Nemergut, Matúš. 2024. Machine Translation Quality Based on TER Analysis from English into Slovak. In: L10N Journal 2(3), pp. 60–86.

Machine Translation Quality Based on TER Analysis from English into Slovak

Matúš Nemergut

Comenius University in Bratislava

nemergut.matus@gmail.com

Abstract

Translators are facing increased demand for their services and are increasingly required to utilize tools that can help them save time and increase their efficiency. Neural machine translation (NMT) has become the leading technology in the translation industry, and its utilization promises efficiency gains. However, this is not as straightforward as it may seem, and actual efficiency gains depend on a number of variables: the quality of the NMT output, the translator's skills, time, and the effort the translator expends on post-editing. This paper aims to analyze the number of edits required for the NMT output to meet quality requirements and determining the acceptability threshold of the neural machine translation output for post-editing based on this number. From a methodological perspective, the study uses TER, an automatic machine translation evaluation metric calculating the smallest edit distance required, to assess the number of edits needed. By analyzing samples from two experiments, it was found that with the TER score between 39% and 42.5%, i.e., when 39-42.5% of the machine translation output needs editing, post-editing ceases to be beneficial, and it is more efficient for the machine translation output with such score to be translated from scratch.

Keywords: post-editing, automatic machine translation evaluation metrics, TER, acceptability threshold for post-editing

1 Introduction

Modern technological advances are influencing and changing the set of demands users have for technology. It is expected to be innovative, fast, high quality, and to get even more innovative, faster, and of higher quality with every update. In translation technology, the past three decades have brought about some significant developments: statistical machine translation (SMT), computer-assisted translation (CAT) tools, and, most recently, neural machine translation (NMT), which, since its introduction in 2016, has been the most dominant among machine translation (MT) paradigms (Rothwell et al. 2023). Munková et al. (2019, in: Hudecová et al., 2021) note that the market has also

changed. Over the last decade, demand in the translation services sector has grown so significantly that it has far outstripped its supply. The increased demand for translation services places an invisible burden on translators to increase their efficiency and thus meet the demands of the market. The use of NMT can bring about efficiency increases, but the actual increase is dependent upon the quality of the output as well as the time and effort required to edit it. This work aims to investigate the effort required to edit an NMT output and to determine its acceptability threshold for post-editing based on TER analysis.

2 Machine Translation Post-Editing (MTPE)

Hudecová et al. (2021) note that the use of machine translation is becoming commonplace today, not only in the sphere of providing language services but also among non-professionals. The significant increase in demand in the translation industry over the last decade, which exceeded its supply (ibid.), underlines the fact that the question of machine translation is shifting from whether to use it to how best to use it, according to Koehn (2020, in: Hudecová et al. 2021). Hudecová et al. (2021) continue: "The suitability of utilizing machine translation is relative" (p. 193), and if we want it to be of high quality, accuracy, and fidelity, it needs to undergo further editing, i.e., *post-editing*.

Vieira (2019) explains that post-editing as an activity has been carried out since the beginning of the existence of MT technologies, but only in recent years it has evolved significantly as a service, practice, and research topic. In the early days, post-editing was seen as a step towards closing the gap between MT's purpose — a fully automatic translation of high quality — and its far insufficient output; for a long time, a model prevailed in which people supported the machine rather than the other way around. Vieira adds that there has been a significant improvement in MT integration in the field of professional translation in recent years, especially regarding CAT tools, changing the model of human-supported MT to computer/machine-supported human translation. However, he also sees the introduction of many technologies into CAT tools to have entailed some blurring of the boundaries between technology and support in the translation process, resulting not only in terminological inconstancy but also in ambiguity in post-editing procedures and tasks.

2.1 Error typology

The primary paradigm in the MTPE process is *static*—that is, MT is generated first and then it is statically edited in a separate step (Vieira, 2019). During editing, the post-editor works with three texts: (i) the source text, (ii) the MT output that is not

processed, and (iii) the MT output that is post-edited, i.e., the target text (Pavlíková 2022).

Situating MTPE within the models of the translation process, Absolon (2018) notes that MTPE is a combination of plain translation and revision (dual). This is due to the nature of errors, which require the post-editor to be flexible and able to shift from a process to process. After reading the MT output, the post-editor must assess its quality to know whether to edit the output or re-translate it. The ability to make quick decisions is therefore key — the work of a post-editor is always about deciding whether to continue with a given process or shift to another one.

The most common actions in MTPE are related to correcting punctuation, word order, number and gender correspondence, or incorrectly translated expressions (Pavlíková, 2022).

For the alternative paradigm in the MTPE process, Vieira (2019) uses the term *interactive*: there is interaction between the translator and the MT system during the production of the target text. He explains that such interaction may look like MT is completing or predicting the text of the human translation as it is being written or reciprocally reacting to and learning from the translator's edits on the fly and adds that while no tendency for post-editor's actions to speed up has been observed in static MTPE, there is such a tendency in interactive MTPE. At the same time, however, he notes that most productivity-focused research only marginally examines the quality of MTPE products and that the results are no longer clear-cut there, because while some research speaks of increased quality of interactive MTPE over static MTPE, other speaks otherwise.

2.2 MTPE vs. From-Scratch Translation

As such, MTPE research only addresses the issue of MTPE product quality in a complementary way in the context of comparing the productivity of from-scratch translation with MTPE (Vieira, 2019). Despite some negative perception of MT in the marketplace (Pavlíková 2022), Vieira (2019) writes about Screen's 2019 research (in: Vieira 2019), comparing a from-scratch translation with a post-edited translation, which concludes that the products of both procedures are largely equivalent; thus, along with other research, supporting the use of MT in professional practice. He also notes that with the advent of NMT, which is currently the cutting-edge MT technology, come the challenges of its post-editing: the higher fluency of the outputs of these systems makes identification of errors and their correction in mono- and bilingual texts more difficult. In a study comparing post-editing of phrase-based SMT and NMT, Yamada (2019, in: Vieira 2019) found that although the output of NMT contained fewer errors and the result of its post-editing was of higher quality, translation students achieved a lower error correction rate during post-editing but put more effort into it.

Vieira adds that research by Jia et al. (2019, in: Vieira 2019) concludes that, depending on the genre of the text, NMT post-editing may require less effort; with no difference in the quality of the target text between NMT post-editing and from-scratch translation, post-editing proved faster only in specialized texts.

Hudecová et al. note that "post-editing represents a specific skill" (2021, p. 195) and it may not be the case that a professional translator is also good at post-editing and vice versa (ibid.). Vieira (2019) recommends post-editing to be preferably done by professional translators, as those less experienced will hardly be aided by NMT in improving their performance and reminds that forming a proper understanding of the role of MT in professional translation is the task of translation study programs.

MTPE also raises the question of how to measure the efforts made by the post-editor. According to Koponen (2016, in: Hudecová et al. 2021), 3 indicators were defined for this purpose:

- *temporal indicator;*
- *technical indicator*—measured using automatic metrics;
- *cognitive indicator*—i.e., the efforts that the post-editor perceives to have made.

Although MTPE as an activity is in some cases demonstrably more productive, there is no known way by which it would be possible to determine "where post-editing is worthwhile and where it is not" (Hudecová et al. 2021, p. 195).

2.3 Translation Edit Rate (TER) and Translation Efforts

As mentioned above, three indicators have been defined to measure translation efforts, namely temporal, technical, and cognitive.

The aspect of time is prominent today, especially in the commercial sphere — do Carmo (2020), drawing on ISO 18587:2017, writes about time as one of the key determinants of MTPE. And since it is relatively easy to measure, it is at the center of much of the translation effort measurements. It should not, however, be left there alone. Since MTPE involves revision of the text, it is important to understand the effort required for the changes made. This is addressed by the technical indicator. When referred to, it can be a measurement of the actions performed with the mouse and keyboard on the one hand. These are recorded using various tools. At the same time, the technical indicator is the edit distance that needs to be performed on the hypothesis. That's where TER is utilized, and TER scores are calculated for these purposes. The lower the TER score, the less effort is required to post-edit a given MT output. For example, if the score is 0.4, i.e., 40%, it means that 40% of the MT output had to be changed to make it satisfactory. The third indicator is cognitive, and the most difficult to measure. It involves cognitive processes that cannot be directly seen or measured because they take place in the brain. These include reading, understanding and comparing the text, or following post-

editing guidelines. For instance, measuring the length of eye fixation or pupil dilation is one way of measuring the cognitive effort expended in MTPE. According to some, MTPE generates a greater cognitive load than translation itself; thus, it is important to take the cognitive dimension of MTPE into account when measuring translation efforts (O'Brien, 2022).

The methodological focus of this work is the evaluation of post-editing effort based on TER, therefore, the measurement of cognitive effort is not within the scope of this work.

3 Similar research

Given the complexity of examining translation efforts, one can encounter a variety of methods used in practice. The intention is to determine the threshold of usability and efficiency of MTPE in relation to from-scratch translation through TER analysis.

Guerrero (2020) works with the hypothesis that 50% edit distance is too high as an MT output acceptability indicator. Moreover, the results of her research show that at an edit distance between 30% and 40%, on a scale[3] of 1–4, MT outputs are already mostly rated as 2 by professional post-editors. Based on the comments of these raters, 2 is closer to the unacceptability of the MT output.

Also based on interviews with R. Tihlárik (2023), who at the time had been the director of a translation and localization services provider for 26 years, it can be stated that the current trend in the localization industry is for clients to consider the TER score of 50% as the limit; this means that the translator has to edit up to half of the entire MT output during MTPE, which in many cases can be more laborious than translating the text from scratch, but they get paid an MTPE rate that is partial to the rate for from-scratch translation.

Research conducted by e.g. Gueberof (2008, 2012), Guerrero (2003), or O'Brien (2006) on SMT has already shown a positive effect of MTPE on translator's productivity compared to translating from scratch or editing so-called "fuzzy matches"[4] and, in some cases, also on the quality of the final product (Temizöz 2013; O'Brien 2006).

In order to build on these conclusions today, it is necessary to consider the research findings of Koponen (2016), in which she emphasizes the existence of a relationship between the effort expended in MTPE and the characteristics of the source text or the MT error rate, the influence of the type of editing on the overall effort expended in MTPE, or the varying speed of post-editors themselves. She further talks about the trend to perceive the editing effort of longer sentences as higher, even if the number of

[3] A Likert scale was used, i.e., a scale used to measure behavior, opinions, and attitudes (scribbr.com, 2020).
[4] *Fuzzy match* represents a condition where a segment of the source text is partially identical to another already translated segment in the translation memory (thelanguageoflocalization.com, 2018).

edits is relatively low, or the effect of sentence length on the time required to post-edit it.

Stefaniak (2020) notes that although it might seem obvious that MT output with lower TER score will also require less time to post-edit and vice versa, this is not necessarily the case. This is a complex issue, and the research results are not sufficient to draw conclusions.

For example, Krings (2001, in: Temizöz 2013) has observed that the effort expended in post-editing a medium-quality MT output is higher than the effort expended in post-editing a low-quality MT output; he's attributed this to the fact that a medium-quality MT output contains a large number of elements that need to be extensively compared between the source text, the target text, and the MT output. He's compared post-editing of a low-quality MT output to the process of standard human translation, in which only the source and target texts are worked with, resulting in a lower cognitive load (Temizöz 2013).

It is important to note that the aforementioned research has been carried out on SMT systems; but as has already been mentioned, NMT systems produce smoother outputs, and thus, based on Yamada's observation (2019, in: Vieira 2019) that "although the output of NMT contained fewer errors and the result of its post-editing was of higher quality, translation students achieved a lower error correction rate during post-editing, but put more effort into it", it can be inferred that Krings' results (2001, in: Temizöz, 2013) can be applied to NMT as well.

Stefaniak's (2020) research on the English-Polish language pair has also shown no clear correlation between the TER score and the time required for post-editing a given MT output. These results are in line with the findings of Gaspari et al. (2014, in: Stefaniak 2020), who have noted that if there is a correlation between TER, but also BLEU or METEOR, and the time required for MTPE, it is only very weak (Stefaniak 2020).

However, in the same research, Krings (2001, in: Temizöz 2013) also found that MT outputs with higher quality assurance (QA) scores were post-edited faster (Temizöz 2013).

Additionally, O'Brien (2011), on the basis of her research, also on SMT, has preliminarily concluded that there is a correlation between TER and MTPE productivity and that TER can be a good indicator of MTPE speed for a set of segments.

Thus, Stefaniak's (2020) research has indeed revealed a correlation: the correlation between the TER score and productivity gains. In the research conducted, with a mean TER score of 0.39 (39%) and a median TER score of 0.375 (37.5%), the average MTPE speed vs. the average from-scratch translation speed was as follows: MTPE = 0.325 words/second (the median 0.295 words/second); from-scratch-translation = 0.215 words/second (the median 0.205 words/second). These results are in line with the

research of de Gibert Bonet (2018, in: Stefaniak 2020), who has set the TER score threshold up to which productivity gains occur at 0.33 (i.e., 33%) for the Spanish-English language pairs (Stefaniak 2020).

In their research on the English-Spanish language pairs, Parra Escartín and Arcedillo (2015) have reported an increase in MTPE productivity relative to from-scratch translation for the TER score ≤ 0.21 (i.e., 21%). However, for this research, it is important to note that only two translators were observed, the test set also contained 75%–100% fuzzy matches from a translation memory, and the MT tool they used to generate the machine translation was their home-grown tool that they had used and fine-tuned for three years (Parra Escartín and Arcedillo, 2015).

4 Methodology, research questions, and research methods

As mentioned previously, measuring the effort expended in MTPE consists of several aspects. As the methodological focus of this work, the technical aspect examined through TER analysis was chosen, and the goal was to determine the threshold of the acceptability of MT output for post-editing. In other words, the research tries to determine:

- what is the TER score at which it is more efficient to translate the source text from scratch than to post-edit its machine-translated form.

Zhechev (2014) notes that when comparing from-scratch translation and translation using MT, it is important to set a baseline. This is not easy if the translators do not translate and post-edit the same segment.

In the experiments, MA students majoring in philology with a specialization in translation and interpreting at Comenius University did not translate and post-edit the same segments because there would have been no way of solving the problem of them remembering what had already been translated/post-edited. Therefore, each of them translated one part of the text and post-edited the other. This eliminated the differences that might arise from the different levels of experience of the post-editors with MTPE and allowed for comparison of the productivity rates of each separately.

At the same time, according to Roberts (2007, in: Temizöz 2013), comparing from-scratch translation and MTPE is only justified if comparing outputs of the same person (Temizöz, 2013).

After defining the objectives and setting the measurement parameters, a decision was made to conduct two experiments: a pilot one—on a smaller sample and with students many of whom had no previous experience with MTPE—and a main one—on a slightly

larger sample and with students enrolled in Machine Translation Post-Editing course, in the middle of semester.

For the first experiment, an e-learning text from the environment of a transportation company was selected. The text was deemed adequate as it represented a common localization task with the need for MTPE, was heterogeneous with regard to its structure (running text, titles, numbering, paragraphs, choice selection, repetitions), and included different terms and abbreviations. It was slightly adapted so that it did not contain any elements characteristic of the industry that could reduce its comprehensibility. A glossary of terms to be adhered to was prepared, and the post-editors were instructed on which terms were designated as DNT *(Do Not Translate)*. The text contained a total of 516 words divided into 53 segments. The text with the glossary was then uploaded to the CAT tool Phrase, which the university had a license for and was accessible to all students. In Phrase, the text was translated by an MT system, in this case DeepL, and two files were created from the translated text: one preserved the MT output in the first half, this half was intended for MTPE, and the translation of the second half, which was intended for from-scratch translation, was deleted. In the second file, it was done the other way around to obtain data related to post-editing of the entire text. For each student, a separate project was prepared in Phrase, containing a source text with one half ready for MTPE and the other for from-scratch translation, and an attached glossary. In addition to translating from scratch and doing MTPE, the task of the post-editors was to record the time taken for each activity. After completing one part, they were not able to return to it again, thus the times they recorded were final. Nine post-editors took part in this pilot experiment.

For the second experiment, a support article for an iPhone stuck during transfer from a previous iPhone was chosen. The text had to be different, as some of the post-editors took part in both experiments. This time the text was rather homogeneous with regard to its structure but required greater attention to the syntax due to its instructive nature. The text was not edited, nor was it necessary to create a glossary for it. The text contained 566 words divided into 33 segments. The procedure was the same and a project was prepared for each post-editor. The experiment took place on university grounds during a class on Machine Translation Post-Editing (a compulsory elective course for MA students majoring in philology with a specialization in translation and interpreting, taught at the Department of British and American Studies at Comenius University). It was attended by 12 post-editors, all students of this course. The experiment was conducted in the middle of the semester; all post-editors had already had some experience with MTPE. In this case, they recorded the time themselves as well. They were not able to return to the completed task, the times were final.

The second step was to analyze the data using the TER metric. The program for TER analysis, along with the expertise required for its evaluation, was provided by exe a.s.

localization. Segments that were post-edited were analyzed and their TER scores were determined. Since the TER score of each segment is calculated based on the word count, it is important to factor this in when calculating the mean TER score of the entire post-edited section. Therefore, the weighted mean of each segment was calculated and exe's instructions for TER analysis evaluation was further followed. A table was created and the results of TER analysis for individual segments were entered into the table, along with the results for the entire post-edited part. Subsequently, both the individual times the post-editors took to perform MTPE and the times they recorded when translating from scratch were added into the table. These were compared and used to determine whether, in some cases, MTPE took more time than-from scratch translation, and if so, what was the TER score at which it happened. When evaluating the main experiment, an interesting phenomenon was observed: the results of one of the post-editors, who is known, to be actively working as a translator while still studying, differed substantially from all other post-editors. Thus, one more research question was formulated: Will the results of a professional post-editor be similar to the results of this experienced student post-editor? The professional post-editor was chosen on the basis of their familiarity with the translated subject. They proceeded in the same way: they post-edited and translated the text from the second experiment, i.e., the same text the experienced student post-editor post-edited and translated, and they recorded their times.

4 Machine Translation Quality Based on TER Analysis

Firstly, the outputs of the pilot experiment were analyzed, in which 9 post-editors participated. Five of them did MTPE first (they post-edited the first half of the text, which contained 230 words of MT output) and then they translated the second half of the text from scratch. The remaining four translated the first half of the text from scratch first and then they did MTPE (they post-edited the second half of the text, which contained 216 words of MT output). The procedure of having part of the post-editors post-edit one half of the text and the rest post-edit the other half was chosen in order to get as comprehensive a view of MTPE and from-scratch translation as possible.

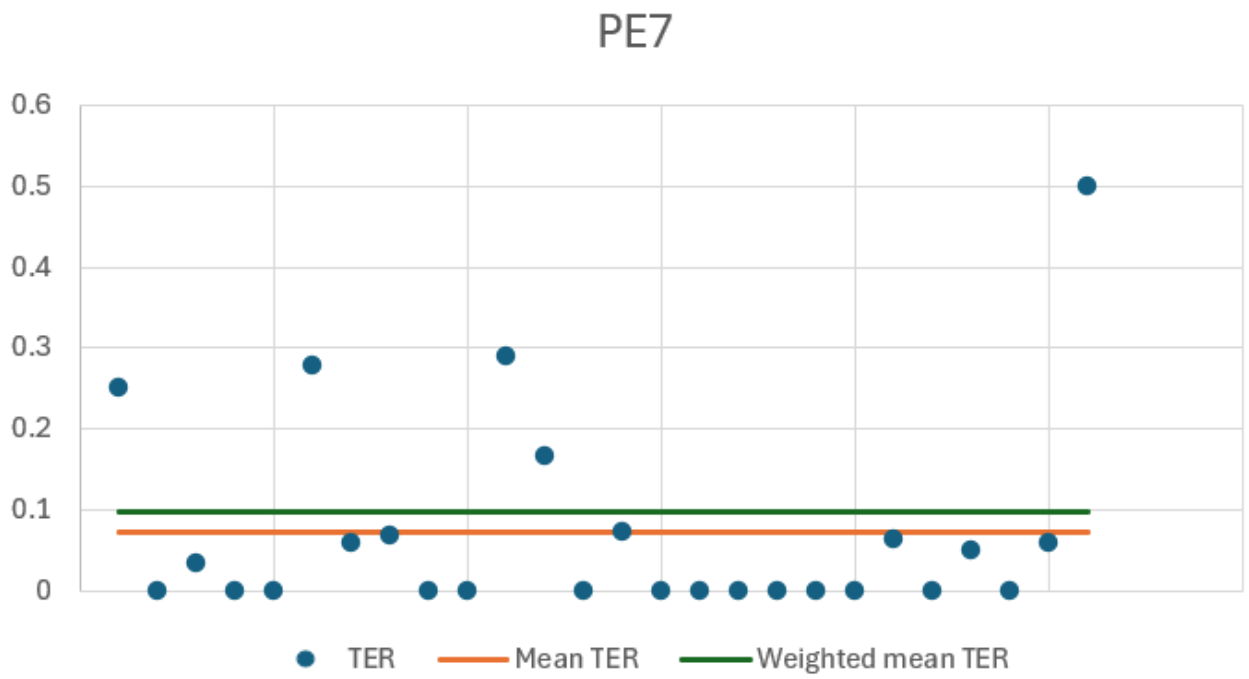
From-scratch translation was not faster than MTPE in any of the 9 cases. The variance of from-scratch translation (shown in Table 1) ranged from 20 minutes and 38 seconds to 1 hour, 1 minute, and 9 seconds; the variance of MTPE ranged from 10 minutes and 3 seconds to 31 minutes and 52 seconds.

Table 1. ***Results overview (pilot experiment): time***

Post-editor	MTPE	From-scratch
PE1	21 min 57 sec	44 min 53 sec
PE2	10 min 42 sec	61 min 9 sec
PE3	31 min 52 sec	43 min 44 sec
PE4	15 min	40 min
PE5	26 min	36 min
PE6	19 min 43 sec	43 min 27 sec
PE7	10 min 3 sec	27 min 34 sec
PE8	17 min 32 sec	54 min 26 sec
PE9	11 min	20 min 38 sec

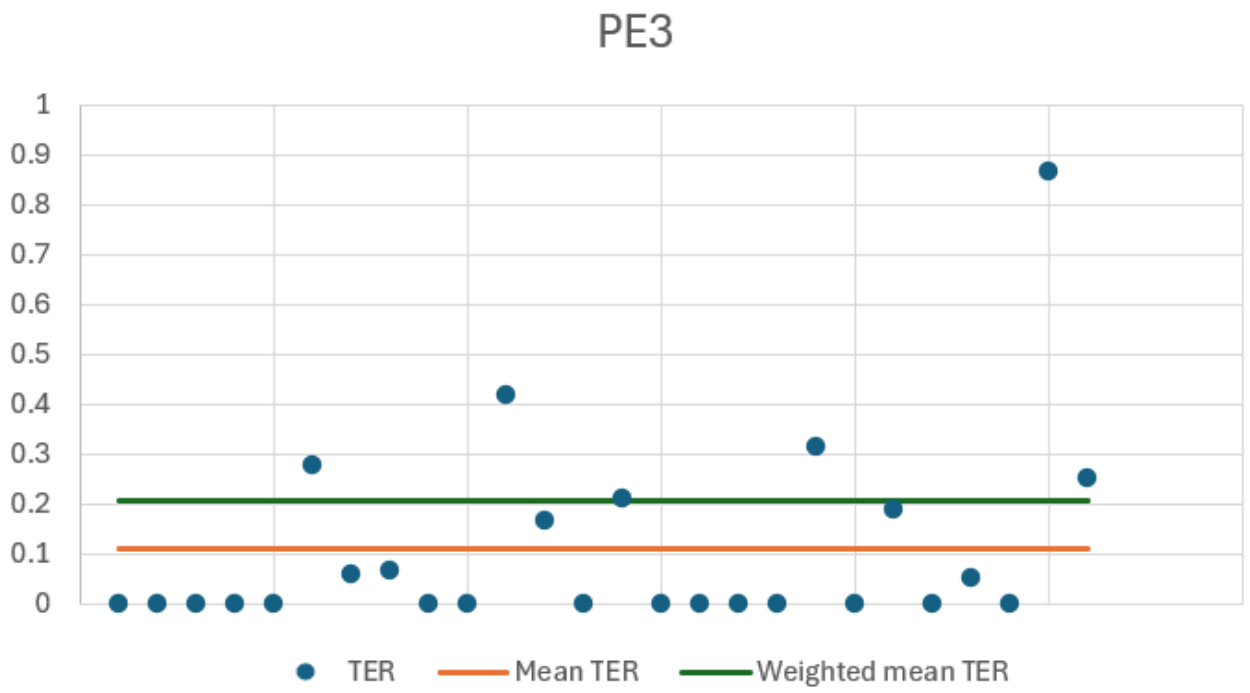
The fastest to post-edit was PE7 with a time of 10 minutes and 3 seconds. PE7 post-edited the MT output with a word count of 230 words and the mean TER score was 7% (see Figure 1). The weighted mean TER score, i.e., the score that takes into account the number of words in each segment and their contribution (weight) to the overall mean score, was 10% in this case. It took this post-editor 27 minutes and 34 seconds to translate from scratch, almost three times the time required for MTPE.

Figure 1. ***PE7's TER Analysis Chart***



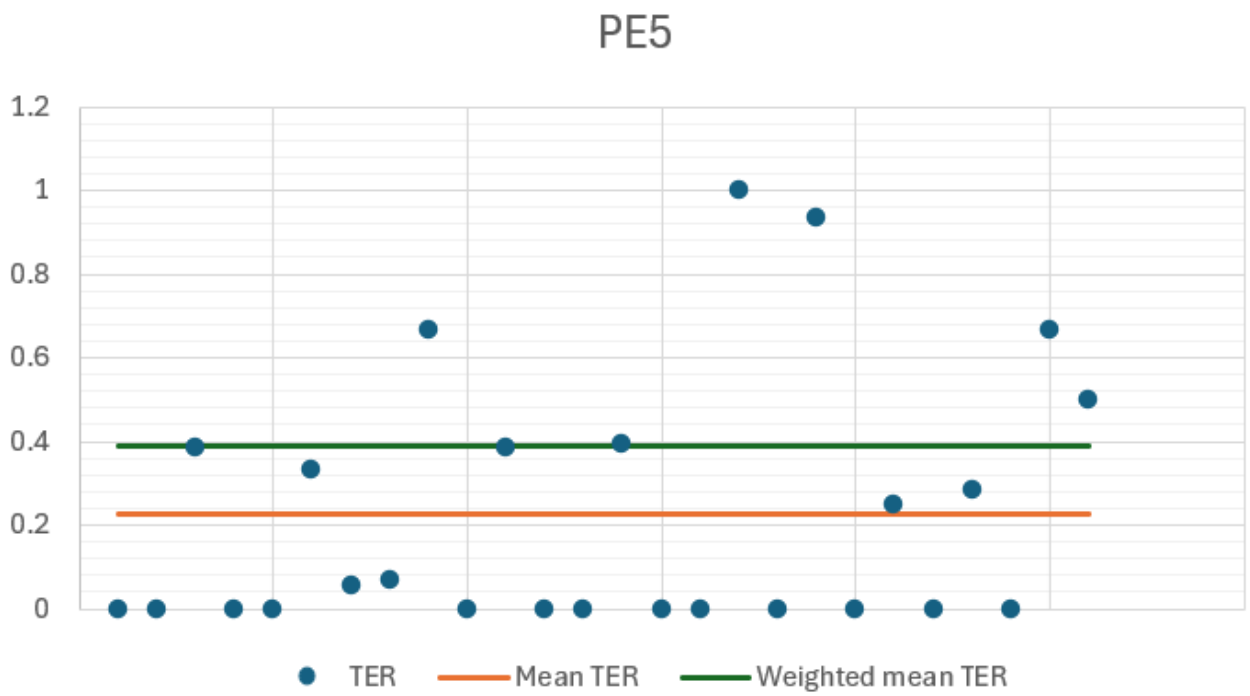
On the other hand, the slowest to post-edit was PE3 with a time of 31 minutes and 52 seconds. PE3 also posted the MT output with a word count of 230 words and the mean TER score was 11% (see Figure 2). The weighted mean TER score was 21%. From-scratch translation took this post-editor 43 minutes and 44 seconds, only about one third more compared with MTPE.

Figure 2. ***PE3's TER Analysis Chart***



The highest TER score of 23% (see Figure 3) with the weighted mean TER score of 39% were recorded by PE5, who took 26 minutes to do MTPE of 230 target words and 36 minutes to translate from scratch.

Figure 3. ***PE5's TER Analysis Chart***



The analysis showed that in none of the cases did the time required for MTPE occur to be longer than the time required for translation from scratch, i.e., in all cases it was demonstrated within the possible measurements (see Tables 2 and 3) that MTPE was (at least time-wise) more efficient than from-scratch translation. Thus, up to the highest recorded mean TER = 23% (weighted mean TER = 39%), there was no evidence that from-scratch translation was more efficient than MTPE.

Secondly, it can be observed that identification of and defining the threshold of the acceptability of MT output for post-editing was not possible in this experiment; one could only assume from the analysis that it must be at TER > 23% (weighted mean TER > 39%), as this is the highest TER score recorded where MTPE is still more efficient than from-scratch translation. Therefore, if at TER = 23% (weighted mean TER = 39%) MTPE is more efficient than from-scratch translation, the threshold of the acceptability of MT output for post-editing must be higher than this recorded TER score.

At the same time, the data does not indicate the existence of a trend that puts the time required for MTPE in a direct linear relationship with the TER score, or the time required for from-scratch translation.

Table 2. ***Results overview (pilot experiment): time, mean TER, weighted mean TER***

Post-editor	MTPE	Mean TER	Weighted mean TER	From-scratch
PE1	21 min 57 sec	9%	12%	44 min 53 sec
PE2	10 min 42 sec	5%	8%	61 min 9 sec
PE3	31 min 52 sec	11%	21%	43 min 44 sec
PE4	15 min	17%	21%	40 min
PE5	26 min	23%	39%	36 min
PE6	19 min 43 sec	6%	7%	43 min 27 sec
PE7	10 min 3 sec	7%	10%	27 min 34 sec
PE8	17 min 32 sec	12%	20%	54 min 26 sec
PE9	11 min	7%	12%	20 min 38 sec

Table 3. *Ascending order of post-editors in each area (pilot experiment)*

PEMT	Mean TER	Weighted mean TER	From-scratch
PE7	PE2	PE6	PE9
PE2	PE6	PE2	PE7
PE9	PE7	PE7	PE5
PE4	PE9	PE1	PE4
PE8	PE1	PE9	PE6
PE6	PE3	PE8	PE3
PE1	PE8	PE3	PE1
PE5	PE4	PE4	PE8
PE3	PE5	PE5	PE2

Afterwards, the main experiment, in which 12 post-editors participated, was analyzed. While the pilot experiment involved students with relatively little or even no experience with MTPE, the main experiment involved students enrolled in Machine Translation Post-Editing course in the middle of the semester; in this way, we wanted to ensure that all post-editors had the same minimum background when it came to MTPE. The setup was the same: six post-editors did MTPE first (they post-edited the first half of the text, which contained 236 words of MT output) and then they translated the second half of the text from scratch. The remaining six post-editors translated the first half of the text from scratch first and then they did MTPE (they post-edited the second half of the text, which contained 272 words of MT output).

From the analysis, it is already evident at a glance that post-editors in this experiment were faster than post-editors in the pilot experiment; the variance of from-scratch translation (shown in Table 4) ranged from 18 minutes and 35 seconds to 35 minutes and 18 seconds, and the variance of MTPE ranged from 9 minutes to 21 minutes and 40 seconds.

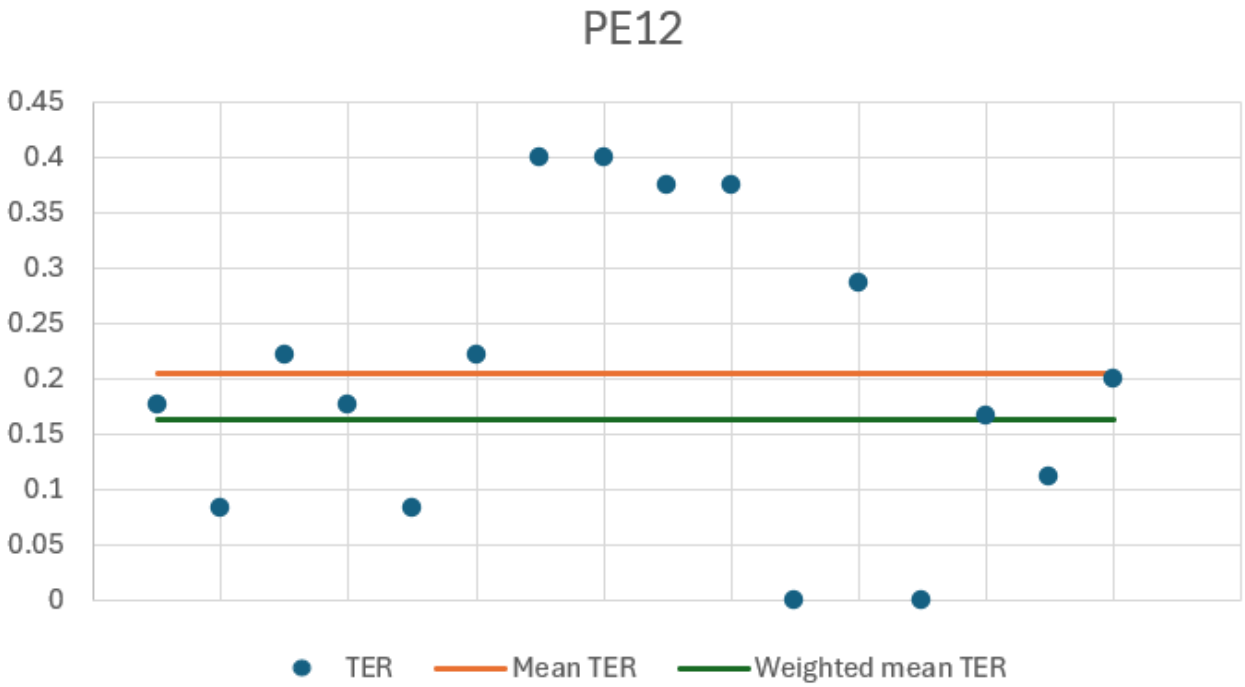
Table 4. ***Results overview (main experiment): time***

Post-editor	MTPE	From-scratch
PE1	21 min 40 sec	18 min 35 sec
PE2	15 min	25 min
PE3	16 min 34 sec	21 min 40 sec
PE4	11 min	27 min
PE5	21 min 40 sec	22 min
PE6	14 min	32 min 45 sec
PE7	13 min 51 sec	21 min 46 sec
PE8	14 min 48 sec	21 min 40 sec
PE9	17 min	27 min
PE10	14 min 30 sec	35 min 18 sec
PE11	17 min	33 min
PE12	9 min	25 min

The higher experience of the post-editors with MTPE could be partly responsible for this acceleration (as evidenced by the higher recorded values of the mean TER score compared to the pilot experiment), but there's no data to confirm this assumption. Moreover, if the post-edited text was the same in both experiments (some of the post-editors took part in both experiments, therefore the text had to be different), it would be possible to speak of a greater experiential complex of the post-editors. In this way, it can be assumed that the improvement in the temporal data was most likely due to the text itself, which was more general and comprehensible compared to the first text.

The fastest to post-edit was PE12, who took only 9 minutes to post-edit the 236-word MT output. The mean TER score (shown in Figure 4) was 20%, the weighted mean TER score was 16%. From-scratch translation took this post-editor 25 minutes.

Figure 4. ***PE12's TER Analysis Chart***



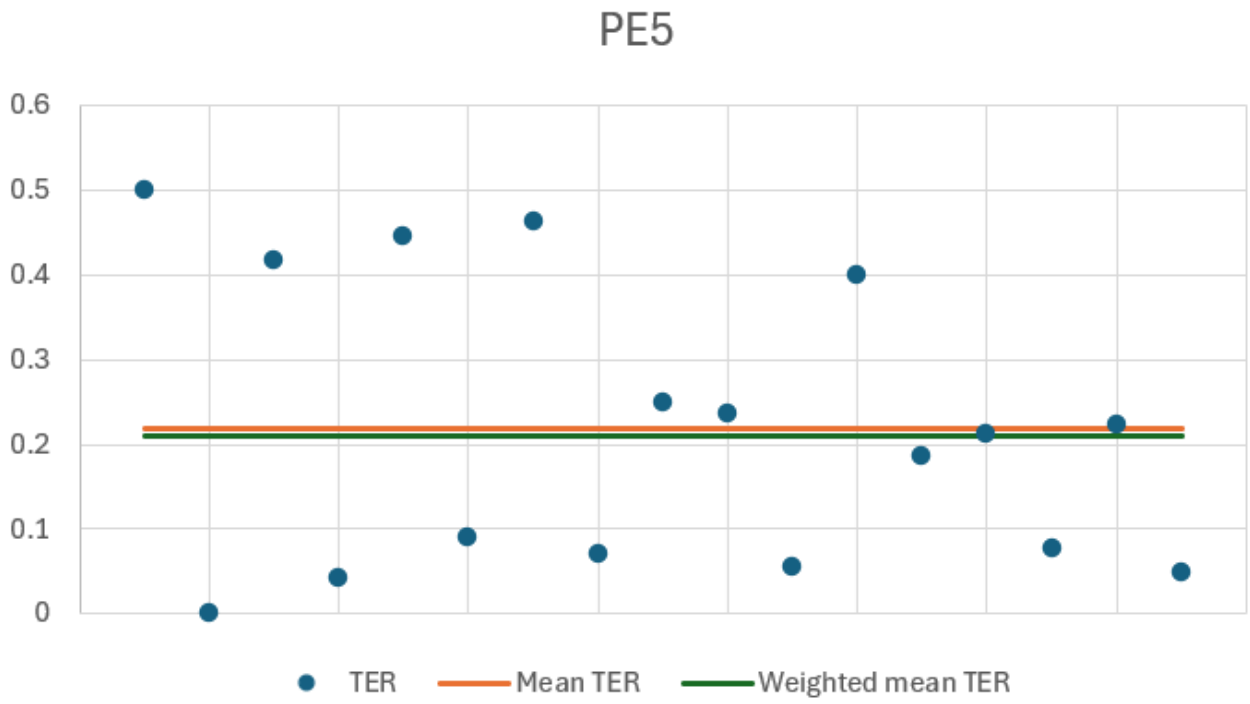
Similar results can be seen for almost all of the other post-editors (shown in Table 5); the time required for MTPE is lower than the time required for from-scratch translation, the differences between these times are heterogeneous and substantial, and there is no evidence of a tendency for the time required for MTPE to increase in direct proportion to the TER score.

Table 5. ***Results overview (main experiment): time, mean TER, weighted mean TER, sorted in ascending order by time required for MTPE***

Post-editor	MTPE	Mean TER	Weighted mean TER	From-scratch
PE12	9 min	20%	16%	25 min
PE4	11 min	11%	14%	27 min
PE7	13 min 57 sec	8%	7%	21 min 46 sec
PE6	14 min	17%	26%	32 min 45 sec
PE10	14 min 30 sec	7%	12%	35 min 18 sec
PE8	14 min 48 sec	9%	10%	32 min 40 sec
PE2	15 min	9%	11%	25 min
PE3	16 min 34 sec	27%	27%	21 min 40 sec
PE9	17 min	15%	12%	27 min
PE11	17 min	17%	17%	33 min
PE5	21 min	22%	21%	22 min
PE1	21 min 40 sec	44%	41%	18 min 35 sec

For PE5, the difference between the times required for MTPE and from-scratch translation was only 1 minute (MTPE took 21 minutes and from-scratch translation took 22 minutes), indicating that MTPE was only slightly more efficient than from-scratch translation. The mean TER score (shown in Figure 5) was 22%, the weighted mean TER score was 21%.

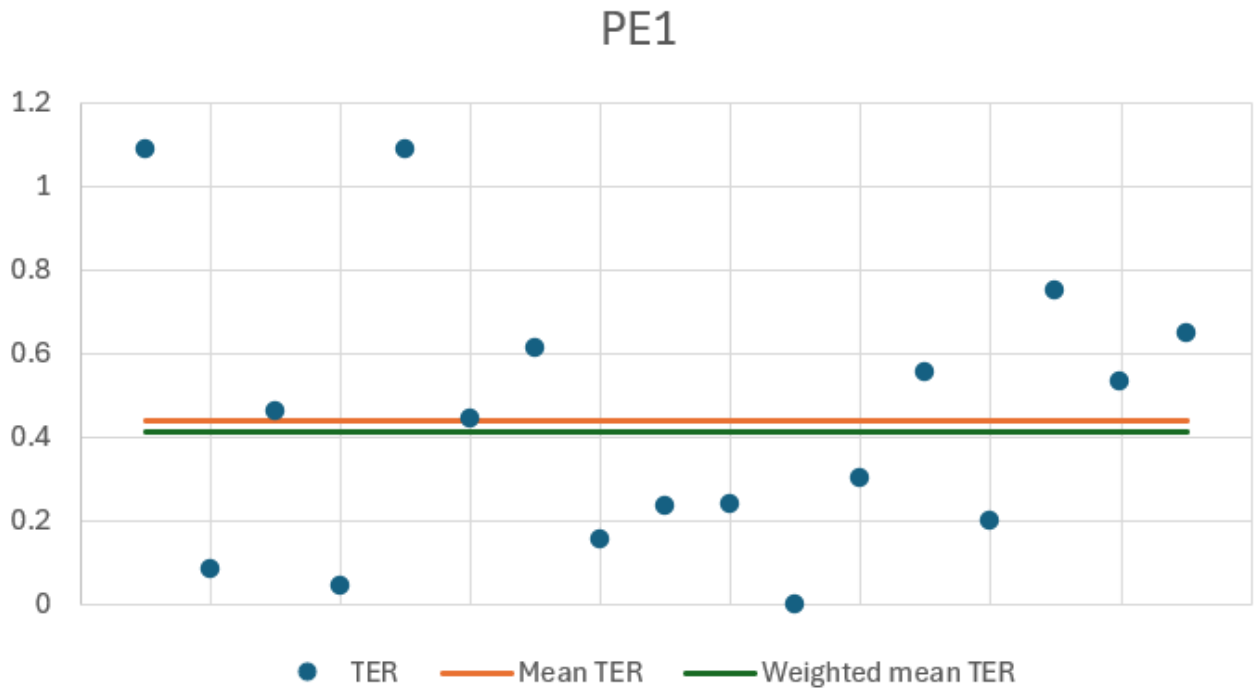
Figure 5. ***PE5's TER Analysis Chart***



With this post-editor, a hint of a tendency can already be seen that if the TER score were to rise to an even higher level, it could mean an increase in the time required for MTPE such that there would be a change in the efficiency of both processes in favor of from-scratch translation. However, it cannot be determined what TER score would be the tipping point based on this post-editor, because the only TER score recorded is the one at which MTPE is still more efficient than from-scratch translation. And since the recorded results do not show the existence of a correlation between the time required for MTPE and the TER score, there's no way of determining – based on the results of PE5 – even a hypothetical TER score where the change in efficiency would occur.

PE1's results are fundamentally different from all other post-editors. First of all, a higher MTPE time can be seen when compared to from-scratch translation, as well as a higher TER score rate. MTPE of the 272-word MT output took PE1 21 minutes and 40 seconds, which is 3 minutes longer than it took this post-editor to translate from scratch (18 minutes and 35 seconds). The recorded mean TER score was 44%, the weighted mean TER score was 41% (shown in Figure 6).

Figure 6. ***PE1's TER Analysis Chart***



What is significant about PE1's results is not only the highest recorded mean TER score, along with the weighted mean TER score (shown in Figure 7), and the higher MTPE time compared to from-scratch translation (shown in Figure 8), but also that both the MTPE time was the highest of all post-editors and the from-scratch translation time was the lowest of all post-editors. The MTPE time was 6 minutes above average, the from-scratch translation time was 8 minutes below average.

Figure 7. ***Chart comparing TER scores (main experiment)***

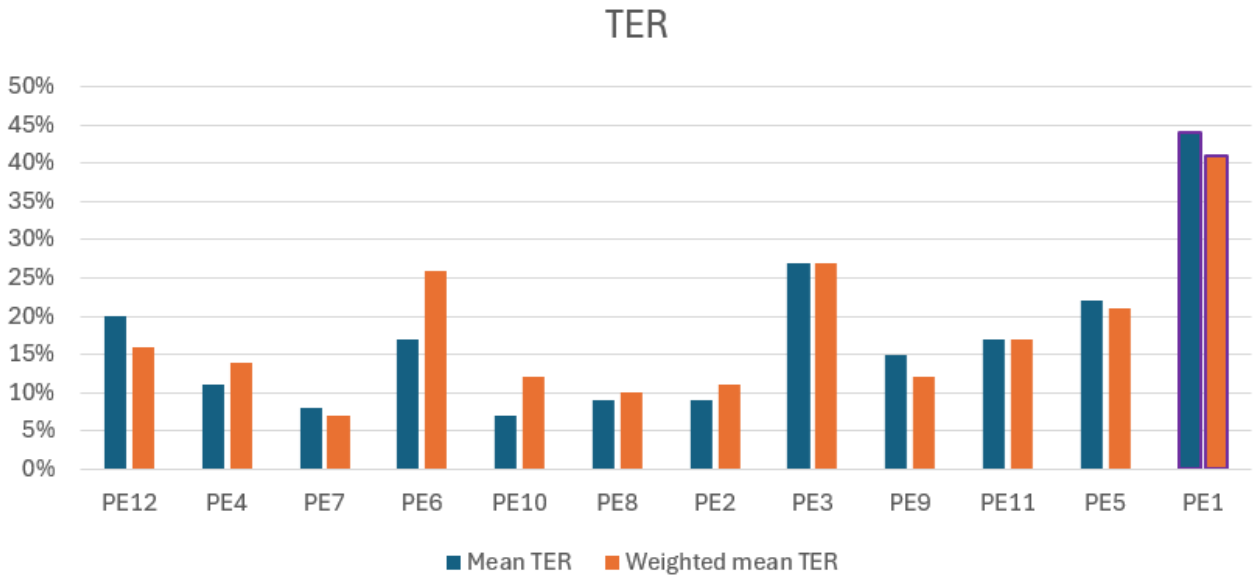
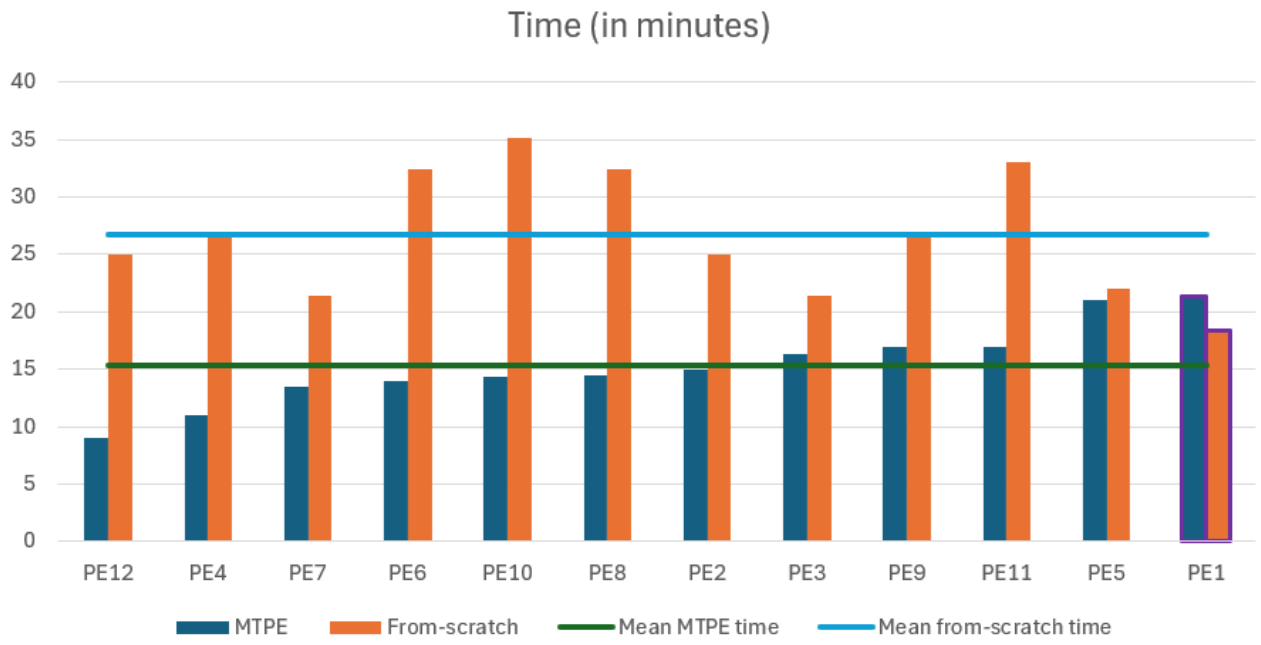


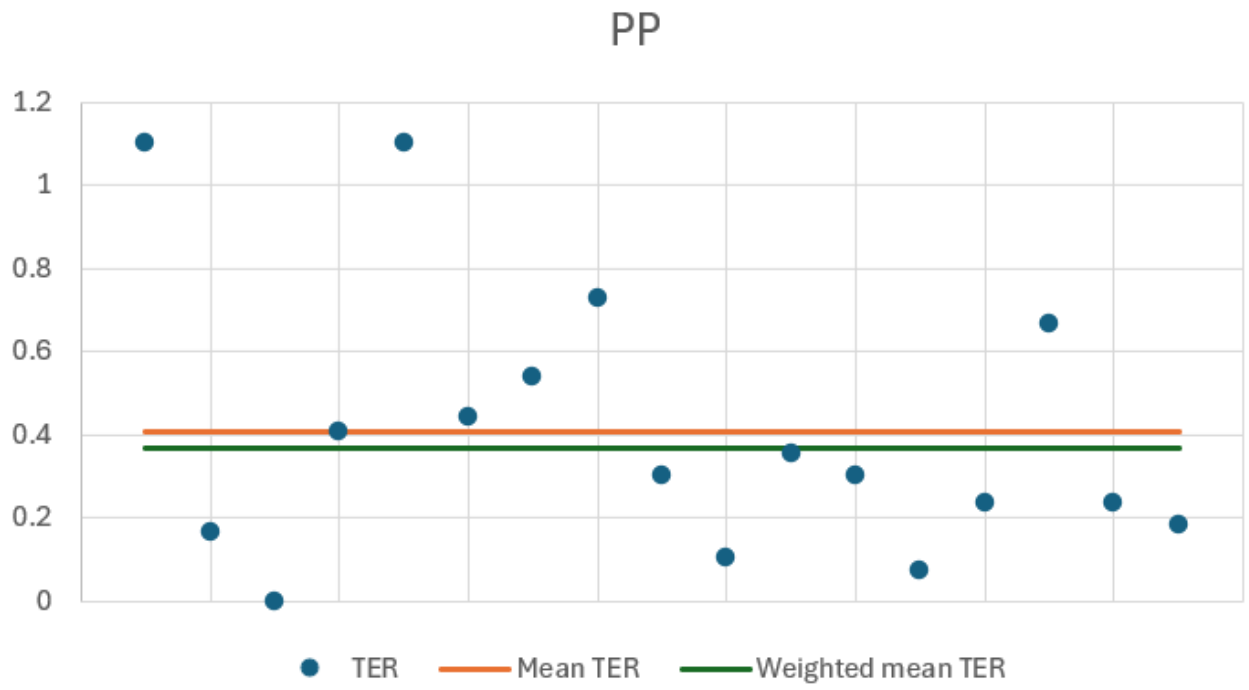
Figure 8. ***Chart comparing MTPE and from-scratch translation times (main experiment)***



On this basis, it is believed that the difference between PE1 and the other post-editors is due to their experiential complex, in which they surpass the other post-editors (PE1 have been active as a translator for a long time). It was decided to verify this assumption by assigning the same task to a professional translator/post-editor.

The professional translator/post-editor (PP) did the same task as PE1translating the first half of the text from scratch, then performing MTPE on the second half of the text, which contained 272 words of MT output. MTPE took them 22 minutes and 24 seconds, which is only 44 seconds more than it took PE1, and from-scratch translation took PP 17 minutes and 8 seconds, which is again just one minute and 27 seconds less than it took PE1. The mean TER score (shown in Figure 9) was 41% (PE1's TER score was 44%) and the weighted mean TER score was 37% (PE1's weighted mean TER score was 41%).

Figure 9. ***Professional translator and post-editor's TER Analysis Chart***



Based on the nearly identical temporal results and the comparable TER scores of PP and PE1 (shown in Table 6), it can be concluded that PE1's experiential complex exceeds that of the remaining post-editors and causes the difference between their results and the results of the remaining post-editors.

Table 6. ***PE1's and PP's results overview: time, mean TER, weighted mean TER***

Post-editor	MTPE	Mean TER	Weighted mean TER	From-scratch
PE1	21 min 40 sec	44%	41%	18 min 35 sec
PP	22 min 24 sec	41%	37%	17 min 8 sec

5 Discussion

Firstly, as mentioned above, from the analysis of the pilot experiment it can be concluded that MTPE was faster than from-scratch translation in all cases, i.e., it demonstrated higher temporal efficiency within the range of possible measurements compared to from-scratch translation in each case (see Table 1 in the previous section). The highest recorded mean TER scores in this experiment were TER = 23% (weighted mean TER = 39%) (see Figure 3 in the previous section). These results are in line with

the findings of Escartín and Arcedillo (2015), who have measured productivity gains for TER ≤ 0.21 (i.e., 21%). The remaining values are shown in Table 2 in the previous section.

In this experiment, it was not possible to identify and define a threshold of the acceptability of MT output for post-editing, as even at the highest recorded TER (23%), MTPE was more efficient than from-scratch translation. Based on this result, it might be assumed that the acceptability threshold must therefore be above the TER score of 23% (or 39% for the weighted mean TER).

At the same time, the measured results do not show a direct correlation between the time required for MTPE and the TER score, as has already been stated by Stefaniak (2020) in her research on English-Polish language pairs, nor between the time required for MTPE and the time required for from-scratch translation. Koponen's (2016) conclusions have shed light on the variation in the measured values: all of these are influenced by the characteristics of the source text, the type of editing, and the varying speed of the post-editors themselves.

The main experiment yielded comparable results: in almost all cases, MTPE was more efficient (faster) than from-scratch translation. The mean TER scores were higher despite the more general text, which can be interpreted to be due to the greater level of experience of the post-editors, especially their increased ability to identify errors in the MT output, which is also more and more difficult with the increasing quality of NMT systems, according to the findings of Yamada (2019, in: Vieira 2019). The participants were MA students at Comenius University majoring in philology with a specialization in translation and interpreting enrolled in Machine Translation Post-Editing course. Although it could not be guaranteed that all post-editors had the same experience, by selecting students enrolled in this course, it was ensured that all had comparable minimum skills. The argument about more experienced post-editors can also be supported by the temporal data, which is lower compared to the pilot experiment, i.e., post-editors in the main experiment post-edited and translated faster. Similarly with the results of the pilot experiment regarding a correlation between the time required for MTPE and the TER score, it can be concluded that such a correlation was not demonstrated here. Gaspari et al. (2014, in: Stefaniak 2020) have stated that such a correlation is weak, if it exists at all.

Results that already show a departure from the rest were found for two post-editors: PE5 and PE1. Although PE5's MTPE time was still lower than from-scratch translation time, the difference was only one minute (see Table 4 in the previous section). The recorded mean TER score was 22% (weighted mean TER = 21%) (see Table 5 or Figure 5 in the previous section). As mentioned earlier, it can be assumed that if the TER scores were to increase further, it could result in an increase of time required for MTPE in such a way that it could cause a change in the efficiency of MTPE vs. from-scratch translation in favor of from-scratch translation. However, since the results do not show a

correlation between the TER score and the time required for MTPE, there's no way of determining – even hypothetically – what TER score would be the threshold representing the tipping point in efficiency. The recorded TER score—at which MTPE was consistently more efficient—is consistent with the findings of Escartín and Arcedillo (2015), who, for English-Spanish language pairs, have set the threshold for the efficiency of MTPE versus from-scratch translation at TER ≤ 0.21 (i.e., 21%).

The most interesting were the results of PE1 (experienced student), whose MTPE time was the only one to be higher than the from-scratch translation time: MTPE = 21 minutes and 40 seconds; from-scratch translation = 18 minutes and 35 seconds (see Table 5 in the previous section). The times also represented the thresholds recorded within this sample: the MTPE time was the highest among all post-editors, the from-scratch translation time the lowest (see Table 5 or Figure 8 in the previous section). PE1 also had the highest recorded TER score, which significantly exceeded the scores recorded for the remaining post-editors (see Figure 7 in the previous section): the mean TER score of PE1 was 44% (weighted mean TER score was 41%) (see Figure 6 in the previous section).

These findings formed the basis for the belief that the results demonstrate PE1's higher experiential complex relative to the other post-editors, as PE1 is known, while still studying, to be actively working as a translator in a company that is among the best in the country in terms of its specialization.

To confirm or refute this assumption, an additional research question was therefore asked: Are the results of a professional post-editor similar to the results of PE1?

The results of the professional post-editor are almost identical to those of PE1 (see Table 6 in the previous section), and it can be concluded that PE1 possesses a significantly higher experiential complex compared to the other post-editors: the professional post-editor's TER score indicates that the MT output contained a significant number of flaws that needed to be identified and corrected, which PE1 handled significantly better than the other post-editors; PE1's from-scratch translation time was only one minute higher than the professional post-editor's, and yet it was still 8 minutes lower than average (see Figure 8 in the previous chapter).

The differences in the measured values between the professional post-editor and PE1 support both Stefaniak's (2020) and Gaspari et al.'s (2014, in: Stefaniak 2020) statements that the relationship between TER score and MTPE time is not directly proportional, or that no clear correlation between the two has been demonstrated.

Based on these claims and on the average of scores of PE1 and the professional post-editor, it can be concluded that at mean TER = 42.5% (weighted mean TER = 39%), MTPE is less efficient than from-scratch translation.

Stefaniak (2020), in her research on English-Polish language pairs, has noted a productivity increase in MTPE even at TER = 39%, therefore, it can be concluded that the threshold of the acceptability of MT output for post-editing is in the interval of TER = 39%–42.5%.

This is consistent with Guerrero's (2020) hypothesis that 50% edit distance is too high as an MT output acceptability indicator; and also with another conclusion in which she has argued that MT output with an edit distance between 30% and 40% is mostly rated as 2 on the acceptability scale (from 1 to 4) for post-editing, which corresponds to the threshold of the acceptability of MT output for post-editing in the interval of TER = 39%–42.5%

This finding is also slightly higher than that of de Gibert Bonet (2018, in: Stefaniak 2020), who, for Spanish-English language pairs, has determined the TER score threshold value up to which productivity gains in MTPE occur to be 33%.

6 Conclusion

The aim of this study was to determine the threshold of the acceptability of MT output for post-editing based on TER analysis. It compares the efficiency of MTPE against from-scratch translation and seeks to determine the threshold at which the efficiency scales tip towards from-scratch translation.

The work defines the concept of machine translation post-editing, presents approaches to it and the process itself, and includes an insight into the similar research carried out on other language pairs. Towards the end of the first part of this work, the research questions are introduced and the methodology of this work as well as the research methods are described.

The second, practical part presents the results of the research together with tables and graphs. The results are elaborated on in the discussion. The research showed that there was probably no clear correlation between the time required for post-editing a machine translation output and the TER score which would place these variables in a directly proportional relationship; furthermore, the comparison with a professional translator identified and confirmed a significantly higher experiential complex of one post-editor from the main group (PE1), whose results alone produced a deviation from the otherwise observed trend; finally, based on the results of these two post-editors, it was concluded that the threshold of the acceptability of MT output for post-editing lies in the interval of TER = 39%–42.5% These findings are in line with those of Stefaniak (2020) and Guerrero (2020).

From a practical point of view, these findings can serve as a guide for translation agencies or individual translators on how to approach, for example, ordering/accepting machine translation post-editing jobs, what expectations are realistic for

translators/agencies, as well as how to assess the value of their own effort, work, or time. From a didactic point of view, these findings can serve as an impetus for educational institutions that train translators to respond to the needs and trends of the market and to offer their students more courses that better prepare them for the realities of the translation profession.

In the future, this research could be enriched by the inclusion of language quality assurance (LQA) and the participation of a larger number of professional translators.

Nemergut, Matúš. 2024. Machine Translation Quality Based on TER Analysis from English into Slovak. In: L10N Journal 2(3), pp. 60–86.

Bibliography

Absolon, Jakub. 2018. Strojový preklad a posteditovanie. PhD dissertation. Nitra: DTS FA Constantine the Philosopher University in Nitra.

Bhandari, Pritha, and Nikolopoulou, Kassiani. 2020. What Is a Likert Scale? | Guide & Examples. https://www.scribbr.com/methodology/likert-scale/. Accessed on: 4 April 2024.

do Carmo, Félix. 2020. 'Time is money' and the value of translation. In: Translation Spaces. 9(1): pp. 35-57.

Escartín, Carla Para, and Arcedillo, Manuel. 2015. A Fuzzier Approach to Machine Translation Evaluation: A Pilot Study on Post-editing Productivity and Automated Metrics in Commercial Settings. In: Babych, Bogdan, Eberle, Kurt, Lambert, Patrik, Rapp, Reinhard, Banchs, Rafael E., and Costa-Jussà, Marta R. (eds.), Proceedings of the Fourth Workshop on Hybrid Approaches to Translation (HyTra). Beijing: Association for Computational Linguistics. pp. 40-45.

Guerrero, Lucía. 2020. In Search of an Acceptability/Unacceptability Threshold in Machine Translation Post-Editing Automated Metrics. In: Proceedings of the 14th Conference of the Association for Machine Translation in the Americas (Volume 2: User Track). Virtual: Association for Machine Translation in the Americas. pp. 32-47.

Hudecová, Elena, Stahl, Jaroslav, Benková, Lucia, and Munková, Daša. 2021. Porovnanie strojového, posteditovaného a ľudského prekladu technickej dokumentácie zo slovenčiny do nemčiny. In: Slovenská reč. 86(2): pp. 192-207.

Koehn, Philipp. 2020. Neural Machine Translation. Cambridge: Cambridge University Press.

Koponen, Maarit. 2016. Machine Translation Post-editing and Effort: Empirical Studies on the Post-Editing Process. PhD dissertation. Helsinki: University of Helsinki.

O'Brien, Sharon. 2006. Machine-Translatability and Post-Editing Effort: An Empirical Study using Translog and Choice Network Analysis. PhD dissertation. Dublin: School of Applied Language and Intercultural Studies, Dublin City University.

O'Brien, Sharon. 2011. Towards Predicting Post-Editing Productivity. In: Machine Translation. 25(3): pp. 197–215.

O'Brien, Sharon. 2022. How to Deal with Errors in Machine Translation: Postediting. In: Kenny, Dorothy (ed.), Machine Translation for Everyone: Empowering Users in the Age of Artificial Intelligence. Berlin: Language Science Press. pp. 105–120.

Pavlíková, Diana. 2022. Porovnanie strojového a humánneho prekladu terminológie. In: L10N Journal. 1(1): pp. 64–101.

Rothwell, Andrew, Moorkens, Joss, Fernández-Parra, Maria, Drugan, Joanna, and Austermuehl, Frank. 2023. Translation Tools and Technologies. Abingdon, Oxon; New York: Routledge.

Stefaniak, Karolina. 2020. Evaluating the Usefulness of Neural Machine Translation for the Polish Translators in the European Commission. In: Martins, André, Moniz, Helena, Fumega, Sara, Martins, Bruno, Batista, Fernando, Coheur, Luisa, Parra, Carla, Trancoso, Isabel, Turchi, Marco, Bisazza, Arianna, Moorkens, Joss, Guerberof, Ana, Nurminen, Mary, Marg, Lena, and Forcada, Mikel L. (eds.) Proceedings of the 22nd Annual Conference of the European Association for Machine Translation. Lisboa: European Association for Machine Translation. pp. 263-269.

Temizöz, Özlem. 2013. Postediting Machine Translation Output and its Revision: Subject-Matter Expert Experts versus Professional Translators. PhD dissertation. Tarragona: Universitat Rovira i Virgili.

Nemergut, Matúš. 2024. Machine Translation Quality Based on TER Analysis from English into Slovak. In: L10N Journal 2(3), pp. 60–86.

Vieira, Lucas Nunes. 2019. Post-Editing of Machine Translation. In: O'Hagan, Minako (ed.), The Routledge Handbook of Translation and Technology. London and New York: Routledge. pp. 319-335.

Walker, Julie. 2018. Term of the Week: Fuzzy Match. https://www.thelanguageoflocalization.com/2018/08/01/term-of-the-week-fuzzy-match/. Accessed on: 21 April 2024.

Zhechev, Ventsislav. 2014. Analysing the Post-Editing of Machine Translation at Autodesk. In: O'Brien, Sharon, Balling, Laura Winther, Carl, Michael, Simard, Michel, and Specia, Lucia (eds.), Post-editing of Machine Translation: Processes and Applications. Newcastle upon Tyne: Cambridge Scholars Publishing. pp. 2-23.