The interaction effect between source text complexity and machine translation quality on the task difficulty of NMT post-editing from English to Chinese: A multi-method study

Yanfang Jia
Hunan Normal University
yanfang.jia@hotmail.com
https://orcid.org/0000-0001-5111-3185

Binghan Zheng
Durham University
binghan.zheng@durham.ac.uk
https://orcid.org/0000-0001-5302-4709
(corresponding author)

Abstract

This study explores the interaction effect between source text (ST) complexity and machine translation (MT) quality on the task difficulty of neural machine translation (NMT) post-editing from English to Chinese. When investigating human effort exerted in post-editing, existing studies have seldom taken both ST complexity and MT quality levels into account, and have mainly focused on MT systems used before the emergence of NMT. Drawing on process and product data of post-editing from 60 trainee translators, this study adopted a multi-method approach to measure post-editing task difficulty, including eye-tracking, keystroke logging, quality evaluation, subjective rating, and retrospective written protocols. The results show that: 1) ST complexity and MT quality present a significant interaction effect on task difficulty of NMT post-editing; 2) ST complexity level has a positive impact on post-editing low-quality NMT (i.e., post-editing task becomes less difficult when ST complexity decreases); while for post-editing high-quality NMT, it has a positive impact only on the subjective ratings received from participants; and 3) NMT quality has a negative impact on its post-editing task difficulty (i.e., post-editing task becomes less difficult when MT quality goes higher), and this impact is stronger when ST complexity increases. This paper concludes that both ST complexity and MT quality should be considered when testing post-editing difficulty, designing tasks for post-

editor training, and setting fair post-editing pricing schemes.

Keywords: source text complexity, machine translation quality, post-editing, task difficulty, multi-method approach

1. Introduction

Due to the advancement and application of machine translation (MT) technology, MT post-editing (PE) has now been provided as an independent service in today's translation market with its own international service standard (ISO 2017: 18587). It is also the most widely adopted set-up nowadays in the professional context in the translation industry (TAUS, 2019). As a relatively new task mode, the potential value and cognitive process of PE are still largely under-investigated but have gained increasing attention both from academia and industry. The recently emerged paradigm of Neural Machine Translation (NMT) has greatly advanced MT quality, especially in the aspects of fluency or readability of translation output, when compared to the once-dominant Statistical Machine Translation (SMT) (Sennrich et al., 2016; Junczys-Dowmunt et al., 2016). However, recent studies show that NMT also brings new challenges to post-editors by producing unpredictable errors hidden in its fluent texts, which make it more difficult to identify and correct translation errors during PE (Yamada, 2019; Vieira, 2019).

Investigating the factors impacting the task difficulty of PE and its measurements is important for testing post-editing difficulty, designing tasks for post-editor training, and setting reasonable post-editing pricing schemes. Among such factors, source text (ST) complexity and MT quality are usually regarded as major intrinsic factors contributing to the task difficulty of PE. However, previous studies (e.g., Krings, 2001; O'Brien, 2006; Daems et al., 2017; Castilho et al., 2018) have rarely taken both factors into account when investigating the human effort exerted during PE, making it difficult to disentangle the role each factor plays in PE. Besides this, these studies have focused predominately on PE of MT approaches before NMT between Indo-European languages, leaving PE of NMT between English and Chinese under-researched.

This study explores the impact of ST complexity and NMT quality on the task difficulty of NMT post-editing from English to Chinese by adopting a multi-method approach (Halversen, 2017), including data collected from eye-tracking, keystroke logging, subjective rating, retrospective protocols, and translation quality evaluation. We aim to address the following two questions: (1) Do NMT quality and ST complexity have an interaction effect on the task difficulty of NMT PE? and (2) If Yes, how do they affect the impact of each other on the task difficulty of PE?

2. Task difficulty of PE

From the cognitive perspective, task difficulty is a concept specific to a task and a person (Dahl, 2004:39), referring to “the degree of cognitive load, or mental effort, required to identify a problem solution” (Gallupe et al., 1988: 280). Measuring task difficulty, therefore, concerns whether a task is easy or difficult for a person performing the task, which is inherently subjective and personal. Cognitive load in the present study refers to the demand on cognitive resources that a PE task imposes upon a post-editor, whereas cognitive effort is the actual amount of cognitive resources that a post-editor used to finish the PE task. Following Sun (2015), “task difficulty” is used as a common and cover term, and will be investigated with respect to the following two aspects: identifying the potential causal factors of PE task difficulty, and measuring its task difficulty.

2.1. Causal factors of PE task difficulty

Cognitive load theory (CLT; Sweller, 1988) is adopted in the present study as a theoretical foundation to explain the causal factors of PE difficulty. According to Paas and Van Merriënboer (1994: 353), cognitive load is “a multidimensional construct representing the load that performing a particular task imposes on the cognitive system of an individual”，and can be divided into intrinsic cognitive load, extraneous cognitive load, and germane cognitive load. The intrinsic cognitive load is immutable and originates from the difficulty level imposed by the inherent nature of the material or task and the expertise of the individual performing the task. The extraneous cognitive load is not constant and should ideally be reduced by improving the usability of the tools or optimizing the way the information is presented. Intrinsic and extraneous cognitive loads add up to determine the total amount of cognitive load imposed by the task to be completed, while germane cognitive load refers to the cognitive resources devoted to learning for schema construction (Sweller et al., 2011).

As a problem solving rather than a learning process, a PE task mainly includes intrinsic and extraneous cognitive loads. The intrinsic cognitive load for PE is determined by the efforts needed to process the ST and the MT output, and the post-editor’s expertise. For manual translation, the intrinsic cognitive load of translation difficulty is primarily decided by ST complexity (Liu et al., 2019); but for PE, post-editors are offered two sources of information (i.e., ST and MT) with different functions. As long as the MT output is not so poor that the post-editor decides to translate everything from scratch, or not so good that the post-editor could accept unedited raw MT, it is safe to say that a PE process entails evaluation of MT output, correction of MT errors, and translation from scratch at different levels. The extraneous cognitive load in PE is caused by external factors such as the user interface and working environment where a PE task is performed, which is gaining increasing interest in usability and ergonomics research (Kappus & Ehrensberger-Dow, 2020).

2.2. ST complexity and PE effort

Previous research on the association between ST and PE effort for Rule-based MT (RBMT) and Example-based MT (EBMT) systems shows that ST with more Negative Translation Indicators (NTIs) (e.g., ambiguity, coordination, ellipsis and gerunds) will result in more cognitive effort, a higher number of edits (Aikawa et al., 2007), and longer time on the task (O'Brien, 2004， 2006). However, these studies have not controlled for the corresponding MT quality for the ST used. As an ST with more NTIs can easily lead to lower MT quality, what those studies examined was actually the difference in cognitive effort when post-editing an ST with more NTIs paired with a lower-quality MT, versus an ST with fewer NTIs coupled with a higher-quality MT. Some other studies investigating the association between ST features and PE effort indicators have not controlled for the quality of MT outputs either. As Aziz et al. (2014) reflected, the PE effort found to be associated with the specific linguistic patterns of ST may be caused by the low-quality MT output of these ST features.

Eye-tracking studies demonstrate differences in how cognitive resources are allocated to ST and target text (TT) during PE tasks. Most of these studies have shown that less visual attention (e.g., total fixation duration) is allocated to the ST area (Carl et al., 2011; Daems et al., 2017). For example, Carl et al. (2011) found that total fixation duration on TT was much longer than on ST during PE. In Mesa-Lao (2014), however, mixed results are reported with more fixations on TT for 4 out of 6 PE tasks. As Mesa-Lao only mentions that some STs were not of similar complexity levels, without providing detailed information on ST complexity and corresponding MT quality levels, it is impossible to further interpret his findings.

2.3. MT quality and PE effort

MT quality has commonly been perceived as the key factor deciding PE effort; however, the argument has not been unanimously borne out by empirical studies. The results of previous studies devoted to the impact of MT have been reported with some inconsistencies. Krings (2001) employed human raters to evaluate RBMT output sentences using a five-point Likert scale. He found that RBMT post-editing speed was faster for higher-rated segments. Krings also observed that the correlation between MT quality and PE effort in terms of attention distribution was not always linear. PE effort was the highest, in many cases, for the medium-quality sentences rather than for the lowest quality ones, with more and greater dispersion of attention distributed across the ST, MT and TT for the medium-quality ones. In addition, he also found that the level of MT quality seemed to have an impact on how attention and effort were allocated to ST processing. However, the ST complexity levels were not controlled for by Krings, making it unclear whether ST complexity itself has affected how the ST is processed.

Other studies have found that MT quality, measured by different automatic evaluation metrics, tends to be negatively correlated with PE effort. For example, lower MT quality as measured by GTM (General Text Matcher) and TER (Translation Edit Rate) in O'Brien (2011), and by BLEU (Bilingual Evaluation Understudy), METEOR (Metric for Evaluation of Translation with Explicit Ordering) and TER in Gaspari et al. (2014), were found to be associated with longer PE time and longer total fixation duration, all of which suggest greater cognitive effort invested by post-editors. Sanchez-Torron & Koehn (2016) assessed how MT quality indicated by BLEU can impact PE effort by using ST of similar complexity. They found that PE for MT output with higher BLEU scores led to better final product quality and reduced the overall PE effort in terms of PE time and operations. In Vieira (2016), higher MT quality as indicated by higher METEOR scores correlates with post-editors' lower cognitive effort as indicated by lower average fixation duration. In addition, MT quality in terms of the number and types of errors was also found to be associated with PE effort exerted. Daems et al. (2017) shows that the overall MT errors were negatively associated with fixation count, number of production units and the HTER (Human-targeted Translation Edit Rate) score, and positively correlated with average pause ratio. They also report that different error types affect different PE effort indicators.

All the above studies and those studies reviewed in Temizöz (2012) take MT quality as the primary factor impacting on PE effort, without controlling the corresponding ST complexity levels. However, post-editing MT with the same errors or automatic evaluation scores may involve different levels of effort, when paired with STs with different complexity levels. In addition, most of these studies investigated MT systems before NMT became the dominant paradigm; therefore, the results may not be replicable to research on PE of NMT.

3. Research design

3.1. Participants

Sixty MA students in Translation (two males and fifty-eight females) with an average age of 24 years (range=22-26, *SD*=1.9 years) were recruited as participants from two Chinese universities. They were all native Chinese speakers with English as their L2. All participants had similar levels of L2 proficiency, passed the Test for English Majors Band 8 (TEM8)[1], but had no professional translation experience. They all had roughly the same level of experience in using MT systems as additional resources during translation, but had never received formal training in PE. To compensate for their work, participants received two academic credits from their university for taking part in a PE training session prior to the experiment, and were given a memory disk as the reward for their participation. All participants were touch-typists and had normal or corrected to normal vision. They were told that anonymity and confidentiality would be ensured, and they all signed a consent form before each experiment. The research was approved by the Ethics Committee of Hunan University.

3.2. Materials

3.2.1. ST selection

Four English news texts[2] (128-145 words, coded as ST_1, ST_2, ST_3 and ST_4), two with high complexity and two with low complexity, were carefully selected as STs for this research. ST_1, ST_2 and ST_4 were selected from *newsela.com*, a website providing newspaper articles at different levels of complexity, and ST_3 from *the Times,* a British daily national newspaper. Featuring news topics for general readers, the four texts are self-contained, requiring no additional context for comprehension and translation.

Four sets of measurements, comprising readability level, word frequency, syntactic complexity, and subjective evaluation, were adopted to measure the ST complexity. As can be seen from Figure 1, in terms of readability indexes, ST_1 and ST_2 are appropriate for 7 and 8 years of schooling respectively, while ST_3 and ST_4 are appropriate for 18 and 16 years of schooling for successful comprehension respectively. Flesch Reading Ease scores show that ST_1 and ST_2 are much easier to read than ST_3 and ST_4. Word frequency tests indicate that ST_1

and ST_2 contain a smaller proportion of low frequency words than ST_3 and ST_4. Sentence syntax similarity values as measured by the Coh-Metrix automatic text analysis tool (version 3.0) indicate that ST_1 (0.165/0.239) and ST_2 (0.168/0.144) present lower complexity than ST_3 (0.057/0.044) and ST_4 (0.022/0.015). Nine freelance translators were recruited to rate the levels of translation difficulty on a nine-point Likert-type scale, with 1 being "extremely easy" and 9 "extremely difficult". The results show that ST_1 and ST_2 were rated to be easier for translation than ST_3 and ST_4. In summary, ST_1 and ST_2 are tested as less complex and less difficult texts for translation than ST_3 and ST_4.

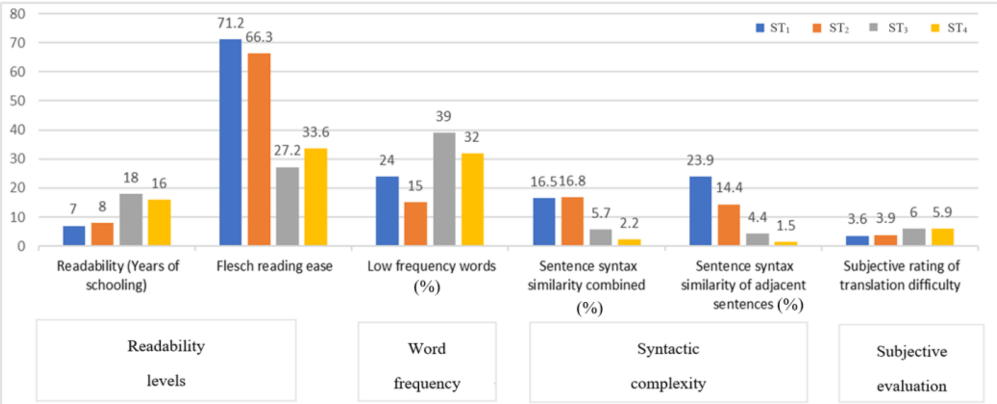

Figure 1. Summary of ST complexity from four sets of measurements

3.2.2. MT output selection

The four STs were firstly pre-translated by five online NMT engines: Google Translate, Baidu Translate, Bing Translate, Systran, and Youdao Translate. The MT outputs were then assessed by TAUS's (2013) fluency and adequacy criteria on a 4-point Likert scale, with "1" being incomprehensible and "4" being flawless for fluency, and "1" being extremely inadequate and "4" being fully adequate for adequacy. The nine freelance translators recruited to assess the MT outputs using this scale were not participants for the main experiment.

Based on the evaluation results, the outputs of Google Translate and Systran were kept for the PE experiments (available upon request). Kendall's W was used to measure the level of agreement among the nine raters. For both fluency and adequacy of the above two NMT outputs (see Table 1), the responses for Kendall's W fall between 0.71-0.90, $p<.001$, indicating a significant, strong agreement among raters. All evaluators rated the quality of MT output from Google to be higher than that from Systran in both fluency and adequacy for all four texts, with all the differences in the average scores being statistically significant.

Table 1. The inter-rater agreement and mean scores for Google and Systran MT outputs

MT Adequacy	**Text**	**Mean**	**Sd.**	**Min**	**Max**	**Kendall's W**	**Chi-Square**	**Sig.**
Google	ST_1	3.33***	0.37	2.57	3.82	0.765	422	*p*<.001
Systran		1.94	0.44	1.19	2.58			
Googles	ST_2	3.36***	0.4	2.57	3.89			
Systran		2.46	0.53	1.62	3.29			
Google	ST_3	3.32***	0.43	2.62	3.68			
Systran		1.65	0.15	1.46	1.82			
Google	ST_4	3.36*	0.33	3.08	3.92			
Systran		2.47	0.79	1.77	3.77			
MT Fluency	Text	Mean	Sd.	Min	Max	Kendall's W	Chi-Square	Sig.
Google	ST_1	3.11***	0.44	2.52	3.69	0.799	453	*p*<.001
Systran		1.79	0.49	1.19	2.51			
Google	ST_2	3.37***	0.48	2.23	3.39			
Systran		2.28	0.54	1.55	3.21			
Google	ST_3	3.3***	0.38	2.98	3.79			
Systran		1.51	0.23	1.21	1.78			
Google	ST_4	3.33*	0.38	3.02	3.88			
Systran		2.39	0.86	1.78	3.88			

3.3. Experiment settings

The eye movements of the PE processing for all participants were recorded by an Eyelink 1000 plus (1000Hz) eye-tracker, connected to a 23-inch LCD monitor as the presentation screen. The screen resolution was set at 1280*1024 pixels. A nine-point calibration was applied to guarantee the precision of the gaze data. The English STs were displayed in the upper window of Translog-II, with Times New Roman Typeface set at 16 points, and double line spacing. The Chinese MT output and final target texts were displayed in the lower window, with SimSun Typeface set at 16 points, also with double line spacing.

3.4. Experiment procedure

Each participant carried out two PE tasks：post-editing one MT output with high quality (coded as MT_H), and the other MT output with low quality (coded as MT_L). The two PE tasks are from different STs to reduce potential learning effect. The order of the four STs and the sequence of the two PE tasks were balanced across the sixty participants in a Latin square design. There was no time constraint on all tasks.

Each participant filled in a pre-task questionnaire concerning their educational and language backgrounds, and their attitudes towards MT and PE. They first carried out a warm-up task and were instructed to post-edit the assigned MT outputs and deliver final products of publishable quality. To eliminate the impact of background knowledge on the task difficulty of PE, a piece of short English news briefing the background of each ST was provided for participants before each task. Right after finishing each task, participants were asked to rate the task difficulty subjectively; and after finishing their tasks, participants were asked to comment in writing regarding the problems and difficulties they had come across during the PE tasks. Participants could choose to take a ten-minute break between the two tasks. The experiment procedure is shown in Figure 2, with the complete session for each participant lasting roughly two hours.

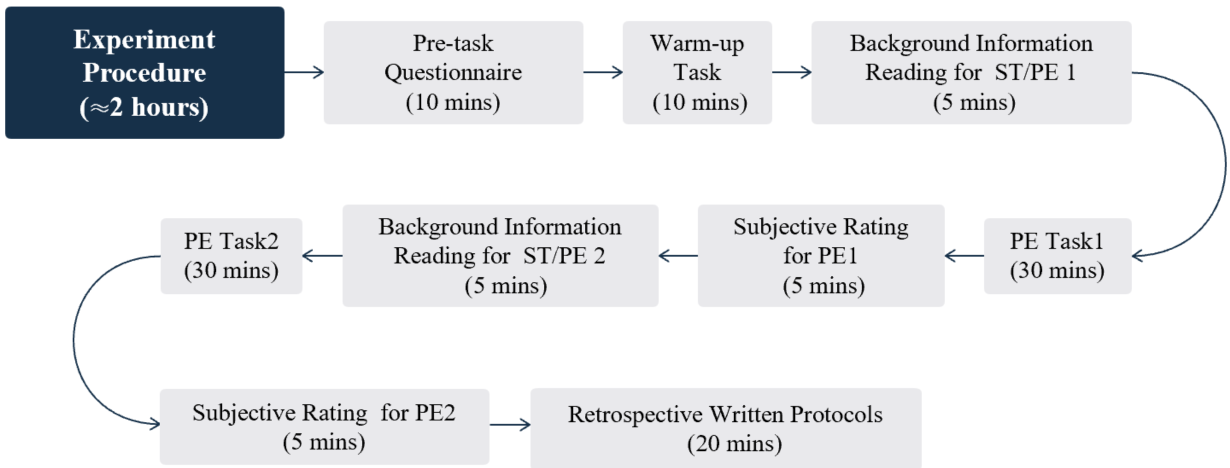

Figure 2. The flow chart of experiment procedure

3.5. Quality of the eye-tracking and key-logging data

The quality of eye-tracking and key-logging data collected from the sixty participants was assessed prior to the data analysis. For the eye-tracking data, gaze data with the average fixation duration (AFD) below 200ms and the ratio of the total gaze time on the screen (GTS) divided by the total task time considerably below sample mean (1.5 SD below sample mean), were eliminated. The samples left were all with GTS above 30% (cf. Hvelpund, 2011; Vieira, 2016). In addition, two abnormal eye-tracking sessions and two corrupted key-logging sessions were excluded; and one session from ST_2 was randomly removed in order to balance the final data points for each task. As a result, 96 valid PE sessions across four STs pre-translated by two NMT engines were selected for further analysis (see Table 2). The percentage of valid data is 80%.

Table 2. Valid data sets left for final analysis ("x" represents the data points being excluded)

NMT engine	**Google translate (high quality)**				**Systran translate (low quality)**			
ST	Low		High		Low		High	
Complexity	ST_1	ST_2	ST_3	ST_4	ST_1	ST_2	ST_3	ST_4
AFD	xx	x	x		xx		x	x
GTS		x	xx	xx		xx	x	x
AFD+GTS	x			x				
Corrupted logging data					x			x
Abnormal data						x	x	
Randomly excluded		x						
Final data points	12	12	12	12	12	12	12	12
	24		24		24		24	

3.6. Data preparation and statistical data analysis

The statistical analysis was conducted using the Linear Mixed Effects Regression (LMER) models provided in the lme4 package (Bates et al., 2014) of the statistical software R (version 3.6.3). The standard errors, effect sizes and significance values were calculated by the software package lmerTest (Kuznetsova et al., 2016). The effects of the models were plotted, by applying *effects* package (Fox et al., 2017). As fixed effects, the main effects of NMT quality levels (MT_H for MT with high-quality, and MT_L for MT with low-quality), ST complexity levels (ST_H for ST with high-complexity, and ST_L for ST with low-complexity) and their interaction were entered into the model. The random effect was the participants.

In the LMER models, the dependent variables investigated were: (1) Subjective rating scores; (2) Processing time; (3) Total fixation duration on ST; (4) Total fixation duration on TT; (5) Pause to word ratio; (6) Total number of editing operations; and (7) Total number of errors.

To eliminate the potential effects of ST length, the following dependent variables were normalized by number of tokens in ST, comprising: processing time, total fixation duration on ST, total fixation duration on TT, total number of editing operations and total number of errors. Subjective rating and pause to word ratio were not normalized by ST tokens, because subjective rating was based on the task difficulty of the whole task and pause to word ratio had already taken the ST length into account.

We applied Skewness and Kurtosis tests to verify that all dependent variables were normally distributed, and checked the residual plots to ensure that the homogeneity of variance for each model was not violated. Square-root or log-transformation was used to transform those variables with Skewness or Kurtosis greater than 1 or smaller than -1, depending on which method produced better normal distribution. To analyze the errors in PE output, the customized error categories of the Multidimensional Quality Metrics framework (MQM, Lommel, 2018) were adopted. The analysis was carried out by two College English teachers who had over 10 years of experience in rating English–Chinese translation examinations. The present study focuses mainly on the overall quality of the final post-edited product, thus total number of errors was calculated as an indicator of the overall PE quality. Unless otherwise stated, MT hereafter refers to NMT.

4. Results

4.1. Subjective rating

The subjective rating presents the participants' subjective perception towards the task difficulty after finishing each PE task, with "1" being the least and "9" the most difficult. The interaction effect between MT quality and ST complexity on subjective rating is plotted in Figure 3 and shows no significance (p>.05). ST complexity showed a consistent, positive impact[3] on PE for both MT_H and MT_L. PE-MT_L for ST_H (6.29) was taken to be significantly more difficult than for ST_L (5.21) (t=2.82, p<.01). Similarly, PE-MT_H for ST_H (4.75) was considered to be slightly more difficult than for ST_L (4.33), but the difference was not significant (t=1.09, p>.05). MT quality had a negative impact on the subjective rating for PE difficulty for both ST_H and ST_L. For ST_H, PE-MT_L (6.29) was perceived to be significantly more difficult than PE-MT_H (4.75) (t=4.01, p<.001); for ST_L, PE-MT_L (5.21) was rated more difficult than PE-MT_H (4.33) (t=2.33, p=.058).

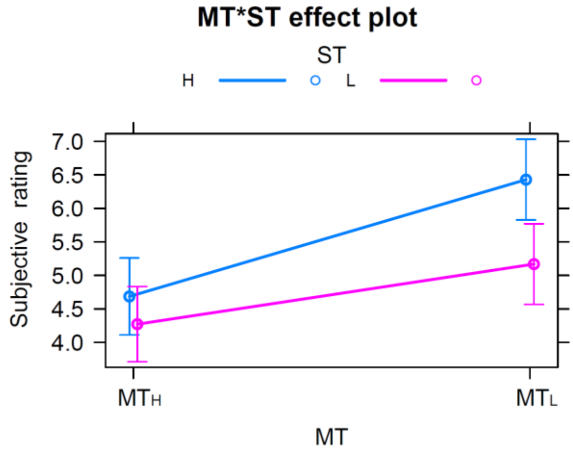

Figure 3. Interaction effect between MT quality and ST complexity on subjective rating

4.2. Processing time

Processing time is the time taken to finish each PE task in millisecond (ms). The longer time it takes, the more cognitive effort is expected to be exerted. As shown in Figure 4, regarding processing time, the interaction effect between MT quality and ST complexity was significant (p<.001). ST complexity has a positive impact on PE-MT_L. PE-MT_L for ST_L (6984 ms) was significantly faster than for ST_H (8846 ms) (t=-2.7, p<.01). For PE-MT_H, the impact of ST complexity was negative. PE-MT_H for ST_H (4950 ms) was faster than ST_L (6287 ms), but the difference was not statistically significant (t=-1.9, p=.052). MT quality demonstrated a negative impact on processing time and this impact was significant only for ST_H, with PE-MT_H (4950 ms) being significantly faster than PE-MT_L (8846 ms) (t=-5.7, p<.001). For ST_L, PE-MT_H (6287 ms) took slightly less time than PE-MT_L (6984 ms) (t=-1, p>.05).

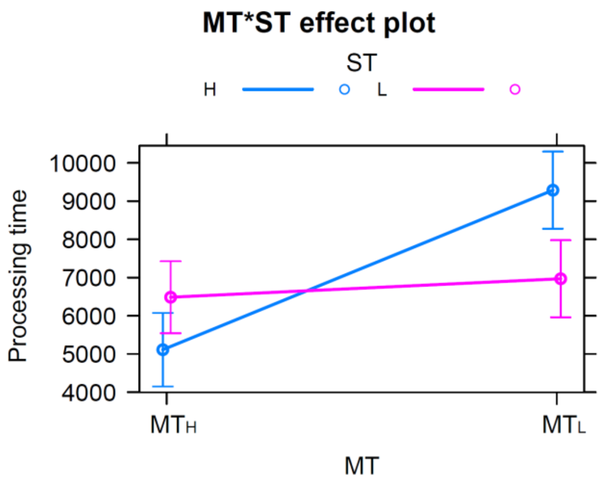

Figure 4. Interaction effect between MT quality and ST complexity on processing time

4.3. Pause to word ratio

Pause to word ratio is calculated by dividing the total number of pauses in a task by the number of tokens in the ST. Higher pause to word ratio indicates more cognitive effort exerted (Lacruz and Shreve, 2014). As plotted in Figure 5, the interaction effect between MT quality and ST complexity on pause to word ratio was significant (p<.05). Pause to word ratio during PE-MT_L for ST_H and for ST_L was almost the same (t=0.09, p=.93). PE-MT_H for ST_H led to significantly lower pause to word ratio than for ST_L (t=3.37, p<.05). MT quality had a consistent significant negative impact on pause to word ratio during PE and this effect was stronger for ST_H. For STs of higher complexity, PE-MT_H resulted in significantly lower pause to word ratio than PE-MT_L (t=-7.75, p<.001). For ST_L, PE-MT_H again had the significantly lower pause to word ratio than PE-MT_L (t=-4.48, p<.001).

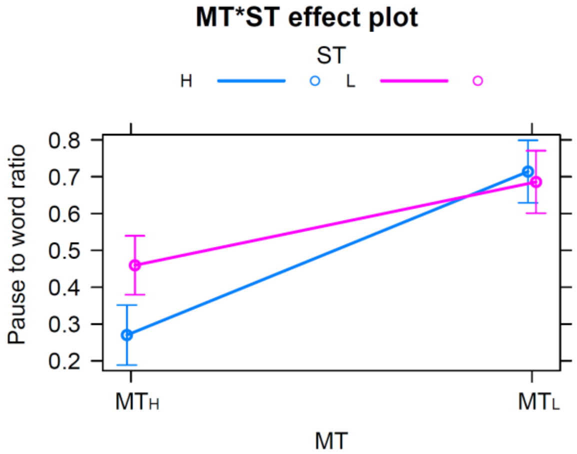

Figure 5. Interaction effect between MT quality and ST complexity on pause to word ratio

4.4. Visual attention allocation

The interaction effect between MT quality and ST complexity on total fixation duration on ST was significant (p<.05) and plotted in Figure 6 (left). ST complexity had a significant positive impact on total fixation duration on ST for PE-MT_L, but not for PE-MT_H. The total fixation duration on ST during PE-MT_H for ST_H (1186 ms) was almost the same as for ST_L (1268 ms) (t=-.199, p=.84). PE-MT_L for ST_H (2284 ms) had significantly longer total fixation duration on ST than for ST_L (1300 ms) (t=2.892, p<.01). MT quality showed a consistent, negative impact on total fixation duration on ST during PE, i.e., processing higher-quality MT costs shorter fixation duration on ST. However, this impact was significant only for ST_H , where significantly shorter total fixation duration on ST was recorded in PE-MT_H (1186 ms) than in PE-MT_L (2284 ms) (t=-3.686, p<.01). For ST_L, total fixation duration on ST during PE-MT_H (1268 ms) was slightly shorter than during PE-MT_L (1300 ms) and the difference was not significant (t=-.16, p=.89).

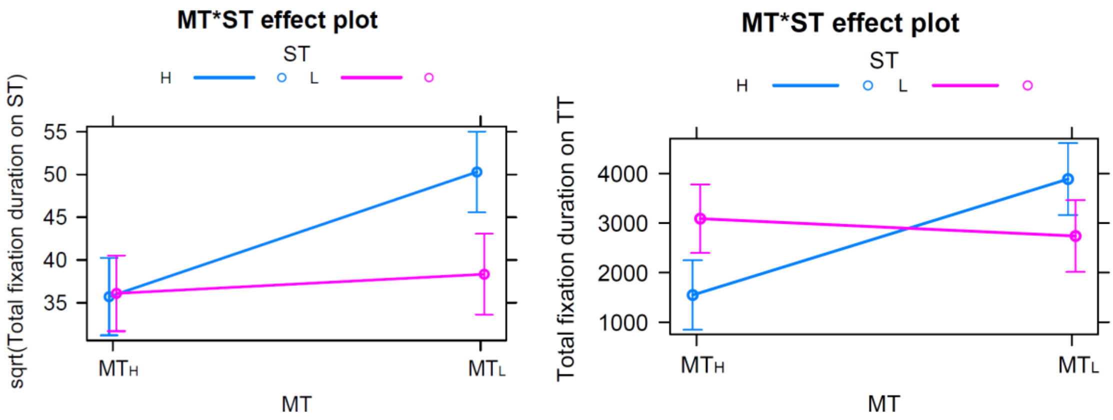

Figure 6. Interaction effect between MT quality and ST complexity on total fixation duration on ST (left) and total fixation duration on TT (right)

The interaction effect between MT quality and ST complexity on total fixation duration on TT was significant (p<.01), with the effect plot shown in Figure 6 (right). ST complexity had a negative effect on total fixation duration on TT during PE-MT$_H$, but a positive effect during PE-MT$_L$. For PE-MT$_L$, the total fixation duration on TT for ST$_H$ (3313 ms) was significantly longer than for ST$_L$ (2274 ms) (t=2.1, p<.05). For PE-MT$_H$, total fixation duration on TT was significantly longer for ST$_L$ (2842 ms) than for ST$_H$ (1325 ms) (t=3.2, p<.01). MT quality demonstrated a significant, negative impact on total fixation duration on TT during PE for ST$_H$, in which PE-MT$_L$ (3313 ms) had significantly longer total fixation duration on TT than PE-MT$_H$ (1325 ms) (t=4.7, p<.001). For ST$_L$, total fixation duration on TT for PE-MT$_H$ (2842 ms) was slightly higher than total fixation duration on TT for PE-MT$_L$ and the difference was not significant (t=.54, p=.85).

4.5. Total number of editing operations

The interaction effect between MT quality and ST complexity on the total number of editing operations in terms of insertions and deletions is presented in Figure 7, with more edits indicating more effort exerted; and this interaction effect was significant (p<.05). For PE-MT$_L$, the total number of editing operations for ST$_H$ (5.25) and for ST$_L$ (5.28) were almost the same (t=-.068, p=.95). For PE-MT$_H$, the total number of editing operations for ST$_L$ (3.1) was significantly higher than for ST$_H$ (1.7) (t=3.52, p<.001). MT quality showed a significant, negative impact on the total number of editing operations during PE, and this impact was stronger for ST_H. For both ST_H and ST_L, PE-MT_H resulted in significantly fewer total number of editing operations than PE-MT_L (t=-8.8, p<.001 for ST_H; t=-5.4, p<.001 for ST_L).

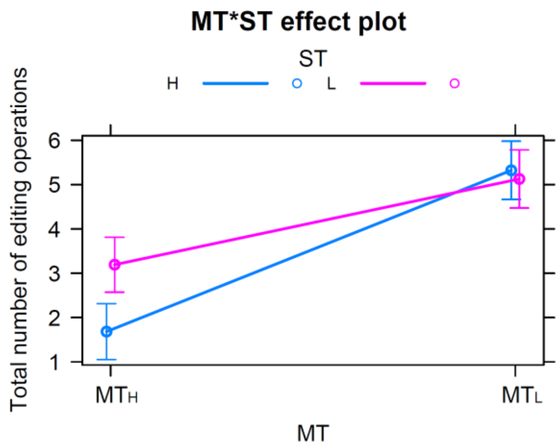

Figure 7. Interaction effect between MT quality and ST complexity on total number of editing operations

4.6. Total number of errors

Finally, the interaction effect between MT quality and ST complexity on the total number of errors is presented in Figure 8 and this effect was not significant (p>.05). For both PE tasks, the translators produced marginally more errors when working on ST_L than on ST_H, with no significant difference. MT quality showed a consistent, negative impact on the total number of errors during PE for both ST_H and ST_L with stronger impact for ST_H. PE-MT_L generated more errors than PE-MT_H (for ST_H, t=2.09, p=.097; for ST_L, t=-1.56, p=.27), but neither impact was significant.

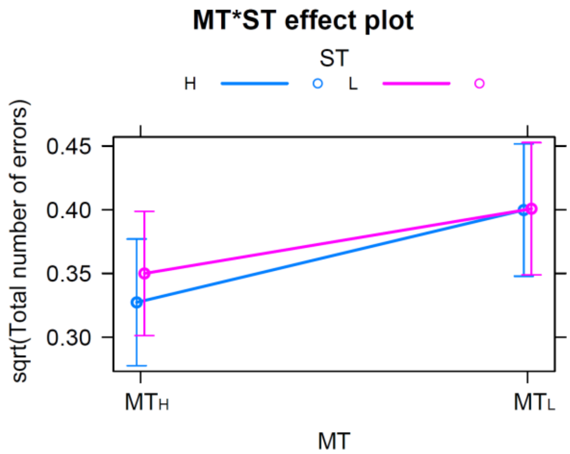

Figure 8. Interaction effect between MT quality and ST complexity on the total number of errors

5. Discussion

ST complexity and MT quality showed an interaction effect on all 7 indicators of PE task difficulty, with 5 indicators being statistically significant (see Table 3). These results suggest that the task difficulty of PE is decided by the combined effects of ST complexity and MT quality. In other words, when participants conduct a PE task, the task difficulty perceived, the processing time spent, the number of pauses produced, the editing operations needed, the fixation duration allocated to ST and TT, and the number of errors produced, are all influenced by factors of ST complexity and MT quality. In the following sections, we will discuss how the two factors interact with each other to impact the overall PE task difficulty. The participants' retrospective reports on problems and difficulties they came across during PE tasks will be applied to support the findings.

Table 3. Interaction effect between MT quality and ST complexity on PE task difficulty

Indicators	Subjective rating	Processing time	Pause to word ratio	Total fixation duration on ST	Total fixation duration on TT	Total number of editing operations	Total number of errors
Interaction effect (*p* value)	>.05	<.01	<.05	<.05	<.01	<.05	>.05

5.1. Effect of ST complexity on PE task difficulty

The impact of ST complexity on PE for MT_L and MT_H were summarized in Table 4, showing that ST complexity has a substantial, positive impact on the task difficulty of PE-MT_L, but not on PE-MT_H.

Table 4. Effect of ST complexity on the PE task difficulty for MT_H and MT_L

ST Complexity	**$ST_L \rightarrow ST_H$**						
Indicators	Subjective rating	Processing time	Pause to word ratio	Total fixation duration on ST	Total fixation duration on TT	Total number of editing operations	Total number of errors

MT_L	↑ **	↑ **	↑	↑ **	↑ *	↓	↓
MT_H	↑	↓	* ↓	↓	↓ **	↓ ***	↓

Note: " ↑ "represents the increase of the value; " ↓ " represents the decrease of the value (* for $p<.05$, ** for $p<.01$, *** for $p<.001$)

When post-editing low-quality machine translation (PE-MT_L), ST complexity displayed a positive impact on 5 out of 7 task difficulty indicators. The increase in ST complexity led to significantly higher subjective rating, processing time, total fixation duration on ST, and total fixation duration on TT. Interestingly, the impact of ST complexity on PE-MT_L was similar to its impact on manual translation as reported in previous studies. Translating high-complexity STs was subjectively rated as more difficult than low-complexity ones (Sun and Shreve, 2014; Liu et al., 2019), took longer processing time (Sun and Shreve, 2014), and had more visual attention on ST and TT (Liu et al., 2019). Although ST complexity seems to have similar impact on PE-MT_L and manual translation, the role ST plays in the two tasks are inherently different.

An ST works both as the reference for correcting MT errors, and the text for translating from scratch in a PE task. According to the retrospective data, 75% of participants reported that, when post-editing a low-quality MT, they had frequently checked the ST for revising the MT output, or for re-translating the segments with critical mistakes from scratch. In other words, when post-editing a text with low-quality MT, participants will have to allocate high cognitive effort to the ST in order to fully understand its meaning either for extensive revision or for manual re-translation without adopting the MT output.

Understanding a high-complexity ST either for PE or for manual translation is cognitively more demanding than understanding a low-complexity ST, as indicated by the significantly longer fixation duration on ST_H. When MT_L was paired with ST_H, 71% of participants expressed their frustration on the PE task, and opined that they might have spent even more effort on the task than on translating from scratch. They reported that the low-quality MT was sometimes very misleading and constantly affected their reading comprehension on the ST. Such experience, however, was not reported by participants when MT_L was paired with ST_L, as reading and understanding low-complexity STs does not require as much effort. PE-MT_L for ST_H is therefore significantly more difficult than for ST_L.

When post-editing high-quality machine translation (PE-MT_H), ST complexity shows a positive, insignificant impact on the subjective rating, with the remaining measurements showing that PE-MT_H for ST_H costs less cognitive effort than for ST_L. The increased cognitive effort exerted for ST_L could have been caused either by ST or MT. The results show that the fixation duration on STs with low and high complexity was approximately the same, implying that ST complexity did not have an impact on cognitive effort exerted in PE-MT_H. According to the retrospective data, 91% of participants reported that they just needed a quick scan of STs for checking the correctness of MT output. Therefore, when MT quality is high, ST reading generally does not require deep cognitive processing regardless of ST complexity level.

On the TT area, however, significantly longer fixation duration was recorded for post-editing ST_L, suggesting that processing these MT outputs requires more cognitive effort. The MT scores for ST_L and ST_H are similarly high. We speculate that the MT output for the ST_L may have contained errors which did not affect the holistic rating of MT quality by human raters but was effortful for correction by participants. Hence, a detailed error analysis using MQM on all four MT outputs was carried out by two professional translators, with the results presented in Table 5.

Table 5. Error analysis on the MT outputs for ST_L and ST_H

Error type	**ST_L**		**ST_H**		**Severity**
	MT_L	MT_H	MT_L	MT_H	
Mistranslation	17	5	16	4	Critical
	5	0	6	2	Minor
Grammar	4	0	3	0	/
Word order	4	3	6	1	/
Omission	0	3	0	0	/
Total	30	11	31	7	/

Table 5 shows that both MT_H outputs (for ST_H and ST_L) are much higher in translation quality with substantially fewer errors than that of MT_L outputs, which can validate the results of our holistic human rating. However, the MT_H for ST_L contains four more errors than MT_H for ST_H. Some errors such as omission may lead to a higher cognitive effort exerted in PE-MT for ST_L. Some recent research also reports that NMT can produce overall high-quality TT, with unpredictable omission and mistranslation "hidden" in the fluent expressions, which are problematic during full PE (e.g., Moorkens, 2018; Yamada, 2019). NMT is also difficult to conceptualize due to complex neural networks behind the system. Although most students did not elaborate how they corrected the errors they came across, two of them particularly reported that they wondered why certain ST words were just omitted unexpectedly in some MT outputs, and that they would not have noticed them had they not checked the STs carefully.

5.2. Effect of MT quality on PE task difficulty

Table 6 shows that MT quality has a negative effect on PE difficulty in general; that is, the higher the MT quality, the lower the difficulty for PE task. This effect becomes stronger when the ST complexity increases.

Table 6. Effect of MT quality on the PE task difficulty indicators for ST_L and ST_H

MT quality	**$MT_L \rightarrow MT_H$**						
Indicators	Subjective rating	Processing time	Pause to word ratio	Total fixation duration on ST	Total fixation duration on TT	Total number of editing operations	Total number of errors
ST_L	↓	↓	↓***	↓	↑	↓***	↓
ST_H	↓***	↓***	↓***	↓**	↓**	↓***	↓

Note: "↑" represents the increase of the value；"↓" represents the decrease of the value; (* for $p<.05$, ** for $p<.01$, *** for $p<.001$)

For low-complexity STs, MT quality has a negative impact on 6 out 7 indicators of PE task difficulty (subjective rating, processing time, pause to word ratio, total fixation duration on ST, total number of editing operations, and total number of errors), with pause to word ratio and total number of editing operations being statistically significant. Compared to post-editing low-quality MT, post-editing high-quality MT was rated to be easier (by participants), with reduced processing time, fixation duration on ST, total editing amount and total number of errors. Fixation duration on TT was about the same between PE-MT_H and PE-MT_L tasks. Daems et al. (2017) found that different types of MT errors could affect different PE effort indicators. The different types of errors which appeared in the MT_L output for ST_L appear to have affected the fixation duration on TT the most but did not evidently affect the other indicators.

For high-complexity STs, MT quality has a negative impact on all 7 indicators, with 6 being statistically significant, indicating that task difficulty of PE decreases significantly when MT quality increases. This is reasonable, as evaluating and revising low quality MT output took more cognitive effort. The negative impact of MT quality on PE difficulty became stronger when the ST was more complex, as indicated by the bigger differences in subjective rating, processing time, pause to word ratio, total number of editing operations, total number of errors between MT_H and MT_L when paired with ST_H. This is mainly because both ST_H and ST_L do not need deep processing and cost high cognitive effort when MT produces high-quality output. On the contrary, in the condition when MT produces low-quality output, processing an ST_H

either for revising MT output or for manual re-translation takes substantially more cognitive effort than processing an ST_L.

5.3. Implications for translation studies and translation market

As the first study investigating how ST complexity and MT quality levels interact to impact the cognitive process of post-editing, our findings will have implications both for translation studies and the translation market.

Methodologically, our study shows that the extent of interaction effects between ST complexity and MT quality on various task difficulty indicators is different. This validates the results of previous studies (e.g., Vieira, 2016; Herbig, 2019) in that "different measures may be more sensitive to different nuances of cognitive effort" (Vieira, 2016:57). Therefore, a multi-method approach could offer a more comprehensive understanding of the cognitive processes of PE and manual translation.

On the side of ST, our findings indicate that, when checking the impact of ST features on PE effort (e.g., O'Brien, 2004, 2006; Aziz et al., 2014), quality of MT outputs should be controlled at a similar level to disentangle the impact of ST from that of MT on the final PE effort. On the side of MT quality, our results support those reported in Krings (2001), Gaspari et al. (2014), Vieira (2016) and Daems et al. (2017), that MT quality has a negative impact on PE effort. None of these studies, however, considered the impact of ST complexity levels. Our findings suggest that it is essential to assess ST complexity if we want to further evaluate the extent of the impact that MT quality imposes on PE effort.

Our results also show that the amount of visual attention allocated to ST and TT areas is significantly impacted by both the given ST complexity and MT quality levels. No previous PE studies have controlled for both the ST complexity and MT quality levels when they looked into the visual attention allocation to ST and TT, which can explain why their results are likely to be inconsistent and difficult to compare (e.g., Carl et al., 2011; Mesa-Lao, 2014; Daems et al., 2017).

For the translation industry, pricing a post-editing task is more challenging than conventional manual translation. A cost-effective operating model for PE pricing is still far from being well-established (TAUS, 2013, 2016). Recent studies have touched on how to improve the MT quality evaluation metrics to better predict PE effort (e.g., Specia and Shah, 2018), but research has not yet considered the potential impact of ST and its interaction with MT quality on PE effort. Our findings indicate that a predictive and fair PE pricing model should factor in both MT quality and ST complexity.

6. Conclusion

This paper investigated the interaction effect between ST complexity and MT quality on the task difficulty of post-editing. The findings can be summarized as follows. Firstly, a significant interaction effect was found between MT quality and ST complexity on most difficulty indicators of PE tasks. Secondly, ST complexity has a substantial, positive impact on the task difficulty of post-editing low-quality MT, similar to its impact on manual translation. However, it has no positive impact on post-editing high-quality MT. This is because ST does not require deep cognitive processing when the MT quality is high enough. Thirdly, MT quality has a negative impact on the task difficulty of post-editing and this effect becomes stronger when the complexity level of ST increases. Therefore, the overall task difficulty of post-editing is decided by both MT quality and ST complexity. They cannot be decoupled from each other and should both be assessed or controlled when developing post-editing pricing schemes or designing post-editing tasks for training purposes.

The results yielded in this study may contribute to the development of training courses and pricing schemes for MT post-editing. However, we are aware that some limitations exist in this study, such as recruitment of a single student participant group and applying only one text type and language pair. Our future research will include professional translators and other text domains for more generalizable results. In addition, our next step of analysis will include detailed eye-tracking data to investigate how different types of ST features and MT errors affect post-editing effort.

Note

1.The Test for English Majors Band 8 is a national English test for English majors in China, which includes tests for listening, reading, writing, translating, proofreading and general knowledge, and requires a candidate to master 13,000 words.
2. Due to the word limit, the four STs and their machine translation are not presented in this paper. Interested readers can request the STs from the first author.
3. A positive impact in this article means that the increase of the value of the independent variable leads to the increase of the value of the dependent variable; while a negative impact means that the increase of the value of the independent variable results in the decrease of the value of the dependent variable.

Acknowledgements

Great thanks go to Prof. Yves Gambier, Dr. Joss Moorkens, the two anonymous reviewers and Dr. Pál Heltai for their comments on the earlier drafts of this article. This research was supported by the Philosophy and Social Science Foundation of Hunan Province (No.20YBQN068).

References

Aikawa, T., Schwartz, L., King, R., Corston-Oliver, M., & Carmen, L. (2007). Impact of controlled language on translation quality and post-editing in a statistical machine translation environment. In B. Maegaard (Ed.), *Proceedings of the MT Summit XI, Copenhagen, Denmark* (pp. 1-7). Copenhagen, Denmark.

Aziz, W., Koponen, M., & Specia, L. (2014). Sub-sentence level analysis of machine translation post-editing effort. In S. O'Brien, L. W. Balling, M. Carl, M. Simard & L. Specia (Eds.), *Post-editing of machine translation: Processes and applications* (pp. 170-199). Cambridge Scholars Publishing.

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2014). *Lme4: linear mixed-effects models using Eigen and S4. R package version 3.1.2*. http://CRAN.R-proje ct.org/package=lme4.

Carl, M., Dragsted, B., Elming, J., Hardt, D., & Jakobsen, A. L. (2011). The process of post-editing: a pilot study. *Copenhagen Studies in Language*, *41*, 131-142.

Castilho, S., Moorkens, J., Gaspari, F., Sennrich, R., Way, A. & Georgakopoulou, P. (2018). Evaluating MT for massive open online courses. *Machine Translation*, *32*, 255–278. https://doi.org/10.1007/s10590-019-09232-x

Daems, J., Vandepitte, S., Hartsuiker, R. J., & Macken, L. (2017). Identifying the machine translation error types with the greatest impact on post-editing effort. *Frontiers in Psychology*, *8*, Article1282. https://doi.org/10.3389/fpsyg.2017.01282

Dahl, Ö. (2004). *The Growth and Maintenance of Linguistic Complexity*. John Benjamins.

Fox, J., Weisberg, S., Friendly, M., & Hong, J. (2017). *Effects: Effect displays for linear, generalized linear, and other models. R package version 4.0-0.* https://cran.r-project.org/web/packages/effect.

Gallupe, R. B., DeSanctis, G., & Dickson, W. G. (1988). Computer-based support for group problem-finding: An experimental investigation. *MIS Quarterly*, *12*(2), 277–296.

Gaspari, F., Toral, A., Naskar, S. K., Groves, D., & Way, A. (2014, October). *Perception vs reality: measuring machine translation post-editing productivity* [Paper presentation]. The third workshop on post-editing technology and practice (WPTP-3), within the eleventh biennial conference of the Association for Machine Translation in the Americas (AMTA-2014). Vancouver, Canada.

Halverson, S. L. (2017). Multimethod approaches. In J. W. Schwieter, A. Ferreira & J. Wiley (Eds.), *The handbook of translation and cognition* (pp. 195-212*)*. Wiley-Blackwell.

Herbig, N., Pal, S., Vela, M., Krüger, A., & Genabith, J. (2019). Multi-modal indicators for estimating perceived cognitive load in post-editing of machine translation. *Machine Translation, 33*, 91–115. https://doi.org/10.1007/s10590-019-09227-8

Hvelplund, K. T. (2011). *Allocation of cognitive resources in translation: An eye-tracking and key-logging study*. [Unpublished PhD dissertation]. Copenhagen Business School.

International Organization for Standardization. (2017). *Translation services - Post-editing of machine translation output - Requirements* (ISO Standard No. 18587:2017). https://www.iso.org/standard/62970.html

Junczys-Dowmunt, M. T., & Dwojak, H. (2016). Is neural machine translation ready for deployment? A case study on 30 translation directions. In *Proceedings of the 9th international workshop on spoken language translation, Seattle, WA.* https://arxiv.org/abs/1610.01108.

Kappus, M., & Ehrensberger-Dow, M. (2020). The ergonomics of translation tools: understanding when less is actually more. *The Interpreter and Translator Trainer*, *14*(4), 386-404. https://doi.org/10.1080/1750399X.2020.1839998

Krings, H. P. (2001). *Repairing texts: Empirical investigations of machine translation post-editing processes*. The Kent State University Press.

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2014). *lmerTest: Tests in linear mixed effects models. R package version 2.0-20.* http://CRAN.R-project.org/package=lmerTest.

Lacruz, I., & Shreve, G. M. (2014). Pauses and cognitive effort in post-editing. In S. O'Brien, L. W. Balling, M. Carl, M. Simard, & L. Specia (Eds.), *Post-editing of machine translation: Processes and applications* (pp. 246–272). Cambridge Scholars Publishing.

Liu, Y., Zheng, B., & Zhou, H. (2019). Measuring the difficulty of text translation: The combination of text-focused and translator-oriented approaches. *Target, 31*(1), 125-149. https://doi.org/10.1075/target.18036.zhe

Lommel, A. (2018). The multidimensional quality metrics and dynamic quality framework. In J. Moorkens, S. Castilho, F. Gaspari, & S. Doherty (Eds.), T*ranslation quality assessment: From principles to practice* (pp. 109–127). Springer.

Mesa-Lao, B. (2014). Gaze behaviour on source texts: An exploratory study comparing translation and post-editing. In S. O'Brien, L. W. Balling, M. Carl, M. Simard, & L. Specia (Eds.), *Post-editing of machine translation: Processes and applications* (pp. 219–245). Cambridge Scholars Publishing.

Moorkens, J. (2018). What to expect from Neural Machine Translation: a practical in-class translation evaluation exercise. *The Interpreter and Translator Trainer*, *12*(4), 375-387. https://doi.org/10.1080/1750399X.2018.1501639

O'Brien, S. (2004). Machine translatability and post-editing effort: How do they relate. *Translating and the Computer, 26*, 1-31.

O’Brien, S. (2006). Controlled language and post-editing. *Multilingual*, *17*(7), 17-19. https://multilingual.com/issues/2006-10-11.pdf

O’Brien, S. (2011). Towards predicting post-editing productivity. *Machine Translation*，*25*(3), 197–215. https://doi.org/10.1007/s10590-011-9096-7

Paas, F., & Van Merriënboer, J. J. G. (1994). Instructional control of cognitive load in the training of complex cognitive tasks. *Educational Psychology Review*, *6*(4), 351-371. https://doi.org/10.1007/BF02213420

Sennrich, R., Haddow, B., & Birch, A. (2016). Neural machine translation of rare words with subword units. *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, Berlin, Germany* (pp. 1715–1725). Association for Computational Linguistics. https://arxiv.org/abs/1508.07909

Sun, S. (2015). Measuring translation difficulty: Theoretical and methodological considerations. *Across Languages and Cultures*, *16*(1), 29–54. https://doi.org/10.1556/084.2015.16.1.2

Sweller, J. (1988). Cognitive load during problem solving: Effects on learning. *Cognitive Science*, *12*(2), 257-285. https://doi.org/10.1207/s15516709cog1202_4

Sweller, J., Ayres, P., & Kalyuga, S. (2011). Cognitive load theory in perspective. In J. Sweller (Ed.), *Cognitive Load Theory* (pp. 237-242). Springer.

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

TAUS. (2019). *A review of the TAUS global content conference in Salt Lake City*. https://www.taus.net/academy/reports.

Temizöz, Ö. (2012). *Machine translation and postediting.* European Society for Translation Studies Research Committee State of the Art Research Reports.

Torrón, M. S., & Koehn, P. (2016). Machine translation quality and post-editor productivity. In S, Green., & L, Schwartz (Eds.), *MT researcher's track, within proceedings of Association for Machine Translation in the Americas (AMTA-2016)* (pp. 16-26). Austin, USA.

Vieira, L. N. (2016). *Cognitive effort in post-editing of machine translation: Evidence from eye movements, subjective ratings, and think-aloud protocols*. [Unpublished doctoral dissertation]. Newcastle University.

Vieira, L. N. (2019.) Post-editing of machine translation. In M. O'Hagan (Ed.), *The Routledge handbook of translation and technology* (pp. 206-318). Routledge.

Yamada, M. (2019). The impact of Google neural machine translation on post-editing by student translators. *The Journal of Specialised Translation*, *31*, 87-106. https://jostrans.org/issue31/art_yamada.php