Towards Predicting Post-editing Effort with Source Text Readability: An Investigation for English-Chinese Machine Translation

Guangrong Dai, Guangdong University of Foreign Studies
Siqi Liu, Guangdong University of Foreign Studies

ABSTRACT

This paper investigates the impact of source text readability on the effort of post-editing English-Chinese Neural Machine Translation (NMT) output. Six readability formulas, including both traditional and newer ones, were employed to measure readability, and their predictive power towards post-editing effort was evaluated. Keystroke logging, self-report questionnaires, and retrospective protocols were applied to collect the data of post-editing for general text type from thirty-four student translators. The results reveal that: 1) readability has a significant yet weak effect on cognitive effort, while its impact on temporal and technical effort is less pronounced; 2) high NMT quality may alleviate the effect of readability; 3) readability formulas have the ability to predict post-editing effort to a certain extent, and newer formulas such as the Crowdsourced Algorithm of Reading Comprehension (CAREC) outperformed traditional formulas in most cases. Apart from readability formulas, the study shows that some fine-grained reading-related linguistic features are good predictors of post-editing time. Finally, this paper provides implications for automatic effort estimation in the translation industry.

KEYWORDS

Neural machine translation, post-editing effort, readability, key logging, student translator.

1. Introduction

In the translation industry, machine translation post-editing (MTPE) has not only become feasible but also essential thanks to the development of neural machine translation (NMT) and the emerging demand for language services. In academic settings, MTPE has also been recognised as a cost-efficient workflow. Several studies on given language pairs and contexts showed that MTPE generally makes translation faster (O'Brien 2007; Lu and Sun 2018), and that the final product quality is equivalent or even better compared with from-scratch translation (Green *et al*. 2013; Jia *et al*. 2019a).

However, certain issues, such as the development of the pricing model and the evaluation of the cost-effectiveness of MTPE, remain to be addressed. While MT quality is considered as evidence, the amount of MTPE effort, "not only the ratio of quantity and quality to time but also the cognitive effort expended" (O'Brien 2011: 198), should also be a prime concern, since it focuses more on the interaction between translators and MT output (Herbig *et al.* 2019). Based on Krings (2001)'s classification of MTPE effort, previous studies have investigated which factors influence temporal, technical and cognitive effort from mainly two aspects: textual features (O'Brien 2005; Tatsumi and Roturier 2010; Koponen *et al.* 2012), and translators' characteristics (Vieira 2014; Daems *et al.* 2017). While these factors proved

https://doi.org/10.26034/cm.jostrans.2024.4723

to be more or less correlated with MTPE effort, the effects of different factors, particularly source text characteristics and MT quality, have often been conflated since many experimental settings did not control one of the factors when the other one is under investigation. Accordingly, it was hard to disentangle their separate contributions. Moreover, most of the previous studies have been carried out in the context of Statistical Machine Translation (SMT), which differs substantially from the currently prevalent NMT (Jia and Zheng 2022). Although NMT achieves state-of-the-art results, it is accompanied by fluent but inadequate errors, which may be overlooked by translators and pose new challenges for MTPE (Castilho *et al.* 2017; Popović 2020; Dai and Liu 2023).

The task of MTPE is to identify and modify errors in the MT output, which is mostly achieved by cross-checking source text and MT. Arguably, it is mostly a reading rather than a writing process, so the focal point of MTPE research should direct to reading-related aspects (Koponen *et al.* 2020:17). As an index of reading difficulty, readability scores have a potential link with MTPE effort. However, readability is mostly discussed in the context of human translation difficulty (Sun 2019:144), while the extent to which source text readability affects MTPE effort is still to be explored. Meanwhile, it is of vital importance to investigate whether such source text features can be used to automatically predict effort, since letting translators themselves evaluate the cost-effectiveness of MTPE would require more time and effort (Daems *et al*. 2017).

Given the aforementioned reasons, this study explores the impact of source text readability on the effort of post-editing English-Chinese NMT output. We also endeavour to predict MTPE time with some reading-related linguistic features. This study is expected to provide evidence for the MTPE pricing model, shed new light on the development of automatic effort estimation and ultimately improve the productivity of MTPE. In particular, it addresses three research questions:

1. How does source text readability impact post-editing effort?
2. What is the predictive power of different readability formulas, including both traditional and newer ones, for post-editing effort?
3. Is it possible to predict post-editing time based on fine-grained reading-related linguistic features?

2. Related research

The three-fold division of MTPE effort, as proposed by Krings (2001), comprises temporal, technical, and cognitive effort and is widely used in MTPE research. *Temporal effort* refers to the time spent during the MTPE process, which is captured by measurements such as processing speed (O'Brien 2011). *Technical effort* involves a series of manual corrections of the MT output, calculated by keystroke logs (Jia *et al.* 2019b). Finally, *cognitive effort* involves "the type and extent of cognitive processes

triggered by the post-editing task" (Krings 2001: 182). While these two types of effort are easier to measure and used more often in the translation industry, cognitive effort cannot be observed directly and is rather confined to academic research. Multiple methods such as think-aloud protocols (Krings 2001), choice network analysis (O'Brien 2005), pause analysis (Toral *et al.* 2018), eye-tracking (Daems *et al.* 2017) and subjective ratings (Vieira 2014) have been introduced to evaluate cognitive effort. These three dimensions of MTPE effort are inherently different and interrelated (Krings 2001:179), and it was found that the correlation between various metrics of effort is rather low (Cumbreño and Aranberri 2021), which underlines the necessity of triangulating data covering different aspects when evaluating MTPE effort.

A crucial question for MTPE research is which factors influence effort. There have been studies investigating whether textual features correlate with MTPE effort. The quality of MT has been considered to be a key variable (Jia and Zheng 2022), but research results are mixed. Krings (2001) found that the correlation between MT quality and MTPE effort is not necessarily linear while other studies suggested that MT quality is negatively correlated with MTPE effort (Tatsumi 2009; O'Brien 2011; Vieira 2014). Apart from viewing general MT quality, MT error classification provides a finer-grained perspective. For instance, errors regarding word order, omission/addition, style, coherence and so forth were found to be strongly correlated with MTPE effort (Popović *et al.* 2014; Daems *et al.* 2017; Qian *et al.* 2022).

While many scholarly endeavours have been devoted to evaluating the impact of MT output, source text features are little explored. In contrast, they have been widely discussed in the human translation setting (Campbell 1999; Hvelplund 2011; Sun 2019). One of the reasons for this imbalanced distribution of research may be that the dominant role of source text has changed in the context of MTPE, and that translators seem to pay more attention to the MT than to the source text (Koglin 2015, Lu and Sun 2018). There is even a view that translators can do MTPE without access to the source text (Koponen and Salmi 2015; Li 2021). Nevertheless, more evidence is needed to draw more robust conclusions about whether such monolingual MTPE works with the less visible errors recurring in NMT. Meanwhile, it should be noted that the impact of the source text is not only confined to the allocation of cognitive resources, but could be extended to the resources allocated to the MT output, since source text features are mirrored in the MT to a certain extent. Therefore, it is still of vital importance to explore the relationship between source text features and MTPE effort.

There has been some relevant research on source text features and MTPE effort, but the results seem to be far from conclusive. O'Brien (2005) discovered that while some source text items that were recognised as negative translatability indicators (NTIs) led to increased effort, the non-NTIs could also increase cognitive processing. Tatsumi and Roturier (2010)

suggested that a complexity score, based on a series of source text features such as sentence length, strongly correlated with technical effort. However, in Aziz *et al*. (2014), the correlation between sentence length and temporal effort was not strong, and Jia *et al.* (2019b) concluded that source text complexity, measured by human ratings, readability scores, word frequency, and non-literalness, did not necessarily affect MTPE effort. The reason for such inconsistencies might be that these studies have chosen different source text features and adopted different ways of evaluating MTPE effort, and that the results were mixed with the effect of MT quality when comparing the effort of post-editing texts of different complexities. In a study with a more rigorous research design, Jia and Zheng (2022) investigated the interaction effect between source text complexity and MT quality. They found that source text complexity had a significant impact on the effort of post-editing low-quality MT. While this study combined four sets of measurements to identify source text complexity, namely readability scores, word frequency, syntactic complexity, and subjective evaluation, the relationships between each dimension of source text complexity and MTPE effort have not been elucidated. Of note, the measurements in this study have a different focus. Readability scores, for instance, focus on reading difficulties, while subjective evaluation concerns translation difficulties. Accordingly, investigating these features separately may provide a more nuanced understanding of the impact of source text on MTPE effort.

Since reading is a significant component of the MTPE process, it is necessary to investigate whether readability, the ease of understanding and processing a text (Nahatame 2021), has an impact on MTPE effort. In previous relevant studies, source text readability was mainly measured using traditional readability formulas including Flesch Reading Ease (Flesch 1948). However, such formulas only examine surface-level linguistic features and fail to give a more in-depth look at text comprehensibility (Graesser *et al.* 2011), limiting its potential use in predicting MTPE effort. In view of such limitations, deeper linguistic features such as text cohesion were taken into account for the development of newer readability formulas, for example, the Coh-Metrix L2 Reading Index (Crossley *et al.* 2008). The newer formulas capture the cognitive process of reading more accurately, and empirical investigations suggested that they are better for estimating text comprehensibility and predicting text processing effort (Crossley *et al.* 2019; Nahatame 2021). However, to the best of our knowledge, such cognitively inspired formulas have yet to be employed for MTPE effort research, and there is little evidence whether or not newer formulas outperform the traditional ones in MTPE effort prediction.

Finally, attempts have been made to build models that automatically predict MTPE effort. Such prediction is usually related to quality estimation (QE), which involves textual feature extraction, annotated scores of MT quality and machine learning algorithms (Specia and Shah 2018: 203). However, QE's relation to actual MTPE effort has yet to be attested (O'Brien 2011;

Tezcan *et al.* 2019). Another concern regarding QE is that the interpretation of such complex models remains "cryptic" to translators (Marg 2016). If translators are guided and paid according to information that they do not really understand, the productivity of MTPE may not necessarily be improved. Conversely, adopting simple linguistic features that have the predictive power for MTPE effort should be more comprehensible to translators. Moreover, it was previously suggested that presenting scores on source text characteristics may be helpful for translators to estimate MTPE time (Tatsumi and Roturier 2010).

As mentioned, MTPE research has been mainly concerned with MT quality, while source text features were largely neglected. Therefore, the current study focuses on the effect of the source text, specifically source text readability, on MTPE effort. Source text readability is evaluated via both traditional and newer readability formulas, while MTPE effort is measured using a combination of keystroke logging, self-report questionnaires and retrospective protocols. Models based on reading-related linguistic features for predicting MTPE time were also developed.

3. Materials and methods

3.1. Participants

Thirty-four first-year Master in Translation and Interpreting (MTI) students (2 males, 32 females; Chinese as their L1 and English as L2), aged 21 to 25 years old, participated in this study. They had a similar level of English proficiency and an average LexTALE test score of 79 (SD=9) indicated that they were advanced English learners (Lemhöfer and Broersma 2012). In addition, they passed the Test for English Majors at Band4 (TEM4) and the China Accreditation Test for Translators and Interpreters (CATTI) Level 3 (translator).[1] Although none of them had worked as professional translators and they had limited MTPE experience, the translation qualification that the participants obtained indicates that they were able to accomplish general translation work. Therefore, the results produced by participants in this study provide implications particularly for novice translators who are new to MTPE. Finally, all participants signed a consent form and were rewarded with 50 yuan for their work.

3.2. Materials

3.2.1. Readability measurement

Three traditional readability formulas and three newer ones were adopted to evaluate the readability scores of the source texts. The traditional formulas are the Flesch Reading Ease (RDFRE) formula (Flesch 1948), the Flesch-Kincaid Grade Level (RDFKGL) formula (Kincaid *et al*. 1975), and the Dale-Chall (DC) formula (Dale and Chall 1948). The RDFRE and RDFKGL measure readability based on word length and sentence length

while the DC relies on the ratio of difficult words. The newer formulas, namely the Coh-Metrix L2 Reading Index (CML2RI) (Crossley *et al.* 2008), the Crowdsourced Algorithm of Reading Comprehension (CAREC) (Crossley *et al.* 2019), and the Sentence BERT Readability Model (SBERT) (Reimers and Gurevych 2019), comprise richer linguistic features such as word overlap, and syntactic similarity (see Choi and Crossley (2022) for more detailed information).

The Coh-Metrix Desktop Tool (McNamara *et al.* 2014) was used to obtain the RDFRE, the RDFKGL and the CML2RI scores, while the DC, the CAREC, and the SBERT scores were acquired via the Automatic Readability Tool for English (ARTE; Choi and Crossley 2022). The RDFKGL, the DC, and the CAREC scores indicate higher text complexity as they increase, while the RDFRE, the CML2RI, and the SBERT scores suggest lower text complexity as they increase.

3.2.2. Source texts selection

Six English news texts from the general domain were selected for the study. ST1, ST3, ST5 and ST6 were from *newsela.com*, a website which provides various adaptations of authentic English news. ST2 and ST4 were from the *multiLing* set of the CRITT TPR-DB (Carl *et al.* 2016). All texts were self-contained and required no specialist knowledge to be post-edited. Under the premise that semantic coherence is preserved, the texts were shortened to 139-150 words. An English native speaker was invited to read them to ensure the comprehensibility of texts. After that, the readability scores were measured, see Table 1.

	ST1	ST2	ST3	ST4	ST5	ST6
Word count	150	148	146	139	150	142
Avg. Sentence length	16.667	13.455	14.6	19.857	18.75	20.286
The RDFRE score	65.302	72.539	61.055	51.96	37.216	50.063
The RDFKGL score	8.292	6.484	8.37	11.016	12.727	11.352
The DC score	7.94	7.4	8.58	8.24	10.64	10.83
The CML2RI score	21.893	19.196	13.476	12.831	10.234	9.588
The CAREC score	0.211	0.1	0.213	0.226	0.237	0.315
The SBERT score	-0.462	-0.173	-0.098	-0.279	-1.036	-0.726

Table 1. Source text readability scores

3.2.3. MT quality assessment

Two second-year MTI students and two second-year MA students in translation were recruited to evaluate the MT outputs. They all had experience in MT error annotation and have passed the CATTI Level 2 (translator). The MT quality evaluation was conducted with TAUS (2019)'s adequacy and fluency approaches. Specifically, the extent to which the source text meaning is expressed in the MT output and the well-formedness

of the MT output were rated separately on a 4-point scale, where "1" represents none/incomprehensible and "4" represents everything/flawless. All the evaluators had training using the scoring rubric. To prevent bias, they were not informed about which MT system was being rated.

Since the focus of the study is to investigate the impact of source text readability on MTPE effort, we believe it is necessary to control the impact of MT outputs. In other words, MT outputs should be of similar quality (Jia and Zheng 2022), so that the difference between the effort spent on editing different texts could be better attributed to readability. A pilot MT evaluation was first conducted on four widely-used NMT engines (Google Translate, DeepL Translate, Youdao Translate and Baidu Translate) translating ST1, ST3, and ST6. As shown in Table 2, Youdao Translate showed the most consistent performances in translating different texts (especially in terms of ST6). Therefore, Youdao Translate was selected for the experiments.

			ST1	ST3	ST6
Fluency score	Google	mean/sd.	3.64/0.14	3.33/0.20	2.90/0.38
		min/max	3.50/3.81	3.17/3.58	2.40/3.20
	DeepL	mean/sd.	3.63/0.10	3.54/0.14	3.15/0.34
		min/max	3.50/3.75	3.42/3.75	2.80/3.60
	Youdao	mean/sd.	3.50/0.14	3.58/0.13	3.59/0.16
		min/max	3.33/3.67	3.40/3.70	3.38/3.75
	Baidu	mean/sd.	3.63/0.15	3.56/0.18	3.30/0.26
		min/max	3.50/3.81	3.33/3.75	3.00/3.60
Adequacy score	Google	mean/sd.	3.73/0.19	3.54/0.21	3.05/0.34
		min/max	3.50/3.94	3.25/3.75	2.60/3.40
	DeepL	mean/sd.	3.64/0.13	3.69/0.08	3.65/0.19
		min/max	3.50/3.81	3.58/3.75	3.40/3.80
	Youdao	mean/sd.	3.56/0.24	3.68/0.15	3.38/0.18
		min/max	3.22/3.78	3.50/3.80	3.13/3.50
	Baidu	mean/sd.	3.62/0.22	3.71/0.11	3.35/0.19
		min/max	3.38/3.88	3.58/3.83	3.20/3.60

Table 2. The results of the pilot MT evaluation

Table 3 presents the evaluation results regarding the MT produced by Youdao Translate. The inter-rater agreement was strong and significant for both fluency (Kendall's $W=0.739$, $p<0.05$) and accuracy (Kendall's $W=0.659$, $p<0.05$). According to the one-way ANOVA pairwise comparison, six texts scored similarly in terms of both fluency ($F=1.105$, $p>0.05$) and accuracy ($F=1.044$, $p>0.05$) with no significant difference, indicating that all texts were of comparable MT quality.

		ST1	ST2	ST3	ST4	ST5	ST6
Fluency score	mean/sd.	3.50/0.14	3.64/0.13	3.58/0.13	3.46/0.14	3.59/0.16	3.59/0.16
	min/max	3.33/3.67	3.46/3.73	3.40/3.70	3.29/3.57	3.38/3.75	3.38/3.75
Adequacy score	mean/sd.	3.56/0.24	3.50/0.12	3.68/0.15	3.54/0.18	3.38/0.18	3.38/0.18
	min/max	3.22/3.78	3.36/3.64	3.50/3.80	3.29/3.71	3.13/3.50	3.13/3.50

Table 3. The evaluation results of the MT produced by Youdao Translate

3.3. Experimental procedures

The MTPE tasks were done within the Translog-II interface (Carl 2012) in November 2022. Before the experiment started, all participants filled in a questionnaire regarding their language, translation and MTPE background. They were asked to do "full post-editing" according to ISO 18587 (2017) and were informed that no external resources such as dictionaries were allowed. There was no time constraint, but participants were required to finish the tasks as soon as possible. They were notified about the layout of the interface, which shows the whole source text in the upper part and the corresponding MT in the lower part (see Figure 1).

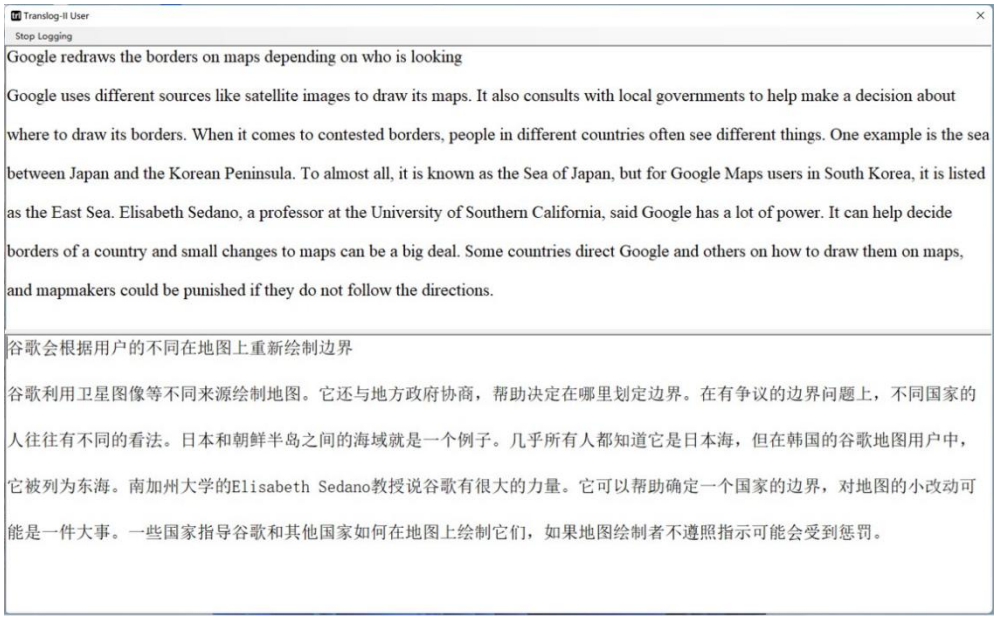

Figure 1. Screenshot of the Translog-II user interface

Each participant finished six tasks and the order of the tasks was balanced across participants in a Latin square design. In order to minimise the impact of fatigue, three tasks were done in the morning session and the other three were done in the afternoon session (Daems *et al.* 2017). All tasks were conducted in the same classroom and took 66 minutes on average (SD=17.413). In the beginning of the morning session, participants were asked to do a warm-up task to get familiar with the interface. After that, they were immediately shown the screen recording of their MTPE process via the "replay" function of Translog-II and were invited to do a

retrospective verbal report concurrently. The recording was played in fast forward mode (two or five times, according to participants' preference) due to time constraints. The participants could freely pause the video or adjust the speed as they commented. There was an outline for the report, based on which participants freely talked about their MTPE patterns, comments on source text and the MT, the difficulties encountered, and the reasons for edits. After the retrospective report, participants rated their subjective cognitive effort. Finally, they took a 5-minute break and proceeded to the main task. All the main tasks followed the same process as the warm-up task did, and there was always a 5-minute break between each task. In the end of the afternoon session, students were asked to take the LexTale test.

3.4. Data processing and statistical analysis

In total, 204 Translog-II xml files that contain the data logged from the MTPE tasks were collected for the study. All the files were uploaded to the CRITT TPR-DB, which can generate different tables and features regarding the MTPE behaviours (Carl *et al.* 2016).

The data analysis was conducted at the textual level with the statistical software R (R Core Team 2022). Specifically, the study adopted Linear Mixed Effects Regression (LMER) models to examine the relationship between readability and MTPE effort. Six null models without a predictor were built via the lme4 package (Bates *et al.* 2015), each with one effort indicator as the dependent variable: 1) total time, 2) the number of keystrokes, 3) average pause time, 4) pause to word ratio, 5) initial pause, 6) subjective cognitive effort. For each dependent variable, we then built six full models separately, each with source text readability (measured by six different formulas respectively) as fixed effect. Forty-two models were built in total, and all of them included participants and texts as random effects.

Before fitting the models, readability scores were *z*-standardised, and the dependent variables that did not follow a normal distribution were transformed via the powerTransform function in the car package (Fox and Weisberg 2019). Subsequently, the processed variables were entered into the models. We then checked whether the residuals of the models were normally distributed. If not, outliers with standardised residuals over 2.5 standard deviations would be removed and the models were refitted (Wu and Ma 2020).

In order to examine whether there is an effect from readability on MTPE effort, we use a log-likelihood ratio test to compare the null models and the full models. Akaike's Information Criterion (AIC) values were adopted to determine the best-fitting full models, with a lower value indicating better performance. Finally, to assess the predictive power of readability, we employed the lmerTest package (Kuznetsova *et al.* 2017) and the MuMIn

package (Bartoń 2023) to measure the significance of the fixed effect and the effect size.

Qualitative data supplements the quantitative data in the current study. 204 retrospective reports were transcribed and coded. Although participants generally talked about every point in the given outline, only the data pertaining to encountered difficulties and comments on the source text and MT were coded, given the research focus and effort constraints.

4. Data analysis and discussion

4.1. Total time

Total time is the total task duration (in millisecond), normalised by the number of words in the source text. According to Table 4, only the difference between the null model and the RDFRE-included model approached significance (χ^2=3.810, p<0.1). The RDFRE performed the best fit to the total time (AIC=1629.7) and showed a marginally significant negative effect (t=-2.357, p<0.1) on total time, which suggests that the more readable the text, the less temporal effort it takes to post-edit.

		Fixed effect					Effect sizes		Log-likelihood ratio test		AIC
	Parameters	*b*	*SE*	95% CI	*t*	*p*	R^2 marginal	R^2 conditional	χ	*p*	
Null Model	Intercept	114.460	3.180	[107.966, 120.955]	35.99	<2e-16 ***	0.000	0.654	-	-	1631.500
Full Model 1	Intercept	114.449	2.862	[108.670, 120.236]	39.982	<2e-16 ***	0.029	0.652	3.810	0.051 .	1629.700
	RDFRE	-3.228	1.369	[-6.448, 0.019]	-2.357	0.0614 .					
Full Model 2	Intercept	114.453	2.944	[108.496, 120.416]	38.872	<2e-16 ***	0.022	0.653	2.617	0.106	1630.900
	RDFKGL	2.810	1.531	[-0.805, 6.403]	1.835	0.121					
Full Model 3	Intercept	114.457	3.096	[108.157, 120.760]	36.97	<2e-16 ***	0.008	0.654	0.819	0.366	1632.700
	DC	1.695	1.804	[-2.543, 5.918]	0.94	0.386					
Full Model 4	Intercept	114.635	3.185	[108.135, 121.136]	35.996	<2e-16 ***	0.000	0.647	0.033	0.856	1648.500
	CAREC	-0.350	1.932	[-4.877, 4.173]	-0.181	0.862					
Full Model 5	Intercept	114.458	3.036	[108.295, 120.624]	37.698	<2e-16 ***	0.014	0.654	1.486	0.223	1632.100
	CML2RI	-2.217	1.697	[-6.198, 1.775]	-1.307	0.243					
Full Model 6	Intercept	114.455	3.009	[108.354, 120.561]	38.04	<2e-16 ***	0.016	0.653	1.798	0.180	1631.700
	SBERT	-2.405	1.648	[-6.275, 1.476]	-1.46	0.199					

Table 4. Summary of the null model and full models for total time (. p<0.1, * p<0.05, ** p<0.01)

The results above reveal that readability might not be an accurate predictor of total time. Although the RDFRE demonstrated a marginally significant prediction, it was not sufficiently reliable. Possible explanations for this phenomenon are related to the MT quality. Firstly, the adopted MT output was of relatively high quality. In this case, according to Jia and Zheng (2022), reading a source text does not generally cause deep cognitive processing. Similarly, retrospective protocols suggest that the MT output alleviated the effect of readability. For example, P32 mentioned that she read the MT first and thought the quality was good, so she only referred to the source text when the MT seemed wrong. P34 commented that MT helped her figure out the meaning of certain words. Meanwhile, the general

MT quality was controlled since the study has a focus on the impact of readability. While the total time involves reading time, the editing time is also considered, which is closely connected with the MT quality. Therefore, the similar quality of MT might lead to similar editing time, contributing to the insignificant difference between total time regarding different texts.

4.2. Total number of keystrokes

The total number of keystrokes includes the number of insertions and deletions, which was normalised by the number of characters in the target text. As presented in Table 5, the differences between the null model and the full models were mostly insignificant. Only the CAREC-included model showed a marginally significant difference (χ^2=2.772, p<0.1), and the CAREC performed the best prediction of the number of keystrokes (AIC=211.3). However, the fixed effect of readability was neither significant nor approached significance in any models.

		Fixed effect					Effect sizes		Log-likelihood ratio test		AIC
	Parameters	*b*	*SE*	95% CI	*t*	*p*	R^2 marginal	R^2 conditional	χ^2	*p*	
Null Model	Intercept	-0.627	0.118	[-0.881, -0.373]	-5.31	0.000 ***	0.000	0.654	-	-	212.100
Full Model 1	Intercept	-0.627	0.118	[-0.881, -0.373]	-5.316	0.000 ***	0.001	0.672	0.020	0.888	214.100
	RDFRE	0.014	0.097	[-0.211, 0.239]	0.141	0.893					
Full Model 2	Intercept	-0.627	0.117	[-0.880, -0.374]	-5.334	0.000 ***	0.002	0.672	0.079	0.779	214.000
	RDFKGL	-0.027	0.096	[-0.251, 0.197]	-0.281	0.788					
Full Model 3	Intercept	-0.627	0.115	[-0.873, -0.380]	-5.453	0.000 ***	0.014	0.672	0.472	0.492	213.600
	DC	-0.065	0.093	[-0.282, 0.151]	-0.701	0.510162					
Full Model 4	Intercept	-0.627	0.102	[-0.841, -0.412]	-6.151	1.38e-05 ***	0.066	0.670	2.772	0.096 .	211.300
	CAREC	-0.144	0.077	[-0.323, 0.034]	-1.882	0.11					
Full Model 5	Intercept	-0.627	0.111	[-0.863, -0.390]	-5.652	6.65e-05 ***	0.031	0.671	1.125	0.289	213.000
	CML2RI	0.098	0.088	[-0.107, 0.303]	1.113	0.309					
Full Model 6	Intercept	-0.627	0.116	[-0.876, -0.377]	-5.397	0.000 ***	0.008	0.672	0.282	0.595	213.800
	SBERT	-0.051	0.095	[-0.271, 0.169]	-0.538	0.610					

Table 5. Summary of the null model and full models for the number of keystrokes (. p<0.1, * p<0.05, ** p<0.01)

The results suggest that readability may not be a good predictor of technical effort. Although the materials varied in terms of reading difficulties, the extent to which the corresponding MT was edited did not differ substantially, indicating a limited impact of readability. Meanwhile, given that the general MT quality was comparable in the current study, it can be speculated that the technical effort is more concerned with the MT quality.

Nevertheless, an observation of estimated coefficients (b) reveals a potential negative relationship between readability and technical effort. To elaborate, the number of keystrokes is likely to decrease when the text becomes more difficult to read. Our results can partially support Jia and Zheng (2022), who consolidated a significant negative impact of source text complexity on the number of editing operations with regards to high-quality MT. We assume that readability has an indirect impact on technical effort,

in that lower readability could lead to increased uncertainties in the MTPE process, and subsequently more restrained editing and reduced number of edits. Additionally, retrospective protocols suggest that such an effect of readability may be modulated by the MT quality. For instance, with regards to ST 5 and ST 6, two texts of relatively lower readability, P23, P28 and P34 all commented that since the MT quality was good in general, they chose to trust and keep the MT when they encountered the things that they were not sure about. However, since the participants had no access to external resources during the tasks, whether this negative relationship still exists without such a restriction requires further investigation.

4.3. Average pause time

Average pause time (APT) is the average time per pause in a session. In line with the previous studies which also investigated the impact of source text features on MTPE cognitive effort, 1000ms was considered as the pause threshold (O'Brien 2006; Jia *et al.* 2019b; Jia and Zheng 2022). As shown in Table 6, only the CAREC-included model differed significantly from the null model (χ^2=5.173, p<0.05) and outperformed other full models in predicting APT (AIC=-1012.7). In addition, the fixed effect of readability was significant in the CAREC-included model (t=2.897, p<0.05).

		Fixed effect					Effect sizes		Log-likelihood ratio test		AIC
	Parameters	*b*	*SE*	95% CI	*t*	*p*	R^2 marginal	R^2 conditional	χ^2	*p*	
Null Model	Intercept	3.184	0.005	[3.173, 3.195]	611.9	<2e-16 ***	0.000	0.623	-	-	-1009.600
Full Model 1	Intercept	3.184	0.005	[3.174, 3.195]	642.992	<2e-16 ***	0.022	0.622	0.931	0.335	-1008.500
	RDFRE	-0.004	0.004	[-0.013, 0.005]	-1.005	0.355					
Full Model 2	Intercept	3.184	0.005	[3.174, 3.195]	647.806	<2e-16 ***	0.025	0.622	1.076	0.300	-1008.600
	RDFKGL	0.004	0.004	[-0.005, 0.013]	1.087	0.32					
Full Model 3	Intercept	3.184	0.005	[3.174, 3.194]	665.096	<2e-16 ***	0.035	0.622	1.598	0.206	-1009.200
	DC	0.005	0.004	[-0.004, 0.013]	1.356	0.225					
Full Model 4	Intercept	3.184	0.004	[3.176, 3.193]	777.471	<2e-16 ***	0.087	0.620	5.173	0.023*	-1012.700
	CAREC	0.008	0.003	[0.001, 0.014]	2.897	0.029 *					
Full Model 5	Intercept	3.184	0.005	[3.175, 3.194]	697.36	<2e-16 ***	0.053	0.622	2.584	0.108	-1010.200
	CML2RI	-0.006	0.003	[-0.014, 0.002]	-1.804	0.123					
Full Model 6	Intercept	3.184	0.005	[3.173, 3.195]	612.394	<2e-16 ***	0.000	0.623	0.015	0.903	-1007.600
	SBERT	-0.001	0.004	[-0.010, 0.009]	-0.122	0.907					

Table 6. Summary of the null model and full models for APT (. p<0.1, * p<0.05, ** p<0.01)

The significant positive relationship between CAREC and APT suggests that participants paused for longer time as the text became harder to read. Vieira (2017) identified three modes of reading involved during MTPE: the first one puts text into working memory for mental processing, the second one concerns specific editing issues, and the third one for revision. Similarly, we believe these three modes can account for the pauses in MTPE. Although the data elicited in the current study has yet to determine which type or types of pauses were prolonged by readability, we assume the impact of readability permeates all three modes of pauses, and particularly the first one, since text processing is one key component of readability (Nahatame

2021). The findings above also indicate that longer APTs are linked with higher cognitive effort. This is somewhat contradictory to Lacruz and Shreve (2014), who claim that APT decreases as cognitive effort increases. However, it should be noted that cognitive effort is itself a complex construct. In Lacruz and Shreve (2014), cognitive effort was indicated by the number of complete editing events, while in this section, the cognitive effort concerns the mental resources for the text understanding and processing.

4.4. Pause-to-word ratio

Pause-to-word ratio (PWR) is calculated by dividing the number of pauses by the number of words in the source text (Lacruz and Shreve 2014). As demonstrated in Table 7, only the difference between the null model and the CAREC-included model approached significance (χ^2=3.091, p<0.1). A marginally significant effect of readability could also be observed in this model (t=-2.02, p<0.1). Meanwhile, it performed a superior prediction of PWR than other formulas (AIC=41.7).

		Fixed effect					Effect sizes		Log-likelihood ratio test		AIC
	Parameters	*b*	*SE*	95% CI	*t*	*p*	R^2 marginal	R^2 conditional	χ^2	*p*	
Null Model	Intercept	-1.149	0.063	[-1.284, -1.014]	-18.24	1.54e-10 ***	0.000	0.564	-	-	42.800
Full Model 1	Intercept	-1.149	0.063	[-1.283, -1.014]	-18.259	1.48e-10 ***	0.001	0.564	0.021	0.884	44.800
	RDFRE	0.007	0.051	[-0.111, 0.126]	0.146	0.889					
Full Model 2	Intercept	-1.149	0.063	[-1.283, -1.014]	-18.325	1.29e-10 ***	0.002	0.564	0.085	0.770	44.700
	RDFKGL	-0.015	0.051	[-0.133, 0.103]	-0.293	0.78					
Full Model 3	Intercept	-1.149	0.062	[-1.281, -1.016]	-18.537	8.29e-11 ***	0.007	0.564	0.294	0.587	44.500
	DC	-0.027	0.050	[-0.143, 0.088]	-0.549	0.603					
Full Model 4	Intercept	-1.149	0.054	[-1.261, -1.036]	-21.35	7.65e-14 ***	0.058	0.562	3.091	0.079 .	41.700
	CAREC	-0.079	0.039	[-0.170, 0.012]	-2.02	0.0918 .					
Full Model 5	Intercept	-1.149	0.061	[-1.280, -1.018]	-18.738	5.4e-11 ***	0.011	0.564	0.492	0.483	44.300
	CML2RI	0.035	0.049	[-0.079, 0.149]	0.716	0.501					
Full Model 6	Intercept	-1.149	0.062	[-1.282, -1.015]	-18.44	1.02e-10 ***	0.005	0.564	0.199	0.656	44.600
	SBERT	-0.023	0.050	[-0.139, 0.094]	-0.45	0.669					

Table 7. Summary of the null model and full models for PWR (. p<0.1, * p<0.05, ** p<0.01)

The results reveal that the predictive power of readability towards PWR is limited. This is not surprising since the pause is only identified between edits, and the number of edits is not mainly decided by the source text readability, as discussed in section 4.2. However, the marginally significant effect of CAREC indicates a potential indirect impact of readability on PWR. The explanation for such an impact is also similar to that in Section 4.2: lower readability may have refrained the subjects from editing, and less edits are linked with lower pause density. Again, this finding is partially consistent with Jia and Zheng (2022), who observed similar but more significant results in the context of high-quality MT.

4.5. Initial Pause

Initial pause (IP) is the pause time before the first edit of the task, which can be considered as the time that translators spent on understanding the text and detecting the mistakes (Cumbreño and Aranberri 2021: 64). Table 8 shows that most full models, except the ones with SBERT and CAREC as fixed effect, differ significantly from the null model. The fixed effect of readability was significant, as assessed by both traditional formulas, namely the RDFRE (t=-3.952, p<0.01) and the RDFKGL (t=2.906, p<0.05), and newer formula, namely the CML2RI (t=-4.644, p<0.01). In the DC-included model, readability had a marginally significant effect (t=2.394, p<0.1). Among the formulas, the CML2RI showed the best performance (AIC=1091.7).

		Fixed effect					Effect sizes		Log-likelihood ratio test		AIC
	Parameters	*b*	*SE*	95% CI	*t*	*p*	R^2 marginal	R^2 conditional	χ^2	*p*	
Null Model	Intercept	15.664	0.767	[14.071, 17.257]	20.42	1.71e-14 ***	0.000	0.591	-	-	1098.500
Full Model 1	Intercept	15.664	0.612	[14.427, 16.900]	25.592	< 2e-16 ***	0.060	0.587	7.398	0.007 **	1093.000
	RDFRE	-1.087	0.275	[-1.737, -0.437]	-3.952	0.009 **					
Full Model 2	Intercept	15.664	0.644	[14.359, 16.969]	24.331	<2e-16 ***	0.049	0.588	5.139	0.023*	1095.300
	RDFKGL	0.985	0.339	[0.187, 1.782]	2.906	0.030 *					
Full Model 3	Intercept	15.664	0.665	[14.311, 17.017]	23.546	<2e-16 ***	0.041	0.588	3.948	0.047*	1096.500
	DC	0.904	0.378	[0.017, 1.791]	2.394	0.057 .					
Full Model 4	Intercept	15.664	0.724	[14.175, 17.152]	21.651	<2e-16 ***	0.019	0.590	1.452	0.228	1099.000
	CAREC	0.607	0.472	[-0.497, 1.711]	1.286	0.248					
Full Model 5	Intercept	15.664	0.598	[14.456, 16.872]	26.196	< 2e-16 ***	0.065	0.586	8.730	0.003 **	1091.700
	CML2RI	-1.127	0.243	[-1.703, -0.552]	-4.644	0.005 **					
Full Model 6	Intercept	15.664	0.719	[14.186, 17.141]	21.79	<2e-16 ***	0.020	0.590	1.622	0.203	1098.800
	SBERT	-0.637	0.465	[-1.724, 0.451]	-1.37	0.223					

Table 8. Summary of the null model and full models for IP (. p<0.1, * p<0.05, ** p<0.01)

The results demonstrate that participants had significantly longer IP for more difficult texts. Since IP can be largely considered as the first mode of reading according to Vieira (2017)'s classification, it can be concluded that readability has a particular impact on the reading for mentally processing the texts, which is consistent with the assumption proposed in Section 4.3.

Of note, shorter IP indicates that participants spent less time on preliminary text processing and error detection, which may lead to the ignorance of MT errors. Although it is not the primary focus of the current study, we would like to provide an example to highlight this issue. Some participants ignored an obvious mistake in ST4, which took the third shortest IP on average (834 ms, normalised by the number of words in source text):

ST: Families Hit with Increase in Cost of Living
MT:美国家庭生活成本上升

ST4 addresses the economic conditions in Britain, while the MT mistakenly added the adjective "美国 (American)" before "家庭 (Families)". Although the following sentence in the source text clearly refers to "British families",

rendering the mistake highly evident, 29% of the students failed to identify this mistranslation. When asked why they did not edit it, participants were surprised to find that they overlooked this error, reporting that they have relaxed their vigilance since the text was not hard to understand and the MT quality was quite good. These findings suggest that translators should exercise caution even when the source text and MT appear to be easily understandable and of good quality, especially as the fluently inadequate errors produced by NMT can still evade detection.

4.6. Subjective cognitive effort

This study applied the adapted NASA Task Load Index (NASA-TLX) (Sun 2012), a multidimensional scale for measuring translation difficulty. We changed the context from human translation to MTPE, and subjects were invited to rate in terms of effort and other five subscales on a 20-point scale. However, for the focus of the current study, only the ratings regarding effort were analysed. As shown in Figure 2, the higher the score, the higher level of effort that participants believed they had exerted.

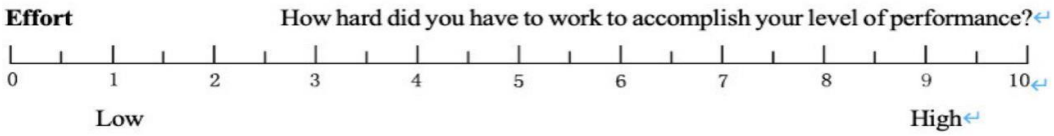

Figure 2. Adapted NASA-TLX subscale

Table 9 suggests that the differences between null model and full models were mostly significant or marginally significant, except the CAREC-included model. The fixed effect of readability was significant in the DC-included model (t=2.977, p<0.01), the CML2RI-included model (t=-2.686, p<0.05), and the SBERT-included model (t=-2.57, p<0.05). A marginally significant effect of readability can also be observed when assessed by RDFRE (t=-2.5, p<0.1) and RDFKGL (t=2.365, p<0.1). The DC performed the best prediction of subjective cognitive effort (AIC=692.7).

		Fixed effect					Effect sizes		Log-likelihood ratio test		AIC
	Parameters	*b*	*SE*	95% CI	*t*	*p*	R^2 marginal	R^2 conditional	χ^2	*p*	
Null Model	Intercept	5.387	0.276	[4.831, 5.944]	19.54	<2e-16 ***	0.000	0.661	-	-	698.200
Full Model 1	Intercept	5.387	0.261	[4.860, 5.914]	20.63	<2e-16 ***	0.012	0.660	4.111	0.043*	696.000
	RDFRE	-0.202	0.081	[-0.394, -0.010]	-2.5	0.053 .					
Full Model 2	Intercept	5.387	0.262	[4.859, 5.916]	20.572	<2e-16 ***	0.012	0.660	3.805	0.051 .	696.400
	RDFKGL	0.196	0.083	[-0.001, 0.394]	2.365	0.0625 .					
Full Model 3	Intercept	5.387	0.259	[4.865, 5.910]	20.797	< 2e-16 ***	0.015	0.660	5.421	0.020*	692.700
	DC	0.220	0.074	[0.074, 0.365]	2.977	0.003 **					
Full Model 4	Intercept	5.387	0.271	[4.840, 5.934]	19.86	<2e-16 ***	0.004	0.661	0.992	0.319	699.200
	CAREC	0.113	0.109	[-0.144, 0.371]	1.044	0.342					
Full Model 5	Intercept	5.387	0.260	[4.862, 5.912]	20.712	<2e-16 ***	0.013	0.660	4.530	0.033*	695.600
	CML2RI	-0.208	0.077	[-0.392, -0.023]	-2.686	0.042 *					
Full Model 6	Intercept	5.387	0.261	[4.861, 5.913]	20.66	<2e-16 ***	0.013	0.660	4.270	0.039*	695.900
	SBERT	-0.204	0.079	[-0.393, -0.015]	-2.57	0.048*					

Table 9. Summary of the null model and full models for subjective cognitive effort (. p<0.1, * p<0.05, ** p<0.01)

The results indicate that participants reported higher cognitive effort with more difficult texts. Among the formulas which significantly predicted subjective cognitive effort, two formulas, namely DC and CML2RI, comprise a similar feature. DC focuses on the percentage of less common words, and CML2RI considers word frequency as one of its major components. Therefore, we assume that infrequent word is one of the key factors that influences participants' perception of MTPE effort. This assumption is in accordance with the retrospective protocols, in which all the participants mentioned unfamiliar words regarding their difficulties during MTPE. Therefore, it can be concluded that readability is a good predictor of subjective cognitive effort, especially when it considers infrequent words.

		Total time	Total number of keystrokes	Average pause time	Pause to word ratio	Initial pause	Subjective cognitive effort
Traditional readability formulas	RDFRE	. √				**	.
	RDFKGL					*	.
	DC					.	** √
Newer readability formulas	CAREC		√	* √	. √		
	CML2RI					** √	*
	SBERT						*

Table 10. Fixed effect of readability (as measured by different formulas) on MTPE effort (. p<0.1, * p<0.05, ** p<0.01, √ the best performance)

Finally, a summary of the fixed effect of readability, measured by different formulas, on MTPE effort is presented in table 10. The formula which shows the best performance in predicting each effort indicator is also marked.

5. Predicting post-editing time with reading-related source text linguistic features

MTPE time has proved to be an economical and convenient effort indicator in the translation industry (Koponen *et al*. 2012). However, our results suggest that the ability of readability formulas in predicting total MTPE time is rather limited. Since previous findings indicate that some linguistic features of the source text have a high correlation with MTPE time (Specia 2011; Green *et al.* 2013; Vieira 2014), the study also explored the predictive power of fine-grained reading-related linguistic features. Specifically, we investigated the relationship between twelve referential cohesion indices in Coh-Metrix and total time by fitting twelve LMER models separately, in which each index being the independent variable, total time the dependent variable, and participants and texts the random effects. Three indices, namely CRFAOa (t=-2.626, p<0.05), CRFCWO1 (t=-4.617, p<0.01) and CRFCWOa (t=-2.749, p<0.05), significantly predict the time.

The CRFAOa represents the global argument overlap, which is the proportion of all possible sentence pairs that share one or more common nouns or pronouns. The CRFCWO1 and CRFCWOa measure content word overlap locally and globally. In other words, the former estimates the proportion of content words that are the same between adjacent sentences, while the latter assesses the overlap between all possible pairs of sentences (Graesser *et al.* 2011). The more words overlap, the higher the indices, and the easier the text is likely to be to understand. The CRFAOa, CRFCWO1 and CRFCWOa of 6 texts are listed in Table 11.

	ST1	ST2	ST3	ST4	ST5	ST6
CRFAOa	0.639	0.436	0.844	0.524	0.393	0.81
CRFCWO1	0.132	0.105	0.134	0.114	0.087	0.119
CRFCWOa	0.116	0.071	0.139	0.059	0.06	0.103

Table 11. The CRFAOa, CRFCWO1 and CRFCWOa of texts

The LMER models suggest that these three indices have a significant negative impact on total time, i.e. total time decreases with the increase of argument/content word overlap. We assume that the repetition of words has facilitated the information processing stage. Our results contradict those of Vieira (2014), who reported that higher type-token ratio, i.e. less repetition of words, led to lower cognitive effort. However, it should be stressed that our results are confined to high-quality MT, while Vieira (2014)'s study covered MT of various quality. Accordingly, the "lack of fluency" problems arising from words repetition might have been reduced in the current study.

6. Conclusion

Given that reading is a crucial aspect of MTPE, the study investigated the impact of source text readability on MTPE effort. The quantitative and qualitative data show that readability has a significant effect on cognitive effort, particularly on IP and subjective cognitive effort. The impact of readability on temporal effort and technical effort appears to be limited and indirect, possibly due to the assistance of high-quality NMT. Of note, while the impact of readability on MTPE effort can be statistically significant, the effect sizes of the models suggest that this impact may be relatively weak.

Regarding the predictive power of readability formulas, the results indicate that they can predict MTPE effort to a certain degree. Nevertheless, no single formula was able to predict all the effort indicators, highlighting the need to combine different formulas in effort prediction. Newer formulas, particularly the CAREC, outperformed traditional formulas in most instances, which may be explained by the fact that the former consider deeper linguistic features. Our findings also reveal that it is promising to adopt

formulas which concern translation-related linguistic features such as translation entropy to automatically predict MTPE effort in the future.

In addition to the readability formulas, the study also applied fine-grained reading-related linguistic features to predict MTPE time. Referential cohesion indices, such as content word overlap, were confirmed to be effective predictors. Therefore, it is recommended that translators utilise these automatically generated features to obtain an estimation of MTPE time, and that future QE models can take such features into account.

There is no denying that some limitations exist in the current study. Firstly, the number and variety of subjects and texts could be expanded to provide more evidence for the translation industry. Secondly, eye-tracking data would enable a finer-grained analysis regarding the impact of readability on MTPE effort. It is also true that controlling the MT quality may have limited the effect of readability on MTPE effort. However, given the limited time and effort available, the current study represents the best efforts of the researchers. For future research, combinations of multiple MT quality levels and readability scores would certainly provide a more comprehensive picture. Finally, the effectiveness of the prediction models developed in the study should be tested in the real MTPE settings. These limitations and suggestions for future research will be taken into account to yield more comprehensive and generalisable results in future studies.

Acknowledgements

This research was supported by the National Social Science Fund of China ("神经网络机器翻译质量提升研究"/"A Study on Quality Improvement of Neural Machine Translation", Grant reference: 22BYY042). The authors would like to thank our participants and evaluators for their valuable time. Heartfelt gratitude is extended to the editors, the anonymous reviewers, and Dr. Jiajun Qian for their constructive comments and insightful feedback.

References

- **Aziz, Wilker, Koponen, Maarit and Lucia Specia** (2014). "Sub-sentence level analysis of machine translation post-editing effort." Sharon O'Brien, Laura Winther Balling, Michael Carl, Michel Simard and Lucia Specia (eds) (2014). *Post-editing of machine translation: Processes and applications*. Cambridge: Cambridge Scholars Publishing,170-199.

- **Bartoń, Kamil** (2023). MuMIn: Multi-Model Inference. R package version 1.47.5, <https://CRAN.R-project.org/package=MuMIn>.

- **Bates, Douglas, Martin Mächler, Ben Bolker and Steve Walker** (2015). "Fitting Linear Mixed-Effects Models Using lme4." *Journal of Statistical Software* 67(1), 1–48.

- **Campbell, Stuart** (1999). A cognitive approach to source text difficulty in translation. *Target* 11(1), 33-63.

- **Carl, Michael** (2012). "Translog-II: A program for recording user activity data for empirical reading and writing research." Nicoletta Calzolari *et al.* (eds) (2012). *Proceedings of the Eighth International Conference on Language Resources and Evaluation*. Eurpean Language Resources Association, 4108-4112.

- **Carl, Michael, Schaeffer, Moritz and Srinivas Bangalore** (2016). "The CRITT translation process research database." Michael Carl, Srinivas Bangalore and Moritz Schaeffer (eds) (2016). *New directions in empirical translation process research.* Cham: Springer, 13-54.

- **Castilho, Sheila, Joss Moorkens, Federico Gaspari, Iacer Calixto, John Tinsley and Andy Way** (2017). "Is neural machine translation the new state of the art?" *The Prague Bulletin of Mathematical Linguistics* 108, 109-120.

- **Choi, Joon S., and Scott A. Crossley** (2022). "Advances in Readability Research: A New Readability Web App for English." Paper presented at the *2022 International Conference on Advanced Learning Technologies* (Bucharest, Romania, 01-04 July 2022).

- **Crossley, Scott A., Greenfield, Jerry, and Danielle S. McNamara** (2008). Assessing text readability using cognitively based indices. *Tesol Quarterly* 42(3), 475-493.

- **Crossley, Scott A., Allen, David B. and Danielle S. McNamara** (2011). "Text readability and intuitive simplification: A comparison of readability formulas." *Reading in a foreign language* 23(1), 84-101.

- **Crossley, Scott A., Skalicky, Stephen, and Mihai Dascalu** (2019). "Moving beyond classic readability formulas: New methods and new models" *Journal of Research in Reading* 42(3-4), 541-561.

- **Cumbreño, Cristina and Nora Aranberri** (2021). "What do you say? Comparison of metrics for post-editing effort." Michael Carl (ed.) (2021). *Explorations in Empirical Translation Process Research.* Cham: Springer, 57-59.

- **Daems, Joke, Sonia Vandepitte, Robert J. Hartsuiker and Lieve Macken** (2017). "Identifying the machine translation error types with the greatest impact on post-editing effort." *Frontiers in Psychology* 8, Article 1282.

- **Dai, Guangrong and Siqi Liu** (2023). "Neural machine translation: Advancements and challenges" 外语教学（*Foreign Language Education*）1, 82-89.

- **Dale, Edgar, and Jeanne S. Chall** (1948). "A formula for predicting readability: Instructions." *Educational research bulletin* 27, 37-54.

- **Flesch, Rudolph** (1948)**.** "A new readability yardstick." *Journal of applied psychology* 32, 221-233.

- **Fox John and Sandford Weisberg** (2019). *An R Companion to Applied Regression*, *Third edition*. Thousand Oaks, CA: Sage publication.

- **Graesser, Aurthur C. , McNamara, Danielle S. and Jonna M. Kulikowich** (2011). "Coh-Metrix: Providing multilevel analyses of text characteristics." *Educational researcher* 40(5), 223-234.

- **Green, Spence, Heer, Jeffrey and Christopher D. Manning** (2013). "The efficacy of human post-editing for language translation." *Proceedings of the SIGCHI conference*

on human factors in computing systems. Association for Computing Machinery, 439-448.

- **Herbig, Nico, Santanu Pal, Mihaela Vela, Antonio Krüger and Josef van Genabith** (2019). "Multi-modal indicators for estimating perceived cognitive load in post-editing of machine translation." *Machine Translation* 33(1), 91-115.

- **Hvelplund, Kristian T.** (2011). *Allocation of cognitive resources in translation: An eye-tracking and key-logging study*. PhD thesis. Copenhagen Business School.

- **ISO 18587** (2017). *Translation services — Post-editing of machine translation output — Requirements*. Geneva: International Organization for Standardization

- **Jia, Yanfang, Carl, Michael and Xiangling Wang** (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.

- **Jia, Yanfang, Carl, Michael and Xiangling Wang** (2019b). "Post-editing neural machine translation versus phrase-based machine translation for English–Chinese." *Machine Translation* 33(1), 9-29.

- **Jia, Yanfang and Binghan Zheng** (2022). "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." *Across Languages and Cultures* 23(1), 36-55.

- **Kincaid, J. Peter, Robert P. Fishburne Jr., Richard L. Rogers and Brad S. Chissom** (1975). *Derivation Of New Readability Formulas (Automated Readability Index, Fog Count And Flesch Reading Ease Formula) For Navy Enlisted Personnel.* Millington, TN: Naval Technical Training Command, Research Branch.

- **Koglin, Arlene** (2015). "An empirical investigation of cognitive effort required to post-edit machine translated metaphors compared to the translation of metaphors." *Translation & Interpreting* 7, 126-141.

- **Koponen, Maarit, Luciana Ramos, Wilker Aziz and Lucia Specia** (2012). "Post-editing time as a measure of cognitive effort." Sharon O'Brien, Michel Simard and Lucia Specia (eds) (2012). *Workshop on Post-editing Technology and Practice*. Association for Machine Translation in the Americas.

- **Koponen, Maarit, Brian Mossop, Isabelle S. Robert and Giovanna Scocchera** (eds) (2020). *Translation Revision and Post-editing: Industry Practices and Cognitive Processes*. London: Routledge.

- **Koponen, Maarit and Leena Salmi** (2015). "On the correctness of machine translation: A machine translation post-editing task." *The Journal of Specialised Translation* 23, 118-136.

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

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

- **Lacruz, Isabel and Gregory M. Shreve** (2014). "Pauses and cognitive effort in post-editing." Sharon O'Brien *et al.* (eds) (2014). *Post-editing of machine translation: Processes and applications*. Cambridge: Cambridge Scholars Publishing, 246-272.

- **Lemhöfer, Kristin and Mirjam Broersma** (2012). "Introducing LexTALE: A quick and valid Lexical Test for Advanced Learners of English." *Behavior Research Methods* 44, 325-343.

- **Li, Mei** (2021). "Impact of source texts on translators in machine translation post-editing output." 外语教学（*Foreign Language Education*）4, 94-99.

- **Lu, Zhi and Juan Sun** (2018). "An eye-tracking study of cognitive processing in human translation and post-editing." 外语教学与研究*(Foreign Language Teaching and Research)* 5, 760-769.

- **Marg, Lena** (2016). "The Trials and Tribulations of Predicting post-editing productivity." Nicoletta Calzolari *et al.* (eds) (2016). *Proceedings of the Tenth International Conference on Language Resources and Evaluation*. European Language Resource Association: 23-26.

- **McNamara, Danielle S, Arthur C. Graesser, Philip M. McCarthy and Zhiqiang Cai** (2014). Automated Evaluation of Text and Discourse with Coh-Metrix. Cambridge: Cambridge University Press.

- **Nahatame, Shingo** (2021). "Text Readability and Processing Effort in Second Language Reading: A Computational and Eye-Tracking Investigation." *Language learning* 71 (4), 1004-1043.

- **O'Brien, Sharon** (2005). "Methodologies for Measuring the Correlations between Post-Editing Effort and Machine Translatability." *Machine Translation* 19(1), 37-58.

- **O'Brien, Sharon** (2006). "Pauses as indicators of cognitive effort in post-editing machine translation output." *Across Languages and Cultures* 7(1), 1-21.

- **O'Brien, Sharon** (2007). "An empirical investigation of temporal and technical post-editing effort." *Translation and Interpreting Studies." The Journal of the American Translation and Interpreting Studies Association* 2(1), 83-136.

- **O'Brien, Sharon** (2011). "Towards predicting post-editing productivity." *Machine Translation* 25(3), 197-215.

- **Popović, Maja, Arle Lommel, Aljoscha Burchardt, Eleftherios Avramidis and Hans Uszkoreit** (2014). "Relations between different types of post-editing operations, cognitive effort and temporal effort". Mauro Cettolo, Marcello Federico, Lucia Specia and Andy Way (eds) (2014). *Proceedings of the 17th Annual conference of the European Association for Machine Translation*. European Association for Machine Translation: 191-198.

- **Popović, Maja** (2020). "Relations between comprehensibility and adequacy errors in machine translation output." Raquel Fernández and Tal Linzen (eds) (2020). *Proceedings of the 24th Conference on Computational Natural Language Learning*. Association for Computational Linguistics: 256-264.

- **Qian, Jiajun, Weiqing Xiao, Yan Li and Xia Xiang** (2022). "Impact of neural machine translation error types on translators' allocation of attentional resources:

Evidence from eye-movement data". 外语教学与研究*(Foreign Language Teaching and Research)* 5, 750-761.

- **R Core Team** (2022). "R: A language and environment for statistical computing." R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.

- **Reimers, Nils and Iryna Gurevych** (2019). "Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks." *https://arxiv.org/abs/1908.10084*

- **Specia, Lucia** (2011). "Exploiting objective annotations for minimising translation post-editing effort." Mikel L. Forcada, Heidi Depraetere and Vincent Vandeghinste (eds) (2011). *Proceedings of the 15th Annual conference of the European Association for Machine Translation*. European Association for Machine Translation

- **Specia, Lucia and Shah Kashif** (2018). "Machine translation quality estimation: Applications and future perspectives." Joss Moorkens, Sheila Castilho, Federico Gaspari and Stephen Doherty (eds) (2018). *Translation Quality Assessment*. Cham: Springer, 201-235.

- **Sun, Sanjun** (2012). *Measuring difficulty in English-Chinese translation: Towards a general model of translation difficulty*. PhD thesis. Kent State University.

- **Sun, Sanjun**. (2019). "Measuring Difficulty in Translation and Post-editing: A Review." Li Defeng, Lei Victoria Lai Cheng and He Yuanjian (eds) (2019). *Researching Cognitive Processes of Translation*. Singapore: Springer, 139-168.

- **Tatsumi, Midori** (2009). "Correlation between automatic evaluation metric scores, post-editing speed, and some other factors." Paper presented at the *Proceedings of Machine Translation Summit XII: Posters* (Ottawa, Canada, 26-30 August 2009).

- **Tatsumi, Midori and Johann Roturier** (2010). "Source text characteristics and technical and temporal post-editing effort: what is their relationship." Ventsislav Zhechev (ed.) (2010). Proceedings of the Second Joint EM+/CNGL Workshop: Bringing MT to the User: Research on Integrating MT in the Translation Industry. Association for Machine Translation in the Americas: 43-52.

- **TAUS** (2019). *Adequacy/Fluency Guidelines.* https://info.taus.net/quality-evaluation-using-adequacy-and-fluency-approaches (consulted 19.12.2022)

- **Tezcan, Arda, Hoste, Veronique and Lieve Macken** (2019). "Estimating post-editing time using a gold-standard set of machine translation errors." *Computer Speech & Language* 55, 120-144.

- **Toral, Antonio, Wieling, Martijin and Andy Way** (2018). "Post-editing Effort of a Novel With Statistical and Neural Machine Translation." *Frontiers in Digital Humanities* 5, Article 9.

- **Vieira, Lucas N.** (2014). "Indices of cognitive effort in machine translation post-editing." *Machine Translation 28*(3), 187-216.

- **Vieira, Lucas N.** (2017). "Cognitive effort and different task foci in post-editing of machine translation: A think-aloud study." *Across Languages and Cultures* 18(1), 79-105.
- **Wu, Shiyu and Zheng Ma** (2020). "How is Chinese reading affected by under-specification and over-specification? Evidence from self-paced reading experiments." *Journal of Pragmatics 155*, 213-233.

Biography

Guangrong DAI (Ph.D) is a Professor at the School of Interpreting and Translation Studies, Guangdong University of Foreign Studies, China. His main research interests cover corpus translation studies, NLP and MTPE. He is also interested in new technologies and their affordances as well as pedagogical theories that facilitate the teaching of those technologies. His blog is http://blog.sciencenet.cn/u/carldy.

ORCID: 0000-0001-7785-8484
E-mail: carldy@163.com

Siqi Liu is an MA student at the School of Interpreting & Translation Studies, Guangdong University of Foreign Studies, China. Her research interests include translation/post-editing process research and corpus-assisted translation teaching. She is also passionate about doing translation education research from interdisciplinary perspectives, for instance, using psychology.

ORCID: 0000-0002-1856-3376
E-mail: 20211210023@gdufs.edu.cn

Notes

[1] The CATTI Level 3 (translator) certificate recipient should be able to complete general translation work. For Level-2 certificate recipient, s/he should be able to independently undertake translation work with a certain degree of difficulty in this speciality (http://www.catticenter.com/cattiksjj/1848).