Efforts of Post-Editing Literary Texts Using Google NMT: An Eye-Tracking Study on L1–L2 Translation

First Author
Affiliation / Address line 1
Affiliation / Address line 2
Affiliation / Address line 3
email@domain

Second Author
Affiliation / Address line 1
Affiliation / Address line 2
Affiliation / Address line 3
email@domain

Abstract

Despite the growing research interest in applying machine translation and post-editing to literary texts, the translator's cognitive process of engaging with translation technologies in such tasks is largely underexplored. This article reported an experimental study using eye-tracking and keystroke logging technologies to compare the temporal, technical, and cognitive aspects of efforts expanded by trainee translators between from-scratch translation and post-editing tasks against the background of translating Chinese novels into English. The results indicate that post-editing largely saved time and lightened technical and cognitive effort as compared to translation from scratch, providing promising possibilities of aiding forward translation process with machine translation systems.

1 Introduction

Despite the drastic improvement in machine translation (MT) output quality, MT systems still generate various categories of errors (Daems *et al.*, 2017a), leading to a quality gap between raw MT outputs and publishable translation products. As a method to bridge such a gap and raise translation productivity, machine translation post-editing (MTPE, henceforth PE) came into play and has become an established workflow in the language service industry (Nitzke & Hansen-Schirra, 2021).

Numerous empirical studies investigated the process and product of PE. These studies have offered evidence that PE is able to improve productivity without compromising product quality for both general and technical non-literary translation tasks (Garcia, 2011; Guerberof-Arenas, 2014; Plitt & Masselot, 2010; Yang *et al.*, 2021). However, this batch of studies rarely used literary texts as experimental materials. This is partly because literary translation was traditionally considered the most challenging translation type, and was long regarded as an unreachable domain for machine translation (Hadley *et al.*, 2022a).

This assumption has been challenged recently, especially after the application of NMT systems and more recently, of LLM-based generative models to the translation practice. There is an increasing interest in exploring the possibilities and restrictions of introducing PE, among other types of technology-aided translation workflows, into the context of literary translation (Guerberof-Arenas & Toral, 2020, 2022, 2023; Hadley *et al.*, 2022b; Taivalkoski-Shilov, 2019; Toral & Way, 2018). This strand of studies mostly investigated the translation direction of back translation (i.e., from a translator's second language (L2) to her first language (L1)), which is a traditionally recommended direction for translators (Samuelsson-Brown, 2010). However, in some countries, forward translation (or L1–L2 translation) is widely accepted as a market reality due to the lack of target-language native speakers working as translators (Horcas-Rufián, 2022; Pokorn, 2005). In China, for example, even much of Chinese literature is translated into English by native Chinese speakers (see Wu (2021, pp. 325–344) for a detailed list of published English translations of Chinese novels in recent decades). Since studies have proved that translation direction will affect both the translation process and product (Chang, 2011; da Silva *et al.*, 2017; Pavlović & Jensen, 2009),

the circumstance of L1–L2 translation and post-editing of literary texts should also be studied within more language pairs and literary genres.

Against this background, this paper aims to explore post-editing literary texts by contrasting post-editing and from-scratch translation into translators' L2 with a special focus on the efforts expanded by the translators.

2 Related Works

2.1 Comparing Efforts of Non-Literary PE to From-Scratch Translation

The construct of PE effort (Krings, 2001) provides an analytical framework for not only the PE process, but also "a more general discussion of effort in translation and associated activities" (Lacruz, 2017, p. 386). The construct is divided into three measurements: temporal effort, cognitive effort, and technical effort. Temporal effort and technical effort can be measured directly: Temporal effort was simply measured by the time spent on a post-editing task. Technical effort is a dimension of translators' psychical cost, which is caused by "purely mechanical operations" (Krings, 2001, p. 179), including deletion, insertion, and reordering. Cognitive effort, however, can only be observed through indirect approaches such as think-aloud protocols, keyboard pause measurements, and eye-tracking technology.

Under this analytical framework, there have been a lot of studies comparing the PE process with from-scratch human translation (HT). For non-literary texts, PE was often observed to be faster than HT using both SMT (Carl *et al.*, 2015; Daems *et al.*, 2017b; Nitzke & Oster, 2016; Plitt & Masselot, 2010; Sun *et al.*, 2023) and NMT (Jia *et al.*, 2019b; Yang *et al.*, 2021) systems. There are some cases where no significant difference in the temporal effort was found. For example, Carl *et al.* (2011) reported only a slight difference in post-editing English news texts into Danish as compared to HT. Jia *et al.* (2019a) also found no significant differences in post-editing general texts, despite that there was a significant improvement in processing speed for post-editing domain-specific texts, suggesting that there might be an effect brought by text domains.

Concerning technical effort, results were consistent with the number of keyboard insertions (or text production activities): HT was observed to involve more insertions than PE (Jia *et al.*, 2019b; Nitzke, 2018; Nitzke & Oster, 2016), whereas the situation is more complex regarding keyboard deletion activities: sometimes PE was found to cost more deletions (e.g., Nitzke, 2018), while sometimes no significant difference was observed (e.g., Koglin, 2015). Jia *et al.* (2019b) found significantly more deletions in SMTPE than translation from-scratch, but no significant difference was observed for NMTPE and from-scratch translation.

Cognitive effort is regarded as the core of PE effort (Krings, 2001). In recent years, eye-tracking technology has been increasingly adopted to measure cognitive effort in the studies of PE (Moorkens, 2018), and was proven to be well-correlated with other objective measures of cognitive effort (Vieira, 2017). Using this method, Daems *et al.* (2017b) investigated the HT and PE of news for professional and trainee translators, using average fixation duration and average fixation counts as indicators of cognitive effort. They found longer average fixation duration and more fixation counts in HT tasks, with no significant difference across participants' levels of translation expertise. Sun *et al.* (2023) used average fixation duration, fixation counts, and pupil size as indicators to gauge cognitive effort in Chinese–English (L1–L2) general-domain translation tasks, finding significant decreases in all three measures for post-editing than from-scratch translation tasks. Along with the analysis of cognitive efforts, previous studies revealed differences in the allocation of visual resources across task modalities. For example, in Daems *et al.* (2017b), translators invested more effort in processing target text than processing the source text, with a more prominent difference in PE tasks, which is also supported by Carl *et al.* (2011).

2.2 Post-Editing Literary Texts

As mentioned before, the research interest in literary texts has only raised in recent years. Research on this topic includes studies on the PE process and product using either general-domain or self-trained literary-adapted MT systems, as well as discussions on user perceptions and ethical issues.

Toral *et al.* (2018) compared the efforts under three task types: HT, SMTPE, and NMTPE, using self-trained, literary-targeted SMT and NMT systems. They found that PE was lower in temporal effort, and PE resulted in fewer but longer keyboard pauses compared to HT. In terms of technical effort, PE involved fewer input operations but more erasing and navigation activities, with an overall decrease in the total number of keystrokes. Guerberof-Arenas and Toral (2020) focused on the creative aspects of PE products as well as the target readers' reception. In their study, HT was

found to present a higher creativity score compared to PE and raw MT outputs. As for the survey concerning reading experience, HT also scored higher on the scales of narrative engagement and translation reception, while PE ranked slightly higher in the enjoyment of reading. They did a follow-up study to address some limitations in the previous work, conducting a small-scale experiment to compare the level of creativity manifested in HT, PE, and raw MT products (Guerberof-Arenas & Toral, 2022). In this study, HT was found to boast the highest level of creativity, followed by PE and MT. Reductions in (objective measure of) technical and cognitive effort were found when using PE, while there was no significant difference in temporal effort. The findings were also supported by the study by Vieira *et al.* (2023), in which a general-domain NMT system was used to post-edit science fiction.

Alongside the works exploring the processes of literary PE, some researchers began to investigate the perceptions of literary translators on the new workflow and discuss some of the ethical issues that may arise. After the experiment reviewed above (Toral *et al.*, 2018), Moorkens *et al.* (2018) reported a survey concerning the participants' perceptions of MT and PE. Most participants preferred HT to PE, and NMTPE to SMTPE, complaining that PE would reduce their creativity and make them feel constrained, especially with sentence-level segmentation of the PE interface. Taivalkoski-Shilov (2019) discussed ethical issues regarding machine(-assisted) translation of literary texts, in which the ethical concerns in product quality, process quality, as well as voice and "noise" (i.e., the translation style of a specific MT system) were commented thoroughly. Kenny and Winters (2020) explored the impact of NMTPE on the textual voice, i.e., the translator's style, of an English–German professional translator, and found that the PE version somewhat diminished his textual voice compared to this translator's from-scratch translated version.

3 Methodology

Against the background of Chinese–English literary translation, the present study followed a one-factor repeated measures design to answer the following research question:

For translation trainees, what are the differences between the effort spent on the from-scratch translation and machine translation post-editing of novels?

3.1 Participants

Ethical approval was gained prior to the experiment. 31 translation trainees (27 females and 4 males) with an average age of 24.16 years (range 23-31 years, SD=1.99 years) took part in this study and were paid for their participation.

All of them were second-year postgraduate students majoring either in translation studies or translation and interpreting practice from a renowned university in China. All participants are native speakers of Chinese with English as their second language. All had passed the Test for English Majors in China-Band 8 (TEM-8) with an average score of 72 (range 60-84, SD=6.33), which proved their L2 competence as proficient users. All participants received more than 48 hours of Chinese–English written-translation training by the time of the experiment, but none of them had worked as professional translators or had taken any PE course before the experiment. All but one of the participants had taken courses on English literature and culture before, while 26 participants had taken courses on literary translation. All of them had experience translating literary texts from Chinese into English in examinations.

3.2 Materials

To retain the literariness of literary works, we did not make any adaptation to the original excerpt of the novels selected as experimental materials. The openings of two novels (Passage A and Passage B) were selected as source texts. Each excerpt was spilt into two sequential segments (Text A1, Text A2, Text B1, Text B2) for different task modalities (HT and PE), which helped to ensure that texts were comparable in linguistic and literary features across modality levels. When we separate source passages, we tried to make the four texts comparable in terms of word count, the number of characters, lexical variability (calculated by type/token ratios, TTR), and the number of lines (see Table 1), all of which are indicators of text complexity. As a result, each separated task consisted of an around-200-character source text, the length of which was similar to that of the translation tasks in the finals of the university's literary translation courses.

Table 1. Text Profiles

Text	Text A1	Text A2	Text B1	Text B2
Words(characters)	115(178)	122(194)	117(186)	145(216)
Lines	11	10	9	10
Lexical variety	72%	70%	71%	71%

Passage A was chosen from *Renshijian* by Xiaosheng Liang, while Passage B was selected from Tong Su's *Mi*. There are two reasons why we chose the two books as source materials: Firstly, the two books were both of high quality: *Renshijian* was awarded the 10th Mao Dun Literature Prize, one of the most prestigious literature prizes in China, while *Mi* enjoys positive critical reviews by the public, and was reprinted for many times since its first publication; Secondly, none of our participants has read either of the two books before the experiment.

Currently, there are no NMT systems built specifically for literary texts that can be easily accessed by freelance translators. Therefore, we chose free online NMT systems to produce raw MT output in an effort to fit the accessibility of MT systems in real life. We chose Google Translate, which is considered the most widely-used free online NMT system for general domains at the time of the experiment.

3.3 Experiment Settings

The experiment was carried out in an eye-tracking lab illuminated only by artificial light. Participants' eye movements were recorded by a Tobii TX300 Eye Tracker at a sampling rate of 300 Hz, while their keystroke activities were recorded by Translog-II (Carl, 2012).

The materials were presented on a 23'' LCD monitor connected to the eye tracker with a screen resolution of 1920 × 1080 pixels. On the presentation screen, we use the Translog-II interface to present source texts and type in target texts. To conform to the participants' work habits, the target window of the Translog-II interface was set to the right, i.e., the source texts were displayed in the left window of the interface in Simsun font size 20. The target texts were produced in the right window in Times New Roman font size 20. The texts were all single-paged to ensure that no scrolling was needed during the experiment. Tobii Studio software (version 3.4.8) was used to export screen recordings and eye data for data analysis.

3.4 Procedure

The participants were tested individually. Firstly, they were given a brief introduction to the eye-tracker and some notifications during the experiment. All tasks were executed with touch-typing in case the overly-long time spent on looking at the keyboard affected the eye data quality. The participants were asked to stay static as far as possible but were allowed to check the keyboard when necessary. After that, the participants were asked to read the instructions in which they were informed that they should regard this task as a publishable literary translation task so that they must guarantee the quality of their translation products, that the target readers are native English speakers with no or little knowledge of Chinese culture, and that there was no time limit for the tasks. The tasks lasted for 1.09 hours on average (warm-up exercises and the breaks between tasks excluded).

The participants were also instructed on the methods to end the task and save the logfiles. Then they were permitted to read an introduction to the two books used as the source of research materials, which helped them to be informed of the narrative backgrounds and settings. After that, they were asked to perform a typing exercise to get familiar with the layout of the keyboard to ensure better touch-typing performance, which was an English copying task with the same interface setting as experimental tasks.

After the participants were correctly positioned, a five-point calibration was executed. The tasks started when an acceptable calibration result was reached. There were in total four aforementioned source texts: two were for HT (Text A2 and Text B1) and two were for PE (Text A1 and Text B2). The two source texts for the PE task were machine-translated and were pasted into the target window in advance.

After each task, participants needed to manually save the keystroke logging data in XML format, and stop the Tobii Studio Screen Recording element to save the eye data.

3.5 Data Analysis

The quality of eye data was assessed before conducting data analysis. Three criteria proposed by Hvelplund (2014) were applied in this study: a) mean fixation duration (MFD); b) gaze time on the screen (GTS); and c) gaze sample to fixation percentage (GSF). Thresholds for the three criteria adopted by Cui and Zheng (2021) were followed in the present study: values within one standard deviation of the mean were considered acceptable in a certain measure,

and data that satisfied at least two out of the three measures were deemed valid. As a result, the data obtained from 19 participants were included in the data analysis.

The data were analysed using R 4.1.3 (R Core Team, 2024). All data were fitted in linear mixed effects regression (LMER) models using the lme4 package (Bates *et al.*, 2015). The models will be simplified by dropping random variables if a convergence issue occurs (Winter, 2019). Statistical significance was assessed by the lmerTest package (Kuznetsova *et al.*, 2017) using the Satterthwaite approximation to estimate the degree of freedom. The coefficient of determination (R^2) value was calculated by the MuMIn package (Bartoń, 2009).

4 Results

In this study, the independent variable is translation modality with two levels, namely, HT and PE. The dependent variables were: a) average processing time, as a measure of temporal effort; b) average insertion activities and c) average deletion activities, as measures of technical effort; and d) mean fixation duration and e) fixation counts, as measures of cognitive effort. A summary of means and standard deviations is reported in Table 2.

Table 2. Summary of Experiment Results

Dimension	Measurement		Mean (SD)	
			HT	PE
Temporal effort	Average processing time (s)		9.81 (2.83)	5.72 (2.59)
Cognitive effort	Mean fixation duration (ms)	AOI ST	243.16 (38.77)	206.58 (23.40)
		AOI TT	327.37 (59.44)	272.37 (27.55)
	Fixation counts	AOI ST	1102.29 (440.41)	544.87 (213.37)
		AOI TT	1660.03 (536.25)	1478.00 (615.47)
Technical effort	Average Insertions		9.00 (1.80)	1.80 (1.23)
	Average Deletions		2.90 (1.56)	1.65 (1.15)

4.1 Temporal Effort

In this study, average processing time, measured by the average time spent on processing a source-text word (total task time / number of source-text words), is used as an indicator of temporal effort.

In the LMER model built for average processing time, the fixed effect included the effect of modality, while random intercepts for participants and passages were included. As a result, the average processing time in HT tasks is significantly longer than that in PE tasks (b= -4.09, t = -11.67, $p < 0.001$, r^2_c =0.81).

4.2 Cognitive Effort

We use mean fixation duration (MFD) and fixation counts (FC) to indicate cognitive effort. The Translog-II interface was divided into two areas of interest (AOIs): source text (AOI ST) and target text (AOI TT), so that we may observe the difference between different processing stages.

In this part, we fit two LMER models with dependent variables being mean fixation duration and fixation counts, respectively. Fixed effects included the main effects of modality and AOI and the interaction effect of modality by AOI.

In the model for MFD, random intercepts for participants, as were random slopes for participants varying by modality were included. In the model for FC, random intercepts for participants and passages, as were random slopes for participants varying by modality were included. The data were transformed using a logit transformation in advance.

4.2.1 Mean Fixation Duration

The LMER analysis showed a significant main effect of modality (b= -0.07, t = -5.90, $p < 0.001$, r^2_c =0.77), and a significant main effect of AOI (b= 0.13, t = 12.12, $p < 0.001$). No significant interaction was found between the two factors (b = -0.007, t = -0.47, p =0.64). The results indicate that the MFD of HT tasks were significantly longer than PE tasks, and the MFD in AOI TT was significantly longer than in AOI ST. The difference between AOIs manifested a larger effect size for PE tasks (β=-0.20) than HT tasks (β=-0.13).

MFDs in each AOI were also analysed separately. The MFD in HT tasks is significantly longer than in PE tasks (b= 0.07, t = 5.90, $p < 0.001$) in AOI ST. Also, seeing AOI TT only, the MFD in HT tasks is significantly longer than in PE tasks (b= 0.08, t = 6.50, $p < 0.0001$).

4.2.2 Fixation Counts

According to the LMER analysis, the interaction effect between modality and AOI is significant (b = 0.25, t = 6.33, $p < 0.001$, r^2_c =0.77). There are more

fixation counts in AOI TT than in AOI ST and the difference is statistically significant in both HT (b = -0.18, t = -6.52, p < 0.001) and PE (b = -0.43, t = -15.46, p < 0.001) tasks, with the difference being more prominent in PE tasks.

Again, we observe the AOI ST and AOI TT separately. In AOI ST, HT tasks involved significantly more fixations than PE tasks (b = 0.31, t = 10.84, p < 0.001). However, in AOI TT, this difference only approached significance (b = 0.05, t = 1.93, p = 0.058).

4.3 Technical Effort

In this part, we use two indicators to measure technical effort: the average number of characters inserted per ST word (Ins) and the average number of characters deleted per ST word (Del) recorded by Translog-II. The insertion and deletion counts were extracted from the logfile and calculated using Python 3.11.

In the LMER model for Ins, the fixed effect included the effect of modality, while random intercepts for participants and passages were included. The results showed that Ins in HT tasks is significantly larger than that in PE tasks (b= -7.20, t = -24.36, p < 0.01, r^2_c =0.89).

In the model for Del, random intercepts for participants and passages, as were random slopes for participants varying by modality were included. Results show that the number of Del in HT tasks is significantly larger than that in PE tasks (b= -1.24, t = -4.13, p < 0.001, r^2_c =0.75).

5 Discussion

In the present study, we found that for trainee translators, from-scratch translation is temporally, technically, and cognitively more effortful than PE when translating literary texts from Chinese to English.

For temporal and cognitive efforts, our results are consistent with previous studies on non-literary L2–L1 PE (Daems *et al.*, 2017b; Jia *et al.*, 2019b; Yang *et al.*, 2021), non-literary L1–L2 PE (Sun *et al.*, 2023), and a literary L2-L1 PE study using literary-adapted MT (Toral *et al.*, 2018). However, as was mentioned before, there are some studies on literary PE where not much difference in the temporal effort was observed (Guerberof-Arenas & Toral, 2022; Vieira *et al.*, 2023). One possible explanation for this inconsistency is that the translation direction in our study (L1–L2) is different from previous studies. L1–L2 from-scratch translation has been proven to be significantly more time-consuming than L2–L1 translation in the Chinese–English pair (Chang, 2011). It is possible that post-editing from L1 to L2 saves more time than from L2 to L1.

We use the number of tokens inserted per ST word (Ins) and the number of tokens deleted per ST word (Del) to measure technical effort, and observed a decrease in both Ins and Del measures for the case of PE, while the difference in means is larger for the Ins measure. The reduction in Ins is consistent with previous works, while previous works found inconsistent results in terms of Del value as is reviewed above. As the previous scholarship on the PE process focused on various target languages, it is difficult to explain the inconsistency, as the number of characters typed in or deleted largely relies on the features of a certain target language, for example, the average word length, the frequency of functional-word use, etc. The finding in our study, though, suggests an overall reduction in participants' technical effort in PE.

Both the mean fixation duration and fixation counts in HT are larger than in PE, suggesting that HT is more cognitively demanding than PE on a global level. When analysing the eye-tracking data in ST area and TT area separately, it was found that the translators expanded higher level of cognitive effort in the ST area in HT tasks than PE tasks, indicating that the PE workflow can effectively save the cognitive resources invested in the reading comprehension of ST, which is in line with previous research in non-literary PE processes (Carl *et al.*, 2011; Daems *et al.*, 2017b; Nitzke & Oster, 2016).

As for the TT area, it was found that the mean fixation duration was longer in HT tasks than in PE, while only a marginally significant difference across task modalities was found in terms of fixation counts, with more fixations during HT tasks. Our findings suggest that processing target texts in HT tasks is also more cognitively effortful than in PE tasks. Comparing our findings with previous works, Daems *et al.* (2017b) documented more fixation counts and shorter mean fixation duration on the target-text area for trainees in PE tasks. Carl *et al.* (2011) reported a significantly higher level of total gaze time and fixation counts on the TT area in PE tasks than in HT tasks. In Carl *et al.*'s (2011) study, total gaze time was defined as "the combined duration of fixations" (Hvelplund, 2011, p. 21). We could therefore estimate the value of mean fixation duration in Carl *et al.* (2011), and found that in the TT area, the mean value of MFD in HT tasks (376.41ms) is longer than in PE tasks (253.57ms) in that study. It seems that the longer total gaze time on the TT area in PE tasks in that study consists of a combination of a large number of short fixations. That is to say, our results are consistent with works in L2-L1 non-literary translation in terms of mean fixation duration but are different concerning fixation counts.

The explanation could be manifold. We drew one possible cause from the screen record: in HT tasks, many of our participants went through the drafted target text for many times in the revision and monitoring phase, while this tendency is not much observed in PE tasks. In PE tasks, the revision phase majorly

consists of consecutively checking the TT sentence with the corresponding ST sentence. It is still unclear why there is a behavioural difference in the revision phase, but it is possible that they were too engaged in checking the semantic accuracy of PE products to pay attention to the coherence on a textual level, or the participants were constrained by the structure of raw MT output, as reflected in translators' subjective reports in previous works (Guerberof-Arenas & Toral, 2022; Moorkens *et al.*, 2018), and gave up the attempt to make extra changes in the last phase.

Considering the distribution of visual resources, it was found that the TT area received more fixations and longer mean fixation durations in both HT and PE tasks than the ST area, which means that processing the target text was more effortful than source text for both HT and PE workflows. We also found that the tendency that the TT area received more fixations is more prominent in PE tasks. These are mostly in line with the literature which discussed post-editing non-literary texts using SMT (Carl *et al.*, 2011; Daems *et al.*, 2017b). The results indicate that in the HT mode, producing and revising the target text is more cognitively consuming than reading the source text for translation. On the other hand, in the PE mode, reading and revising raw MT output is also more effortful than the reading comprehension of the ST. When post-editing non-literary texts, as Carl *et al.* (2011) assumed, the participants possibly only consulted the ST very briefly to check the accuracy and adequacy of the MT output, which made most of the participants' cognitive resources allocated to the TT area. Our findings seem to provide evidence that the role ST plays in literary PE is somewhat similar to that in non-literary PE.

6 Conclusion and Future Works

This study sets out to discover the potential of applying machine translation post-editing to L1–L2 literary translation. An overall reduction in efforts was discovered when adopting the workflow of PE. The findings in the present study brought up possibilities to introduce NMT systems as an effective aid in L1–L2 literary translation, especially for novice translators.

The conclusions drawn from the present work have certain limitations. The first major limitation is that our participants were all postgraduate translation students. Currently, it is unknown whether the results obtained in the present study can be generalized to professional literary translators, which surely calls for more studies with suitable-sized professional translators preferably holding positive opinions on technology use.

Secondly, in the present study, we only selected novels as the genre of our research materials. However, there are considerable distinctions between different genres of literary works. The results obtained in the present study should be cautiously treated when trying to expand to other kinds of literary translation tasks. Also, as Toral and Way (2018, p. 274) put it, novels are "far from a monolithic domain" when discussing literary MT. We admit that the two source texts we selected may not cover all kinds of contemporary Chinese novels, and more nuanced categorizations could be used in future works.

References

Bartoń, Kamil. (2009). MuMIn: multi-model inference [R Package]. Retrieved from https://CRAN.R-project.org/package=MuMIn

Bates, Douglas, Mächler, Martin, Bolker, Ben, & Walker, Steve. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software, 67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01.

Carl, Michael. (2012). *Translog-II: a Program for Recording User Activity Data for Empirical Translation Process Research*. Paper presented at the The Eighth International Conference on Language Resources and Evaluation, Istanbul, Turkey.

Carl, Michael, Dragsted, Barbara, Elming, Jakob, Hardt, Daniel, & Lykke Jakobsen, Arnt. (2011). The Process of Post-Editing: A Pilot Study. *Copenhagen Studies in Language, 41*, 131–142.

Carl, Michael, Gutermuth, Silke, & Hansen-Schirra, Silvia. (2015). Post-editing machine translation: A usability test for professional translation settings. In Aline Ferreira & John W. Schwieter (Eds.), *Psycholinguistic and Cognitive Inquiries into Translation and Interpreting* (pp. 145–174). Amsterdam: John Benjamins.

Chang, Vincent Chieh-Ying. (2011). Translation Directionality and the Revised Hierarchical Model: An Eye-Tracking Study. In Sharon O'Brien (Ed.), *Cognitive explorations of translation* (pp. 154–174). London: Continuum.

Cui, Yixiao, & Zheng, Binghan. (2021). Consultation behaviour with online resources in English-Chinese translation: an eye-tracking, screen-recording and retrospective study. *Perspectives, 29*(5), 740–760. https://doi.org/10.1080/0907676X.2020.1760899.

da Silva, Igor, Alves, Fabio, Schmaltz, Marcia, Pagano, Adriana, Wong, Derek, Chao, Lidia, Leal, Ana, Quaresma, Paulo, & Silva, Gabriel. (2017). Translation, Post-Editing and Directionality: A Study of Effort in the Chinese-Portuguese Language Pair. In Arnt Lykke Jakobsen & Bartolomé Mesa-Lao (Eds.), *Translation in Transition: Between cognition, computing and technology* (pp. 107–134). Amsterdam: John Benjamins.

Daems, Joke, Vandepitte, Sonia, Hartsuiker, Robert J., & Macken, Lieve. (2017a). Identifying the Machine Translation Error Types with the Greatest Impact

on Post-editing Effort. *Frontiers in Psychology, 8*, 1282. https://doi.org/10.3389/fpsyg.2017.01282.

Daems, Joke, Vandepitte, Sonia, Hartsuiker, Robert J., & Macken, Lieve. (2017b). Translation Methods and Experience: A Comparative Analysis of Human Translation and Post-editing with Students and Professional Translators. *Meta, 62*(2), 245–270. https://doi.org/10.7202/1041023ar.

Garcia, Ignacio. (2011). Translating by post-editing: is it the way forward? *Machine Translation, 25*(3), 217–237. https://doi.org/10.1007/s10590-011-9115-8.

Guerberof-Arenas, Ana. (2014). Correlations between productivity and quality when post-editing in a professional context. *Machine Translation, 28*(3), 165–186. https://doi.org/10.1007/s10590-014-9155-y.

Guerberof-Arenas, Ana, & Toral, Antonio. (2020). The impact of post-editing and machine translation on creativity and reading experience. *Translation Spaces, 9*(2), 255–282. https://doi.org/10.1075/ts.20035.gue.

Guerberof-Arenas, Ana, & Toral, Antonio. (2022). Creati vity in translation: Machine translation as a const raint for literary texts. *Translation Spaces, 11*(2), 184–212. https://doi.org/10.1075/ts.21025.gue.

Guerberof-Arenas, Ana, & Toral, Antonio. (2023). To be or not to be: a translation reception study of a literary text translated into Dutch and Catalan using machine translation. *arXiv preprint*. https://doi.org/10.48550/arXiv.2307.02358.

Hadley, James Luke, Taivalkoski-Shilov, Kristiina, Teixeira, Carlos S. C., & Toral, Antonio. (2022a). Introduction. In James Luke Hadley, Kristiina Taivalkoski-Shilov, Carlos S. C. Teixeira, & Antonio Toral (Eds.), *Using Technologies for Creative-Text Translation* (pp. 1–17). New York: Routledge.

Hadley, James Luke, Taivalkoski-Shilov, Kristiina, Teixeira, Carlos S. C., & Toral, Antonio (Eds.). (2022b). *Using Technologies for Creative-Text Translation*. New York: Routledge.

Horcas-Rufián, Sara. (2022). L2 translation practice in Spain: Report on a survey of professional translators. *Translation & Interpreting, 14*(1), 121–141. https://doi.org/10.12807/ti.114201.2022.a07.

Hvelplund, Kristian Tangsgaard. (2011). *Allocation of Cognitive Resources in Translation: An Eye-tracking and Key-logging Study*. Frederiksberg: Samfundslitteratur.

Hvelplund, Kristian Tangsgaard. (2014). Eye tracking and the translation process: reflections on the analysis and interpretation of eye-tracking data. *MonTI. Monographs in translation and interpreting*, 201–223. https://doi.org/10.6035/MonTI.2014.ne1.6.

Jia, Yanfang, Carl, Michael, & Wang, Xiangling. (2019a). How does the post-editing of neural machine translation compare with from-scratch translation? A product and process study. *The Journal of Specialised Translation*(31), 60–86. Retrieved from https://jostrans.soap2.ch/issue31/art_jia.pdf

Jia, Yanfang, Carl, Michael, & Wang, Xiangling. (2019b). Post-editing neural machine translation versus phrase-based machine translation for English–Chinese. *Machine Translation, 33*(1), 9–29. https://doi.org/10.1007/s10590-019-09229-6.

Kenny, Dorothy, & Winters, Marion. (2020). Machine translation, ethics and the literary translator's voice. *Translation Spaces, 9*(1), 123–149. https://doi.org/10.1075/ts.00024.ken.

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*(1), 126–141.

Krings, H.P. (2001). *Repairing Texts: Empirical Investigations of Machine Translation Post-editing Processes* (G.S. Koby, Trans.). Kent, OH: Kent State University Press.

Kuznetsova, Alexandra, Brockhoff, Per B., & Christensen, Rune H. B. (2017). lmerTest Package: Tests in Linear Mixed Effects Models. *Journal of Statistical Software, 82*(13), 1–26.

Lacruz, Isabel. (2017). Cognitive Effort in Translation, Editing, and Post-editing. In John W. Schwieter & Aline Ferreira (Eds.), *The Handbook of Translation and Cognition* (pp. 386–401). Hoboken: Wiley.

Moorkens, Joss. (2018). Eye tracking as a measure of cognitive effort for post-editing of machine translation. In Callum Walker & Federico M. Federici (Eds.), *Eye Tracking and Multidisciplinary Studies on Translation* (pp. 55–69). Amsterdam: John Benjamins.

Moorkens, Joss, Toral, Antonio, Castilho, Sheila, & Way, Andy. (2018). Translators' perceptions of literary post-editing using statistical and neural machine translation. *Translation Spaces, 7*, 240–262. https://doi.org/10.1075/ts.18014.moo.

Nitzke, Jean. (2018). *Problem solving activities in post-editing and translation from scratch: A multi-method study*. Berlin: Language Science Press.

Nitzke, Jean, & Hansen-Schirra, Silvia. (2021). *A short guide to post-editing*. Berlin: Language Science Press.

Nitzke, Jean, & Oster, Katharina. (2016). Comparing Translation and Post-editing: An Annotation Schema for Activity Units. In Michael Carl, Srinivas Bangalore, & Moritz Schaeffer (Eds.), *New Directions in Empirical Translation Process Research: Exploring the CRITT TPR-DB* (pp. 293–308). Cham: Springer.

Pavlović, Nataša, & Jensen, Kristian T H. (2009). Eye tracking translation directionality. In Anthony Pym & Alexander Perekrestenko (Eds.), *Translation research projects 2* (pp. 93–109). Tarragona: Intercultural Studies Group.

Plitt, Mirko, & Masselot, François. (2010). A Productivity Test of Statistical Machine Translation Post-Editing in a Typical Localisation Context. *The Prague Bulletin of Mathematical Linguistics, 93*(1), 7–16. https://doi.org/10.2478/v10108-010-0010-x.

Pokorn, Nike K. (2005). *Challenging the Traditional Axioms: Translation into a non-mother tongue*. Amsterdam: John Benjamins.

R Core Team. (2024). R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.R-project.org

Samuelsson-Brown, Geoffrey. (2010). *A Practical Guide for Translators* (5th ed.). Bristol: Multilingual Matters.

Sun, Juan, Lu, Zhi, Lacruz, Isabel, Ma, Lijun, Fan, Lin, Huang, Xiuhua, & Zhou, Bo. (2023). An eye-tracking study of productivity and effort in Chinese-to-English translation and post-editing. In Isabel Lacruz (Ed.), *Translation in Transition : Human and Machine Intelligence* (pp. 57–82). Amsterdam: John Benjamins.

Taivalkoski-Shilov, Kristiina. (2019). Ethical issues regarding machine(-assisted) translation of literary texts. *Perspectives, 27*(5), 689–703. https://doi.org/10.1080/0907676X.2018.1520907.

Toral, Antonio, & Way, Andy. (2018). What Level of Quality Can Neural Machine Translation Attain on Literary Text? In Joss Moorkens, Sheila Castilho, Federico Gaspari, & Stephen Doherty (Eds.), *Translation Quality Assessment: From Principles to Practice* (pp. 263–287). Cham: Springer.

Toral, Antonio, Wieling, Martijn, & Way, Andy. (2018). Post-editing Effort of a Novel With Statistical and Neural Machine Translation. *Frontiers in Digital Humanities, 5*, 9. https://doi.org/10.3389/fdigh.2018.00009.

Vieira, Lucas Nunes. (2017). How do measures of cognitive effort relate to each other? A multivariate analysis of post-editing process data. *Machine Translation, 30*(1), 41–62. https://doi.org/10.1007/s10590-016-9188-5.

Vieira, Lucas Nunes, Zelenka, Natalie, Youdale, Roy, Zhang, Xiaochun, & Carl, Michael. (2023). Translating science fiction in a CAT tool: machine translation and segmentation settings. *The International Journal of Translation and Interpreting Research, 15*, 216–235. https://doi.org/10.12807/ti.115201.2023.a11.

Winter, Bodo. (2019). *Statistics for Linguists: An Introduction Using R*. New York: Routledge.

Wu, Yun. (2021). *改革开放以来中国当代小说英译研究* [A Study on English Translations of Chinese Novels since the Reform and Opening-up]. Hangzhou: Zhejiang University Press.

Yang, Yanxia, Wang, Xiangling, & Yuan, Qingqing. (2021). Measuring the usability of machine translation in the classroom context. *Translation and Interpreting Studies, 16*(1), 101–123. https://doi.org/10.1075/tis.18047.yan.