Machine Translation (2019) 33:9–29
https://doi.org/10.1007/s10590-019-09229-6

Post-editing neural machine translation versus phrase-based machine translation for English–Chinese

Yanfang Jia[1] · **Michael Carl**[2] · **Xiangling Wang**[1]

Received: 14 July 2018 / Accepted: 21 February 2019 / Published online: 8 March 2019

Abstract
This paper aims to shed light on the post-editing process of the recently-introduced neural machine translation (NMT) paradigm. Using simple and more complex texts, we first evaluate the output quality from English to Chinese phrase-based statistical (PBSMT) and NMT systems. Nine raters assess the MT quality in terms of fluency and accuracy and find that NMT produces higher-rated translations than PBSMT for both texts. Then we analyze the effort expended by 68 student translators during HT and when post-editing NMT and PBSMT output. Our measures of post-editing effort are all positively correlated for both NMT and PBSMT post-editing. Our findings suggest that although post-editing output from NMT is not always significantly faster than post-editing PBSMT, it significantly reduces the technical and cognitive effort. We also find that, in contrast to HT, post-editing effort is not necessarily correlated with source text complexity.

Keywords Neural machine translation · Phrase-based statistical machine translation · Temporal effort · Technical effort · Cognitive effort · Human assessment

✉ Xiangling Wang
xl_wang@hnu.edu.cn

Yanfang Jia
yanfangjia@hnu.edu.cn

Michael Carl
mcarl6@kent.edu

[1] Hunan University, Changsha, China

[2] Kent State University, Kent, OH, USA

1 Introduction

Neural machine translation (NMT) has recently gained great interest in both academia and industry. NMT outperforms phrase-based statistical systems (PBSMT) for many language pairs, in terms of automatic evaluation metrics (Bahdanau et al. 2015; Bojar et al. 2016; Junczys-Dowmunt et al. 2016; Castilho et al. 2018) and human evaluation scores (Castilho et al. 2018; Klubicka et al. 2017). However, we still have little knowledge of how post-editors work with NMT, the differences between post-editing NMT output and post-editing PBSMT, and the potential advantages and challenges of post-editing NMT output. This paper aims to address these issues by assessing the accuracy and fluency of NMT output, as compared to PBSMT output based on human evaluation, and by investigating the temporal, technical, and cognitive effort exerted by the translators during NMT post-editing, as compared to during PBSMT post-editing and human/from-scratch translation (HT). As NMT is still far from perfect, studies on NMT post-editing are not only an effective method of assessing NMT quality, they will also increase our understanding of the potential of NMT post-editing.

This paper thus focuses on the process of post-editing NMT output from English to Chinese, a language pair posing great challenges for MT (Wu et al. 2006; Suo et al. 2012). We first compare the quality of NMT and PBSMT output based on human assessment. We then compare temporal effort and typing behaviors for NMT and PBSMT post-editing and for HT.

2 Related research

2.1 Post-editing effort: SMT vs. from-scratch translation

In a seminal work studying post-editing processes, Krings (2001) defines the standard for measuring post-editing effort from three separate but inter-related dimensions: temporal, technical, and cognitive effort. Temporal effort is the amount of time spent post-editing the MT output. Technical effort involves purely mechanical operations, including deletions, insertions and mouse movements. Cognitive effort refers to the "type and extent of those cognitive processes that must be activated to remedy a given deficiency in a machine translation" (ibid, p. 179). Krings claims that a combination of the three efforts determines the acceptability of post-editing MT when compared to HT. With the help of keylogging tools, temporal and technical effort can be measured, while cognitive effort can only be indirectly investigated, most usually through pause analysis or gaze data.

In the last decade, many studies have investigated the effort involved in post-editing and compared it with HT. Post-editing domain-specific texts has often been found to be faster than HT (O'Brien 2007; Guerberof 2009; Groves and Schmidtke 2009; Tatsumi 2009; Plitt and Masselot 2010). However, for more general text types, post-editing is not always found to be superior in speed. For example, Daems et al.

(2017) reported post-editing to be significantly faster, while Carl et al. (2011) found no significant increase in speed. Screen (2017) reported that post-editing general information text does not reduce the processing time from HT.

The idea behind measuring technical effort is that MT post-editing is expected to reduce the labour caused by typing. Compared to studies on temporal effort, there have not been many studies focusing on this aspect. Koglin (2015) found post-editing news text in English to Brazilian Portuguese requires both fewer insertions and deletions than HT. Carl et al. (2011) showed that post-editing tasks necessitate more deletions, navigation keystrokes and mouse clicks, but fewer insertions when compared to HT. These studies concur that post-editing requires fewer insertions than HT, but the results for deletions are inconsistent.

Regarding cognitive effort, in the early studies, Krings (2001) employed think-aloud protocols (TAPs) to indicate the cognitive effort involved in post-editing. The employment of eye-tracking and keylogging in translation process research has greatly expanded our ability to understand reading and writing processes during translation. The allocation of cognitive resources to the source text and target text, which is indicated by the records of gaze data, is very different in post-editing and HT (Carl et al. 2011; Balling and Carl 2014; Mesa-Lao 2014; Carl et al. 2015; Daems et al. 2017; da Silva et al. 2017). These studies conclude that fixations during post-editing occur more frequently on the target text, whereas fixations during HT tend to be concentrated more on source text.

Pauses between keystrokes during typing are generally agreed to be an effective indicator of cognitive effort in language production including translation (Butterworth 1980; Schilperoord 1996; Jakobsen 1998, 2002; Hansen 2002; Alves 2006). Both Butterworth (1980) and Schilperoord (1996) contend that the number and duration of pauses measured in language production can be connected to processing effort. Pause duration and pause density have both been adopted by researchers to indicate cognitive effort in translation process studies. Longer and more frequent pauses signal higher cognitive effort. However, the results based on pause metrics seem to be far from conclusive. Based on pause duration, some studies (e.g. Koglin 2015) have demonstrated that post-editing triggers shorter total pause duration than HT, while others (e.g. Screen 2017) have showed the opposite results. More recently, Lacruz and Shreve (2014) have introduced the pause-to-word ratio (number of pauses per word) (PWR) as a measurement of cognitive effort. A higher PWR value signals more cognitive effort. Lacruz et al. (2014) have observed that PWR correlates strongly with human-targeted translation edit rate (HTER) (Snover et al. 2006), a metric measuring technical post-editing effort based on the minimum number of edit steps between raw MT and its corresponding post-edited version. Schaeffer et al. (2016) have found that PWR also correlates highly with a gaze-based translation difficulty index (TDI) (Mishra et al. 2013), using the large scale multilingual dataset in the TPR-DB (Carl et al. 2016). This study shows that the values of PWR in post-editing tend to be significantly lower than for HT.

The quality of the MT output is—without doubt—one of the key elements determining how much effort the post-editing task requires temporally, technically and cognitively. The above-mentioned studies are mainly based on SMT, which is also

by far the most widely studied MT method in post-editing process research, whereas NMT post-editing has been scarcely investigated prior to the current special issue.

2.2 Post-editing effort: NMT vs. SMT

The translation quality of NMT systems has been shown to be better than PBSMT in several recent studies (Bojar et al. 2016; Castilho et al. 2018; Klubicka et al. 2017). However, it is not yet clear how post-editors benefit from the potential quality improvement of NMT. To the best of our knowledge, only two studies so far have compared the effort involved in post-editing NMT with that of PBSMT. Castilho et al. (2018) discussed the temporal and technical effort involved in post-editing SMT and NMT performed by three post-editors for English to German, Greek, and Portuguese and by two for English to Russian. They found that the total number of keystrokes were reduced for NMT post-editing in all the four language pairs, while the temporal effort was only slightly reduced for English to German, Greek and Portuguese with NMT, but not for English to Russian. Shterionov et al. (2018) compared the productivity of HT with post-editing NMT and SMT from English to German, Spanish, Japanese, Italian and Chinese performed by three translators. Their results showed that (1) post-editing both NMT and SMT was faster than translating from-scratch, and (2) post-editing NMT was faster for most of the translators for all the language pairs except for English to Chinese. However, both studies involved only three post-editors/translators and the authors have not investigated cognitive effort in the post-editing process.

This article extends this work by investigating the temporal, technical and cognitive effort in HT, and post-editing of NMT and PBSMT outputs produced by Google Translate for English to Chinese. Sixty-eight Chinese first-year Master's-level translation students participated in this study, post-editing and translating (from scratch) two English source texts into Chinese. Detailed information on the participants' profiles and research design will be introduced in the following sections.

3 Experimental texts and their MT quality

3.1 Experimental texts

Two British newspaper texts in the general domain (Jensen 2009, 2011; Carl et al. 2016), were selected for this research. Text 1 consists of 148 words and 11 segments. Text 2 contains 140 words and seven segments. Text 2 is of higher complexity than Text 1 as indicated by human rating of text difficulty, readability level, word frequency and the number of non-literal expressions (idiom, metonyms and metaphors) (see Jensen 2011, pp. 88–93 for an elaborate account of how the complexity of the two source texts was measured). Jensen (2009) investigates the impact of source text complexity on HT. In this study, we check whether the text of higher complexity also has an impact on MT quality and post-editing effort. Care was taken that there were no corresponding Chinese translations of the English source

Table 1 Rating scales and operational definitions used for quality evaluation

Category	Rating scales and operational definitions
Fluency	4. Flawless Chinese: refers to a perfectly flowing text with no errors
	3. Good Chinese: refers to a smoothly flowing text even when a number of minor errors are present
	2. Disfluent Chinese: refers to a text that is poorly written and difficult to understand
	1. Incomprehensible Chinese: refers to a very poorly written text that is impossible to understand
Accuracy	4. Everything: all the meaning in the source is contained in the translation, no more, no less
	3. Most: almost all the meaning in the source is contained in the translation
	2. Little: fragments of the meaning in the source are contained in the translation
	1. None: none of the meaning in the source is contained in the translation

texts online, as the participants were free to use any online resources to complete their tasks. The two English source texts were pre-translated into Chinese by the free online MT engine, Google Translate. The output of Google's NMT system was obtained in May of 2017 and the one for Google's PBSMT output was obtained in April of 2012 during a previous experiment carried out within TPR-DB.

3.2 MT quality: PBSMT vs. NMT

3.2.1 Quality evaluation criteria and procedure

The PBSMT and NMT output of the 18 segments from Text 1 and Text 2 were evaluated by nine first-year Master in Translation and Interpreting (MTI) students, who were not part of the post-editing and HT experiments introduced below. They all had no previous experience of MT quality evaluation. The adequacy and fluency criteria developed within TAUS's Dynamic Quality Evaluation Framework[1] were employed to assess the translations. The nine raters were not informed of the origin of the MT output. They were instructed to (1) assess the fluency of the PBSMT and NMT output without the source texts provided, and (2) assess the adequacy of PBSMT and NMT output with the source texts. Both fluency and accuracy were rated on a 4-point Likert-type scale as introduced in Table 1.

Before the quality evaluation process, the nine raters received an assessment brief, including detailed explanation of the scales and operational definitions of the two criteria with commented scoring examples. They then evaluated five test segments to familiarise with the evaluation criteria of fluency and accuracy. Finally, the NMT and PBSMT output for the 18 segments were shown in two Excel spreadsheets. The raters had to first give the fluency scores for the two versions of the

[1] https://taus.net/academy/best-practices/evaluate-best-practices/adequacy-fluency-guidelines (accessed May 2018).

Table 2 Average scores of the 9 raters for fluency and accuracy (**p < 0.01 ***p < 0.001)

MT	Fluency			Accuracy		
	Text 1	Text 2	Text 1 + Text 2	Text 1	Text 2	Text 1 + Text 2
NMT	2.86 ± 0.85**	3.14 ± 0.61**	2.98 ± 0.76***	2.9 ± 0.77**	3.4 ± 0.52**	3.06 ± 0.69***
	κ = 0.19	κ = 0.159		κ = 0.158	κ = 0.115	
PBSMT	1.97 ± 0.66	2.06 ± 0.92	2.01 ± 0.75	2.08 ± 0.63	2.32 ± 0.84	2.17 ± 0.7
	κ = 0.072	κ = 0.172		κ = 0.034	κ = 0.141	

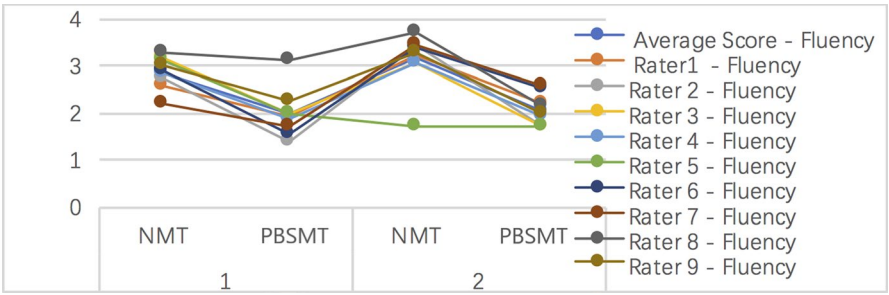

Fig. 1 The 9 raters' scores for fluency for the Text 1 and Text 2

18 segments in the first spreadsheet, after which they provided accuracy scores for these segments in the second spreadsheet.

3.2.2 Quality evaluation results

Table 2 contrasts the average scores of the nine raters for fluency and adequacy of PBSMT and NMT output for the 18 segments across the two texts. To check inter-annotator agreement for fluency and accuracy on the four scales, we computed Fleiss' kappa coefficient (Fleiss 1971) (also presented in Table 2 and denoted by κ). The results show that the nine raters had only slight agreement ($0\% < \kappa \leq 20\%$) for both fluency and accuracy. Despite the low inter-rater agreement, most evaluators (eight out of nine) marked NMT quality higher than PBSMT in both fluency and accuracy and the differences in the average scores were all significant.

For fluency, the NMT segments scored about 3 (Good Chinese) (2.86 for Text 1, 3.14 for Text 2 and 2.98 for the 2 texts combined), while the raw PBSMT output scored around 2 (Disfluent Chinese) (1.97 for Text 1, 2.06 for Text 2 and 2.01 for the 2 texts combined). The difference between NMT and PBSMT in fluency was significant for Text 1 ($p < 0.01$), Text 2 ($p < 0.01$) and the 2 texts combined ($p < 0.001$). Figure 1 shows that eight out of nine raters ranked NMT as more fluent than PBSMT for both texts. Only Rater 5's average scores for the NMT and PBSMT segments were about the same (1.7).

For accuracy, the NMT output also significantly outperformed the PBSMT output for both texts. NMT scored around 3 (Most) (2.9 for Text 1, 3.4 for Text 2 and 3.08

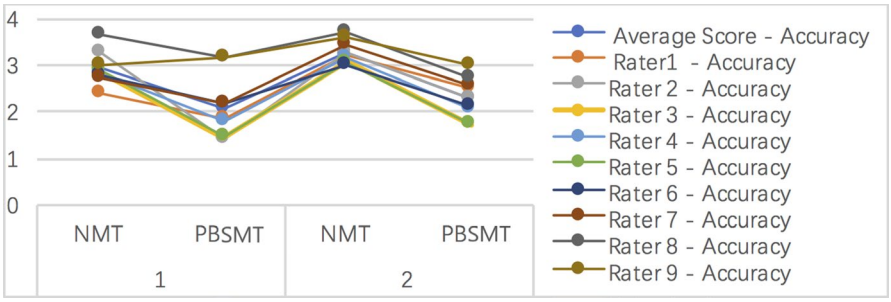

Fig. 2 The 9 raters' scores for accuracy for Text 1 and Text 2

for the 2 texts combined). PBSMT scored about 2 (Little) (2.08 for Text 1, 2.32 for Text 2, and 2.17 for the 2 texts combined). Except for Rater 9, all the raters scored NMT accuracy higher than PBSMT accuracy (Fig. 2). For Text 1, Rater 9's average score for PBSMT (3.18) was slightly higher than for NMT (3.0).

The following four segments show how NMT outperformed PBSMT in accuracy and fluency (Table 3). The underline shows how the source text parts were translated in the corresponding PBSMT and NMT systems, respectively. Recent studies have reported that NMT produces better reordering than PBSMT (Bentivogli et al. 2016; Toral and Sánchez-Cartagena 2017), producing translations that are more natural and accurate. We also found this to be true in our segments, for example in T1S4, where the translation of the modifier adverbial phrase *following the trial* was reordered in the NMT output which was placed before the verb phrase "被判有 (was found guilty of)", making it fit well within the Chinese syntactic structure. The PBSMT translation just followed the source segment order, mistakenly translating it into a modifier-object structure (长期试验后的谋杀罪名/after a long trial's murder).

In addition, we found that the NMT output was more accurate in terms of tense and voice. In T2S7, the present perfect tense of *have increased* was translated into perfect Chinese by adding "了" after "增加(increase)", and the passive voice *could be considered* in T1S11 was also properly translated into "被认为是" in Chinese through "被". However, its PBSMT counterpart did not generate the right tense and voice markers, leading to completely inaccurate Chinese translations. We also notice that the NMT system takes the broader context of the text into consideration and selects the most relevant translation. In T2S7, *bills* was translated properly into 账单(bill), which perfectly matches the context where the suppliers have increased the price, while the PBSMT translated it into "法案(act)" meaning a formal decision made by the legislature, which is an incorrect lexical choice. Wu et al. (2016) also reported that Google NMT reduced translation errors by about 60% compared to it PBSMT output based on human evaluation results. We plan to analyse the error types in detail as part of future research.

Furthermore, the human quality evaluation results also reveal that the MT output of Text 2 (the more complex one) was more accurate and fluent than Text 1 (the less complex one) for both PBSMT and NMT. The differences between Text 1 and Text 2 in fluency and accuracy were more obvious for NMT than PBSMT. The source

Table 3 Examples of the PBSMT and NMT output of the source text segments

Source text segment	PBSMT output	NMT output
T1S4: Yesterday, he was found guilty of four counts of murder following a long trial	昨天, 他被发现犯有四项长期试验后的谋杀罪名 Back translation: Yesterday, he was found guilty of four counts after the long experiment of the murder	昨天, 他被长期审判后被判犯有四项谋杀罪 Back translation: Yesterday, he was found guilty of four counts the murder after the long trial
T1S11: All of them could be considered a burden to hospital staff	他们都可以考虑到医院的工作人员的负担 Back translation: They could all consider the burden of the hospital staff	所有这些都可能被认为是医院工作人员的负担 Back translation: All of these people could be considered a burden to the hospital staff
T2S1: Families hit with increase in cost of living	家庭生活费用增加命中 Back translation: The cost of living of families increase in life	家庭生活成本上涨 Back translation: The cost of living of families increases
T2S7: Five out of the six largest suppliers have increased their customers' bills	五出的六个最大的供应商增加其客户的法案 Back translation: Five out's six suppliers increase their customers' act	六大供应商中有五家增加了客户账单 Back translation: Five out of the six largest suppliers have increased their customers' bills

texts complexity measures introduced by Jensen (2009) do not necessarily correlate with MT quality. Therefore, more complex text indicated by measures of source text complexity, which are tailored for HT, do not necessarily also imply worse MT quality.

4 Post-editing and HT process

4.1 Participants' profile

In total, 68 first-year MTI students participated in this study. They were all native Chinese speakers with English as their second language, aged 22 to 26 years old. The participants had very similar English proficiency. They all had minimal professional translation experience. None of them had prior experience of post-editing. They had all passed the test for English Majors Band 8 (TEM8)[2] or its equivalent before the experiment. All students were undertaking an advanced translation course at the time of the tasks, established for the first-year MTI students, taught by the authors (the first and third authors) during the spring semester (February to June) every year at a Chinese University. Thirty of them were from the course in the semester 2017, and the other 38 were from this course in the semester 2018. This study was approved by the Ethics Committee of the College of Foreign languages at Hunan University. The participants all signed the informed consent form before the experiment. The participants' profile is summarized in Table 4.

4.2 Post-editing and HT procedures

The data from the post-editing tasks were collected from the 38 MTI students of the advanced translation course in May 2018. The whole class was divided into two homogenous groups based on their English proficiency (measured in TEM8 scores). There were 19 students in each of Group1 (G1) and Group 2 (G2). G1 post-edited the NMT output of Text 1 and the PBSMT output of Text 2. G2 post-edited the NMT output of Text 2 and the PBSMT output of Text 1. The TAUS post-editing guidelines (TAUS 2016) for publishable quality were provided for guidance and participants were told that full post-editing for publishable quality was expected. The post-editing tasks were carried out with Translog-II (Carl, 2012). Students also filled a questionnaire after each task regarding their perception of post-editing speed, cognitive effort, quality of the MT output, and of their own translations. The questionnaire results will be reported at a later juncture.

The data for the HT tasks used in this study were collected from another experiment, completed by the 30 MTI students also from this advanced translation course in May 2017. They were also divided into two homogenous groups based on English language proficiency. There were 15 students in each group, named Group 3 (G3) and Group 4 (G4). Text 1 was translated from scratch by G3 and Text 2 was

[2] The Test for English Majors Band 8 is a national English test for English majors in China, which requires a candidate to master about 13,000 words.

Table 4 Participants' profile

Educational background	Number	Age	English language proficiency	Post-editing experience	Professional translation experience
First-year MTI	68	22–26 years old	TEM8 or the equivalent	No	Minimal

translated from scratch by G4. Their keystrokes were also logged by Translog-II. For the post-editing tasks, there were six students who only finished one text. Furthermore, two participants' logging data had to be discarded due to problems when saving the final logging data. Those participants' data were excluded. In total, 90 tasks remained for data analysis, including 30 HT tasks, 30 NMT post-editing tasks, and 30 PBSMT post-editing tasks (Table 5).

5 Results and discussion

5.1 Statistical analysis

The 90 translations were first manually aligned using the YAWAT tool (Germann 2008) and then processed into a set of CRITT TPR-DB tables (Carl et al. 2016). The analysis was carried out at segment (SG) level, by concatenating all 90 SG-tables. Each line in an SG table encodes approximately 55 different features which summarize properties of the SL and TL segment and the translation production process. The data were then analysed in the R statistical environment (R Core Team 2014) by fitting linear mixed-effects models with the lme4 package (Bates et al. 2014). This statistical method was chosen over traditional factorial designs because it includes both fixed and random effects in the linear mixed effects models (LMER), which compensate for the weak control of variables in naturalistic translation tasks (Balling 2008).

In our study, the random effect was always the participant whose individual difference may influence the data and the source text segment whose inherent difference may influence the data. As the main objective of this study was to investigate the difference in temporal, technical and cognitive effort during HT, post-editing PBSMT and post-editing NMT for the two texts (Text 1 and Text 2), the fixed effects were always task (HT, post-editing PBSMT and post-editing NMT) and text (Text 1 and Text 2) with interaction. The dependent variables were temporal, technical and cognitive effort. We checked whether the impact of task on the dependent variables

Table 5 Data used for final data analysis

Task type	Text 1	Text 2	Total
NMT post-editing	G1 (15)	G2 (15)	30
PBSMT post-editing	G2 (15)	G1 (15)	30
HT	G3 (15)	G4 (15)	30

varied between the two texts. The detailed results and discussions are introduced in the following section.

5.2 Temporal effort

We measure temporal effort by the average token processing time in milliseconds (ms), by dividing the total processing duration of a segment by the number of source text tokens in a segment (DurTokS). The interaction effect of task and text on DurTokS is plotted in Fig. 3. The results show that the participants spent significantly less time post-editing both NMT and PBSMT than translation from scratch for both texts. We also found that post-editing NMT reduced the average processing time per token by about 30% ($p = 0.058$) for Text 1 and almost 50% for Text 2 ($p < 0.05$) as compared to post-editing PBSMT. We assume that this is due to the better translation quality of the NMT output, as compared with PBSMT output. Interestingly, post-editing (NMT and PBSMT) of the text with higher complexity (Text 2) was quicker than Text 1 with lower complexity, which is also consistent with the better MT translation quality of Text 2. However, we found that in HT participants spent more time on the more complex text (Text 2) than the less complex one (Text 1). This supports Jensen's (2009, p. 62) claims that "(the source text complexity indicated by) a set of objective indicators can to some extent account for the degree of difficulty experienced by translators when translating a text".

Our results are in agreement with Shterionov et al. (2018) and Jia et al. (2019). Shterionov et al. report that post-editing both PBSMT and NMT is significantly faster than HT. Jia et al. show that post-editing NMT from English to Chinese is faster than HT for both general and domain-specific texts, although the result is only significant for domain-specific texts. However, our finding contradicts Shterionov's

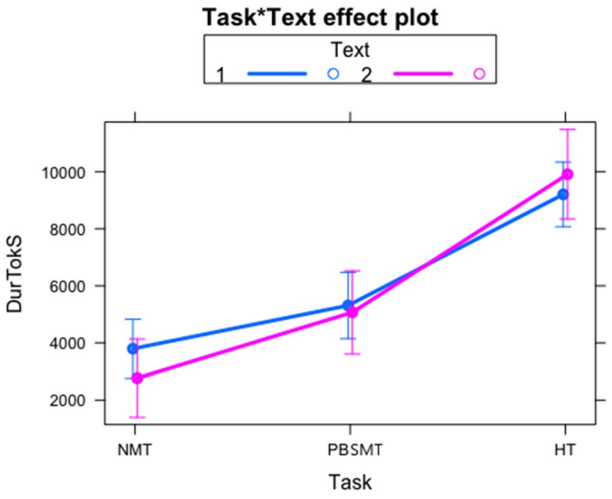

Fig. 3 Effect plot of interaction effect between task (HT, post-editing NMT and post-editing PBSMT) and text (Text 1 and Text 2) on average duration per token in ms (DurTokS)

et al. finding that post-editing NMT is faster than PBSMT for the English-Chinese language pair. Although Shterionov et al. report that post-editing NMT from English to German, Spanish, Japanese and Italian is faster for most of their participants than post-editing PBSMT, it is slower than post-editing PBSMT for most participants when it comes to English–Chinese language pair. Castilho et al. (2018) also report that post-editing NMT is not faster for all language pairs investigated in their study (English to German, Greek, Portuguese, and Russian). However, as there were only three translators/post-editors in these two studies, the results may be due to participants' individual differences. Previous studies have shown mixed results with respect to processing time of post-editing SMT output as compared with HT. While in contrast with Carl et al. (2011), Screen (2017) and da Silva et al. (2017), our finding is consistent with O'Brien (2007), Guerberof (2009), Plitt and Masselot (2010) and Daems et al. (2017) in that post-editing SMT is significantly faster than HT. Our results show that post-editing NMT output is more efficient than post-editing SMT output.

5.3 Technical effort

We measure technical effort by the number of keystrokes, which can be separated into insertions and deletions. We first check the total number of insertions and deletions. Then, we check the numbers of insertions and deletions separately.

5.3.1 Total keystrokes

The total number of insertions and deletions within each segment was divided by the number of tokens in each segment (Insdeltoks). Figure 4 shows the interaction effect of task and text on average keystrokes per token. The results reveal that post-editing NMT and PBSMT output produced significantly fewer keystrokes than HT for both texts, and that post-editing NMT required the fewest keystrokes. Post-editing NMT generated about one keystroke per token fewer for Text 1 ($p < 0.01$) and around two fewer than post-editing PBSMT for Text 2 ($p < 0.001$). Similarly to the results for temporal effort, post-editing NMT and PBSMT both generated fewer keystrokes per token for the more complex text (Text 2) with better MT quality, while the participants produced more keystrokes per token when translating the text of higher complexity (Text 2) from-scratch as compared to the less complex one (Text 1). Again, the more difficult text requires more keystrokes only for HT but not for post-editing.

5.3.2 Deletions

The interaction effect of task and text on average deletions per token (DelTokS) is presented in Fig. 5. For both texts, working with PBSMT generated the highest number of deletions among the three kinds of tasks, while post-editing NMT generated more deletions in Text 1 but fewer deletions in Text 2 than HT. The difference in average deletions per token between post-editing PBSMT and HT was significant

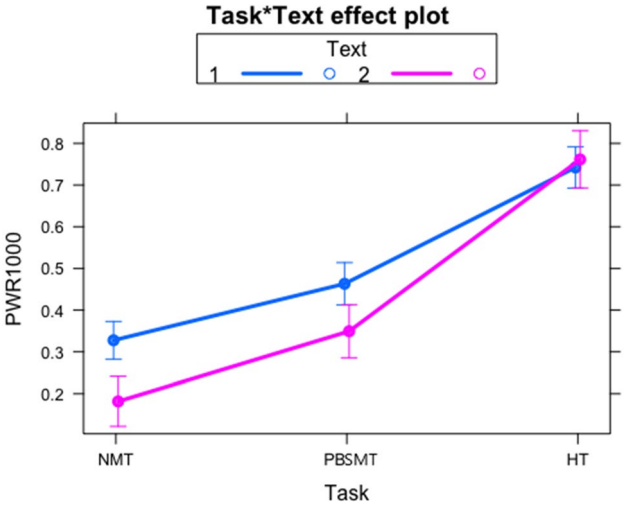

Fig. 4 Effect plot of interaction effect between task (HT, post-editing NMT and post-editing PBSMT) and text (Text 1 and Text 2) on average keystrokes per token (Insdeltokes)

for both texts ($p < 0.05$ for Text 1 and $p < 0.001$ for Text 2). Compared with HT, post-editing NMT produced more deletions in Text 1 ($p = 0.06$), but fewer deletions in Text 2 ($p = 0.3$), but neither difference was significant. More translations of the PBSMT output were deleted than those of the NMT output, which is likely to be due to the lower quality of the PBSMT output.

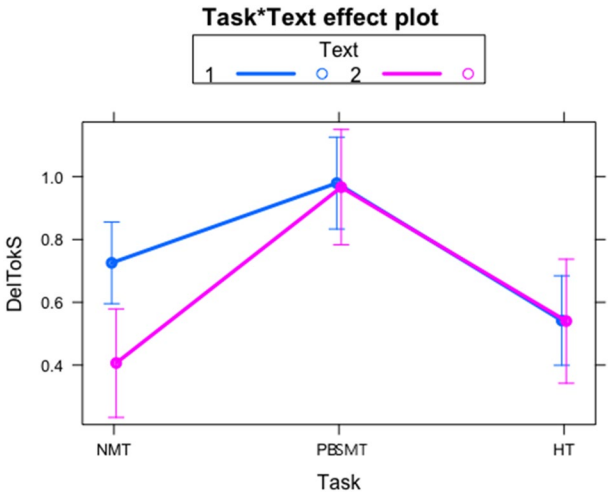

Fig. 5 Effect plot of interaction effect between task (HT, post-editing NMT and post-editing PBSMT) and text (Text 1 and Text 2) on average deletions per token (DelTokS)

5.3.3 Insertions

Figure 6 presents the interaction effect plot of task and text on average insertions per token. The general analysis shows that for both texts post-editing NMT and PBSMT needed significantly fewer insertions per token than HT, and post-editing NMT required fewer insertions per token than PBSMT. For Text 1, post-editing NMT needed about four insertions per token fewer than HT ($p < 0.001$) and almost one insertion per token fewer than post-editing PBSMT ($p < 0.01$). The situation was similar in Text 2, where post-editing NMT reduced the number of insertions per token by about five as compared with HT ($p < 0.001$) and by about one compared with post-editing PBSMT ($p < 0.001$).

We find that post-editing significantly reduced the participants' technical effort measured by typing activity. Our finding that participants made significantly more deletions and fewer insertions when post-editing PBSMT than HT is in line with Carl et al. (2011), but only partly consistent with Koglin (2015), who report that post-editing Google PBSMT requires both fewer insertions and deletions than HT. Our results corroborate those of Castilho et al. (2018), finding that post-editing NMT requires less technical effort than post-editing PBSMT. In our study, we show in addition that the number of deletions and insertions correlate with the quality of the NMT and the PBSMT output. Better MT quality leads to fewer typing activities.

5.4 Cognitive effort

Butterworth (1980) and Schilperoord (1996) argue that the number and the duration of pauses during language production are linked to cognitive effort. We use two pause metrics to measure cognitive effort during the post-editing and translating

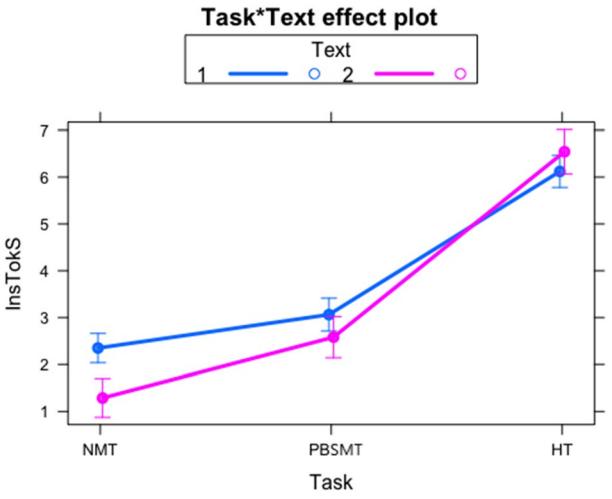

Fig. 6 Effect plot of interaction effect between task (HT, post-editing NMT and post-editing PBSMT) and text (Text 1 and Text 2) on average insertions per token (InsTokS)

process, based on Butterworth's (1980) and Schilperoord's (1996) argument that both the number and duration of pauses generated in language production can be linked to processing effort. The *pause duration per token* is based on the idea that longer pauses indicate higher cognitive effort. The *pause to word ratio* (PWR, Lacruz and Shreve 2014), referred to hereafter as *pause density* is based on the idea that more pauses indicate more cognitive effort. As these two metrics have not been used together at the same time in the same study, we also want to check whether the results of them point in the same direction. We adopted a pause threshold of 1000 ms which has been frequently used in previous studies (Jakobsen 1998; Krings 2001; O'Brien 2006; Lacruz et al. 2012).

5.4.1 Cognitive effort: pause duration

The *pause duration per token* (Pause1000) was computed by dividing the total pause time per segment by the number of tokens in the source text segment. We used Pause1000 as dependent variable and task and text with interaction as explanatory factors. The interaction effect of task and text is illustrated in Fig. 7.

The results are similar to those presented in Sect. 5.1 and 5.2. They show that participants paused significantly longer when translating from scratch than when post-editing both NMT and PBSMT for both texts. Post-editing NMT shortened the pause duration per token than by 35% for Text 1 ($p < 0.01$) and almost 50% for Text 2 ($p < 0.01$). Our findings on pause time support Koglin (2015) in that post-editing Google PBSMT triggers a shorter pause time than HT. The findings are also consistent with Jia et al. (2019) in that post-editing Google NMT from English to Chinses costs significantly shorter pause time than HT. However, the results contradict

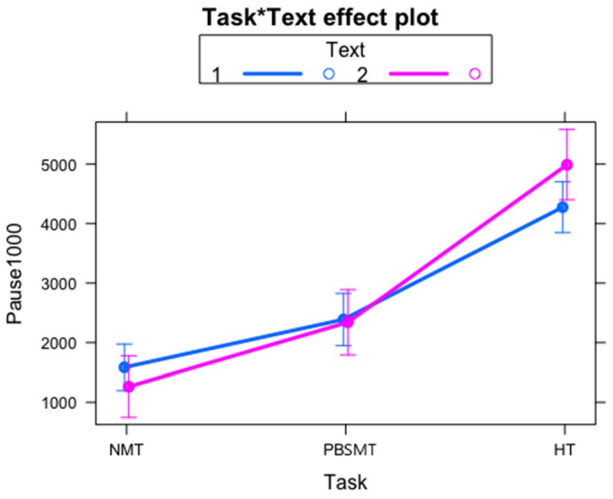

Fig. 7 Effect plot of interaction effect between task (HT, post-editing NMT and post-editing PBSMT) and text (Text 1 and Text 2) on pause duration per token (Pause1000)

Screen (2017) who suggests that post-editing Google PBSMT involves longer pause time than HT.

During HT, participants paused longer per token for the more complex text (Text 2) than the less complex one (Text 1). However, shorter pause times were produced when post-editing NMT output for Text 2 (the more complex one) than for Text 1 (the less complex one), while the pause time per token for post-editing PBSMT of Text 1 and 2 was more or less the same. This observation confirms our previous findings on temporal and technical effort that text of higher complexity seems to be cognitively more effortful to translate only for HT but not necessarily for post-editing.

5.4.2 Cognitive effort: pause density

Pause density (PWR) was calculated by dividing the total number of production pauses that occurred during the translation of a segment by the total number of tokens in the ST segment (Lacruz and Shreve 2014). The pause density is expected to be higher when translators exert more effort. We define PWR1000 as pause density where each pause lasts 1000 ms or longer. We used PWR1000 as dependent variable and task and text with interaction as explanatory factors. The effect is presented in Fig. 8. Post-editing both PBSMT and NMT significantly lowered the number of pauses per token for both texts, as compared with HT. The pause density for post-editing NMT (PWR1000 = 0.33) was significantly lower than for post-editing PBSMT (PWR1000 = 0.46) ($p < 0.001$) for Text 1. This effect is quite consistent in Text 2, where post-editing NMT (PWR1000 = 0.18) caused significantly fewer pauses than post-editing PBSMT (PWR1000 = 0.35) ($p < 0.001$). The results also show a similar picture as before. Both PBSMT and NMT post-editing triggered significantly lower pause density for the more complex text (Text 2) than the less

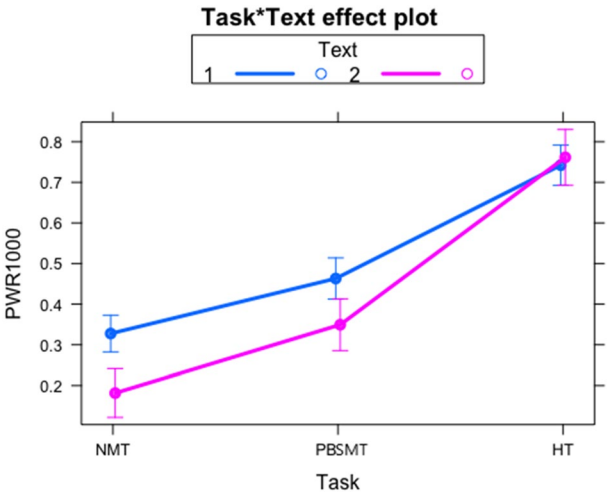

Fig. 8 Effect plot of interaction effect between task (HT, post-editing NMT and post-editing PBSMT) and text (Text 1 and Text 2) on pause density (PWR1000)

complex one (Text 1). Translating the more complex text (Text 2) (PWR1000 = 0.76) from scratch caused slightly higher pause density than the less complex one (Text 1) (PWR = 0.74).

The results on pause density are in line with Schaeffer et al. (2016), who also report PWR scores that are significantly lower when post-editing SMT than HT based on the multilingual data in CRITT TPR-DB. Our results also support Jia's et al. finding that post-editing Google NMT from English to Chinese causes lower pause density than HT. As post-editing NMT output required both shorter pause duration per token and lower pause density than post-editing PBSMT, it is reasonable to suggest that NMT is a more promising approach for post-editing news texts from English to Chinese than PBSMT and HT. Although post-editing NMT was significantly faster than post-editing PBSMT for only one text, it has significantly reduced the participants' cognitive effort for both texts. This finding seems to validate the argument of Krings (2001) and O'Brien (2011) that the productivity of post-editing should not only be determined by the processing time involved but also the cognitive effort. Our results also confirm that temporal effort, technical effort and cognitive effort each seem to have their own role in explaining the translation and post-editing process. What's more, the results of the two pause metrics, one based on pause length and the other one based on number of pauses, allow for the same conclusion that HT is cognitively more effortful than both PBSMT and NMT post-editing, and PBSMT post-editing is cognitively more effortful than NMT post-editing.

5.5 Correlation between the measures of post-editing effort

In this section, we compute the Pearson's correlations in R for the 18 segments, between the measures of temporal (DurTokS), technical (DelTokS, InsTokS, and InsDelToks) and cognitive effort (PWR1000 and Pause1000). A correlation matrix is shown in Table 6. Results for NMT are in the upper part (red) and the ones for PBSMT are in the lower part (blue). We see that all the behavioural measures have a modest-to-strong positive correlation for both NMT and PBSMT post-editing. The correlations between the measures for NMT are generally higher than the ones for PBSMT. Higher post-editing time per token (DurTokS) is correlated with more

Table 6 Correlation matrix between the behavioural measurements of temporal, technical and cognitive effort for NMT (in bold) and for PBSMT (in italics)

PBSMT	NMT					
	DurTokS	DelTokS	InsTokS	InsDelToks	PWR1000	Pause1000
DurTokS	–	**0.73**	**0.61**	**0.68**	**0.68**	**0.66**
DelTokS	*0.63*	–	**0.70**	**0.83**	**0.70**	**0.59**
InsTokS	*0.48*	*0.72*	–	**0.98**	**0.83**	**0.58**
InsDelToks	*0.56*	*0.86*	*0.98*	–	**0.85**	**0.62**
PWR1000	*0.62*	*0.67*	*0.69*	*0.73*	–	**0.79**
Pause1000	*0.63*	*0.41*	*0.44*	*0.46*	*0.70*	–

deletions and insertions per token and longer and more pauses per token. The two pause metrics for cognitive effort also have a very strong correlation both for both NMT (r = 0.79) and PBSMT (r = 0.7).

6 General discussion and conclusion

This paper investigates whether NMT facilitates post-editing. To this end, we have compared the fluency and accuracy of Google's PBSMT and NMT output and compared post-editing effort with human (from scratch) translation for the English-Chinese translation of two news texts. Our main findings are: (1) NMT produced better translations in terms of fluency and accuracy than PBSMT, (2) post-editing NMT reduced the temporal, technical and cognitive effort, as compared with PBSMT and HT, (3) all measures of post-editing effort were positively correlated, and (4) the source text complexity measures tailored for HT only impact the effort expended during HT, but not the quality of MT output and post-editing effort.

However, there are some limitations in this study should be addressed in future research. Our findings are based on post-editing news texts of the NMT and PBSMT output generated by Google-Translate by the student translators from English to Chinese. Therefore, future studies could include additional text domains, MT engines and language pairs to allow for more generalizable results. What's more, as previous studies show that the behaviors of student translators during post-editing process can be different from professionals' (e.g. Moorkens and O'Brien 2015; Yamada 2015), participants with varying levels of expertise are ideally to be included in the future research.

In the next step of our study, we intend to check the specific features of the NMT output that help alleviate the post-editors' cognitive effort as compared with the PBSMT output. Besides, we will also further investigate the source text features that impact MT quality and post-editing effort. The qualitative questionnaires, which could provide us with more understanding of how post-editors work with NMT and PBSMT, have not been analysed due to the scope of this study. Therefore, the questionnaires will be analysed in detail to check the translators' attitude towards NMT post-editing and NMT quality. Furthermore, the perceived post-editing effort indicated by the questionnaires will be analysed to see its correlation with the objective measurements of the effort.

Acknowledgements Thanks goes to our participants, annotators and raters for their precious time. Particular gratitude is extended to Prof. Yves Gambier for his comments on the earlier drafts of this article. We are also grateful to Dr. Joss Moorkens and the anonymous reviewers for their constructive comments. This research was supported by the Social Science Foundation of Hunan Province (17ZDB005), China Hunan Provincial Science & Technology Foundation ([2017]131).

References

Alves F (2006) A relevance-theoretic approach to effort and effect in translation: discussing the cognitive interface between inferential processing, problem-solving and decision-making. In: Proceedings of

the international symposium on new horizons in theoretical translation studies, Hong Kong, China, pp 1–12
Bahdanau D, Cho K, Bengio Y (2015) Neural machine translation by jointly learning to align and translate. In: Proceedings of the 6th international conference on learning representations, San Diego, California, USA
Balling LW (2008) A brief introduction to regression designs and mixed-effects modelling by a recent convert. In: Göpferich S, Jakobsen AL, Mees IM (eds) Looking at eyes: eye tracking studies of reading and translation processing. Copenhagen studies in language, vol 36, pp 175–192
Balling L, Carl M (2014) Production time across language and tasks: a large-scale analysis using the CRITT translation process database. In: Schwieter J, Ferreira A (eds) The development of translation competence: theories and methodologies from psycholinguistics and cognitive science. Cambridge Scholar Publishing, Cambridge, pp 239–268
Bates D, Maechler M, Bolker B, Walker S (2014) lme4: linear mixed-effects models using Eigen and S4. R package version 3.1.2. http://CRAN.R-project.org/package=lme4
Bentivogli L, Bisazza A, Cettolo M, Federico M (2016) Neural versus phrase-based machine translation quality: a case study. In: Proceedings of the 2016 conference on empirical methods in natural language processing, Austin, Texas, pp 257–267
Bojar O, Chatterjee R, Federmann C et al (2016) Findings of the 2016 conference on machine translation. In: Proceedings of the first conference on machine translation (WMT 2016), Berlin, Germany, pp 131–198
Butterworth B (1980) Evidence from pauses in speech. In: Butterworth B (ed) Language production, vol 1. Speech and talk. Academic Press, London, pp 155–176
Carl M (2012) Translog-II: a program for recording user activity data for empirical reading and writing research. In: Proceedings of the eighth international conference on language resources and evaluation (LREC12), pp 4108–4112
Carl M, Dragsted B, Elming J, Hardt D, Jakobsen AL (2011) The process of post-editing: a pilot study. In: Sharp B, Zock M, Carl M, Jakobsen AL (eds) Proceedings of the 8th international NLPCS Workshop, Copenhagen, Denmark, pp 131–142
Carl M, Gutermuth S, Hansen-Schirra S (2015) Post-editing machine translation: a usability test for professional translation settings. In: Schwieter J, Ferreira A (eds) Psycholinguistic and cognitive inquiries in translation and interpretation studies. John Benjamins, Amsterdam, pp 145–174
Carl M, Schaeffer M, Bangalore S (2016) The CRITT Translation process research database. In: Carl M, Bangalore S, Schaeffer M (eds) New directions in empirical translation process research: exploring the CRITT TPR-DB. Springer, Cham, pp 13–54
Castilho S, Moorkens J, Gaspari F, Sennrich R, Way A, Georgakopoulou P (2018) Evaluating MT for massive open online courses. Mach Transl 32(3):255–278
da Silva IAL, Alves F, Schmaltz M et al (2017) Translation, post-editing and directionality: a study of effort in the Chinese-Portuguese language pair. In: Jakobsen AL, Mesa-Lao B (eds) Translation in transition. John Benjamins, Amsterdam, pp 91–117
Daems J, Vandepitte S, Hartsuiker RJ, Macken L (2017) Translation methods and experience: a comparative analysis of human translation and post-editing with students and professional translators. Meta 62(2):245–270
Fleiss JL (1971) Measuring nominal scale agreement among many raters. Psychol Bull 76(5):378–382
Germann U (2008) Yawat: yet another word alignment tool. In: Proceedings of the 46th annual meeting of the association for computational linguistics: demo session, Columbus. Columbus, OH, pp 20–23
Groves D, Schmidtke D (2009) Identification and analysis of post-editing patterns for MT. In: Proceedings of the twelfth machine translation Summit, August 26–30, Ottawa, ON, Canada, pp 429–436
Guerberof A (2009) Productivity and quality in MT post-editing. In: Proceedings of MT Summit XII—workshop: beyond translation memories: new tools for translators MT, Ottawa, Ontario, Canada
Hansen G (2002) Zeit und Qualität im Übersetzungsprozess. In: Hansen G (ed) Empirical translation studies: process and product. Samfundslitteratur, Copenhagen, pp 29–54
Jakobsen AL (1998) Logging time delay in translation, LSP texts and the translation process. In: Copenhagen working papers, pp 73–101
Jakobsen AL (2002) Translation drafting by professional translators and by translation students. In: Empirical translation studies: process and product. Copenhagen studies in language vol 27, pp 191–204

Jensen KTH (2009) Indicators of text complexity. In: Göpferich S, Jakobsen AL, Mees IM (eds) Behind the mind: methods, models and results in translation process research. Copenhagen studies in language, vol 36. Samfundslitteratur, Copenhagen, pp 61–80
Jensen KTH (2011) Allocation of cognitive resources in translation—an eye-tracking and key-logging study. Dissertation, Copenhagen Business School
Jia Y, Carl M, Wang X (2019) How does the post-editing of neural machine translation compare with from-scratch translation? A product and process study. J Specialised Transl 3:60–86
Junczys-Dowmunt M, Dwojak T, Hoang H (2016) Is neural machine translation ready for deployment? A case study on 30 translation directions. In: Proceedings of the 9th international workshop on spoken language translation, Seattle, Washington
Klubička F, Toral A, Sánchez-Cartagena VM (2017) Fine-grained human evaluation of neural versus phrase-based machine translation. Prague Bull Math Linguist 108:121–132
Koglin A (2015) An empirical investigation of cognitive effort required to post-edit machine translated metaphors compared to the translation of metaphors. Transl Interpret 7(1):126–141
Krings HP (2001) Repairing texts: empirical investigations of machine translation post-editing processes. The Kent State University Press, Kent, OH
Lacruz I, Shreve GM (2014) Pauses and cognitive effort in post-editing. In: O'Brien S, Winther Balling L, Carl M, Simard M, Specia L (eds) Post-editing of machine translation: processes and applications. Cambridge Scholars Publishing, Newcastle-upon-Tyne, pp 246–272
Lacruz I, Shreve GM, Angelone E (2012) Average pause ratio as an indicator of cognitive effort in post-editing: a case study. In: Proceedings of the AMTA 2012 workshop on post-editing technology and practice (WPTP 2012), San Diego, CA, pp 21–30
Lacruz I, Denkowski M, Lavie A (2014) Cognitive demand and cognitive effort in post-editing. In: Proceedings of the third workshop on post-editing technology and practice, Vancouver, Canada
Mesa-Lao B (2014) Gaze behaviour on source texts: an exploratory study comparing translation and post-editing. In: O'Brien S, Winther Balling L, Carl M, Simard M, Specia L (eds) Post-editing of machine. Cambridge Scholars Publishing, Newcastle, pp 219–245
Mishra A, Bhattacharyya P, Carl M (2013) Automatically predicting sentence translation difficulty. Proceedings from the 51st annual meeting of the association for computational linguistics (vol 2: short papers). Sofia, Bulgaria, pp 346–351
Moorkens J, O'Brien S (2015) Post-editing evaluations: trade-offs between novice and professional participants. In: Proceedings of the 18th annual conference of the European Association for Machine Translation (EAMT 2015), Antalya, Turkey, pp 75–81
O'Brien S (2006) Pauses as indicators of cognitive effort in post-editing machine translation. Across Lang Cult 7(1):1–21
O'Brien S (2011) Towards predicting post-editing productivity. Mach Transl 25(3):197–215
O'Brien S (2007) An empirical investigation of temporal and technical post-editing effort. Transl Interpret Stud 2(1):83–136
Plitt M, Masselot F (2010) A productivity test of statistical machine translation post-editing in a typical localisation context. Prague Bull Math Linguist 93:7–16
R Core Team (2014) R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org
Schaeffer M, Carl M, Lacruz I, Aizawa A (2016) Measuring cognitive translation effort with activity units. Baltic J Mod Comput 4(2):331–345
Schilperoord J (1996) It's about time. Temporal aspects of cognitive processes in text production. Rodopi, Amsterdam
Screen B (2017) Machine translation and Welsh: analysing free statistical machine translation for the professional translation of an under-researched language pair. J Spec Transl 28:317–344
Shterionov D, Superbo R, Nagle P, Casanellas L, O'Dowd T (2018) Human versus automatic quality evaluation of NMT and PBSMT. Mach Transl 32(3):217–235
Snover M, Dorr B, Schwartz R, Miccuilla L, Makhoul J (2006) A study of translation edit rate with targeted human annotations. In: Proceedings of association for machine translation in the Americas, Cambridge, MA, USA, pp 223–231
Suo J, Yu B, He Y, Zang G (2012) Study of ambiguities of English–Chinese machine translation. Appl Mech Mater 157:472–475
Tatsumi M (2009) Correlation between automatic evaluation scores, post-editing speed and some other factors. In: Proceedings of MT Summit XII, Ottawa, Canada, pp 332–339

TAUS (2016) TAUS post-editing guidelines. https://www.taus.net/think-tank/articles/postedit-articles/taus-post-editing-guidelines. Accessed 2 May 2016

Toral A, Sánchez-Cartagena VM (2017) A multifaceted evaluation of neural versus phrase-based machine translation for 9 language directions. arXiv:1701.02901

Wu X, Cardey S, Greenfield P (2006) Realization of the Chinese BA-construction in an English-Chinese machine translation system. In: Proceedings of the fifth SIGHAN workshop on Chinese language processing, pp 79–86

Wu Y, Schuster M, Chen Z et al (2016) Google's neural machine translation system: bridging the gap between human and machine translation. CoRR. arXiv:1609.08144

Yamada M (2015) Can college students be post-editors? An investigation into employing language learners in machine translation plus post-editing settings. Mach Transl 29(1):49–67