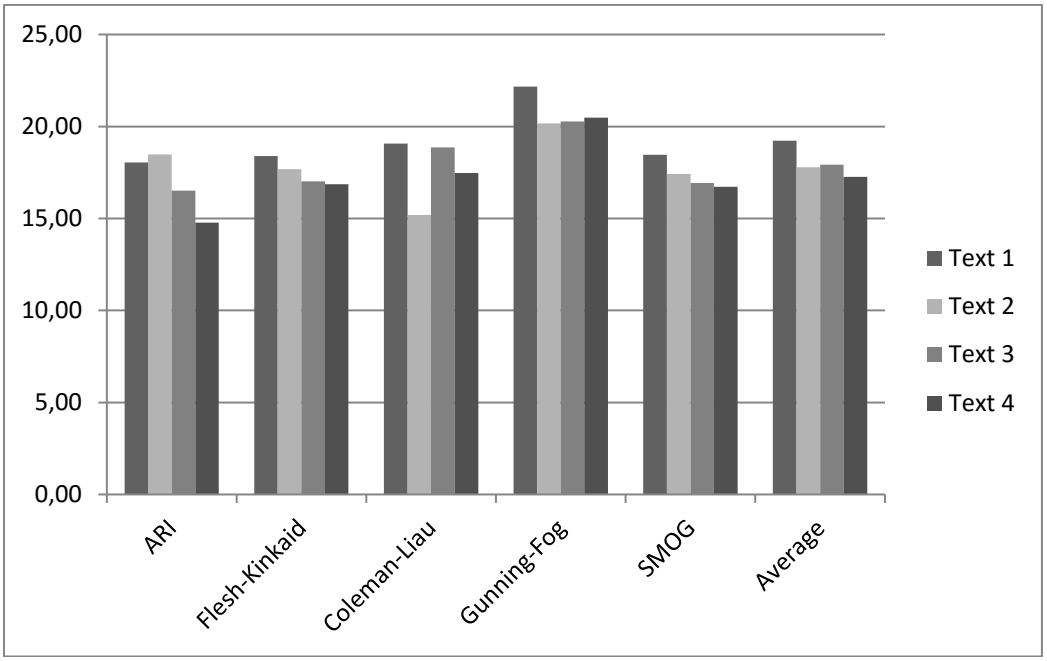

Figure 1. Scores of five readability indexes for four experimental texts (in years of schooling).

The results show that it is necessary to have at least 14 years of schooling to comprehend any of the four experimental texts. The maximum number of years of schooling to comprehend a text is that of Text 1 according to the Gunning-Fog index (22 years). Considering all five indexes altogether, ca. 19.23, 17.79, 17.92 and 17.26 years on average are necessary to comprehend Texts 1, 2, 3 and 4 respectively. The greatest variations in these results occur in Coleman-Liau score for Text 2 and ARI score for Text 4. These results indicate that less years of schooling are necessary for a reader to understand Texts 2 and 4 respectively in comparison to the other texts. The standard deviations for both Coleman-Liau and ARI are more than one year of schooling when comparing to the other readability indexes (Flesch-Kincaid: *M*=17.49, *SD*=0.70; Coleman-Liau: *M*=17.65, *SD*=1.79; SMOG: *M*=17.38, *SD*=0.77; Gunning-Fog: *M*=20.78, *SD*=0.94; ARI: *M*=16.95, *SD*=1.68). As participants had been studying for 19.26 years on average, the selected texts seemed to be adequate for their presumed level of text comprehension.

The other two readability indexes used by Jensen (2009), Flesch Reading Ease and LIX, return numeric values. Flesch Reading Ease returns a value from 0 to 100 using the following rationale: the lower the value, the more difficult to comprehend the text. However, some texts returned unexpected values, particularly Text 4, machine-translated from Chinese, which returned a negative value, i.e., below 0 (zero). These numbers can indicate that these texts would be so difficult to understand that even a person who already holds a BA would struggle to

comprehend them. The texts that presented lower values such as the negative ones were excluded early from the analysis. In contrast, LIX returns values that are to be understood the other way round: the lower the score, the easier to comprehend the text. Figures 2-4 display the scores of LIX, Flesch Reading Ease and Flesch Reading Ease from Coh-Metrix-Port, respectively.

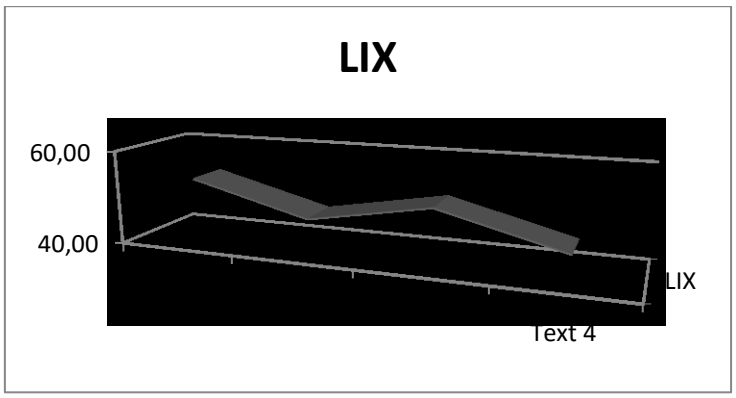

Figure 2. LIX scores of the experimental texts.

LIX scores range from 45 to lesser than 55. According to LIX parameters, Text 2 is at the top border of Medium category (40-45), while the other texts (T1, T3, and T4) fall into the Difficult category (50-55).

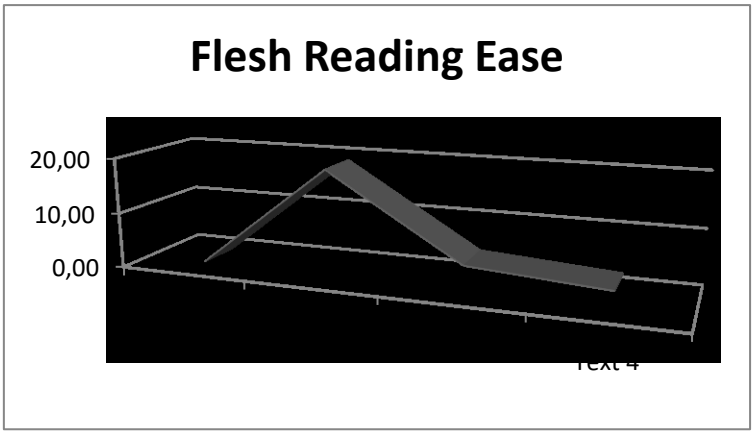

Figure 3. Flesch Reading Ease scores of the experimental texts.

Flesch Reading Ease classified all four texts into the Very difficult category, which ranges from 0 to 20. Text 2 is the only one that presents a score close to a new category, Difficult, with a score of 19.10. These results are consistent with those of the LIX formula.

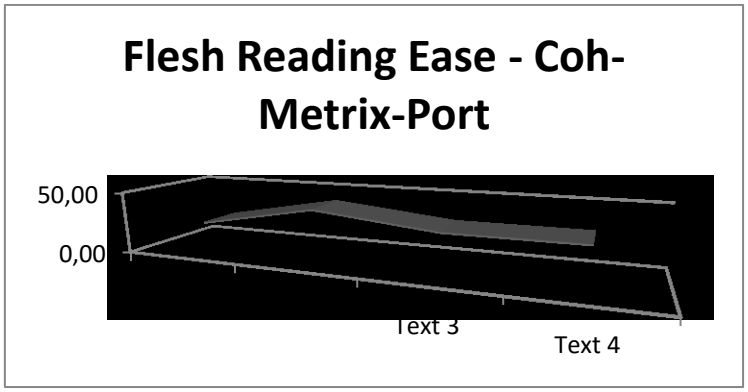

Figure 4. Flesch Reading Ease adapted to Coh-Metrix-Port scores of the experimental texts.

The same pattern found using English Flesch Reading Ease is found in the Portuguese adapted version of Flesch Reading, that is, Text 2 presents the highest score, which means it features the lowest level of text difficulty.

Having analyzed the experimental texts by using a pool of eight readability indexes as a first indicator of text complexity, Section 4.2 shows the results concerning the second indicator of text complexity, word frequency.

4.2. Word frequency

As Nation and Waring (1997, p. 8) state, "[a]lthough the language makes use of a large number of words, not all of these words are equally useful. One measure of usefulness is word frequency, that is, how often the word occurs in normal use of the language".

Based on the *Brown Corpus*, developed by W. N. Francis and H. Kucera (Brown University), the relation between vocabulary size and text coverage can be summarized as in Table 2. Numbers in the second column show the proportion of a text that is covered by bands of high word frequency presented in the first column.

Table 2. Vocabulary size and text coverage in the *Brown corpus*.

Vocabulary size	Text coverage
1,000	72.0%
2,000	79.7%
3,000	84.0%
4,000	86.8%
5,000	88.7%
6,000	89.9%
15,851	97.8%

Source: Francis and Kucera (1982) and Kucera (1982), as cited in Nation and Waring (1997, p. 9).

Another relation of text coverage percentage and vocabulary size is reported in a study carried out by Hirsh and Nation (1992), who analyzed three novels for teenagers. These authors indicate different numbers of coverage rate as shown in Table 3.

Table 3. Percentage text coverage of various vocabulary sizes in three short novels.

	2,000	**2,000 + proper nouns**	**2,600**	**5,000**	**7,000**
The Pearl	89.7	92.5	95.0	97.0	98.0
Alice	91.9	95.0	97.0	98.0	99.0
The Haunting	90.2	94.9	97.0	98.0	99.0
Mean	90.6	94.1	96.3	97.7	98.7

Source: Hirsh and Nation (1992, p. 691). Adapted.

The numbers of text coverage rate in Tables 2 and 3 are slightly different from each other and from the 95% coverage rate suggested by many authors (NATION, 1990; LIU NA; NATION, 1985; LAUFER, 1989) for adequate reading comprehension. According to Francis and Kucera (1982) and Kucera (1982), for instance, the words belonging to the top 5,000 of the *Brown Corpus* cover 88.7% of a text while the 5,000 most frequent words in Hirsh and Nation's (1992) study have a 97.7% coverage rate. However, based on the fact that Kucera's (1982) and Francis and Kucera's (1982) findings are taken from a bigger corpus than Hirsh and Nation's (1992) findings, we assume that a vocabulary size of 5,000 words and a text coverage of 88.7% are more suitable for assessment of word frequency in machine-translated texts. A rate as high as the ones found in novels written for teenagers may be biased, since this kind of text would concentrate on using words that are frequent to the target audience, who presumably do not have much experience as readers nor the necessary years of schooling to comprehend difficult texts. It would be unreasonable to concentrate on a high coverage rate such as Hirsh and Nation's (1992) because the potential participants of the experiment under scrutiny are expected to have more experience in reading and more years of schooling.

Figure 5 shows the coverage rate of the experimental texts. The word frequency that serves as a basis for this analysis is extracted from Davies and Preto-Bay (2008). This publication was designed for language learning based on the assumption that allowing students to use frequency as a guide in their learning can help them learn a language. According to Davies (2014, p. 105) "[t]he dictionary is based on the actual frequency of words in the 20 million words in the texts from the 1900s in the *Corpus do Português*", thus it can be used to check word frequency as used by Jensen (2009). Words such as proper names were considered less frequent words because they were not listed in the frequency dictionary.

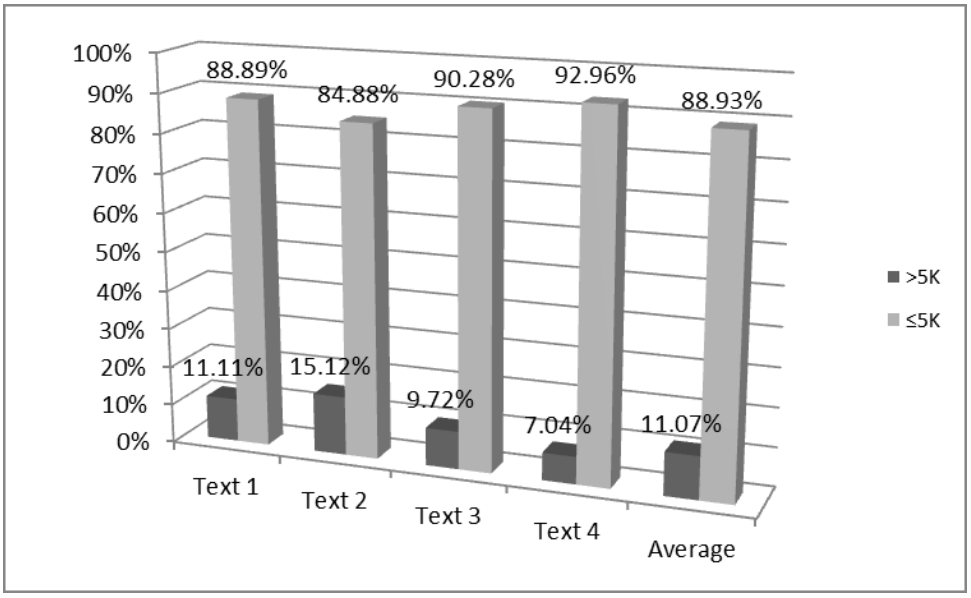

Figure 5. Word frequency of experimental texts.

Figure 5 indicates that the least frequent words represent from 7.04% in Text 4 to 15.12% in Text 2. In contrast, the most frequent words, which occur in Davies and Preto-Bay's (2008) dictionary, cover from 84.88% in Text 2 to 92.96% in Text 4. This number is close to the aforementioned 95% coverage rate upon which some authors (Nation, 1990; Liu and Nation, 1985; Laufer, 1989) have agreed for adequate reading comprehension. On average, the coverage rate for the four experimental texts is 88.93%, which is slightly higher than the 88.7% text coverage of the top 5.000 words in the *Brown corpus*.

4.3. A subjective indicator of text difficulty: post-editors' perceptions

Nineteen undergraduate students of language with a focus on translation were asked to assess the level of difficulty of the four experimental texts. Using a Likert scale, students had to choose one option from five alternatives (1-Very easy, 2-Easy, 3- Neither easy nor difficult, 4-Difficult, 5-Very difficult) to answer two questions for each experimental text: 1) How would you rate your level of difficulty to understand this text? and 2) How would you rate your level of difficulty to perform the monolingual post-editing of this text? To answer this second question, the students first performed each of the four monolingual post-editing tasks. It is necessary to emphasize that these students are helping to determine if the text selection made is valid. They are not going to take part in the definitive experiment, but they have a profile similar to the group of students that is going to perform the tasks.

Table 4 provides the level of difficulty the post-editors reported they had to comprehend and to perform the monolingual post-editing of the experimental texts. Each experimental text (T1, T2, T3, and T4) is followed by a letter, C or P, which mean comprehension and post-editing, respectively.

Table 4. Descriptive statistics of post-editors' perception of text difficulty[6].

	Minimum	Maximum	Median	Mean	Mode	Standard Deviation
T1C	1	3	3	2.53	3	0.77
T1P	1	4	3	2.58	3	0.84
T2C	1	4	2	2.53	2	1.02
T2P	2	4	3	2.68	3	0.67
T3C	2	5	4	3.68	4	1.05
T3P	2	5	4	4.11	4	0.94
T4C	1	5	2	2.79	2	1.27
T4P	1	5	4	3.37	4	1.16

As Table 4 indicates, students tend to evaluate differently the level of difficulty to comprehend and to perform the task. The greatest minimum-maximum variation occurs for Text 4, with scores ranging from 1 to 5 both for comprehension and post-editing. Mode also presents the greatest variation in Text 4, ranging from score 2 to comprehend to score 4 to post-edit the text. Although there is variation in level of difficulty to comprehend and to perform the monolingual post-editing of Texts 1 and 2, the results are similar in median, mode and mean. The variation in the post-editor's perception of difficulty to comprehend and to post-edit Text 3 occurs only in the standard deviation and the mean; all values of the other measures (minimum, maximum, median, and mode) are the same.

4.4. Statistical analysis

Principles of standard scores, i.e., the distance a value is above or below the mean in standard deviation units, were applied to have all eight readability indexes with different formulas and standard deviations in the same standard scale and then perform correlation tests with the other indicators of text complexity and difficulty. Correlation tests were applied to assess how readability indexes, word frequency, and post-editors' perception of text difficulty are related to each other. Correlation coefficients range from -1 to 1. Exact coefficients of ± 1

[6] Mean is the average of all numbers in a set of numbers; median is the middle value in a sequence of ordered numbers; mode is the most frequent number within a set of numbers.

indicate a perfect degree of association between two indicators, while an exact coefficient of 0 (zero) indicates no association. Coefficients within the [-1,0] and [0,+1] intervals are weaker the closer they reach 0 and stronger the closer they reach -1 or +1. Positive coefficients mean direct associations, i.e., the higher an indicator, the higher the other indicator. Negative coefficients mean inverse associations, i.e., the higher an indicator, the lower the other indicator.

Free software R environment (available at https://www.r-project.org/) was used to apply the principles of standard score and to perform correlation tests. Significance was set at $p < 0.05$ to reject the null hypothesis (H_0) as follows:

H_o: There is no relationship between the indicators of text complexity and difficulty analyzed in this paper.

The correlation results are represented by the formula "$r_s(df)=r_s$ coefficient, p", in which r_s stands for Spearman's correlation coefficient, df stands for degrees of freedom, and p stands for p value. The degrees of freedom were obtained as number of pairwise cases, i.e., the number of analyzed texts (4) minus 2, and number of post-editors (19) minus 2.

Spearman's correlation coefficient was first calculated individually to test the mean standard score of each readability index. Results pointed to a perfect uphill (positive) linear relationship between the mean standard score and Flesch-Kincaid [$r_s(2)=1$, $p=1.00$] and SMOG [$r_s(2)=1$, $p=0.08$] indexes, and a strong uphill (positive) relationship between the mean standard score and ARI [$r_s(2)=0.8$, $p=0.33$] and LIX [$r_s(2)=0.8$, $p=0.33$] indexes. The correlation coefficient pointed to a weak uphill (positive) linear relationship of the mean standard score with Coleman-Liau index [$r_s(2)=0.4$, $p=0.75$] and a weak downhill (negative) linear relationship of the mean standard score with Flesch Reading Ease-Coh-Metrix-Port index [$r_s(2)=-0.4$, $p=0.75$]. However, no relationship was found of the mean standard score with the last two readability indexes: Gunning-Fog $r_s(2)=0.2$, $p= 0.92$ and Flesch Reading Ease $r_s(2)=-0.2$, $p= 0.92$.

The correlation coefficient between the mean standard score of the first indicator (readability indexes) and the >5K band (uncommon words) of the second indicator (word frequency) showed a strong uphill (positive) linear relationship between the mean standard score of readability indexes and this specific band [$r_s(2)=0.8$, $p=0.33$]. In contrast, the correlation coefficient between the standard score and the ≤5K band (common words) pointed to a strong downhill (negative) linear relationship [$r_s(2)=-0.8$, $p=0.33$].

The correlation between mean standard score of the readability indexes and post-editor's perceptions resulted in a strong downhill (negative) linear relationship in both difficulty to comprehend, with $r_s(2)=-0.74$, $p=0.26$, and difficulty to post-edit, with $r_s(2)=-0.8$, $p=0.33$. The correlation between >5K band word frequency and post-editor's perceptions showed a strong downhill (negative) linear relationship of this band with difficulty to comprehend ($r_s= -0.74$, $p=0.26$) and a moderate downhill (negative) relationship with the difficulty to post-edit the texts ($r_s= -0.6$, $p=0.42$). Additionally, the ≤5K band showed a strong uphill (positive) relationship with the post-editor's perception of difficulty to comprehend [$r_s(2)= 0.74$, $p=0.26$] and a moderate (positive) relationship with their perception of difficulty to post-edit these texts ($r_s(2)= 0.6$, $p=0.42$).

Based on the *p* values, the null hypothesis was confirmed in all aforementioned analyses since all correlations proved to be statistically non-significant ($p < 0.05$), i.e., the correlations may be of interest, but they are not statistically significant. This could be due to the small sample of texts we analyzed herein.

The only correlation coefficient result that almost proved to be statistically significant was the one between mean post-editor's difficulty to comprehend and mean post-editor's difficulty to post-edit the texts ($r_s(2)=0.95$, $p=0.05$), with a strong uphill (positive) linear relationship. The null hypothesis was close to rejection in this case (5,1%). As the difficulty to comprehend the texts increases, the difficulty to post-edit them also increases, with results pointing to a nearly perfect relationship.

Differently from the analysis of the relationship between mean post-editor's perceptions of difficulty to comprehend and mean post-editor's difficulty to post-edit the texts, the correlation coefficient of all 19 post-editor's perceptions of difficulty to comprehend and difficulty to post-edit all texts [$r_s(17)=0.75$, $p=0.00$] showed a strong uphill (positive) linear relationship, with results being statistically significant, i.e., they reject the null hypothesis entirely. It also holds true for the relationship of these variables for each text individually: T1=> $r_s(17)=0.79$, $p=0.00$; T2=> $r_s(17)=0.74$, $p=0.00$; T3= $r_s(17)= 0.71$, $p= 0.00$; T4= $rs(17)= 0.82$, $p=0.00$. This means that when each post-editor's perception of difficulty to comprehend a text increases, their perception of difficulty to post-edit it also increases.

5. Concluding remarks

We posed some research questions regarding the selection of machine-translated texts into Portuguese to be used in a monolingual post-editing experiment. Some hypotheses emerged from these questions, and they were tested to confirm if the application of readability indexes, word frequency and post-editor's perception of difficulty to comprehend and to post-edit the texts could be used to select such texts.

By applying readability indexes to select texts other than English for monolingual post-editing, this study provides some evidence that they can be used to predict level of text complexity for experiments using Brazilian Portuguese as the target language. This holds true for the application of the adapted formula of Flesch Reading Ease to Portuguese, since the same pattern is observed in both formulas (Portuguese and English Flesch Reading Ease). This seems to partially confirm our first hypothesis: most of the applied readability indexes point out that machine-translated texts into Portuguese for experimental purposes feature the same level of difficulty. However, only half of them showed a strong or perfect relationship between the texts and the mean standard score, and two of them indicated no relationship whatsoever. Due to the small sample of texts, all these relationships proved to be statistically non-significant.

By challenging the notions of word frequency and word familiarity, as proposed by Jensen (2009), we also tested a hypothesis that a large Brazilian Portuguese corpus could be used to compare word frequency to word familiarity. Although being non-significant statistically, finding a strong correlation downhill (negative) of word frequency (from a list based on *Corpus do Português)* and the post-editor's perception of difficulty to comprehend, we sought to prove that this could be a way to compare word frequency to word familiarity and that needs further investigation.

The uncommon words (>5K band) of this list strongly correlated positively with the mean standard score of the readability indexes regarding level of post-editing difficulty of the experimental texts, while the common words (≤5K) strongly correlated negatively with the mean standard score. This points out that readability indexes can be used in conjunction with word frequency to assess post-editing and translation difficulty level of texts, not only for experimental setting purposes but also for selecting texts to be translated or post-edited in class and for selecting texts to be used in even Portuguese language classrooms and for producing textbooks according to years of schooling.

Within TPR, many researchers select texts for experiments by judging text difficulty themselves. However, this practice can lead them to compare tasks featuring too easy texts to tasks featuring too difficult texts. Then, one question arises: How can text difficulty be assessed without the researchers assessing the text themselves prior to an experiment? In this paper, we hypothesized that potential participants with a profile similar to those who will take part in an experiment could previously assess text difficulty to validate text selection made by using the other two indicators of complexity, readability indexes and word frequency. This concern was solved by asking foreign language students with a focus on translation to rate their perception of difficulty to comprehend the text before performing a post-editing task and to indicate their perception of difficulty to post-edit the text after performing such a task.

We analyzed the correlation of the mean post-editor's perceptions of difficulty to comprehend and their mean perception of difficulty to post-edit each text (T1, T2, T3, and T4) with the mean standard score. Both correlation coefficients proved that these perceptions and the mean standard score have a strong, negative relationship. This seems to indicate that the easiness to comprehend and to post-edit a text and the mean standard score are somehow related, but in an unexpected way: when the mean standard score decreases, the post-editor's perception of difficulty increases.

This also holds true for the correlation coefficients between the uncommon words of *Corpus do Português* and the mean post-editor's perceptions of difficulty to comprehend and to post-edit each text (T1, T2, T3, and T4). Both indicated a negative relationship of these words with the mean post-editor's perceptions of difficulty to comprehend and to post-edit the texts, the relationship post-editor's perception of difficulty to comprehend being strong, while relationship with post-editor's difficulty to post-edit being moderate. Once again, correlation existed but in an unexpected way (negative): a decrease in the uncommon words followed by an increase in the post-editor's perception of difficulty to comprehend and to post-edit the texts. In other words, decreasing the amount of uncommon words makes post-editor's perception of difficulty to comprehend and to post-editing the texts increase. In contrast, the correlation for the common words was the one expected (positive).

We also sought to answer the question: "How can we assess the relationship between the easiness to comprehend and the easiness to post-edit experimental texts?" The question arose because authors such as Mishra, Bhattacharyya and Carl (2013, p. 346) tend to claim "translation is not merely a reading activity." Translation surely involves more than just

reading, but the relationship between both activities (translation and reading) is a strong one. Besides, this was the only relationship that was very close to significance regarding the analysis of texts (5.1%).

By using a subjective measure, we hope to have provided insights into the difference between text complexity and text difficulty, since readability indexes as well as word frequency only point to text complexity. To have an idea of text difficulty and to investigate the relationship between easiness to comprehend and easiness to translate or post-edit texts, researchers should better assess the translators' or post-editors' perception of a text's difficulty for comprehension and execution of a task, rather than judging the text themselves.

Although readability indexes showed similar scores and seemed adequate for the post-editors on the basis of the post-editors' years of schooling, the texts imposed some challenges to them, especially Texts 3 and 4. Some of these challenges are related to less frequent words such as "obsolescência" ("obsolescence" in English) in Text 3, the word "Enforcement" that remains untranslated in English in Text 4, and the word "flocagem" ("flocking" in English) in Text 1. However, the discussion concerning word familiarity and word frequency needs more investigation since, in this analysis, words such as proper names, which usually are easily understood in context, were considered less frequent words because they were not listed as high frequency words in the list we used. When asking post-editors' perception, researchers can also ask them which words they do not know and compare their different perceptions of these words. Therefore, it is possible to predict not only text complexity but also text difficulty to comprehend, to translate or to post-edit as well as word familiarity, and thus to relate this word familiarity to word frequency. This could help to understand and to get around the limitations of word frequency, as mentioned by Jensen (2009), such as participants' different levels of familiarity with a word, no correlation between the occurrence of a word in high frequency band and the participants' difficulty to understand or translate it, and no correlation between the occurrence of a word in a high frequency band and the participants' easiness to translate it.

Despite the moderate to strong correlation coefficients found in the analysis of the relationship held between the indicators, these results have to be seen with caution since most *p* values indicated that they are not statistically significant. When selecting texts, researchers can preselect more texts (say, ten texts) and apply the same methodology herein described to source texts or to machine-translated texts and then ask some potential participants to assess

their perception of level of difficulty. By doing this, researchers can become more confident about their findings and select more adequate experimental texts among all the analyzed texts.

Moreover, the present results can be applied to the translation classroom. Teachers can use the same methodology to select texts to be translated or post-edited. They can encourage students to analyze the texts themselves before performing a task, so students can become aware of the potential effort to be invested on a task or the real effort invested on the task after performing it.

Acknowledgments

This research was funded by the Coordination for the Improvement of Higher Education Personnel (CAPES) in Brazil.

We are thankful to researchers Otávio Lima and Rodrigo Castro for their help in statistics. We also thank the reviewers for their valuable suggestions.

References

ALMEIDA, D. M.; ALUÍSIO, S. M. **Manual de uso do Coh-Metrix-Port 1.0.** Technical report NILC-TR-09-05, São Carlos, SP. 2009. Available at: http://nilc.icmc.usp.br/manual.pdf. Accessed on: 17 Aug. 2014.

ALUÍSIO, S.; SCARTON, C. E.; ALMEIDA, D. M. Análise de Inteligibilidade de textos via ferramentas de Processamento de Língua Natural: adaptando as métricas do Coh-Metrix para o Português. In: BRAZILIAN SYMPOSIUM IN INFORMATION AND HUMAN LANGUAGE TECHNOLOGY (STIL), 7., São Carlos, Brasil, 8-11 Sept. 2009. **Proceedings…** São Carlos: Universidade de São Paulo. 2009. Available at: http://www.nilc.icmc.usp.br/til/stil2009_English/Proceedings/stil/Scarton-57783_1.pdf. Accessed on: 12 May 2015.

ALVES, F.; KOGLIN A.; MESA-LAO, B.; MARTÍNEZ, M. G., FONSECA, N. B. L., SÁ, A. M., GONÇALVES, J. L., SZPAK, K. S.; SEKINO, K.; AQUINO, M. Analysing the impact of interactive machine translation on post-editing effort. In: CARL, M.; BANGALORE, S.; SCHAEFFER, M. (ed.). **New Directions in Empirical Translation Process Research**. Cham: Springer International Publishing, 2016. p. 77-94. http://dx.doi.org/10.1007/978-3-319-20358-4_4

ALVES, F.; PAGANO, A.; DA SILVA, I. A. A New window on translators' cognitive activity: Methodological issues in the combined use of eye tracking, key logging and retrospective protocols. In: MEES, I.; ALVES, F.; GÖPFERICH, S. (ed.). **Methodology, technology and innovation in translation process research**: a tribute to Arnt Lykke Jakobsen. Copenhagen: Samfundslitteratur, 2010. p. 267-292.

ALVES, F. **Zwischen Schweigen und Sprechen**: Wie bildet sich eine transkulturelle Brücke?: eine psycholinguistisch orientierte Untersuchung von Übersetzungsvorgängen zwischen portugiesischen und brasilianischen Übersetzern. Hamburg: Dr. Kovac. 1995.

CAMPBELL, S. A cognitive approach to source text difficulty in translation. **Target**, v. 11, n. 1, p. 33-63, 1999. http://dx.doi.org/10.1075/target.11.1.03cam

CAMPBELL, S. Critical structures in the evaluation of translations from Arabic into English as a second language. **The Translator**, v. 6, n. 2, p. 211-229, 2000. http://dx.doi.org/10.1080/13556509.2000.10799066

CAMPBELL, S.; HALE, S. What makes a text difficult to translate? In: ANNUAL ALAA CONGRESS, 23, Brisbane, Australia 30 June-03 July, 1998. **Refereed Proceedings**… 1999. Available at: http://pandora.nla.gov.au/nph-wb/20001105130000/http://www.cltr.uq.edu.au/alaa/proceed/camphale.html. Accessed on: 15 Sep. 2015.

CAMPBELL, S.; HALE, S. The interaction between text difficulty and translation accuracy. **Babel**, v. 48, n. 1, p. 14-33, 2002. http://dx.doi.org/10.1075/babel.48.1.02hal

CARL, M.; BANGALORE, S.; SCHAEFFER, M. **New Directions in Empirical Translation Process Research**. Cham: Springer International Publishing, 2016. http://dx.doi.org/10.1007/978-3-319-20358-4

DAVIES, M.; PRETO-BAY, Ana M. R. A frequency dictionary of Portuguese: core vocabulary for learners. Nova York/Londres: Routledge, 2008. p. 322-323.

DAVIES, M. The Corpus do português and the frequency dictionary of Portuguese. In: BERBER-SARDINHA, T.; FERREIRA, T. L. S. B. (ed.). **Working with Portuguese corpora**. Londres/Nova York: Bloomsbury Publishing Plc., 2014.

DAVISON, A.; GREEN, G. M. **Linguistic complexity and text comprehension:** readability issues reconsidered. Hillsdale, N.J.: L. Erlbaum Associates, 1988.

FONSECA, N. B. L. **Pós-edição monolíngue**: uma análise de indicadores do dispêndio de esforço temporal, técnico e cognitivo. 2016. 194 f. Tese (Doutorado em Estudos Linguísticos) – Faculdade de Letras, Universidade Federal de Minas Gerais, Belo Horizonte.

FRANCIS, W.N.; Kucera H. Frequency Analysis of English Usage. Boston: Houghton Mifflin Company. 1982.

HIRSH, D.; NATION, P. What vocabulary size is needed to read unsimplified texts for pleasure? **Reading in a Foreign Language**, v. 8, n. 2, p. 689-696, 1992.

JENSEN, K. T. H. Indicators of text complexity. **Copenhagen Studies in Language**, Copenhagen, v. 37, p. 61-80, 2009.

KRINGS, H. P. **Repairing texts**: empirical investigations of machine translation post-editing processes. Kent/Ohio/Londres: The Kent State University Press, 2001.

KÖNIGS, F. G. Der Vorgang des Übersetzens: Theoretische Modelle und praktischer Vollzug. Zum Verhältnis von Theorie und Praxis in der Übersetzungswissenschaft. **Lebende Sprachen**, v. 31, n. 1, p. 5-12, 1986. http://dx.doi.org/10.1515/les.1986.31.1.5

KÖNIGS, F. G. Was beim Übersetzen passiert. Theoretische Aspekte, empirische Befunde und praktische Konsequenzen. **Die Neueren Sprachen**, v. 86, n. 2, p. 162-185, 1987.

KÖNIGS, F. G. **Beim Übersetzen schreibt man- übersetzt man auch beim Schreiben?:** ein psycholinguistisch orientierter Vergleich zweier fremdsprachlicher Produktionsprozesse bei fortgeschrittenen deutschen Spanischlernern. Bochum: Ruhr Universität.1989.

KUCERA, H. The mathematics of language. **The American Heritage Dictionary**. 2. ed. Boston: Houghton Mifflin, 1982. p. 37-41.

LÄUBLI, S.; FISHEL, M.; MASSEY, G.; EHRENSBERGER-DOW, M.; VOLK, M. Assessing post-editing efficiency in a realistic translation environment. In: MT SUMMIT XIV WORKSHOP ON POST-EDITING TECHNOLOGY AND PRACTICE, 14. Nice, 2 Sept. 2013. **Proceedings…** Allschwil: European Association for Machine Translation, 2013. p. 83-91.

LAUFER, B. What percentage of text-lexis is essential for comprehension? In: LAUREN, C.; NORDMAN, M. (ed.), **Special language**: from humans to thinking machines. Clevedon, England: Multilingual Matters, 1989. p. 316-323.

LAUKKANEN, J. Affective and attitudinal factors in translation processes. **Target,** v. 8, n. 2, p. 257-274, 1996. http://dx.doi.org/10.1075/target.8.2.04lau

MANN, W. C.; THOMPSON. S. A. Rhetorical Structure Theory: a theory of text organization. **Technical Reports RS-87-190**. Los Angeles: Information Sciences Institute, 1987. Available at: http://www.sfu.ca/rst/pdfs/Man Marina n_Thompson_1987.pdf. Accessed on: 23 Aug. 2014.

MARTINS, T. B. F.; GHIRALDELO, C. M.; NUNES, M. G. V.; OLIVEIRA JR., O. N. Readability formulas applied to textbooks in Brazilian Portuguese. **Notas do ICMSC**, n. 28, p. 1-11, 1996. Available at http://www.nilc.icmc.usp.br/nilc/download/Reltec28.pdf. Accessed on: 15 Aug. 2014.

MCLAUGHLIN, G. SMOG grading: a new readability formula. **Journal of Reading**, v. 12, n. 8, p. 639-646, 1969.

MISHRA, A.; BHATTACHARYYA, P.; CARL, M. Automatically predicting sentence translation difficulty. In: ANNUAL MEETING OF THE ASSOCIATION FOR COMPUTATIONAL LINGUISTICS, 51., Sofia, Bulgaria, 4-9 Aug. 2013. **Proceedings…** Sofia: Association for Computational Linguistics, 2013, p. 346-351. Available at: http://www.aclweb.org/anthology/P13-2062. Accessed on: 14 Oct. 2014.

MOORKENS, J., O'BRIEN, S.; DA SILVA, I. A. L.; FONSECA, N. B. de L.; ALVES, A. Correlations of perceived post-editing effort with measurements of actual effort. Machine Translation, v. 29, n. 3-4, p. 267-284, 2015. http://dx.doi.org/10.1007/s10590-015-9175-2

NA, L.; NATION, I. S. P. Factors affecting guessing vocabulary in context. **RELC Journal**, v. 16, n. 1, p. 33-42, 1985. http://dx.doi.org/10.1177/003368828501600103

NATION, P.; WARING, R. Vocabulary size, text coverage, and word lists. In: SCHMITT, N.; FERREIRA, T. L. S. B. (ed.). **Vocabulary**: description, acquisition, pedagogy. Nova York: Cambridge University Press, 1997. p. 6-19.

NATION, I. S. P. **Teaching and learning vocabulary**. Boston: Heinle and Heinle Publishers, 1990.

O'BRIEN, S. Pauses as indicators of cognitive effort in post-editing machine translation output. **Across Languages and Cultures**, v. 7, n. 1, p. 1-21, 2006. http://dx.doi.org/10.1556/Acr.7.2006.1.1

O'BRIEN, S. The borrowers: researching the cognitive aspects of translation. **Target**, v. 25, n. 1, p. 5-17, 2013. http://dx.doi.org/10.1075/target.25.1.02obr

RODRIGUES, C. A abordagem processual no estudo da tradução: uma meta-análise qualitativa. **Cadernos de Tradução**, Florianópolis, v. 2, n. 10, p. 23-57. 2002.

SALDANHA, G.; O'BRIEN, S. **Research methodologies in translation studies**. Manchester: St Jerome Publishing, 2013.

SUN, S.; SHREVE, G. Measuring translation difficulty: an empirical study. **Target**, v. 26, n. 1, p. 98-127, 2014. http://dx.doi.org/10.1075/target.26.1.04sun

Appendixes

Machine-translated text 1

Voo mais eficiente na formação

Muitas pessoas estão muito familiarizados com a formação V utilizada por bandos de aves migratórias, e os cientistas determinaram que este é um modo eficiente de viagens que ajuda os pássaros economizar energia, especialmente em longas viagens migratórias. Mas o mesmo conceito está a ser considerada para melhorar a eficiência de aviões comerciais.

Entre os fabricantes de aviões, a Airbus é uma das empresas que procuram as vantagens de flocagem comercial. "Em uma formação em V de 25 animais, cada um pode obter uma redução da fricção induzida por até 65 por cento e aumentar a sua gama de 7 por cento (...)."

Source text available at:

http://www.ecogeek.org/computing-and-gadgets/3906-more-efficient-flight-in-formation

Machine-translated text 2

Cd velhas pode ser re-utilizado no tratamento de esgotos

Há uma boa chance de que tem sido um longo tempo desde que você comprou ou até mesmo usou um CD ou DVD em vez de arquivos de música digital ou streaming online. Com todos os CDs lá fora destinado a se tornar o lixo eletrônico, um novo método de tratamento de esgoto pode dar-lhes uma nova vida, em vez de vê-los acabam no aterro.

"Discos ópticos são baratos, facilmente disponíveis, e muito utilizada", diz Din Ping Tsai, um físico da Universidade Nacional de Taiwan, onde um grupo de pesquisadores surgiu com a idéia de usar os discos para limpar águas residuais.

Source text available at:

http://www.treehugger.com/clean-technology/old-cds-could-be-re-used-sewage-treatment.html

Machine-translated text 3

Obsolescência planejada: O que fazer com celular

A evolução contínua da obsolescência mercado de tecnologia faz com que milhares de telefones celulares. Na verdade, a cada ano não é mais usado em Espanha mais de 20 milhões de terminais.

O que devemos fazer com esses telefones que você não usa mais?

Nunca jogue-os no lixo, pois eles contêm poluentes orgânicos e materiais também porque os telefones celulares e equipamentos eletrônicos (desperdícios e) têm valor em si mesmos. Bem não pode considerar tratado como lixo, são equipamentos ou peças electrónica do mesmo, o que pode re-entrar no mercado de reuso podem ser reciclados ou materiais recuperados.

Source text available at:

http://catedraecoembesdemedioambiente.blogspot.com.br/2013/11/obsolescencia-programada-que-hacer-con.html

Machine-translated text 4

Enforcement ainda é insuficiente para tratar de questões ambientais e de saúde

No ano passado, as pessoas têm visto uma série de relatórios sobre os efeitos da poluição na saúde todos os dias, incluindo a água, o solo ea poluição do ar e outros riscos para a saúde ambiental, bem como uma variedade de questões de segurança alimentar chocantes.

A preocupação pública com o rápido aquecimento marca um ponto de viragem na causa de proteção ambiental de China. Mas vamos? Preocupações com a saúde pode trazer alguma mudança? Para resolver estes problemas no futuro o que precisamos?

Source text available at:

https://www.chinadialogue.net/article/show/single/ch/6926-Stronger-enforcement-won-t-be-enough-to-solve-China-s-environment-and-health-problems

Artigo recebido em: 17.02.2016 Artigo aprovado em: 24.06.2016