Cognitive effort in human translation and machine translation post-editing processes
A holistic and phased view

Yu Wang and Ali Jalalian Daghigh
Faculty of Languages and Linguistics, Universiti Malaya

Due to the inaccuracies of Machine Translation (MT), Post-Editing (PE) is inevitable. This poses questions over whether human effort to polish an MT is worthwhile or whether it would be more efficient to translate manually. However, to date, fewer attempts have been made to compare the cognitive effort in the PE process and the sub-phases (orientation, drafting, and revision) of PE with that of Human Translation (HT). To fill this gap, the current study aims to investigate and compare cognitive effort in HT and PE processes in translation from Chinese to English. Data were collected via eye-tracking and keyboard-logging approaches from 25 participants recruited to fulfil three HT and three PE tasks respectively. The comparison of cognitive effort was made from the processes of HT and PE, and their different sub-phases. The study reveals a significant difference in cognitive effort, orientation duration, and drafting duration between HT and PE.

**Keywords:** machine translation, post-editing, cognitive effort, translation process

1. **Introduction**

The emergence of Neural Machine Translation (NMT) has changed the traditional workflow and generates fast translation through web browsers and mouse clicks. However, translation machines rarely produce perfectly accurate translations, and therefore, it is still vital for a human translator to edit and polish the machine-translated text. The position of post-editing (PE) in the translation industry poses questions over human-machine interaction. Many translators have asserted that the texts translated by machine programs are far from perfect and it would be quicker to translate manually from scratch (Durban 2011). Generally, humans invented machine translation (MT) to make the translation task easier,
with less effort: therefore, machine translation is supposed to be a substitute for manual labour. Yet, human effort is still required to prevent blunders and mistranslations caused by machines. Hence, a comparison relating to the human cognitive effort should be made in order to shed light on the difference between human translation (HT) and PE. In addition, the changing role of the human translator under the scenario of MT also demands a study to better understand how the translator’s cognitive effort differs in the PE process compared to translating from scratch. While HT demands the translator to translates from scratch, from the source language comprehension to target language reproduction; technically, when the translator is conducting the PE task, he or she is more of a post-editor than a translator. Presumably, when acting as a post-editor, the translator is supposed to handle a butcher’s cleaver skilfully, since his or her task is to render the machine translated texts rather than to deal with the two languages. Thus, research on how much effort the translator makes in PE compared with traditional HT is necessary to understand the difference between the HT and PE processes.

While limited in number, although there have been attempts by researchers to compare HT and PE (Guerberof-Arenas 2014; Läubli et al. 2019; Zhechev 2012), the comparison has been made mainly from the perspectives of speed and quality. While speed is the easiest variable to measure (Krings 2001), and quality is also relatively easy to measure (Sanchez-Torron and Koehn 2016), cognitive effort cannot be gauged directly. This accounts for one of the main limitations into researching the cognitive effort in the HT and PE processes. Studies show that PE saves time compared to HT (Plitt and Masselot 2010; Sousa, Aziz, and Specia 2011). However, a faster PE process may not necessarily indicate lower cognitive effort, since speed is more of a temporal than a cognitive indicator (Krings 2001). The increase of productivity depends on such factors as the difficulty of the experimental text and the translators’ ability and language proficiency (Koponen 2016). Thus, the key is to investigate how much cognitive effort a translator exerts in each translation to make the PE process differ from HT. Just as Krings (2001, 61) points out, speed and productivity do not inform “how post-editing occurs as a process, how it is distinguished from conventional translation, and what demands it makes on post-editors.” By measuring cognitive effort, it is possible to learn about the effort the human puts into processing and producing language (Sjørup 2013).

The current research compares the cognitive effort of the three sub-phases between HT and PE, which has not been closely examined by previous studies. Generally, the HT process consists of three phases: the reading of the source text, the reproduction of the target language, and the revision of the final target text – a division which Jakobsen (2002) set as orientation, drafting, and revision. In PE, although the MT engine has already produced the output of the target text,
a translator may still need to read the source text and then edit or revise the machine-translated text to achieve a perfect target text. In this case, the three sub-phases also apply to the PE process. Hence, this study intends to examine the three sub-phases and investigate how translators use their cognitive effort in different stages of HT and PE, as well as how cognitive effort in these stages varies to make the whole process of PE differ from that of HT.

2. Related research

2.1 Cognitive effort in the human translation process

Cognitive effort is such a recurring topic in Translation Process Research that some researchers even consider Translation Process Research as Cognitive Translatology (Martín 2010), which refers to the various studies on the cognitive process of translation. Difficult as it is for researchers to gauge the cognitive effort directly, the application of advanced technologies such as eye-tracking and keyboard-logging has opened a window into the translator’s inner activities, which enables the researchers to explore how translators employ their effort to make the translation happen. Eye-tracking records the translator’s eye movements, reflecting the cognitive process based on the Eye-Mind Hypothesis that what is being fixed on by the eye is also being cognitively processed by the mind simultaneously (Just and Carpenter 1980). Keyboard-logging software registers the language production activities, such as deletion, insertion, and substitution of words or phrases in real-time: that is, the typing and editing of texts. Since the introduction of cognitive approaches into translation process research, eye-tracking and keyboard-logging have been widely used in the area of Human Translation Process Research with different focuses, ranging from translation units (Carl and Kay 2011) to metaphor translation (Koglin 2015; Koglin and Cunha 2019), directionality (Ferreira et al. 2016; Whyatt 2019), and searching of online resources (Whyatt, Witczak, and Tomczak 2021), to name a few.

Despite various studies into the HT process, little work has been done to compare the cognitive effort in HT with that in PE, let alone the comparison of cognitive effort in the sub-phases between HT and PE. The cognitive effort in HT is measured mainly by fixation, including fixation duration, average fixation duration, and fixation count. These parameters reflect the level of difficulty of processing and extraction of information (Just and Carpenter 1976); therefore, they are used as indicators of cognitive effort in the translation process. Sjørup (2013) investigated metaphor in the translation process using eye-tracking and keyboard logging. Comparing the eye data in metaphor areas with that in non-
metaphor areas, the study reported a significantly longer fixation duration and first-pass fixation time in metaphor areas, indicating increased effort in translating metaphors. Another study, conducted by Dragsted (2010), compared the fixation counts between novice translators and professional translators and found that professional translators had a significantly lower number of fixations than their counterparts, showing that novice translators expended a higher level of cognitive effort in carrying out the translation task. Ferreira et al. (2016) carried out an experiment using eye-tracking to explore the cognitive effort with translation directionality and found a significantly longer average fixation duration in indirect translation compared with direct translation, suggesting an allocation of more effort in the indirect translation process. Whyatt et al. (2016) collected data through eye-tracking and keyboard-logging to study the cognitive effort in translation and paraphrasing. They reported a significantly longer average fixation duration in the source text area than in the target text, suggesting a more demanding process in dealing with the source text in both translating and paraphrasing. These studies confirmed that fixation can reflect different cognition behaviours and cognitive changes. However, these indicators, being mainly derived from eye movement, hardly provide a deep understanding of the cognitive effort involved in the translation process. The current study, therefore, in addition to adopting fixation duration and fixation count, combines other indicators, including keyboard activities and pauses, to compare the cognitive effort in HT and PE processes. As well as investigating HT processes, researchers have also devoted themselves to studying PE process, for which the following section presents a review.

2.2 PE effort

Studies on PE effort have mainly focused on temporal and technical aspects (e.g., Alves et al. 2016; Carl et al. 2011; O’Brien 2007; Koponen 2012; Guerberof-Arenas 2012). Temporal effort is gauged using PE time: that is, the total time duration required by the PE task. Technical effort is mainly examined using Translation Edit Rate (TER) or keystroke activities. Cognitive effort is examined with the aid of some alternative approaches, mostly involving eye-tracking, either alone or combined with other methods. Koponen et al. (2012) investigated the correlation between temporal and cognitive effort. By selecting sentences with similar lengths and editing distances, they compared the editing time of the sentences and found that shorter PE times were related to errors of low cognitive ranking. The results indicated that the temporal indicator was an effective way to study cognitive effort.

Similarly, Cumbreño and Aranberri (2019) studied the error types in PE belonging to different levels of cognitive difficulty and correlated them with time,
keyboard data, and perceived cognitive effort. They found that PE time was similar among all the error types but was different in terms of keystrokes and perceived cognitive effort, suggesting that the result of each dimension separately only offers a limited understanding of the effort in the PE process.

Vieira (2014) combined eye-tracking with a subjective rating scale to predict cognitive effort in PE. By analysing the individual and textual features, he found that the Meteor metric (a score assessing how a machine translation matches a human reference translation) had a significant correlation with all the parameters of cognitive effort, suggesting that Meteor had higher predictive power for cognitive effort in PE.

Alves et al. (2016) combined eye-tracking with keyboard logging to investigate the impact of interactive machine translation on the PE effort under experimental conditions and the correlation of PE effort with Translation Edit Rate (TER) scores. They compared the PE process with interactive machine translation with the process without interactive MT and observed a significantly lower edit rate relating to fixation data and edit rate when using the interactive mode, suggesting low cognitive effort in the PE process.

The above studies focused on the cognitive effort in the PE process and offered various approaches to measure cognitive effort. However, based on the three-part division of PE effort proposed by Krings (2001) – namely temporal, technical, and cognitive effort – the studies cannot fully account for the effort in the whole PE process in general, as only one or two aspects were considered. We hypothesize that the three dimensions also apply to HT, since it also involves temporal effort (time duration to finish the task), cognitive effort (mental effort to do the translation), and technical effort (the editing of the translation draft) (Krings 2001). Moreover, hardly any comparisons are made between all three of these dimensions. The current study, therefore, aims to compare the effort between HT and PE processes from all three dimensions to tap into the cognitive effort in both HT and PE.

2.3 Pause

In addition to the above measurements, a handful of studies have also used pause as an indicator to measure cognitive effort in the translation process. O’Brien (2006) was the first researcher to hypothesize that pause could be used as an indicator to study PE. O’Brien proposed the Pause Ratio measurement, which was calculated by dividing the total duration of a sentence by the duration of the pauses of the sentence. She used quantitative methods to compare the PE of two sets of sentences: sentences with Controlled Language and sentences without Controlled Language. However, no significant difference in cognitive effort was found between the two sentence types when Pause Ratio was used as a measurement.
Based on the findings from O’Brien, Lacruz et al. (2012) proposed another metric, namely the Average Pause Ratio (APR), to measure pauses in PE. In addition to pause duration, APR also took into consideration word length, which was computed by dividing the average length of a pause by the average time needed to edit a word. The APR metric was tested with different pause patterns and found to be higher for less cognitively demanding segments than for those that were more cognitively demanding (Lacruz, Shreve, and Angelone 2012). APR was further confirmed with another metric, namely the Pause-to-Word Ratio (PWR), which is computed by dividing the total number of pauses in a segment by the total word counts in that segment (Lacruz and Shreve 2014). The two metrics were correlated with cognitive effort in PE: an increase in APR accompanied a decrease in cognitive effort. In contrast, an increase in PWR accompanied an increase in cognitive effort (Lacruz et al. 2016).

Although APR and PWR were initially adopted to gauge the cognitive effort in PE, we assume that they also apply to translation, as the translation process involves many pauses when translators are typing. To date, two researchers have studied the pauses in HT and PE processes; however, they used only one of the above metrics. Vanory et al. (2019) used APR as one feature to investigate the correlation between translation process data and translation product data. They reported that APR as a process feature was correlated with the number of errors, the translation entropy, and the syntactic equivalence in the final translation product. Later, Jia et al. (2019) took PWR as a parameter to compare the cognitive effort between HT and PE in English–to–Chinese translation, and concluded that the translation task and text type had a significant influence on PWR. Since both APR and PWR have been shown to be more valid indicators to examine cognitive effort, compared to Pause Ratio, both APR and PWR were adopted in this study, together with other indicators, to investigate the cognitive effort in HT and PE.

In sum, the current study contributes to the body of literature in three main ways. First, a series of comprehensive indicators are adopted to examine the effort in the translation process, in contrast to previous studies, which have considered only one or two aspects of cognitive effort. Second, all three sub-phases (orientation, drafting, and revision) are investigated to gain a more in depth understanding of the differences between HT and PE, which was neglected by the previous studies. Last but not least, this research examines the Chinese–to–English translation direction, which has received little attention in Translation Process Research. To address the research gaps, this study is designed to answer the following two questions:

1. To what degree does the cognitive effort in the whole process of HT differ from that of PE in Chinese–to–English translation?
2. How does the cognitive effort in sub-phases of HT differ from that of PE in Chinese–to–English translation?

3. Method

3.1 Research design

An experiment was conducted to investigate the cognitive effort in the HT and PE processes. The experiment obtained ethical approval from Universiti Malaya. Empirical data were collected from eye-tracking and keyboard logging to investigate the cognitive effort in the translation process. The experimental facilities were a laptop with Translog-II and a portable Tobii Pro Fusion eye-tracker (Figure 1) with a sampling frequency of 250 Hz. Translog-II divided the laptop screen in half (Figure 2). The upper half was used to present the source texts, and the lower half was used to perform the manual translation and PE tasks. Before the experiment, five-point calibration was made to ensure the data quality.

Figure 1. Tobii pro fusion with laptop
(source: https://www.tobiipro.com)

Figure 2. Translog II interface
3.2 Participants

A total of 25 professional translators from translation agencies in China were recruited for the experiment. The homogeneity of their language proficiency and translation ability was assured. The translators all held BA degrees in translation or related fields such as the English Language. All of them were native Chinese speakers and had passed the TEM (Test for English Major) Band 8, which is the highest-level English proficiency test in China. Participants all held CATTI (China Accreditation Test for Translators and Interpreters), which is considered the requirement to practice translation in China. They had at least three years’ full-time translation experience. Since the translation agencies did not offer PE services, none of them had professional PE experience. Participants were recruited through advertisements on social media. They signed a letter of consent after recruitment. Participants were informed in advance that their participation was anonymous. No time constraints were imposed for the task, although it was recommended to participants that they should try to finish all the six texts within 2.5 hours to ensure the validity of the experiment (see Sun and Shreve 2014). Since participants’ consultation of online resources yields fixation data beyond the source text and target text areas of Translog-II, which undermines the data quality (Hvelplund 2011), translation aids such as online dictionaries and online inquiry were not provided.

3.3 Materials

A total of six texts were used as the experimental materials (Table 1). Three Chinese texts (H1, H2, H3) were used for HT, and another three texts (P1, P2, P3) were used for PE. The texts for PE were translated by Google Translate in November 2021. The length and readability of the texts were controlled before the experiment. Each text consisted of approximately 145 Chinese characters, which was the appropriate length to fit into the upper half of the Translog-II window without scrolling. The readability of these texts was tested via the Chinese Coh-Metrix¹ system and CRIE² (Sung et al. 2016). The word complexity, the lexical complexity, and the average sentence length were calculated, which proved that the texts used for HT and PE had similar readability levels. A Latin Square Design was used to balance the sequential effect of the experiment material on the participants (Bradley 1958; Richardson 2018).

¹ Website: http://210.240.188.161/Chinese_CohMetrix/index.php
² China Readability Index Explorer, CRIE 3.0.
http:// http://www.chinesereadability.net/CRIE/?LANG=CHT
Table 1. Experimental texts

<table>
<thead>
<tr>
<th>Text</th>
<th>Characters</th>
<th>PWC</th>
<th>PLC</th>
<th>ASL</th>
<th>RG</th>
</tr>
</thead>
<tbody>
<tr>
<td>H1</td>
<td>142</td>
<td>27.81%</td>
<td>47.65%</td>
<td>8.89</td>
<td>3</td>
</tr>
<tr>
<td>H2</td>
<td>146</td>
<td>30.15%</td>
<td>48.27%</td>
<td>7.08</td>
<td>3</td>
</tr>
<tr>
<td>H3</td>
<td>146</td>
<td>27.82%</td>
<td>47.78%</td>
<td>8.22</td>
<td>4</td>
</tr>
<tr>
<td>P1</td>
<td>148</td>
<td>30.76%</td>
<td>45.63%</td>
<td>9.11</td>
<td>3</td>
</tr>
<tr>
<td>P2</td>
<td>144</td>
<td>29.27%</td>
<td>48.86%</td>
<td>8.33</td>
<td>3</td>
</tr>
<tr>
<td>P3</td>
<td>143</td>
<td>31.03%</td>
<td>49.18%</td>
<td>7.27</td>
<td>4</td>
</tr>
</tbody>
</table>

PWC: Percentage of Word Complexity; PLC: Percentage of Lexical Complexity; ASL: Average Sentence Length; RG: Readability Grade.

3.4 Data quality

Raw gaze data derived from the eye-tracking and keyboard logging were screened and adjusted. These data were uploaded to the CRITT-DB and then generated into tables that could be downloaded for data analysis. In total, 25 participants’ data were recorded; however, two participants’ data were discarded due to technical problems. One of them changed the target text format in editing, which resulted in messy eye data in the logging file. Another participant’s logging data was incomplete due to the missing gaze data in several sentences in two texts. As a result, a total of 23 participants’ data were used for analysis. Further data checking for each parameter was then undertaken to ensure the data quality. Following Hvelplund (2014), this study set the threshold of Gaze Time on Screen, calculated by dividing the total fixation duration by total task time, at above 56.6% (one SD below the mean). Since useful information cannot be extracted from fixations below 80 ms and fixations longer than 800 ms are a sign of distraction (Conklin, Pellicer-Sánchez, and Carrol 2018), the threshold of fixation duration was set between 100 ms and 800 ms (Pellicer-Sánchez 2016).

3.5 Data analysis

R language software, an open-source language and environment for statistical computing, was used to analyse the quantitative data (R Core Team 2021). Before conducting the statistical analysis, the Shapiro-Wilk test was performed in R to check the normality of data distribution. When data were normally distributed

---

3. CRITT-DB (Center for Research and Innovation in Translation and Translation Technology-Database) is a database of recorded text processing sessions in translation. Website: https://sites.google.com/site/centretranslationinnovation/tpr-db?authuser=0
(\(p > 0.05\)), a paired T-test was used to investigate the difference between HT and PE. When data were not normally distributed (\(p < 0.05\)), a non-parametric statistical hypothesis test, namely the Wilcoxon signed-rank test, was used to test the difference between the two tasks.

4. Findings

4.1 Cognitive effort in HT and PE: Whole process

The cognitive effort was measured using the following indicators: total task time, total fixation duration, fixation count, pauses, and the total amount of keyboard activities.

4.1.1 Total task time

Total task time was the time duration from the time stamp when a translator began the translating activity to the time stamp when the translator finished the task. The duration was also recorded by Translog software from the participant clicking the start button until the time when the logging file was saved. The Shapiro-Wilk test result showed that the values of total task time in the HT task (\(p < 0.05\)) and PE (\(p < 0.05\)) were not normally distributed. Thus, a Wilcoxon signed-rank test was performed in R.

Table 2. Total task time (ms) in HT and PE

<table>
<thead>
<tr>
<th>Task</th>
<th>(M)</th>
<th>(SD)</th>
</tr>
</thead>
<tbody>
<tr>
<td>HT</td>
<td>613584.6</td>
<td>195808.5</td>
</tr>
<tr>
<td>PE</td>
<td>464,942.5</td>
<td>239,614.3</td>
</tr>
</tbody>
</table>

We found a significant difference in total task time between HT and PE (see Table 2 and Figure 3). The total task time in HT (\(M = 613584.6, SD = 195808.5\)) was significantly longer than that in PE (\(M = 464942.5, SD = 239614.3\)): \(W = 1051, p < 0.05, d = 0.55\).
4.1.2 Total fixation duration

The second indicator we investigated was total fixation duration. Total fixation duration was the sum of the total reading duration in the source text and the target text. It was also the sum of the fixation duration in the source and target text areas. Total fixation duration was calculated by adding together the value of total reading time in the source text and total reading time in the target text (Table 3).

Table 3. Total fixation duration (ms) in HT and PE

<table>
<thead>
<tr>
<th>Task</th>
<th>M</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>HT</td>
<td>416721.5</td>
<td>190995.7</td>
</tr>
<tr>
<td>PE</td>
<td>307236.0</td>
<td>198497.6</td>
</tr>
</tbody>
</table>

Data in HT and PE were not normally distributed ($p < 0.05$): therefore, the Wilcoxon test in R was performed to compare the total fixation duration between HT and PE. We found a significant difference in total fixation duration between the two tasks (see Table 3 and Figure 4). The total fixation duration in HT ($M = 416721.5$, $SD = 190995.7$) was significantly longer than that in PE ($M = 307236.0$, $SD = 198497.6$): $W = 1200$, $p < 0.05$, $d = 0.58$.  

Figure 3. Total task time in HT and PE
4.1.3 Fixation count

The third aspect we investigated was fixation counts. Fixation count was the total number of times that the participant’s gaze fixated on the text and was calculated by adding the fixation counts on all the texts in one task (Table 4).

Table 4. Fixation count in HT and PE

<table>
<thead>
<tr>
<th>Task</th>
<th>M</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>HT</td>
<td>1167</td>
<td>468</td>
</tr>
<tr>
<td>PE</td>
<td>980</td>
<td>499</td>
</tr>
</tbody>
</table>

The fixation counts in HT and PE were normally distributed ($p > 0.05$): therefore, a paired sample t-test was conducted to compare the fixation counts in HT and PE. We found a significant difference between HT and PE (Figure 5). The fixation counts in HT ($M = 1167, SD = 468$) were significantly higher than those in PE ($M = 980, SD = 499$): $t(68) = 2.6423, p < 0.05, d = 0.35$.

Different from the above three indicators, the next two indicators we examined were pauses and keystrokes, which were specifically related to language production.
4.1.4 Pauses

A pause is a stop or break in typing. In this research, the number and duration of pauses were calculated by adding the corresponding values in all the texts together in one task (Table 5).

Table 5. Pause in HT and PE

<table>
<thead>
<tr>
<th></th>
<th>HT</th>
<th>PE</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>M</td>
<td>SD</td>
</tr>
<tr>
<td>Pause Duration (ms)</td>
<td>391524.5</td>
<td>157825.6</td>
</tr>
<tr>
<td>Pause Number</td>
<td>100</td>
<td>28.14</td>
</tr>
<tr>
<td>APR</td>
<td>0.67</td>
<td>0.22</td>
</tr>
<tr>
<td>PWR</td>
<td>1.00</td>
<td>0.29</td>
</tr>
</tbody>
</table>

As shown in Table 5, the pause duration in HT ($M=391524$, $SD=157825$) was longer than that in PE ($M=380730$, $SD=200365.3$), and the pauses in HT ($M=100$, $SD=28.14$) outnumbered those in PE ($M=32.1$, $SD=27.18$). We divided pause duration by pause number and found that the average pause duration in HT (3.9s) was approximately three times shorter than in PE (11.8s). This means that, in HT, there was a large cluster of short pauses, which may signify higher cognitive effort (Lacruz, Shreve, and Angelone 2012). However, it was not sufficient to
generalize because large numbers of long pauses have also been connected with higher cognitive effort (Lacruz, Shreve, and Angelone 2012). We therefore adopted APR and PWR measures (see Lacruz, Denkowski, and Lavie 2014). We computed APR by dividing the average time per pause in one task by the average time per word, and calculated PWR by dividing the total number of pauses by the total number of words. We found that the average APR in HT was 11 times lower than that in PE. In contrast, the average PWR in HT was three times higher than that in PE. These differences can be seen in Figure 6.

![APR and PWR in HT and PE](image)

Figure 6. APR and PWR in HT and PE

Since the data were not normally distributed \((p < 0.05)\), the Wilcoxon test was performed to compare the difference in the two indicators between HT and PE. The output showed that the APR value in PE \((M=7.92, SD=13.90)\) was significantly higher than that in HT \((M=0.67, SD=0.22)\), \(W=1, p<0.05, d=0.86\). The PWR value in HT \((M=1.00, SD=0.29)\) was significantly higher than that in PE \((M=0.30, SD=0.26)\), \(W=1820, p<0.05, d=0.85\).

4.1.5 Keyboard activities

The last aspect we examined was keyboard activities. Keyboard activities involve the number of deletions, number of insertions, and total number of keyboard actions, derived by adding the numbers of insertions and deletions in all sessions in one task together (Table 6).
Table 6. Keyboard activities in HT and PE

<table>
<thead>
<tr>
<th>Keyboard</th>
<th>Task</th>
<th>M</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Insertion</td>
<td>HT</td>
<td>643.46</td>
<td>109.10</td>
</tr>
<tr>
<td></td>
<td>PE</td>
<td>124.41</td>
<td>129.35</td>
</tr>
<tr>
<td>Deletion</td>
<td>HT</td>
<td>135.15</td>
<td>88.93</td>
</tr>
<tr>
<td></td>
<td>PE</td>
<td>122.27</td>
<td>131.31</td>
</tr>
<tr>
<td>Total</td>
<td>HT</td>
<td>778.6</td>
<td>191.75</td>
</tr>
<tr>
<td></td>
<td>PE</td>
<td>246.6</td>
<td>259.53</td>
</tr>
</tbody>
</table>

As data in these groups were not normally distributed ($p<0.05$), the Wilcoxon signed-rank test was performed. We found differences in keyboard activities between HT and PE. Participants made significantly more insertions in HT ($M=643.46, SD=109.10$) than in PE ($M=124.41, SD=129.35$): $W=1830, p<0.05, d=0.86$. Similarly, the total number of keyboard activities in HT ($M=778.6, SD=191.75$) was significantly larger than that in PE ($M=246.6, SD=259.53$): $W=1806, p<0.05, d=0.84$. Although the number of deletions in HT was larger than that in PE, only a marginally significant difference between HT ($M=135.15, SD=88.93$) and PE ($M=122.27, SD=131.31$) was found in the two tasks ($W=1163.5, p=0.06, d=0.23$).

4.2 Cognitive effort in HT and PE: Phased view

CRITT divides the translation process into three phases: initial orientation, drafting, and revision (Carl, Dragsted, and Jakobsen 2011; Carl 2012; Dragsted and Carl 2013). In this study, we adopt this subdivision to compare HT and PE in detail. The orientation phase refers to the duration of the translator’s reading activities, whether a detailed reading or skimming of the source text before the translator begins translation. It starts from the initial fixations on the source text and lasts until the input of the first keystroke takes place. The drafting phase is the process of actual translation activity, lasting from the input of the first keystroke until the last token of the source text is registered. In this process, the translator finishes the first draft of the target text with keyboard activities such as insertions or deletions. The revision phase is the translator’s return to the first draft to make changes or modifications after the last token of the source text is registered. Data from these three phases are shown in Table 7 and Figure 7. These data were not normally distributed; therefore, the Wilcoxon test was performed to compare the difference in the three phases between HT and PE.
### Table 7. Fixation duration of sub-phases in HT and PE (ms)

<table>
<thead>
<tr>
<th>Phase</th>
<th>Task</th>
<th>M</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Orientation</td>
<td>HT</td>
<td>1762.28</td>
<td>14717.81</td>
</tr>
<tr>
<td></td>
<td>PE</td>
<td>69612.08</td>
<td>50868.63</td>
</tr>
<tr>
<td>Drafting</td>
<td>HT</td>
<td>509795.1</td>
<td>148828.8</td>
</tr>
<tr>
<td></td>
<td>PE</td>
<td>317649</td>
<td>209120.7</td>
</tr>
<tr>
<td>Revision</td>
<td>HT</td>
<td>83932.28</td>
<td>90623.38</td>
</tr>
<tr>
<td></td>
<td>PE</td>
<td>76364.88</td>
<td>72460.08</td>
</tr>
</tbody>
</table>

#### Figure 7. Duration of sub-phases in HT and PE (s)

4.2.1 **Orientation**

The duration of the orientation phase was calculated using the time stamp of the registration of the first keystroke in Translog. It can be seen clearly in Table 7 that the orientation duration differed between tasks. The R language output showed
that the orientation duration of PE ($M = 69612.08, SD = 50868.63$) was significantly longer than that of HT ($M = 117622.88, SD = 14717.81$): $W = 73, p < 0.05, d = 0.80$. Participants spent a long time reading the source text and preparing to produce the target text in PE (Figure 7).

4.2.2 Drafting

Although Nitzke and Oster (2016) stated that drafting is a typical human translation phase because the translation draft is manual in HT and machine-made in PE, we included the drafting phase in both HT and PE to investigate the difference when translators were working with a manual draft and a machine-made draft. We believed that even if the first translation draft was done by a machine, the participant still needed more than one round to post-edit the text. A translator might edit machine-translated texts more than once from the first token until the last token. Hence, we regarded the first-round editing of MT texts, from the typing of the first keystroke until the last token was edited, as the drafting phase of PE. The downloaded table recorded the time stamps when the drafting process began and the time stamps when revision began. We calculated the duration of the drafting phase by subtracting the revision starting time stamps from the drafting starting time stamps (Table 7). We found that the drafting duration of HT ($M = 509795.1, SD = 148828.8$) was significantly longer than that of PE ($M = 317649, SD = 209120.7$): $W = 1589, p < 0.05, d = 0.64$. The findings showed that the participants spent more time working with the first draft of the target text in HT tasks than in PE (Figure 7). In PE, the first draft of the target text was already there in the target area; therefore, less typing resulted in a shorter duration of the drafting phase.

4.2.3 Revision

The duration of the revision phase was calculated by subtracting the beginning time stamps of revision from the total duration of the whole session (Table 7). We found a slightly longer revision duration in HT ($M = 83932.28, SD = 90623.38$) than in PE ($M = 76364.88, SD = 72460.08$); however, the difference was not significant ($W = 928, p > 0.05, d = 0.01$).

5. Discussion and conclusion

Differences between HT and PE were found in all the indicators. In terms of total task time, we found a significantly longer duration in HT compared with that in PE, which suggests a time saving and an increase in speed in PE. This result is in line with most of the previous findings: for instance, the findings reported by
Plitt and Masselot (2010); Sousa et al. (2011); Läubli et al. (2019), and Stasimioti and Sosoni (2021). However, as discussed earlier, speed and productivity, the tip of the iceberg, indicate effort temporally. Hence, an increase in speed in PE partially suggests that the PE task requires less effort than HT.

Similarly, we found a significantly longer total fixation duration and significantly higher fixation counts in HT compared to PE, indicating that the PE process was less demanding. These findings are in accordance with those reported by Stasimioti and Sosoni (2020), who found that HT had a significantly longer fixation duration and a larger number of fixation counts. Our study also partially confirms the finding of Daems et al. (2017), who reported a significantly shorter average fixation duration in PE; however, this was only in source texts by student translators. By contrast, these findings do not support the research conducted by Carl et al. (2011), who reported no significant difference in the average gaze time between HT and PE.

In terms of pause, we found a significantly higher APR and lower PWR in PE. These findings support those reported by Schaeffer et al. (2016) and Jia et al. (2019), who found that PE had a significantly lower PWR than HT. Based on the conclusion reached by Lacruz et al. (2012) and Lacruz and Shreve (2014), who argued that higher PWR and lower APR are indicators of higher cognitive effort, our findings imply that the cognitive effort in HT is significantly higher than that in PE.

Moreover, we found a significant difference between HT and PE regarding keyboard activities. In HT, participants made more insertions and deletions than in PE. In this sense, the PE task required less cognitive effort because post-editors do less typing and mouse clicking. Also, the result corroborates the findings of Green, Heer, and Manning (2013); Screen (2017), and Stasimioti and Sosoni (2020), who found that the average keyboard activities involved in HT were significantly more than those in PE. Since pause and keyboard activities are measurements of the text production phase in translation, our findings conclude that PE is less cognitively demanding than HT in target text production. Taking all the above indicators together, the results unanimously indicate that HT is more cognitively demanding than PE.

The difference between HT and PE was also found in the three sub-phases. Participants spent longer fixation durations reading the source texts and preparing themselves for the editing task in PE, implying that the orientation phase of PE is more cognitively demanding. It has been stated that in PE, the translators skip the orientation phase and begin the revision work directly (Mesa-Lao 2014; Stasimioti and Sosoni 2020); however, our findings on the orientation phase show that this is not case. For the drafting phase, our study shows that the fixation duration of the drafting phase in HT is longer than that in PE, suggesting a higher cog-
nitive effort in the drafting phase of HT in comparison with PE. This finding can be explained by the fact that in PE, the machine has already produced the target texts; thus, a shorter drafting process is entailed when the PE task is carried out. In the revision phase, the translators spent a slightly longer fixation duration in revising the manually translated target texts; however, the results in the revision phase did not show any significant difference between HT and PE. This finding again contradicts the claim that translators tend to ignore the final revision phase in PE (Mesa-Lao 2014; Stasimioti and Sošoni 2020). In our study, translators all demonstrated a tendency to read the source texts and revise the final target texts regardless of whether they were post-editing or translating from scratch.

Through the investigation of cognitive effort, this research provides a deeper insight into the cognitive differences between HT and PE, especially the differences among the three sub-phases. As this research operates in the empirical paradigm, the translators’ behaviour data collected in this experiment enrich the translation process data repository and extend our understanding of translators’ mental activities in dealing with the natural translation process. The findings based on the Chinese–English language pair in this research have referential significance for translation process research on other language pairs. Admittedly, limitations still exist. Although there has been no commonly agreed saturation in translation process research, more participants may be needed to reach a more sound and valid generalization. The findings of the study are also based on the indirect investigation of translators’ mental processes in translation. Perhaps more direct methods such as fMRI would greatly help researchers to understand the cognitive effort in the translation process. Since this research is a comparative study of cognitive effort in HT and PE, further studies could be conducted to explore the factors influencing cognitive effort, such as the effect of time pressure on the translation process, or to examine the distribution of translators’ cognitive resources in different segments in both HT and PE.

Acknowledgements

The authors would like to thank Prof. Wang Yunhua at Zhengzhou Sias University and Miss Zhai Xiaojuan from Translink Co., Ltd. for their academic support. Thanks also to Tobii China for the technical support.
References


Cognitive effort in human translation and machine translation post-editing processes


Résumé

En raison des imprécisions inhérentes à la traduction automatique (TA), la post-édition (PE) devient inévitable. Cette situation soulève la question de la pertinence de l’effort humain pour améliorer une TA par rapport à l’efficacité d’une traduction manuelle. Cependant, jusqu’à présent, peu de recherches ont été menées pour comparer l’effort cognitif impliqué dans le processus de PE et ses sous-phases (orientation, rédaction et révision) avec celui de la traduction humaine (TH). Afin de combler cette lacune, la présente étude vise à examiner et à comparer l’effort cognitif dans les processus de TH et de PE lors de la traduction du chinois vers l’anglais. Les données ont été recueillies à l’aide de méthodes de suivi oculaire et d’enregistrement de frappes au clavier auprès de 25 participants sollicités pour effectuer trois tâches de TH et trois tâches de PE respectivement. La comparaison de l’effort cognitif a été réalisée en prenant en compte les processus de TH et de PE, ainsi que leurs différentes sous-phases. L’étude met en évidence une différence significative en termes d’effort cognitif, de durée d’orientation et de durée de rédaction entre la TH et la PE.

Mots-clés : traduction automatique, post-édition, effort cognitif, processus de traduction

Address for correspondence

Ali Jalalian Daghigh
Department of English Language
Faculty of Languages and Linguistics
Universiti Malaya
50603 Kuala Lumpur
Malaysia
jalalian@um.edu.my
https://orcid.org/0000-0001-7418-7048

Co-author information

Yu Wang
Department of English Language
Faculty of Languages and Linguistics
Universiti Malaya
siaswangyu@gmail.com

Publication history

Date received: 25 June 2022
Date accepted: 13 March 2023