Translation Studies, Winter 2024, Vol. 21, No. 84, pp. 54–73: Scientific Research Paper
DOR: 20.1001.1.17350212.1402.21.4.8.0

A Comparison of the Quality and Speed of Post-Edited Machine Translation and Human-Initiated Translation by Novice Translators[1]

S. Hossein Arjani[2] & Masoud Jamshidiha[3]

Abstract

The present study compares post-editing and translation from scratch by novice translators in terms of quality and speed in the English-Persian language pair. Moreover, it investigates the factors that affect the cognitive and temporal aspects of post-editing effort. To that end, 10 B.A. students of English Translator Training were briefed to translate a short English news text by post-editing the raw MT output provided in Persian, while eight students translated the same text from scratch, with both groups performing the task in the online CAT tool MateCat. The performance of the participants was monitored and screen-recorded to compare the two groups' speed in completing the task. Furthermore, the two groups' translations were evaluated analytically and holistically by three evaluators. It was found that the post-editors were significantly faster, and their TL texts were of a considerably higher quality. After completing the task, the participants were asked to fill out questionnaires to provide insight into the cognitive and temporal post-editing efforts. The responses indicated that the grammatical errors present in the raw MT output were the most important factor affecting the temporal post-editing effort, while finding 'proper equivalents', and balancing usage of MT output were reported to a lesser extent, and correcting zero-width non-joiner was reported by only one post-editor. The most widely reported issue related to cognitive effort was the concern that post-editing could adversely affect the creativity of the translator, while finding 'proper equivalents', balancing usage of MT output, and producing an easily readable target text, among others, were also reported to be causes for cognitive effort, albeit with less frequency.

Keywords: CAT tools, Machine translation post-editing (MTPE), Novice translator, Post-editing effort, Translation quality

---

1. This paper was received on 23.10.2023 and approved on 02.01.2024.

2. Corresponding Author: Assistant Professor, Department of English Translation Studies, Faculty of Persian Literature and Foreign Languages, Allameh Tabataba'i University, Tehran, Iran; email: arjani@atu.ac.ir

3. M.A. in Translation Studies, Allameh Tabataba'i University, Tehran, Iran; email: masoudjamshidiha@gmail.com

1. Introduction

Machine translation systems were initially made to automate translation and remove humans from the translation process (Hutchins & Somers, 1992). However, given that this goal has not yet become feasible, an alternative was suggested in the form of post-editing. Post-editing is "usually understood as a human being (normally a translator) comparing a source text with the machine translation and making changes to it to make it acceptable for its intended purpose" (Koby, 2001, p. 1).

One of the primary questions inspiring much of the research on post-editing is how post-editing differs from translation from scratch, as well as the extent of the benefits provided by post-editing, if any. A crucial concept to this question is post-editing effort, which is often measured to better understand the post-editing process and compare it to translation from scratch. The majority of research on this topic, however, is usually dedicated to quantitative measurement of post-editing effort, and factors that affect post-editing effort have not received enough attention.

In spite of the rapid advancements in translation technology, which necessitates frequent examination of technology-related topics, not much research has been done on the topic of post-editing in Persian language.

The present study is an attempt to address these two gaps, namely the lack of research on the factors that influence post-editing effort, and the scarcity of literature on post-editing in the context of the English-Persian translation. Additionally, this research intends to contribute to the literature on the quality and speed of post-editing, specifically when novice translators are concerned. In order to do so, it will compare post-editing and translation from scratch by translation students in the English-Persian language pair, and investigate the factors that inform post-editing effort. Accordingly, the following research questions were formulated:

1. How do post-editing and translation from scratch by novice translators compare in terms of speed and quality?

2. What are the factors that affect the temporal and cognitive post-editing effort?

2. Review of Literature

Several studies have examined the differences between post-editing and translation from scratch. Generally, post-editing is found and thought to be quicker than translation from scratch (e.g., see Aranberri et al., 2014; Daems et al., 2017; Jia et al., 2019; Koehn, 2009; Läubli et al., 2019; Plitt & Masselot, 2010; Skadiņš et al., 2011). However, there does not appear to be any consensus on the amount of the speed advantage post-editing can provide since different variables seem to be involved.

Findings on quality are mixed, but several studies have found the quality of post-edited translations to be comparable or even better than translations prepared from scratch (e.g., see Carl et al., 2011; Garcia, 2010; Jia et al., 2019; Lee & Liao, 2011; Plitt & Masselot, 2010). Plitt and Masselot (2010) compared MT post-editing and translation from scratch in English to French, Italian, German, and Spanish, and found that texts translated from scratch contained more mistakes than those prepared through post-editing. Carl et al. (2011) compared the quality of post-edited translation and from scratch translation in the English-Danish language pair and found the quality of post-edited translation to be slightly higher than translation from scratch. Jia et al. (2019) compared translation from scratch and post-editing in the English-Chinese language pair and found the quality of post-edited texts to be comparable to those translated from scratch.

Another topic of interest is whether the translation and work experience of the post-editor has an effect on the extent of the potential benefits gained from post-editing. Several studies (e.g., see Aranberri et al., 2014; Daems et al., 2017; Lee & Liao, 2011) have found post-editing to be more beneficial for students and novice translation when compared to professional or more competent translator. Daems et al. (2017) compared the effect of post-editing on M.A. students and professional

translators and found that post-editing decreased the cognitive load of translation for students, while no such effect was observed in professionals. Aranberri et al. (2014) compared the effect of post-editing on the productivity of professional translators and lay users, and found that post-editing increased the productivity of both groups, but had a stronger effect on lay users. Lee and Liao (2011) investigated the effects of post-editing in a study with two groups of students who had different levels of proficiency and concluded that post-editing had a more significant effect for students who were less proficient and could "level the playing" between the two groups (Lee & Liao, 2011, p. 142).

Literature on post-editing effort mostly seems to be focused on measuring post-editing effort through quantitative means (e.g., see Alves, Koglin, et al., 2016; Alves, Szpak, et al., 2016; Nitzke & Oster, 2016). Perhaps, the most influential work about post-editing effort is Krings' (2001) seminal categorization of the components that comprise post-editing effort, in which he proposes that post-editing effort has three constituents (pp. 178–179):

1. Temporal post-editing effort: The time it takes to edit raw MT output into a finalized translation comparable to human translation.

2. Cognitive post-editing effort: The mental processes involved in editing raw MT output into a finalized TL text.

3. Technical post-editing effort: Operations such as addition, deletion, re-ordering, etc. which are required to carry out post-editing.

Temporal effort is the easiest to measure, since measuring time is an easy and straight-forward task, but measuring technical and cognitive effort can be more complicated. Measuring cognitive effort in particular is difficult since one cannot have direct and unmediated access to the translator's mind. Krings (2001) himself suggests that think aloud protocol can serve as a way of measuring cognitive effort. However, as O'Brien (2005) has stated, think aloud protocol is not a perfect

solution, since it can interfere in the translation process, especially because it can affect cognitive and temporal post-editing effort, which makes its use rather counterintuitive.

Cognitive effort can be measured through other means as well. O'Brien (2007) studied the use of eye-tracking for examining translators' interaction with translation memory tools and found eye-tracking to be an effective method for researching translation process when used as a measure of cognitive load. Eye-tracking has also been used in the post-editing research as well. Many different studies have utilized eye-tracking as a way of measuring cognitive effort in post-editing (e.g., see Alves, Koglin, et al., 2016; Alves, Szpak, et al., 2016; Nitzke & Oster, 2016; Vieira, 2014). While this method may not interfere with the translation process to the extent that think aloud protocol does, eye-tracking could still add a cognitive pressure to the students which would affect the process of translation, and was therefore not used in the present study.

Technical post-editing effort has been measured through different methodologies. One of the most widely used methods is keystroke logging (Cumbreño & Aranberri, 2021). Tools developed specifically for the use of keystroke logging in translation research include PET (Aziz et al., 2012), a software developed for performing and researching post-editing, and Translog-II (Carl, 2012), a software dedicated to translation research which can perform key-stroking and eye-tracking as the translator performs the translation task. Several studies have used keystroke logging for investigating post-editing effort, often alongside eye-tracking (e.g., see Alves, Koglin, et al., 2016; Alves, Szpak, et al., 2016; Nitzke & Oster, 2016). Given that the present study primarily focused on qualitative investigation of post-editing effort (in contrast with the majority of studies on the topic), keystroke logging was not used in the present study. Moreover, most software dedicated to keystroke logging do not have a natural translation interface

and have poor support for Persian language, adding another factor affecting the data. For these reasons, technical post-editing effort has not been measured in the present study.

3. Methodology

The study was done in two phases. In the first phase, a controlled experiment was conducted, and in the second phase, the different data gathered in the first phase were analyzed through different methodologies and triangulated.

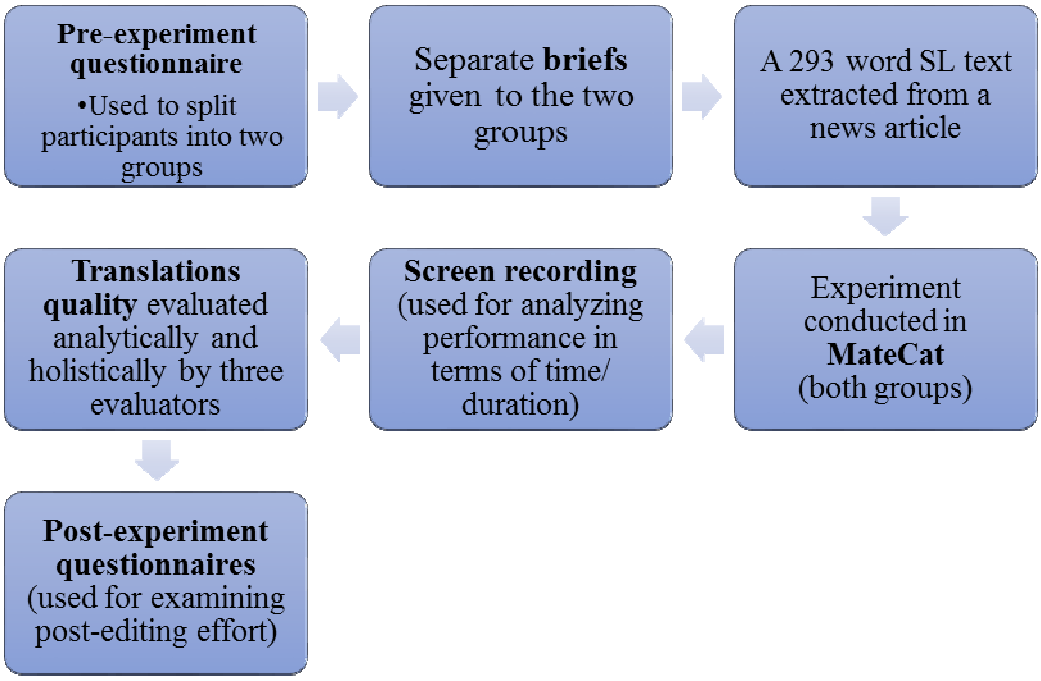

Figure 1. Research Process Overview

Figure 1 illustrates an overview of the workflow of the research process. In the first phase, in order to compare post-editing and translation from scratch, 18 senior students were selected from the B.A. in English Translator Training Program at Allameh Tabataba'i University. Initially, 21 students participated in the experiment, but three had to be removed from the data analysis for non-conformity to the task guidelines and other issues. The students were asked to fill out a pre-experiment questionnaire about their background in using MT. The results were used to split the students into two groups. Ten students were selected to post-edit an

SL text (the experimental group), and eight were chosen to translate the same SL text from scratch (the control group). Separate briefs were prepared for each group, and both groups were asked to translate a 293-word SL text extracted from a news article.

Both groups were instructed to translate the SL text inside the online CAT tool MateCat and provide their finalized work in a word document. Experimental group participants (post-editors) produced the TL using MT, while control group participants (translators) were not permitted access to MT.

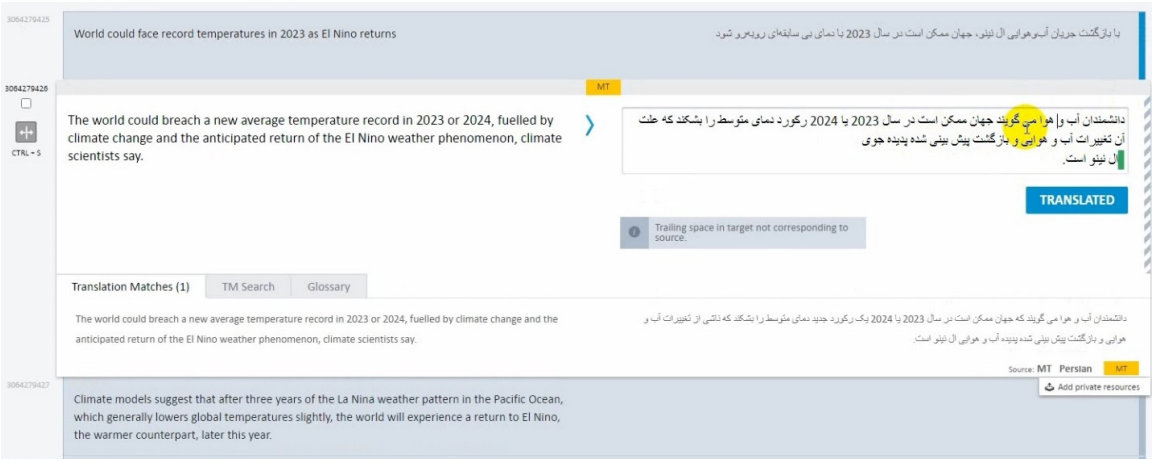

Figure 2. Sample Screenshot of a Screen-recording of a Post-editor Editing a Segment in MateCat

Figure 2 shows a post-editor using MateCat. Note the MT suggestion provided by MateCat.

Figure 3 shows a post-editor using MateCat. Note that MateCat has not provided any MT suggestions.

Both groups had access to online resources such as dictionaries, search engines, etc. and were briefed to: avoid adding or omitting any information to/from the text; preserve the style and tone of the SL; observe Persian language's conventions; and tailor the target text for a general audience. Additionally, post-editors were asked to use as much of MT output as possible.

Figure 3. Sample Screenshot of a Screen-recording of a Translator Editing a Segment in MateCat

The performance of both groups were monitored (in person) and screen-recorded (without their knowledge, but with *ex post facto* consent) to analyze their performance in terms of time. The translations produced by all participants were used to compare the quality of post-edited translations with translations prepared from scratch. In addition, the participants were asked to fill out post-experiment questionnaires to examine the post-editing effort and the participants' perceptions around the task. To ensure the validity of the questionnaires, comments from two expert reviewers were sought and applied.

In the second phase, the translations produced by the participants were evaluated analytically using a slightly modified version of the analytical method proposed by Eyckmans et al. (2009) which consists of an evaluation grid comprised of the following types of errors: 'meaning or sense', 'misinterpretation', 'vocabulary', 'calque', 'register', 'style', 'grammar', 'omission', 'addition', 'spelling', and 'punctuation' (pp. 92–93). Each error type has a fixed penalty which was subtracted from a top score of 50. The modifications were done based on a pilot evaluation by the research authors to adapt the method to the present study.

Moreover, they were validated after consulting with a translation studies PhD holder.

In order to ensure evaluation quality and consistency, the translations were evaluated by three different evaluators, of whom two were post-graduate students with professional translation experience outside the university, and the other evaluator was a graduate alumnus with professional experience as a translator outside the university. Moreover, Pearson correlation coefficient was calculated for all possible pairings of evaluators to ensure inter-rater reliability.

In addition to the analytical evaluation, the evaluators also provided a holistic score (from 1 to 10) and a description for each translation based on their own impression of the quality of the translation. After the evaluators had graded the translations, they were informed of the nature of the experiment (which was not previously disclosed to them) and asked to guess which TL texts were post-edited and which were translated from scratch.

4. Results and Discussion

The screen-recording indicated that, on average, the experimental group participants (hereafter post-editors) required only 57.79% of the time required by the control group participants (hereafter translators) to translate the SL text. The post-editors took 36 minutes and 58 seconds to complete the task on average (excluding small layout/formatting adjustments done in Word after exporting the file from MateCat), while translators required 1 hour, 3 minutes, and 59 seconds on average to accomplish the same.

One-tailed t-test between the time spent by the post-editors and the translators showed a t-value of -2.88 and a p-value of .005 which indicates a statistically significant difference between the speeds of the two groups.

The post-editors also spent much less time searching during the procedure. On average, post-editors required only 43.13% of the time the translators required for

searching. The average searching time of the post-editors was 6 minutes and 9 seconds, compared to 14 minutes and 15 seconds among translators. While this may have been expected based on the findings on the overall time spent on the task by each group, it is worth noting that the searching time of the post-editors comprised a smaller portion of the overall time they spent on the task (12.48% in post-editors vs. 18.84% in translators), meaning that post-editing had a more extreme impact on searching time than the overall time spent on the task. This could indicate that the use of MT can potentially decrease the need for searching. However, this may also be interpreted as a sign of over-reliance on MT, potentially hindering further recourse to resources and/or creativity.

The analytical quality evaluation revealed that post-editing resulted in higher scores.

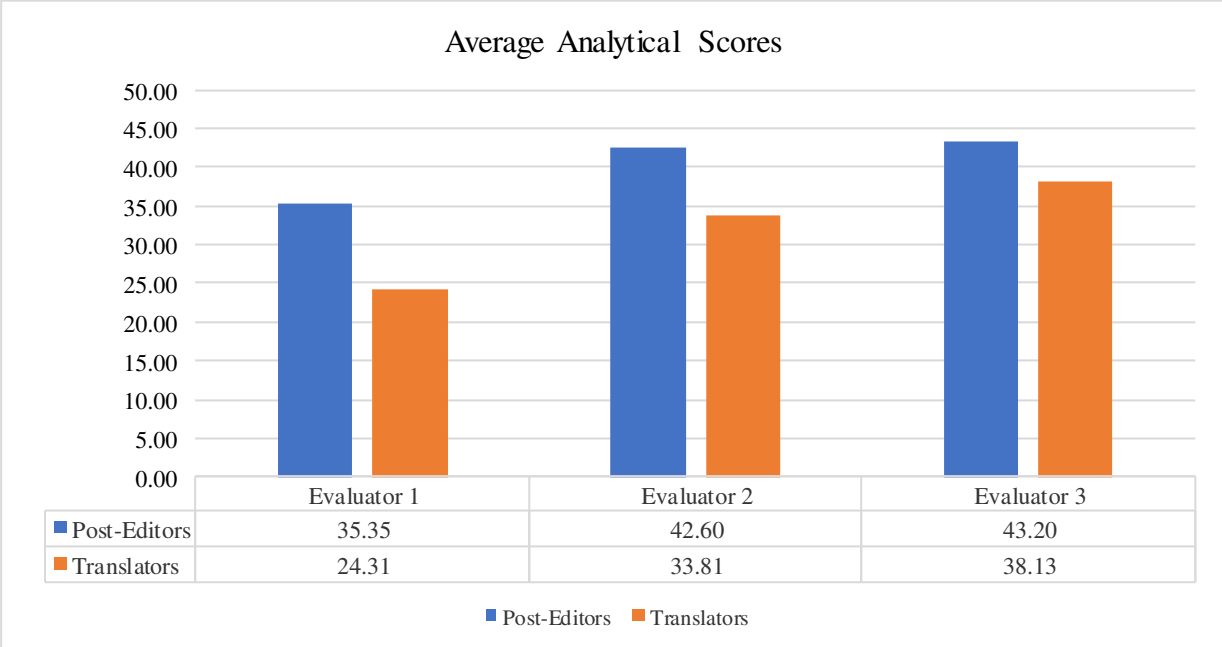

Figure 4. *Average Analytical Score Given to Each Group by the Three Evaluators*

As seen in Figure 4, the post-editors attained a significantly higher average score than the translators based on the analytical evaluation undertaken by all three evaluators.

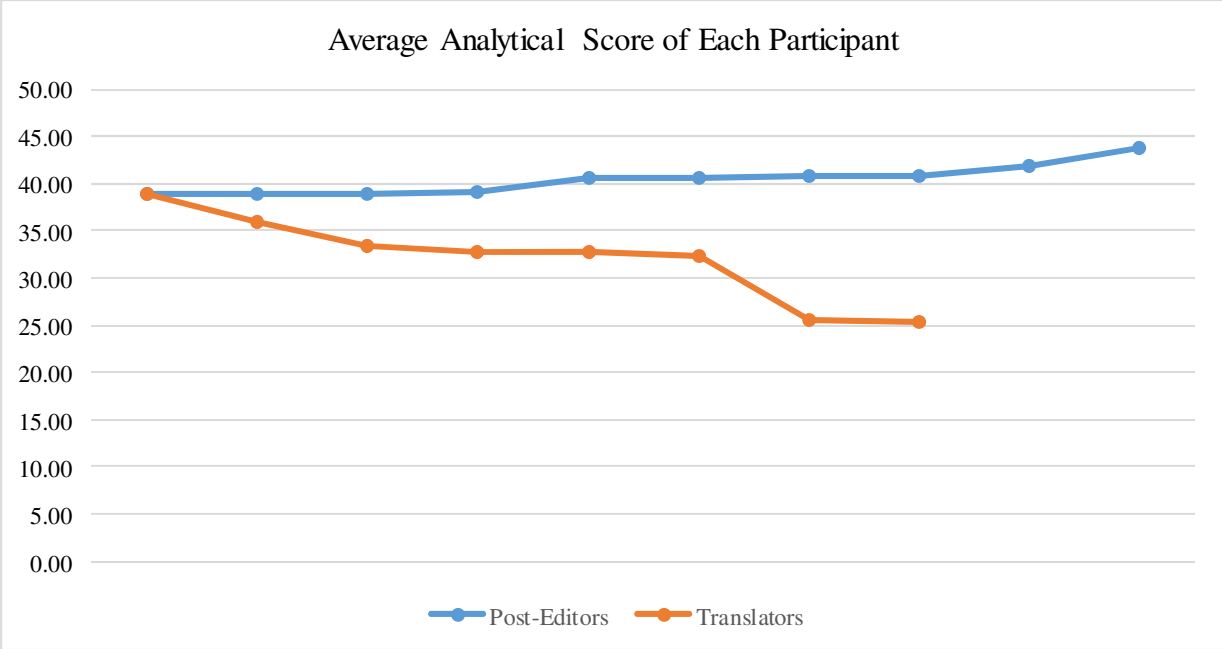

Figure 5. *Average Analytical Score of Each Participant*

As illustrated by Figure 5, the higher average analytical scores attained by the post-editors were not caused by outliers. In fact, comparing the average of the scores provided by the three evaluators for each participant, only one of translators attained a score that was not smaller than the lowest score attained by the post-editors. This indicates a stark and consistent difference that eliminates the need for testing statistical significance.

In order to ensure that the scores provided by the different evaluators had inter-rater reliability, the Pearson correlation coefficient of the different evaluators' scores were calculated for all the different pairings of evaluators. The lowest correlation coefficient score was 0.748 (with the other pairings scoring even higher), which indicates that the scores had high inter-rater reliability.

Generally, all three evaluators found significantly fewer instances of most error types in the post-edited TL texts compared to those translated from scratch.

The most notable exception to this was 'calque' which was more common in post-edited texts. This may have been caused by over-reliance on MT. Other exceptions were 'register', 'addition', and 'punctuation' errors, which had similar frequency in both groups.

The higher number of 'calque' errors in post-edited texts may be explained by the recurring nature of particular calque errors in post-edited texts, such as the example below:

> "El Nino is normally associated with *record breaking temperatures* at the global level.
>
> «ال نینو معمولاً *رکوردشکنی دمای* **متوسط** در سطح جهانی را به همراه دارد.

In this sentence, the MT had rendered 'record breaking temperature' as 'رکوردشکنی دما' (shown in italic) which can be considered 'calque'. Of the post-editors, 90% had preserved this rendering in their final file, while none of the translators had rendered this sentence in this way.

In the same example, 'متوسط', meaning 'average' was added to the text (shown in bold) by the participant and was not present in the SL text, which is an 'addition error'. While this particular rendering was written by a post-editor, the results suggest that post-editing is not likely to increase or decrease 'addition' errors, since addition errors did not originate from raw MT output.

> The last eight years were the world's *eight hottest* on record—reflecting the longer-term warming trend driven by greenhouse gas emissions.
>
> هشت سال گذشته، *هشت تا* از گرم‌ترین سال‌های ثبت شده در جهان بود که نشان‌دهندهٔ روند گرمایش طولانی‌مدت ناشی از انتشار گازهای گلخانه‌ای است.

In the example above, 'هشت تا' was identified as a 'register' error. Register errors were very few, and did not have much significance in the quality of the translations by either group.

World could face *record* temperatures in 2023 as El Nino returns

با بازگشت ال نینو، جهان ممکن است در سال ۲۰۲۳ با دمای ***بی‌سابقه ای*** روبرو شود

In this example, 'بی سابقه ای' is an instance of error in the use of zero-width non-joiner, one of the modern features of Persian orthography. This constitutes a punctuation error. The proper rendering would have been 'بی‌سابقه‌ای'. This error was present in the raw MT output and fully preserved in three post-edited texts, and partially preserved (as 'بی سابقه‌ای' and 'بی‌سابقه ای') in two. A similar rendering was only seen in one translated text, but this was mainly due to translators selecting a different wording for the corresponding English section in most cases. Even though recurring punctuations errors like this were observed in the raw MT output, the overall number of punctuation errors found in the TL texts were similar between the groups, indicating that MT is unlikely to cause a major disadvantage in punctuation.

Lastly, one of the three evaluators found more 'style' errors in post-edited texts, in contrast to the other evaluators. This discrepancy may be partially due to the effect of the evaluator's subjectivity in the identification and classification of errors.

> [A]lthough climate change *has fuelled* extreme temperatures even in years without the [El Nino] phenomenon.
>
> اگرچه تغییرات اقلیمی حتی در سال‌هایی که این پدیده رخ نداد هم *باعث* افزایش دمای شدیدی *شده بود*.

For instance, the rendition of 'has fuelled' as 'باعث ... شده بود' in the example above was identified as a style error by the evaluator in question but as a grammar error by another evaluator.

A notable observation about errors was that recurring errors were a more frequently occurring phenomenon in post-edited texts, especially in the case of calque errors where this was observed with the most frequency, such as the 'calque' instance discussed above. These recurring errors were often errors that were present in the raw MT output which were then kept in the post-edited text. It is worth noting that the texts which were translated from scratch contained more errors overall, but these errors were often unique errors, and the recurring errors were fewer in texts translated from scratch.

The holistic scores followed a similar pattern. The scores were once again higher in the post-edited texts.

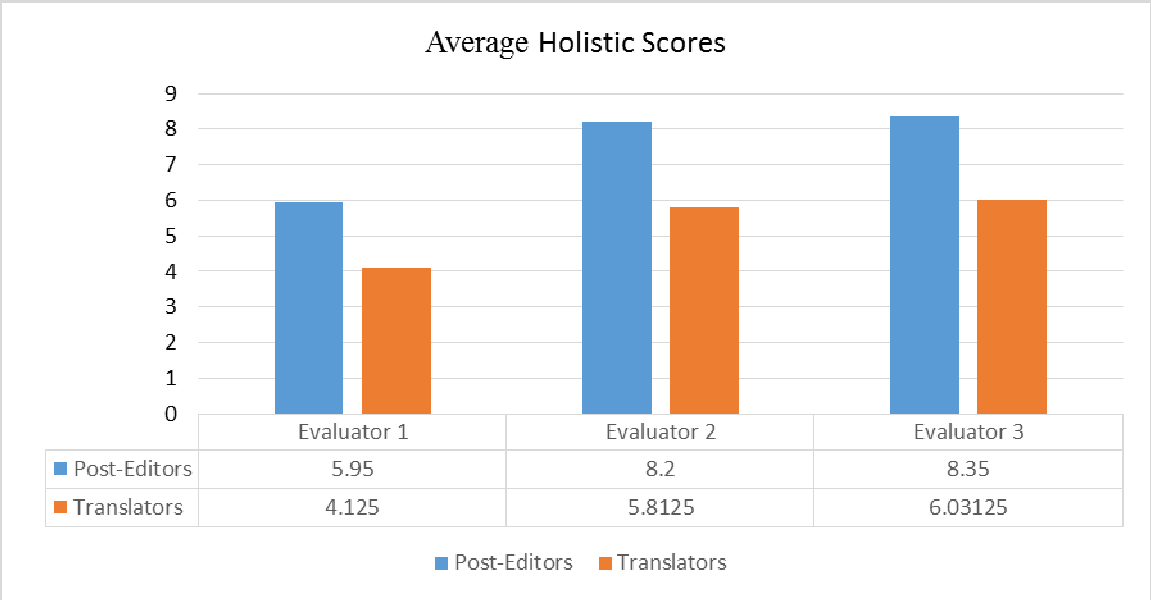

Figure 6. *Average Holistic Score Given to Each Group by the Three Evaluators*

As shown in Figure 6, the post-editors attained a significantly higher average score than the translators based on the holistic evaluation undertaken by all three evaluators.

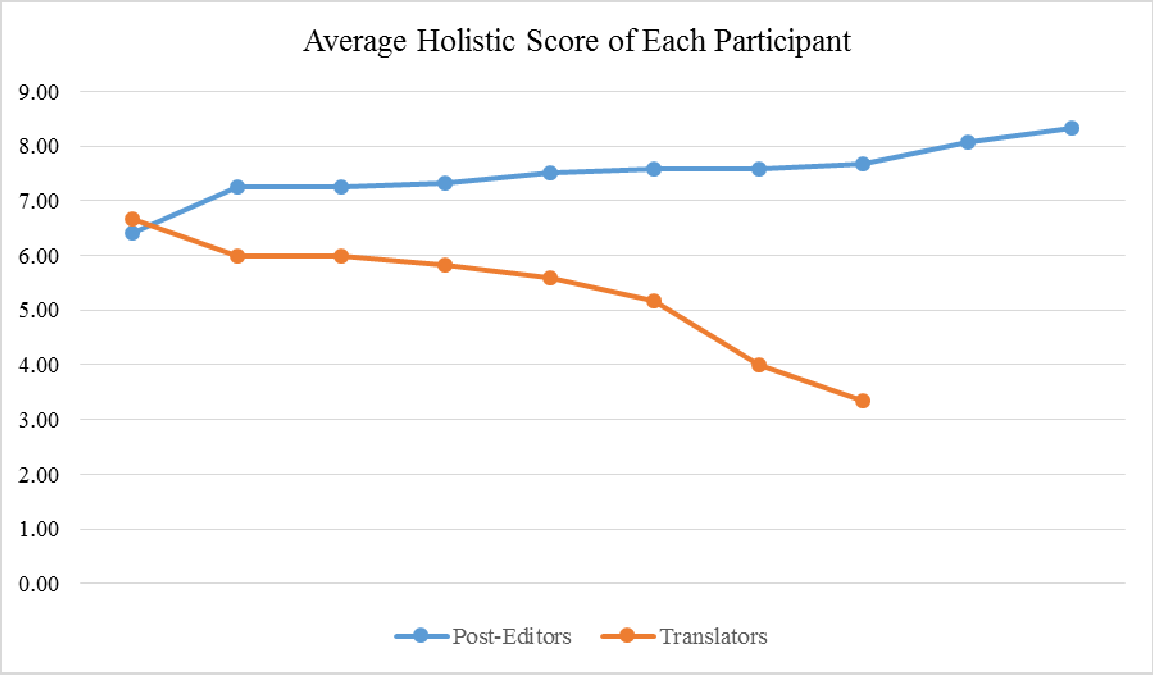

Figure 7. *Average Analytical Score of Each Participant*

As indicated in Figures 7, the higher average holistic scores attained by the post-editors were not caused by outliers either. Only one of the translators outscored the lowest grade attained by the post-editors. As in the case of the holistic score, the difference observed between the two groups are so stark and consistent that no further testing is required for evaluating statistical significance.

Pearson correlation coefficient of the scores provided by the different evaluators' scores were calculated for the holistic scores, and it was found that the lowest correlation coefficient was 0.762, with other pairings scoring higher, which indicates high inter-rater reliability in the holistic scores.

The holistic descriptions provided by all three evaluators indicated that post-edited TL texts were more fluent and accurate on average.

The methods used to compare the quality of translations produced by the post-editors and the translators both indicated that post-editing was superior in terms of quality. However, it is also important to note that the participants of this study were B.A. translator training students, and the results of this study are only

applicable to novice translators and may not necessarily hold true for professional translators and different contexts.

As explained earlier, the evaluators were asked to guess which TL texts were translated and which were post-edited. Evaluator 1 was able to guess correctly in 72.22% of instances, Evaluator 2's guesses were correct in 66.66% of the instances, and Evaluator 3's guesses were correct in 27.77% of the instances. Evaluators 1 and 2 primarily based their guesses on the similarity of the post-edited texts, while Evaluator 3 based their guesses on the assumption that translations produced from scratch would be of a higher quality. The similarity noted by Evaluators 1 and 2 is also in line with the observation that the post-edited texts contained recurring mistakes often originating from raw MT output.

Based on the questionnaires filled out by the participants after the experiment, it was found that the translators found the text to be slightly more difficult than the post-editors. This could either be attributed to MT making the task easier for the post-editors or interpreted as the post-editors being more competent in general. However, even if this difference was to be solely attributed to the translational competence of the post-editors, the difference was too small to invalidate the results of the study.

The participants' perceived level of success was similar between the two groups, which contradicted the quality evaluations. However, the inherent subjectivity of perceived success means that the different participants' perceptions may not be necessarily comparable.

Among translators, the most commonly cited difficulty of the translation text was finding or choosing the 'proper' or 'correct' equivalents.

According to the post-editors, grammatical issues had the most significant effect on the temporal post-editing effort. In addition, two factors were also reported to affect temporal post-editing effort with similar frequency: 'finding the "right"

equivalents' and balancing the use of MT. Lastly, proper usage of zero-width non-joiner was reported to have an effect on temporal post-editing effort by only one participant.

The factor which was reported to have the most significant effect on the cognitive effort of post-editing was the concern that the use of MT could affect the post-editor's creativity. Additionally, three factors were also reported to affect the cognitive post-editing effort to a similar degree: 'finding the "right" equivalents', balancing use of MT, and producing an easily readable text as the final product. The first two of these factors were also reported to affect the temporal post-editing effort. Finally, several factors were only reported by one participant each, namely: fixing punctuation errors, ensuring all mistakes and problems are fixed without altering meaning, fixing incoherent sentences, and the concern that using MT might prevent translators from producing translations with the highest quality they are capable of.

5. Conclusion

The present study was able to demonstrate that post-editing has the potential to speed up the translation process and increase the quality of translations, at least for novice translators. In this regard, the findings of the present study are in line with other studies which found post-editing to be beneficial to novice translators (e.g., see Carl et al., 2011; Garcia, 2010; Jia et al., 2019; Lee & Liao, 2011; Plitt & Masselot, 2010). However, as noted before, it is crucial to note that the performance of novice translators is not necessarily generalizable to translators with higher levels of translational competence. The effect of experience and level of expertise on the benefits attained from post-editing warrant further research. The results also indicated that post-editors spent much less time consulting resources than translators, which could indicate that post-editing can decrease the post-editors' need (or their perceived need) for other resources.

It was also found that the factor which affected temporal and cognitive post-editing effort the most, respectively, were the grammatical issues present in raw MT output, and the concern that MT might limit the translators' creativity. Other factors reported by multiple participants included finding the right balance in the use of MT and 'finding the "right" equivalent' which were reported to affect both the temporal and cognitive aspect of post-editing with similar frequency; moreover, producing an easily readable text was also reported to cause cognitive effort with the same frequency.

Considering how effective post-editing was found to be in the present study, post-editing can be a significant topic for future research. The effect of other variables such as experience, the peculiarities of Persian language, as well as different text types on the post-editing process and product, and the pedagogical implications of post-editing are all relatively unexplored areas warranting more research.

The results also have implications in translator training. Training in the use of MT and post-editing rarely receives enough attention in translation teaching. The improvements brought by the use of MT and the fast pace of advancements in translation technology warrant a more serious consideration of post-editing in pedagogy.

Works Cited

Alves, F., Koglin, A., Mesa-Lao, B., Martínez, M. G., de Lima Fonseca, N. B., de Melo Sá, A., Gonçalves, J. L., Szpak, K. S., Sekino, K., & Aquino, M. (2016). Analysing the impact of interactive machine translation on post-editing effort. In M. Carl, S. Bangalore, & M. Schaeffer (Eds.), *New directions in empirical translation process research: Exploring the CRITT TPR-DB* (pp. 77–94). Springer International Publishing. https://doi.org/10.1007/978-3-319-20358-4_4

Alves, F., Szpak, K., Gonçalve, J., Sekino, K., Aquino, M., Castro, R., Koglin, A., & Fonseca, N. (2016). Investigating cognitive effort in post-editing: A relevance-

theoretical approach. *Language Science Press*, *2*, 109–142. https://doi.org/10.17169/langsci.b108.296

Aranberri, N., Labaka, G., Diaz de Ilarraza, A., & Sarasola, K. (2014). Comparison of post-editing productivity between professional translators and lay users. In S. O'Brien, M. Simard, & L. Specia (Eds.), *Proceedings of the 11th conference of the Association for Machine Translation in the Americas* (p. 20–33). Association for Machine Translation in the Americas.

Aziz, W., Castilho, S., & Specia, L. (2012). PET: A tool for post-editing and assessing machine translation. In In N. Calzolari, K. Choukri, T. Declerck, M. U. Doğan, B. Maegaard, J. Mariani, A. Moreno, J. Odijk, & S. Piperidis (Eds.), *Proceedings of the eighth international conference on language resources and evaluation* (pp. 3982–3987). European Language Resources Association.

Carl, M. (2012). Translog-ii: A program for recording user activity data for empirical reading and writing research. In In N. Calzolari, K. Choukri, T. Declerck, M. U. Doğan, B. Maegaard, J. Mariani, A. Moreno, J. Odijk, & S. Piperidis (Eds.), *Proceedings of the eighth international conference on language resources and evaluation* (pp. 4108–4112). European Language Resources Association.

Carl, M., Dragsted, B., Elming, J., Hardt, D., & Jakobsen, A. L. (2011). The process of post-editing: A pilot study. In B. Sharp, M. Zock, M. Carl, & A. L. Jakobsen (Eds.), *Proceedings of the 8th international NLPCS workshop* (pp. 131–142). Copenhagen Studies in Language.

Cumbreño, C., & Aranberri, N. (2021). What do you say? Comparison of metrics for post-editing effort. In M. Carl (Ed.), *Explorations in empirical translation process research* (pp. 57–79). Springer International Publishing. https://doi.org/10.1007/978-3-030-69777-8_3

Daems, J., Vandepitte, S., Hartsuiker, R., & Macken, L. (2017). Translation methods and experience: A comparative analysis of human translation and post-editing with students and professional translators. *Meta*, *62*(2), 245–270. https://doi.org/10.7202/1041023ar

Eyckmans, J., Anckaert, P., & Segers, W. (2009). The perks of norm-referenced translation evaluation. In C. V. Angelelli & H. E. Jacobson (Eds.), *Testing and assessment in translation and interpreting studies* (pp. 73–94). John Benjamins Publishing Company.

Garcia, I. (2010). Is machine translation ready yet. *Target*, *22*(1), 7–21.

Hutchins, W. J., & Somers, H. L. (1992). *An introduction to machine translation*. Academic Press.

Jia, Y., Carl, M., & Wang, X. (2019). How does the post-editing of neural machine translation compare with from-scratch translation? A product and process study. *The Journal of Specialised Translation*, 31, 60–86.

Koby, G. S. (2001). Editor's introduction: Post-editing of machine translation output: Who, what, why, and how (much). In H. P. Krings, *Repairing texts: Empirical*

investigations of machine translation post-editing processes (pp. 1–24). Kent State University Press.

Koehn, P. (2009). A process study of computer-aided translation. *Machine Translation*, *23*, 241–263. https://doi.org/10.1007/s10590-010-9076-3

Krings, H. P. (2001). *Repairing texts: Empirical investigations of machine translation post-editing processes*. Kent State University Press.

Läubli, S., Amrhein, C., Düggelin, P., Gonzalez, B., Zwahlen, A., & Volk, M. (2019). Post-editing productivity with neural machine translation: An empirical assessment of speed and quality in the banking and finance domain. In M. Forcada, A. Way, B. Haddow, & R. Sennrich (Eds.), *Proceedings of machine translation summit XVII: Research* (pp. 267–272). European Association for Machine Translation.

Lee, J., & Liao, P. (2011). A comparative study of human translation and machine translation with post-editing. *Compilation & Translation Review*, *4*(2). 105–149.

Nitzke, J., & Oster, K. (2016). Comparing translation and post-editing: An annotation schema for activity units. In M. Carl, S. Bangalore, & M. Schaeffer (Eds.), *New directions in empirical translation process research: Exploring the CRITT TPR-DB* (pp. 293–308). Springer International Publishing. https://doi.org/10.1007/978-3-319-20358-4_14

O'Brien, S. (2005). Methodologies for measuring the correlations between post-editing effort and machine translatability. *Machine Translation*, *19*, 37–58. https://doi.org/10.1007/s10590-005-2467-1

O'Brien, S. (2007). Eye-tracking and translation memory matches. *Perspectives*, *14*(3), 185–205. https://doi.org/10.1080/09076760708669037

Plitt, M., & Masselot, F. (2010). A productivity test of statistical machine translation post-editing in a typical localisation context. *The Prague Bulletin of Mathematical Linguistics*, *93*, 7–16. https://doi.org/10.2478/v10108-010-0010-x

Skadiņš, R., Puriņš, M., Skadina, I., & Vasiļjevs, A. (2011). Evaluation of SMT in localization to under-resourced inflected language. In M. L. Forcada, H. Depraetere, & V. Vandeghinste (Eds.), *Proceedings of the 15th annual conference of the European Association for Machine Translation* (pp. 35–40). European Association for Machine Translation.

Vieira, L. N. (2014). Indices of cognitive effort in machine translation post-editing. *Machine Translation*, *28*, 187–216. https://doi.org/10.1007/s10590-014-9156-x