Current evidence of post-editese: differences between post-edited neural machine translation output and human translation revealed through human evaluation

Michael Farrell[1][0000-0002-7138-6639]

[1] IULM University, Milan, Italy
michael.farrell@iulm.it

Abstract. The experiment reported in this paper is a follow-up to one conducted in 2017/2018. The new experiment aimed to establish if the previously observed lexical impoverishment in machine translation post-editing (MTPE) has become more marked as technology has developed or if it has attenuated. This was done by focusing on two n-grams, which had been previously identified as MT markers, i.e., n-grams that give rise to translation solutions that occur with a higher frequency in MTPE than is natural in HT. The new findings suggest that lexical impoverishment in the two short texts examined has indeed diminished with DeepL Translator.
The new experiment also considered possible syntactic differences, namely the number of text segments in the target text. However no significant difference was observed.
The participants were asked to complete a short questionnaire on how they went about their tasks. It emerged that it was helpful to consult the source language text while post-editing, and the original unedited raw output while self-revising, suggesting that monolingual MTPE of the two chosen texts would have been unwise.
Despite not being given specific guidelines, the productivity of the post-editors increased. If the ISO 18587:2017 recommendation of using as much of the MT output as possible had been strictly followed, the MTPE would have been easier to distinguish from HT. If this can be taken to be generally true, it suggests that it is neither necessary nor advisable to follow this recommendation when lexical diversity is crucial for making the translation more engaging.

Keywords: post-editese, machine translation post-editing, neural machine translation, human translation, human machine translation output evaluation, DeepL Translator, Google Translate.

1 Introduction

Several researchers have reported the existence of features of post-edited machine translation output (MTPE) that distinguish it from human translated text (HT), defined as post-editese. By way of example, Castilho et al. looked at literary texts Google-

Proceedings of the International Conference HiT-IT 2023, pages 52–63,
7-9 July 2023, Naples, Italy
https://doi.org/10.26615/issn.2683-0078.2023_005

translated from English into Brazilian Portuguese and found evidence for post-editese in one of the two texts examined [1]; Volkart et al. found that post-edited machine translation was not only lexically poorer than human translation, but also less dense and less varied in terms of translation solutions [12]; Toral found that MTPE was simpler and more normalized and had a higher degree of interference from the source language than HT [11]; and Castilho et al. found evidence of post-editese features, especially in light post-edited texts and in certain domains [2]. By contrast, on the other hand, Daems et al. found no proof of the existence of post-editese, either perceived or measurable [5].

The experiment described in this paper is a follow-up to an experiment carried out for two consecutive years (2017 and 2018) with postgraduate university students of translation (IULM University, Milan). In the previous experiment, half of the students did an unaided human translation (HT) and the other half post-edited machine translation output (MTPE). Comparison of the texts produced in 2017/18 showed that certain turns of phrase, expressions and choices of words occurred with greater frequency in MTPE than in HT (MT markers), making it theoretically possible to design tests to tell them apart. To verify this, the author successfully carried out one such test in 2018 on a small group of six professional translators [6].

The primary aim of the new experiment described in this paper was *not* to show that MTPE generally results in an increase in productivity, which is well documented elsewhere, but to see if it is still possible to detect MT markers in MTPE, despite the advances in MT technology since 2018, and if it is also still possible to distinguish MTPE from HT simply by comparing the number of these markers found in each kind of text. The students were also asked to provide various details of how they went about their tasks.

2 Design and methods

Two short extracts from English-language Wikipedia entries were taken for the experiment: one about Slovakia (262 words) and one about the Euromaidan civil unrest in Ukraine (263). Besides being the same genre as used in the previous experiment, Wikipedia articles were chosen since they are likely to be less challenging for an MT engine than classic works of literature but more problematic than the boilerplate-style texts which are often considered to lend themselves best to machine translation.

The first text was selected since it contained the bigram *there are* four times, and the other because it contained the monogram *people* (used as the plural of the word person and not as the singular noun meaning populace) six times. These were the first two short extracts that contained at least four examples of the chosen n-gram in the space of approximately 250 words that the author came across while searching randomly through Wikipedia.

These two n-grams had been identified in the 2017/18 experiment as among the best MT markers, that is n-grams which were translated with a highly statistically significantly greater number of correct translation solutions in HT than in MTPE, and *there*

are was the specific bigram used in the above mentioned successful test to distinguish HT from MTPE carried out in 2018 on a small group of six professional translators [6].

Forty-two postgraduate students of translation (IULM University, Milan) were divided into two groups and worked from English into Italian. Group A (21 students) translated the Slovakia text and post-edited the machine-translated Ukraine text, and group B (21 students) translated the Ukraine text and post-edited the machine-translated Slovakia text.

The pay-for version of DeepL Translator was chosen as the MT engine for this experiment for two main reasons:

1. The week before, the students had machine translated several short extracts (250 words approx.) from Wikipedia entries using different free online MT engines to compare the quality of their raw output, and an overwhelming majority had judged DeepL Translator to be the best for this genre (87%).
2. In a recent survey among professional translators [7], the MT engine most used by the respondents who declared that they use MT at some point in their workflow turned out to be DeepL Translator (183 users). Its nearest rival Google Translate was only chosen by just over half that number (93 users). The majority of DeepL Translator users surveyed (102) stated that they use the pay-for version.

The students were deliberately not given any post-editing guidelines but were told that they should transform the machine translated output into a text of the same quality as a human translation for publication (full post-editing). Both the post-editors and the translators were told that the task was urgent and should be completed in the shortest possible time without compromising on quality. They were also told that the objective of the experiment was to compare the average time taken for each task. They were not told beforehand that their final texts would be analysed for traces of post-editese. The latter was in reality the primary reason for the experiment.

The students were allowed to use any dictionaries and reference material they liked, including Wikipedia itself, and even to ask for advice on individual problems from friends and colleagues not involved themselves in the experiment in a way that would not disturb the others, for example via WhatsApp. The intention was to recreate something as near as possible to normal working conditions. They were however instructed not to use MT engines in any way to prevent the translators from turning their task into a second post-editing assignment. This unfortunately goes against the aim of recreating real working conditions since it was found in the aforementioned recent survey that just over 69% of professional translators use MT in some way during their workflow, but not necessarily to translate the whole text for subsequent post-editing [7].

The files for translation and post-editing were provided as word processor documents and the translators and post-editors worked in Microsoft Word. The task was presented in this way so that the students were not influenced by the segmentation imposed by CAT or post-editing tools. It has in fact been observed that machine translation output normally has the same number of segments as the original language text, whereas translators who are working without a CAT tool sometimes organize the translated text into a different number of sentences. This can be verified by taking the first 26 sentences of Chapter 3 of *The Adventures of Pinocchio* by Carlo Collodi [3],

machine-translating them with Google Translate and comparing the output with the 1926 translation by Carol Della Chiesa [4]. The raw MT output is also organized into 26 sentences, whereas Della Chiesa's translation has 28. It could be argued that it is rather obvious that there will be the same number of segments in a machine-translated text as in the original, but some machine translation engines today work at larger-than-segment level, notably ModernMT and possibly also DeepL Translator [7]. To make comparisons in this experiment, it was decided to count the number of segments created using the default segmentation rules of the two most used CAT tools according to the previously mentioned recent survey, Trados Studio and memoQ [7].

The students were also asked to complete a short questionnaire after they delivered their files to report some details of how they went about their tasks.

Unfortunately, one student misunderstood the instructions and translated and post-edited the same text; his work was discarded since the results of one of the two tasks were probably influenced by having done the other. Another student was not a native Italian speaker; her work was discarded since her translation choices may have been affected by her native language. Yet another student delivered a damaged file; it was however possible to evaluate the undamaged one. And one student did not deliver their files at all. In the end, 20 post-edited Ukraine texts, the same number of post-edited Slovakia texts, 19 translated Ukraine texts and 18 translated Slovakia texts were analysed.

Most of the variables measured in this paper are non-numeric, non-parametric, categorical variables which can only take on a limited number of values. For this reason, when possible, the widely used chi-square ($\chi 2$) test was chosen for the statistical analyses. The significance level was set to .05, as per convention, to ensure a 95% confidence level, and the online chi-square test calculator provided by Jeremy Stangroom was used [9]. The results are reported in the format required by the American Psychological Association (APA) [10].

3 Results and discussion

3.1 Time comparison

As expected, and as is commonly reported, it took less time on average to post-edit the MT output than to translate the same text from scratch (Table 1).

Table 1. Translation and post-editing times in minutes and seconds.

Text	Length of text (words)	Translation time (mean ± SD)	Post-editing time (mean ± SD)	Productivity increase (%)
Slovakia	262	37:34 ± 5:28	22:59 ± 8:06	39.50%
Ukraine	263	38:47 ± 10:00	20:01 ± 7:18	47.99%

The productivity increase was calculated as the translation time minus the post-editing time, divided by the translation time. This was then multiplied by one hundred to obtain a percentage.

3.2 Slovakia text n-gram

The source text contained the bigram *there are* four times. DeepL Translator translated the bigram with *ci sono* three times and *vi sono*[1] once. Table 2 shows the translation solutions the translators and post-editors chose. The number shown is the overall number of occurrences of the n-gram indicated in all the texts of the given type (18 translated texts, 1 raw output and 20 post-edited texts). Since the number of texts in each category is different, the overall percentage number of occurrences should be considered when making comparisons.

Table 2. Italian translation solutions in HT, raw MT output and MTPE for *there are*

	Translation		Raw output		MTPE	
abbonda	1	1.39%				
ci sono	36	50.00%	3	75.00%	52	65.00%
è caratterizata da	1	1.39%				
è possibile ammirare	1	1.39%				
è possibile trovare	1	1.39%				
è ricca di					1	1.25%
esistono	4	5.56%				
presenta	1	1.39%			1	1.25%
si possono trovare	7	9.72%				
si ritrovano					1	1.25%
si trovano	6	8.33%			1	1.25%
sono presenti	6	8.33%			5	6.25%
troviamo	1	1.39%				
vengono offerti	1	1.39%				
vi sono	5	6.94%	1	25.00%	19	23.75%
vi trovano	1	1.39%				
Totals	72	100%	4	100%	80	100%

From Table 2, it is evident that there is less variety in the solutions the post-editors chose since they are clearly primed by the raw output. This difference is statistically significant, as can be verified using the contingency table below (Table 3).

[1] Equivalent to *ci sono* but higher in register.

Table 3. Lexical variety contingency table for *there are*.

	Translation	MTPE
ci sono	36	52
vi sono	5	19
Other n-gram	31	9
(χ2 (2, N = 152) = 22.82, p < .05)		

The fact the raw output already contained two different translation solutions was unexpected (Table 2). Indeed, the same text translated by Google Translate contained the same solution all four times (*ci sono*). The presence of alternative translation solutions in DeepL Translator raw output is discussed in more detail below under *Degree of naturally occurring lexical variety in DeepL Translator raw output*.

In one of the previous experiments reported in 2018, a 273-word text containing five occurrences of *there are* was given to three professional translators for translation, and Google-translated and given to another three for full post-editing. None of the translators translated *there are* with *ci sono*, whereas all the post-editors left at least one occurrence of *ci sono*. Therefore, if the 2018 texts are split into two sets on the basis of how many times *ci sono* was chosen as the translation solution, it is possible to identify the MTPE with 100% accuracy. The same method of splitting the 2023 texts into two sets according to the number of occurrences of *ci sono* results in five misattributed texts. In other words, the translations are identifiable with 13/18 = 72.22% accuracy and the MTPE, with 15/20 = 75% accuracy.

3.3 Ukraine text n-gram

The source text contained the monogram *people*, used as the plural of the word *person*, six times. The raw output from DeepL Translator contained the monogram *persone* seven times since a demonstrative pronoun plus adjective in one of the source text sentences (*those killed*) was resolved into a noun plus adjective (*persone uccise*). Seven of the translators chose to do the same (Table 4). The number shown in Table 4 is the overall number of occurrences of the n-gram indicated in all the texts of the given type (19 translated texts, 1 raw output and 20 post-edited texts). Since the number of texts in each category is different, the overall percentage number of occurrences should be considered when making comparisons.

Table 4. Italian translation solutions in HT, raw MT output and MTPE for *people*

	Translation		Raw output		MTPE	
17	1	0.75%				
cittadini	1	0.75%				
coloro che erano stati uccisi					1	0.71%
coloro che furono uccisi	1	0.75%				
coloro che sono stati uccisi	1	0.75%				
folla	4	3.01%				
gente	3	2.26%			1	0.71%
individui					2	1.43%
Maidan	1	0.75%				
manifestanti					1	0.71%
morti	1	0.75%				
persone	104	79.20%	6	85.71%	114	81.43%
persone uccise	7	5.26%	1	14.29%	16	11.43%
presenti					1	0.71%
protestanti					1	0.71%
tutti coloro che erano stati uccisi	2	1,50%				
uccisioni					1	0.71%
vittime	1	0.75%			1	0.71%
vittime uccise	6	4.51%				
Totals	133	100%	7	100%	140	100%

From Table 4, it is again evident that there is less variety in the solutions the post-editors chose, although perhaps a little less so. However, the difference is again statistically significant, as can be verified using the contingency table below (Table 5).

Table 5. Lexical variety contingency table for *people*

	Translation	MTPE
persone	104	114
persone uccise	7	16
Other n-gram	22	10
($\chi 2$ (2, N = 273) = 8.31, p < .05).		

The method described above of dividing the texts into two sets according to the number of occurrences of *persone* results in six misattributed texts. In other words, the human translations and MTPE may be identified with 14/20 = 70.00% accuracy.

3.4 Degree of naturally occurring lexical variety in DeepL Translator raw output

To measure the degree of lexical variety naturally produced by DeepL Translator in its raw output, two longer texts were machine-translated, containing a number of MT markers equal to the number of students involved times the number of MT markers

found in each of the two shorter texts translated/post-edited in the main experiment (18 x 4 = 72 for *there are*; 19 x 6 = 114 for *people*). These longer texts were put together by taking whole paragraphs rich in the n-gram concerned from several Wikipedia articles and pasting them all into a single document. The raw MT output from this experiment was found to be less lexically impoverished than in the equivalent experiment reported in 2018, at least as regards the two n-grams studied. In the case of the first MT marker considered (*there are*), the number of translation solutions in the raw output from DeepL Translator (8) is quite a lot smaller than the number used by the human translators (14), and the distribution of the HT solutions is more even. However, the most chosen solution (*ci sono*[2]), had exactly the same frequency in the HT and the raw output (50%). In the similar experiment reported in 2018, DeepL Translator had translated *there are* with *ci sono* 90% of the time.

Regarding the second MT marker (*people*, as the plural of *person*), the translation solutions in the raw output (11) were only slightly less numerous than those chosen by the human translators (13) but the solutions themselves were often quite different.

Due to the presence of a lot of very low frequency translation solutions and translation solutions occurring only in the HT and not in the raw output and vice versa (zero values), meaningful chi-square statistical analysis is unfortunately not possible.

By way of comparison, the same longer texts were also fed to Google Translate, whose raw output showed much clearer signs of lexical impoverishment (only 3 solutions for the first MT marker and 7 for the second).

3.5 Task questionnaire

The students completed a short questionnaire after they delivered their files. They were first of all asked how they had done the translation. The majority wrote their translations in a new Microsoft Word file (Table 6).

Table 6. How the translation was done in Microsoft Word

	Number of replies
New empty Microsoft Word file	26
Overwrite original text	12
Create two column table	3
Write underneath, then delete original text	1

They were then asked what reference material they had used while translating or post-editing (Table 7).

[2] Variants required for grammatical reasons, such as *ci siano* (subjunctive tense), were considered to be the same solution.

Table 7. Use of reference material while translating or post-editing, multiple answers were allowed.

	Translating	Post-editing
Online dictionaries, encyclopaedias or websites	42	40
Asked a colleague for help	4	1
Physical, printed reference material	0	0
No reference material	0	2

The results show quite clearly that print dictionaries are a thing of the past.

The students were instructed not to use MT engines in any way to prevent the translators from turning their task into a second post-editing assignment. It is clear from the tables above that the same kinds of materials were used for both processes. Two post-editors did not refer to any external reference material.

The students were asked to assess how useful it was to be able to refer to the source language text while post-editing (8.00 ± 1.89 SD points out of 10) and to the original unedited raw output during the self-revision of their post-editing (6.12 ± 3.05 SD points out of 10). These results substantially confirm the ISO 18587:2017 definition of post-editing as involving three texts: the source text, the MT output and the final target text [8]. They also suggest that monolingual post-editing would have been ill advised in the case of the texts chosen.

Another question the students were asked was if they would have used MT in some way during their task if it had been allowed (Table 8).

Table 8. Number of translators and post-editors who would have used MT if it had been allowed

	Number of replies
Never	5
Only during the post-editing	0
Only during the translation	24
Both during the post-editing and the translation	13

3.6 Syntactic differences

3.6.1 Slovakia text segmentation

There were nine segments in the original text and in the machine translated text before post-editing. Table 9 shows the number of translators and post-editors who either split or joined segments at least once.

Table 9. Number of translators and post-editors who joined or split segments in the Slovakia text

Segments	N. translation (18)	N. post-editing (20)
Split/join	9	7
No split/join	9	13
(χ2 (1, N = 38) = 8.74, p < .05).		

The difference is not statistically significant. So, the translators and post-editors felt equally free to split/join segments.

3.6.2 Ukraine text segmentation

There were fifteen segments in the original text and in the machine translated text before post-editing. Table 10 shows the number of translators and post-editors who either split or joined segments at least once.

Table 10. Number of translators and post-editors who joined or split segments in the Ukraine text

Segments	n. translation (18)	n. post-editing (20)
Split/join	6	4
No split/join	14	16
(χ2 (1, N = 40) = 0.53, p < .05).		

The difference is again not statistically significant. So, the translators and post-editors felt equally free to split/join segments.

4 Conclusion

These conclusions are drawn on the basis of two short texts of only one genre, which limits the generality of the findings to some extent. This limitation is however inevitable since the experiment was carried out as part of an undergraduate degree course and only a limited amount of time could be devoted to it.

In the case of the particular texts used in this experiment, the priming received from the raw output led to MTPE that is distinguishable from HT with a success rate of between 70 and 75%, which is however down from the 100% success rate observed in the 2017/2018 experiment. On the basis of these results, we are led to conclude that the lexical impoverishment phenomenon is indeed attenuating with DeepL Translator. It is however apparent that the results would have been different with Google Translate, which produces raw output with clearer lexical impoverishment, as was mentioned in section 3.4 above.

Despite not being given any particular post-editing guidelines, there was an increase in productivity of between 39.50 and 47.99%. Some of the translators and post-editors chose exactly the same translation solutions for the n-grams studied as were found in

the raw output in precisely the same places. Therefore, the solutions in the raw output are acceptable. Consequently, if the post-editors had strictly applied the ISO 18587:2017 post-editing recommendation to use as much of the MT output as possible [8], the post-editors would not have altered these solutions making the MTPE even easier to distinguish from HT. If we can generalize these results, this fact, together with the increase in productivity, suggests that, in the case of texts where lexical uniformity would make the translation less interesting to read and less intellectually stimulating, such as in the fields of marketing, advertising, literature, journalism, education, entertainment, and creative writing in general, it is neither necessary nor advisable to apply this ISO 18587:2017 recommendation.

Another way of avoiding lexical impoverishment may be to avoid MTPE entirely and use MT as a tool during the translation process, for example in one of the various ways that emerged from the previously mentioned survey among professional translators, such as *for inspiration* or *as a dictionary* [7]. However, this would almost certainly not lead to anything like the increase in productivity achieved with MTPE.

The students found it useful to refer to the source language text during the post-editing and to the original unedited raw output during the self-revision of their post-editing. This suggests that monolingual post-editing in the case of the texts chosen would have been ill advised.

No significant difference was found in the number of segments in the target texts. Evidently, the translators and post-editors felt equally free to split and join segments during their task. Obviously, since some of the post-editors chose not to alter the segmentation of the raw output and therefore found it acceptable, if the previously mentioned ISO 18587:2017 recommendation [8] had been strictly applied, none of the post-editors would have changed the segmentation thus making MTPE more distinguishable from HT syntactically. However, it would be interesting to repeat this experiment asking the participants to use a CAT or post-editing tool to see if they feel equally empowered to join and split segments.

References

1. Castilho, S., Resende, N. Post-Editese in Literary Translations. In Machine Translation for Conquering Language Barriers, Special Issue (2022).
2. Castilho, S., Resende, N., Mitkov, R: What Influences the Features of Post-editese? A Preliminary Study. In: Proceedings of the Human-Informed Translation and Interpreting Technology Workshop (HiT-IT 2019), pp. 19–27. Incoma Ltd., Shoumen, Bulgaria.. (2019).
3. Collodi, C. Le avventure di Pinocchio. Storia di un burattino. Libreria Editrice Felice Paggi (1883).
4. Collodi, C. The Adventures of Pinocchio (Translation by Carol Della Chiesa). The Macmillan Company (1926).
5. Daems, J., De Clercq, O., Macken, L.: Translationese and post-editese: How comparable is comparable quality? Linguistica Antverpiensia New Series - Themes in Translation Studies 16:89–103 (2017).
6. Farrell, M.: Machine Translation Markers in Post-Edited Machine Translation Output. In: Proceedings of the 40th Conference Translating and the Computer, pp, 50–59. AsLing: The International Association for Advancement in Language Technology (2018).

7. Farrell, M: Do translators use machine translation and if so, how? Results of a survey held among professional translators. Presented at the 44th Conference Translating and the Computer. Preprint of peer-reviewed paper awaiting publication, DOI:10.13140/RG.2.2.33996.69768, (2022).
8. International Organization for Standardization. ISO 18587:2017: Translation services – Post-editing of machine translation output – Requirements (2017).
9. Stangroom, J.: Chi-Square Test Calculator, www.socscistatistics.com/tests/chisquare2/default2.aspx, last accessed 2023/04/23.
10. Stangroom, J.: How to Report a Chi-Square Test Result (APA), www.socscistatistics.com/tutorials/chisquare/default.aspx, last accessed 2023/04/23.
11. Toral, A.: Post-editese: an Exacerbated Translationese. In: Proceedings of Machine Translation Summit XVII: Research Track, pp. 273–281. European Association for Machine Translation, Dublin, Ireland (2019).
12. Volkart, L., Bouillon, P: Studying Post-Editese in a Professional Context: A Pilot Study. In: Proceedings of the 23rd Annual Conference of the European Association for Machine Translation, pp. 71–79. European Association for Machine Translation, Ghent, Belgium (2022)