transLogos 2023 Vol 6 Issue 1
Ekinci, Sena, pp. 98–116
https://dx.doi.org/10.29228/transLogos.56

The Impact of Error Annotation on Post-Editing of Subtitles: An Investigation into Effort and Product

Sena EKİNCİ*

This paper aims to investigate the effect of error annotation on post-editing effort and post-edited product. The study also attempts to highlight the significance of quality evaluation, particularly error annotation, which, I believe, is a useful method for learning how to work with machine translation (MT). In order to accomplish these goals, ten translation students were divided into two groups—a control group and a treatment group—in an experimental study. The control group post-edited the machine-translated subtitles of an educational video while the treatment group performed a quality evaluation prior to the task of post-editing the same content. Temporal and technical effort data (Krings 2001) of students were gathered to measure whether there was a significant difference between the two groups. In addition, the end products were examined to see if quality evaluation had a different impact on the post-editing decisions of the treatment group compared to the control group. The results show that there is a significant difference in temporal effort between the two groups—the treatment group completing the post-editing task faster—and the control group expended more technical effort than the treatment group, though the difference was not significant. The treatment group also displayed a tendency to use MT and edit more efficiently than the control group.
Keywords: machine translation; post-editing; quality evaluation; error annotation; subtitling

1. Introduction

Even though the history of machine translation (MT) dates to the development of computers in the 1940s, it has only recently become widely used in both professional and non-professional contexts. Language service providers (LSPs) have incorporated machine translation post-editing (MTPE) into their processes with the rise of neural machine translation (NMT) in recent years. A prevalent method in the translation industry today is post-editing, which speeds up the translation process since MT produces high-quality translations in some industries, particularly in the technical domain.

* Research assistant at Istanbul 29 Mayıs University.
E-mail: sekinci@29mayis.edu.tr; ORCID ID: https://orcid.org/0000-0003-1358-9443.
(Received 9 January 2023; accepted 10 June 2023)

On the other hand, the growth of online platforms, including those for entertainment (Netflix, Amazon Prime, etc.) as well as educational content (Coursera, Udemy, etc.), has accelerated the demand for translation. During my discussions with translators, I made the observation that the audiovisual translation (AVT) industry is reluctant to use MT due to technical problems of subtitling and dubbing, despite short turnaround times brought on by high demand in translation. Prior to this study, I conducted an interview with representatives of two LSPs that offer AVT to gain insights about the use of MT in the workflows.[1] The results revealed that neither company uses MT in AVT, since "There are certain limitations imposed by alternative subtitling software, rendering the use of MT impractical"[2] and "MT can only be used for plain texts and provides poor quality because of the technical issues."[3] While the results confirm my observation, it is safe to say that the reasons put forth by industry representatives encompass generalizations that disregard the critical aspects of context, domain, and purpose. This situation shows that despite the rise of MT, there is still a limited understanding of how to work with it.

We are witnessing a shift in the responsibilities of professional translators and translator candidates. For a very long time, 'post-editor' has been envisioned as the future title for translators. In order to prepare translation students for the future, some higher education institutions that are aware of this predicament have added some MT courses into their curricula. However, there is currently a lack of up-to-date educational guidelines that provide practical studies on how to work with MT. In this regard, and due to their lack of knowledge of MTPE, translators and post-editors tend to make arbitrary changes to MT output, which takes longer than anticipated to complete their tasks. Therefore, I argue that this problem can be resolved by learning how to assess MT quality before starting the task of post-editing. This procedure could shorten the time spent post-editing and encourage translators and post-editors to work with the output more efficiently.

Taking all these into consideration, the present study intends to examine the impact of error annotation on post-editing effort and post-edited product and of the integration of MT into

[1] All interviews were conducted in confidentiality, and the names of interviewees are withheld by mutual agreement.
[2] Interview with LSP-2, May 4, 2020.
[3] Interview with LSP-1, May 1, 2020.

subtitling in the field of education.[4] The reason behind this choice of domain is that there are already several studies on the integration of MT into MOOCs (Massive Open Online Courses) showing good results for the quality of MT (Ruiz Costa-jussà et al. 2015; Castilho et al. 2018); thus, this might be a contribution to the literature. Moreover, less is required for educational content than for movies and television shows in terms of synchronization and timecodes.

The remainder of this paper is organized as follows: Section 2 will present a review of the literature on the integration of MT into AVT. Section 3 will describe the methodology of an experiment performed in line with the purposes of the study. Section 4 will present the results of the mentioned experiment. Section 5 will encompass a discussion on the limitations of the study, offer suggestions for future work, and provide concluding remarks.

2. The Integration of MT into AVT

The rise of digital media platforms, including Netflix, Amazon Prime, and Hulu, as well as massive open online course (MOOC) platforms like Coursera, Udemy, and edX, has led to an increased demand for translation services. As a result, translation scholars have become increasingly interested in the field of AVT. This section provides a literature review of the previous studies on the integration of MT into AVT, especially focusing on quality evaluation and post-editing effort. Studies on the integration of MT into the subtitling process date back to the late 1990s, when rule-based and transfer-based systems predominated. An MT system for translating business news captions between English and German in real-time was proposed by Eric Nyberg and Teruko Mitamura (1997). Additionally, Popowich et al. (2000) presented ALTo, an MT system that uses lexical resources to translate closed captions of American television broadcasts.

The developments in MT over the past ten years have had a huge impact on AVT. MT has distinguished itself in research on not only subtitles but also other AVT modes such as voice-over (Martín-Mor and Sánchez-Gijón 2016) and audio-description (Fernández-Torné and Matamala 2016). The effectiveness of MT systems specifically designed for subtitling, as well as the productivity and cognitive effort of post-editors, have been the main topics of research

[4] This study derives from the author's master's thesis (see Ekinci 2022) submitted at Istanbul 29 Mayıs University under the supervision of Işın Öner.

on subtitle translation. Before educational content, the first studies on MT and AVT were about film/TV subtitles.

In order to convert DVD subtitles, Armstrong et al. (2006) created an example-based machine translation system (EBMT). Volk et al. (2010) developed SMT systems to translate TV subtitles into English, Danish, Norwegian, and Swedish. Their systems produced excellent results (ibid.).

Despite its challenging nature, MT should be used in AVT due to the rising demand, according to Burchardt et al. (2016). They suggested conducting quality evaluation studies so that the quality could be improved, and accordingly, this will help to build a bridge between MT and AVT.

Sheila C. M. de Sousa, Wilker Aziz, and Lucia Specia (2011) examined post-editing effort for the translation of DVD subtitles in their study, where the post-editing effort is measured in terms of time. According to the study, post-editing the subtitles was 40% faster than translating from scratch. In addition, 69% of the translation generated by their MT system required little or no post-editing.

In terms of time spent and keylogging data, Koponen et al. (2020) compared the productivity of 12 professional subtitlers in both post-editing and translation from scratch. Although participants' time spent on the tasks and their rate of editing varied, the study indicated that overall, MTPE was "slightly faster" and required fewer keystrokes. They also stated that subtitle translators are "affected by the visual context of the video" (123).

In addition to the research on film/TV subtitles, there are several studies on subtitles of educational content, which is the focus of the present study. Castilho et al. (2017) introduced TraMOOC (Translation for Massive Open Online Courses), a European Union funded project that provides MT service customized for MOOCs from English into 11 languages. The authors note that while the primary objective of MOOCs is to enhance accessibility to knowledge and training for individuals who may not have access to traditional educational institutions, the presence of language barriers serves as a limiting factor for the intended audience. Therefore, TraMOOC intends to "develop high quality MT of the multifarious text genres typically included in MOOCs" (9361). Another study by Castilho et al. (2018), whose methodology inspired the present study, evaluated the MT quality using mixed-domain and in-domain data

from various sources to train PBSMT and NMT systems, with participants conducting tasks such as error annotation, post-editing, and giving adequacy and fluency scores and ranking.

The projects known as EMMA (European Multiple MOOC Aggregator) and transLectures (Transcription and Translation of Video Lectures) are noteworthy for their ability to expand the reach of online educational content to a wider audience. The objective of TransLectures is to produce transcriptions and translations for video lectures using automatic speech recognition (ASR) and MT technologies (Silvestre-Cerdà et al. 2012). EMMA offers automated transcription and translation services for MOOCs, incorporating cross-language support through the use of ASR and SMT (Brouns et al. 2015). These projects highlight the significant amount of educational material available through MOOCs and the necessity for their translation. Given that certain projects were initiated prior to the advent of the NMT approach and the proliferation of online educational platforms, it is likely that the need for translation services has escalated even more today.

Lastly, it is important to have a look at the industry side of the subject. As stated, there has been a significant rise in the need for translation services, leading the subtitling industry to explore cost-efficient strategies to meet this demand. Lindsay Bywood, Panayota Georgakopoulou, and Thierry Etchegoyhen (2017, 494) claim that technological advancements in soundwave representation and autotime functions have occurred in the subtitling industry, yet the use of translation memories (TMs) or MT has not been widely adopted in their workflows. The primary impediment to the integration of MT into subtitling is attributed to technical challenges such as time-coding and character limit, as mentioned in Section 1. AVT tools, such as MateSub, have the capability to incorporate MT and enable users to select a suitable style guide, such as that of Netflix. Additionally, these tools can provide translators with a warning if they surpass the character limit, allowing them to promptly revise their translation. Moreover, certain tools facilitate the automatic execution of time coding through sound waves. Thus, it is my contention that the integration of MT technologies and novel technical advancements can expedite both the technical and textual sides of subtitling. The study is primarily concerned with post-editing effort and product quality; therefore, technical issues will not be the central focus of the study.

3. Methodology

This section provides the methodology adopted in the experiment, where ten students were divided into two groups: The control group only performed the post-editing task, while the treatment group, in addition to post-editing, conducted a pre-task quality evaluation. The objective was to investigate whether the quality evaluation had any influence on the post-editing effort and the quality of the post-edited outputs.

The study uses Hans P. Krings's (2001) model for measuring post-editing effort, where he defines three measurements: temporal effort, technical effort, and cognitive effort. The first two measurements will be used as the post-editing effort indicators in this study. Being "the most easily measured" according to Krings (2001, 178), temporal effort refers to the time spent during post-editing. On the other hand, technical effort is related to the changes made by post-editors. The methods for measuring these efforts will be explained in the following sections.

3.1 Research Questions

In line with the purposes of the experiment, the research questions are as follows: (i) Is there a significant difference between the two groups in terms of temporal effort? (ii) Is there a significant difference between the two groups in terms of technical effort? (iii) Do annotators only correct the errors that they have annotated while post-editing?

3.2 Research Design

The study uses a mixed-method approach, which could be defined as using "several methods to collect or analyze data" (Saldanha and O'Brien 2014, 23). In this study, mixed-method approach specifically refers to the utilization of both qualitative and quantitative methodologies. Although translation researchers have been dealing with statistics and measurements, especially in cognitive effort studies, translation / post-editing cannot be reduced only to numbers. As Joss Moorkens (2018a) and Lucas Nunes Vieria (2015) suggest, relying only on quantitative data poses challenges in empirical research, and adopting a mixed-method approach can offer a viable solution to this problem. Therefore, this study benefits from commentaries and questionnaires as well as time, effort, and quality measurements.

The experiment was carried out in the translation technologies laboratory at Istanbul 29 Mayıs University in two separate sessions for the treatment and control groups. The experiment had seven stages:

1. Pre-experiment questionnaire
2. A training on quality evaluation metrics of the TAUS Dynamic Quality Framework (TAUS DQF) and its Quality Evaluation feature on the platform
3. The task of quality evaluation
4. A training on the Productivity feature of the TAUS DQF
5. Post-editing brief
6. The task of post-editing
7. Post-experiment questionnaire

The control group was exempt from stages 2 and 3, while the treatment group completed all stages. The participants were informed about this road map of the study prior to the start of the experiment.

3.3 Participants

The selected group of participants were third- and fourth-grade undergraduate students who were invited to participate in the experiment through email invitations. Ten students in total accepted to take part in the experiment. The reason behind choosing students as the participants of this experiment was to "empower the students, and help them understand the strengths and weaknesses of this new technology" (Moorkens 2018b, 375).

The participants' native language is Turkish. They have been pursuing a bachelor's degree in Translation Studies in the language pair English–Turkish and have successfully finished the courses Translation Technologies and Machine Translation. They are proficient with CAT tools and have fundamental knowledge of MT and post-editing procedures. The study focuses on a branch of MT that even students who have completed a course on machine translation are unfamiliar with, which is why this particular set of students was selected. Including participants in the study who do not have a basic understanding of MT could have rendered the findings dubious.

3.4 Online System and Tools

To machine translate the source text, an NMT model was trained on Google AutoML specifically for this study. The corpus contained TED talks and TED-Ed videos about psychology. The bilingual transcripts were aligned using Abbyy Aligner, and all the TMX files were uploaded to the model to be trained. The trained model obtained a 21.81 BLEU score (Papineni et al. 2002), performing marginally better than Google NMT with a performance improvement of 0.43, indicating that translations performed using it will be more closely related to the reference dataset.

For quality evaluation, the harmonized TAUS DQF-MQM (Multidimensional Quality Metrics) was chosen as the metric for the experiment since it is less detailed than MQM itself, which facilitates carrying out an evaluation in a specific amount of time. Besides, its quality evaluation tool has a user-friendly interface and is easy-to-use for students. This feature is also useful for researchers since it automatically generates results.

Post-editing effort can be measured using various tools, depending on research expectations and the intended outcome. The experiment aimed to measure participants' time spent on segments and changes, and a comfortable environment was tried to be arranged for the experiment. Many CAT tools were considered, but TAUS Productivity emerged as the best choice due to its clear interface to perform post-editing and comprehensible results section.

The pre- and post-experiment questionnaires were given to the participants via Google Forms. Lastly, a screen-recorder software was used to record all the experiment processes for further analysis.

3.5 Source Content

The source material was obtained from the "Introduction to Clinical Psychology" course on the MOOC platform, edX,[5] which offers a wide variety of courses on several subjects. The source video taken from this course was titled "Introduction to Anxiety and Mood Disorders" and serves as an introduction to a course module. It briefly discusses the definition of mental illness, how it is diagnosed, and the topics that will be covered in this module of the course. The video was 2.13 minutes long, and its transcript consisted of 331 words. Considering the

[5] https://www.edx.org/.

students' limited experience in quality evaluation and the fact that the treatment group would be performing error annotation and post-editing concurrently, it was decided to select a source of manageable length that would not disrupt their concentration and effort.

4. Results

This section encompasses the results of the pre- and post-experiment questionnaires, the findings related to post-editing effort in terms of temporal and technical effort, and a comparative analysis of the post-edited products with error annotation.

4.1 Pre-Experiment Questionnaire

A pre-experiment questionnaire was completed by students, asking them about their professional experience in translation, post-editing, and subtitling. Three out of 10 students had experience in translation, but none of them had enough experience to distinguish them from their peers. The questionnaire also asked about MT knowledge, with all students stating they knew what post-editing meant and offering different definitions to describe the term even though they had the same educational background. All students also mentioned that they did not know what an error annotation meant.

Students' attitudes towards MT were also assessed. 50% of the students said that they used MT "always," and 40% of them used it "sometimes." On the other hand, only 10% of students thought that "machine translation could be used in any project," while 80% believed it "produced quality translation only in specific domains." According to 90% of the students, "MT will be used widely, but the need for human touch will remain," indicating a hopeful outlook for their future line of work.

4.2 Findings for Temporal Effort

The study analyzed the temporal effort of the two groups based on the time spent in seconds. The analysis was conducted using Jamovi,[6] a free and user-friendly statistical spreadsheet. Mean, median, standard deviation (SD), and standard error (SE) values were

[6] The jamovi project (2022). jamovi (Version 2.3) [Computer Software], https://www.jamovi.org.

displayed in table 1. Due to the non-normal distribution of the data and the small sample size, both Student's t-test and non-parametric Mann-Whitney analysis were performed.

Table 1. Control and treatment group descriptives for temporal effort

	Group	**Mean**	**Median**	**SD**	**SE**
Temporal Effort	Control	1699	1554	678	303
	Treatment	941	944	150	67.3

The results of the Student's t-test indicated a significant difference ($t(8) = 2.44$, $p = .040$) in temporal effort scores between the control group (M= 1699, SD = 678) and the treatment group (M= 941, SD = 150). Additionally, the median latencies in the control and treatment groups were 1554 and 944 seconds, respectively. The distributions in the two groups differed significantly ($U = 1.00$; $p = 0.016$).

4.3 Findings for Technical Effort

The technical efforts of both the treatment and control groups were analyzed by means of calculating the edit distance using the Levenshtein distance. Like the temporal effort analysis, the technical effort was also analyzed through Jamovi. Table 2 presents the mean, median, SD, and SE values.

Table 2. Control and treatment group descriptives for technical effort

	Group	**Mean**	**Median**	**SD**	**SE**
Technical Effort	Control	282	240	164	73.3
	Treatment	221	196	86.6	38.7

Due to the same reasons mentioned in section 4.2, both the Student's t-test and the non-parametric Mann-Whitney analysis were carried out. In response to the second research question, the Student's t-test revealed that there was no significant difference ($t(8) = 0.738$, $p = .482$) in technical effort scores between the control group (M= 282, SD = 164) and the treatment group (M= 221, SD = 86.6). Also, based on the median latencies observed in the control and

treatment groups, which were 240 and 196, respectively, there was no significant difference in the distributions between the two groups (U = 11.00; p = 0.841).

Although the outcome was not statistically significant, the treatment group exhibited a lower degree of editing in the MT output in comparison to the control group. The precise cause of this finding remains unclear; however, it is plausible that the treatment group displayed a greater degree of trust in MT compared to the control group.

4.4 Comparative Analysis of Post-Edited Products with Error Annotation

Firstly, it is important to observe how many errors each participant found in the output. Table 3 displays the total number of errors assigned by five students for each error category. It is very clear that there are variations in the total number of errors among the evaluators. However, this observation aligns with the comments provided by the students who did not identify any significant issues in the output, resulting in fewer assigned error categories. To further analyze the choices made by the participants, this section reviews selected sentences alongside the corresponding post-edited output from both the treatment and control groups.

Table 3. Distribution of the numbers of errors

Error Types	**S1**	**S2**	**S3**	**S4**	**S5**
Accuracy	6	8	3	1	2
Fluency	11	4	11	6	2
Terminology	2	4	0	0	0
Style	2	5	1	2	1
Locale convention	0	0	0	0	0
Design	0	0	0	0	0
Verity	0	0	0	0	0
Total	21	21	15	9	5

There were 15 segments in the source content. Since discussing all of the segments would exceed the limits of this paper, this section focuses on the analysis of three selected problematic segments. The post-editing choices made by both groups are examined along with the errors categorized by the treatment group. Furthermore, general observations are provided regarding the rest of the tasks. The students enumerated from S1 to S5 were the members of the treatment group, while the students from S6 to S10 took part in the control group.

Segment 5

Source Text:

It's been a long and difficult process.

Machine Translation:

Uzun ve zor bir süreç oldu.

This segment exhibits a fluency error due to the omission of the pronoun "it," resulting in potential comprehension difficulties for readers. While not a mistranslation, the sentence lacks clear contextual direction. This error was identified by annotators S1 and S3 under the accuracy category, with subcategories of undertranslation and omission. S2, on the other hand, focused on grammar, noting that the perfect tense was translated to the past tense in Turkish, which is not technically incorrect given that Turkish does not have a perfect tense. Annotators S4 and S5 did not identify any errors in the sentence.

Regarding post-editing, most of the students in the treatment group post-edited the segment as "Bu, uzun ve zor bir süreç oldu." Only one student in the treatment group did not make any edits. Conversely, it should be noted that a total of three students refrained from making any revisions to this particular segment in the control group. S7 edited the text to read "Bu uzun ve zor bir süreç" by removing the word "oldu," resulting in a slight alteration in meaning. Meanwhile, the post-edited segment of S9 resulted in the phrase "Uzun ve zor bir süreç söz konusu," which also caused a slight shift in meaning.

Segment 9

Source Text:

The second classification system is the International Classification of Diseases or the ICD and this is widely used in the UK and Europe.

Machine Translation:

İkinci sınıflandırma sistemi, Uluslararası Hastalık Sınıflandırması veya ICD'dir ve bu İngiltere ve Avrupa'da yaygın olarak kullanılmaktadır.

The segment contains a fluency error wherein the flow is impacted by the frequent use of conjunctions. Dividing the segment into two would be one of the most effective ways to fix the problem. Only S2 and S3 identified this problem as a fluency error.

Three students in the treatment group post-edited this fluency error in a variety of ways. S1 did not eliminate any conjunctions, but instead changed the sentence structure to correct the fluency problem. S3 wrote the final clause of the phrase as a relative clause. S4 split the sentence in half. On the other hand, S6 applied the same strategy as S1 in the control group. No adjustments were made by S7 or S8 to fix the fluency mistake. Editing decisions made by S9 and S10 resulted in mistranslations.

Segment 10

Source Text:

Why am I telling you this?

Machine Translation:

Bunu sana neden anlatıyorum?

The MT output contains an obvious fluency-related grammatical register error. When the plural form of "you" (*size*) was intended, MT translated "you" as the informal "you" (*sana*). One annotator highlighted the identical problem in the commentary section but labeled it as a "style" error. On the other hand, the other four evaluators annotated this as a fluency error, specifically mentioning that it was about grammatical register.

While post-editing, four students in the treatment group post-edited this part as "Bunu size niye anlatıyorum?," and S1 post-edited it as "Peki, bunu sizlere niye anlatıyorum?," where s/he added a conjunction, meaning "so" or "well," in the beginning of the sentence, which was in fact a plausible choice to link this sentence with the rest of the speech. S1 annotated an additional style error along with fluency attached to "sana." Three students from the control group, on the other hand, likewise post-edited the part as "Bunu size niye anlatıyorum?" One of the other two students edited the word "size" in this section while changing the verb's case to "anlatmaktayım." Since the Productivity feature does not include a commentary section, it is unclear why the editing was made, and there is no contextual justification similar to S1's process; however, there might be an issue of over-editing. Another student in the control group went down a similar route by changing "sana" to "size" and the verb to "söylüyorum," which resulted in yet another arbitrary change.

In terms of error annotation, the annotators performed efficiently given that this was their first time conducting such a project. However, the outcomes also demonstrated that error

annotation is a very personal and subjective process. Additionally, in response to the third research question, it was observed that annotators made edits that they had not initially annotated. The post-experiment questionnaire has provided some valuable insight into this phenomenon.

Regarding post-editing, both groups exhibited equal performances in certain cases, while in others the treatment group outperformed the control group. The control group displayed a tendency towards arbitrary changes and over-editing, which could not be justified stylistically or based on the context. This tendency was not limited to the provided segments but extended to the rest of the segments as well, affirming the findings regarding technical effort, where the control group made more changes compared to the treatment group. In terms of the use of the MT output, the treatment group made more effective attempts at editing the segments.

4.5 Post-Experiment Questionnaire

The treatment group and control group were asked to comment on their post-editing processes, while the treatment group also wrote a commentary on their task of annotating errors.

The students in the treatment group responded to the questions similarly. Firstly, all students found error annotation challenging, even though they spotted the errors. This finding aligns with previous studies conducted by Işın Öner and Senem Öner Bulut (2021) and Moorkens (2018b). Students stated that "each category looks alike"[7] (S1) and they "had to check the criteria definitions quite often" (S3). On the other hand, S1 also suggested that if the descriptions of error categories were more detailed, annotation might have been easier to conduct. They mentioned that they had tried to select the correct error category (S3, S4), but they sometimes "failed" (S2, S3). Two students also stated that they found error annotation helpful in order to group the mistakes and improve the output.

In their commentaries about the task of post-editing, four students in the treatment group mentioned how error annotation affected their post-editing process. S3 said that the error annotation process had him/her remember the errors that s/he could not have found if s/he had not annotated errors in the previous task. S5 stated that the fact that they had annotated errors in advance made editing the text easier. S1 and S2 agreed that error annotation was harder than

[7] Students wrote their commentaries in English. Throughout this paper, they are presented exactly as they are.

post-editing. S1 also mentioned that the experience for him/her in terms of post-editing made him/her realize the aspects s/he had not realized before.

On the other hand, the control group focused on different aspects compared to the treatment group in their post-editing process. Students admitted that they looked up terminology they did not understand or whose meaning they were unsure about (S6, S7, S8, S9). S9 also stated that "MT processed a lot" and translated better compared to the past except for a few mistakes. S8 mentioned that "it was not a useful opportunity to see the previous segment during the post-editing process in terms of consistency and making the same decisions." S6 related the errors in the output to the fact that "the syntax of Turkish is so much different than English."

5. Conclusion

This paper sought to investigate the impact of error annotation on post-editing effort and post-edited product. The aim was to highlight the significance of quality evaluation, particularly error annotation, which was argued to be one of the best methods to learn how to work with MT. Among various domains and industries, the focus was specifically on AVT since it is an area that could greatly benefit from MT but, paradoxically, utilizes it to a limited extent.

To investigate the impact of error annotation, three questions were posed to measure the post-editing effort in terms of both technical and temporal effort. In response to the first research question, there was a significant temporal effort difference between the two groups, with the treatment group finishing the task faster. The answer to the second research question is that although there was no significant difference in technical effort between the two groups, the treatment group's technical effort was lower than that of the control group. The third research question revealed that students did not initially annotate all the errors that they post-edited during the task.

While it should be noted that drawing generalizations from a single study is not conclusive, these results suggest that error annotation had a positive impact on the treatment group. The treatment group completed the task faster than the control group, spent lower technical effort, and made more conscious changes, which shows that quality evaluation produced desirable outcomes that could be used in training both students and professionals in the future.

It is also important to mention the limitations of the study. First of all, the sample size was small; had there been a larger number of students involved, more information could have been gathered and the process could have been better understood. Secondly, keyboard logging systems were initially considered for measuring technical effort, but due to confidentiality concerns, their applicability proved to be limited. Keyboard logging data might have given greater details about the technical procedures connected to their temporal effort.

This study may offer suggestions for further research in the future. The same experiment might be carried out with graduate students or experienced translators/post-editors, yielding results that could be compared to those of the current study. This experiment could also be conducted using different domains or including other industries within different language pairs. These potential studies could display all sides of error annotation and its impacts.

References

Armstrong, Stephen, Andy Way, Colm Caffrey, and Marian Flanagan. 2006. "Improving the Quality of Automated DVD Subtitles via Example-Based Machine Translation." In *Proceedings of Translating and the Computer*, 1–13. London: Aslib. https://aclanthology.org/2006.tc-1.9.

Brouns, Francis, Nicolás Serrano Martínez-Santos, Jorge Civera, Marco Kalz, and Alfons Juan. 2015. "Supporting Language Diversity of European MOOCs with the EMMA Platform." In *Proceedings of the European MOOC Stakeholder Summit 2015*, edited by M. Lebrun, M. Ebner, I. de Waard, M. Gaebel, 157–165. https://research.ou.nl/en/publications/supporting-language-diversity-of-european-moocs-with-the-emma-pla.

Burchardt, Aljoscha, Arle Lommel, Lindsay Bywood, Kim Harris, and Maja Popović. 2016. "Machine Translation Quality in an Audiovisual Context." *Target* 28 (2): 206–221. doi:10.1075/target.28.2.03bur.

Bywood, Lindsay, Panayota Georgakopoulou, and Thierry Etchegoyhen. 2017. "Embracing the Threat: Machine Translation as a Solution for Subtitling." In "Translation of Economics and the Economics of Translation," edited by Łucja Biel and Vilelmini Sosoni. Special Issue, *Perspectives* 25 (3): 492–508. doi:10.1080/0907676x.2017.1291695.

Castilho, Sheila, Federico Gaspari, Joss Moorkens, and Andy Way. 2017. "Integrating Machine Translation into MOOCs." In *Proceedings of EDULEARN17 Conference*, edited by L. Gómez Chova, A. López Martínez, and I. Candel Torres, 9360–9365. Barcelona, Spain: IATID. doi:10.21125/edulearn.2017.0765.

Castilho, Sheila, Joss Moorkens, Federico Gaspari, Rico Sennrich, Andy Way, and Panayota Georgakopoulou. 2018. "Evaluating MT for Massive Open Online Courses." In "Human Evaluation of Statistical and Neural Machine Translation," edited by Andy Way and Mikel L. Forcada. Special Issue, *Machine Translation* 32 (3): 255–278. doi:10.1007/s10590-018-9221-y.

de Souza, Sheila C. M., Wilker Aziz, and Lucia Specia. 2011. "Assessing the Post-editing Effort for Automatic and Semi-Automatic Translation of DVD Subtitles." In *Proceedings of Recent Advances in Natural Language Processing*, edited by Galia Angelova, Kalina Bontcheva, Ruslan Mitkov, and Nikolai Nikolov, 97–103. https://aclanthology.org/R11-1014.

Ekinci, Sena. 2022. "The Effect of Error Annotation on Post-editing Effort and Post-edited Product: An Experimental Study on Machine-Translated Subtitles of Educational Content." Master's thesis, Istanbul 29 Mayıs University.

Fernández-Torné, Anna, and Anna Matalama 2016. "Machine Translation in Audio Description? Comparing Creation, Translation and Post-editing Efforts." *SKASE*

Journal of Translation and Interpretation 9 (1): 64–87. http://www.skase.sk/Volumes/JTI10/pdf_doc/05.pdf.

Koponen, Maarit, Umut Sulubacak, Kaisa Vitikainen, and Jorg Tiedemann. 2020. "MT for Subtitling: User Evaluation of Post-editing Productivity." In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, edited by André Martins, Helena Moniz, Sara Fumega, Bruno Martins, Fernando Batista, Luisa Coheur, Carla Parra, Isabel Trancoso, Marco Turchi, Arianna Bisazza, Joss Moorkens, Ana Guerberof, Mary Nurminen, Lena Marg, and Mikel L. Forcada, 115–124. Lisboa, Portugal: European Association for Machine Translation. https://aclanthology.org/2020.eamt-1.13.

Krings, Hans P. 2001. *Repairing Texts: Empirical Investigations of Machine Translation Post-editing Processes.* Edited by Geoffrey S. Koby. Kent: The Kent State University Press.

Martín-Mor, Adrià, and Pilar Sánchez-Gijón. 2016. "Machine Translation and Audiovisual Products: A Case Study." *The Journal of Specialised Translation*, no. 26, 172–186. https://jostrans.org/issue26/art_martin.pdf.

Moorkens, Joss. 2018a. "Eye Tracking as a Measure of Cognitive Effort for Post-editing of Machine Translation." In *Eye Tracking and Multidisciplinary Studies on Translation*, edited by Callum Walker and Federico M. Federici, 55–70. Amsterdam: John Benjamins. doi:10.1075/btl.143.04moo.

———. 2018b. "What to Expect from Neural Machine Translation: A Practical In-class Translation Evaluation Exercise." *The Interpreter and Translator Trainer* 12 (4): 375–387. doi:10.1080/1750399X.2018.1501639.

Nunes Vieira, Lucas. 2015. "Cognitive Effort in Post-Editing of Machine Translation: Evidence from Eye Movements, Subjective Ratings, and Think-Aloud Protocols." PhD diss., Newcastle University.

Nyberg, Eric, and Teruko Mitamura. 1997. "A Real Time MT System for Translating Broadcast Captions." In *Proceedings of the Sixth Machine Translation Summit*, 51–57. https://aclanthology.org/1997.mtsummit-papers.2.

Öner, Işın, and Senem Öner Bulut. 2021. "Post-Editing Oriented Human Quality Evaluation of Neural Machine Translation in Translator Training: A Study on Perceived Difficulties and Benefits." *transLogos* 4 (1): 100–124. doi:10.29228/transLogos.33.

Papineni, Kishore, Salim Rukos, Tod Ward, and Wei-Jing Zhu. 2002. "BLEU: A Method for Automatic Evaluation of Machine Translation." In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, 311–318. Philadelphia. doi:10.3115/1073083.1073135.

Popowich, Fred, Paul McFetridge, Davide Turcato, and Janine Toole. 2000. "Machine Translation of Closed Captions." *Machine Translation* 15 (4): 311–341. https://www.jstor.org/stable/20060451.

Ruiz Costa-jussà, Marta, Lluis Formiga, Oriol Torrillas, Jordi Petit, and José Adrián Rodríguez Fonollosa. 2015. "A MOOC on Approaches to Machine Translation." *The International Review of Research in Open and Distributed Learning* 16 (6): 174–205. doi:10.19173/irrodl.v16i6.2145.

Saldanha, Gabriela, and Sharon O'Brien. 2014. *Research Methodologies in Translation Studies.* New York: Routledge.

Silvestre-Cerdà, J. A., M. A. del Agua, G. Garcés, G. Gascó, A. Giménez, A. Martínez, A. Pérez, I. Sánchez, N. Serrano, R. Spencer, J. D. Valor, J. Andrés-Ferrer, J. Civera, A. Sanchis, and A. Juan. 2012. "transLectures." In *Online Proceedings of Advances in Speech and Language Technologies for Iberian Languages, IBERSPEECH '12*, Madrid, Spain. https://riunet.upv.es/handle/10251/37290.

Volk, Martin, Rico Sennrich, Christian Hardmeier, and Frida Tidström. 2010. "Machine Translation of TV Subtitles for Large Scale Production." In *Proceedings of the Second Joint EM+/CNGLWorkshop "Bringing MTto the User: Research on Integrating MTin the Translation Industry,"* edited by Ventsislav Zhechev, 53–62. Denver, Colorado, USA: Association for Machine Translation in the Americas. doi:10.5167/uzh-36755.