Does a Formal Post-Editing Training Affect the Performance of Novice Post-Editors? An Experimental Study

Post-Editing Eğitimi, Acemi Post-Editörlerin Performansını Etkiliyor mu? Deneysel Bir Çalışma

Volkan Dede*
Independent Researcher
Elena Antonova-Ünlü**
Hacettepe University

Abstract
Machine translation systems led to the creation of a new role for translators: the post-editor. With the birth of neural machine translation systems, the demand for post-editing has been increasing in the recent years, and it has now become a common service given by language service providers and professional translators. Such a change in the landscape of the translation industry might evolve the translation training programs worldwide. It is still heavily discussed whether post-editing and translation skills overlap, and post-editing courses are now included into the curriculum by several translation departments. We set out to investigate whether post-editing training influences the performance of student post-editors in order to explore the necessary background and skills in post-editing tasks. We measured productivity parameters and quality of the final outputs produced by two groups of participants, one of which was previously trained on post-editing. Our results show that, the experimental and control groups did not differ significantly from each other in terms of productivity. There was also little to no difference when we evaluated the post-edited outputs produced by both groups against a reference text using automatic machine translation evaluation metrics. However, we detected a statistical significance between the groups when we analyzed the number of errors in the final output. The post-editors in the experimental group were more aware of the typical errors of machine translation engines.
Keywords: machine translation, post-editing, translator training, translation curriculum

Öz
Makine çevirisi sistemleri, çevirmenler için yeni bir rolün oluşumuna yol açmıştır: post-editör. Nöral makine çevirisi sistemlerinin doğuşuyla post-editing hizmeti için talep son yıllarda artmaktadır ve artık dil hizmeti sağlayıcıları ile profesyonel çevirmenler tarafından sağlanan yaygın bir hizmet haline gelmiştir. Çeviri endüstrisindeki bu değişim, dünya genelindeki çeviri eğitimi programlarında köklü bir değişime yol açabilir. Post-editing ve çeviri becerilerinin birbiriyle ne ölçüde benzeştiği hâlâ tartışmalıdır ve bazı çeviri departmanlarının müfredatına post-editing dersleri eklenmiştir. Bu çalışmada, post-editing projelerinde gerekli arka planı ve becerileri incelemek için post-editing eğitiminin öğrenci post-editörlerin performansını etkileyip etkilemediği araştırılmıştır. Biri post-editing konusunda eğitilen iki katılımcı grubunun

Çankaya University *CUJHSS* (ISSN 1309-6761), December 2022, 16/2:131-148
https://dergipark.org.tr/en/pub/cankujhss. DOI: 10.47777/cankujhss.1106326
Submitted: April 20, 2022; Accepted: October 10, 2022.
*Dr., Independent researcher; ORCID#: 0000-0002-9691-2391; volkandde@gmail.com
**Assoc. Prof. Dr., Dept of Translation and Interpreting, Hacettepe University
ORCID#: 0000-0001-8544-6500; elenaunlu@gmail.com

sunduğu nihai çıktıların kalitesi ve üretkenlik parametreleri ölçülmüştür. Sonuçlar, deney ve control gruplarının üretkenlik bakımından birbirinden anlamlı şekilde farklı olmadığını göstermiştir. Post-editing uygulanan çıktılar, otomatik makine çevirisi değerlendirme yöntemleri kullanılarak referans metinle karşılaştırıldığında da neredeyse hiç fark gözlenmemiştir. Fakat nihai çevirideki hata sayısı analiz edildiğinde gruplar arasında istatistiksel olarak anlamlı bir fark görülmüştür. Deney grubundaki post-editörler, makine çevirisi motorlarının tipik hatalarını daha kolay fark etmiştir.
Anahtar kelimeler: makine çevirisi, post-editing, çevirmen eğitimi, çeviri müfredatı

Introduction and Literature Review

Machine translation refers to the use of computers instead of humans for translation and has been around since the Cold War era, when first studies into automatic or mechanical translation between English and Russian were made. Since the development of the first machine translation engine that was able to translate 60 Russian sentences into English (Hutchins), several types of engines have emerged: rule-based, phrase-based, statistical, and more recently, neural machine translation.

While rule-based machine translation engines were based on language-specific grammar and syntax rules that were manually fed into the machine, statistical engines were trained on preexisting corpora of bilingual texts with hopes to produce outputs that sounded more human. This human-like fluency, however, was not fully achieved until the introduction of neural machine translation in 2016, first announced by Google (Le and Schuster), owing to the fact that the technology behind neural engines was able to mimic the human brain (Thames).

With the birth of neural machine translation, machine translation has become a reality in the translator's workspace. Machine translation and post-editing are being increasingly integrated into the workflows of translation agencies, and most agencies have begun to promote their post-editing services. Although earlier surveys by several institutions such as the Translation Automation User Society (TAUS) and American Translators Association (ATA) reported the uncommon use of machine translation post-editing services in the translation market (Six; *TAUS Research-Postediting in Practice*), a more recent survey conducted in 2015 by Common Sense Advisory reported that post-editing moved from eight position to seventh position among the services grown (as cited in Aranberri). There are more recent surveys about language service providers providing machine translation services or translators providing post-editing services but these reports are privately available to the members of organizations such as the ones cited above. However, it's no doubt that the birth of neural machine translation has evolved the perception of translation in a layperson's mind and the translation industry. In a popular online blog on language industry, *Slator*, Diño reported that research into machine translation systems was at the highest amount in 2018, which suggests a willingness of the industry towards improving machine translation systems and making them a conventional part of the translation workflow.

Post-editing is the act of editing/improving the machine translation output. It is not a new term as it was even mentioned in the infamous ALPAC report of 1966 (Şahin), which caused the studies on machine translation systems to halt in the United States because the report indicated that machine translation systems were unsuccessful and developing one would be more expensive than using human translators. Post-editing is only now becoming a common task for a translator, and research into various aspects of post-editing such as cognitive effort, editing time, or whether it is similar to translation or not, has skyrocketed. Although many researchers have demonstrated that post-editing differs from translation in many ways (O'Brien; Rico and Torrejón), others have suggested that the features of a post-editing task depend on many factors: the text type, the machine translation system used, the language pair, and the competence of the translator/post-editor (Aranberri).

The translator/post-editor is thought by the industry to be natural post-editors when it's not always the case. A study by Aranberri exploring first-time post-editors reported that translators who post-edit for the first time tend to over-edit the machine translation output and make preferential changes. However, many industrial guidelines for post-editing (Massardo et al.; *Post-Editing Machine Translation Training*) strongly recommend that a post-editor should use as much of the raw machine translation output as possible, or else it would be easier to translate it from scratch. TAUS's (Massardo et al.) much-referenced basic guidelines for post-editing are as follows:

> Guidelines for achieving quality similar or equal to human translation:
> - Aim for grammatically, syntactically and semantically correct translation.
> - Ensure that key terminology is correctly translated and that untranslated terms be- long to the client's list of "Do Not Translate" terms.
> - Ensure that no information has been accidentally added or omitted.
> - Edit any offensive, inappropriate or culturally unacceptable content.
> - Use as much of the raw MT output as possible.
> - Basic rules regarding spelling, punctuation and hyphenation apply.
> - Ensure that formatting is correct. (Massardo et al. 17)

However, these guidelines can be regarded as too vague (Aranberri) or in some cases, they can be too detailed (Allen). Such guidelines are commonly used with specific "task descriptions" for each project, analyzing the quality of the machine translation output and skimming it for general errors that repeat through the text – so that the translator is able to pay their attention to these errors. These task descriptions also include the client requirements such as client-specific style or terminology. Furthermore, it's advised to train the post-editors because a regular translator unaware of what a typical post-editing process entails would be unsuccessful during such a task (as shown in Aranberri). An analysis by Rico and Torrejón proposes three main categories of competences required for a successful post-editor:

Linguistic skills:
Communicative and textual competence in at least two languages and cultures
Cultural and intercultural competence
Subject area competence
Instrumental competence:
MT knowledge
Term management
MT dictionary maintenance
Basic programming skills
Core competences:
Attitudinal or psycho-physiological competence
Strategic competence (Rico and Torrejón 169)

Such a categorization can be taken as a basis in training translators and/or translators-to-be for machine translation post-editing processes. Thus, it's clear that machine translation and post-editing competences should be added to the curricula of translation departments as it's becoming a reality and a new role for the contemporary translator today.

O'Brien was the first one to suggest a course content for post-editing teaching. The paper, published in 2002, outlines the main competences a post-editor should have, much like the one above, and suggests an outline for a syllabus of such a module. O'Brien maintains that a good post-editor would double or triple their daily translation output by post-editing machine translation. It's also put forward in the paper that teaching post-editing would make the translators give up their negative attitudes towards machine translation and embrace it. O'Brien's paper further indicates that post-editing differs fundamentally from traditional translation and it may even be possible for non-translators to become post-editors. Post-editing does not only differ from translation itself but also from editing or revision as the errors made by a machine and a human will differ considerably. It can also depend on the type of machine translation system used, and at the time of the aforementioned paper, neural machine translation systems were not a reality. It's repeatedly indicated that errors of neural machine translation are much more ambiguous, hidden, and slier than that of the other systems, statistical and rule-based machine translation. The skills proposed by O'Brien for a successful post-editor adds to the above ones the following: pre-editing/controlled language skills (to make the text suitable for the machine translation system in order to get a much more accurate raw machine translation output). O'Brien's suggestions for a module on post-editing consists of theoretical and practical ones, the former including introduction classes to post-editing, machine translation technology, controlled language, terminology management, linguistics, and programming skills while the latter includes practical post-editing courses using different text types and machine translation systems. It's also proposed to include this module "in the last part of an undergraduate translator training programme, or, even more ideally, in a post-graduate programme" (O'Brien 105) as post-editing would require advanced translation skills.

O'Brien's unique research has been followed by few papers so far. Post-editing training is unfortunately not a topic of interest among the researchers. It's only in the recent years that researchers have begun to focus on how to teach post-editing to students and professional translators. Until then, the teaching of post-editing had been included into translation technology and machine translation classes (Kenny and Doherty; Gaspari et al.; Austermuehl; Balkul).

Depraetere's 2010 paper titled "What counts as useful advice in a university post-editing training context? Report on a case study," aimed to identify the post-editing guidelines that need to be highlighted in a teaching context. The researcher, using the aforementioned post-editor skills proposed by O'Brien, identifies the basic competences a post-editor must have and asks trainee translators to post-edit a text for analysis. The researcher emphasizes that the post-edited text is more similar to the source text compared to a human translation, which is in contrary to studies demonstrating the higher quality of post-edited texts against human translations. In Depraetere's context, the students abided by the post-editing guidelines and did not attempt to over-edit the text, yet some students failed to notice some significant errors in the raw machine translation output. This paper differs from other similar ones as the students enrolled in this study were able to strictly follow the guidelines and use as much of the raw output as possible despite the fact that it resulted in less-than-perfect target texts. The author attributed this to the lack of experience on the students' part. In the conclusion part, the author addressed the aspects of machine translation that needed to be taught to students such as the typical errors made by a given machine translation system (as stated above) and she warned against the possibility of students trusting the machine translation engine too much.

Another similar study conducted later by Koponen reported on the experiences gained by a teacher and students from a machine translation and post-editing course offered at the University of Helsinki. Emphasizing that some of the post-editing skills are shared with "traditional human translation, such as source and target language proficiency, subject area knowledge, text linguistic skills, cultural and intercultural competence, as well as general documentation and research skills," (Koponen) the author once more added that the task of post-editing differed from traditional translation and revision processes and suggested that there were skills that were specific for post-editing tasks. In this paper, there are also new additions to the aforementioned post-editing skills: the skill to "learn to learn" as suggested by Pym (as cited in Koponen) or "learn how to pick up any new software quickly," which means that it's necessary for a post-editor to evaluate the machine translation software offered. The ideal post-editor should also be able to quickly evaluate the usability of machine translation outputs as it will greatly affect their productivity. Koponen's course on post-editing focuses on the history and theory of machine translation systems and post-editing, controlled language and pre-editing, post-editing guidelines, machine translation quality evaluation, and post-editing skills. An interesting addition in this course is the use of post-editing without a source text, which can be regarded as unlikely in a regular translation workflow as post-

editing inherently requires the presence of a source text, or it's basically editing. Based on the reflective essays evaluating the course written by the students, Koponen concludes that students tended to have negative attitudes towards machine translation, but the course turned their perception of machine translation into a positive one. With this paper, the above-mentioned notion by O'Brien (2002) that a course on machine translation and post-editing would make translators embrace machine translation seems to be proven.

In the context of Turkey, there have only been two researchers studying the use of post-editing in Turkish at the time of writing this paper: Temizoz and Şahin. In his paper, Şahin reports on a quantitative study exploring the use of machine translation post-editing for a subject-specific translation course. Şahin's is the first paper investigating post-editing teaching in a Turkish translator training context. This study included 15 senior (fourth year) translation students, who did not have much post-editing experience before the class, from a private university in Turkey. The author used the basic guidelines proposed by TAUS as the guidelines to be used by the trainees. Şahin's work on post-editing is unique in that it also explores how background research before the post-editing task could affect the understanding of the text to be post-edited. The background research would allow the post-editor to easily detect the mistakes and thus, can be an essential part of a machine translation post-editing course. According to the survey results, the students in this study were frustrated by the post-editing task at first but through practice, they became accustomed to it, which again proves that the negative perceptions towards machine translation could be eliminated by integrating these concepts into translator training. The final conclusion of this study was that the quality of the post-edited and translated texts were no different from each other, which is similar to what has been reported by similar studies (Depraetere et al.; Daems et al.) demonstrating that the difference in quality tends to be minimal.

Although not explicitly focusing on the training of post-editing skills, Temizoz's article is also worth mentioning for it may be the only article exploring the productivity and quality of post-editing in the Turkish context. The author investigates whether professional translators and subject-matter experts who carry out translation tasks perform differently for the post-editing of a technical text. The findings of this study indicate that post-editing quality is similar between translators and subject-matter experts. It is also demonstrated that the engineer-translators enrolled in the study performed better with regard to terminological choices. The author concludes that although a degree in translation studies does not necessarily mean a higher quality post-edited text, expertise in the subject matter is a critical factor for post-editing quality. Temizoz's article adds to the above-mentioned ones which report insignificant results with regard to post-editing quality.

Although the articles in the literature, particularly in Turkey, are all unique in that post-editing is a particularly under-researched area, nearly all of them were conducted before the birth of neural machine translation, which fundamentally changed the translation industry and the translation/post-editing practice itself,

and thus should be studied separately. As mentioned before, these papers report contradictory results in terms of post-editing effort and the behaviors of post-editors, which could be attributed to the use of non-neural machine translation systems in particular language pairs, especially in the case of English-Turkish.

The aim of this experimental study is to determine whether a formal training on post-editing is enough at the undergraduate level by way of examining the post-editing effort and time as productivity parameters and quality of first-time post-editor students selected from a translation department in Turkey. This study also produces valuable results for a particularly under-researched language, Turkish, in terms of machine translation and post-editing, despite the advances of popular machine translation systems in this relatively free-structured language.

Hypotheses

Three different hypotheses were tested in the present study:

1. There would be significant differences between the treatment and control groups with the treatment subjects performing better in terms of productivity.
2. The treatment group would be more successful at identifying and correction errors while the control group would tend to trust the machine translation output more as judged by the total number of errors left in the final translation.
3. The quality of the post-edited texts by the treatment group would obtain better results in traditional automatic machine translation evaluation scores than the control group.

Methodology

Either to validate or reject the above-stated hypotheses we performed an experiment where two groups of undergraduate students post-edited a technical text with the treatment group getting a brief training on machine translation systems and post-editing.

Participants

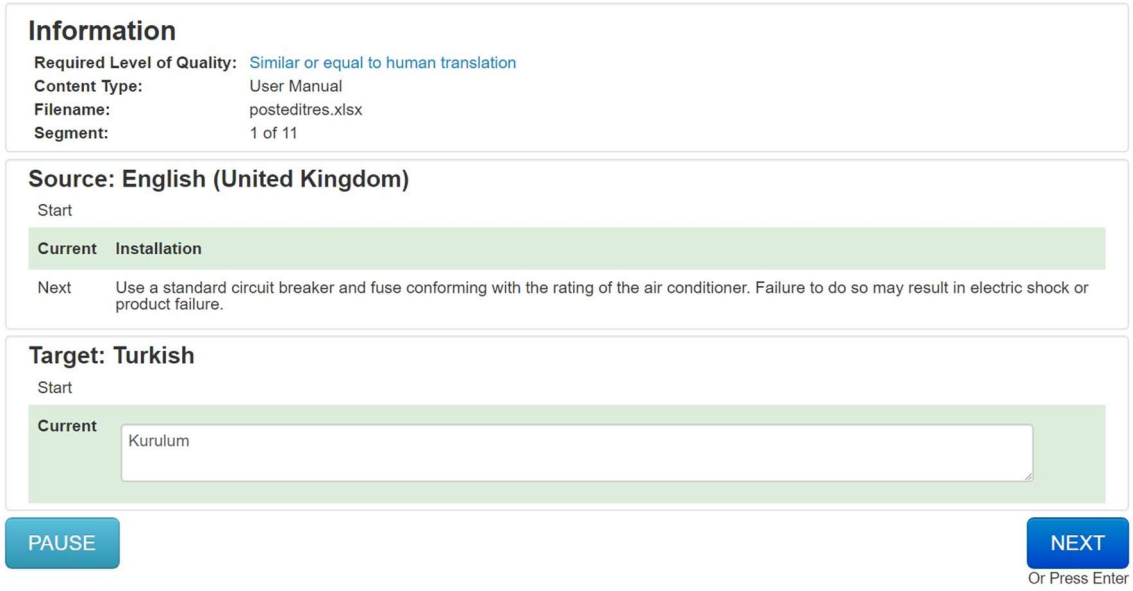
The participants of the present study were chosen from the students taking the Editing and Proofreading on Translation course offered at the Department of English Translation and Interpretation of Hacettepe University. A total of 23 students were present at the time of the first part of the study, which consisted of a survey exploring the background of the students with regard to their academic success, professional translation and post-editing experience, knowledge of and attitude towards machine translation. The survey detailed the purposes of the study and featured a consent part where the students agreed to take part in both parts of the study (the questionnaire and the post-editing task). In the end, there were a total of 20 students who gave consent to participating in both parts and who eventually comprised the sample of the present study.

Following the completion of the questionnaire, the students were instructed on how to use the online system where the experiment would be conducted. Then, half of the students (10/20) were randomly assigned to the control group. The remaining students, comprising an experimental group, listened to a brief course on machine translation systems and post-editing, which provided a general overview of machine translation systems, post-editing and related guidelines, and a step-by-step approach on how to perform post-editing. The students were instructed to post-edit according to the TAUS guidelines for "achieving quality similar or equal to human translation," (Massardo et al.) which has been cited above.

Online system

The Dynamic Quality Framework platform provided by TAUS[1] was used as the online tool where the post-editing task would be carried out (Figure 1). The participants were instructed beforehand on how to use the tool and what they should or should not do. For instance, if they had to leave their computer in the middle of the task, they were told to use the pause feature and resume the task at a later time. Use of online tools such as dictionaries was permitted.

Figure 1. Screenshot of the post-editing task performed on the TAUS DQF system.



This publicly available tool was chosen for its easiness of use and its statistical features. The tool has three main task types, which are productivity, quality evaluation, and ranking engines. The productivity feature was used to test the productivity hypothesis for which the post-editing time and effort of all students participating in the second part of the study were recorded in real time. The productivity feature demonstrates these two parameters in seconds (time) and percentage (effort). The percentage expresses how much effort was required to edit the machine translation output with 0% representing that no effort (no change) was needed.

[1] For more information about the DQF tool please visit: http://dqf.taus.net

The quality analysis feature was utilized for the first part of the quality analysis, which is detailed below.

Machine translation engine

The text that was used for the post-editing task was translated using Google's public translation engine, Google Translate. As this engine is one of the most popular machine translation engines in Turkey, no particular analysis was deemed necessary. Furthermore, Temizoz's study chose to use Google Translate after conducting a quality analysis. Other tools supporting the language pair of English-Turkish tend to provide poorer results compared to the outputs produced by Google Translate. Google Translate was also the most commonly used translation engine by the participants according to the survey results.

Technical text

A publicly available manual for air-conditioners was chosen to be translated and post-edited. The source text was in English and contained 239 words divided into 11 segments. Special attention was paid to the fact that the sentences selected from the user manual contained minimal amount of terminology and would lead to undesirable machine translation outputs. When a given segment led to a high-quality translation, the source text was slightly manipulated. There was also a typo in one of the original sentences which was not edited to see its effect on the translation and the post-editing performance of the students.

Quality analysis

Machine translation outputs are usually evaluated by human evaluators or automatic machine translation evaluation metrics. We used both methods in our study for different purposes.

Firstly, the total number of errors in the source text and the target texts produced by the participants was assessed. The DQF tool was utilized for its quality evaluation feature. This feature allows the researchers to identify the number of errors in a given text and classify it according to TAUS's own error typology called MQM (*Harmonized DQF-MQM Error Typology*). The main types of error include accuracy, fluency, style, and terminology. The rest (e.g. design or formatting errors) were deemed unrelated to the type of errors that could be observed in the present study and therefore excluded.

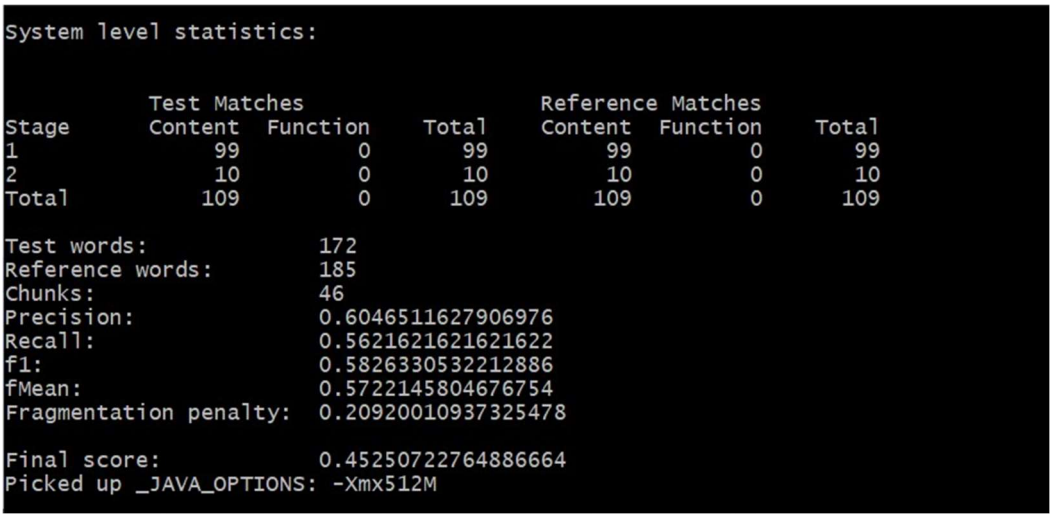
For the automatic evaluation part, we used several different Java or Python-based software. It is important to note that Turkish outputs compared using these tools tend to get lower scores due to the nature of the Turkish language. The main logic behind these metrics is that they calculate the number of matches between a machine translation output (also called hypothesis or candidate) and a reference translation of the same source text. All work different from each other with regard to how they calculate these matches and express the scores. For example, BLEU (Papineni et al.) is the metric that is widely used in the industry. It calculates each n-gram match and sequence of n-gram matches. Thus, for a higher score to be obtained, the machine translation output would have to follow the same sequence of words as in the reference text.

The other scores used for analysis in the present study were TER (Translation Edit Rate) (Snover et al.), METEOR (Denkowski and Lavie), and CharCut (Lardilleux et al.). TER measures the number of edits (deletions, insertions, and substitutions) that have to be made in the machine translation output to reach the reference text. METEOR, on the other hand, works in a similar way to BLEU but it can take into account exact matches as well as stem words, function words, synonyms, and paraphrases. METEOR supports several languages, including Turkish with limited capacity (i.e., it supports only stem words). We assumed that METEOR would provide better results compared to BLEU and TER thanks to its additional capabilities and its suitability for languages like Turkish. Finally, CharCut, a machine translation metric measuring character-based matches that has strong correlations with human evaluator judgements, was also employed to evaluate the quality of the machine translation output and final translations in the present study.

For two of these four metrics, a higher score/percentage means a higher quality text (BLEU and METEOR) whereas for TER and CharCut a higher score/percentage indicates a higher amount of difference between the candidate and reference, and hence, lower quality.

To calculate these scores, we utilized the Java-based software, multeval[2] (Clark et al.) for BLEU and TER and the original repositories for CharCut[3] and METEOR[4] (Figure 2). All entries were tokenized and lowercased, and punctuation was disregarded and/or manually removed during the analysis.

Figure 2. Screenshot of a METEOR analysis.



Statistical analysis

Two-sample t-test was employed to see if there were any statistical differences between the experimental and control groups in terms of time spent during post-editing and edit effort. Statistical significance threshold was set as $p < 0.05$. All statistical analyses were conducted in R (R Core Team).

[2] https://github.com/jhclark/multeval
[3] https://github.com/alardill/CharCut
[4] https://github.com/cmu-mtlab/meteor

Results and discussion

Findings of the questionnaire

A total of 20 students completed the survey, the first part of the study. 45% of the participants were female and 55% were male. With the exception of 1 student (fourth-year), the remaining students were studying their 3rd year at the department. 63.2% of the students had a GPA in the range of 3.00-3.50 while the rest had a GPA below 3.00 out of 4.00, which suggested that the majority of the students were academically successful. As the editing class was an elective course, it could feature students from different years with different experiences and training, therefore, it was necessary to investigate if they had taken the technical translation course offered during the sixth semester (third-year) but we found that only 2 students had taken the course. Except for 1 student, none of the students were providing professional translation services or had previously conducted post-editing. When asked if they heard the term "post-editing" before, the majority of the participants (57.9%) answered yes while a considerable number of students hadn't heard of post-editing (42.1%). Despite their unfamiliarity with post-editing practices, a staggering 75% of the participants reported using a machine translation engine. However, four of the students entered online dictionary names (e.g. Tureng, Zargan) when they were asked to name the machine translation engines they commonly used, which suggested that there was some confusion among the students with regard to the concept of machine translation. It is also important to note that there are currently no technology courses offered, focusing on machine translation technologies, at the department at the time of writing this paper. Still, 14 students reported the use of Google Translate with one student also indicating the use of MateCat, an open-source web-based computer assisted translation tool supported with Google's technologies (including Translate).[5] To determine the attitude of the students towards machine translation technologies, we explored if they thought that machine translation had the potential of replacing human translators in the future. While 30% of the students reported that they did not think machine translation would ever replace human translators, 40% highlighted the importance of translators catching up with the new advances in the industry or machine translation technologies could pose a threat for them. The remaining 25% indicated that machine translation would replace them in some areas but not all while one student indicated that translators should not use machine translation technologies at all. The final question revealed that 85% of the students were not satisfied with the current curriculum offered at the department – specifically, they did not think that the bachelor's programme was consistent with the current developments in the translation industry.

Comparison of post-editing time and effort

For the post-editing task, we first had to evaluate the technical text we were going to use for the post-editing practice. We used the aforementioned automatic quality analysis procedures using the publicly available Turkish

[5] For more information on MateCat please refer to https://www.matecat.com.

translation of the source text as the reference. In the end, the machine translation evaluation metrics reported similar results with a BLEU score of 25.3 (Table 1), consistent with the only study on post-editing in the Turkish language providing such a score (Temizöz). The scores in Table 1 indicates that the machine translation output was of poor to moderate quality with little similarity to the reference text. However, the scores should not be interpreted literally as they are dependent on many factors, including the language of the target text. As we assumed, METEOR provided a better result for the machine translation output, indicating moderate quality.

Table 1. Quality of the machine translation output as assessed by common machine translation evaluation metrics.

Metric	Score for the machine translation output
BLEU	25.3
METEOR	45.25
TER	58.9
CharCut	35%

A total of 11 students finished the second task in the study despite the initial number of 20 students who completed the survey: 6 in the experimental group and 5 in the control group. We used the measurements provided by the TAUS DQF tool for the statistical analysis of any differences between the two groups.

We expected intragroup consistency with regard to post-editing time and effort. Although the majority of the subjects included in the test seemed to have performed consistently with the rest within each group, there were outlier subjects who finished the task too quickly or too slowly (Table 2). For the time parameter, the t-test showed no statistical significance with a p-value of 0.46 ($t = 0.78$; $df = 5.69$).

Table 2. Time spent on post-editing.

Participant	E1	E2	E3	E4	E5	E6	C1	C2	C3	C4	C5
PE time (sec)	674	748	688	1541	2316	712	1165	1184	674	1147	3985

(E: Experimental; C: Control)

Despite the insignificance, the results reveal that students with prior post-editing tend to take less time with the post-editing task, which indicates higher productivity. However, we expected the contrary with the control group taking much less time as they would trust the machine translation engine more. Still, if we were to judge these results with regard to a traditional understanding of productivity, the experimental group would be considered more productive and the post-editing training would prove to be effective. We attributed this

difference to the fact that treated students' familiarity with the task and knowledge of machine translation may have given them confidence in performing the task. Due to the lack of any post-test communication, we cannot make definitive suggestions related to the background of the performance. We cannot directly compare our results with the literature as no such experimental study has been conducted but other studies (Garcia) have reported insignificant results with regard to post-editing time when compared with translation from scratch.

We also assumed that there would be significant differences between the groups with regard to edit effort (average number of changes made). We collected the mean effort per segment for each participant, a higher percentage indicating more effort, then calculated the mean effort of both groups per segment (Table 3).

Table 3. Average edit effort per segment (%).

Segment	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**Mean**
Experiment	32.33	10.16	30.83	31.33	29.33	63.16	24.83	39	14.66	19.66	21	28.75
Control	0	14.8	13.2	15.8	53.4	64.2	14.8	38.2	10.2	22.2	24.8	24.69

Overall, we observed a higher amount of effort in the experimental group with more instances of zero effort in the control group. However, the difference was not statistically significant ($p = 0.58$; $t = 0.55$; $df = 18.27$). Still, the small difference in means in favour of the experimental group might indicate that the trained group post-edited the text more thanks to their prior knowledge of the typical errors of a neural machine translation system. On the other hand, the control group may have trusted the machine translation result more.

The inconsistent results and the insignificant differences may have stemmed from a sampling error, where more students with a GPA below 3.00 were randomized to the control group (n=3) while there was only one (n=1) such participant in the treatment group. Overall, we could not determine a correlation between any of the variables (machine translation usage, having taken the technical translation course, or academic year) and the outlier results in each group.

Our analysis indicates that post-editing training did not significantly affect the performance of first-time post-editors with regard to edit effort and speed. Both groups perform with little difference, which could suggest that traditional translation and editing (the students in our sample received an editing class) skills are sufficient for post-editing tasks. Nevertheless, our small sample constitutes limitations against making any conclusive suggestions. Our results still imply that there is a small difference between groups for each task with a higher productivity indicated for the treatment group who received post-editing training.

Quality analysis

As previously mentioned, we applied two separate quality analysis methods to determine the quality of the final outputs. First, we calculated the number of errors in the machine translation output and the final translations produced by the participants and compared them. Then, we investigated if there was any difference between the outputs of the two groups according to the scores obtained from the automatic machine translation evaluation metrics.

Number of errors

In the original source text, there were a total of 21 errors in four categories: accuracy (10) [mistranslation, addition or omission, etc.], fluency (2) [spelling, punctuation, etc.], terminology (3) [wrong terminology or terminology inconsistent with the company guidelines, the latter not considered in the present study], and style (6) [awkward style, localization errors etc.]. We then used the same quality analysis tool to determine the number of errors left or added in the post-edited outputs (Table 4).

Table 4. Total number of errors in the final output per participant.

Participant	E1	E2	E3	E4	E5	E6	C1	C2	C3	C4	C5
Total errors (n)	4	8	9	9	10	10	15	15	17	9	18

The results indicate that the outputs produced by the control group are high in number, confirming our second hypothesis that the control outputs would be lower in quality compared to the experimental group as judged by the number of errors identified in the final text. The t-test also returned significant results, indicating that the difference was not due to mere coincidence with a p-value of 0.01 ($t = 3.56$; $df = 6.61$), consistent with similar studies (Garcia).

It seems that the experimental group identifies more errors and post-edits the text more to the human quality compared to the control group, consistent with the higher edit effort mentioned above. Therefore, we can argue that it is important to teach students the typical errors produced by machine translation engines and the common guidelines on post-editing, which instruct the post-editor on how to proceed in a typical post-editing task. The experimental group appears to have complied with the TAUS guidelines on post-editing for human quality while the control group left more errors unedited. During the manual analysis, we observed some instances of introducing errors that were not in the original text, which may be attributed to the unadvanced translation skills. Similarly, there is still a large number of errors left in the outputs produced by the experimental group, so the final translation is not perfect or close to human quality. It was unfortunate to see simple spelling or punctuation errors introduced into the target text. Awkwardness in terminology or style was not post-edited as much as we expected. For instance, "conflict" between neighbors was translated as "*çatışma*," which is much more aggressive than the desired term "*anlaşmazlık*," but most students left the original translation as it is. However, more obvious awkward translations (e.g. one word was translated as a swear word) were edited.

Quality according to evaluation metrics

We wanted to use the evaluation metrics typically used for the evaluation of machine translation outputs to compare the results of the two groups. As previously mentioned, these metrics do not provide definitive outcomes but rather illustrate the general quality of a target text compared to a reference text. We used the publicly available reference text to compare the post-edited outputs. The tool used for BLEU and TER scores (multeval) had the capacity to compare multiple outputs. Therefore, in comparing the BLEU and TER scores of the outputs, we used the mean calculated by the multeval software (Table 5). However, for METEOR and CharCut, we had to individually run the software for each subject (Table 6).

Table 5. Average BLEU and TER scores for the final translations.

Group	BLEU	TER
Original output	25.3	58.9
Experiment	25.7	60.4
Control	27.0	60.2

Table 6. Average CharCut and METEOR scores for the final translations.

Participant	E1	E2	E3	E4	E5	E6	C1	C2	C3	C4	C5
CharCut	33%	34%	35%	35%	42%	43%	33%	34%	34%	37%	39%
METEOR	0.42	0.38	0.48	0.44	0.48	0.45	0.52	0.45	0.44	0.40	0.50

Note: Original CharCut and METEOR scores were 35% and 0.45, respectively.

Despite the statistical significance with regard to the number of errors in the final outputs described above, the automatic evaluation metrics did not detect any considerable difference between the outputs of experiment and control groups. Surprisingly, BLEU scores indicated a better-quality text for the control group, and for both groups, TER scores revealed that the text actually worsened compared to the reference text. The human evaluation did not detect any particular deterioration in the quality of the texts despite some additional errors introduced by the post-editors.

We expected METEOR and CharCut scores to be more accurate with regard to the actual quality of the text, particularly for the Turkish language. The mean scores for the experiment and control groups were 37.0% and 35.4% for CharCut and 0.44 and 0.46 and for METEOR. When compared with the original score, the CharCut average indicates a worse final output for both groups while the METEOR score has improved for the experiment group but deteriorated for the control group. However, we could not detect any statistical significance for any of the metrics ($p = 0.46$ for METEOR and $p = 0.46$ for CharCut).

During the manual evaluation in the first part, we observed some relatively free translations, particularly in the experimental group; however, it's not likely that a few free translations could lead to such inconsistent results. It is, however, worth mentioning that the reference text available online was probably translated by a professional translator, and we cannot reasonably expect 3rd-year translation students with limited skills to produce translations similar to that of a professional translator. As all of these metrics calculate the number of matches between two texts, it is likely that a more literal but correct translation might have been scored lower compared to the more client-specific and suitable reference text.

All in all, the findings obtained from the automatic evaluation software mostly indicated poorer results for both groups. Still, we found statistical significance with regard to the number of errors in the final translations.

Conclusion and Future Work

In this paper we set out to evaluate the effect of post-editing training on the post-editing performance of students in an experimental setting. Two of the three hypotheses were rejected: the experimental and control groups did not differ significantly from each other in terms of productivity when one of the groups was trained on post-editing and the typical errors of neural machine translation systems. There was also little to no difference when we evaluated the post-edited outputs produced by both groups against a reference text using automatic machine translation evaluation metrics. However, we detected a statistical significance between the groups when we analyzed the number of errors in the final output. The post-editors in the experimental group were aware of the typical errors of machine translation engines and were also instructed to post-edit according to TAUS guidelines. The control group, however, had a number of errors that was closer to the number in the original source text. This significant finding highlights the importance of post-editing training and is consistent with previous research in the field emphasizing the difference between post-editing and translation/editing. The main limitation of the present study includes the limited sample size. Future studies may include a larger sample size to validate our findings. We also need studies on the correlation of automatic metrics with human judgements in order to reliably compare the results we obtain. There is also little research on post-editing in the Turkish language; thus, studies exploring any aspect of post-editing in the Turkish context are very much welcome.

Works Cited

Allen, Jeffrey. "Post-Editing." *Computers and Translation: A Translator's Guide*, edited by Harold Somers, John Benjamins, 2003, pp. 297–318.

Aranberri, Nora. "What Do Professional Translators Do When Post-Editing for the First Time? First Insight into the Spanish-Basque Language Pair." *HERMES - Journal of Language and Communication in Business*, no. 56, 2017, pp. 89–110, https://doi.org/10.7146/hjlcb.v0i56.97235.

Austermuehl, Frank. *Future (and Not-so-Future) Trends in the Teaching of Translation Technology*. http://revistes.uab.cat/tradumaticaElscontingutsdela revistaestansubjectesaunallicènciaCreativeCommons. Accessed 23 Dec. 2018.

Balkul, Halil İbrahim. *Türkiye'de Akademik Çeviri Eğitiminde Çeviri Teknolojilerinin Yerinin Sorgulanması: Müfredat Analizi ve Öğretim Elemanlarının Konuya Ilişkin Görüşleri Üzerinden Bir Inceleme*. Sakarya University, 2015.

Clark, Jonathan H., et al. "Better Hypothesis Testing for Statistical Machine Translation: Controlling for Optimizer Instability." *Acl-2011*, 2011, pp. 176–81, https://doi.org/10.1057/dev.2008.5.

Daems, Joke, et al. "Translation Methods and Experience: A Comparative Analysis of Human Translation and Post-Editing with Students and Professional Translators." *Meta: Journal Des Traducteurs*, vol. 62, no. 2, 2017, p. 245, https://doi.org/10.7202/1041023ar.

Denkowski, Michael, and Alon Lavie. "Meteor 1.3: Automatic Metric for Reliable Optimization and Evaluation of Machine Translation Systems." *Proceedings of the Sixth Workshop on Statistical Machine Translation*, 2011, pp. 85–91, https://doi.org/10.1080/00288306.2004.9515087.

Depraetere, Ilse, et al. *Post-Edited Quality, Post-Editing Behaviour and Human Evaluation: A Case Study*. 2014, pp. 78-108., https://hal.archives-ouvertes.fr/halshs-01060447.

Diño, Gino. "Google, Facebook, Amazon: Neural Machine Translation Just Had Its Busiest Month Ever | Slator." *Slator*, 2018, https://slator.com/technology/google-facebook-amazon-neural-machine-translation-just-had-its-busiest-month-ever.

Garcia, Ignacio. "Is Machine Translation Ready Yet?" *Target*, vol. 22, no. 1, 2010, pp. 7–21, https://doi.org/10.1075/target.22.1.02gar.

Gaspari, Federico, et al. "A Survey of Machine Translation Competences: Insights for Translation Technology Educators and Practitioners." *Perspectives: Studies in Translatology*, vol. 23, no. 3, 2015, pp. 333–58, https://doi.org/10.1080/0907676X.2014.979842.

Hutchins, W. John. *Machine Translation: Past, Present, Future*. Ellis Horwood; Halsted Press, 1986.

Kenny, Dorothy, and Stephen Doherty. *Statistical Machine Translation in the Translation Curriculum: Overcoming Obstacles and Empowering Translators Statistical Machine Translation in the Translation Curriculum: Overcoming Obstacles and Empowering Translators*. no. April 2018, 2014, https://doi.org/10.1080/1750399X.2014.936112.

Koponen, Maarit. "How to Teach Machine Translation Post-Editing? Experiences from a Post-Editing Course." *Proceedings of 4th Workshop on Post-Editing Technology and Practice (WPTP4)*, 2015.

Lardilleux, Adrien, et al. *CHARCUT: Human-Targeted Character-Based MT Evaluation with Loose Differences to Cite This Version: HAL Id: Hal-01726326 C HAR C UT: Human-Targeted Character-Based MT Evaluation with Loose Differences*. 2018.

Le, Quoc v., and Mike Schuster. "A Neural Network for Machine Translation, at Production Scale." *Google AI Blog*, 27 Sept. 2016, https://ai.googleblog.com/2016/09/a-neural-network-for-machine.html.

Massardo, Isabella, et al. *MT POST-EDITING GUIDELINES*. TAUS Signature Editions, 2016.

O'Brien, Sharon. "Teaching Post-Editing: A Proposal for Course Content." *Proceedings of the 6th EAMT Workshop: Teaching Machine Translation*, European Association for Machine Translation, 2002, https://aclanthology.org/2002.eamt-1.11.

Papineni, Kishore, et al. "BLEU: A Method for Automatic Evaluation of Machine Translation." *40th Annual Meeting of the Association for Computational Linguistics*, 2002, pp. 311–18, https://doi.org/10.1002/andp.19223712302.

Post-Editing Machine Translation Training. https://www.sdltrados.com/learning/training/post-editing-machine-translation.html. Accessed 24 Dec. 2018.

R Core Team. *R: A Language and Environment for Statistical Computing*. 2018, https://www.r-project.org/.

Rico, Celia, and Enrique Torrejón. "Skills and Profile of the New Role of the Translator as MT Post-Editor." *Revista Tradumàtica: Tecnologies de La Traducció*, vol. 2012, no. 10, 2012, pp. 166–78, http://revistes.uab.cat/http://revistes.uab.cat/tradumatica.

Şahin, Mehmet. "Using MT Post-Editing for Translator Training." *Tralogy*, II: 6, 2011.

Six, Shawn E. *Summary of the ATA Translation and Interpreting Services Survey | The Chronicle*. 2014, http://www.atanet.org/chronicle-online/featured/summary-of-the-ata-translation-and-interpreting-services-survey/#sthash.6MTcHkO3.h9ME1ijM.dpbs.

Snover, Matthew, et al. "A Study of Translation Edit Rate with Targeted Human Annotation." *Proceedings of Association for Machine Translation in the Americas*, 2006, pp. 223–31, https://doi.org/10.1.1.129.4369.

TAUS Research-Postediting in Practice. 2010, http://taus-website-media.s3.amazonaws.com/images/stories/pdf/benchmarch-data-for-postediting-practices-globally.pdf.

Temizöz, Özlem. "Postediting Machine Translation Output: Subject-Matter Experts versus Professional Translators." *Perspectives: Studies in Translatology*, vol. 24, no. 4, 2016, pp. 646–65, doi.org/10.1080/0907676X.2015.1119862.

Thames, Jonathan. "Machine Translation." *LanguageSolutions*, 24 June 2019, https://langsolinc.com/machine-translation.