Type of the Paper (Article)

Post-editese in Literary Translations

Sheila Castilho[1*], Natália Resende[2,]

**both authors contributed equally to this work*

[1] School of Computing, Dublin City University; sheila.castilho@adaptcentre.ie
[2] School of Computing, Dublin City University; natalia.resende@adaptcentre.ie
* Correspondence: natalia.resende@adaptcentre.ie

Abstract: In the present study, we investigate the post-editese phenomenon, i.e., the unique features that set machine translated post-edited texts apart from human-translated texts. We use two literary texts, namely, the English children's novel by Lewis Carroll *Alice's Adventures in Wonderland* (AW) and Paula Hawkins' popular book *The Girl on the Train* (TGOTT) translated from English into Brazilian-Portuguese to investigate whether the post-editese features can be found on the surface of the post-edited (PE) texts. In addition, we examine how the features found in the PE texts differ from the features encountered in the human-translated (HT) and machine translation (MT) versions of the same source text. Results revealed evidence for post-editese for TGOTT only with PE versions being more similar to the MT output than to the HT texts.

Keywords: post-editing; machine translation; Portuguese; English; translationese; post-editese

1. Introduction

One of the biggest challenges for machine translation (MT) currently is to handle creative texts, such as literature, marketing content, etc., as these text types tend to contain a large amount of non-literal language, such as sarcasm, metaphor, irony and ambiguous elements of language that are likely to result in a word-by-word translation, thus compromising the rendering of the source text in the target language [1]. However, with the advent of neural MT systems (NMT), researchers in the field of artificial intelligence have identified a window of opportunity to translate creative texts more efficiently {2,3], as NMT systems are reported to outperform their predecessor, statistical MT systems, because they are able to learn the similarity between words and consider the context of the entire sentence, rather than just n-grams {4].

While a number of studies have investigated whether post-editing the MT output for literature might help literary translators in terms of productivity [e.g., 5, 2,3], translators' perception of MT is that the system is less useful for creative texts [5] than for other text types. In accordance with this, one study attempting to quantify creativity in MT and post-edited (PE) literary texts investigated whether the translation modes impact the reader experience [1]. The study has shown that human translation (HT) scores higher for creativity than PE translations, although for reading experiences related to emotional engagement and narrative presence, no statistically significant differences between HT, MT and PE have been found. These results suggest that MT might have just started to become a tool to be considered when translating creative texts, but it is still an open question whether there are characteristics typical of PE literary texts and whether these characteristics possibly make them less creative than HT texts. For that reason, more research on the MT output and PE for this textual domain is necessary.

In this work, we focus on the quest for the typical features of PE literary texts and the differences between PE texts from other comparable translated texts (MT and HT), that is, *post-editese* features. We believe that researching the features of the PE literary texts and contrasting them with HT texts, the raw MT output and their source texts will allow us to obtain a better understanding of the processes involved during the PE task and the influence of technology on the translation product of literary texts. In addition, we believe that awareness of these features can inform translators regarding the challenges they will face when using technology for translating creative texts.

1.1 The Present Study

According to Chesterman [6], the search for universal patterns lie into two categories: i) the search for universal patterns in translations through the comparison of features extracted from translated texts with features extracted from their source texts, as well as ii) the search for patterns in translations through comparisons of features extracted from translations and comparable (i.e., same text genre) non-translations in the same language. Chesterman [6] calls the search for universal patterns in translations using source texts as *S-universals* (S for source) and the search for universal patterns in translations using comparable non-translations *T-universals* (T for target). As our quest for the post-editese phenomenon involves capturing the differences between PE texts from other comparable translated texts (MT and HT), we focus on the quest for the features that have been associated in the literature [7-10] with the hypothetical T-universal features, namely, simplification, explicitation and convergence.

The idea of T-universals is also associated with the idea of *translationese* [18] which is the term used to refer to the language typical of translated texts that causes strangeness in the readers. Thus, in the present study, we adopt the term *translationese features* when referring to the T-universals examined. Following the rationale behind the extraction and analyses of the translationese features as described by Baker [7], linguistic features are extracted from our corpus composed of two literary texts, namely, the English children's novel by Lewis Carroll *Alice's Adventures in Wonderland* and Paula Hawkins' popular book *The Girl on the Train*, using a set of computational analyses with the purpose of identifying the existence of post-editese, i.e., features that are typical of PE texts. All features extracted from our corpus are compared between the HT version of the source text, the MT version of the source text and nine PE versions of the same MT output. As all translation versions originate from the same source text, we also extract features from the source to examine how much of the source text features is maintained in the translated versions.

Before presenting our methodology and the results of our experiments in detail, the next section presents an overview of the research in the field of translation studies addressing the features of translated texts as opposed to non-translated texts, as well as recent research focusing on the quest for post-editese features.

.

2. The Phenomenon of Post-Editese

In the field of translation studies, results of a number of research papers [e.g. 11-15] have shown that translated texts are statistically different from texts originally written in a certain language. Research has shown, for instance, that translated texts present less varied vocabulary and simpler syntax as reflected by lower type-token, i.e., lower lexical richness, and shorter mean sentence length than original texts [16,17, 13]. Research has also shown that translated texts tend to be more similar to each other than non-translated texts [17]. These differences are the product of the translation process that produces an interlanguage, the so-called *translationese*, that is, the language typical of translated texts [18], regardless of the source and target languages. According to Volansky et al.[13], the *translationese* phenomenon is the product of two coexisting forces that translators have to

cope with two during the translation process: the fidelity to the source text and the fluency in the target language. These two forces result in the strangeness of translated texts, that is, result in the translationese phenomenon.

Inspired by Toury's [19] norms of translation, Baker [7,8] proposes to investigate the linguistic and stylistic features of translated texts by looking for universal patterns that distinguish translated texts from non-translated texts using comparable corpora, naming these universal patterns as *Translation Universals*. Translation Universals are hypotheses of linguistic features common to all translated texts regardless of the source and target languages. The *translation universal* features proposed by Baker are: Simplification, Explicitation, Normalisation (or Conservatism) and Levelling out (or Convergence, as named by Corpas et al. [11]).

The hypotheses raised by Baker [7] on the characteristics common to all translated texts have aroused the interest of several researchers in the field of translation studies to investigate whether translationese features are manifested on the surface of translated texts. More recently, as the increased need for translation productivity in a globalised society resulted in the post-editing of the MT output, a number of studies [e.g., 20-22] from the natural language processing and MT fields have been discussing and investigating whether there are universal patterns typical of PE texts. Hence, the focus of attention has shifted from the typical features of HT texts to the typical features of PE texts.

Within the literature on translationese features, although several studies have shown that computers can distinguish, to a high degree of accuracy, between translations and originals [11; 24, 24, 13, 14], it is still unclear whether the same differences can be found between HT and PE texts. In contrast, the literature in the field of MT has shown some evidence that there might be differences between MT output and its PE version and HT texts. Several studies have shown, for instance, that the MT output differs from HT texts in terms of lexical variety. Vanmassenhove et al. [25] found that current MT systems processes cause a general loss in terms of lexical diversity and richness when compared to HTs. Thus, this loss in vocabulary range in the MT output may influence the product of PE translations, resulting consequently in differences between PE and HT texts.

Another example is the study from Culo and Nitzke [26] who found that terminology of PE texts is closer to MT output than to HT. The work of Groves and Schmidtke [27] also provides a clue to the existence of the post-editese phenomenon. The researchers compared the raw MT output produced by Microsoft's Treelet MT engine [28] with its PE counterpart, for English–German and English–French. They found that in the English-German corpus, there were many cases of changes in case and gender of nouns, removal of commas and pronouns such as the German pronoun *sie* and insertion of the determiner *die*. Similarly, in the English-French corpus, they found edits involving the deletion and insertion of the French function word *de*. Stylistic changes were also observed such as changes in words with the same meaning. The edits common to both corpora were: edits involving punctuation with removal or insertion of commas, changes in Part-of-Speech (determiners) and other structural changes: adjuncts and prepositional phrases and, in a smaller proportion, changes in terminology.

Despite the studies evidencing differences between MT output and PE texts and PE texts and HT texts, the study by Daems et al. [20] did not find evidence for the existence of post-editese. It was in this paper that the term "*post-editese*" was introduced, which the researchers define as "the expected unique characteristics of a PE text that set it apart from a translated text". The study investigated whether humans are able to distinguish PE from HT texts, and whether a supervised machine learning model could distinguish HT from PE texts. The results showed that neither humans nor the machine could distinguish between the translation modalities.

Contrary to the results reported by Daems et al. [2] Castilho et al. [22] found evidence for the existence of post-editese while investigating the features of PE texts in a corpus composed of HT, MT and PE texts in two domains: News and Literature. The authors also tested whether the PE level, the translators' experience, as well as the text domains influence the magnitude of the post-editese features. To this end, professional

translators and student translators PE the MT outputs of two different domains, namely news and literature, in the two different modalities of post-editing: full PE, in which more modifications were allowed, and light PE, in which translators were asked to use as much of the MT output as possible. The results revealed evidence of post-editese features as PE texts were found to be more similar to the raw MT output and source texts rather than to the HT texts.

Toral [21] has also found evidence for the manifestation of post-editese in PE texts. The author investigated the post-editese phenomenon using a set of computational analyses of a corpus composed of several datasets containing HTs and the PE texts, including different language directions and domains. The author found that the PE texts are simpler and have a higher degree of interference from the source language than HTs.

Considering this unclear scenario showing mixed results which leaves room for further discussion, in this article, we investigate the features of PE literary texts by comparing the features extracted from them with the features extracted from the raw MT output and the HT version of the source texts. As outlined previously, since translationese phenomenon has been found by a number of studies, we hypothesise here that the post-editese phenomenon will be found on the surface of PE literary texts as well, although manifested differently when compared to the translationese phenomenon emerging from HT texts.

Inspired by Gellerstam's [18] definition of *translationese* we define in the present study post-editese as follows:

Post-editese is the difference between the characteristics of human-translated texts (HT) and the post-edited (PE) versions, in relation to the raw MT output.

We propose to extract and analyse a series of linguistic features that have come to define the post-editese phenomenon in MT, that is, the unique characteristics of PE texts that set them apart from HT texts. Our quest for post-editese features in literary texts is guided by an overarching research question:

RQ: What are the characteristics of the PE literary texts?

In order to answer that, we use the rationale behind three translationese features as described by Baker [7], namely, simplification, explicitation and convergence. Thus, two sub-questions are posed:

RQ1- *Are the PE versions closer to the human translation (HT) or to the raw MT text (MT) and source (source) in terms of the translationese features?*

RQ2- *Which translationese features (as described by Baker [7]) can also support the post-editese hypothesis?*

Based on the results encountered in the literature, we hypothesise that post-editese will be manifested as PE texts being closer to the MT output and source texts, than HT texts are from either source texts or MT output. If we confirm our hypothesis, i.e., if we observe differences in features between the PE and HT texts, then we assume we have evidence for the existence of the post-editese phenomenon. Moreover, due to the difference in the genre of the two book excerpts (see section 3.1), we hypothesise that the degree of these differences will vary between the PE and HT from these two books excerpts, where one will require more edits than the other.

In the next subsections, we present the translationese features that will be addressed in the present study. The examination of these features along with findings reported in the post-editese literature [20-22] guide our experiments and analysis. Based on the results of our experiments, we discuss how our study can contribute to the quality of post-edited literary texts.

2.1 Simplification

According to Baker ([8], p. 181-182), simplification is "the tendency to simplify the language used in translation" and "involves making things easier for the reader". For Baker, as translators tend to split long sentences into smaller ones to facilitate text comprehension, simplification can be reflected by differences in the number of sentences and sentence length, as well as in punctuation, as "punctuation tends to be changed in translation in order to simplify and clarify". Moreover, simplification can be determined by comparing the vocabulary range and information load of the translated and original texts.

In the present study, the manifestation of the simplification feature in PE texts is investigated by calculating and comparing the lexical density (content words/words ratio), lexical richness (type/ token ratio), differences in punctuation between HT, MT and PE texts as well as sentence count and mean sentence length (in words and characters).

2.2 Explicitation

According to Baker ([8], p.180), explicitation means that "there is an overall tendency to spell things out rather than leave them implicit in translation". Therefore, HT texts tend to be longer than original texts in the same language. Moreover, HT texts tend to follow the source in using pronouns even when they are optional in the target language [13]. This is the case of the language pair studied here: English does not allow subject omission, while for PT-BR an explicit subject is optional as tense, person and number information expressed by the subject can also be inferred from the structure of the verbs [29]. In order to investigate explicitation phenomena and its manifestation as post-editese, we test whether PE texts are longer than MT and HT texts, and whether the amount of personal pronouns is different between the source, HT, MT and PE texts.

2.3 Convergence

Translated texts tend to be more similar to each other than non-translated texts [7, 8, 11]. For Baker ([8], pg. 177), convergence "simply means that we can expect to find less variation among individual texts in a translation corpus than among those in a corpus of original texts". Therefore, we investigate whether the convergence hypothesis holds true for PE texts when compared to source, HT and MT texts. We compute convergence by calculating variance scores for the features extracted from source, HT, MT and PE texts.

3. Procedures and Post-editese Features

In this section, we first describe the corpus used in our experiments, the post-editing process, and the features we consider to investigate the existence of post-editese phenomenon in the PE literary texts. In addition, we describe the experiments carried out to extract the features previously outlined. We first run a series of automatic metrics to investigate the differences in terms of edits between PE texts and the raw MT output and differences between the MT and HT. For the automatic metrics, we use the MultEVAL[1] tool which provides the (h)TER metric scores. For feature extraction, we create ad hoc programs using Python programming language.

3.1. Corpus

As mentioned previously, our corpus consists of two book excerpts: The children's novel by Lewis Carroll Alice's adventures in Wonderland (AW) and the popular novel The Girl on the Train (TGOTT) by Paula Hawkins. The AW test set is available in the Opus

[1] https://github.com/jhclark/multeval

Corpus [30] from which 250 sentences (5920 tokens) were selected from the source. The excerpt from the TGOTT, both the original in English and its human translation, is freely available online, from which 260 lines (5155 tokens) were selected. We chose the AW test set for two reasons: first, because it was the test set used in our previous work [22] and so it enables us to draw some correlations between the present study and this previous one; second, because it is a fantasy genre which contains metaphors, idioms and irony, thus its translation involves more creativity on the part of the translators and post-editors to adapt its rich language to target language [1]. In this text genre, not only the plot is important, but rather the author's individual use of language, i.e., the author's style. These characteristics allow us to contrast with the TGOTT test set because it is a thriller genre containing a more descriptive language of the plot, where action prevail over the author's language style. The source texts are in English and the translation versions of the source were in Brazilian Portuguese (PT-BR).

3.2 Translators, Tools and Guidelines

Tools: The source texts were translated using Google Translate (GT) from English into PT. The AW test set was translated in March 2020, while the TGOTT test set was translated in September 2020. The tool used for the post-editing task was the PET tool [31], and no time constraints were set for the task. A warm-up task for the translators to get acquainted with the tool and guidelines was set up. Translators were encouraged to ask questions about the tool and/or guidelines if needed.

Translators: We hired nine Brazilian professional translators to post-edit the MT output of the source texts. Translators filled out a questionnaire with questions about their background experience in translation and post-editing. Results of this questionnaire show that all translators have professional training in translation, ranging from professional experience, bachelors and masters. Their professional experience with translation ranges from 2 to +5 years. Although a few of the translators have translated novels, some of them have translated short literary texts during their training. Regarding experience with post-editing, only one translator reported not doing post-editing professionally. Moreover, over 60% of the translators reported to use MT for their daily work and like it.

Guidelines: Translators were given specific guidelines and were asked to follow them thoroughly. The first guideline was on how to use the tool, explaining all the functions and features, and the user interface. The post-editing guidelines instructed the translators to perform post-editing to achieve publishable professional quality translations, and not to look for the original translation of the source texts (both books have been translated into Brazilian Portuguese). Since the PET tool segments the source and MT by sentence, that is, each source sentence corresponds to one target sentence, in the translation of novels it is not that uncommon to have some cases of many-to-1 (more than 1 source sentence translated as 1 target sentence) or 1-to-many (1 source sentence translated as more than 1 target sentence). Therefore, the task guidelines instructed translators how to deal with merging or splitting sentences of the source text, and left it up to translators to decide when or whether they would like to do it.

4. Results and Discussion

4.1 Automatic Metrics

In order to measure the distance between the PE and the HT, and the distance between the PE versions and the MT output, we compute the automatic metric (h)TER ([32], p.225), which is "the minimum number of edits needed to change a hypothesis so that it exactly matches one of the references". The types of edits (h)TER accounts for are insertion, deletion, substitution of lexical items, and shifts of word sequences. The higher the

score for this metric, the greater is the difference between the text types. Table 1 shows the (h)TER scores using HT as reference and MT and PE texts as hypotheses, while Table 2 shows (h)TER score with PE texts as references and MT as the hypothesis.

Table 1. (h)TER scores using HT as reference, showing how far the PE versions and MT are from the HT.

Translation Type	AW	TGOTT
MT	47.8	59.2
T1	46.6	59.8
T2	45.4	59.4
T3	49.7	59.2
T4	46.9	60.6
T5	46.1	60.4
T6	47.2	59.5
T7	45.4	60.0
T8	48.6	51.5
T9	46.5	58.6
PE average	**46.9**	**58.7**

[1] Tables may have a footer.

We observe that for the AW test set, MT shows a 47.8 (h)TER score and for the TGOTT test set a 59.2 (h)TER score meaning that it would need a great amount of post-editing to make the raw MT output closer to the HT in both test sets.

We note that the PE versions obtained a reduced average score, 46.9 for AW and 58.7 for TGOTT, indicating that human intervention tends to distance the PE versions from the MT output. However, it is interesting to see that for T3 and T8, the (h)TER scores for the AW test set are even higher (49.7 and 48.6 respectively) compared to other PE texts which contradicts our hypothesis that with more human interference, the closer the PE versions would get to the HT. For the TGOTT test set, while the PE versions of T3, T8 and T9 have lower or the same (h)TER scores than the MT output, T1, T2, T4,T5 and T7's PE versions present higher (h)TER scores, suggesting they are more distant from the HT than the other translators. Nonetheless, it is evident from Table 1 that (h)TER scores calculated for PE texts are close to (h)TER scores calculated for MT output. These results suggest that while PE texts are distant from the HT texts, they are close to the MT output as the scores obtained indicate that translators did not add many edits to the MT output. This result is better exemplified in Table 2, which shows the amount of edits performed by the translators in the MT output.

Table 2. (h)TER scores using HT as reference, showing how far the PE versions and MT are from the HT.

Translation Type	AW	TGOTT
T1	19.7	05.5
T2	21.7	09.5
T3	39.1	11.9
T4	23.8	10.0
T5	19.4	07.2
T6	23.5	11.4
T7	19.6	07.9

T8	34.2	22.0
T9	30.3	13.5
PE average	**25.7**	**10.98**

[1] Tables may have a footer.

We observe that PE is performed very lightly for the TGOTT test set by all the translators as indicated by the low average h(TER) score obtained (10.98). T8 is the one who most interferes in the MT output (more post-editing performed) with a 22.0 (h)TER score. We hypothesise that due to the fact that the TGOTT test set contains more descriptive language rather than creative language, the translators were more prone to accept the MT output without editing it, as the MT output did not compromise the meaning of the source text. For the AW, we note that more post-editing is performed in comparison with the TGOTT test set, especially by T3 (40.6). Interestingly, T3's PE version is more distant from the HT in comparison with the other translators, as seen in Table 1, even though more post-editing was performed. We hypothesise that, even though T3 performs more PE, T3's lexical choices might not be the same as the original translation (HT).

4.2 Simplification

In this section, we present the descriptive analysis of the results to examine the simplification hypothesis. This descriptive analysis is based on the averages calculated for each of the features extracted sentence by sentence, namely, lexical density, lexical richness and sentence length (in words and characters). The inferential statistics indicating significant differences between source and translation versions and between translation versions themselves (HT, MT and PE texts) for each of these features as well as for punctuation feature are presented in section 4.5.

4.2.1 Lexical Richness

In order to compare the vocabulary range of PE text with the source, HT and MT texts, we calculate type-token ratio (TTR) sentence by sentence for all texts from each of the corpus. TTR is calculated as the number of types (number of unique lexical items in the text), divided by the number of total tokens (all lexical items in the text). The simplification hypothesis claims that texts originally written in a language present higher lexical richness than the comparable translated texts in the same language. However, because literature domain may involve more verbal artistry - e.g. paraphrase of figurative language and metaphors in the target language [7] - we hypothesise that the difference between the translated versions (HT, MT, and PE) might be lower. Specifically, in relation to the PE texts, we hypothesise they contain lower lexical richness than the originals and the HT texts based on the assumption that they will follow the MT pattern which might contain less varied vocabulary as pointed out in literature [25].

Table 3. Average Lexical richness scores. The higher the score, the more varied the vocabulary range.

Translation Type	**AW**	**TGOTT**
Source	0.93	0.93
HT	0.95	0.94
MT	0.95	0.95
T1	0.95	0.95
T2	0.95	0.95
T3	0.95	0.95
T4	0.94	0.95

T5	0.95	0.95
T6	0.95	0.95
T7	0.95	0.90
T8	0.95	0.95
T9	0.95	0.95
PE average	**0.95**	**0.95**

[1] Tables may have a footer.

We observed in Table 3 that source texts contain less varied vocabulary than the HT and PEs. For this finding, we share the same rationale described in Castilho et al. [22]. As PT-BR contains more verbal forms than English, these forms increased the number of types per verb root. We found, for instance, 120 occurrences of auxiliary verbs in the HT version, but only 37 in the original texts in the AW test set. Thus, we assume that, when rendering the original message in the target language, translators might have used more lexical resources increasing, consequently, the number of types in the translated texts.

Regarding the differences between HT and PE versions, we note that the PE versions present, on average, slightly less (0.95) lexical richness than the HT (0.94) in the TGOTT test set, and the same lexical richness in the AW test set (PE 0.95 vs. HT 0.95). Therefore, although the AW and TGOTT test sets differ in terms of amount of edits, according to the h(TER) scores where more PE was performed in the AW test set (Tables 1 and 2), translators tended to keep the vocabulary range of the MT output which, in turn, seems not to be greatly different from the HT.

Nonetheless, it is important to note that the averages suggest that PE versions are close to the MT in terms of lexical richness, especially in the TGOTT test set (both averages are 0.95). This is not the case in AW test set where we cannot observe differences between HT and PE texts. Thus, it seems that TGOTT confirms the post-editese hypothesis (which states that PE version will be closer to the MT and more distant from the HT) and AW reject it.

4.2.2 Lexical density

To compare the information load of the texts, that is, the information that is carried in content words (Nouns, verbs, adjectives and adverbs) between original text and the all translated versions of our corpus (HT, MT and PE texts), we extract lexical density features by calculating the ratio of the number of content words (nouns, verbs, adjectives, adverbs) to the total number of words sentence by sentence for all texts in the corpus. We use spaCy[2] part-of-speech tagger library available in Python programming language for PT language.

In this experiment, we exclude auxiliary verbs. As lower lexical density is a way of building redundancy and making a text simpler, the simplification hypothesis claims that HT texts present lower lexical density than comparable non-translated texts. We expect that the MT version will be similar to the source with lower lexical density compared to the HT, and consequently, the PE versions will follow the MT output as the PE versions originate in the MT, meaning PE versions will present lower lexical density than the HT. Table 4 shows the average lexical density scores for both test sets.

Table 4. Average lexical density scores, where the higher the score, the higher the ratio of the number of content words.

Translation Type	**AW**	**TGOTT**

[2] https://spacy.io/

Source	0.46	0.44
HT	0.44	0.49
MT	0.44	0.47
T1	0.44	0.47
T2	0.44	0.47
T3	0.44	0.48
T4	0.44	0.47
T5	0.44	0.47
T6	0.43	0.44
T7	0.43	0.47
T8	0.44	0.48
T9	0.43	0.44
PE average	**0.43**	**0.47**

[1] Tables may have a footer.

In both test sets shown in Table 4, we observe differences between the source texts and the HT and MT versions, where the source shows higher lexical density scores in the AW test set, but lower scores in the TGOTT test set. This is probably due to the differences between the languages.

Regarding the PE versions, it is possible to observe that, in the TGOTT test set, on average, the lexical density of PE versions (0.47) is closer to the MT (0.47) than to the HT (0.49), suggesting the post-editese hypothesis is confirmed[3]. In the AW test set, we did not find any clear pattern as all translated versions present very close lexical density scores on average (HT and MT 0.44, PE 0.43). In a closer examination of the translated versions for the AW test set, we found that the amount of adjectives, adverbs, nouns and verbs are very similar between HT, MT and PEs, thus resulting in close lexical density averages. We speculate that the pattern convergence between the PE versions and HT in the AW test set is due to the characteristics of the domain style. In order to maintain the amount of information of the source and author's style, the translated texts tend not to vary much in terms of lexical choices. Interestingly, the MT lexical density average is also very close to the HT in both test sets. As we observed for the lexical richness averages (Table 4), even though more PE was performed in the AW test set (Tables 1 and 2), translators kept the lexical choices of the MT output which in turn is not different from the HT. This result suggests that the MT output preserves the amount of content words of the original texts. This might be an indication of the MT output quality in relation to the preservation of the information load in the target language.

4.2.3 Sentence Count (SC) and Sentence Length (SL)

SC and SL are calculated by simply counting the total number of sentences and the sentence length (in words and characters) sentence by sentence. As mentioned previously, because translations tend to be simplified, the simplification hypothesis expects translated texts to have a higher number of sentences and that those sentences will be shorter than the sentences in the source texts [7]. Regarding the PE versions, we expect them to be closer to the MT, by showing lower sentence count and longer sentences compared to the HT. Table 5 shows the total sentence count and mean sentence length in words, while Table 6 shows mean sentence length in characters.

[3] Even though we note for T3 and T8 lexical density increases 0.01 point.

Table 5. Total Sentence Count and Mean Sentence Length (in words)

Translation Type	AW				TGOTT			
	count	longest	shortest	length (mean)	count	longest	shortest	length (mean)
Source	250	131	1	23.68	260	70	2	16.54
HT	262	128	1	21.63	261	78	2	16.80
MT	250	117	1	22.53	260	72	2	16.34
T1	252	119	1	22.07	260	72	2	16.33
T2	263	94	1	20.68	260	72	2	16.37
T3	260	124	1	20.84	260	73	2	16.41
T4	253	118	1	21.75	270	59	2	15.65
T5	251	119	1	22.51	260	72	2	16.39
T6	259	116	1	21.48	262	81	2	16.59
T7	258	122	1	21.72	260	74	1	16.46
T8	276	90	1	19.29	262	75	2	16.10
T9	269	125	1	20.83	260	74	2	16.38
PE average	**260. 1**	**-**	**-**	**21.24**	**261.5**	**-**	**-**	**16.3**

Table 6. Total Sentence Count and Mean Sentence Length (in characters)

Translation Type	AW				TGOTT			
	count	longest	shortest	length (mean)	count	longest	shortest	length (mean)
Source	250	693	8	123.38	260	371	10	88.88
HT	262	722	9	122.76	261	396	8	91.59
MT	250	688	8	126.55	260	391	9	90.00
T1	252	695	9	125.47	260	391	9	90.23
T2	263	526	8	120.25	260	391	9	90.19
T3	260	709	9	119.73	260	393	9	90.59
T4	253	686	9	122.36	270	312	9	86.13
T5	251	705	8	128.40	260	395	9	90.29
T6	259	704	9	122.08	262	427	9	90.95
T7	258	710	9	124.38	260	396	9	91.09
T8	276	527	7	110.55	262	398	8	88.39
T9	269	726	8	118.58	260	401	11	90.65
PE average	**260.1**	**-**	**-**	**121.3**	**261.5**	**-**	**-**	**89.9**

From Tables 5 and 6, we note that the source presents fewer sentences in both test sets when compared to the HT. The source presents longer sentences when compared to HT in the AW test set, but on average the same length in the TGOTT test set.

Regarding the comparison among the translated versions, from Table 5 we can see that , for the TGOTT test set, the HT has slightly more sentences (261) compared to the MT (260) and roughly the same amount compared to the average of the PE versions[4] (261.5). In relation to the sentence length, the PE versions present roughly the same sentence length in words and characters (16.3, 89.9) compared to the MT (13.34, 90.0), but they are slightly shorter compared to the HT (16.80, 91.89).

It is interesting to note that even though T8 and T9 have the same number of words for the shortest sentence (2 words - see Table 5), the number of characters for that same sentence are quite different (8 and 11 respectively - see Table 6):

EN: "Now look."

HT: "Veja só."

T8: "Veja só."

T9: "Agora veja."

This might explain the difference among translators not only for sentence length in words and characters, but also the differences between the lexical density, lexical richness and (h)TER scores. In this example, although the lexical density would be the same for both "Veja só" e "Agora veja" because "veja" is a verb in both cases and "só" and "agora" adverbs in both cases, as well as the type/token ratio would be 1 for both sentences (as 2 types divided by 2 tokens equals 1) the edits of T9 would be reflected the (h)TER scores with HT as reference.

For the AW test set, the MT version has the same number of sentences as the source (250), while the PE versions present more sentences on average (260.1) than the MT (250), being closer to the HT (262). Because the HT and the PE versions present more sentences than the MT, the average sentence length for the HT and the PE versions is lower when compared to the MT in words (PE 21.4, HT 21. 63, MT 23.56) and characters (PE 121.3, HT 122.76, MT 126.55). In contrast, again, HT and the PE versions patterns tend to converge both in terms of sentence count, i.e., post-editors and translators tend to split source sentences into more sentences as well as in terms of sentence length in words and characters as they tend to shorten the original sentences.

Thus, the results for sentence count and length seem not to confirm the post-editese hypothesis for both test sets as we *do not* observe a pattern where the PE versions are closer to the MT.

4.2.4 Punctuation

According to Baker [8] translated texts tend to have different punctuation marks when compared to the originals. In our corpus, we test for the most common punctuation marks such as question (?) and exclamations marks (!), colon (:), semi-colon (;) ellipsis (...), coma (,) parentheses (()), dash (-), double dash (--)[5], and full stop (.). We expect translated versions will differ from source text as translators tend to modify the punctuation marks in order to adapt the text to the punctuation system of the target language. Specifically in relation to the translation versions, we expect that the PE versions will follow the MT but will present differences in punctuation when compared to the source, HT and MT texts. Table 7 shows the total punctuation count for both test sets.

Table 7. Total punctuation count for both test sets.

[4] Apart from T4, who presents a few more sentences (270).

[5] Double dash is generally used to represent the break of a speech.

AW	**?**	**!**	**.**	**…**	**;**	**,**	**(**	**)**	**:**	**-**	**--**
Source	37	127	147	0	40	525	22	22	55	0	37
HT	39	125	208	15	39	493	22	22	63	14	8
MT	38	126	175	7	40	473	21	20	79	33	0
T1	38	117	171	24	41	495	22	22	78	8	0
T2	37	127	238	25	41	503	22	21	75	8	0
T3	38	120	251	31	41	567	22	21	64	1	0
T4	42	123	164	5	36	467	22	22	78	19	7
T5	37	127	164	6	40	483	22	21	70	22	1
T6	37	128	197	13	37	484	22	22	77	15	0
T7	39	127	208	19	45	508	23	23	67	10	0
T8	39	123	221	16	25	502	21	22	72	12	0
T9	34	125	265	23	26	610	22	22	51	1	1
PE Av.	**37.89**	**124.11**	**208.78**	**18**	**36.89**	**513.22**	**22**	**21.78**	**70.22**	**10.67**	**1**
TGOTT	**?**	**!**	**.**	**…**	**;**	**,**	**(**	**)**	**:**	**-**	**--**
Source	6	0	266	5	21	370	4	4	14	0	11
HT	7	0	268	5	26	381	2	2	10	17	0
MT	6	0	266	5	21	357	5	5	9	11	0
T1	6	0	266	5	21	357	3	3	8	7	0
T2	6	0	266	5	16	361	5	5	9	14	0
T3	6	0	266	5	22	357	5	5	7	11	0
T4	6	0	277	5	10	364	5	5	9	11	0
T5	6	0	266	5	21	356	5	5	9	11	0
T6	6	0	268	5	21	366	5	5	13	11	0
T7	5	0	270	6	23	361	5	5	9	8	0
T8	6	1	267	5	23	361	5	5	10	11	0
T9	5	0	270	6	24	388	5	5	10	6	0
PE Av.	**5.92**	**0.08**	**268**	**5.17**	**20.75**	**364.92**	**4.50**	**4.50**	**9.75**	**9.83**	**0.92**

From Table 7 we note that there are differences in punctuation counts between translation versions and source text and that the PE versions tend to follow the punctuation of the MT output. A qualitative analysis of the texts presented in Table 8 and Table 9 exemplifies these differences.

One example of differences in punctuation from original to translations is the use of ellipsis (...). Although very used in the HT and the PE versions (and a few times in the MT output) does not show in the original, which prefers the use of two dashes to indicate an abruptly unfinished thought. We see in Table 9 that while four translators (T1, T2,T8, and T9) decide to modify the single dash given by the MT output to ellipsis, five translators (T3, T4,T5,T6 and T7) keep the dash given in the MT output. Interestingly, we note that the lack of the comma after the word “Rome" present in the source but absent in the MT output, is followed by all the PE versions, while kept in the HT. We also note that T1 and T8, as in the HT version, decide to split the sentence into two, while all the other translators keep everything in a single sentence.

Table 8. Example of differences in punctuation for the AW test set between original, HT, MT and PEs, where the source uses -- to indicate an abruptly unfinished thought,

while MT translates into one – and HT uses the more common ellipsis in PT.

S	London is the capital of Paris, and Paris is the capital of Rome**,** and Rome--no, that's all wrong, I'm certain!
MT	Londres é a capital de Paris, e Paris é a capital de Roma e Roma - não, está tudo errado, tenho certeza!
HT	Londres é a capital de Paris, e Paris é a capital de Roma, e Roma... Não, está tudo errado, tenho certeza!
T1	Londres é a capital de Paris, e Paris é a capital de Roma e Roma… Não, está tudo errado, tenho certeza!
T2	Londres é a capital de Paris, e Paris é a capital de Roma e Roma… não, está tudo errado, tenho certeza!
T3	Londres é a capital de Paris, Paris é a capital de Roma e Roma - não, está tudo errado, tenho certeza!
T4	Londres é a capital de Paris, e Paris é a capital de Roma, e Roma - não, está tudo errado com certeza!
T5	Londres é a capital de Paris, e Paris é a capital de Roma e Roma - não, está tudo errado, tenho certeza!
T6	Londres é a capital de Paris, e Paris é a capital de Roma, e Roma - não, está tudo errado, tenho certeza!
T7	Londres é a capital de Paris, e Paris é a capital de Roma e Roma - não, está tudo errado, tenho certeza!
T8	Londres é a capital de Paris, e Paris é a capital de Roma, e Roma… Não, está tudo errado, tenho certeza!
T9	Londres é a capital de Paris, e Paris é a capital de Roma, e Roma... não, está tudo errado, tenho certeza!

Another example of difference in punctuation in the AW test set is shown in Table 9. We see that while the source shows a colon, the HT decides to split the sentence into two sentences. This change corroborates Baker's hypothesis that translations tend to modify the punctuation to simplify the sentences. Interestingly, since the MT version follows the source, all the PE versions follow the MT in this case and decide to also use the colon, making the MT and PE versions closer to the source.

Table 9. Example of differences in punctuation for the AW test set between original, HT, MT and PEs, where the source uses a colon, while the HT version decides to split the sentence in two.

S	**And so it was indeed: she was** now only ten inches high, and her face brightened up at the thought that she was now the right size for going through the little door

into that lovely garden.

HT **E de fato estava. Agora** ela tinha somente 25 centímetros de altura e o seu rosto iluminou-se com a idéia de que agora ela tinha o tamanho certo para passar pela portinha para aquele amável jardim.

MT **E assim foi, de fato: ela** agora tinha apenas dez centímetros de altura, e seu rosto se iluminou ao pensar que agora ela estava do tamanho certo para passar pela portinha daquele lindo jardim.

T1 **E assim foi, de fato: ela** agora tinha apenas vinte e cinco centímetros de altura, e seu rosto se iluminou ao pensar que agora ela estava do tamanho certo para passar pela portinha daquele lindo jardim.

T2 **E assim foi, de fato: ela** agora tinha apenas vinte e cinco centímetros de altura, e seu rosto se iluminou ao pensar que agora ela estava do tamanho certo para passar pela portinha daquele lindo jardim.

T3 **E assim foi, de fato: agora** ela tinha apenas dez centímetros de altura, e seu rosto se iluminou ao pensar que estava do tamanho certo para passar pela portinha daquele lindo jardim.

T4 **E era isso mesmo de fato: ela** agora tinha apenas trinta centímetros de altura, e seu rosto se iluminou ao pensar que agora ela estava do tamanho certo para passar pela portinha daquele lindo jardim.

T5 **E assim foi, de fato: ela** agora tinha apenas dez polegadas de altura, e seu rosto se iluminou ao pensar que agora ela estava do tamanho certo para passar pela portinha daquele lindo jardim.

T6 **E assim foi, de fato: ela** agora tinha apenas vinte e cinco centímetros de altura, e seu rosto se iluminou ao pensar que agora ela estava do tamanho certo para passar pela portinha que dava para aquele lindo jardim.

T7 **E assim foi, de fato: ela** agora tinha apenas vinte e cinco centímetros de altura, e seu rosto se iluminou ao pensar que agora ela estava do tamanho certo para passar pela portinha daquele lindo jardim.

T8 **E assim foi, de fato: ela** agora tinha apenas dez centímetros de altura, e seu rosto se iluminou ao pensar que agora estava do tamanho certo para passar pela portinha daquele lindo jardim.

T9 **E assim foi, de fato: ela** agora tinha apenas cerca de vinte e cinco centímetros de altura, e seu rosto se iluminou ao pensar que agora ela estava do tamanho certo para passar pela portinha até aquele lindo jardim.

4.3 Explicitation

4.3.1 Length Ratio

According to Baker [7], translated texts tend to be longer than original texts in the same language. We test whether this is the case for PE texts, i.e., whether there are differences between the length ratio of the PE versions and the HT and MT versions.

We expect to find that the PE versions will be longer than the MT based on the assumption that translators tend to interfere on the MT output adding more information to explicit things that are implicit in the MT output. In the same vein, we believe that the HT will be longer than the PE versions and the MT versions.

Table 10 shows the average length in characters for all translation types and average for all the PE versions combined. Table 11 displays the length ratios obtained for all comparisons made between the translated versions (HT, MT, and PEs).

Table 10. Average length (characters) per translation type for both test sets.

Translation Type	AW	TGOTT
Source	123.38	88.89
HT	122.76	91.59
MT	126.55	90.00
T1	125.47	90.23
T2	120.25	90.19
T3	119.72	90.60
T4	122.36	86.12
T5	128.40	90.29
T6	122.08	90.95
T7	124.37	91.09
T8	110.33	88.40
T9	118.58	90.65
PE average	**121.28**	**89.84**

Table 11. Length Ratio for both test sets. Ratios closer to 0 means that the second text (MT in the first and second rows, and PEs in the third row) is closer to the first text (HT in the first and third rows, and PE in the second row). A positive ratio means that the first texts are longer, while a negative ratio means the first texts are shorter.

Translation Type	AW	TGOTT
HT x MT	-0.03	0.02
PEs x MT	-0.04	0.00
HT x PEs	0.01	0.02
source x MT	-0.03	0.01

In Table 10, we see that the mean sentence length of all translated versions is longer than the mean sentence length of the source, in the TGOTT test set. We also note there are no differences between MT and PE texts as reflected by ratio 0.00 and that HT is longer than the MT and PE texts, thus confirming the post-editese hypothesis.

For the AW test set, we can see that the HT version has fewer characters than the source (also confirmed in Table 11 by the positive ratio of 0.01). This happens because the

HT has split the text into more sentences (as seen in Tables 5 and 6) thus reducing the number of characters per sentence. In contrast, the MT has more characters than the HT (MT 126.55 vs. HT 122.76) and the source (123.38), since the MT keeps the same number of sentences of the original text (Tables 5 and 6), resulting in more characters per sentence. It is worth mentioning once again that the reason the MT present more characters than the source and HT text is due to the differences between the languages. Regarding the differences between the PE versions and HT in AW test set, we note that the PEs tend to split texts into more sentences similarly to the HT, and consequently, the average sentence length of the PE versions is shorter than the MT, being close to the average sentence length of the HT, contradicting, therefore, the post-editese hypothesis.

4.3.2 Personal Pronoun Ratio (PPR)

To test if translated texts tend to follow the original in using personal pronouns (PPs) even when they are optional in the target language, we calculate the difference in the number of PPs between original and translated texts, divided by the count in the original [21], and also between the translated versions.[6] While we expect the original source texts to have a higher personal pronoun ratio since they might be optional in PT-BR, our post-editese hypothesis is that the MT version will be closer to the original as the systems tend to produce a word by word translation, having more PPs than the HT, and, consequently, that the PE versions will be closer to the MT, having more PPs than the HT. Table 12 shows the total number of PPs for both test sets, while Table 13 shows the PP ratio.

Table 12. Personal Pronoun count per test set.

Translation Type	AW	TGOTT
Source	562	456
HT	351	175
MT	366	213
T1	350	215
T2	351	217
T3	273	212
T4	343	229
T5	368	219
T6	331	213
T7	354	287
T8	238	192
T9	321	221
PE average	**328**	**222.78**

Table 13. Personal Pronoun Ratio per test set. Ratios closer to 0 are closer to the original. A positive ratio means that the first variable contains more PPs, while negative ratio means the first variable contains fewer PPs.

Translation	AW	TGOTT

[6] The PPs in English were: I, you, he, she, it, we, they, one, me, him, her, us, them, my, your, his, our, their, mine, hers, its, ours, theirs, oneself, myself, yourself, himself, herself, itself, ourselves, themselves. The PPs in Portuguese were: eu, me, mim, comigo, tu, te, ti, contigo, você, ele, ela, lhe, se, ele, ela, si, consigo, nos, nós, conosco, vós, vos, convosco, vocês, eles, elas, lhes, meu, minhas, meus, minhas, teu, tua, teus, tuas, dele, deles, dela, delas, nosso, nossos, nossa, nossas, vosso, vossos, vossa, vossas.

Type		
source x HT	0.38	0.62
source x MT	0.35	0.53
source x PEs	0.42	0.51
HTxMT	-0.04	-0.22
HTxPEs	0.07	-0.27
MTxPEs	0.10	-0.05

Table 12 shows a higher pronoun count for the source text than for the HT text in both test sets. Table 13 shows a positive pronoun ratio for both test sets when comparing HT and source texts (AW 0.38 and TGOTT 0.62). These results indicate that, compared to the HT version, the original contains more PPs as expected due to the differences between the languages.

For the AW test set, the MT versions are closer to the source (as reflected by the pronoun count (366) in Table 12 and the lower ratio (0.35) in Table 13) than the HTs (351 and 0.38, respectively). Regarding the ratio between the translated versions, we note that the MT has more PPs than both HT (-0.004) and PE (0.10), while the HT presents more PPs than the PE versions (0.07), contradicting the post-editese hypothesis. Interestingly, the ratio difference between MT vs PE is higher than PE vs HT. This is due to the fact that the MT is closer to the original, i.e. the MT tends to keep the number of pronouns of the original text indicating that the MT produces a word-by-word translation, while for the PE version, the translators tend to cut repetitive and unnecessary use of pronouns in Portuguese language. In contrast, in the TGOTT test set, it is noticeable that HT and PE are dissimilar in terms of PPs count (175 vs. 222 respectively) and PP ratio (0.353 vs. -0.222), confirming, therefore, our post-editese hypothesis for this feature which states that the PE has more PPs than the HT.

Overall, we can observe that while in the AW test set the number of personal pronouns tends to be closer to the HT, in the TGOTT test the number of personal pronouns in the PE versions tends to be close to the MT.

4.4 Convergence

According to Baker [7], translated texts tend to be more similar to each other than the original texts in the same language are similar to each other. To investigate this hypothesis, we compare the (dis)similarity within the translated texts (HT, MT and the PE versions). To compute the variance scores, it was only possible to select the features that were extracted sentence by sentence as the score provides an indication of the variance of the features within a set of values obtained for each of feature examined from each of the test sets. The features selected were: sentence length (SL), lexical richness (LR), lexical density (LD). Table 14 displays the variance of scores obtained within each of the test sets.

Table 14. Variance scores within feature scores extracted from each of the text types of the AW test set. The higher the variance score, the higher the dissimilarity within the test sets. Higher scores in bold.

Features	AW			
	source	HT	MT	PEs
LD	0.007	**0.008**	**0.008**	**0.008**
LR	**0.007**	0.005	0.005	0.005
SL (words)	**420.290**	378.360	360.110	302.023
SL (characters)	**12.458.810**	12.187.650	11.506.830	9.924.300

Table 15. Variance scores within feature scores extracted from each of the text types of the TGOTT test set. The higher the variance score, the higher the dissimilarity within the test sets. Higher scores in bold.

Features	TGOTT			
	source	HT	MT	PEs
LD	0.008	0.009	**0.010**	**0.010**
LR	**0.007**	0.005	0.006	0.006
SL (words)	118.840	**127.280**	117.960	117.960
SL (characters)	3468.820	**3794.680**	3634.140	3609.670

In the AW test set (Table 14), higher variance scores are found within all features from the source texts, except for the lexical density whose variance scores computed for all text types are very close to each other. This result supports the convergence hypothesis that predicts more variance within the set of non-translated texts (in this case the source text) than within the set of translated texts. Conversely, in the TGOTT test set (Table 15), the source texts vary slightly more than the within translated texts for lexical richness features only (source 0.007 vs. HT0.005 and MT/PE 0.006). For all the other features, higher variance scores are found for the MT and the PE versions (lexical density) and the HT (sentence length in words and characters). In other words, there is no clear pattern on variance scores within the features of the TGOTT test set. Thus, these results partly show the convergence hypothesis, which states that non-translated texts tend to vary more than translated texts, as they are found for the AW test set only. However, it is interesting to observe that, for both the AW and TGOTT test sets, variance scores obtained from the features within the MT and the PE versions are very similar, indicating that they vary to a similar extent in terms of lexical density, lexical richness, sentence length and sentence count.

4.5 Statistical Analysis

We computed *t*-tests to investigate the (dis)similarities between the texts types in order to confirm or reject the post-editese hypothesis for each simplification feature individually. The t-tests were computed to compare the distributions of the features extracted sentence by sentence between text types, namely, lexical richness, lexical density and sentence length (words and characters) as well as for punctuation feature. We first calculated the average of the nine PE texts sentence by sentence for each of the features analysed and then computed the *t*-test comparing the averaged PE texts with the source, HT and MT texts. The p-values obtained from the features extracted from the AW test set are shown in Table 16 and p-values obtained from the features extracted from the TGOTT test set are shown in Table 17.

Table 16. P-values for differences between text types from the AW test set, computed using t-test.

Features	AW				
	source x HT	HT x MT	PEs x MT	PEs x HT	PEs x source
LD	**0.00**	0.41	**0.04**	0.44	**0.00**
LR	**0.00**	0.87	0.59	0.57	**0.00**
SL (words)	**0.50**	0.92	0.73	0.66	**0.18**
SL (characters)	**0.50**	0.79	0.94	0.7	**0.63**
Punctuation	0.67	0.47	0.4	0.5	0.52

Table 17. P-values for differences between text types from the TGOTT test set, computed using t-test.

Features	TGOTT				
	source x HT	HT x MT	PEs x MT	PEs x HT	PEs x source
LD	**0.00**	**0.01**	0.36	**0.02**	**0.00**
LR	**0.01**	0.48	0.23	0.28	**0.00**
SL (words)	0.71	0.55	0.87	0.53	0.49
SL (characters)	0.52	0.69	0.72	0.72	0.18
Punctuation	0.51	0.21	0.34	0.21	0.62

Considering the P-values for the AW (Table 16) and for the TGOTT test sets (Table 17), we observe that statistical significant differences are found only within certain features for certain text comparisons. In the AW test set, none of the texts are significantly different for feature sentence length in words and characters (all P>0.05), but source texts significantly differ from HT in terms of lexical density and lexical richness (p<0.01). Source texts also significantly differ from the PE versions (P<0.01). This is an expected result, as pointed out previously, due to the differences between the languages of the source text and the language of the target texts. In contrast, we found a marginally significant difference (p<0.05) between the MT and the PE versions in relation to the lexical density feature in the AW test set. This is a surprising result since both MT and the PE versions present the same lexical density average score (see Table 4). This result reveals that, despite the similarity of the average lexical density between the MT and the PE versions, their distributions differ significantly, which suggests that translators interfered in the lexical choices of the MT output to improve its overall quality for publication purposes.

In the TGOTT test set, the MT and the PE versions do not differ significantly in any of the features examined suggesting less interference from the translators in the lexical lexical range and sentence length of the MT output. This similarity between MT and PE texts in TGOTT test was also revealed by the (h)TER scores in Tables 1 and Table 2. However, we can see a statistically significant difference between the HT and the PE versions in lexical density feature indicating that translators followed the lexical choices from the MT output, resulting in a distance from the HT lexical choices.

Therefore, we can see that the post-editese hypothesis was confirmed for the TGOTT test set for simplification feature lexical density only as reflected by a statistically significant difference between PE and HT texts for this feature. For all other features, although PE and HTs are not significantly different, we can see that PE and MT are not significantly different either, that is, we observe a convergence between them. Therefore, the post-editese hypothesis is confirmed partly for all other features. For the AW test set, in contrast, was not confirmed in any of the features examined, especially because we see significant differences between PE and MT output.

Regarding the differences in punctation, we note from Tables 16 and 17 that, although it is not possible to confirm the post-editese hypothesis for punctuation feature as there are no statistically significant differences between the text types (all P>0.05), the p-values obtained reveal that some distributions are more similar to each other. In TGOTT, we observe that the distribution of the punctuation counts of the PE versions is more similar to the distribution of the MT counts (as reflected by a greater p-value (p=0.34) than to the HT as reflected by a lower p-value (p=.21). For the AW is it possible to observe the inverse pattern, that is, PE versions are more similar to HT as reflected by a greater p-value (p=0.50) than to the MT (p=0.40). In addition, the distribution of the counts of the HT and MT is very different in the TGOTT test set (p=0.21) than in AW as indicated by greater p-value (AW p=0.47).

5. General discussion and Conclusions

In the present study, we investigate the existence of post-editese features in a corpus composed of excerpts from two different literary books: *Alice's Adventures in Wonderland* and *The Girl on the Train*. While the former contains a rich language style as the author plays on words, introducing puns, metaphors, the latter contains simple and relatively straightforward language where action and emotion prevail over the author's writing style.

In order to answer our RQ1 "*Are the PE versions closer to the HT or to the MT and source in terms of the translationese features?"*, we use the rationale behind the hypothetical features described by Baker [7] namely *simplification, explicitation* and *convergence.* Examining these features allowed us to investigate the differences between the HT and the PE versions to investigate the existence of *post-editese* phenomenon. Table 18 shows a summary of our findings.

Table 18. manifestation of post-editese per feature

Features	**Post-editese (HT vs PE)**	
	AW	TGOTT
LR	Not confirmed	Not confirmed
LD	Not confirmed	**Confirmed**
SC	Not confirmed	Not confirmed
SL	Not confirmed	Not confirmed
punct	Not confirmed	Not confirmed
LGHT R	Not confirmed	**Confirmed**
PPs	Not confirmed	**Confirmed**
Convergence	**Confirmed**	**Partly confirmed**

Regarding ***simplification***, from Table 18 we see that the *post-editese* hypothesis is not supported for the lexical richness (LR), sentence count (SC), sentence length (SL) and punctuation. Statistically significant differences between PE and HT texts is only observed for lexical density feature in TGOTT test set. Thus, the *post-editese* hypothesis, we find that, indeed, PE texts are different from the HT text. We also confirmed that the PE versions are closer to the MT, but in the TGOTT test set only. However, the same is not true for the AW test set as I did not find statistically significant differences between PE and HT texts in any of the simplification features examined.

Regarding sentence count, the post-editese hypothesis is not confirmed for the both test sets, as we found that the PE texts were similar to the HT texts. Finally, regarding punctuation, the qualitative analysis has revealed that the HT punctuation differs from the source punctuation both in TGOTT and AW, as punctuations were used by translators to split sentences and simplify the text. We also note that punctuation in PE tends to follow the MT punctuation more closely than the HT. However, we did not confirm the post-editese hypothesis for punctuation feature as significant differences in punctuation counts were not found between the texts types in any of the comparisons made.

Taking the results of simplification into consideration, our findings show a mixture of results as some simplification features were confirmed only for TGOTT, but none for the AW. Thus, regarding the question of whether they are good features to support the post-editese hypothesis (RQ2 *"which translationese features (as described by Baker [7]) can also support the post-editese hypothesis?") our findings show that, lexical richness, sentence length, sentence count and punctuation are not good indicators of the existence of post-editese in our corpus.*

As regards to the ***explicitation*** features, post-editese is confirmed for both features, i.e., length ratio and PPs ratio for the TGOTT test set, but not for the AW test set. Taking the results of explicitation into consideration, we can answer (RQ2) that length ratio and personal pronoun ratio were good indicators of the existence of post-editese hypothesis, but there is a difference between text sub-genres.

Finally, the ***convergence*** feature confirms post-editese for all features since we observe that, PE variance scores are similar to MT variance in both test sets. Thus convergence in our study is a good indicator of post-editese (RQ2).

Considering the results of all features together, we can note that post-editese was not confirmed for most of the features within the AW test set, but it was confirmed for more features in the TGOTT test set. Nonetheless, our findings show that the post-editese phenomenon is manifested on the surface of the post-edited texts as there are differences between those and HT versions. These differences are manifested in terms of the proximity or distance from the source and MT versions. While PE texts from the TGOTT test set are closer to the MT output in a series of features, the features extracted from the HT texts are more distant from the source and MT versions.

The major contribution of this work is the answer to our overarching research question *"What are the characteristics of the PE literary texts?"*. Our findings demonstrate that there is a clear difference between the literary genres: while literary texts whose author's style is full of figurative language pose a harder challenge to the MT system, texts that emphasise action over language style are less challenging. We validate this assumption based on our observations that AW involved more edits than the TGOTT test set, suggesting that the MT output is capable of expressing the meaning of the source text more efficiently than for the AW. Moreover, we find a more visible pattern in terms of features for the TGOTT test set when compared to the AW which, in turn, is unstable in terms of pattern manifestation. This allowed us to confirm our post-editese hypothesis for some features in the TGOTT but for none in the AW. Thus, based on our results, the main characteristics of PE literary texts is that they are similar to the MT output in terms of lexical density, use of pronouns and sentence length. However, this scenario can be blurred by the sub-genre of the literary text.

Further analysis in the different literary genres is necessary in order to answer our research question more comprehensively, and so, the question whether there are characteristics of PE literary texts that possibly make them less creative than HT texts remains open. Nevertheless, based on our results, we assume that literary creativity in PE texts may be compromised, as shown by the Guerberof-Arenas and Toral [1], due to the influence of the MT lexical and syntactic choices on the translators' choices. As seen, the MT output performs a translation that tends to be as equivalent as possible to their source texts. It is possible that when post-editing the raw MT output translators are primed by the MT choices even though they were instructed to change the text to achieve a high quality translation for publication standards, thus resulting in a PE text similar to the MT output. Consequently, this effect pushes them to converge to an equivalence with both the MT output, resulting in the manifestation of similar features and in the distortion of the writer's language style. At the same time, this result may also indicate that NMT systems are achieving good quality literary translations, especially for literary texts in which action prevail over the author's style, as translators did not need to interfere in the MT output in a great extent in order to obtain a high standard translation comparable to a high standard human translation.

Altogether, our results show that, when post-editing, translators should be aware of the priming effect of the raw MT output on their lexical and syntactic choices. The PE proximity to the MT output may result in distortion of the writers' style, consequently, influencing the final product of the post-edited texts. This is therefore the major challenge translators face when post-editing literary texts.

It is noteworthy that we are aware of the limitations of our study. Although all translators were professional, with more than 2 years of experience, not all of them were literary translators, which might mean they would not be as experienced in effectively

commanding the tone, author's style and creativity when modifying the MT output to adapt into linguistic framework of the target language. Since we can assume that Google translate provided 'good' quality translation based on the (h)TER scores (seen by the number of low edits), these translators who were not experienced with literary texts could have accepted the MT output, and kept most of the system's lexical and syntactic choices, resulting in fewer differences.

Another limitation is that the PET tool used for the translation might have restricted and biased the translators in not using the 1-to-many or many-to-1 option, that is, splitting or joining sentences, even though the guidelines allowed translators to do that. We speculate that, perhaps, if the translation task was set up in a word processor file, translators would feel freer to split/join sentences, and it would have given us different results. Finally, our study dealt with an unbalanced number of translated versions, with nine post-edited texts but only one human translation text. Thus, a more balanced dataset with more human translations from the same text could provide us more data that could allow us to run robust statistical analysis providing, consequently, more evidence for the existence or not of the post-editese phenomenon.

Therefore, with further study in the literary genres and post-editese, we will be able to collect more characteristics of PE literary texts which will be relevant to inform translators regarding other challenges they will face when using technology for translating different creative texts.

Data Availability Statement: In this section, please provide details regarding where data supporting reported results can be found, including links to publicly archived datasets analyzed or generated during the study. Please refer to suggested Data Availability Statements in section "MDPI Research Data Policies" at https://www.mdpi.com/ethics. You might choose to exclude this statement if the study did not report any data.

Acknowledgments: We would like to thank the professional translators for providing us with the post-edited versions for both corpora. This research was conducted with the financial support of the innovation programme under the Marie Sklodowska-Curie grant agreement No 843455 and also the Irish Research Council (GOIPD/2020/69). Science Foundation Ireland under Grant Agreement No. 13/RC/2106_P2 at the ADAPT SFI Research Centre at Dublin City University. ADAPT, the SFI Research Centre for AI-Driven Digital Content Technology, is funded by Science Foundation Ireland through the SFI Research Centres Programme.

Conflicts of Interest: "The authors declare no conflict of interest."

References

1. Ana Guerberof-Arenas and Antonio Toral. 2020. The Impact of Post-editing and Machine Translation on Creativity and Reading Experience. Translation Spaces. John Benjamins, 2020 9: 255–282.

2. Antonio Toral, Martijn Wieling, and Andy Way. 2018. 'Post-Editing Effort of a Novel with Statistical and Neural Machine Translation'. Frontiers in Digital Humanities 5: 1–11. https://doi.org/10.3389/fdigh.2018.00009.

3. Antonio Toral and Andy Way. What Level of Quality Can Neural Machine Translation Attain on Literary Text? In J. Moorkens, S. Castilho, F. Gaspari and S. Doherty (eds) *Translation Quality Assessment: From Principles to Practice* 1 (2018): 263-287.

4. WU, Y. et al. 2016. Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation. Computation and Language, v.1, 2016. https:// arxiv.org/pdf/1609.08144.pdf.

5. Joss Moorkens, Antonio Toral, Sheila Castilho, and Andy Way. 2018. 'Translators' Perceptions of Literary Post-Editing Using Statistical and Neural Machine Translation'. Translation Spaces 7(2): 240–62. https://doi.org/10.1075/ts.18014.moo.

6. Chesterman, A. Beyond Particular. In. Mauranen, A,; Kujamäki, P. Translation Universals. Do they exist? John Benjamins, 48, 2004, p.33-49.

7. Mona Baker. 1993. Corpus linguistics and translation studies: Implications and applications. In Gill Francis and Elena Tognini-Bonelli, editors, Text and Technology: In Honour of John Sinclair, John Benjamins Publishing Company, Netherlands, pages 233–252.

8. Mona Baker. 1996. Chapter corpus-based translation studies: The challenges that lie ahead. In Terminology, LSP and Translation: Studies in Language Engineering, in Honour of Juan C. Sager. Amsterdam: John Benjamins Publishing Company, page 175186.

9. Laviosa, S., 1996. *How Comparable Can 'Comparable Corpora' Be?*. John Benjamins Publishing Catalog. Available at: https://benjamins.com/catalog/target.9.2.05lav.

10. Mauranen, A. and Kujama ̈ki, P. (eds). (2004). Translation universals: Do they exist?. Philadelphia: John Benjamins.

11. Gloria Corpas, Pastor Ruslan Mitkov, and Viktor Pekar. 2008. Translation universals: Do they exist? corpus-based NLP study of convergence and simplification. In. Proceedings of the AMTA.

12. M. Koppel. and N. Ordan. 2011. Translationese and its dialects. Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies. Portland, Oregon: Association for Computational Linguistics, pp. 1318–26. http:// www.aclweb.org/anthology/P11-1132.

13. Vered Volansky, Noam Ordan, and Shuly Wintner.2013.On the features of translationese. Digital Scholarship in the Humanities 30 (1):98–118.https://doi.org/10.1093/llc/fqt031.

14. Marco Baroni and Silvia Bernardini. 2006. A new approach to the study of translationese: machine- learning the difference between original and trans- lated text. *Literary and Linguistic Computing* 21(3):259–274.

15. Resende, N. C. A. Testing the validity of translation universals by employing comparable corpora and NLP techniques. In. Historical Corpora: Challenges and Perspectives. Jost Gippert & Ralf Gehrke (eds.). Tubingen: Narr Verlag, 2015.

16. Stig Johansson. 1995. Mens sana in corpore sano: on the role of corpora in linguistic research. The European English Messenger4:19–25.

17. Laviosa, S. 1998. Core patterns of lexical use in a comparable corpus of English lexical prose. Meta, 43 (4):557–570.

18. Martin Gellerstam. 1986. Translationese in Swedish novels translated from English. In Wollin, L. and Lindquist, H. Translation Studies in Scandinavia. CWK Gleerup, Lund, volume 4, pages 88–95.

19. Toury, G. (1980). In Search of a Theory of Translation. The Porter Institute for Poetics and Semiotics, Tel Aviv University, Tel Aviv.

20. Joke Daems, Orphée De Clercq, and Lieve Macken. 2017. Translationese and post-editese: How comparable is comparable quality? Linguistica Antverpiensia New Series - Themes in Translation Studies 16:89–103.

21. Antonio Toral. 2019. Post-editese: an exacerbated translationese. In Proceedings of Machine Translation Summit. Dublin, Ireland.

22. Castilho, S.; Resende, N.C.A.; Mitkov, R. 2019. What Influences Post-editese features? A preliminary study. *Proceedings of the second workshop on Human-Informed Translation and Interpreting Technology (HiT-IT 2019). 5-6September, 2019, Varna, Bulgaria. http://rgcl.wlv.ac.uk/wp-content/uploads/2019/11/HiT-IT2019-proceedings.pdf*

23. Iustina Ilisei, Diana Inkpen, Gloria Corpas Pastor, and Ruslan Mitkov. 2010. Identification of translationese: a machine learning approach. In Gelbukh, A. F. (ed.), Proceedings of CICLing-2010:11th International Conference on Computational Linguistics and Intelligent Text Processing, volume 6008 of Lecture Notes in Computer Science. Pages 503–511.

24. Iustina Ilisei and Diana Inkpen. 2011. Translationese traits in Romanian newspapers: a machine learning approach. International Journal of Computational Linguistics and Applications 2(12).

25. Eva Vanmassenhove, Dimitar Shterionov, and Andy Way. 2019. Lost in translation: Loss and decay of linguistic richness in machine translation. In Proceedings of MT Summit XVII. Dublin, Ireland.

26. Oliver Culo and Jean Nitzke. 2016.Patterns of terminological variation in post-editing and of cognate use in machine translation in contrast to human translation. In Proceedings of the19th Annual Conference of the European Association for Machine Translation. Pages 106–114.https://www.aclweb.org/anthology/W16-3401.

27. Declan Groves and Dah Schmidtke. 2009. Identification and Analysis of Post-Editing Patterns for MT. In. Proceedingsof Machine Translation Summit XII: Commercial MT User Program.

28. Quirk, C., Menenzes, A., & Cherry, C. (2005). Dependency Treelet Translation: Syntactically Informed Phrasal SMT. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics* (pp. 271–279).

29. Noam Chomsky. 1993. In *Lectures on Government and Binding: The Pisa Lectures*. Holland: Foris Publi- cations.Reprint. 7th Edition, Berlin and New York: Mourton de Gruyter.

30. Jörg Tiedemann. 2012. Parallel data, tools and interfaces in OPUS. In Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC-2012). European Languages Resources Association (ELRA), Istanbul, Turkey, pages 2214–2218.http://www.lrec-conf.org/proceedings/lrec2012/pdf/463Paper.pdf.

31. Wilker Aziz, Sheila Castilho, Lucia Specia. 2012. PET: a Tool for Post-editing and Assessing Machine Translation. In Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12). Istanbul, Turkey

32. Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In AMTA 2006: Proceedings of the 7th Conference of the Association for Machine Translation in the Americas. Cambridge, MA, USA, pages 223–231.