

How does the post-editing of neural machine translation compare with from-scratch translation? A product and process study

Yanfang Jia, Hunan University

Michael Carl, Kent State University

Xiangling Wang, Hunan University

ABSTRACT

This study explores the post-editing process when working within the newly introduced neural machine translation (NMT) paradigm. To this end, an experiment was carried out to examine the differences between post-editing Google neural machine translation (GNMT) and from-scratch translation of English domain-specific and general language texts to Chinese. We analysed translation process and translation product data from 30 first-year postgraduate translation students. The analysis is based on keystroke logging, screen recording, questionnaires, retrospective protocols and target-text quality evaluations. The three main findings were: 1) post-editing GNMT was only significantly faster than from-scratch translation for domain-specific texts, but it significantly reduced the participants' cognitive effort for both text types; 2) post-editing GNMT generated translations of equivalent fluency and accuracy as those generated by from-scratch translations; 3) the student translators generally showed a positive attitude towards post-editing, but they also pointed to various challenges in the post-editing process. These were mainly due to the influence of their previous translation training, lack of experience in post-editing and the ambiguous wording of the post-editing guidelines.

KEYWORDS

Post-editing process, translation quality, neural machine translation, text types.

1. Introduction

In the last decade, machine translation (MT) has been increasingly adopted by the translation industry as an effective solution to the globally ever-increasing demands for translation that from-scratch human translation cannot satisfy. Unfortunately, raw MT output cannot always meet the end user's expectations in terms of translation quality, thus making MT plus post-editing a necessary and standard practice. Compared to statistical machine translation (SMT), the recently developed neural machine translation (NMT) paradigm is found to have greatly advanced the state of the art, by improving translation quality, as measured mainly by automatic evaluation metrics (Bahdanau *et al.* 2014; Sennrich *et al.* 2016; Bojar *et al.* 2016; Junczys-Dowmunt *et al.* 2016), although the human evaluation results can sometimes be mixed (Castilho *et al.* 2017; Popović 2017; Klubička *et al.* 2017). It is reasonable to expect post-editing of NMT to be a more promising approach to adopt than post-editing of SMT, although the post-editing process of NMT has scarcely been investigated.

In addition, despite the increasing demand in the translation market for post-editing services and post-editors (Lommel and DePalma 2016), professional translators are found to be reluctant to take post-editing jobs due to their negative perceptions of MT quality and post-editing work, while

student translators seem to show greater potential to become future post-editors to fill this gap (Garcia 2010; Yamada 2015). In China, according to a report of the Translation Association of China (TAC) in 2016, there were more than 72,495 companies providing language-related services at the end of 2015, among which around 7,400 companies specialised in language and translation services, not including the undocumented small businesses and freelance translators and interpreters. To meet the increasing need for translators in the language service industry, the National Degree Committee under the State Council of China launched its “Master in Translation and Interpreting” (MTI) programme in 2007, aiming to train professionally competent translators and interpreters as demanded by the market. By 2016, 206 colleges and universities had been authorised to enrol MTI students.

This progress highlights the effort the translation education systems in China had made in response to market needs. However, problems arose with the expansion of recruitment, and among these the lack of qualified teachers was one of the most urgent. As most of the existing teachers are not translation professionals, they are not capable of offering professional translation training to the MTI students. The integration of the latest technology, such as computer-aided translation tools and machine translation, makes teaching even more challenging. The TAC 2016 report identified that courses related to translation technologies were greatly needed in the MTI programme.

To date, no systematic post-editing training courses have been incorporated into the MTI programme. A more profound insight into the nuances and complexities of this relatively new task for the English-Chinese language pair, especially with respect to how it differs from traditional from-scratch human translation, could, therefore, facilitate the translator training process. It would then be possible to adjust the current curriculum and better prepare students for the job market.

The present study investigates the differences between the post-editing of GNMT and human from-scratch translation in terms of differences in the translation process and in the translation product. The study examines how MTI students carry out post-editing tasks involving different text types. It seeks to address the following three questions: 1. What are the differences in the translation process and product quality between post-editing of NMT and from-scratch translation? 2. What is the impact of text types on these differences? 3. What are the MTI students’ opinions concerning the differences between post-editing of NMT and from-scratch translation?

2. Related research

Post-editing is “the task of editing, modifying and/or correcting pre-translated text that has been processed by an MT system from a source

language into a target language” (Allen 2003: 297), and the post-edited text should meet “the end user’s expected quality levels” (TAUS 2013a).

Over the last decade, many studies have explored the differences between post-editing and from-scratch translation from various perspectives. Processing speed is one of the most frequently investigated factors in these comparisons, and this is also an issue of primary concern for the industry. Post-editing domain-specific texts is constantly found to be faster than from-scratch translation (O’Brien 2007; Groves and Schmidtke 2009; Tatsumi 2009; Plitt and Masselot 2010). For general language texts, however, post-editing is not always faster. Daems *et al.* (2017) found the post-editing of news texts to be significantly faster, while other studies reported no significant increase in speed when post-editing news texts (e.g. Carl *et al.* 2011) and general information texts (e.g. Screen 2017). Temporal aspects are important but do not provide information on “how post-editing occurs as a process, how it is distinguished from conventional translation, what demands it makes on post-editors, and what kind of acceptance it receives from them” (Klings 2001: 61). Therefore, Klings (2001) argues that the feasibility of post-editing compared to human translating should not be determined by processing time alone. O’Brien (2011: 198) also claims that post-editing productivity means “not only the ratio of quantity and quality to time but also the cognitive effort expended; and the higher the effort, the lower the productivity”.

Research into the cognitive aspects of post-editing is, therefore, necessary for a better understanding of the post-editing process and how it compares to from-scratch translation. Klings (2001: 179) defined cognitive effort as the “type and extent of those cognitive processes that must be activated to remedy a given deficiency in a machine translation”. Klings employed think-aloud protocols (TAPs) and claimed that post-editing entailed more verbalization effort than from-scratch translation. The introduction of eye tracking and keylogging into translation process research has greatly extended our ability to understand the translators’ reading and writing process during translation at any given point in time. Da Silva *et al.* (2017) found no significant difference in processing time between post-editing and from-scratch translation, but they noticed a significant difference in cognitive effort based on eye-tracking metrics. Records of gaze data reveal that the reading time of, and thus presumably also the allocation of cognitive resources to, the source text and target text is very different in post-editing compared to from-scratch translation (Mesa-Lao 2014; Carl *et al.* 2015; Daems *et al.* 2017; da Silva *et al.* 2017). These studies indicate that fixations during post-editing seem to focus more on the target text, and those during from-scratch translation tend to be centred more on the source text (see e.g. Carl *et al.* 2015: 165).

Beside gaze data, pauses between keystrokes during typing are generally agreed to be an effective indicator of cognitive effort in the translation process (Jakobsen 1998, 2002; Hansen 2002; Alves 2006). Longer pause

duration and larger pause density both signal higher cognitive effort. However, the results from comparisons between post-editing and from-scratch translation based on pause metrics seem to be far from conclusive. Koglin (2015) found that post-editing news texts from English into Brazilian Portuguese triggered shorter total pause duration than from-scratch translation, which contradicts Screen (2017) who reported post-editing general language texts from English into Welsh contained longer total pause duration. Based on the observation that the density of short pauses during post-editing is a good indicator of cognitive effort, Lacruz and Shreve (2014) introduced the pause to word ratio (number of pauses per word) (PWR) to measure cognitive effort. Based on the multilingual data in the Translation Process Research Database (TPR-DB) (Carl *et al.* 2016), Schaeffer *et al.* (2016) found that PWR correlated strongly with a gaze-based translation difficulty index (TDI), and that the values of PWR for SMT post-editing were significantly lower than for from-scratch translation. As these studies have mainly investigated the cognitive process of SMT post-editing as compared to from-scratch translation, more extensive research on the cognitive process of NMT post-editing, which remains rather unclear, is highly necessary.

In addition to the post-editing process, the quality of the post-edited product is also a matter of concern, as the time and cognitive effort saved in post-editing is only worthwhile if its final product is not compromised, as compared to from-scratch translation. Fiederer and O'Brien (2009) showed that English to German post-edited domain-specific texts were superior in accuracy and fluency when compared to those translated from scratch, although inferior in style. Based on the error types developed by the Localization Industry Standards Association (LISA), post-editing supply chain management content from English into Spanish resulted in fewer errors than when translating from scratch (Guerberof 2009). These results are in line with those of Garcia (2010), who reports that, according to the Australian NAATI test criteria, post-edited output was favoured by the evaluators. Carl *et al.* (2011) also found post-editing news texts from English into Danish led to a modest improvement in quality compared to from-scratch translation. Similar results were obtained by Green *et al.* (2013), who reported that post-editing Wikipedia articles improved product quality in comparison to texts translated from scratch from English into Arabic, French, and German. There seems to be a tendency for post-editing to deliver comparable or even better translation quality in comparison to from-scratch translation.

Finally, an aspect of equal consideration in the current study is how Chinese MTI students without post-editing experience perceive the differences between post-editing and from-scratch translation, including the strategies they adopt, the challenges they encounter and their attitude towards post-editing. Investigations into these aspects have far-reaching educational implications, as a better understanding and clear awareness of machine

translation and post-editing may prepare the student translators for the challenges of working with evolving technologies in the future.

The above-mentioned studies are inspiring but predominantly involve SMT post-editing with English and other alphabetic Indo-European languages, while NMT post-editing with logographic languages such as Chinese are barely investigated. This paper, therefore, endeavours to bridge this gap by analysing how English-Chinese NMT post-editing differs from from-scratch translation in terms of both process and product. An assessment of the impact of text types on these differences and the students' perceptions of the post-editing process is also undertaken.

3. Materials and methods

3.1. Participants' profile

The data collection was carried out in 2017 from February to June. Thirty first-year MTI students at a Chinese university participated in this study. They all specialised in translation and were all enrolled on an advanced translation course. There were four males and 26 females, aged 22 to 26 years. They all had the same language background with Chinese as L1 and English as L2 and also a very similar level of English language proficiency. Twenty-seven of them had passed the Test for English Majors at Band 8 (TEM8)¹, and three had passed at Band 4 (TEM4). They all had very limited professional translation experience. None of them had any professional experience or formal training in post-editing. In order to compare the participants' performance in post-editing of NMT and in from-scratch translation, the participants were divided into two groups (G1 and G2), based on their scores in their two most recent translation tests. This served to ensure that the students' level of translation ability was broadly the same between the two groups, which consisted of 15 students each.

3.2. Materials

In order to address the research questions, two English domain-specific texts and two English general language texts were selected, ranging from 142 to 156 words in length. The domain-specific texts were a patient information leaflet (ST1) and a dishwasher manual (ST2). The general language texts were two promotional brochures for two beverage brands, Coca-Cola (ST3) and Starbucks (ST4), respectively. All texts were self-contained and required no additional context to understand. There were some specialised words in the domain-specific texts, but the participants could use online dictionaries to obtain direct Chinese translations. We also checked that no Chinese translations of the English texts could be found on the Internet, as the participants were allowed to access the Internet to consult external resources during the test to replicate their customary everyday translation scenarios. The four English texts were pre-translated by Google's Neural Machine Translation system (GNMT) (May 2017).

Translation briefs were provided for each text, which instructed the participants about the target audience, the purpose, and the quality expectations of the target text. The four texts were all translated and post-edited for external dissemination. Post-editing guidelines for publishable quality developed by TAUS (2016) were provided for the post-editing tasks.

A pre-task questionnaire was used to collect information concerning the participants' educational and professional backgrounds as translators and post-editors, and to enquire specifically about their attitudes to MT and post-editing. On completion of the experiment, participants were asked to complete a post-task questionnaire, comprising 15 questions, to check their perception of post-editing speed, mental effort, translation quality, and their opinions on the provision for post-editing skills training in the curriculum.

3.3. Experimental procedures

The experiment was part of an advanced translation course for MTI students during the second semester of their first-year postgraduate studies in May 2017. The participants all signed an informed consent form approved by the Ethics Committee of the College of Foreign Languages at Hunan University.

Each of the four texts was translated from scratch by one group and post-edited by the other group. A combination of questionnaires, a keystroke logging tool (Translog-II), a screen recorder (BB FlashBack), and retrospective written reports was used for triangulation purposes. There were no time constraints and the participants all used their own laptops during the experiment. The aim was to have the translators work under conditions that were as close as possible to their natural working environment, with access to their usual browser preferences, dictionaries, and input methods. They also had access to the Internet during the experiment for information searching and external resources consultation. The experiment was carried out in two sessions, with a one-week interval between them.

The first session started with a general introduction to MT, post-editing and the TAUS post-editing guidelines. This was followed by the pre-task questionnaire. Then, each student translated one short text and post-edited one short text to become familiar with the functions of BB FlashBack and the Translog-II Interface. The actual experimental tasks in session one consisted of two domain-specific texts (ST1 and ST2). First, the corresponding translation briefs for each task and post-editing guidelines were provided to the students. Then, ST1 was translated from scratch by G1 and post-edited by G2, while ST2 was translated from scratch by G2 and post-edited by G1. Session two included two general language texts (ST3 and ST4). At the beginning of the second session, the students again received the translation briefs and post-editing guidelines for their session two tasks. Then, ST3 was translated from scratch by G1 and post-edited by

G2, while ST4 was translated from scratch by G2 and post-edited by G1. Table 1 shows the study's experimental set-up.

	Session1(Domain-specific texts)		Session2(General language texts)	
Text	ST1	ST2	ST3	ST4
From-scratch translation	G1	G2	G1	G2
Post-editing	G2	G1	G2	G1

Table1. The experimental set-up.

After the translation and post-editing tasks, the students viewed the recordings of the translation and post-editing process of the four tasks they had undertaken in the first and second sessions on BB FlashBack. Then, they were asked to provide a retrospective written report on the differences between from-scratch translation and post-editing, including the strategies they adopted, the challenges they came across and their attitudes towards post-editing. This was followed by the post-test questionnaire mentioned above.

3.4. Data exclusion

For each participant, logging data was collected for the two post-editing tasks and the two from-scratch translation tasks. Some sessions were discarded due to corrupted logging data. In addition, the recordings of BB FlashBack showed that, when asked to translate from-scratch, six students were found to have machine-translated the whole texts first and then copied the MT output to the Translog-II target window. These tasks were, therefore, excluded. In total, 99 tasks were left for data analysis, including 44 from-scratch translation tasks (22 from the general language texts and 22 from the domain-specific texts) and 55 post-editing tasks (27 from general language texts and 28 from the domain-specific texts). Table 2 shows the tasks used for final data analysis.

	Domain-specific texts	General language texts	Total
From-scratch translation	22	22	44
Post-editing	28	27	55
Total	50	49	99

Table 2. Tasks used for final data analysis.

3.5. Analysis

The final Translog-II xml files were first manually aligned using the YAWAT tool (Germann 2008), after which they were processed into a set of tables containing more than 200 features describing the process and product of the translation in detail (Carl *et al.* 2016). The data analysis was carried out

at the segment level using the concatenated SG-tables which contain information concerning translated segments. Then, the data were loaded into R, a statistical software package (R Core Team 2014). Linear mixed effects analyses were performed on our data, using the lme4 package (Bates *et al.* 2014). The main reason for choosing this statistical method over traditional factorial designs including both fixed and random effects in the linear mixed effects models (LMER) is that it compensates for the weak control of variables in naturalistic translation tasks (Balling 2008). We built five LMER models altogether.

For all five models, the random effects were always the participant and the source-text sentence, as differences associated with these factors may have a systematic impact on the data. The dependent variables of the five models were 1) processing time per word, 2) pause density, 3) pause duration per word, 4) average fluency score and 5) average accuracy score. For models 1) and 2), the fixed effects were task (from-scratch translation and post-editing) and text type (domain-specific texts and general language texts). For models 4) and 5), the fixed effects were output type (GNMT output, post-editing, and from-scratch translation) and text type. We first checked whether there was a significant main effect and then assessed the interaction of the fixed effects. The fixed factors each had at least two levels, but a significant main, or interaction effect of the LMER model, would not specify whether all or only some of the possible comparisons between factor levels were significant. Post-hoc follow-up comparisons were, therefore, employed by constructing additional LMER models, through redefining the reference level against which the other factor levels are compared, to check the relevant comparisons between factor levels in detail. The results of the five LMER models will be discussed in the following sections.

4. Data analysis and discussion

4.1. Process: processing time

The first dependent variable in our LMER is processing time per word (DurTokS), calculated by dividing the total processing time of each sentence by the total number of words in the source text sentence. The effect is plotted in Figure 1. Overall, post-editing took less time than from-scratch translation for both text types. However, this effect was significant only for domain-specific texts ($p < 0.05$), where the time needed per word was more than 3 seconds lower compared to from-scratch translation. Furthermore, the results reveal that translating domain-specific texts from scratch took slightly longer than translating general language texts from scratch, whereas post-editing domain-specific texts took much less time than post-editing general language texts ($p = 0.068$).

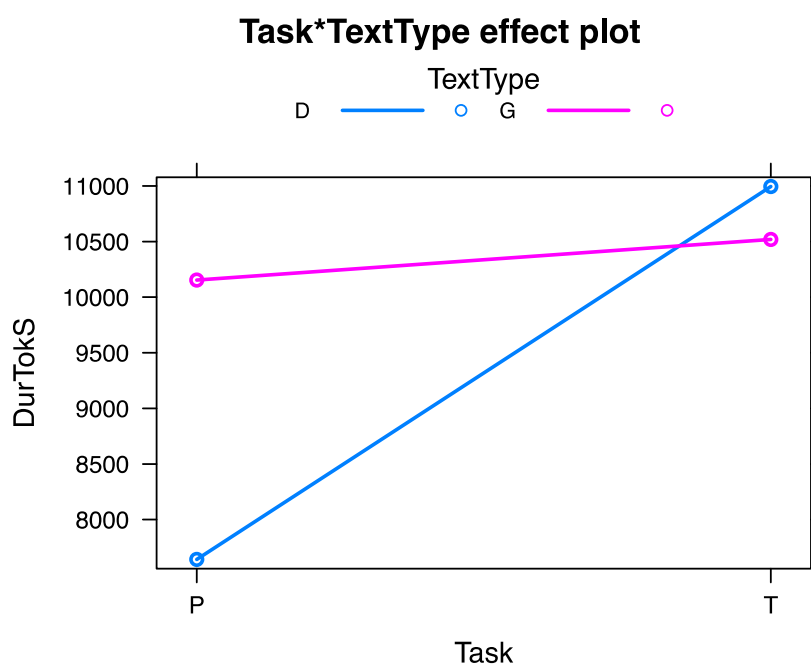


Figure 1. Effect plot of the interaction effect between task (post-editing and from-scratch translation, P and T, respectively) and text type (domain-specific texts and general language texts, D and G, respectively) on processing time per word (in ms) (DurTokS).

These results support the findings of some previous studies which found that post-editing is faster than from-scratch translation. However, the extent of the time-saving effect varies with different text types. This effect is more evident for domain-specific texts, which is in line with findings from previous research (O’Brien 2007; Guerberof 2009; Plitt and Masselot 2010) and with the translation industry’s use of MT post-editing for technical translation. For general language texts, in contrast with Daems *et al.* (2017), our results correspond to Carl *et al.* (2011), Screen (2017) and da Silva *et al.* (2017) showing that post-editing does not significantly reduce overall time. These comparisons should, however, be made with caution since the participants involved in these studies were more experienced translators than those recruited for our study, who had neither professional translation nor post-editing experience.

The difference in the time-saving effect resulting from text types may be explained by the linguistic differences of the two text types. Domain-specific texts contain relatively more terms than general language texts, which may require a considerable amount of time for a translator to retrieve from online resources during from-scratch translation. Domain-specific texts are, moreover, also syntactically more formulaic and less complex than general language texts. As inexperienced translators tend to treat translation more like a lexical task (Tirkkonen-Condit 1990), with the previously translated texts readily available, post-editing may be much quicker than having to translate all the lexical information from scratch. As the students in the current study had no former experience in post-editing, it is reasonable to

expect that they may become much quicker at post-editing both text types, after gaining more experience and receiving proper training.

4.2. Process: cognitive effort

The cognitive processing in from-scratch translation and post-editing was compared in terms of pause density (PWR) (Lacruz and Shreve 2014) and pause duration per word. Although the operationalization of pause thresholds tends to be arbitrary, 1000ms was adopted here to ensure comparability with some previous studies (Jakobsen 1998; Krings 2001; O'Brien 2006; Lacruz *et al.* 2012).

4.2.1. Cognitive effort: pause density

The second dependent variable in our LMER was pause density (PWR), measured by dividing the total number of pauses in a segment by the number of words in the source-text segment. The pause density is expected to be higher when translators exert more effort. This effect is demonstrated in Figure 2.

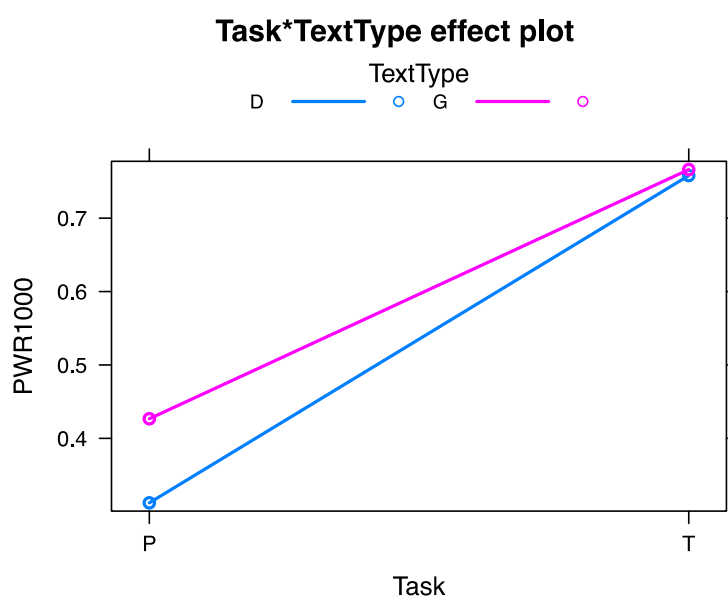


Figure 2. Effect plot of the interaction effect between task (post-editing and from-scratch translation, P and T, respectively) and text type (domain-specific texts and general language texts, D and G, respectively) on pause density (PWR1000).

The overall results show that both fixed effects (task and text type) had a significant effect on pause density. First, task was a significant predictor ($p < 0.001$), with post-editing involving significantly fewer pauses per word than from-scratch translation. For both text types, PWR was significantly lower during post-editing than during from-scratch translation. Second, text type had a significant effect on pause density ($p < 0.01$), with general language texts requiring significantly more pauses per word than domain-specific texts. Finally, there was also a significant interaction effect between task and text type ($p < 0.05$). The difference in pause density between post-

editing and from-scratch translation for general language texts was smaller than the difference in pause density between post-editing and from-scratch translation for domain-specific texts. Post-editing domain-specific texts was also found to trigger significantly fewer pauses per word than post-editing general language texts ($p < 0.001$). There was little difference in pause density when translating these two text types from scratch. These findings are in line with those of Schaeffer *et al.* (2016), who also reported that PWR scores were significantly lower during post-editing compared to from-scratch translation.

4.2.2. Cognitive effort: pause duration per word

The third dependent variable in our LMER is the average pause duration per word (Pdur1000), which indicates the total pause duration per sentence divided by the number of words in the source text sentence. The effect is presented in Figure 3.

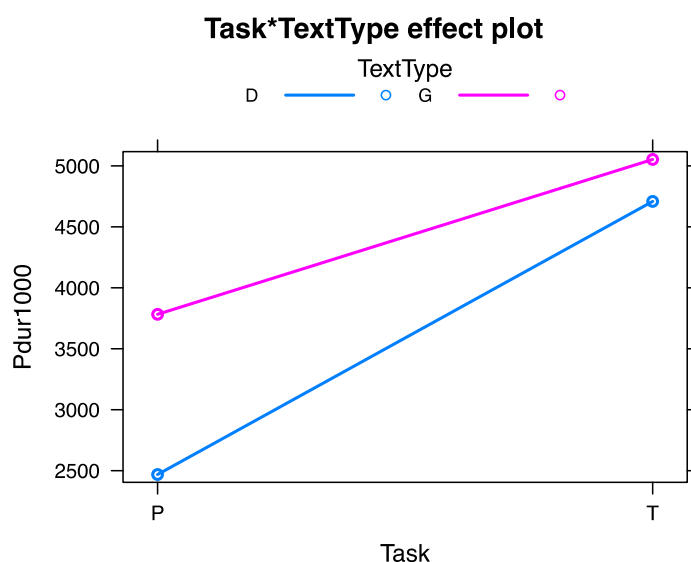


Figure 3. Effect plot of the interaction effect between task (post-editing and from-scratch translation, P and T, respectively) and text type (domain-specific texts and general language texts, D and G, respectively) on pause duration per word (Pdur1000).

The overall results reveal that the main factor task had a significant effect on pause duration per word. Post-editing caused significantly shorter pause duration per word than from-scratch translation ($p < 0.001$). This effect is significant for both domain-specific and general language texts. Post-editing domain-specific texts needed 2240ms less in pause time per word than from-scratch translation ($p < 0.01$). Post-editing general language texts took 1270ms less in pause time than from-scratch translation ($p < 0.05$). The main factor text type was also a significant predictor, with general language texts leading to longer pauses than domain-specific texts ($p < 0.05$). However, this effect is only significant for post-editing. Post-editing domain-specific texts resulted in significantly shorter pause time per word than general language texts ($p < 0.01$). Translating domain-specific texts from

scratch also led to less pause time per word than general language texts, but the difference was not significant. These results support Koglin (2015) but contradict Screen (2017).

Interestingly, our findings for pause density and pause duration indicate that post-editing and from-scratch translation involve different pause behaviours. Post-editing triggers shorter pauses and lower pause density, suggesting that post-editing was, for our participants, cognitively less demanding than from-scratch translation irrespective of the text types the participants worked with. These findings can be explained using relevance theory (Sperber and Wilson 1986), which views translation as a process of searching for interpretive resemblance between the source text and the corresponding target text (Gutt 1991). The principle of relevance regulates the cognitive effort spent and cognitive effect achieved, as “the human being’s cognitive environment searches for the generation of the maximum cognitive effects possible while spending the minimum processing effort necessary to achieve this end” (Alves and Gonçalves 2013: 109). As one of the objectives of post-editing is to produce target texts with human-like quality by making full use of the raw MT output to increase productivity, when machine translation is found to be good enough to fulfil the cognitive effects, the participants will most likely stop investing more cognitive effort in looking for alternative possible translations for the source text units. When translating from scratch, in most cases the translators may generate several translations for a chunk of source text and then select the one that optimally realises the cognitive effects. Post-editing may, therefore, save the cognitive effort needed to make decisions when there are multiple choices for certain source text words as well as the time required for consulting external resources.

In this study, no significant difference was found in translation time between post-editing and from-scratch translation for general language texts, but a statistically significant difference was observed in cognitive effort indicated by both pause duration per word and pause density. This finding corresponds to results obtained by da Silva *et al.* (2017) and the argument of Krings (2001) and O’Brien (2011) that temporal effort and cognitive effort each seem to have their own strength in explaining the post-editing and translation processes.

4.3. Product: quality

As the time and effort saved during the post-editing process is worthwhile only on condition that the quality of the final product is not compromised, the output quality of the GNMT as well as that of the post-editing and from-scratch translations is analysed in this section.

4.3.1. Data selection

The 99 tasks included 952 sentences altogether. Having all these sentences manually evaluated would be very time-consuming, so due to time restrictions text ST2 and text ST3 were selected as representatives of the domain-specific texts and the general language texts, respectively. Both of these texts contained 11 sentences, so there were 22 sentences altogether. The data from the first three participants of each group, who finished the tasks of the two texts with no missing translations, were selected, namely the translations of participants P1, P2, P4, P17, P18 and P19. The quality of the raw GNMT output was also evaluated. The total number of sentences to be evaluated was 154 (i.e. seven versions of the 22 source sentences), although a larger sample would be statistically more convincing.

4.3.2. The evaluators' profile

Four native Chinese-speaking evaluators participated in the evaluation task. Two evaluators were professional translators from two universities. They both had approximately 10 years of translation experience and had taught translation courses at undergraduate and graduate levels. The other two evaluators were PhD candidates specialising in translation studies. They had also worked as teaching assistants at both undergraduate and graduate levels. All the evaluators had rich and extensive experience in translation evaluation in this language combination.

4.3.3. Quality evaluation criteria and procedure

The adequacy and fluency criteria developed within TAUS's Dynamic Quality Evaluation Framework were employed to assess the translations (TAUS 2013b). The operational definition of each category can be found in Table 3.

Category	Rating scales and operational definition	Examples:
Fluency	4. Flawless Chinese: A perfectly flowing text with no errors.	ST: For plastic items not so marked, check the manufacturer's recommendations. TT(P17):对于无标志的塑料制品,请核对制造商的相关说明。 Back translation: For plastic items not so marked, please check the manufacturer's corresponding recommendations.
	3. Good Chinese: A smoothly flowing text even when a number of minor errors are present.	ST: Load sharp knives with the handles up to reduce the risk of cut-type injuries. TT(P18):尖锐刀具的手柄应朝上放置,降低刀口造成损害的风险。

		Back translation: For those sharp knives, load them with the handles up to reduce the risk of cut-type damage.
	2. Disfluent Chinese: A text that is poorly written and difficult to understand.	ST: For plastic items not so marked, check the manufacturer's recommendations. TT(MT):对于塑料物品不是这样标志, 检验生产者推荐。 Back translation: For plastic items not like this sign, test producer recommendations.
	1. Incomprehensible Chinese: A very poorly written text that is impossible to understand.	ST: We offer the world a portfolio of drinks brands that anticipate and satisfy people's desires and needs. TT(P04):我们提供了在世界上饮料品牌一个作品集, 满足了人们欲望需求。 Back translation: We offered in the world drink brands a sample reel, satisfied people desire need.
Accuracy	4. Everything: All the meaning in the source is contained in the translation, no more, no less.	ST: For plastic items not so marked, check the manufacturer's recommendations. TT(P02):对于没有这样标记的塑料制品, 请查看制造商的建议。 Back translation: For plastic items not so marked, please check the manufacturer's recommendations.
	3. Most: Almost all the meaning in the source is contained in the translation.	ST: For plastic items not so marked, check the manufacturer's recommendations. TT(P17): 对于无标志的塑料制品,请核对制造商的相关说明。 Back translation: For plastic items with no marks, please check the manufacturer's corresponding instructions.
	2. Little: Fragments of the meaning in the source are contained in the translation.	ST: Locate sharp items so that they are not likely to damage the door seal. TT(MT): 把锋利的物品拿出, 以免损坏门盖。 Back Translation: Take out sharp items, to avoid damaging the door panel.
	1. None: None of the meaning in the source is contained in the translation.	ST: Load sharp knives with the handles up to reduce the risk of cut-type injuries. TT (P02) : 用手柄装上锋利的刀, 以减少切割伤害型。 Back Translation: Use the handles to install the knives to reduce cut damage type.

Table 3. Operational definition of rating categories used in quality assessment.

Each evaluator was presented with a source sentence as well as seven candidate translations in randomised presentation order. Of the seven

versions, three were translated from scratch, three were post-edited, and one was the GNMT output. The evaluators were not provided with any information about the origin of the translations. While both source and target sentences were always visible, the evaluators were instructed to first assess fluency by relying on the target sentences only, and then to evaluate adequacy by comparing the source and target sentences. Finally, they were asked to select the best translation out of the seven candidates.

4.3.4. Inter-rater reliability

We measured the inter-rater reliability with Fleiss's kappa (Fleiss 1971). The resulting kappa scores for fluency and accuracy were 0.0334 and 0.0744, respectively. Both scores indicate that the raters' agreement was only slightly above chance. This result is in line with the general impression of the raters that they found the quality of the seven versions sometimes difficult to inter-distinguish. Assessing translation quality is known to be an extremely complicated and subjective task, and low agreement between raters was also found in previous studies. Vieira (2016) reported low agreement for fluency scores and the reliability rates found in Carl *et al.* (2011) were also only marginally better than by chance.

The effects of output type (raw GNMT, post-editing, and from-scratch translation) and text type (domain-specific and general language) on average fluency (AVEFluency) and average accuracy (AVEAccuracy) are investigated below in Sections 4.3.5 and 4.3.6, respectively. Given the framing of the categories, it was assumed that the distance between each of the four levels of fluency and accuracy in Table 3 was approximately equal, so that the scores of the 4 raters could be averaged, as suggested by previous studies which have adopted similar (e.g. Fiederer and O'Brien 2009) and the same (e.g. Vieira 2016) translation quality evaluation methods to this study.

4.3.5. Fluency

The fourth LMER model was built with the average fluency score (AVEFluency) as the dependent variable. The effect is shown in the plot in Figure 4. For the domain-specific text, the average fluency score for the post-editing output (3.25) was significantly higher than the score for the GNMT output (2.88) ($p < 0.01$). The output of from-scratch translation (3.31) also scored significantly higher than the GNMT output (2.88) ($p < 0.001$). There was no significant difference between from-scratch translation and post-editing in the average fluency score for the domain-specific text. For the general language text, only post-editing (3.33) scored significantly higher than the GNMT output in the average fluency score (3.0) ($p < 0.05$). In addition, there was no statistically significant difference between the output of from-scratch translation (3.19) and post-editing (3.33) for the general language text. It seems reasonable to conclude, therefore, that the output of post-editing and from-scratch translation are

equally fluent, both with scores above 3.0 (Good Chinese) for both text types. This finding also indicates that post-editing significantly improved the raw GNMT quality in terms of fluency irrespective of which text type the participants worked with.

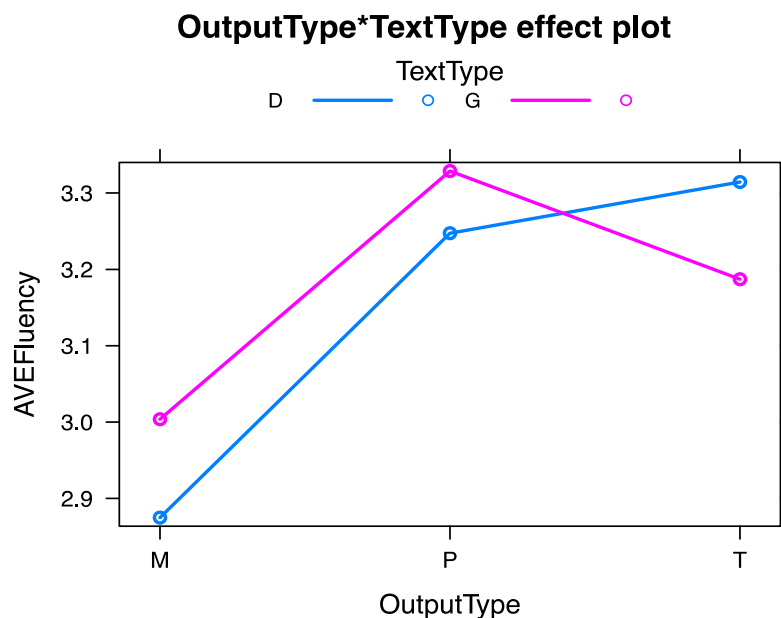


Figure 4. Effect plot of the interaction effect between output type (GNMT output, post-editing and from-scratch translation, M, P and T, respectively) and text type (domain-specific text and general language text, D and G, respectively) on the average fluency score (AVEFluency).

4.3.6. Accuracy

In the last LMER model, the average accuracy score (AVEAccuracy) was evaluated as the dependent variable. The effect is presented in Figure 5. It was found that, for the domain-specific text, post-editing output (3.2) scored significantly higher than the GNMT output (2.76) ($p < 0.01$) on the average accuracy score. The output of from-scratch translation (3.29) also scored significantly higher than the GNMT output ($p < 0.001$) on the average accuracy score. There was no significant difference in the average accuracy score between the output of from-scratch translation (3.29) and post-editing (3.2) for the domain-specific text. For the general language text, there was no significant difference in the average accuracy score between the output of GNMT (3.03), post-editing (3.19) and from-scratch translation (3.16). This is likely due to the fact that “most information [was] included” in the GNMT output, as the relatively high AVEAccuracy (3.03) score suggested.

In this study, it was noted that, for both fluency and accuracy, the MT output of the general language text scored higher than that of the domain-specific text, but post-editing general language texts was found to be more time-consuming and cognitively more demanding. This could be because the participants spent more time and effort refining the language and style of

the general language texts than of the domain-specific texts. This was indicated by the retrospective reports, which are discussed in more detail in Section 4.4.

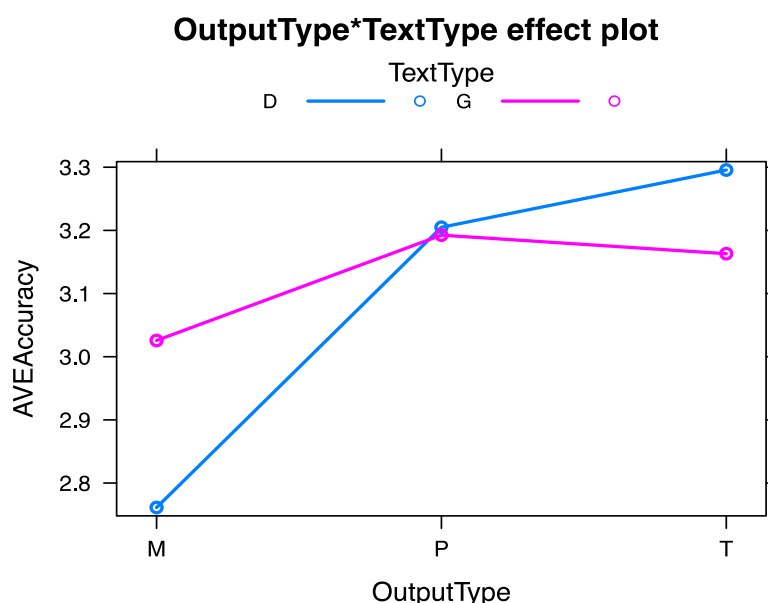


Figure 5. Effect plot of the interaction effect between output type (GNMT output, post-editing and from-scratch translation, M, P and T, respectively) and text type (domain-specific text and general language text, D and G, respectively) on the average accuracy score (AVEAccuracy).

4.3.7. Best translation selected by the raters

The raters were instructed to choose the best translation out of the seven candidates. In some cases, they chose two best translations when they found them comparable. As shown in Figure 6, the raters showed a clear preference for sentences translated from scratch (HT) over post-edited (PE) ones: 60.2% to 33.7%. This preference was more remarkable for the domain-specific text: 69.2% (HT) to 25% (PE). For the general language text, it was 50% (HT) to 43.5% (PE). Interestingly, as reported above, there was no significant difference in accuracy and fluency between post-editing and from-scratch translation. Post-edited sentences even scored slightly higher in some cases. For the domain-specific text, post-edited sentences scored slightly higher than those translated from-scratch in terms of accuracy but lower in terms of fluency. For the general language text, post-editing scored slightly higher both in fluency and accuracy. Therefore, the data were checked in detail to explore what might have caused this preference.

Raters selected those sentences with the highest scores in terms of both fluency and accuracy to be the best translation. In cases where the accuracy and fluency scores were not consistent, the raters generally valued accuracy over fluency. There were altogether 15 instances where several candidates

were awarded the same scores for fluency and accuracy. In 11 out of the 15 instances, the raters selected the sentences translated from scratch as the best translation over the post-edited ones. This finding indicates that the preference in these situations may be due to criteria other than fluency and accuracy.

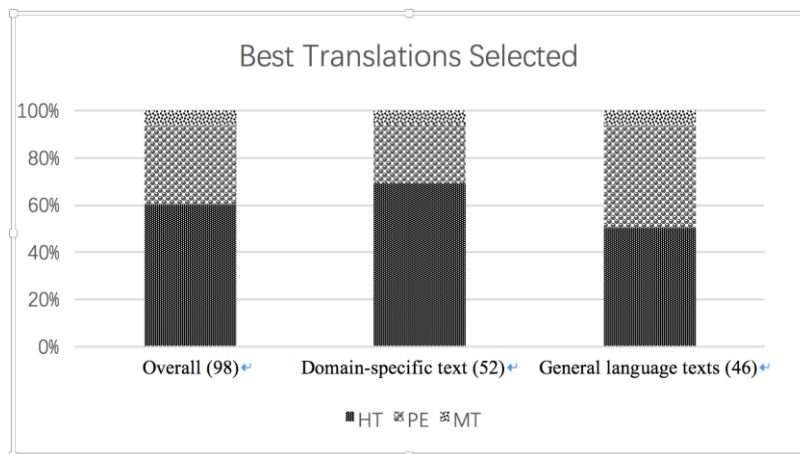


Figure 6. Best translations selected by the four raters.

Similar findings have been reported by Fiederer and O'Brien (2009), where raters also preferred sentences translated from scratch over post-edited ones. Fiederer and O'Brien speculated that style might have carried more weight than clarity and accuracy in their study. This may also be the case in our study, as the post-editing guidelines state that the post-edited output should be "stylistically fine though the style may not be as good as that achieved by a native-speaking human translator" (TAUS 2016). The from-scratch translation output may be stylistically more refined compared with the post-editing output. In the future, we intend to include style as a criterion in quality evaluation to see whether there is any correlation between the style score and the best translation selected. In addition, retrospective interviews or think-aloud protocols (TAPs) about the raters' evaluation process may also offer more persuasive explanations for this phenomenon.

4.4. Students' interpretations of the post-editing process

The pre-test questionnaire contained questions about students' attitude towards post-editing before the experiment. Based on their former experience with on-line MT engines, only one participant thought MT was of good quality. Fifteen thought the raw MT outputs were generally of average quality. Fourteen thought they were of bad or very bad quality. In terms of speed, twenty-two participants supported the idea that post-editing could increase translation speed. Twenty-five agreed or strongly agreed that post-editing would provide them with new sources of work and new professional skills. Most of the participants had a very positive attitude towards post-editing before the experiment, although they all had almost

no former post-editing experience. This may result from the introductory class before the experiment, which gave them general knowledge of MT, post-editing and the industry demand for post-editors.

The post-experiment questionnaire explored the participants' perceptions of the speed, mental effort, and quality of post-editing NMT. The qualitative data of the retrospective report is incorporated into the results of the post-experiment questionnaire to help understand their choices in the questionnaire. All of the students were convinced that post-editing was faster than from-scratch translation after the experiment. Twenty-four claimed that they felt post-editing was mentally less demanding than from-scratch translation. In the retrospective report, most of them mentioned that, by providing them with NMT output, post-editing saved them time and effort due to not having to type the whole translation or not having to consult external resources to check all lexical information they did not know. They found this was especially helpful for domain-specific texts. Those who felt post-editing was as tiring as from-scratch translation thought the raw MT output could be misleading. They believed correcting these mistakes sometimes took more effort than translating them from scratch. As far as quality was concerned, twenty-six students thought that their from-scratch translation output was of higher quality than the post-edited output. All of them believed that post-editing was more helpful for domain-specific texts than for general language texts.

In addition, the students found from-scratch translation and post-editing very different in many aspects. Eleven students mentioned that there was much less room for them to show their creativity and linguistic skills and also less freedom during post-editing, especially for the general language texts. They also claimed that they would pay more attention to the integrity and cohesion of their translation during from-scratch translation, as their former translation training programmes always asked them to produce flawless and elegant translations. When post-editing, however, they focused mainly on the accuracy of their translations, as the post-editing guidelines asked them to use as much of the MT output as possible. In addition, six students pointed out that their post-editing output was more literal than those translated from scratch. This might be a reason why most of the students thought their own translation was better in quality compared to their post-editing output. This may also explain why all of them considered that learning post-editing skills was necessary, which indicated that they found post-editing and from-scratch translation very different in terms of the translation skills and strategies involved. Finally, twenty-nine students agreed or strongly agreed that systematic post-editing training courses should be included in the regular MTI curriculum.

The retrospective report also revealed that the students came across different kinds of challenges when post-editing. Eleven students mentioned that they found it difficult to decide which MT translations were to be retained and which ones were to be corrected during the post-editing

process. Making good use of the MT output was, therefore, quite challenging for them. On the one hand, many of them felt compelled to perfect all the flaws in the MT output. On the other hand, seven students found that they sometimes relied too much on the MT output and missed MT errors. In addition, ten students expressed a desire for the post-editing guidelines to offer more detailed and explicit guidance.

5. Conclusion

Based on qualitative and quantitative data from key logging, screen recording, retrospective reports, questionnaires and quality evaluations, the research questions posed at the beginning of the paper are addressed.

First, post-editing NMT was generally faster than from-scratch translation, but this effect was only significant for domain-specific texts. Post-editing triggered significantly lower pause density and shorter pause duration than from-scratch translation for both text types, which indicates that post-editing is cognitively less demanding than from-scratch translation. In addition, for from-scratch translation, domain-specific texts took more time and led to only slightly lower pause density and shorter pause duration per word as compared to general language texts. Post-editing domain-specific texts, however, took much less time and significantly lower pause density and shorter pause duration than post-editing general language texts. Therefore, these results indicate that the impact of text type on the from-scratch translation process is different from its impact on the post-editing process.

Second, the quality evaluation results imply that the translation products of NMT post-editing and from-scratch translation were comparable in terms of fluency and accuracy, both for domain-specific texts and general language texts. However, this conclusion is only tentative, given the relatively small amount of data (66 translated sentences and 66 post-edited sentences from 6 translators) used for quality evaluation. We intend to increase the sample size in subsequent studies to see whether these results still hold true. The results also indicate that post-editing remarkably improved the quality of the raw NMT output. In future studies, a detailed analysis of the errors in the NMT, post-editing and from-scratch translation output could reveal more specific information concerning how post-editing improves NMT output and whether the students changed more than the post-editing guidelines required.

Finally, the students generally demonstrated a positive attitude towards NMT post-editing, although they were not totally accustomed to this new way of translating. They found post-editing involved challenges, translation skills and strategies that were different from from-scratch translation. As expertise research found that expert skills and knowledge cannot be easily transferred from one domain to another (Ericsson 2006), it is highly recommended that post-editing training programmes should be added to

universities' programmes for translator training. It is reasonable to expect that, through systematic training, the student translators will benefit more from post-editing, in terms of saving processing time and reducing cognitive effort, as compared to from-scratch translation. Future studies should also involve professional post-editors to obtain more profound insights into post-editing expertise.

There are a number of limitations to this study. First, eye-tracking data would be more informative in measuring the cognitive processes of post-editing and from-scratch translation. Eye-tracking data would be helpful in analysing which specific characteristics in the source texts and in the NMT output require more cognitive effort. This might better illustrate and explain the finding that the cognitive processes of post-editing and from-scratch translation varied for different text types. Second, although four raters were employed, quality evaluation cannot be absolutely objective and the sample size was small. These limitations will be carefully considered in future research.

Acknowledgements

Thanks go to our participants, annotators and raters for their precious time. Particular gratitude is extended to Prof. Yves Gambier and Dr. Bingham Zheng for their comments on the earlier drafts of this article. Finally, we are also grateful to Dr. Lucas Nunes Vieira and the anonymous reviewer who provided thorough, constructive and thought-provoking comments. This research was supported by the Social Science Foundation of Hunan Province (17ZDB005), China Hunan Provincial Science & Technology Foundation ([2017]131) and the Foundation of Hunan University (531107024014).

References

- **Allen, Jeffery** (2003). "Post-editing." Harold Somers (ed.) (2003). *Computers and translation: A translator's guide*. Amsterdam: John Benjamins, 297-318.
- **Alves, Fabio** (2006). "A relevance-theoretic approach to effort and effect in translation: Discussing the cognitive interface between inferential processing, problem-solving and decision-making." *Proceedings of the International Symposium on New Horizons in Theoretical Translation Studies*. Hong Kong: Chinese University of Hong Kong Press, 1-12.
- **Alves, Fabio and José Luiz Gonçalves** (2013). "Investigating the conceptual-procedural distinction in the translation process: a relevance-theoretic analysis of micro and macro translation units." *Target* 25(1), 107-124.
- **Bahdanau, Dzmitry, Cho, Kyunghyun and Yoshua Bengio** (2014). "Neural machine translation by jointly learning to align and translate." *arXiv preprint arXiv:1409.0473*.
- **Balling, Laura** (2008). "A brief introduction to regression designs and mixed-effects modelling by a recent convert." Jakobsen Arnt Lykke, Susanne Göpferich and Inger M.

Mees (eds) (2008). *Looking at eyes: Eye-tracking studies of reading and translation processing*. Copenhagen: Samfundslitteratur, 175–191.

- **Bates, Douglas, Maechler, Martin, Bolker, Ben, Walker, Steven, Christensen, Rune Haubo Bojesen, Singmann, Henrik, Dai, Bin, Scheipl, Fabian, Grothendieck, Gabor and Peter Green** (2014). "lme4: Linear mixed-effects models using Eigen and S4." *R package version 1.7*, 1-23. <https://CRAN.R-project.org/package=lme4> (consulted 21.11.2018).
- **Bojar, Ondřej, Chatterjee, Rajen, Federmann, Christian, Graham, Yvette, Haddow, Barry, Huck, Matthias, Jimeno Yepes, Antonio, Koehn, Philipp, Logacheva, Varvara, Monz, Christof, Negri, Matteo, Névél, Aurélie, Neves, Mariana, Popel, Martin, Post, Matt, Rubino, Raphael, Scarton, Carolina, Specia, Lucia, Turchi, Marco, Verspoor, Karin and Marco Zampieri** (2016). "Findings of the 2016 Conference on Machine Translation (WMT16)." *Proceedings of the First Conference on Machine Translation. Volume 2: Shared Task Papers, Berlin 2016*, 131–198. <http://www.aclweb.org/anthology/W16-2200> (consulted 20.11.2018).
- **Carl, Michael, Dragsted, Barbara, Elming, Jakob, Hardt, Daniel and Arnt Lykke Jakobsen** (2011). "The Process of Post-Editing: A pilot study." Bernadette Sharp, Michael Zock, Michael Carl and Arnt Lykke Jakobsen (eds) (2011). *Proceedings of the 8th International NLPCS Workshop. Special Theme: Human-Machine Interaction in Translation*. Copenhagen: Samfundslitteratur, 131-142.
- **Carl, Michael, Gutermuth, Silke and Silvia Hansen-Schirra** (2015). "Post-editing Machine Translation: Efficiency, strategies and revision processes in professional translation settings." Aline Ferreira and John Schwieter (eds) (2015). *Psycholinguistic and Cognitive Inquiries into Translation and Interpreting*. Amsterdam/Philadelphia: John Benjamins Publishing Company, 145-174.
- **Carl, Michael, Schaeffer, Moritz and Srinivas Bangalore** (2016). "The CRITT translation process research database." Michael Carl, Srinivas Bangalore and Moritz Schaeffer (eds) (2016). *New directions in empirical translation process research*. Cham: Springer, 13-54.
- **Castilho, Sheila, Moorkens, Joss, Gaspari, Federico, Calixto, Iacer, Tinsley, John and Andy Way** (2017). "Is neural machine translation the new state of the art?" *The Prague Bulletin of Mathematical Linguistics* 108, 109–120.
- **da Silva, Igor A. Lourenço, Alves, Fabio, Schmaltz, Márcia, Pagano, Adriana, Wong, Derek, Chao, Lidia, Leal, Ana Luísa V., Quaresma, Paulo, Garcia, Caio and Gabriel Eduardo da Silva** (2017). "Translation, post-editing and directionality: A study of effort in the Chinese-Portuguese language pair." Arnt Lykke Jakobsen and Bartolomé Mesa-Lao (eds) (2017). *Translation in transition: Between cognition, computing and technology*. Amsterdam: John Benjamins, 91-117.
- **Daems, Joke, Vandepitte, Sonia, Hartsuiker, Robert J. and Lieve Macken** (2017). "Translation methods and experience: A comparative analysis of human translation and post-editing with students and professional translators." *Meta* 62, 245-270.
- **Ericsson, K. Anders** (2006). "An introduction to Cambridge Handbook of Expertise and Expert Performance: Its development, organization, and content." K. Anders Ericsson, Neil Charness, Paul J. Feltovich and Robert R. Hoffman (eds) (2006). *The Cambridge handbook of expertise and expert performance*. Cambridge: Cambridge University Press, 3-20.

- **Fiederer, Rebecca and Sharon O'Brien** (2009). "Quality and machine translation: A realistic objective?" *The Journal of Specialised Translation* 11, 52-74.
- **Fleiss, Joseph L.** (1971). "Measuring nominal scale agreement among many raters." *The Prague Bulletin of Mathematical Linguistics* 76(5), 378–382.
- **Garcia, Ignacio** (2010). "Is machine translation ready yet?" *Target* 22(1), 7-21.
- **Germann, Ulrich** (2008). "Yawat: Yet Another Word Alignment Tool." *ACL-08. Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies. HLT Demo Session (Companion Volume), Columbus, Ohio 2008*. The Association for Computational Linguistics, 20-23. <http://www.aclweb.org/anthology/P/P08/P08-40.pdf> (consulted 20.11.2018).
- **Green, Spence, Heer, Jeffrey and Christopher D. Manning** (2013). "The efficacy of human post-Editing for language translation." *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (ACM). Association for Computing Machinery, 439-448. <https://dl.acm.org/citation.cfm?id=2470718> (consulted 21.11.2018).
- **Groves, Declan and Dag Schmidtke** (2009). "Identification and analysis of post-editing patterns for MT." *MT Summit XII – The twelfth Machine Translation Summit International Association for Machine Translation hosted by the Association for Machine Translation in the Americas*. Association for Machine Translation in the Americas, 429-436. <http://www.mt-archive.info/MTS-2009-Groves.pdf> (consulted 21.11.2018).
- **Guerberof, Ana** (2009). "Productivity and quality in MT post-editing." Marie- Laurie Gerber, Pierre Isabelle, Roland Kuhn, Nick Bemish, Mike Dillinger and Marie-Josée Goulet (eds) (2009). *Beyond Translation Memories Workshop. MT Summit XII. The twelfth Machine Translation Summit. International Association for Machine Translation, Ottawa, August 26-30*. Association for Machine Translation in the Americas. <http://www.mt-archive.info/MTS-2009-Guerberof.pdf> (consulted 11.12.2017).
- **Gutt, Ernst-August** (1991). *Translation and Relevance: Cognition and Context*. Manchester: St Jerome.
- **Hansen, Gyde** (2002). *Empirical Translation Studies: Process and Product (Copenhagen Studies in Language 27)*. Copenhagen: Samfundslitteratur.
- **Jakobsen, Arnt Lykke** (1998). "Logging time delay in translation." Gyde Hansen (ed.) (1998). *LSP Texts and the Process of Translation*. (Copenhagen Working Papers in LSP) Copenhagen: Copenhagen Business School, 71-101.
- — (2002). "Translation drafting by professional translators and by translation students, in empirical translation studies: Process and product." Gyde Hansen (ed.) (2002). *Empirical Translation Studies: Process and Product*. (Copenhagen Studies in Language Series 27) Copenhagen: Samfundslitteratur, 191-204.
- **Junczys-Dowmunt, Marcin, Dwojak, Tomasz and Hieu Hoang** (2016). "Is neural machine translation ready for deployment? A case study on 30 translation directions." <https://arxiv.org/abs/1610.01108>
- **Klubička, Filip, Toral, Antonio and Víctor M. Sánchez-Cartagena** (2017). "Fine-grained human evaluation of neural versus phrase- based machine translation." *The Prague Bulletin of Mathematical Linguistics* 108, 121–132.

- **Koglin, Arlene** (2015). "An empirical investigation of cognitive effort required to post-edit machine translated metaphors compared to the translation of metaphors." *Translation & Interpreting* 7, 126-141.
- **Krings, Hans P.** (2001). *Repairing Texts: Empirical Investigations of Machine Translation Post-editing Processes* (Geoffrey Koby, ed.). Kent: Kent State University Press.
- **Lacruz, Isabel, Shreve, Gregory M. and Erik Angelone** (2012). "Average pause ratio as an indicator of cognitive effort in post-editing: A case study." Sharon O'Brien, Michel Simard and Lucia Specia (eds) (2012). *Proceedings of the AMTA 2012 Workshop on Post-editing Technology and Practice, San Diego, 28 October*. Association for Machine Translation in the Americas, 21-30. <http://www.mt-archive.info/10/AMTA-2012-WS-WPTP.pdf> (consulted 05.12.2018).
- **Lacruz, Isabel and Gregory M. Shreve** (2014). "Pauses and Cognitive Effort in Post-Editing." Sharon O'Brien, Laura Winther Balling, Michael Carl, Michel Simard and Lucia Specia (eds) (2014). *Post-editing of machine translation: Processes and applications*. Cambridge: Cambridge Scholars Publishing, 246-272.
- **Lommel, Arle R. and Donald A. DePalma** (2016). *Post-Editing Goes Mainstream: How LSPs Use MT to Meet Client Demands*. Cambridge MA: Common Sense Advisory. http://www.commonsenseadvisory.com/Portals/default/Knowledgebase/ArticleImages/1605_R_IP_Postediting_goes_mainstream-extract.pdf (consulted 07.11.2018).
- **Mesa-Lao, Bartolomé** (2014). "Gaze behaviour on source texts: An exploratory study comparing translation and post-editing." Laura Winther Balling, Michael Carl, Michael Simard and Lucia Specia (eds) (2014). *Post-editing of machine translation: Processes and applications*. Newcastle: Cambridge Scholars Publishing, 219-245.
- **O'Brien, Sharon** (2006). "Pauses as indicators of cognitive effort in post-editing machine translation output." *Across Languages and Cultures* 7, 1-21.
- — (2007). "An empirical investigation of temporal and technical post-editing effort." *Translation and Interpreting Studies* 2, 83-136.
- — (2010). "Introduction to post-editing: Who, what, how and where to next." Paper presented at *The Ninth Conference of the Association for Machine Translation in the Americas* (Denver, Colorado 31 October – 4 November 2010). <http://www.mt-archive.info/10/AMTA-2010-OBrien.pdf> (consulted 21.11.2018).
- **O'Brien, Sharon** (2011). "Towards predicting post-editing productivity." *Machine Translation*, 25(3), 197-215.
- **Plitt, Mirko and François Masselot** (2010). "A productivity test of statistical machine translation post-editing in a typical localisation context." *The Prague Bulletin of Mathematical Linguistics* 93, 7-16.
- **Popović, Maja** (2017). "Comparing language related issues for NMT and PBMT between German and English." *The Prague Bulletin of Mathematical Linguistics* 108, 209-220.
- **R Core Team** (2014). "R: A language and environment for statistical computing." R Foundation for Statistical Computing. Vienna. <http://www.R-project.org/> (consulted 10.11.2018).

- **Schaeffer, Moritz, Carl, Michael, Lacruz, Isabel and Akiko Aizawa** (2016). "Measuring cognitive translation effort with activity units." *Baltic Journal of Modern Computing* 4, 331-345.
- **Screen, Benjamin** (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* 28, 317-344.
- **Sennrich, Rico, Haddow, Barry and Alexandra Birch** (2016). "Neural machine translation of rare words with subword units." *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 1715-1725. <http://www.aclweb.org/anthology/P16-1162> (consulted 21.11.2018).
- **Sperber, Dan and Deirdre Wilson** (1986). *Relevance: Communication and Cognition*. Oxford: Blackwell.
- **Tatsumi, Midori** (2009). "Correlation between automatic evaluation metric scores, post-editing speed and some other factors." *MT Summit XII – The twelfth Machine Translation Summit International Association for Machine Translation hosted by the Association for Machine Translation in the Americas*. Association for Machine Translation in the Americas, 332-339. <http://www.mt-archive.info/MTS-2009-Tatsumi.pdf> (consulted 21.11.2018).
- **TAUS** (2013a). *Translation Technology Landscape Report*. <https://www.taus.net/thinktank/reports/translate-reports/taus-translation-technology-landscape-report> (consulted 01.04.2017).
- – (2013b). *Adequacy/Fluency Guidelines*. <https://taus.net/academy/best-practices/evaluate-best-practices/adequacy-fluency-guidelines> (consulted 10.05.2017).
- – (2016). *TAUS Post-editing Guidelines*. <https://www.taus.net/think-tank/articles/postedit-articles/taus-post-editing-guidelines> (consulted 04.04.2017).
- **Tirkkonen-Condit, Sonja** (1990). "Professional vs. non-professional translation: A think-aloud protocol study." Michael Halliday, Alexander Kirkwood, John Gibbons and Howard Nicholas (eds) (1990). *Learning, keeping and using language: Selected papers from the eighth World Congress of Applied Linguistics*. Amsterdam: John Benjamins, 381-394.
- **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, Masaru** (2015). "Can college students be post-editors? An investigation into employing language learners in machine translation plus post-editing settings." *Machine Translation* 29, 49-67.

Biographies

Yanfang Jia is a PhD candidate at Hunan University (China). She has been using eye tracking and keylogging to study the cognitive aspects of human translation and post-editing, with a special focus on measuring and predicting cognitive effort in post-editing and human translation.



E-mail: yanfangjia@hnu.edu.cn

Michael Carl is a Professor at Kent State University (USA) and Director of the Center for Research and Innovation in Translation and Translation Technology (CRITT). He has studied Computational Linguistics and Communication Sciences in Berlin, Paris and Hong Kong and obtained his PhD degree in Computer Sciences from the Saarland University/Germany. His current research interest is related to the investigation of human translation processes and interactive machine translation.



E-mail: mcarl6@kent.edu

Xiangling Wang is a Professor at Hunan University and Director of Centre for Studies of Translation, Interpreting and Cognition. She obtained her PhD degree in Translation Studies from Hunan Normal University in China. She's currently hosting a few projects both at national level and state level on translation process and interactive machine translation.



E-mail: xl_wang@hnu.edu.cn

Notes

¹ The Test for English Majors Band 4 and Band 8 are national English tests for English majors in China, which require a candidate to master 8,000 and 13,000 words, respectively.