

Article

Post-Editese in Literary Translations

Sheila Castilho ^{*,†} and Natália Resende ^{*,†} 

School of Computing, Dublin City University, Dublin 9, D09 E432 Dublin, Ireland

* Correspondence: sheila.castilho@adaptcentre.ie (S.C.); natalia.resende@adaptcentre.ie (N.R.)

† These authors contributed equally to this work.

Abstract: In the present study, we investigated the post-editeese phenomenon, i.e., the unique features that set machine translated post-edited texts apart from human-translated texts. We used 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). Both literary texts were Google translated from English into Brazilian Portuguese to investigate whether the post-editeese features can be found on the surface of the post-edited (PE) texts. In addition, we examined 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-editeese 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-editeese

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 (e.g., [2,3,5]) have investigated whether post-editing the MT output for literature might help literary translators in terms of productivity, translators' perception of MT is that the system is less useful for creative texts [5] than for other text types. In accordance with this, Guerberof-Arenas and Toral [1], while attempting to quantify creativity in MT and post-edited (PE) literary texts, investigated whether the translation modes impact the reader experience. Their 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 post-editing for this textual domain is necessary.

In this work, we focused 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) for the detection of post-editeese features [6]. 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



Citation: Castilho, S.; Resende, N. Post-Editese in Literary Translations. *Information* **2022**, *13*, 66. <https://doi.org/10.3390/info13020066>

Academic Editor: Ivan Dunder

Received: 7 December 2021

Accepted: 24 January 2022

Published: 28 January 2022

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Copyright: © 2022 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

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.

According to Chesterman [7], the search for universal patterns in translation lies in 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 [7] 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-editease 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 [8–11] with the hypothetical T-universal features, namely, simplification, explication, and convergence.

The idea of T-universals is also associated with the idea of translationese [12], 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 adopted the term translationese features when examining the T-universals. Following the rationale behind the extraction and analyses of the translationese features as described by Baker [8], linguistic features were extracted from 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-editease, i.e., features that are typical of PE texts. All features extracted from our corpus were compared between the HT version of the source text, the MT version of the source text and nine (9) PE versions of the same MT output. As all translated versions originate from the same source text, we also extracted features from the source text as a reference data source.

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-editease features.

2. The Phenomenon of Post-Editease

In the field of translation studies, results of a number of research papers, e.g., [13–17], 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 [15,18,19]. Research has also shown that translated texts tend to be more similar to each other than non-translated texts [19]. 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 [12], regardless of the source and target languages. According to Volansky et al. [15], the translationese phenomenon is the product of two coexisting forces that translators have to cope with 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, they result in the translationese phenomenon.

Inspired by Toury's [20] norms of translation, Baker [8,9] proposed 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. Therefore, translation universals are hypotheses of linguistic features common to all translated texts regardless of the source and target languages. The hypothetical universal features in translations proposed by

Baker are simplification, explicitation, normalization (or conservatism), and leveling out (or convergence, as named by Corpas et al. [13]).

The set of hypotheses raised by Baker [8] 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 globalized society resulted in the post-editing of the MT output, a number of studies, e.g., [6,21–23], 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 [13,15,16,24], 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 HT. 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-editeese 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. [6] did not find evidence for the existence of post-editeese. It was in this paper that the term post-editeese 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. [6], Castilho et al. [22] found evidence for the existence of post-editeese 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-editeese features. To this end, professional translators, and student translators post-edited the MT outputs of two different domains, namely news and literature, in the two different levels 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-editeese 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] also found evidence for the manifestation of post-editeese in PE texts. The author investigated the post-editeese 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 HT.

Considering this unclear scenario showing mixed results which leaves room for further discussion, in this article, we investigated 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 the phenomenon has been found by several studies, we hypothesized here that the post-edited phenomenon would 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 [12] definition of translationese, we defined in the present study post-edited as follows:

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

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

To answer that, we use the rationale behind three translationese features as described by Baker [8], 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 [8]) can also support the post-edited hypothesis?

Based on the results encountered in the literature, we hypothesized that post-edited would 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 confirmed our hypothesis, i.e., if we observed differences in features between the PE and HT texts, then we assumed we have evidence for the existence of the post-edited phenomenon. Moreover, due to the difference in the genre of the two book excerpts (see Section 3.1), we hypothesized that the degree of these differences would vary between the PE and HT from these two excerpts, where one would require more edits than the other.

In the next subsections, we present the translationese features that are addressed in the present study. The examination of these features along with findings reported in the post-edited literature [6,21,22] 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 ([9], pp. 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 ([9], 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 [15]. 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-editease, we tested whether PE texts are longer than MT and HT texts, and whether the number 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 ([9], p. 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 investigated whether the convergence hypothesis holds true for PE texts when compared to source, HT, and MT texts. We computed convergence by calculating variance scores for the features extracted from source, HT, MT, and PE texts.

3. Procedures and Post-Editease Features

In this section, we first describe the corpus used in our experiments, the post-editing process, and the features we considered to investigate the existence of post-editease phenomenon in the PE literary texts. In addition, we describe the experiments carried out to extract the features previously outlined. Next, we present a series of automatic metrics computed 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 used the MultEVAL (<https://github.com/jhclark/multeval>, 27 November 2021) tool which provides the (h)TER metric scores. For feature extraction, we created ad hoc programs using the Python programming language.

3.1. Corpus

As mentioned previously, our corpus consisted 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 dataset is available in the Opus 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 dataset for two reasons: first, because it was the dataset used in our previous work [22] and so it enabled 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, and thus its translation involves more creativity on the part of the translators and post-editors to adapt its rich language to the 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 allowed us to contrast with the TGOTT dataset because it is a thriller genre containing a more descriptive language of the plot, where action prevails over the author’s language style. The source texts were in English, and the translated versions of the source were in Brazilian Portuguese (PT-BR).

3.2. Tools, Translators and Guidelines

Tools: The source texts were translated using Google Translate (GT) from English into PT. The AW dataset was translated in March 2020, while the TGOTT dataset was translated in September 2020. We did not observe any impact of time differences in the quality of the translations. 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 were allowed to go back and modify the post-edited sentences as they wished, and we used the final post-edited version in our experiments.

Translators: Nine Brazilian professional translators were hired 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 showed that all translators had professional training in translation, ranging from professional experience to bachelor's and master's degrees. Professional experience with translation ranged from 2 to +5 years. Although a few translators had translated novels, some of them had 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.

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 on how to handle 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

To measure the distance between the PE and the HT, and the distance between the PE versions and the MT output, we computed the automatic metric (h)TER, which is “the minimum number of edits needed to change a hypothesis so that it exactly matches one of the references” ([32], p. 225). 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
PE1	46.6	59.8
PE2	45.4	59.4
PE3	49.7	59.2
PE4	46.9	60.6
PE5	46.1	60.4
PE6	47.2	59.5
PE7	45.4	60.0
PE8	48.6	51.5
PE9	46.5	58.6
PE average	46.9	58.7

We observed that for the AW dataset MT showed a 47.8 (h)TER score and for the TGOTT dataset 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 datasets.

Table 2. (h)TER scores using the PE versions as references, showing the number of edits performed in MT output.

Translation Type	AW	TGOTT
PE1	19.7	05.5
PE2	21.7	09.5
PE3	39.1	11.9
PE4	23.8	10.0
PE5	19.4	07.2
PE6	23.5	11.4
PE7	19.6	07.9
PE8	34.2	22.0
PE9	30.3	13.5
PE average	25.7	10.98

We noted that the PE versions obtained a lower 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, for PE3 and PE8, the (h)TER scores for the AW dataset were even higher (49.7 and 48.6, respectively) compared to other PE texts which contradicted our hypothesis that with more human interference, the closer the PE versions would be to the HT. For the TGOTT dataset, while the PE versions of PE3, PE8, and PE9 had lower or the same (h)TER scores than the MT output, PE1, PE2, PE4, PE5, and PE7's PE versions presented 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 the TGOTT 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 number of edits performed by the translators in the MT output.

We observed that PE was performed very lightly for the TGOTT dataset by all the translators as indicated by the low average h(TER) score obtained (10.98). PE8 is the one that most interferes in the MT output (more post-editing performed) with a 22.0 (h)TER score. We hypothesized that, because the TGOTT dataset contained 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 noticed that more post-editing is performed in comparison with the TGOTT dataset, especially by PE3 (39.1). Interestingly, although more post-editing was performed in PE3, i.e., there was more human interference in this PE version, it was more distant from the HT than the other PE versions in the AW dataset as seen in Table 1. We hypothesized that, even though PE3 performs more post-editing, the lexical choices from PE3 might not be the same as the HT.

4.2. Simplification

In this section, we presented the descriptive analysis of the results to examine the simplification hypothesis. This descriptive analysis was 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 statistically significant differences between source and translated versions (HT, MT, and PE texts) and between translated versions themselves for each of these features as well as for punctuation feature are presented in Section 4.5.

4.2.1. Lexical Richness

To compare the vocabulary range of PE texts with the source, HT, and MT texts, we calculated the type–token ratio (TTR) sentence by sentence for all texts from each of the datasets. TTR was 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 hypothesized that the difference between the translated versions (HT, MT, and PE) might be lower. Specifically, in relation to the PE texts, we hypothesized that 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].

We observe in Table 3 that source texts contained less varied vocabulary than the HT and PEs. For this finding, we shared 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 dataset. Thus, we assumed 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.

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
PE1	0.95	0.95
PE2	0.95	0.95
PE3	0.95	0.95
PE4	0.94	0.95
PE5	0.95	0.95
PE6	0.95	0.95
PE7	0.95	0.90
PE8	0.95	0.95
PE9	0.95	0.95
PE average	0.95	0.95

Regarding the differences between HT and PE versions, we noted that the PE versions presented, on average, slightly more (0.95) lexical richness than the HT (0.94) in the TGOTT dataset, and the same lexical richness in the AW dataset (PE 0.95 vs. HT 0.95). Therefore, although the AW and TGOTT datasets differed in terms of the number of edits, according to the h(TER) scores where more PE was performed in the AW dataset (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 dataset (both averages are 0.95). This is not the case in AW dataset, where we cannot observe differences between HT and PE texts. Thus, although it seemed that TGOTT might confirm the post-edite hypothesis, we did not find any statistically significant difference when comparing lexical richness between HT and PE texts (See Section 4.5).

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 extracted 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 used the spaCy (<https://spacy.io/>, accessed on 27 November 2021) part-of-speech tagger library available in Python programming language for PT language.

In this experiment, we excluded auxiliary verbs as we focused only on verbs carrying information. 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 expected that the MT version would be similar to the source with lower lexical density compared to the HT, and consequently, the PE versions would follow the MT output as the PE versions originated in the MT, meaning PE versions would present lower lexical density than the HT. Table 4 shows the average lexical density scores for both datasets.

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
Source	0.46	0.44
HT	0.44	0.49
MT	0.44	0.47
PE1	0.44	0.47
PE2	0.44	0.47
PE3	0.44	0.48
PE4	0.44	0.47
PE5	0.44	0.47
PE6	0.43	0.44
PE7	0.43	0.47
PE8	0.44	0.48
PE9	0.43	0.44
PE average	0.43	0.47

In both datasets shown in Table 4, we observed differences between the source texts and the HT and MT versions, where the source showed higher lexical density scores in the AW dataset, but lower scores in the TGOTT dataset. This is probably due to the differences between the English and Portuguese languages.

Regarding the PE versions, it is possible to observe that, in the TGOTT dataset, on average, the lexical density of PE versions (0.47) was closer to the MT (0.47) than to the HT (0.49), suggesting the post-edite hypothesis is confirmed (even though we note for PE3 and PE8 lexical density increased 0.01 point). In the AW dataset, we did not find any clear pattern as all translated versions presented 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 dataset, we found that the number of adjectives, adverbs, nouns, and verbs were 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 dataset was due to the characteristics of the domain style. 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 was also very close to the HT in both datasets. As we observed for the lexical richness averages (Table 4), even though more PE was performed in the AW dataset (Tables 1 and 2), translators kept the lexical choices of the MT output which in turn was 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 expected translated texts to have a higher number of sentences and that those sentences would be shorter than the sentences in the source texts [7]. Regarding the PE versions, we expected 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.

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
PE1	252	119	1	22.07	260	72	2	16.33
PE2	263	94	1	20.68	260	72	2	16.37
PE3	260	124	1	20.84	260	73	2	16.41
PE4	253	118	1	21.75	270	59	2	15.65
PE5	251	119	1	22.51	260	72	2	16.39
PE6	259	116	1	21.48	262	81	2	16.59
PE7	258	122	1	21.72	260	74	1	16.46
PE8	276	90	1	19.29	262	75	2	16.10
PE9	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
PE1	252	695	9	125.47	260	391	9	90.23
PE2	263	526	8	120.25	260	391	9	90.19
PE3	260	709	9	119.73	260	393	9	90.59
PE4	253	686	9	122.36	270	312	9	86.13
PE5	251	705	8	128.40	260	395	9	90.29
PE6	259	704	9	122.08	262	427	9	90.95
PE7	258	710	9	124.38	260	396	9	91.09
PE8	276	527	7	110.55	262	398	8	88.39
PE9	269	726	8	118.58	260	401	11	90.65
PE average	260.1	-	-	121.3	261.5	-	-	89.8

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

Regarding the comparison among the translated versions, from Table 5 we can see that, for the TGOTT dataset, the HT had one more sentence (261) compared to the MT (260)

and roughly the same number compared to the average of the PE versions (261.5) (apart from PE4, which presented a few more sentences (270)). In relation to the sentence length, the PE versions presented roughly the same sentence length in words and characters (16.3, 89.9) compared to the MT (16.34, 90.0), but they were slightly shorter compared to the HT (16.80, 91.59).

It is interesting to note that even though PE8 and PE9 had the same number of words for the shortest sentence (2 words—see Table 5), the number of characters for that same sentence were quite different (8 and 11, respectively—see Table 6):

EN: “Now look”.

HT: “Veja só”.

PE8: “Veja só”.

PE9: “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 features and the (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 being 1 for both sentences (as 2 types divided by 2 tokens equals 1), the edits of PE9 would be different as reflected by the (h)TER scores with HT as reference.

For the AW dataset, the MT version had 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 presented more sentences than the MT, the average sentence length for the HT and the PE versions was 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 tended to converge both in terms of sentence count, i.e., post-editors and translators tended 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-edite hypothesis for both datasets as we did 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 tested for the most common punctuation marks such as question (?), exclamation marks (!), colon (:), semi-colon (;), ellipsis (...), comma (,), parentheses (()), dash (-), double dash (--) (double dash is generally used to represent the break of a speech), and full stop (.). We expected translated versions would differ from source text as translators tend to modify the punctuation marks to adapt the text to the punctuation system of the target language. Specifically in relation to the translated versions, we expected that the PE versions would follow the MT but would present differences in punctuation when compared to the source, HT, and MT texts. Table 7 shows the total punctuation count for both datasets.

Table 7. Total punctuation count for both datasets.

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
PE Av.	37.89	124.11	208.78	18	36.89	513.22	22	21.78	70.22	10.67	1

Table 7. Cont.

TGOT	?	!	;	,	()	:	-	—
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
PE Av.	5.78	0.11	268.44	5.22	20.11	363.44	4.78	4.78	9.33	10	0

From Table 7, we noted that there were differences in punctuation counts between translated versions and source text and that the PE versions tended to follow the punctuation of the MT output. A qualitative analysis of the texts presented in Tables 8 and 9 exemplifies these differences.

Table 8. Example of differences in punctuation for the AW dataset 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!
PE1	Londres é a capital de Paris, e Paris é a capital de Roma e Roma . . . Não, está tudo errado, tenho certeza!
PE2	Londres é a capital de Paris, e Paris é a capital de Roma e Roma . . . não, está tudo errado, tenho certeza!
PE3	Londres é a capital de Paris, Paris é a capital de Roma e Roma—não, está tudo errado, tenho certeza!
PE4	Londres é a capital de Paris, e Paris é a capital de Roma, e Roma—não, está tudo errado com certeza!
PE5	Londres é a capital de Paris, e Paris é a capital de Roma e Roma—não, está tudo errado, tenho certeza!
PE6	Londres é a capital de Paris, e Paris é a capital de Roma, e Roma—não, está tudo errado, tenho certeza!
PE7	Londres é a capital de Paris, e Paris é a capital de Roma e Roma—não, está tudo errado, tenho certeza!
PE8	Londres é a capital de Paris, e Paris é a capital de Roma, e Roma . . . Não, está tudo errado, tenho certeza!
PE9	Londres é a capital de Paris, e Paris é a capital de Roma, e Roma . . . não, está tudo errado, tenho certeza!

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) it does not show in the original, which preferred the use of two dashes to indicate an abruptly unfinished thought. We see in Table 9 that while four translators (PE1, PE2, PE8, and PE9) decided to modify the single dash given by the MT output to ellipsis, five translators (PE3, PE4, PE5, PE6, and PE7) kept the dash given in the MT output. Interestingly, we noted that the lack of the comma after the word “Rome” present in the source, but absent in the MT output, was followed by all the PE versions, while kept in the HT. We also noted that PE1 and PE8, as in the HT version, decided to split the sentence into two, while all the other translators keep everything in a single sentence.

Another example of difference in punctuation in the AW dataset is shown in Table 9. We see that while the source showed a colon, the HT decided 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 followed the MT in this case and decided 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 dataset 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.
PE1	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.
PE2	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.
PE3	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.
PE4	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.
PE5	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.
PE6	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.
PE7	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.
PE8	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.
PE9	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 [8], translated texts tend to be longer than original texts in the same language. We tested whether this is the case for PE texts, i.e., whether there were differences between the length ratio of the PE versions and the HT and MT versions.

We expected to find that the PE versions would 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 believed that the HT would be longer than the PE versions and the MT versions.

Table 10 shows the mean sentence 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. Mean sentence length (characters) per translation type for both datasets.

Translation Type	AW	TGOTT
Source	123.38	88.89
HT	122.76	91.59
MT	126.55	90.00
PE1	125.47	90.23
PE2	120.25	90.19
PE3	119.72	90.60
PE4	122.36	86.12
PE5	128.40	90.29
PE6	122.08	90.95
PE7	124.37	91.09
PE8	110.33	88.40
PE9	118.58	90.65
PE average	121.28	89.84

Table 11. Length ratio for both datasets. 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 × MT	−0.03	0.02
PEs × MT	−0.04	0.00
HT × PEs	0.01	0.02
source × MT	−0.03	0.01

In Table 10, for the TGOTT dataset, we saw that the mean sentence length of all translated versions was longer than the mean sentence length of the source. We also noted there were no differences between MT and PE texts as reflected by ratio 0.00 and that HT was longer than the MT and PE texts, thus confirming the post-editese hypothesis.

For the AW dataset, we can see that the HT version had fewer characters than the source. This happened because the HT had 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 had more characters than the HT (MT 126.55 vs. HT 122.76) and the source (123.38), since the MT kept 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 for the MT presenting more characters than the source is due to the differences between English and Portuguese. Regarding the differences between the PE versions and HT in AW dataset, we noted that the PE texts tended to split texts into more sentences similarly to the HT, and consequently, the average sentence length of the PE versions was 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 tended to follow the original in using personal pronouns (PPs) even when they were optional in the target language, using *ad hoc* Python scripts, we extracted the frequency of all PPs in Portuguese from each dataset to 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. (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.) While we expected the original source texts to have a higher personal pronoun ratio since they might be optional in PT-BR, our post-editese hypothesis was that the MT version would be closer to the original as the systems tended to produce a word-by-word translation, having more PPs than the HT, and, consequently, that the PE versions would be closer to the MT, having more PPs than the HT. Table 12 shows the total number of PPs for both datasets, while Table 13 shows the PP ratio.

Table 12. Personal pronoun count per dataset.

Translation Type	AW	TGOTT
Source	562	456
HT	351	175
MT	366	213
PE1	350	215
PE2	351	217
PE3	273	212
PE4	343	229
PE5	368	219
PE6	331	213
PE7	354	287
PE8	238	192
PE9	321	221
PE average	328	222.78

Table 13. Personal pronoun ratio per dataset. 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 Type	AW	TGOTT
source × HT	0.38	0.62
source × MT	0.35	0.53
source × PEs	0.42	0.51
HT × MT	−0.04	−0.22
HT × PEs	0.07	−0.27
MT × PEs	0.10	−0.05

Table 12 shows a higher pronoun count for the source text than for the HT text in both datasets. Table 13 shows a positive pronoun ratio for both datasets 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 dataset, the MT versions were 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 noted the MT had more PPs than both HT (−0.004) and PE (0.10), while the HT presented more PPs than the PE versions (0.07), contradicting the post-editese hypothesis. Interestingly, the ratio difference between MT vs. PE was higher than PE vs. HT. This was probably because the MT was closer to the original, i.e., the MT tended 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 dataset, 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-edited hypothesis for this feature which states that the PE has more PPs than the HT.

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

4.4. Convergence

According to Baker [8], translated texts tend to be more similar to each other than original texts in the same language are similar to each other. Baker named this feature as “leveling out”, also known in the literature as “convergence” [13], which concerns “the tendency of a translated text to gravitate towards the center of a continuum” [8] p. 177. There is some evidence in the literature that variance (a statistical measure of heterogeneity) is consistently lower on lexical density and lexical richness for translated texts compared to texts originally written in a certain language. We calculate this feature for post-edited versions as an attempt to investigate whether this hypothesis also holds true for the context of the post-edited research, that is, whether there are differences in terms of homogeneity and heterogeneity between PE texts and the other translated versions (HT and MT). This feature has not received much attention in the literature; moreover, we are not aware of previous work testing this hypothesis in the post-edited research. However, we believe that this feature can reveal whether MT texts tend to push the features to drift toward “the center of any continuum rather than move towards the fringes” and whether translators tend to reproduce this tendency.

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 heterogeneity of the features within a set of values obtained for each feature examined from each of the datasets. The features selected were sentence length (SL), lexical richness (LR), and lexical density (LD). Table 14 displays the variance of scores obtained within each of the datasets.

Table 14. Variance scores within feature scores extracted from each of the text types of the AW dataset. The higher the variance score, the higher the dissimilarity within the datasets. 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

In the AW dataset (Table 14), higher variance scores were 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 dataset (Table 15), the source texts vary slightly more than the within translated texts for lexical richness features only (source 0.007 vs. HT 0.005 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 dataset. Thus, these results partially showed the convergence hypothesis, which states that non-translated texts tend to vary more than translated texts, as they are found for the AW dataset only. However, it is interesting to observe that, for both the AW and TGOTT datasets, 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.

Table 15. Variance scores within feature scores extracted from each of the text types of the TGOTT dataset. The higher the variance score, the higher the dissimilarity within the datasets. 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

4.5. Statistical Analysis

We computed *t*-tests to investigate the (dis)similarities between the text types in order to confirm or reject the post-edite hypothesis for each simplification feature individually. The *t*-tests were computed in language R 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 analyzed 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 dataset are shown in Table 16 and *p*-values obtained from the features extracted from the TGOTT dataset are shown in Table 17.

Table 16. *p*-values for differences between text types from the AW dataset, computed using *t*-test.

Features	AW				
	Source × HT	HT × MT	PEs × MT	PEs × HT	PEs × 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 dataset, computed using *t*-test.

Features	TGOTT				
	Source × HT	HT × MT	PEs × MT	PEs × HT	PEs × 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 datasets (Table 17), we observed that statistically significant differences were found only within certain features for certain text comparisons. In the AW dataset, none of the texts had significant differences for the 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 differed from the PE versions ($p < 0.01$). This was 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 dataset. This was a surprising result since both MT and the PE versions presented the same lexical density average score (see Table 4). This result revealed that, despite the similarity of the average lexical density between the MT and the PE versions, their distributions differed 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 dataset, the MT and the PE versions did not differ significantly in any of the features examined, suggesting less interference from the translators in the 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 2. However, we can see a statistically significant difference between the HT and the PE versions in the distribution of 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-editeuse hypothesis was confirmed for the TGOTT dataset 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 observed a convergence between them. Therefore, the post-editeuse hypothesis was confirmed partially for all other features. For the AW dataset, in contrast, post-editeuse hypothesis was not confirmed in any of the features examined, especially because we see significant differences between PE and MT output.

Regarding punctuation, we noticed from Tables 16 and 17 that, although it was not possible to confirm the post-editeuse hypothesis for punctuation feature as there were no statistically significant differences between the text types (all $p > 0.05$), the p -values obtained comparing the distributions of punctuation counts between the text types revealed that some distributions were more similar to each other than other distributions. In TGOTT, we observed that the distribution of the punctuation counts of the PE versions was 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 ($p = 0.21$). For AW, it was possible to observe the inverse pattern, that is, PE versions were 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 was more different in the TGOTT dataset ($p = 0.21$) than in AW as indicated by greater p -value (AW $p = 0.47$).

5. General Discussion and Conclusions

In the present study, we investigated the existence of post-editeuse 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.

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 used 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-editeuse phenomenon. Table 18 shows a summary of our findings.

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

Table 18. Manifestation of post-editease per feature.

Features	Post-Editease (HT vs. PE)	
	AW	TGOTT
Lexical Richness	Not confirmed	Not confirmed
Lexical Density	Not confirmed	Confirmed
Sentence count	Not confirmed	Not confirmed
Sentence length	Not confirmed	Not confirmed
Punctuation	Not confirmed	Not confirmed
Length Ratio	Not confirmed	Confirmed
Personal Pronoun	Not confirmed	Confirmed
Convergence	Confirmed	Partially confirmed

Regarding sentence count, the post-editease hypothesis was not confirmed for either dataset, as we found that the PE texts were similar to the HT texts. Finally, regarding punctuation, the qualitative analysis 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 noted that punctuation in PE tends to follow the MT punctuation more closely than the HT. However, we did not confirm the post-editease hypothesis for the punctuation feature as significant differences in punctuation counts were not found between the text types in any of the comparisons made.

Taking the results of simplification into consideration, our findings showed 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-editease hypothesis (RQ2 “*which translationese features (as described by Baker [8]) can also support the post-editease hypothesis?*”) our findings showed that, lexical richness, sentence length, sentence count, and punctuation might be good indicators of the existence of post-editease, but they are not good indicators of the existence of post-editease in our corpus.

As regards to the explicitation features, post-editease is confirmed for both features, i.e., length ratio and PPs ratio for the TGOTT dataset, but not for the AW dataset. 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-editease hypothesis, but there was a difference between text sub-genres.

Finally, the convergence feature confirmed post-editease for all features since we observed that PE variance scores are similar to MT variance in both datasets. Thus, convergence in our study is a good indicator of post-editease (RQ2).

Considering the results of all features together, we note that post-editease was not confirmed for most of the features within the AW dataset, but it was confirmed for more features in the TGOTT dataset. Nonetheless, our findings showed that the post-editease phenomenon was manifested on the surface of the post-edited texts as there were differences between those and HT versions. These differences were manifested in terms of the proximity or distance from the source and MT versions. While PE texts from the TGOTT dataset were closer to the MT output in a series of features, the features extracted from the HT texts were 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 was 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 emphasize action over language style are less challenging. We validated this assumption based on our observations that AW involved more edits than the TGOTT dataset, suggesting that the MT output can express the meaning of the source text more efficiently than for the AW. Moreover, we found a more visible pattern in terms of features for the TGOTT dataset when compared to the AW which, in turn, was unstable in terms of pattern manifestation. This allowed us to confirm our post-editease 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 were

that they were 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 of whether there are characteristics of PE literary texts that possibly make them less creative than HT texts remain open. Nevertheless, based on our results, we assumed that literary creativity in PE texts may be compromised, as shown by the Guerberoﬀ–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, meaning they would not be as experienced in effectively commanding the tone, author's style, and creativity when modifying the MT output to adapt into the 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. This unbalanced dataset could have biased the results for the feature convergence as we combined 9 PE texts, while for the source, HT, and MT, we computed the variance scores within the set of translated sentences of each translated version. Thus, a more balanced dataset with more human translations from the same text would 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-edited phenomenon.

Therefore, with further study in the literary genres and post-edited, 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.

Author Contributions: Conceptualization, N.R. and S.C.; methodology, N.R. and S.C.; validation, N.R. and S.C.; formal analysis, N.R. and S.C.; investigation, N.R. and S.C.; resources, N.R. and S.C.; data curation, N.R. and S.C.; writing—original draft preparation, N.R. and S.C.; writing—review and editing, N.R. and S.C.; visualization, N.R. and S.C.; project administration, N.R. and S.C.; funding acquisition, N.R. All authors have read and agreed to the published version of the manuscript.

Funding: ADAPT: the Science Foundation Ireland Research Centre for AI-Driven Digital Content Technology at Dublin City University, is funded by the Science Foundation Ireland through the SFI Research Centres Programme (Grant 13/RC/2106_P2). This project was partially funded by the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 843455.

Institutional Review Board Statement: The study was conducted according to the guidelines of the Declaration of Helsinki, and approved by the Ethics Committee of Dublin City University (protocol code: DCUREC/2019/110, date of approval: 21 June 2019).

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: Not applicable.

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 Skłodowska-Curie grant agreement No 843455 and 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 Centre Programme.

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

References

1. Ana Guerberof-Arenas and Antonio Toral. *The Impact of Post-Editing and Machine Translation on Creativity and Reading Experience. Translation Spaces*; John Benjamins: Amsterdam, The Netherlands, 2020; Volume 9, pp. 255–282.
2. Toral, A.; Wieling, M.; Way, A. Post-Editing Effort of a Novel with Statistical and Neural Machine Translation. *Front. Digit. Humanit.* **2018**, *5*, 9. [CrossRef]
3. Toral, A.; Way, A. What Level of Quality Can Neural Machine Translation Attain on Literary Text? In *Translation Quality Assessment: From Principles to Practice*; Moorkens, J., Castilho, S., Gaspari, F., Doherty, S., Eds.; Springer: Berlin, Germany, 2018; Volume 1, pp. 263–287.
4. Wu, Y.; Schuster, M.; Chen, Z.; Le, Q.V.; Norouzi, M.; Macherey, W.; Krikun, M.; Cao, Y.; Gao, Q.; Macherey, K.; et al. Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation. *arXiv* **2016**, arXiv:1609.08144.
5. Moorkens, J.; Toral, A.; Castilho, S.; Way, A. Translators' Perceptions of Literary Post-Editing Using Statistical and Neural Machine Translation. *Transl. Spaces* **2018**, *7*, 240–262. [CrossRef]
6. Daems, J.; de Clercq, O.; Macken, L. Translationese and post-edited: How comparable is comparable quality? *Linguist. Antwerp. New Ser.-Themes Transl. Stud.* **2017**, *16*, 89–103. [CrossRef]
7. Chesterman, A. Beyond Particular. In *Translation Universals. Do They Exist?* Mauranen, A., Kujamäki, P., Eds.; John Benjamins: Amsterdam, The Netherlands, 2004; Volume 28, pp. 33–49.
8. Baker, M. Corpus linguistics and translation studies: Implications and applications. In *Text and Technology*; Francis, G., Tognini-Bonelli, E., Eds.; John Sinclair, John Benjamins Publishing Company: Amsterdam, The Netherlands, 1993; pp. 233–252.
9. Baker, M. Chapter corpus-based translation studies: The challenges that lie ahead. In *Terminology, LSP and Translation: Studies in Language Engineering*; Juan, C.S., Ed.; John Benjamins Publishing Company: Amsterdam, The Netherlands, 1996; p. 175186.
10. Laviosa, S. *How Comparable Can Comparable Corpora Be?* John Benjamins Publishing Catalog.: Amsterdam, The Netherlands, 1996; Available online: <https://benjamins.com/catalog/target.9.2.05lav> (accessed on 20 January 2022).
11. Mauranen, A.; Kujamäki, P. (Eds.) *Translation Universals: Do They Exist?* John Benjamins: Philadelphia, PA, USA, 2004.
12. Gellerstam, M. Translationese in Swedish novels translated from English. In *Translation Studies in Scandinavia*; Wollin, L., Lindquist, H., Eds.; CWK Gleerup: Lund, Sweden, 1986; Volume 4, pp. 88–95.
13. Corpas, G.; Mitkov, P.R.; Pekar, V. Translation universals: Do they exist? Corpus-based NLP study of convergence and simplification. In Proceedings of the AMTA, Waikiki, Hawaii, USA, 21–25 October 2008.
14. Koppel, M.; Ordan, N. Translationese and its dialects. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, Portland, OR, USA, 19–24 June 2011; Association for Computational Linguistics: Portland, OR, USA, 2011; pp. 1318–1326.
15. Volansky, V.; Ordan, N.; Wintner, S. On the features of translationese. *Digit. Scholarsh. Humanit.* **2013**, *30*, 98–118. [CrossRef]
16. Baroni, M.; Bernardini, S. A new approach to the study of translationese: Machine-learning the difference between original and translated text. *Lit. Linguist. Comput.* **2006**, *21*, 259–274. [CrossRef]
17. Resende, N.C.A. Testing the validity of translation universals by employing comparable corpora and NLP techniques. In *Historical Corpora: Challenges and Perspectives*; Gippert, J., Gehrke, R., Eds.; Narr Verlag: Tübingen, Germany, 2015.
18. Johansson, S. Mens sana in corpore sano: On the role of corpora in linguistic research. *Eur. Engl. Messenger* **1995**, *4*, 19–25.
19. Laviosa, S. Core patterns of lexical use in a comparable corpus of English lexical prose. *Meta* **1998**, *43*, 557–570. [CrossRef]

20. Toury, G. The Porter Institute for Poetics and Semiotics. In *Search of a Theory of Translation*; Tel Aviv University: Tel Aviv, Israel, 1980.
21. Toral, A. Post-editsese: An exacerbated translationese. In *Proceedings of the Machine Translation Summit*, Dublin, Ireland, 19–23 August 2019.
22. Castilho, S.; Resende, N.C.A.; Mitkov, R. What Influences Post-editsese features? A preliminary study. In *Proceedings of the Second Workshop on Human-Informed Translation and Interpreting Technology (HiT-IT 2019)*, Varna, Bulgaria, 5–6 September 2019. Available online: <http://rgcl.wlv.ac.uk/wp-content/uploads/2019/11/HiT-IT2019-proceedings.pdf> (accessed on 20 January 2022).
23. Ilisei, I.; Inkpen, D.; Pastor, G.C.; Mitkov, R. Identification of translationese: A machine learning approach. In *Proceedings of the CICLing-2010: 11th International Conference on Computational Linguistics and Intelligent Text Processing*, Iași, Romania, 21–27 March 2010; Gelbukh, A.F., Ed.; *Lecture Notes in Computer Science*. Volume 6008, pp. 503–511.
24. Ilisei, I.; Inkpen, D. Translationese traits in Romanian newspapers: A machine learning approach. *Int. J. Comput. Linguist. Appl.* **2011**, *2*, 319–332.
25. Vanmassenhove, E.; Shterionov, D.; Way, A. Lost in translation: Loss and decay of linguistic richness in machine translation. In *Proceedings of the MT Summit XVII*, Dublin, Ireland, 19–23 August 2019.
26. Culo, O.; Nitzke, J. Patterns of terminological variation in post-editing and of cognate use in machine translation in contrast to human translation. In *Proceedings of the 19th Annual Conference of the European Association for Machine Translation*, Riga, Latvia, 30 May–1 June 2016; pp. 106–114. Available online: <https://www.aclweb.org/anthology/W16-3401> (accessed on 20 January 2022).
27. Groves, D.; Schmidtko, D. Identification and Analysis of Post-Editing Patterns for MT. In *Proceedings of the Machine Translation Summit XII: Commercial MT User Program*, Ottawa, OT, Canada, 26–30 August 2009.
28. Quirk, C.; Menenzes, A.; Cherry, C. Dependency Treelet Translation: Syntactically Informed Phrasal SMT. In *Proceedings of the 43th Annual Meeting of the Association for Computational Linguistics*, Ann Arbor, MI, USA, 25–30 June 2005; pp. 271–279.
29. Chomsky, N. *Lectures on Government and Binding: The Pisa Lectures, Reprint*, 7th ed.; Foris Publications, Mouton de Gruyter: Berlin, Germany; New York, NY, USA, 1993.
30. Tiedemann, J. Parallel data, tools and interfaces in OPUS. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC-2012)*, Istanbul, Turkey, 21–27 May 2012; European Languages Resources Association (ELRA): Istanbul, Turkey, 2012; pp. 2214–2218. Available online: http://www.lrec-conf.org/proceedings/lrec2012/pdf/463_Paper.pdf (accessed on 20 January 2022).
31. Aziz, W.; Castilho, S.; Specia, L. 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, 23–25 May 2012.
32. Snover, M.; Dorr, B.; Schwartz, R.; Micciulla, L.; Makhoul, J. A study of translation edit rate with targeted human annotation. In *Proceedings of the AMTA 2006 7th Conference of the Association for Machine Translation in the Americas*, Cambridge, MA, USA, 8–12 August 2006; pp. 223–231.