Translation vs Post-editing of NMT Output: Measuring effort in the English-Greek language pair

Maria Stasimioti stasimioti@ionio.gr
Department of Foreign Languages, Translation and Interpreting, Ionian University, Corfu, 49100, Greece
Vilelmini Sosoni sosoni@ionio.gr
Department of Foreign Languages, Translation and Interpreting, Ionian University, Corfu, 49100, Greece

Abstract

Machine Translation (MT) has been increasingly used in industrial translation production scenarios thanks to the development of Neural Machine Translation (NMT) models and the improvement of MT output, especially at the level of fluency. In particular, in an effort to speed up the translation process and reduce costs, MT output is used as raw translation to be subsequently post-edited by translators. However, post-editing (PE) has been found to differ from both human translation and revision of human translation in terms of the cognitive processes and the practical goals and processes employed. In addition, translators remain sceptical towards PE and question its real benefits. The paper seeks to investigate the effort required for full PE and compare it with the effort required for manual translation, focusing on the English-Greek language pair and NMT output. In particular, eye-tracking and keystroke logging data are used to measure the effort expended by translators while translating from scratch and the effort required while post-editing the NMT output. The findings indicate that the effort is lower when post-editing than when translating from scratch, while they also suggest that experience in PE plays a role.

1. Introduction

In the past fifteen years, the translation industry has seen a growth in the amount of content to be translated and has received pressure to increase productivity and speed at reduced costs. To respond to these challenges, it has turned to Machine Translation (MT). The most common and widely expanding scenario –especially for certain language pairs and domains– involves the use of MT output to be then post-edited by professional translators (Koponen, 2016). This practice –generally termed post-editing of machine translation (PEMT) or simply post-editing (PE)– is increasingly gaining ground (Green et al., 2013; O’Brien et al., 2014; O’Brien and Simard, 2014; Lommel and DePalma 2016; Vieira et al. 2019) not least because of the development of Neural Machine Translation (NMT) models and the subsequent improvement of MT output, especially at the level of fluency (Castilho et al., 2017). In fact, studies have shown that post-editing high-quality MT output can, indeed, increase the productivity of professional translators compared to manual translation, i.e. human translation or translation “from scratch” (cf. O’Brien 2007; Groves and Schmidtke 2009; Tatsumi 2009; Guerberof, 2009; Plitt and Masselot, 2010). However, PE has been found to differ from both human translation and revision of human translation in terms of the cognitive processes and the practical goals and processes employed (Krings, 2001; O'Brien, 2002), while translators approach it with caution and skepticism and question its real benefits (Gaspari et al., 2014; Koponen, 2012; Moorkens, 2018; Vieira and Alonso, 2018). Their skepticism is directly related to the nature of PE which involves "working by correction rather than creation" (Wagner, 1985: 2), to the perception that PEMT is slower than translating from scratch and to the fear that MT is a threat to their profession (Moorkens, 2018) and "might have a de-professionalising effect on translation" (Vieira and Alonso, 2018: 16). It is, thus, particularly interesting to investigate the productivity gains when post-editing NMT output and to measure the cognitive effort expended by post-editors during the PE task and determine whether the translators' skepticism is justified or whether translating by PE is indeed the way forward (Garcia, 2011).

Under the light of the above, the aim of the paper is to investigate the effort required for the full PE of NMT output and compare it with the effort required for manual translation, focusing on the English-Greek language pair. To that end, twelve experienced professional translators are asked to post-edit NMT output of two semi-specialised texts and also manually translate two different comparable texts. Eye-tracking and keystroke logging data are used in order to measure the effort expended by translators while translating from scratch and the effort required while carrying out full PE of the NMT output.

2. Related work

Lately, many studies have showcased the benefits of post-editing MT output, as opposed to translating source texts (STs) from scratch, mainly in the context of non-literary translation (cf. O'Brien 2007; Groves and Schmidtke 2009; Tatsumi 2009; Green et al., 2013; Plitt and Masselot, 2010), but also in the context of literary translation (cf. Genzel et al., 2010; Greene et al., 2010; Jones and Irvine 2013; Besacier, 2014; Toral and Way, 2015; Moorkens et al., 2018). More specifically, several studies have been carried out with a view to estimating the productivity gains when post-editing MT output and measuring the cognitive effort expended by post-editors. In particular, Plitt and Masselot (2010) carried out a productivity test involving PE of MT output compared to traditional human translation in an industrial environment and found that MT helped translators substantially improve their productivity given that MT followed by PE improved throughput on average by 74%, thus reducing translation time by 43%. In a similar study, Zhechev (2014) found that MT followed by PE resulted in substantial productivity gains as compared to translation from scratch.

However, productivity alone does not provide information on "how post-editing occurs as a process, how it is distinguished from conventional translation, what demands it makes on post-editors, and what kind of acceptance it receives from them" (Krings 2001: 61). Therefore, Krings (2001) argues that the feasibility of post-editing compared to human translating should not be determined by processing time alone. O'Brien (2011: 198) also claims that post-editing productivity means "not only the ratio of quantity and quality to time but also the cognitive effort expended; and the higher the effort, the lower the productivity". More specifically, Krings (2001) identifies three categories of PE effort: the temporal effort, which refers to the time taken to post-edit a sentence to a particular level of quality, the technical effort, which refers to keystroke and mouse activities such as deletions, insertions, and text re-ordering and the cognitive effort, which refers to the "type and extent of those cognitive processes that must be activated in order to remedy a given deficiency in a machine translation" (Krings, 2001: 179). Therefore, research into the cognitive aspect of PE is necessary for a better understanding of PE effort and its relation to that of conventional translation. Under that light, a series of studies have tried to investigate the cognitive effort in relation to PE and manual translation (e.g. Carl et al., 2011;

Balling and Carl, 2014; Mesa-Lao, 2014; Elming, Balling and Carl, 2014; Carl, Gutermuth and Hansen-Schira, 2015, Jia et al., 2019).

The above studies compare manual translation with PE of Statistical Machine Translation (SMT) and NMT outputs. The present study's novelty is the focus on the English-Greek language pair for which there are no related studies to date.

3. Experimental setup

As already pointed out, eye-tracking and keystroke logging data were used to measure the temporal, technical and cognitive effort expended by translators while translating from scratch and while carrying out full PE of the NMT output. The translation and PE experiments were carried out in March 2018 at the HUBIC Lab[1] (Raptis and Giagkou, 2016) of the Athena Research Center[2] in Athens. A detailed consent form was signed by all participants prior to the execution of the experiments, while all stored data were fully anonymized in accordance with Greek Law 2472/97 (as amended by Laws 3783/2009, 3917/2011 and 4070/2012).

Twelve Greek professional translators participated in the experiments, in which their eye movements and typing activity were registered with the help of an eye-tracker and specialised software. Their selection followed a call for participation which was sent to the members of the two biggest Greek associations of professional translators, i.e. the Panhellenic Association of Translators[3] (PEM) and Panhellenic Association of Professional Translation Graduates of the Ionian University[4] (PEEMPIP) and was shared on social media. Potential participants expressed their interest for participating in the study by filling in a Google form; they subsequently received an e-mail with details on the aim of the research and guidelines for the translation and PE task along with some educational material (see section 3.2). In addition, they were asked to fill in two questionnaires: a pre-task questionnaire and a post-task questionnaire. The pre-task questionnaire, consisting of 34 questions (22 closed-ended questions and 12 open-ended questions), aimed at defining the profile of the participants and their perception of MT and had to be filled in before the experiment, while the post-task questionnaire, consisting of 15 questions (13 closed-ended questions and 2 open-ended questions), aimed at receiving feedback on translation and PE tasks and had to be filled in after the experiment.

3.1. The participants

As it emerges from Table 1, all the participants were female. Half of them were aged 30 to 40 years old, 33% were aged 40-50 years old and 17% were aged 20-30 years old. The majority of the participants had either an undergraduate degree (42%) or a postgraduate degree (50%), mainly in the translation field (67%). It should also be noted that all participants had normal or corrected to normal vision, two wore contact lenses, and one wore glasses, yet the calibration with the eye-tracker was successful for all twelve.

Gender	Female	100%
	Male	0%
Age group distribution	20-30	17%
	30-40	50%

[1] http://www.hubic-lab.eu/

[2] https://www.athenarc.gr/en

[3] http://www.pem.gr/el/

[4] http://peempip.gr/el/

	40-50	33%
Education level	Undergraduate degree holders	42%
	Postgraduate degree holders	50%
	PhD holders	8%
Degree type	Translation	67%
	Language/Linguistics	25%
	Other	8%

Table 1. Participants' age distribution, education level and degree type

The majority (83%) had at least 5 years of experience in translation (Table 2), while their work involved translation tasks (100%), revision tasks (92%), PE tasks (67%), terminology work (50%) project management (50%), subtitling (33%) as well as other tasks (17%) (Table 3).

Years of experience in translation	1-5 years	17%
	5-10 years	17%
	10-20 years	58%
	> 20 years	8%

Table 2. Participants' years of experience in translation

Tasks involved in participants' work	Translating	100%
	Revising	92%
	Post-editing	67%
	Project Management	50%
	Terminology work	50%
	Subtitling	33%
	Other	17%

Table 3. Tasks involved in participants' work

As far as their experience in PE is concerned, 84% of participants had experience in PE, either 1 year (25%), 2 years (17%), 3 years (17%), 5 years (17%) or over 5 years (8%) of experience in PE (Table 4).

Years of experience in PE	0 years	16%
	1 year	25%
	2 years	17%
	3 years	17%
	4 years	0%
	5 years	17%
	> 5 years	8%

Table 4. Participants' years of experience in PE

However, when they were asked about their workload ratio involving the PE of MT output, more than half replied that PE involved only 1% to 25% of the daily workload. For one of them PE involved 26% to 50% of the daily workload, for another one PE involved 51% to 75% of the daily workload, while for 3 of them PE involved 0% of the daily workload (Table 5).

Participants' workload ratio involving post-editing MT output	0%	25%
	1 - 25%	59%
	26 - 50%	8%
	51 - 75%	8%
	76 - 100%	0%

Table 5. Participants' workload ratio involving post-editing MT output

Although a high percentage of the participants, namely 84%, declared that they had previous experience in PE, only 33% of them had received training in PE, while 83% would be interested in receiving training in PE, saying that they consider it to be either moderately important (58%) or very important (42%) (Table 6). In addition, 75% of the participants stated that they prefer not to use MT in their CAT tools (Table 7).

Participants' view on PE training	Extremely important	0%
	Very important	42%
	Moderately important	58%
	Not important	0%
	Not at all important	0%

Table 6. Participants' view on PE training

Use of MT in participant's work	Yes	25%
	No	75%

Table 7. Use of MT in participants' work

Their answers to these two questions are closely related to their answers about their perception towards PE and MT, since a positive attitude to MT has been found to be a factor in PE performance (de Almeida, 2013; Mitchell, 2015). In particular, their answers regarding their perception towards PE were mixed. Some of them believed that PE is a useful, time-saving and necessary task, going hand in hand with MT and they were willing to add it to their services. However, others were negatively disposed stating that they preferred translation from scratch, that PE made their job harder and that PE rates were not fair. It should be noted that those negatively disposed were mainly translators with many years of experience or translators working predominantly with marketing texts or transcreation. This is in line with the findings of Moorkens and O'Brien (2015), who also observed that attitudes appear to be more negative in the case of experienced translators. As regards their perception towards MT and although the majority pointed out that they prefer not to use MT in their CAT tools, many appeared to recognise the latest developments in the field stating that "[MT] has done huge steps forward in the past years. Definitely here to stay. And to be used more with AI applications"; "MT can offer significant improvements in speed and accuracy when the machine is trained with good quality data", while as far as Google Translate is concerned "[It is] very useful and getting better by the day. I am happy to use it for languages I do not know, I may not always feel 100% positive about it as a professional linguist, but I accept it for what it is". It should be noted that the participants who were negatively disposed to PE were also negatively disposed to MT.

Regarding the translation and PE task difficulty, as this was identified by them in the post-task questionnaire, the participants found both tasks to be neither very easy nor very difficult. The User Interface (Translog II environment), the STs' difficulty and the quality of the MT raw output were among the factors that posed problems to the participants during the translation task and the PE task respectively. There were, also, other reasons that caused difficulties in both tasks such as the inability of the participants to consult dictionaries and external resources.

3.2. Description of the experiment

A Tobii TX-300 eye-tracker[5] and Translog-II software (Carl, 2012) were used to register the participants' eye movements, keystrokes and time needed during the translation and PE tasks they were asked to carry out. The texts (see below) were displayed in 17-point Tahoma font and double spacing on a Tobii TX Display (23'') at 1920 x 1080 pixels and the average viewing distance aimed at was 50-60 cm from the screen.

According to O'Brien (2009) the quality of the eye-tracking data may be affected by several factors, such as participants' optical aids, eye make-up, lighting conditions, noise, unfamiliarity, user's distance from the monitor etc. In an effort to minimize the implications of some of these factors, a controlled environment for the experiment was set up. In particular, a quiet room was selected, blackout blinds were used to reduce the amount of natural light, the same artificial light was used during all experiments, and a fixed chair was used, so that the participants could not easily move about and increase or decrease the distance to the monitor (Hvelplund, 2011).

The experiment consisted of one session for each participant. Before the sessions, the participants were informed by email about the nature of the experiments, the task requirements and the general as well as task-specific guidelines they had to follow. More specifically, the general guidelines they received included the following:

- Your hair should not block your eyes.
- Do not wear mascara.
- Avoid touching your eyes (e.g. rubbing your eyes, removing/wearing eyeglasses, etc.).
- During the translation and PE tasks, look exclusively at the computer screen.
- Try to keep your head as steady as possible.
- External resources (dictionaries, Internet, etc.) cannot be used.

The translation task was a traditional manual translation assignment. Participants were asked to provide their translation in a split-screen window. The ST was displayed at the top half of the screen and the translation at the bottom half, as suggested by previous studies (Hvelplund, 2011; Carl et al., 2011; Mesa-Lao, 2014; Carl et al., 2015). Since all the participants in this study were professional translators, the only guideline provided to them was to produce a text with the same *skopos* (Vermeer, 1989) as that of the original text and of publishable quality.

The PE task was a traditional PE assignment. Participants were asked to fully post-edit the raw output generated by the NMT-core engine. Like in the translation task, the ST was displayed at the top half of the screen and the translation at the bottom half, as suggested by previous studies (Hvelplund, 2011; Carl et al., 2011; Mesa-Lao, 2014; Carl et al., 2015). Translators worked directly on the translation. To facilitate eye-tracking measurements, texts were fully displayed to avoid any need for participants to scroll in either the source (ST) or the target text (TT) window. As opposed to the translation task, they were given detailed guidelines as well as training material in PE. In particular, since previous training and experience in PE was not a prerequisite for participating in the study, the participants received brief training in PE before executing the task. The training included a video, a presentation, as well as some educational material in PE which were sent to them five days before the execution of the tasks. The

[5] The TX-300 eye tracker is an integrated eye tracker that is supplied with a removable 23'' TFT monitor. Its large head movement box allows the subject to move during tracking while maintaining accuracy and precision at a sampling rate of 300 Hz. (https://www.tobiipro.com/product-listing/tobii-pro-tx300/).

guidelines for the full PE of the NMT output were based on the comparative overview of full PE guidelines provided by Hu and Cadwell (2016) as these were proposed by TAUS (2016), O'Brien (2010), Flanagan and Christensen (2014), Mesa-Lao (2013) and Densmer (2014), i.e retain as much raw MT translation/output as possible, the message transferred should be accurate, fix any omissions and/or additions (at the level of sentence, phrase or word), correct mistranslations, correct morphological errors, correct misspellings and typos, fix incorrect punctuation if it interferes with the message, correct wrong terminology, fix inconsistent use of terms, do not introduce stylistic changes.

In an effort to ensure that they had actually studied the material and that there were no questions or doubts, the participants were interviewed prior to the execution of the tasks and were specifically asked about the training material and also about the guidelines they had received.

A warm-up task was completed for human translation before the translation task and a warm-up task for PE before the actual PE task. The participants were informed that data from all texts would be subjected to analysis, although the warm-up texts were used only in order to familiarize the participants with the environment, the tools and the different types of tasks. After the warm-up, the actual experimental tasks followed, which involved the translation of two texts, i.e Text 1 and Text 2 (see below), and the PE of two different texts, i.e Text 3 and Text 4 see below), following the afore-mentioned guidelines. Participants were also asked to carry out both tasks at the speed at which they would normally work in their everyday work as professional translators; therefore, no time constraint was imposed. However, access to either online or offline translation aids was not allowed as it could have led to a reduction in the amount of recorded eye-tracking data.

The English STs used in this study were short educational texts selected from OER Commons[6], which is a public digital library of open educational resources. Six[7] 120 to 140-word long excerpts were selected from various courses on Business Administration and Social Change and the titles of the courses were retained as context information for the participants. The texts were chosen with the following criteria in mind: they had to be semi-specialised and easy for participants to translate or post-edit without access to external resources and they also had to be of comparable complexity. The texts chosen had comparable Lexile®[8] scores per task (between 1000L and 1100L for the translation task and 1300L and 1400L for the PE task), i.e they were suitable for 11th/12th graders (Table 8).

	Text 1 – T1	Text 2 – T2	Text 3 – T3	Text 4 – T4
Lexile® Measure	1000L - 1100L	1000L - 1100L	1300L - 1400L	1300L - 1400L
Number of sentences	8	8	6	7
Mean sentence length	15.38	17.43	28.60	22.67
Word count	123	122	143	136
Characters without spaces	777	713	785	896

Table 8. Lexile® scores for the source texts used in the study

[6] https://www.oercommons.org/

[7] Two texts were used exclusively for the warm-up session and are not included in the ensuing analysis and discussion.

[8] https://la-tools.lexile.com/free-analyze/

The NMT-core engine used to produce the Greek raw MT output for the PE task was Google Translate (output obtained March 24, 2018). The NMT output was evaluated using the BLEU and WER metrics. The BLEU score was calculated using the Tilde Custom Machine Translation toolkit[9]. As it emerges from Table 9, both texts had a very good score as regards BLEU and WER score and PE could be used to achieve publishable translation quality.

Text	**Translation engine**	**BLEU**	**WER**
Economics – Text 3	Google Translate NMT	51.33	37.7
The Endocrine System – Text 4	Google Translate NMT	60.62	34.5

Table 9. Automatic evaluation scores per text

4. Measuring translation and PE cognitive effort

As already pointed out, eye-tracking and keystroke logging data were used to calculate the participants' effort, i.e. the temporal effort, the technical effort and the cognitive effort which was expended during the translation and PE tasks.

4.1. Temporal Effort

According to Carl et al. (2011: 137) "One of the most obvious reasons for engaging in post-editing is the desire to save time". In his study the average time spent on manually translating a text was 7.52 minutes, while the average time spent on post-editing a text was 7.35 minutes. Although that difference was not significant ($p = 0.7118$), Carl et al. considered it "an indication that post-editing may lead to some time saving" (Carl et al., 2011: 137). In our study, we observed a statistically significant difference $t(23) = 3.04$, $p < 0.01$, when comparing the average task time[10] required for the translation task ($M = 9.86$, $SD = 4.53$) and the PE task *(*$M = 7.91$, $SD = 2.48$) (Table 10), resulting, thus, in an average time saving[11] of 19.8%. It is worth noting that the study' s findings corroborate the findings of previous studies which, however, involve different language pairs, MT systems, participants and experimental set-ups. In particular, the 19.8% average time saving percentage is similar to the 25% average time saving reported by Elming et al. (2014). According to Mesa-Lao (2014), who also found that translators in his study were always faster in the PE task, the longer task time in the translation task may be explained by the requirement of the translators to first read the ST (initial orientation phase) before starting to type the translation (drafting phase). When translating from scratch there are three phases: initial orientation (reading), translation drafting and final revision (Mesa-Lao, 2014; Carl et al., 2011). When post-editing, though, most post-editors tend to skip the initial orientation phase, in an effort to save time and they also tend to skip overall the final revision phase after making their changes, since PE is a kind of revision of the machine generated text (Mesa-Lao, 2014). So, according to Mesa-Lao (2014), this lack of a clear orientation phase and revision phase, along with the fact that (in principle) much less typing should be involved in PE when compared to translation, may explain the differences in task times.

Carl et al. (2015) and Jia et al. (2019) measured the average per-word translation and PE time in milliseconds (ms) and also found PE to be faster than translation from scratch. Although the participants in both studies had no previous experience in PE, they needed less time for PE,

[9] https://www.letsmt.eu/Bleu.aspx

[10] It should be noted that the start time of the task was calculated from the moment we opened the project (i.e. when we pressed the "start logging" button) and the task was considered finished when we pressed the "stop logging" button.

[11] Time saved percentage = 100 - average PE time/average translation time*100 (Elming et al., 2014)

leading Carl et al. (2015: 168) to make the assumption that "trained post-editors would even be more efficient in terms of editing times". A similar assumption, i.e. "more post-editing experience will yield a margin of time saving", was made in another previous study (Carl et al., 2011: 138), where also no participant had previous experience in PE. In our case, the majority (84%) of the participants had previous experience in PE (see section 3.1). When measuring the average task time expended by the participants with previous experience in PE and comparing it to the average task time expended by those without previous experience in PE (Table 11), we noticed that the experience in PE had affected the time the participants needed to post-edit the two texts (Text 3 and Text 4). In particular, the average task time expended by the participants with previous experience in PE was 7.07 minutes, while the average task time expended by those without previous experience was 10.42 minutes (Table 11). Although that difference is not significant (*p* = 0.11) –due to the low number of the participants and the number of texts involved in this study– it still indicates that PE experience may lead to lower temporal effort.

Task	Mean	SD
Translation task	9.86	4.53
PE task	7.91	2.48

Table 10. Temporal effort per task: Mean and standard deviation values of the task duration

Task	Participants	Mean	SD
PE	Professionals with experience in PE	7.07	1.34
	Professionals without experience in PE	10.42	4.09

Table 11. Professionals with experience in PE vs professionals without experience in PE: Mean and standard deviation values of the PE task duration

4.2. Technical Effort

Although it goes without saying that translation requires more typing than PE, given that one starts from scratch, it is interesting to compare the technical effort, i.e. the number of keystrokes (insertions and deletions), involved in both activities as the findings are useful in terms of ergonomics related to the translators' overall well-being and acceptance of MT and PE. The study reveals a statistically significant difference *t* (23) = 16.08, *p* < 0.01 between the average keyboard activity in the translation task (*M* = 1195, *SD* = 126) and the PE task (*M* = 458, *SD* = 226) (Table 12). In line with Carl et al. (2011), we noticed that the number of insertions was higher in the translation task, while the number of deletions was higher in the PE task. This can be easily explained by the fact that in the translation task the participants performed the translation from scratch, whereas in the PE task they only corrected the errors from the machine generated output. Interestingly, deletions were quite high in the translation task. This may be (partly) due to the participants' inability to consult external resources, a fact that led them to delete and rewrite words of their own translations in an effort to produce a better translation, as well as due to typos they had to correct while translating.

The experience in PE seems also to have affected the technical effort (Table 13). In particular, in the PE task the average keyboard activity of the participants with previous experience in PE was 438 keystrokes and for those without experience 521 keystrokes. Although the difference between the average keyboard activity is not significant (*p* = 0.52), it indicates that experienced post-editors may perform less keystrokes than those without experience in PE.

Task	**Total number of keystrokes**		**Insertions**		**Deletions**	
	Mean	**SD**	**Mean**	**SD**	**Mean**	**SD**
Translation	1195	126	1039	70	156	62
PE	458	226	239	116	220	111

Table 12. Technical effort per task: Mean and standard deviation values for the total number of keystrokes, insertions and deletions

		Total number of keystrokes		**Insertions**		**Deletions**	
Task	**Participants**	**Mean**	**SD**	**Mean**	**SD**	**Mean**	**SD**
PE	Professionals with experience in PE	438	211	228	110	209	103
	Professionals without experience in PE	521	279	270	140	252	138

Table 13. Professionals with experience in PE vs professionals without experience in PE: Mean and standard deviation values for the total number of keystrokes, insertions and deletions in the PE task

4.3. Cognitive effort

Eye-tracking measures, such as fixation count, fixation duration, gaze time, pupil dilation and saccades, have been lately used for measuring cognitive effort in translation studies (Moorkens, 2018). In particular, an increased number of fixations (Doherty et al., 2010), longer average fixation durations (Carl et al., 2011) and gaze time, i.e. the sum of all fixation durations, (Sharmin et al., 2008) have been used as indicators of particular items requiring more cognitive effort. In the present study and similarly to Mesa Lao (2014), we noticed that the translation task triggered more (*M* = 1284, *SD* = 791) and longer (*M* = 420, *SD* = 70.38) fixations than the PE task (*M*= 1135, *SD* = 429 and *M* = 355, *SD* = 37.75 respectively) (Table 14). The differences in average fixation count (*p* = 0.17) and fixation duration (*t*(23) = 5.46, *p* < 0.01) indicate that the cognitive load is higher in the translation task than in the PE task. Contrary to Carl et al. (2011), who found the average gaze time to be almost the same in the manual translation task and in the PE task, we found in our study a statistically significant difference *t*(23) = 3.27, *p* < 0.01 between the average gaze time in the translation task (*M* = 8.44, *SD* = 3.94) and in the PE task (*M* = 6.62, *SD* = 2.18) (Table 14). Therefore, it is obvious from our findings that PE is less cognitively demanding than translation from scratch. Similarly to our findings in the case of temporal effort (section 4.1) and technical effort (section 4.2), previous experience seems to have also affected the cognitive effort. In particular, a difference in fixation count (*p* = 0.18) and gaze time (*p* = 0.19) was found between the participants with previous experience in PE (*M* = 1020 and *M* = 6.05 respectively) and those without previous experience in PE (*M* = 1480 and *M= 8.31 respectively*), indicating that the cognitive load might be lower for experienced post-editors (Table 15).

	Fixation count		Fixation duration (msec)		Total gaze time (mins)	
Task	**Mean**	**SD**	**Mean**	**SD**	**Mean**	**SD**
Translation	1284	791	420	70.38	8.44	3.94
PE	1135	429	355	37.75	6.62	2.18

Table14. Cognitive effort per task: Mean and standard deviation values of the fixation count, the fixation duration and the gaze time

		Fixation count		Fixation duration (msec)		Total gaze time (mins)	
Task	**Participants**	**Mean**	**SD**	**Mean**	**SD**	**Mean**	**SD**
PE	Professionals with experience in PE	1020	202	345.68	43.99	6.05	1.16
	Professionals without experience in PE	1480	719	358.30	36.39	8.31	3.56

Table 13. Professionals with experience in PE vs professionals without experience in PE: Mean and standard deviation values of the fixation count, the fixation duration and the gaze time in the PE task

Looking at the distribution of visual attention between the ST and TT areas, we noticed that in the translation task the fixation count (*M* = 751, *SD* = 467) and the gaze time (*M* = 4.41, *SD* = 2.28) were higher in the ST areas than in the TT areas (*M* = 533, *SD* = 323 and *M* = 4.03, *SD* = 1.64 respectively) (Table 16) presumably due to more careful reading and understanding of the ST, as well as due to the translators' need not only to feed their brain with input for meaning construction but also to monitor while typing that the TT conveys the meaning of the ST (Carl et al., 2011 and Mesa-Lao, 2014). In line with the findings of previous studies (Mesa-Lao, 2014 and Carl et al., 2011), in the PE task, the fixations (*M* = 386, *SD* = 144) and the gaze time (*M* = 1.95, *SD* = 0.72) on the ST areas decrease considerably, while much of the activity involved in the task takes place in the TT area (*M* = 748, *SD* = 303 and *M* = 4.67, *SD* = 1.43 respectively) (Table 16). According to Elming et al. (2014: 161), this is not surprising since "translation suggestion is already presented for post-editing, so less inspiration from looking at the source is needed". In line with the findings of a previous study (Carl et al., 2011), the number of fixations in the translation task was, in most cases, distributed more evenly on the ST and the TT areas than in the PE task, where the majority of the participants (9 out of 12) had almost twice as many fixations on the TT areas than on the ST areas.

	Fixation count				Total gaze time (mins)			
	ST area		TT area		ST area		TT area	
Task	**Mean**	**SD**	**Mean**	**SD**	**Mean**	**SD**	**Mean**	**SD**
Translation	751	467	533	323	4.41	2.28	4.03	1.64
PE	386	144	748	303	1.95	0.72	4.67	1.43

Table 146. Cognitive effort per task: Mean and standard deviation values of the fixation count and the gaze time per text area

5. Conclusions and Future Work

Although the sample is small, taking into account the length of the texts and the number of participants, our initial study indicates clearly that the effort needed by professional translators when post-editing NMT output is less than the effort required when translating comparable texts from scratch. In particular, the study showed that professional translators needed less time (temporal effort) for post-editing NMT output compared to the time required for translating from scratch, leading, thus, to a time saving of almost 20%. Keyboard activity (technical effort) was almost triple in the translation task, where insertions were more and deletions were less than in the PE task. Furthermore, the analysis reveals a higher cognitive effort in the translation task, with more and longer fixations and higher average gaze time. When translating from scratch, a more careful reading and a better understanding of the ST is evident from the higher fixation count and total gaze time on the ST area. In the PE task, on the other hand, much of the activity took place in the TT area.

Another interesting finding that emerges from the study is that professional translators with experience in PE expend less temporal, technical and cognitive effort during PE from professional translators with no experience in PE. Although the professionals' PE experience is not extensive and although the results are not statistically significant, they are still indicative of the importance that experience can play in the effort required during PE. It is our intention in the future to build on this research by increasing sample sizes and target languages and by complementing the results with a qualitative analysis of the final translation and post-edited products in order to ascertain if (and how) quality is affected. In addition, we aim to study whether translation experience and areas of specialization and expertise may affect the results.

Acknowledgments

The authors would like to thank the HUBIC Lab at the Athena Research Center in Athens where all the experiments were carried out.

References

Balling, Laura and Michael Carl. (2014). Production time across language and tasks: A large scale analysis using the CRITT translation process database. In John W. Schwieter and Aline Ferreira (eds), *The development of translation competence: Theories and methodologies from psycholinguistics and cognitive science*. Cambridge Scholars Publishing, pp. 239–268.

Besacier, Laurent and Lane Schwartz. (2015). Automated translation of a literary work: a pilot study. In *Proceedings of the Fourth Workshop on Computational Linguistics for Literature* (Denver, CO), pp. 114–122.

Carl, Michael. (2012). Translog – II: A program for recording user activity data for empirical reading and writing research. In *Proceedings of the 8th international conference on language resources and evaluation*, European Language Resources Association (ELRA).

Carl, Michael, Barbara Dragsted, Jakob Elming, Daniel Hardt and Arnt Lykke Jakobsen. (2011). The process of post-editing: A pilot study. In *Proceedings of the 8th international NLPCS workshop –*

Special theme: Human-machine interaction in translation. Copenhagen Studies in Language 41. Samfundslitteratur, Copenhagen, pp. 131–142.

Carl, Michael, Silke Gutermuth and Silvia Hansen-Schirra. (2015). Post-editing machine translation: Efficiency, strategies, and revision processes in professional translation settings. In Aline Ferreira and John W. Schwieter (eds) *Psycholinguistic and cognitive inquiries into translation and interpreting*. John Benjamins, Amsterdam, pp 145–174.

Castilho, Sheila, Joss Moorkens, Federico Gaspari, Rico Sennrich, Vilelmini Sosoni, Yota Georgakopoulou, Pintu Lohar, Andy Way, Antonio Valerio, Antonio Valerio Miceli Barone and Maria Gialama. (2017). A Comparative quality evaluation of PBSMT and NMT using professional translators. In *Proceedings of Machine Translation Summit XVI*. Nagoya, Japan.

de Almeida, Giselle. (2013). *Translating the post-editor: An investigation of post-editing changes and correlations with professional experience*. PhD Thesis, Dublin City University.

Densmer, Lee. 2014. Light and Full MT Post-Editing Explained. http://info.moravia.com/blog/bid/353532/Light-and-Full-MT-Post-Editing-Explained.

Doherty, Stephen, Sharon Brien and Michael Carl. (2010). Eye tracking as an MT evaluation technique. *Machine Translation* 24:1-13.

Elming, Jakob, Laura Winther Balling and Michael Carl. (2014). Investigating user behaviour in post-editing and translation using the CASMACAT workbench. In Sharon O’Brien, Laura Winther Balling, Michael Carl, Michel Simard and Lucia Specia (eds.) *Post-editing of machine translation*. Cambridge Scholars Publishing, Newcastle.

Flanagan, Marian and Tina Paulsen Christensen. (2014). Testing post-editing guidelines: how translation trainees interpret them and how to tailor them for translator training purposes. *The Interpreter and Translator Trainer* 8(2):257–275.

Garcia, Ignacio. (2011). Translating by post-editing: Is it the way forward? *Machine Translation*, 25(3): 217-237. http://www.jstor.org/stable/41487495

Gaspari, Federico, Antonio Toral, Sudip Kumar Naskar, Declan Groves and Andy Way. (2014). Perception vs reality: Measuring machine translation post-editing productivity. In *Proceedings of AMTA workshop on post-editing technology and practice*. Vancouver, pp. 60–72.

Genzel, Dimitriy, Jakob Uszkoreit and Franz Och. (2010). Poetic Statistical Machine Translation: Rhyme and Meter. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, MIT, Massachusetts, pp. 158–166.

Green, Spence, Heer, Jeffrey and Christopher D. Manning. (2013). The efficacy of human post-Editing for language translation. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (ACM)*. Association for Computing Machinery, 439-448.

Greene, Erica, Tugba Bodrumlu and Kevin Knight. (2010). Automatic analysis of rhythmic poetry with applications to generation and translation. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, Cambridge, MA, pp. 524–533.

Groves, Declan and Dag Schmidtke. (2009). Identification and analysis of post-editing patterns for MT. *MT Summit XII – The twelfth Machine Translation Summit International Association for Machine Translation hosted by the Association for Machine Translation in the Americas*. Association for Machine Translation in the Americas, pp. 429-436.

Guerberof, Anna. (2009). Productivity and quality in MT post-editing. In Goulet MJ et al. (eds.) *Beyond translation memories workshop*. MT Summit XII, Ottawa. Association for Machine Translation in the Americas.

Jones, Ruth, Ann Irvine. (2013). The (un)faithful machine translator. In *Proceedings of the 7th Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*. Sofia, Bulgaria, pp. 96–101.

Hu, Ke and Patrick Cadwell. 2016. A comparative study of post-editing guidelines. In *Proceedings of the 19th annual conference of the European association for machine translation*, pp. 346–353.

Hvelplund, Kristian Tangsgaard. (2011). *Allocation of cognitive resources in translation: An eye-tracking and key-logging study*. PhD thesis, Copenhagen Business School.

Jia, Yanfang, Michael Carl and Xiangling Wang. (2019). How does the post-editing of neural machine translation compare with from-scratch translation? a product and process study. *The Journal of Specialised Translation* 31:60–86.

Koponen, Maarit. (2016). *Machine translation post-editing and effort: Empirical Studies on the post-editing effort*. PhD Thesis, University of Helsinki.

Koponen, Maarit. (2012). Comparing human perceptions of post-editing effort with post-editing operations. In *Proceedings of the 7th workshop on statistical machine translation*. Montreal, Canada.

Krings, Hans P. (2001). *Repairing texts: Empirical investigations of machine translation post-editing processes*. Geoffrey S. Koby (ed.). Kent, Ohio: Kent State University Press.

Lommel, Arle and Donald A. DePalma. (2016). *Europe's leading role in Machine Translation: How Europe is driving the shift to MT*. Technical report. Common Sense Advisory, Boston.

Mesa-Lao, Bartolomé. (2013). Introduction to post-editing - the CasMaCat GUI. http://bridge.cbs.dk/projects/seecat/material/hand- out_post- editing_bmesalao.pdf.

Mesa-Lao, Bartolomé. (2014). Gaze behaviour on source texts: An exploratory study comparing translation and post-editing. In Sharon O'Brien, Laura Winther Balling, Michael Carl, Michel Simard & Lucia Specia (eds.), Post-editing of Machine Translation, 219–245. United Kingdom: Cambridge Scholars Publishing.

Mitchell, Linda. (2015). *Community post-editing of machine-translated user-generated content. PhD thesis*. Dublin City University.

Moorkens, Joss. (2018). Eye tracking as a measure of cognitive effort for post-editing of machine translation. In Walker Calum and Federico M. Federici (eds.) *Eye tracking and multidisciplinary studies on translation*. John Benjamins, Amsterdam, pp. 55-69.

Moorkens, Joss and Sharon O'Brien. (2015). Post-editing evaluations: Trade-offs between novice and professional participants. In İlknur Durgar El-Kahlout, Mehmed Özkan, Felipe Sánchez-Martínez, Gema Ramírez-Sánchez, Fred Hollowood and Andy Way (eds.) *Proceedings of European Association for Machine Translation* (EAMT) 2015, Antalya, pp. 75–81.

Moorkens, Joss, Antonio Toral, Sheila Castilho and Andy Way. (2018). Translators' perceptions of literary post-editing using statistical and neural machine translation. *Translation Spaces* 7(2): 242-260.

O'Brien, Sharon. (2002) Teaching post-editing: A proposal for course content. In *Proceedings of 6th EAMT workshop on teaching machine translation,* Manchester, UK, pp 99–106.

O'Brien, Sharon. (2007) An empirical investigation of temporal and technical post-editing effort. *Translation and Interpreting Studies (TIS)* 2(1): 83-136.

O'Brien, Sharon. (2009) Eye tracking in translation process research: methodological challenges and solutions. In Inger M. Mees, Fabio Alves & Susanne Göpferich (eds.) *Methodology, technology and innovation in translation process research: A tribute to Arnt Lykke Jakobsen*. Copenhagen studies in language, 38. Samfundslitteratur, Copenhagen, pp. 251-266.

O'Brien, Sharon. (2010). Introduction to post-editing: Who, what, how and where to next. *Paper presented at The Ninth Conference of the Association for Machine Translation in the Americas* (Denver, Colorado 31 October – 4 November 2010).

O'Brien, Sharon. (2011). Towards predicting post-editing productivity. *Machine Translation* 25(3):197-215.

O'Brien, Sharon, Laura Winther Balling, Carl Michael, Michel Simard and Lucia Specia. (2014). *Post-editing of machine translation: Processes and applications*. Cambridge Scholars Publishing, Newcastle.

O'Brien, Sharon and Michel Simard. (2014). Introduction to special issue on post-editing. *Machine Translation* 28(3):159–164.

Plitt, Mirko and François Masselot. (2010). A productivity test of statistical machine translation post-editing in a typical localisation context. *The Prague Bulletin of Mathematical Linguistics* 93:7–16.

Raptis, Spyros and Maria Giagkou. (2016). From capturing to generating human behavior: closing the interaction loop at the hubic lab. In *Proceedings of the 20th pan-hellenic conference on informatics (pci) with international participation*. Partas, Greece: ACM Digital Library, International Conference Proceedings Series.

Sharmin, Selina, Oleg Spakov, Kari-Jouko Räihä, and Arnt Lykke Jakobsen. (2008), Where on the screen do translation students look while translating, and for how long?. In Susanne Göpferich, Arnt Lykke Jakobsen and Inger M. Mees (eds.) *Looking at eyes. Eye-tracking studies of reading and translation processing*. Samfundslitteratur, Copenhagen, pp. 31-51.

Tatsumi, Midori. (2009). Correlation between automatic evaluation metric scores, post-editing speed and some other factors. *MT Summit XII – The twelfth Machine Translation Summit International Association for Machine Translation hosted by the Association for Machine Translation in the Americas.* Association for Machine Translation in the Americas, pp. 332-339.

TAUS. (2016). Taus post-editing guidelines. https://www.taus.net/think-tank /articles/postedit-articles/taus-post-editing-guidelines

Toral, Antonio and Andy Way. (2015). Machine-assisted translation of literary text: A case study. *Translation Spaces* 4(2):241–268.

Vermeer, Hans. (1989). Skopos and commission in translational action. In Andrew Chesterman (ed.) *Readings in translation theory*. Routledge, London, pp.173-187

Vieira, Lucas Nunes. (2016). How do measures of cognitive effort relate to each other? A multivariate analysis of post-editing process data. *Machine Translation* 30: 41-62.

Vieira, Lucas Nunes and Elisa Alonso. (2018). *The use of machine translation in human translation workflows: Practices, perceptions and knowledge exchange*. Report. Institute of Translation and Interpreting.

Vieira, Lucas Nunes, Elisa Alonso and Lindsay Bywood. (2019). Introduction: post-editing in practice – process, product and networks. *The Journal of Specialised Translation* 31:2–13.

Wagner, Emma. (1985). Post-editing systran – A challenge for commission translators. *Terminologie et Traduction* 3:1–7.

Zhechev, Ventsislav. (2014). Analysing the post-editing of machine translation at autodesk. In Sharon O'Brien, Laura Winther Balling, Carl Michael, Michel Simard and Lucia Specia (eds.) *Post-editing of machine translation: Processes and application*. Cambridge Scholars, pp. 2–13.