ARCHIVIO ISTITUZIONALE DELLA RICERCA

Alma Mater Studiorum Università di Bologna
Archivio istituzionale della ricerca

Investigating usability in postediting neural machine translation: Evidence from translation trainees' self-perception and performance

This is the final peer-reviewed author's accepted manuscript (postprint) of the following publication:

Published Version:
Wang, X., Wang, T., Muñoz Martín, R., Jia, Y. (2021). Investigating usability in postediting neural machine translation: Evidence from translation trainees' self-perception and performance. ACROSS LANGUAGES AND CULTURES, 22(1), 100-123 [10.1556/084.2021.00006].

Availability:
This version is available at: https://hdl.handle.net/11585/821353 since: 2021-06-02

Published:
DOI: http://doi.org/10.1556/084.2021.00006

Terms of use:

Some rights reserved. The terms and conditions for the reuse of this version of the manuscript are specified in the publishing policy. For all terms of use and more information see the publisher's website.

This item was downloaded from IRIS Università di Bologna (https://cris.unibo.it/).
When citing, please refer to the published version.

(Article begins on next page)

04 July 2024

This is a postprint, i.e., the authors' manuscript after the corrections made in the course of the peer review process. Hence, it is NOT the is not the version of register, to be found at https://doi.org/10.1556/084.2021.00006. This is the full correct citation:
Wang, X., Wang, T., Muñoz Martín, R., & Jia, Y. 2021. Investigating usability in postediting neural machine translation texts. *Across Languages & Cultures* 22 (1): 100–123.

INVESTIGATING USABILITY IN POSTEDITING NEURAL MACHINE TRANSLATION TEXTS:

EVIDENCE FROM TRANSLATION TRAINEES' SELF-PERCEPTION AND PERFORMANCE

Xiangling WANG[1*], Tingting WANG[1], Ricardo MUÑOZ MARTÍN[2], & Yanfang JIA[3]

[1] Hunan University | [2] MC2 Lab, University of Bologna | [3] Hunan Normal University

Abstract: This is a report on an empirical study on the usability for translation trainees of neural machine translation systems when post-editing (MTPE). Sixty Chinese translation trainees completed a questionnaire on their perceptions of MTPE's usability. Fifty of them later performed both a post-editing task and a regular translation task, designed to examine MTPE's usability by comparing their performance in terms of text processing speed, effort, and translation quality. Contrasting data collected by the questionnaire, keylogging, eyetracking and retrospective reports we found that, compared with regular, unaided translation, MTPE's usefulness in performance was remarkable: (1) it increased translation trainees' text processing speed and also improved their translation quality; (2) MTPE's ease of use in performance was partly proved in that it significantly reduced informants' effort as measured by (a) fixation duration and fixation counts; (b) total task time; and (c) the number of insertion keystrokes and total keystrokes. However, (3) translation trainees generally perceived MTPE to be useful to increase productivity, but they were sceptical about its use to improve quality. They were neutral towards the ease of use of MTPE.

Keywords: Usability; neural machine translation post-editing; perception; effort; eyetracking

* Corresponding author: Prof Wang Xiangling. School of Foreign Languages. Hunan University. 2 Lushan South Road, Yuelu, Changsha, Hunan Province, China.

1. INTRODUCTION

Newly released neural machine translation systems have achieved great progress in improving output quality, when compared to phrase-based machine translation and statistic machine translation systems (Bahdanau *et al.* 2014; Yamada 2019). The advantage of neural machine translation also holds true for Chinese, a logographic language, as measured by automatic or human evaluation (Junczys-Dowmunt *et al.* 2016; Jia, Carl & Wang 2019a). Neural MT systems have experienced a spectacular progress but fully automated high-quality translation remains a distant possibility. Thus, the question remains unanswered whether neural machine translation post-editing is ready for deployment.

Post-editing (PE) is the traditional means for achieving publication-ready translations (Koponen 2016). *Machine translation post-editing* (MTPE, henceforth) is a human-computer interactive mode of revising translations. Advances in machine translation (MT) systems and the growth of translation needs have made MTPE a mainstream set-up in the translation industry (TAUS 2019). Thus, MTPE is an increasingly prevalent working strategy in professional workflow and everyday translation tasks. That is why it has been under the spotlight in several research efforts focusing on both observational and introspective variables.

1.1 Measurable Variables

Output quantity—or, in some research projects, *productivity*—is one of the major concerns of the industry. Several studies have explored MTPE usefulness in performance and reported mixed results: Some studies indicated that MTPE enabled translators to process text faster than when translating without MT aids (O'Brien 2007; Guerberof 2009; Plitt & Masselot 2010) whereas García (2010) and Lee & Liao (2011) found MTPE took as long as, or even slightly longer than, regular translating. These contradictory results in output quantity or productivity may be due to variations in language pairs and text types.

Other studies go deeper to investigate translators' effort expended when at task. De Sousa, Aziz & Specia (2011) studied MTPE effort with human raters, but the scores were extremely complicated and highly variable. O'Brien (2007) used keylogging and reported that MTPE required less temporal and technical effort than *regular, unaided translation* (RUT, henceforth).[1] This finding was supported by Green *et al.* (2013), Läubli *et al.* (2013), Carl, Gutermuth & Hansen-Schirra (2015) and Koglin (2015). Based on interstroke pauses, Jia, Carl & Wang (2019a) found MTPE to be cognitively easier. M TPE triggered significantly fewer and shorter pauses than RUT. In contrast, Screen (2017) found that informants had longer pauses in MTPE.

Besides pauses, and based on Just & Carpenter's (1980) eye-mind assumption, eyetracking metrics are generally assumed to be effective indicators of cognitive effort in translation. More and longer fixations are assumed to be associated to more cognitive effort. The allocation of cognitive resources on source texts and target texts during PE

and RUT is quite different (Carl, Gutermuth & Hansen-Schirra 2015; Daems *et al*. 2017). Based on source-text fixations' higher counts and longer timespans, Da Silva *et al.* (2017) found that RUT was cognitively more effortful.

Studies addressing translation quality have also used different methods. Fiederer & O'Brien's (2009) human raters assigned MTPE texts higher values in accuracy and clarity, but lower ones in style than RUT. Lee & Liao's (2011) found that MTPE was helpful in reducing errors in final versions. Depraetere *et al.* (2014), however, observed a slight (statistically insignificant) decrease in quality in MTPE texts. On the other hand, Screen (2017) and Jia, Carl & Wang (2019a) reported that the quality of MTPE's output can be equivalent to that of RUT, and Screen (2019) found end-users to share the same positive opinion of MTPE text quality. The above research generally suggests that MTPE is useful in improving output quantity without decreasing quality. This advantage, nevertheless, may not be enough if users perceive it otherwise.

1.2 Users' Perceptions

Research on MTPE has been undertaken from various perspectives but informants' perceptions have been relatively sidelined. Guerberof (2013) found that professional translators had mixed feelings towards MTPE due to the uneven quality of MT systems and the types of their texts and projects. Translators—especially, professionals—may perceive that MTPE will yield texts of poorer quality than RUT (Yamada 2015). Rossi (2019) tested an adapted Technology Acceptance Model with a questionnaire and claimed that MT output quality was not related to professional translators' perceived usefulness of MT.

To better understand the factors involved in translators' adoption of MT, Cadwell *et al.* (2016, 2018) carried out two focus-group studies with professional translators. MT was not consistently adopted for all tasks, though they also observed broadly positive attitudes to it. Informants reported several reasons for their (non-)adoption of MT, such as text type, language pair, quality, and trust. Similarly, Briggs (2018) reported that most students used MT tools but had limited trust in the accuracy of their output. Castilho & O'Brien (2017) showed that implementing light PE increased the quality and acceptability of MT output, which further led to higher satisfaction in end-users. Jia, Carl & Wang (2019a) found that Chinese translation trainees generally maintained a positive attitude towards MTPE but found it challenging as well. In view of the centrality of the relationship between computers and translators in our project, we turned to some notions of Human-Computer Interaction in order to contribute to solve the discrepancies between observational and introspective assessments.

1.3 Usefulness and Ease of Use

Usability or 'ease of use' (Miller 1971) is a central concept in Human-Computer Interaction. Usability research emphasizes user-centrality and human factors and it is

reasonable to expect that almost all human activities are prone to be investigated from the point view of usability (Suojanen *et al.* 2015). Research on the usability of MTPE can provide insights for MT-system development, as well as for translator training. Nevertheless, usability tests have been conducted mainly in engineering areas such as system design in the context of marketing, game development and e-learning (e.g., Yousef *et al.* 2015; Travis 2017; Revythi & Tselios 2019; Thorpe *et al.* 2019).

Shackel (1991:24) referred to the usability of a system or equipment as "the capability to be used by humans easily and effectively", with effectiveness, learnability, flexibility, and attitude as operational criteria. Nielsen (1993) suggested that usability is a multi-dimensional construct consisting of learnability, efficiency, memorability, errors, and satisfaction. As extensive studies on usability are conducted, the norm ISO 9241-11: 2018 has made usability an ergonomic requirement for office work with visual display terminals. Here, *usability* is "the extent to which a system, product or service can be used by specified users to achieve specified goals with effectiveness, efficiency and satisfaction in a specified context of use". The three criteria (effectiveness, efficiency and satisfaction) are further divided into seven indicators for measurement.

As illustrated, usability has not been defined consistently across standards and scholars. The sketched criteria are either too broad or extremely complex, with a large number of determinants to measure. In order to be both comprehensive and straightforward, the present study operationally defines *usability* as usefulness + ease of use (Hartson 1998; Lund 2001). In MTPE, *usefulness* indicates the degree to which a translator's performance can be enhanced, both in terms of output quantity and translation quality; *ease of use* measures the effort required to complete a translation task.

The reviewed research projects provided inspiration for our study, but logographic languages are rarely discussed in this realm and they mainly involved alphabetic languages such as French, German, Portuguese and Spanish. Furthermore, most studies mainly investigate the process and product of post-editing statistical MT output, so more research on the latest neural MT post-editing seems in order. This study attempts to explore the usability of neural MT post-editing for English-Chinese. In order to contribute to usability research within Translation Studies, it investigates the usability of MTPE systems (Figure 1), with evidence from both informants' perceptions and their performance through data collected by means of a questionnaire, keylogging, eyetracking and retrospective reports, to attempt to answer the following research questions:

(1) Compared with RUT, what is the *usefulness* of MTPE as measured by output quantity and translation quality in the informants' performance?
(2) Compared with RUT, what is the *ease of use* of MTPE as measured by temporal and typing length and cognitive effort in the informants' performance?
(3) What are translation trainees' perceptions about the usability of MTPE systems as measured by *perceived usefulness* and *perceived ease of use*?

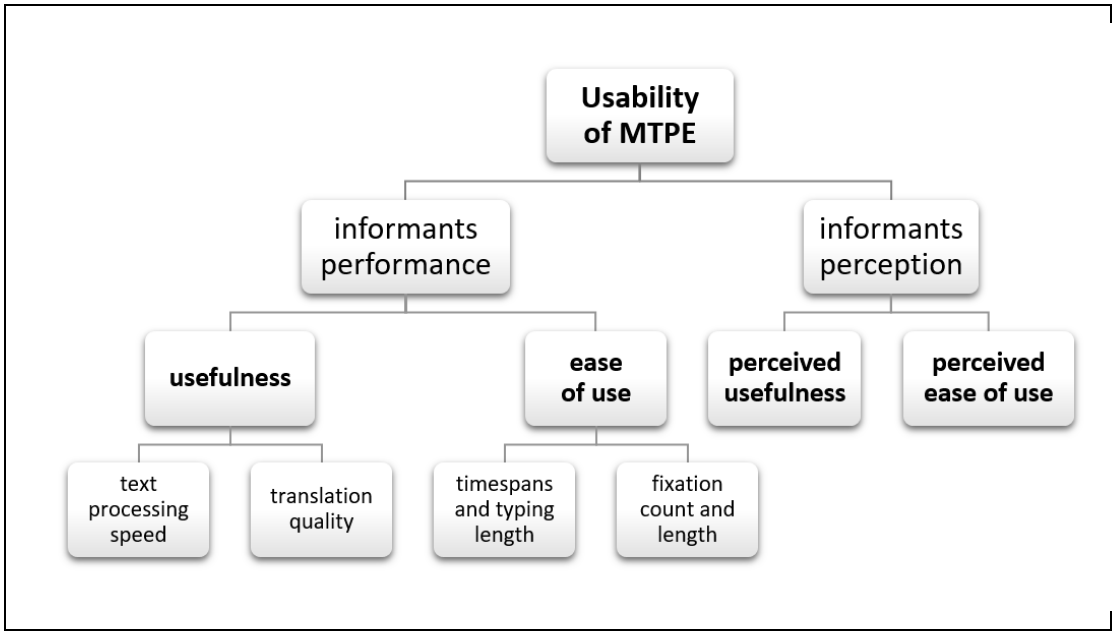

Figure 1. Internal structure of the research construct *Usability of MTPE*

2. METHODOLOGY

2.1 Informants

Translation trainees in universities are the main source of future professional translators in the language service industry. Knowing how they perceive and use MTPE should be useful in translator training. Hence, this study chose translation trainees as the informants. A convenience sampling technique was used in this study for the questionnaire on informants' perception of MTPE usability. A total of 60 postgraduate translation students at Hunan University, China, volunteered to take part in the present study, 48 females and 12 males. They were 23.75 years old on average (range 21–37, SD=2.672). They all had Chinese as L1 and English as L2. None of them had been brought up in a bilingual context. They had learned English for 13.78 years on average (range 8–26, SD=3.370). They all passed the Test for English Majors-Band 8.[2] They had acquired basic notions about MTPE in a 'Computer-Aided Translation' course. None of them had hands-on training in MTPE for professional purposes. Informants P1 to P50 (N=50) volunteered to participate in the experimental tasks to investigate MTPE's usability in performance—the codes were assigned afterwards. They were 45 females and 5 males. Their average age was 23.32 years old (range 21–29, SD=1.392). They had normal or corrected-to-normal vision.

2.2 Materials

Two English general texts were selected (Appendix A): ST1, for MTPE, had 141 words and ST2, for RUT, 142. They were considered similar from the scope of several quantitative indicators (Table 1). Both texts were news excerpts which did not entail

additional knowledge to understand them (informal assessment with excluded classmates), and no online Chinese translations of these texts were found. ST1 was pre-translated with Google Neural Machine Translation system (output obtained May 12th, 2019). Translation briefs introducing the target audience and quality expectations were provided to all informants. Post-editing guidelines (TAUS 2016) were also provided as instructions in the post-editing task.

Our three-part questionnaire (Appendix B) was adapted from the scales for *Perceived Usefulness* and *Perceived Ease of Use* by Davis (1989). They were designed to measure user acceptance of information technology. Part I collected demographic information. Part II contained six items on *perceived usefulness* of MTPE and Part III, six more items on *perceived ease of use* of MTPE. All the 12 items were measured with a 5-point Likert scale, in which 1 meant *strongly disagree* and 5, *strongly agree.*

Table 1. Quantitative profiling of ST1 and ST2

Indicators	**ST1**	**ST2**
Length (in # of words)	141	142
Number of sentences	7	6
U.S. grade level	14	14
Flesch Kincaid Reading Ease	42.1	44.8
Gunning Fog score	15.2	15.8
SMOG index	11.8	11.8
Coleman Liau index	14	13.2
Automated readability index	14.7	14.2
LIX	52 (difficult)	50 (difficult)
Lexile measures	1210L-1400L	1210L-1400L

2.3 Experimental Procedures

The experiment was approved by the Ethics Committee of the College of Foreign Languages at Hunan University. All informants were guaranteed both anonymity and confidentiality. All of them signed an Informed Consent form before the experiment. They were paid a ¥20 reward for their work.

A pilot study was conducted to test the comprehensibility of the questionnaire as well as the performance of the keylogger and eyetracker. Defects spotted in the pilot study were modified. For example, expressions of some items in the questionnaire were refined to avoid ambiguity, and the font size of the texts was enlarged in order to be easily read. The pilot informants were excluded from the formal experiment.

The formal experiment was conducted at the end of a 'Computer-Aided Translation' course in May to June 2019. The questionnaires were distributed to 60 respondents and data were collected from all informants. Thus, the requirement was met that the number of respondents to questionnaire items be at least 5:1 (Bentler & Chou 1987). After filling out the questionnaire, the informants carried out the MTPE and RUT tasks. All 50 informants were tested individually in the eyetracking laboratory. The

site is soundproof and is equipped with steady artificial lighting. Their fixation activities were recorded with an Eyelink 1000 plus eyetracker (1000 Hz).

The informants' typing activities were recorded by Translog-II (Carl 2012). Figure 2 shows the interface set during the post-editing task. The double-spaced source text was displayed on the upper window of the screen in Times New Roman at 14 point size. The neural MT double-spaced Chinese output, with the typeface SimSun at 13 point size, was placed on the lower window. During the RUT task, the lower window was left blank for the target Chinese text. In accordance with translation briefs and TAUS post-editing guidelines, the informants were asked to produce publishable translations and use as much neural MT output as possible during MTPE.

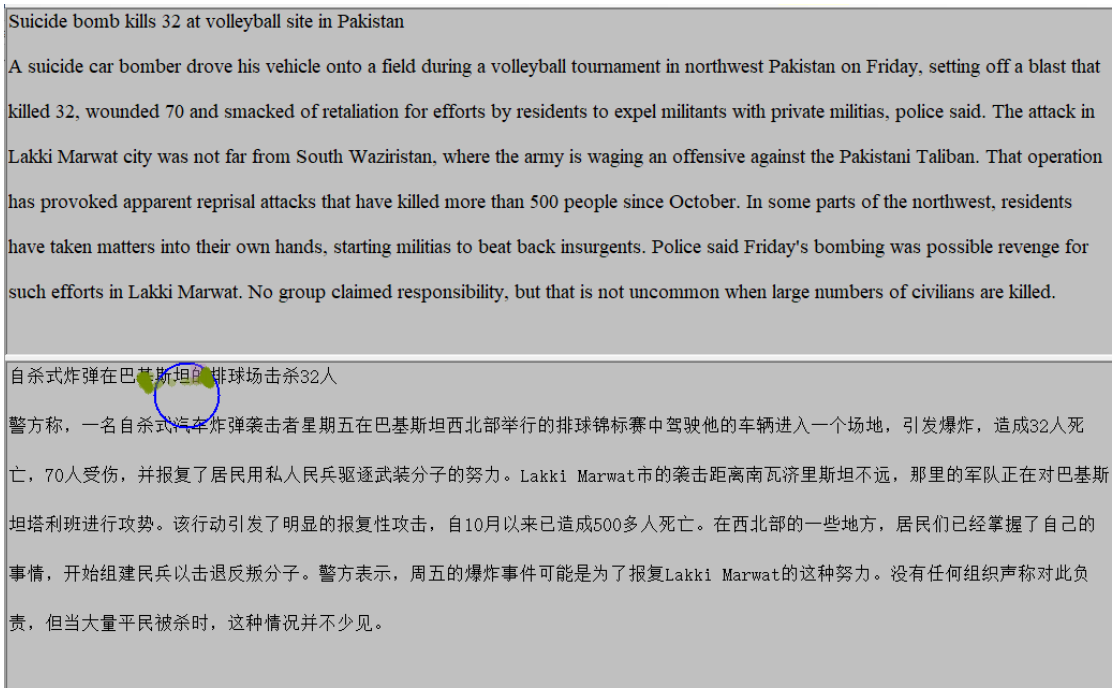

Figure 2. Screenshot of a post-editing session in Translog-II user windows. The circle, dots and the word highlighted in pink in the lower (target text) window represent gaze fixation, gaze samples and GW-mapping respectively.

The informants first completed a warm-up task to get used to the translation interface and keyboard in the lab, and made adjustments when needed. Then, they post-edited the neural MT output of ST1 and translated the ST2 with no MT aid. The two tasks were conducted in random order to avoid sequence effects. Informants had no access to the Internet or other resources so as to prompt their gaze data to stay on the Translog-II interface. No time limits were imposed, and the informants were given time to relax between the two tasks. Once they had completed both of them, their keylogged translation process and eyetracked gaze data were replayed. Then they were asked to write a retrospective report on their feelings during the tasks, as well as the differences between two translation modes, namely MTPE and RUT. Data collected from the keylogger, eyetracker and retrospective report were conflated for analysis.

2.4 Data Exclusion and Quality Assessment of Eyetracking Data

All 60 respondents returned valid questionnaires. In the following tasks, one informant's logged file of RUT (P25_TT1) was lost because of system error. Thus, 99 logged files from 50 informants were correctly saved. The logged files were manually aligned on YAWAT, a tool for word alignment (Germann 2008). Afterwards, final translations and a set of tables containing keylogging and eyetracking data were processed.

The quality of all collected eyetracking data was assessed before data analysis. Two criteria were adopted: *Gaze Time on Screen*, or GTS; and *Mean Fixation Duration*, or MFD (Hvelplund 2011, 2014). The threshold for GTS was one standard deviation below the mean; for MFD, it was 200 ms (Pavlović & Jensen 2009; Hvelplund 2011). The records of 15 informants were discarded because their MTPE or RUT eyetracking data were invalid. In total, 60 questionnaires, 99 final translations, 70 recording data from 35 informants and 50 written retrospective reports were considered for analysis.

3. DATA ANALYSIS AND DISCUSSION

As stated in §2.2, informants' perception of MTPE usability was measured with the questionnaire, validly returned by 60 translation trainees. Cronbach's Alpha scores for the questionnaire's reliability and validity yielded 0.869 for Part II (Perceived Usefulness of MTPE, 6 items); 0.734 for Part III (Perceived Ease of Use of MTPE, 6 items) and 0.839 for all 12 items. All results were above the threshold of 0.7 (Nunnally & Bernstein 1994), which suggests that the questionnaire is reliable. The score in the Kaiser-Meyer-Olkin (KMO) test for sampling adequacy is 0.806 and the Bartlett's Test is at significant level ($p < 0.01$), which means the questionnaire has a good structural validity and was suitable for further analysis.

3.1 Performance: Usefulness

As defined in the study, the usefulness of MTPE was examined and compared with RUT from both text processing speed and translation quality perspectives. Paired t-test and Wilcoxon test were mainly used in §3.1 and 3.2, depending on whether data were normally distributed.

3.1.1 Text processing speed

Text processing speed is one of the primary concerns in the translation industry, for it is closely related to costs in time and human resources. Following Guerberof (2009), text processing speed was measured by the number of ST tokens translation trainees processed per second. The results of the two tasks are plotted in Figure 3. Text processing speed of MTPE was significantly higher than that of RUT ($p<0.01$). Therefore, using MTPE enables translation trainees to improve their output quantity. This finding is in line with informants' perception reflected through the questionnaire and retrospective reports (see §3.3 and 3.4). Since this study focuses on "general language" texts, our

results support Daems (2017) and Lu & Sun (2018) but differ from da Silva (2017) and Jia, Carl & Wang (2019a), which found no significant difference between MTPE and RUT in text processing speed. The inconsistency with da Silva (2017) was probably due to the fact that the participants involved in his experiment were professional translators who had more related experience. On the other hand, the inconformity with Jia, Carl & Wang (2019a) might be explained by the difference in experimental design. In Jia, Carl & Wang's (2019a) experiment, participants had access to the Internet during tasks, so they were able to use online dictionaries or other external resources. In that condition, the advantage of MTPE might be weakened.

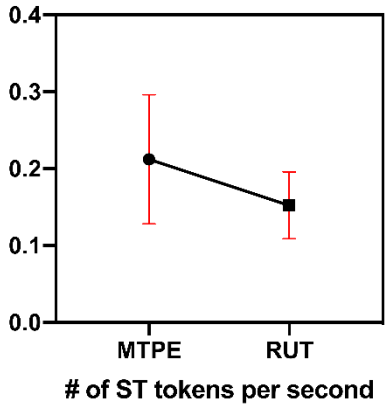

Figure 3. Text processing speed in the two tasks, in terms of number of ST tokens per second.

3.1.2 Translation quality

In performance, the usefulness of MTPE should involve not only improving text processing speed, but also producing a translation quality comparable to that of RUT (Figure 1). This section studies the quality of final MTPE and RUT translations. TAUS's Dynamic Quality Evaluation Framework (TAUS 2013) has two evaluation dimensions, *adequacy* and *fluency*, and was adopted for translation quality assessment (TQA). Both dimensions can be rated from 1 to 4 points, with 1 indicating the poorest, and 4 the best quality, as introduced in Tables 4 and 5. Two PhD candidates in translation studies with L1 Chinese who had worked as teaching assistants in translation courses were invited to assess the quality of the translations. They were provided with reference translations made by a translator with 10 years of professional experience. The nature of each translation, whether of MTPE or RUT origin, was not disclosed.

We measured the inter-rater reliability with Spearman's rank correlation coefficient. The resulting correlation coefficients for adequacy and fluency were 0.553 and 0.368 respectively. The scores show that the agreement between the two raters was moderate or even low. This result is in accordance with our expectations, for assessing translation quality is extremely complicated and raters' score of a translation may be influenced by subjective factors that are difficult to control or probe into. Low reliability between raters was also found in previous researches (e.g., Carl *et al.* 2011; Vieira 2016; Jia, Carl & Wang 2019b). The average scores from raters as to *adequacy* and *fluency* were calculated to measure the translation quality of the two tasks (Figure 4).

Table 4. Rating scales and operational definitions used for assessing *adequacy*

Rating scale	Examples
4. Everything All the meaning in the ST is present in the TT, no more, no less	ST: *A suicide car bomber drove his vehicle onto a field during a volleyball tournament in northwest Pakistan on Friday.* TT: 一名自杀式汽车炸弹袭击者星期五在巴基斯坦西北部举行的排球锦标赛中驾驶他的车辆进入一个场地。(P46_P1) BT: A suicide car bomber drove his vehicle onto a field during a volleyball tournament in northwest Pakistan on Friday.
3. Most Almost all the meaning in the ST is present in the TT	ST: *A suicide car bomber drove his vehicle onto a field during a volleyball tournament in northwest Pakistan on Friday.* TT: 一名自杀式汽车炸弹袭击者星期五在巴基斯坦西北部举行的排球锦标赛中驾驶车辆进入排球场。(P48_P1) BT: A suicide car bomber drove a vehicle onto the volleyball field during a volleyball tournament in northwest Pakistan on Friday.
2. Little Fragments of the meaning in the ST are present in the TT	ST: *At around 10pm a car containing the device was abandoned close to the gates of the County Down courthouse, which is protected by thick security walls.* TT: 晚上 10 点左右，装载炸弹的汽车被遗弃在县级下法院的大门口，汽车被厚厚的安全墙保护住。(P05_T1) BT: At around 10pm, a car containing the bomb was abandoned at the gates of the county's lower courthouse. The car was protected by thick security walls.
1. None The meaning in the ST is not present in the TT	ST: *At around 10pm a car containing the device was abandoned close to the gates of the County Down courthouse, which is protected by thick security walls.* TT: County Down 酒馆大门在厚实的安全墙下显得很隐蔽，大约晚上 10 点时，这辆载满物品的车正好丢弃于此。(P30_T1) BT: Protected by thick security walls, the gates of the pub *County Down* was hidden. At around 10pm the car full of goods was abandoned just here.

Key: ST, source text; TT, target text; BT back translation.

Table 5. Rating scales and operational definitions used for assessing *fluency*

Rating scales	Examples
4. Flawless A perfectly flowing text with no errors	ST: *A suicide car bomber drove his vehicle onto a field during a volleyball tournament in northwest Pakistan on Friday.* TT: 一名自杀式汽车炸弹袭击者于星期五在巴基斯坦西北部举行的排球锦标赛中驾驶汽车进入场地。(P49_P1) BT: A suicide car bomber drove a vehicle onto the field during a volleyball tournament in northwest Pakistan on Friday.
3. Good A smoothly flowing text with a number of minor errors	ST: *A car bomb last night exploded in the Northern Ireland city of Newry, sending the political message that rebel republicans remain intent on attacking the Irish peace process.* TT: 昨夜，北爱尔兰纽瑞市发生一起汽车炸弹爆炸事故。事故传递出反对派共和党决意进一步扰乱爱尔兰和平进程的政治意图。(P21_T1) BT: Last night, a car bomb exploded in the city of Newry, Northern Ireland. The incident sent the political message that the rebel republicans determined to further attack the Irish peace process.
2. Disfluent A poorly written text that is difficult to understand	ST: *A car bomb last night exploded in the Northern Ireland city of Newry, sending the political message that rebel republicans remain intent on attacking the Irish peace process*. TT: 下午 10 点左右，一辆配有装置的汽车被停放在了下议院的大门前，尽管那里有厚重的安全墙保护。(P49_T1) BT: At around 10pm, a car equipped with the device was parked at the gates of the House of Commons, although it was protected by thick security walls.
1. Incomprehensible A very poorly written text that is impossible to understand	ST: *A car bomb last night exploded in the Northern Ireland city of Newry, sending the political message that rebel republicans remain intent on attacking the Irish peace process.* TT: 昨夜，纽里北爱尔兰城发生了一起汽车爆炸，意为反叛共和党为打击爱尔兰和平进程而作出的政治反对。(P24_T1) BT: Last night, a car bomb exploded in the Newry city of Northern Ireland, meaning is rebel republicans' political oppositions for attacking Irish peace process.

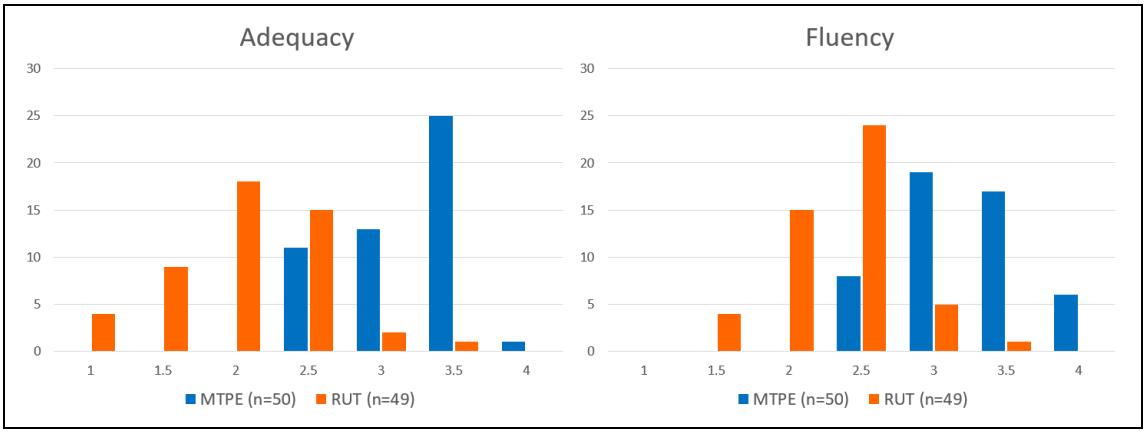

Figure 4. Translation quality assessment of the MTPE and RUT tasks on *adequacy* and *fluency* (RUT from P25 was lost). Scale 1–4, where 1 is the lowest, and 4 the highest, level of quality.

Figure 4 reveals that for both *adequacy* and *fluency,* informants' translate MTPE texts were overall rated higher than RUT texts. Furthermore, no RUT text received a higher *adequacy* score than MTPE texts. Wilcoxon test reveals that the difference in *adequacy* between MTPE (M=3.173) and RUT (M=2.051) is significant (p<0.01). On the whole, results for *fluency* are similar, with three exceptions (P09, P11, P34). The difference in *fluency* between MTPE (M=3.214) and RUT (M=2.337) is also statistically significant (p<0.01). In contrast with Screen (2017) and Jia, Carl & Wang (2019a), which found no significant difference in quality between MTPE and RUT, our results support García (2011) and thus support the usefulness of MTPE in improving translation quality.

The above results may be interpreted as a positive impact of neural MT output. In order to induce informants to exert cognitive effort, the two source texts had been intentionally (albeit intuitively) chosen to be complex, so as to increase the task difficulty levels. The informants, we thought, might thus encounter unknown words during both tasks. However, neural MT output seems to have provided candidate Chinese translations of these words for them, so that they could hypothesize likely translations of their choice. In other words, the informants had no access to any information resources, but the neural MT output seems to have helped them to understand the meaning of unknown words and hence, the *adequacy* of MTPE translations was improved. Besides, Popović (2018) found neural MT systems to be more advanced in processing natural language for translation. Though not perfect, the neural MT candidates helped informants to process long and complex sentences. Thus, the *fluency* of MTPE translations could be better than that of RUT.

3.2 Performance: Ease of Use

This section focuses on the other aspect of usability in performance, namely, ease of use. *Ease of use* measures the efforts required in completing the translation tasks with MTPE (§1.3). This section concentrates on comparing Krings' (2001) temporal effort (typing timespans), technical effort (typing length in keystrokes) and cognitive effort required during MTPE and RUT.

3.2.1 Temporal length

Firstly, we addressed the differences in temporal duration between the two tasks. We distinguished between *initial span* and *total task time*. The initial span may be indicative of more or less previous cognitive elaboration prior to action. It is what Carl, Dragsted & Jakobsen (2011) label *orientational phase* and Carl, Schaeffer & Bangalore (2016:20–21) define as 'the time offset from the beginning of the session until the first keystroke, which coincides with the end of the orientation phase' (end marked as TimeD).[3] The results are displayed in Figure 5.

Translation trainees had longer initial spans in MTPE than in RUT. The initial span data was not normally distributed, so Wilcoxon test was used and the difference was found not significant (p=0.522). This result is in line with Screen (2017), who did not find statistical differences in initial spans between MTPE and RUT. The slightly longer initial spans during MTPE can simply be explained by the fact that translation trainees have to read both the source text and the neural MT output before pressing the first key to post-edit. There was a significant difference between the total times of the two tasks (p<0.01). This reduction of temporal length in total task time supports the finding in text processing speed (§3.1.1) and also suggests that 'post-editing newspaper texts may save time' (Koglin 2015:132).

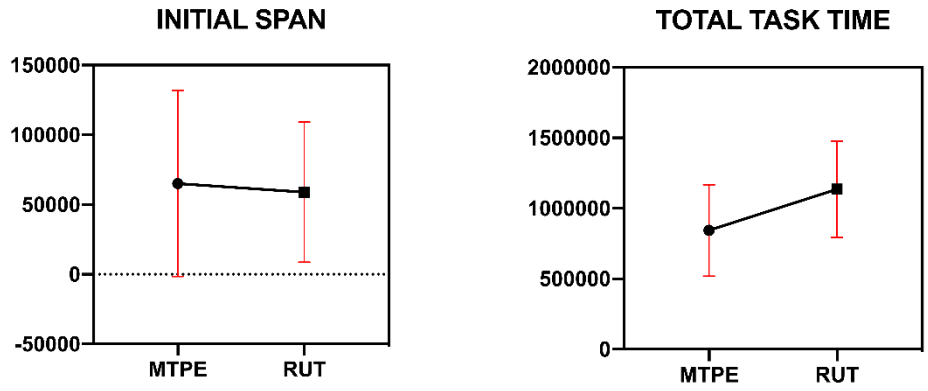

Figure 5. Temporal length of initial spans and total task times in MTPE and RUT, in ms.

3.2.2 Typing length

The experimental task MTPE is assumed to entail no addition of new text except when changes are introduced, often after having deleted some text; that is, with the exception of fixing omissions, the business of MTPE is more substituting language chains than building new text. The main goal of MTPE is fixing errors, and informants might be prone to perform more corrections than they would in RUT circumstances (cf. Mellinger & Shreve 2006). Thus, the typing length during MTPE and RUT was measured by total keystrokes, which could be separated into the number of manually inserted and deleted

characters as recorded by Translog-II (Figure 6). As expected, MTPE significantly reduced informants' insertions and total keystrokes, when compared with RUT (p<0.01). Interestingly, the mean of MTPE deletions (98.00) was higher than that of RUT (88.31), but the difference was not significant (p=0.411). This finding is in accordance with Carl *et al.* (2011) and da Silva *et al.* (2015). It reveals that, MTPE may require more deletions by translators to improve the neural MT output, but also that it still reduces insertion and total keystrokes and thus proves its ease of use.

These results stand to reason. When translator trainees were asked to translate a text from-scratch, they could draft their translation of a segment, not of a single word, in their minds, then type it on the computer quite confidently. Thus, they performed a lot of insertions and few deletions, though they might revise it later. In contrast, when they were asked to post-edit an NMT output to achieve publishable quality, they already had a draft, so the drafting step was omitted. Thus, translator trainees' insertions and total keystrokes could be significantly reduced. However, due to the imperfect nature of NMT output, they had to delete the incorrect NMT suggestions, then type their own translations. Post-editing was mainly performed at word or phrase level rather than segment or sentence level, so their deletions would not significantly increase, compared with RUT.

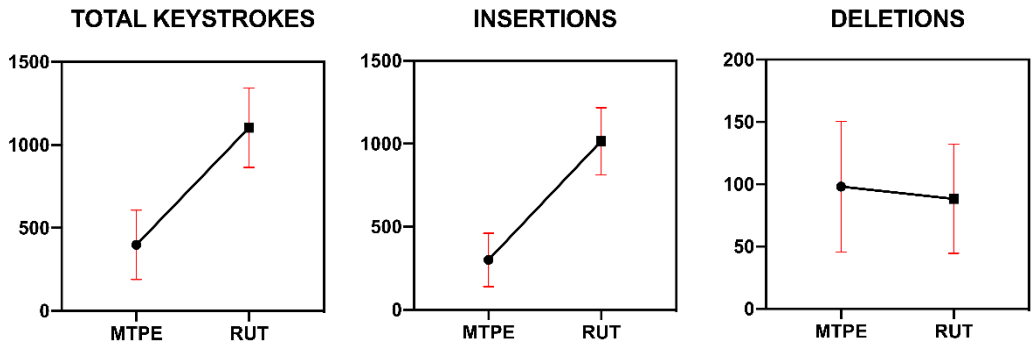

Figure 6. Typing length of MTPE and RUT tasks, measured by the total number of keystrokes, insertions and deletions.

3.2.3 Cognitive effort

According to the eye-mind assumption, 'there is no appreciable lag between what is being fixated and what is being processed' (Just & Carpenter 1980:331). Hence, here the cognitive effort was measured by fixation duration in ms, and fixation counts, as recorded by Eyelink 1000 plus. The eyetracking data were not distributed normally, so a Wilcoxon test was applied. Figure 7 shows that, first, the fixation duration was significantly lower in MTPE (p<0.01). On average, it was reduced by 206.5 ms (24.59%). Second, fixation counts were also significantly lower in MTPE (p<0.01). Contrary to Screen (2017), who found that MTPE was not easier than RUT, the results confirm those by Carl, Gutermuth & Hansen-Schirra (2015), da Silva *et al.* (2015), and Lu & Sun (2018), who also measured cognitive effort by fixation duration and fixation counts.

The results were also in line with O'Brien (2007) and Jia, Carl & Wang (2019a), who used pauses as indicators of cognitive effort. Since translation trainees were provided with neural MT output, they reported that their "cognitive effort spent in comprehending source text was greatly reduced" and they "only needed to modify errors and perfect translations". This is reasonable, since translator trainees did not have to totally rely on the source text to produce a translation, especially to understand the content. A part of cognitive effort needed to understand the source text might be devoted to the NMT suggestions. However, the NMT output was in Chinese and the quality might be acceptable, making it easier for them to read it over and comprehend it.

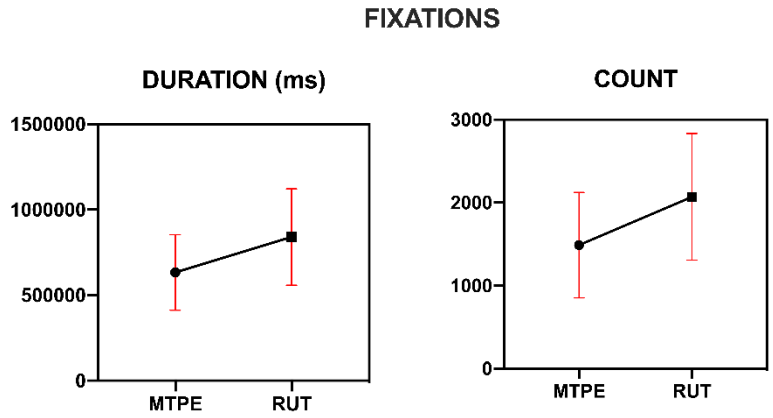

Figure 7. Cognitive effort in MTPE and RUT tasks as measured by fixation duration (in ms) and fixation counts.

3.3 Perception: Perceived Usefulness

Table 6 shows the descriptive analysis of Perceived Usefulness (PU) of MTPE. The mean score of the six items for MTPE in Part II (Perceived Usefulness) is 3.925 (SD=0.275) indicating that translation trainees generally deem it useful. Translation trainees agree or nearly agree in five out of six items (PU1, PU3, PU4, PU5, PU6) concerning MTPE's usefulness in boosting speed, increasing productivity, and enhancing effectiveness. However, PU2—"Using MTPE would improve my translation quality"—gets the lowest mean score among all six items (M=3.40, SD=0.887), and 10 out of 60 informants even had negative values in PU2. This implies that translation trainees are somewhat neutral about the MTPE's usefulness in improving translation quality. In their retrospective reports, some translation trainees frankly expressed that they were more satisfied with RUT for its higher quality. They felt that language of translations produced through MTPE was inflexible, leading to obvious translationese. This result echoes professional translators' skepticism about the quality of MTPE (Yamada 2015), but should be taken with a grain of salt in view that, when assessing RUT quality, informants were evaluating their own output.

Table 6. *Perceived usefulness* of MTPE

Item	N	Min	Max	Mean	Std Dev
PU1	60	3	5	4.13	0.676
PU2	60	2	5	*3.40*	0.887
PU3	60	2	5	4.12	0.739
PU4	60	2	5	3.93	0.756
PU5	60	2	5	3.90	0.796
PU6	60	2	5	4.07	0.778
Avg.	**60**	**2**	**5**	**3.925**	**0.772**

3.4 Perception: Perceived Ease of Use

Table 7 presents the descriptive analysis of *perceived ease of use* of MTPE. Compared with *perceived usefulness*, the mean scores of *perceived ease of use* are much lower for both single items (range 3.32–3.73) and their average (M=3.59, SD=0.143). The results indicate that translation trainees are relatively neutral about the ease of use for learning, operating and interacting MTPE systems. The results can be interpreted by their comments in retrospective report. Many of them expressed that "the output of machine translation was too confusing to post-edit", "they would be misled by the MT output" or "their creativity as well as subjectivity was greatly limited". This is in line with results in Jia, Carl & Wang (2019a), where students felt they had less freedom and less room to show their creativity. They had to invest additional effort to deal with MT output and to adjust themselves to task demands. They found MTPE was not easy to use or to interact with because they were distracted from their routine workflow, and they would hesitate about reading the source text or the MT output first, or about comparing them sentence by sentence.

Table 7. *Perceived ease of use* of MTPE

Item	N	Min	Max	Mean	Std Dev
PEU1	60	1	5	3.60	0.867
PEU2	60	1	5	3.32	0.833
PEU3	60	2	5	3.63	0.712
PEU4	60	2	5	3.58	0.720
PEU5	60	2	5	3.68	0.701
PEU6	60	1	5	3.73	0.756
Avg.	**60**	**1**	**5**	**3.59**	**0.765**

4. CONCLUSION

Our study applied usability research to Translation Studies by investigating the usability of MTPE from both informants' perception and performance perspectives. The research questions—whether (1) post-editing machine-translated texts was more useful and (2) its ease of use was higher than regular, unaided translating; and whether (3) informants perceived them to be so—were addressed by combining data collected with a questionnaire, keylogging, eyetracking and retrospective reports.

First. We tested whether MTPE was useful performance through data collected by a keylogger and translation quality assessment. Compared with regular translation, the text processing speeds of translation trainees were significantly enhanced, although in their retrospective report a few informants said MTPE slowed them down. As to quality, MTPE did better than expected. MTPE translations received higher scores in both adequacy and fluency. Thus, MTPE proved to be useful in performance, in that it was faster and also resulted in translations of higher quality. Some informants linked the enhanced speed with the acceptable neural MT raw output. Therefore, and in contrast with the consensual expectation that MT systems are only appropriate for domain-specific texts, our results reveal that this neural MT system was suitable for translating a "general language" text from English into Chinese.

Second. We measured ease of use of MTPE in performance by timespans and typing length, and by the cognitive effort required during translation tasks, as collected by means of eyetracking. There was no significant difference in initial spans between MTPE and RUT, but the informants' total task times were significantly reduced in MTPE (M=843.6 s). On average, and compared with RUT (M=1,135.2 s), task times were 25.69% shorter. The number of insertions and total keystrokes was significantly reduced. The results show that MTPE was convenient to translators, for they did not need to type every character by themselves. Finally, cognitive effort—measured by fixation durations and counts—was significantly lower in MTPE than in RUT. Our results support the usability of MTPE.

Third. We tapped into the informants' perception of MTPE's usability with a questionnaire. Translation trainees generally perceived MTPE to be useful to improve translating speed and productivity, but they were skeptical about its usefulness to improve translation quality. As in Guerberof (2013), several translation trainees preferred regular translations because some neural MT segments were disastrous. As for *perceived ease of use*, translation trainees gave less favorable scores than in *perceived usefulness*. They found themselves stuck with neural MT output, with little freedom to show their subjectivity. They also felt they had to invest additional effort to decide working procedures. Thus, for them, interacting with neural MT during MTPE seems to have been felt effortful to some extent.

Our study has some limitations. For example, the only group of informants was translation trainees with little formal translation experience. Besides, we studied usability of MTPE in informants' perception and performance separately, without delving into their relationship. Future research projects may conduct comparisons by inviting more informants with different backgrounds, such as translation and post-editing professionals. Moreover, whether and how informants' perception affects their

performance, and vice versa, should be tested as well.

Since the respondents and informants of the study are translation trainees, the findings may have implications for translation training. Teachers and trainers should incorporate MTPE training in their curricula to improve translation trainees' post-editing skills and overall translation expertise. Besides, MT system developers can also get useful data for their system designing and developing.

Note

1. Krings (2001) broke down post-editing effort into three different but related categories, namely *temporal*, *technical* and *cognitive* effort. Here, Krings' temporal and technical efforts are taken to actually be additional, indirect indicators of cognitive effort, rather than independent indicators of different efforts, but clarifying this classical misunderstanding cannot be the goal of this paper. When talking about our own work we will, however, refer to them respectively as *(typing) timespans* and *typing lengths (in keystrokes)*, which is in our view an unequivocal way to refer to what they measure.

2. Test for English Majors-Band 8, or TEM-8, is based on the highest level of standard for English major students in China. It is taken at the eighth term. TEM-8 evaluates students' English ability in listening, reading, writing and translating.

3. Again, this is not the place to develop a theoretical argument about our conceptual or terminological preferences. Suffice it to say that, in most usual understandings of *drafting*, it is done throughout the whole task. Orientation rarely seems to stop with the first keystroke. Furthermore, initial spans nearly never cover the whole ST and some later spans might also be considered of equal nature (i.e., devoted to planning).

REFERENCES

Bahdanau, D., Cho, K. & Bengio, Y. 2014. Neural Machine Translation by Jointly Learning to Align and Translate. https://arxiv.org/abs/1409.0473

Bentler, P. M., & Chou, C. P. 1987. Practical issues in structural modeling. *Sociological Methods & Research* Vol. 16. No. 1. 78–117.

Briggs, N. 2018. Neural Machine Translation Tools in the Language Learning Classroom: Students' Use, Perceptions, and Analyses. *The JALT CALL Journal* Vol. 14. No. 1. 3–24.

Cadwell, P., Castilho, S. & O'Brien, S. 2016. Human Factors in Machine Translation and Post-editing among Institutional Translators. *Translation Spaces* Vol. 5. No. 2. 222–243.

Cadwell, P., O'Brien, S. & Teixeira, C. 2018. Resistance and Accommodation: Factors for the (non-)Adoption of Machine Translation among Professional Translators. *Perspectives.* Vol. 26. No. 3. 301–321.

Carl, M. 2012. Translog-II: A Program for Recording User Activity Data for Empirical Reading and Writing Research. Paper presented at the *Eight International Conference on Language Resources and Evaluation* (Istanbul, Turkey, May 23–25).

Carl, M., Dragsted, B., & Jakobsen, A. L. 2011. A Taxonomy of Human Translation

Styles, *Translation Journal*, Vol. 16, No. 2.

Carl, M. Dragsted, B., Elming, J., Hardt, D. & Jakobsen, A. L. 2011. The Process of Post-editing: A Pilot Study. In: Sharp, B., Zock, M., Carl, M., & Lykke Jakobsen, A. (eds) *Human-Machine Interaction in Translation: Proceedings of the 8th International NLPCS Workshop*. Frederiksberg: Samfundslitteratur. 131–142.

Carl, M., Schaeffer, M. & Bangalore, S. 2016. The CRITT Translation Process Research Database. In: Carl, M., Bangalore, S. & Schaeffer, M. (eds) 2016. *New Directions in Empirical Translation Process Research: Exploring the CRITT TPR-DB*. New York: Springer. 13–54.

Carl, M., Gutermuth, S. & Hansen-Schirra, S. 2015. Post-editing Machine Translation: Efficiency, Strategies and Revision Processes in Professional Translation Settings. In: Ferreira, A. & Schwieter, J. W. (eds). *Psycholinguistic and Cognitive Inquiries into Translation and Interpreting*. Amsterdam: John Benjamins. 145–175.

Castilho, S. & O'Brien, S. 2017. Acceptability of Machine-translated Content: A Multi-language Evaluation by Translators and End-users. *Linguistica Antverpiensia, New Series: Themes in Translation Studies* Vol. 16, 120–136.

Da Silva, I. L., Schmaltz, M., Alves, F., Pagano, A., Wong, D., Chao, L., Leal, A. L. V. Quaresma, P. & Garcia, C. 2015. Translating and Post-editing in the Chinese-Portuguese Language pair: Insights from an Exploratory Study of Key Logging and Eye Tracking. *Translation Spaces* Vol. 4. No. 1. 145–169.

Da Silva, I. Alves, F., Schmaltz, M., Pagano, A., Wong, D., Chao, L., Leal, A. L. V. Quaresma, P., Garcia, C. and da Silva, G. E. 2017. Translation, Post-editing and Directionality: A Study of Effort in the Chinese-Portuguese Language Pair. In: Jakobsen, A. L. & Mesa Lao, B. (eds) *Translation in Transition: Between Cognition, Computing and Technology*. Amsterdam: John Benjamins. 107–134.

Daems, J., Vandepitte, S., Hartsuiker, R. J. & Macken, L. 2017. Translation Methods and Experience: A Comparative Analysis of Human Translation and Post-editing with Students and Professional Translators. *Meta* Vol. 62. No. 2. 245–270.

Davis, F. D. 1989. Perceived Usefulness, Perceived Ease of Use, and User Acceptance of Information Technology. *MIS Quarterly* Vol. 13, 319–340.

De Sousa, S. C., Aziz, W. & Specia, L. 2011. Assessing the Post-editing Effort for Automatic and Semi-automatic Translations of DVD Subtitles. In Angelova. G., Bontcheva, K., Mitkov, R. & Nicolov, N. (eds) *Proceedings of the Recent Advances in Natural Language Processing Conference*. Hissar, Bulgaria. 97–103.

Depraetere, I., de Sutter, N. & Tezcan, A. 2014. Post-edited Quality, Post-editing Behavior and Human Evaluation: A Case Study. In: O'Brien, S., Balling, L. W., Carl, M., Simard, M. & Specia, L. (eds). *Post-editing of Machine Translation: Processes and Applications*. Newcastle: Cambridge Scholars Publishing. 78–108.

Fiederer, R. & O'Brien, S. 2009. Quality and Machine Translation: A Realistic Objective? *The Journal of Specialised Translation* Vol. 11, 52–74.

Garcia, I. 2010. Is Machine Translation Ready yet? *Target* Vol. 22. No. 1. 7–21.

Garcia, I. 2011. Translating by Post-editing: Is it the Way Forward? *Machine Translation* Vol. 25. 217–237.

Germann, U. 2008. Yawat: Yet Another Word Alignment Tool. In *Proceedings of the*

ACL-08: HLT Demo Session (Companion Volume). 20–23.

Green, S., Heer, J. & Manning, C. 2013. The Efficacy of Human Post-editing for Language Translation. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery. 439–448.

Guerberof, A. 2009. Productivity and Quality in MT Post-editing. In *MT Summit XII-Workshop: Beyond Translation Memories: New Tools for Translators MT*. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.575.5398&rep=rep1&type=pdf

Guerberof, A. 2013. What do Professional Translators Think about Post-editing? *The Journal of Specialised Translation* Vol. 19. 75–95.

Hartson, H. R. 1998. Human–Computer Interaction: Interdisciplinary Roots and Trends. *Journal of Systems and Software* Vol. 43. No. 2. 103–118.

Hvelplund, K. T. H. 2011. *Allocation of Cognitive Resources in Translation: An Eye-tracking and Key-logging Study.* Copenhagen: Copenhagen Business School.

Hvelplund, K. T. 2014. Eye Tracking and the Translation Process: Reflections on the Analysis and Interpretation of Eye-tracking Data. In: Muñoz, R. (ed.) *Minding Translation. MonTI* Special Issue 1. 201–223.

ISO 9241-11 (1998. *Ergonomic Requirements for Office Work with Visual Display Terminals (VDTs), Part 11: Guidance on Usability*. Geneva: International Organization for Standardization.

Jia, Y., Carl, M. & Wang, X. 2019a. Post-editing Neural Machine Translation versus Phrase-Based Machine Translation for English–Chinese. *Machine Translation* Vol. 33. No. 1-2. 9–29.

Jia, Y., Carl, M. & Wang, X. 2019b. How Does the Post-editing of Neural Machine Translation Compare with From-scratch Translation? A Product and Process study. *The Journal of Specialised Translation* Vol. 31. 60–86.

Junczys-Dowmunt, M., Dwojak, T. & Hoang, H. 2016. Is Neural Machine Translation Ready for Deployment? A Case Study on 30 Translation Directions. https://arxiv.org/abs/1610.01108

Just, A. & Carpenter, P. 1980. A Theory of Reading: From Eye Fixation to Comprehension. *Psychological Review* Vol. 4, 329–354.

Koglin, A. 2015. An Empirical Investigation of Cognitive Effort Required to Post-edit Machine Translated Metaphors Compared to the Translation of Metaphors. *Translation & Interpreting* Vol. 7. 126–141.

Koponen, M. 2016. Is Machine Translation Post-editing Worth the Effort? A Survey of Research into Post-editing and Effort. *The Journal of Specialised Translation* Vol. 25, 131–148.

Krings, H. P. 2001. *Repairing Texts: Empirical Investigations of Machine Translation Postediting Processes*. Kent, OH: Kent State University Press.

Läubli, S. Fishel1, M., Massey, G., Ehrensberger-Dow, M.& Volk, M. 2013. Assessing Post-editing Efficiency in a Realistic Translation Environment. In: O'Brien, S., Simard, M. & Specia, L. (eds). *Proceedings of MT Summit XIV Workshop on Post-editing Technology and Practice*, Nice. 83–91.

Lee, J. & Liao, P. 2011. A Comparative Study of Human Translation and Machine

Translation with Post-editing. *Compilation and Translation Review* Vol. 4. No. 2. 105–149.

Lu, Z. & Sun, J. 2018. An Eye-tracking Study of Cognitive Processing in Human Translation and Post-editing. *Foreign Language Teaching and Research* Vol. 50. No. 5. 760–769.

Lund, A. M. 2001. Measuring Usability with the USE Questionnaire. *Usability Interface* Vol. 8. No. 2. 3–6.

Mellinger, Ch., & Shreve, G. M. 2006. Match Evaluation and Over-editing in a Translation Memory Environment. In: Muñoz, R. (ed.) *Reembedding Translation Process Research*. Amsterdam: John Benjamins. 131–148.

Miller, R. B. 1971. *Human Ease of Use Criteria and their Tradeoffs. IBM Report TR 00.2185*. Poughkeepsie, NY: IBM Corporation.

Nielsen, J. 1993. *Usability Engineering*. London: Academic Press.

Nunnally, J. & Bernstein, I. 1994. *Psychometric theory 3E*. New York: McGraw-Hill.

O'Brien, S. 2007. Eye-tracking and Translation Memory Matches. *Perspectives* Vol. 14, 185–205.

Pavlović, N. & Jensen, K. T. H. 2009. Eye Tracking Translation Directionality. In: Pym, A. & Perekrestenko, A. (eds) *Translation Research Projects 2.* Tarragona: Universitat Rovira i Virgili. 101–119.

Plitt, M. & Masselot, F. 2010. A Productivity Test of Statistical Machine Translation Post-editing in a Typical Localization Context. *The Prague Bulletin of Mathematical Linguistics* Vol. 93. 7–16.

Popović, M. 2018. Language-related Issues for NMT and PBMT for English–German and English–Serbian. *Machine Translation* Vol. 32. No. 3. 237–253.

Revythi, A., & Tselios, N. 2019. Extension of Technology Acceptance Model by using System Usability Scale to Assess Behavioral Intention to Use E-learning. *Education and Information Technologies* Vol. 24. No. 4. 2341–2355.

Rossi, C. 2019. Uses and Perception of Machine Translation at the European Commission. *The Journal of Specialised Translation* Vol. 31, 177–200.

Screen, B. 2017. Machine Translation and Welsh: Analysing free Statistical Machine Translation for the Professional Translation of an Under-researched Language Pair. *The Journal of Specialised Translation* Vol. 28, 317–344.

Screen, B. 2019. What Effect does Post-editing Have on the Translation Product from an End-user's Perspective? *The Journal of Specialised Translation* 31. 133–157.

Shackel, B., and Richardson, S. J. (eds) 1991. *Human Factors for Informatics Usability*. Cambridge: Cambridge University Press.

Suojanen, T., Koskinen, K., and Tuominen, T. 2015. *User-centered Translation*. Abingdon: Routledge.

TAUS. 2013. *Adequacy/Fluency Guidelines*. https://taus.net/academy/best-practices/evaluate-best-practices/adequacy-fluency-guidelines

TAUS. 2016. *TAUS Post-editing Guidelines.* https://www.taus.net/think-tank/articles/postedit-articles/taus-post-editing-guidelines

TAUS. 2019. *TAUS Keynotes Asia 2019* https://www.taus.net/academy/reports/event-reports/taus-keynotes-asia-2019

Thorpe, A., Nesbitt, K., & Eidels, A. 2019. Assessing Game Interface Workload and usability: A Cognitive Science Perspective. In: *Proceedings of the Australasian computer science week multiconference,* ACSW 2019. ACM. 44:1–44:8.

Travis, D. 2017. *E-commerce Usability: Tools and Techniques to Perfect the On-line Experience*. London: CRC Press.

Vieira, Lucas N. 2016. *Cognitive effort in post-editing of machine translation: evidence from eye movements, subjective ratings, and think-aloud protocols*. PhD Thesis. Newcastle University.

Yamada, M. 2015. Can College Students Be Post-editors? An Investigation into Employing Language Learners in Machine Translation plus Post-editing Settings. *Machine Translation* Vol. 29. No. 1. 49–67.

Yamada, M. 2019. The Impact of Google Neural Machine Translation on Post-editing by Student Translators. *The Journal of Specialised Translation* Vol. 31. 87–106.

Yousef, A. M. F., Chatti, M. A., Schroeder, U., & Wosnitza, M. 2015. A Usability Evaluation of a Blended MOOC Environment: An Experimental Case Study. *The International Review of Research in Open and Distributed Learning* Vol. 16. No. 2. 69–93

Appendix A

ST1

Suicide bomb kills 32 at volleyball site in Pakistan

A suicide car bomber drove his vehicle onto a field during a volleyball tournament in northwest Pakistan on Friday, setting off a blast that killed 32, wounded 70 and smacked of retaliation for efforts by residents to expel militants with private militias, police said. The attack in Lakki Marwat city was not far from South Waziristan, where the army is waging an offensive against the Pakistani Taliban. That operation has provoked apparent reprisal attacks that have killed more than 500 people since October. In some parts of the northwest, residents have taken matters into their own hands, starting militias to beat back insurgents. Police said Friday's bombing was possible revenge for such efforts in Lakki Marwat. No group claimed responsibility, but that is not uncommon when large numbers of civilians are killed.
(Source: *The Independent* Jan. 1st, 2010; Length: 141 words)

Google NMT output of ST1

自杀式炸弹在巴基斯坦的排球场击杀 32 人

警方称，一名自杀式汽车炸弹袭击者星期五在巴基斯坦西北部举行的排球锦标赛中驾驶他的车辆进入一个场地，引发爆炸，造成 32 人死亡，70 人受伤，并报复了居民用私人民兵驱逐武装分子的努力。 Lakki Marwat 市的袭击距离南瓦济里斯坦不远，那里的军队正在对巴基斯坦塔利班进行攻势。该行动引发了明显的报复性攻击，自 10 月以来已造成 500 多人死亡。在西北部的一些地方，居民们已经掌握了自己的事情，开始组建民

兵以击退反叛分子。警方表示，周五的爆炸事件可能是为了报复 Lakki Marwat 的这种努力。没有任何组织声称对此负责，但当大量平民被杀时，这种情况并不少见

(obtained May 12th, 2019)

Back translation of Google NMT output

Suicide bomb killed 32 at a volleyball site in Pakistan

Police said a suicide car bomber drove his vehicle onto a field during a volleyball tournament held in northwestern Pakistan on Friday, setting off an explosion that killed 32, injured 70 and retaliated against the effort by residents to expel militants with private militias. The attack in Lakki Marwat city was not far from South Waziristan, where the army is attacking the Pakistani Taliban. The action provoked apparent retaliatory attacks that have killed more than 500 people since October. In some parts of the northwest, residents have mastered their own affairs and began to form militia to defeat the rebels. Police said Friday's bombing was possible to be in retaliation for such efforts in Lakki Marwat. No organization claimed responsibility for this, but it is not uncommon when a large number of civilians are killed.

ST2

A car bomb last night exploded in the Northern Ireland city of Newry, sending the political message that rebel republicans remain intent on attacking the Irish peace process. There were no immediate reports of injuries in the explosion, which took place as police were evacuating the area around the city's courthouse, which is close to one of Northern Ireland's busiest roundabouts. At around 10pm a car containing the device was abandoned close to the gates of the County Down courthouse, which is protected by thick security walls. A spokeswoman for the Police Service of Northern Ireland said "We don't have any indication that anyone was hurt. Police were in the process of evacuating the area when there was an explosion." Last night's bombing bore the hallmarks of an attack by one of the three dissident republican groups which are still violently active.

(Source: *The Independent* Feb. 23rd, 2010; Length: 142 words)

Appendix B

Questionnaire on Users' Perception of Machine Translation Post-editing Usability

Part I. Demographic Information

Gender

Age

Years of learning English

TEM-8 score

I have worked as a professional or freelance translator (with payment from the client or company) •Yes •No

I have received professional Machine translation post-editing training •Yes •No

Part II. Perceived Usefulness of Machine Translation Post-editing (MTPE)

1. Using MTPE in my translation tasks would enable me to accomplish tasks more quickly.

2. Using MTPE would improve my translation quality.
3. Using MTPE in my translation tasks would increase my productivity.
4. Using MTPE would enhance my effectiveness on translation tasks.
5. Using MTPE would make it easier to do my translation tasks.
6. I would find MTPE useful in my translation tasks.

Part III. Perceived Ease of Use of Machine Translation Post-editing (MTPE)

7. Learning to operate MTPE would be easy for me.
8. I would find it easy to get MTPE to do what I want it to do.
9. My interaction with MTPE is clear and understandable.
10. I find MTPE to be flexible to interact with.
11. It would be easy for me to become skillful at using MTPE.
12. I find MTPE easy to use.