# Word-Based Human Edit Rate (WHER) as an Indicator of Post-editing Effort


**Jie Huang and Michael Carl**

**Abstract** Estimating post-editing effort is essential to identify translation difficulties and decide the payment for post-editors. Keystrokes, fixations, and production duration, as well as lexical and syntactic variations of the translation product, are frequently used as indicators of post-editing effort. This chapter introduces Word-based Human Edit Rate (WHER), a measure derived from HTER, as a new predictor to measure post-editors' effort on the word-level. The original HTER metric calculates an edit distance between MT output and its post-edited version from the minimum number of assumed edit operations. The WHER metric matches these edit operations to the corresponding words in the TT segment and maps them via alignment links to ST words. WHER thus provides the minimum number of expected edit operations for each ST word given the MT output. The chapter describes an experiment in which 21 student translators were invited to post-edit audiovisual texts and their translation processes were recorded with eye-tracking and keystroke-logging devices. After correlating WHER operations with the other common effort indicators derived from the process and product, we find that WHER is a reliable predictor for word-level post-editing effort.

**Keywords** Word-based Human Edit Rate (WHER) · Cognitive effort · Product analysis · Process analysis · Post-editing effort

## 1 Introduction

Several studies using MT for audiovisual texts have shown that the direct MT output cannot meet high-quality standards in the domain of audiovisual products (Armstrong et al. 2006, Melero et al. 2006,\vadjust{\pagebreak} Volk 2008, Bywood

J. Huang (✉)
Renmin University of China, Beijing, China

M. Carl
Kent State University, Kent, OH, USA
e-mail: mcarl6@kent.edu



M. Carl (ed.), *Explorations in Empirical Translation Process Research*, Machine Translation: Technologies and Applications 3,
https://doi.org/10.1007/978-3-030-69777-8_2

et al. 2013, Burchardt et al. 2016). However, MT post-editing can reduce translators' effort and increase productivity (de Sousa et al. 2011, Ortiz-boix and Matamala 2016, DePalma et al. 2019). Estimating post-editing effort relates to the identification of translation difficulty (e.g., Dragsted 2012; see also Vanroy et al. this volume, Chap. 10) and also impacts the pay rates of post-editors (Vieira 2014).

According to Krings (2001), the post-editing effort consists of three aspects: technical, temporal, and cognitive effort. Technical and temporal effort can be measured by the number of keystrokes typed and the time spent (Dragsted 2012, Jia et al. 2019). However, cognitive effort cannot be directly observed but only estimated through the process of reading and writing and also post-edited texts (Campbell 2017). For example, eye-tracking techniques are used to collect reading activities (Koglin 2015, Vieira 2016a), and the lexical and syntactic variations of post-edited texts are measured to reflect the cognitive effort in writing (Nitzke 2019, Vanroy et al. 2019).

This study defines a new product-based measure to assess translation difficulty, the Word-based Human Edit Rate (WHER). WHER is a derivation of HTER which measures the minimum edit distance between MT output and its post-edited version on the sentence level (Snover et al. 2006). Do Carmo (this volume, Chap. 1) inverts the reference and the hypothesis in the computation of HTER which makes the result better interpretable to assess post-editing effort and calls his new measure HER. In this chapter, we learn from do Carmo's practice and extend HER into WHER by mapping the edit operations in the TT words via word-alignment links to the equivalent ST positions. We assess to what extent WHER can be used to indicate post-editing effort as exerted during the process of post-editing. While keystroke data is recorded during the post-editing process to indicate the "real" amount of technical effort, WHER measures the number of minimum TT edit operations per ST word. We, thus, expect to find a correspondence between the minimum possible and the really performed edit operations.

The objective of this study is to assess the extent to which WHER correlates with post-editing behavior and, thus, may be suited to estimate post-editing effort. In addition, the WHER score might point to positions that are difficult to post-edit, which can be helpful to evaluate post-editors' effort without observations from the post-editing process. Another potential of WHER is to predict whether a word is correctly translated by the MT system and thus helps estimate the MT quality. Translation quality estimation (QE) is increasingly important in developing Natural Language Processing (NLP) and MT engines; various models have been created to fulfill the task without human ratings (Martins et al. 2017, Basu et al. 2018, Xenouleas et al. 2019). In the following section, we give an overview of frequently used effort indicators in the translation process and product. In Sect. 3, we present our experimental setting, data collection method, and WHER computing method. The correlation between WHER and several effort indicators of post-editing effort will be discussed in Sect. 4, and the key findings and future scenarios for WHER usage will be summarized in Sect. 5.

## 2 Related Research

Three approaches are commonly used to estimate effort in translation which are sometimes combined to triangulate the data. The first approach is to observe reading and writing activities from keystroke-logging and eye-tracking data (Dragsted 2012, Koglin 2015, Vieira 2016a). The second approach is to estimate the effort from the post-edited results. Lexical and syntactic variations of the TT indicate cognitive effort in translation production, which has been addressed with multiple measures including word-level entropy scores such as HTra and Word Distortion Entropy (HCross) (Carl et al. 2016). The last approach is subjective ratings from human post-editors (de Sousa et al. 2011, Vieira 2016b). As a traditional method, subjective reflection is used to elicit the perceived effort of post-editors during the translation or post-editing task (Moorkens et al. 2015).

However, previous studies found that a single measure from any of the above approaches is not robust enough to explain the cognitive effort (Koponen 2012, Guerberof 2014). Some measures are more sensitive to individual differences than others (Vieira 2016a). Also, there may not be strong correlations between the different approaches to measuring cognitive effort. For example, average fixation duration per sentence and subjective ratings are not strongly associated with each other in correlation tests and principal component analyses (Schaeffer and Carl 2014, Vieira 2016a). That is why multiple approaches are normally used together to measure the post-editing effort. In this section, we present commonly used effort indicators and related studies from the approaches of process and product analysis. As human subjective ratings concern larger segments and are usually not applicable to word-level analysis, the last approach is less relevant to this study and not further discussed.

### *2.1 Process Indicators*

The first approach is to estimate post-editors' effort by their eye movements and keystroke behavior during reading and writing activities. Detailed information including keystrokes (e.g., number of insertions, number of deletions), fixations (e.g., number of fixations on ST/TT, duration of fixations on ST/TT). and duration (e.g., total production duration) are commonly used metrics in translation and writing studies (Mossop 2007, Dragsted 2012, Leijten and Van Waes 2013, Koglin 2015).

Keystroke information, such as the number of insertions and deletions and the total number of all keystrokes, are related to the technical effort of post-editors as they reflect "the actual linguistic changes to correct the machine translation errors" (Koglin 2015: 129). Naturally, the more MT output is modified, the more effort is required during post-editing. Eye movements reflect reading behavior on the ST and TT. According to the eye-mind hypothesis, fixations are usually linked to attention

(Just and Carpenter 1980) such that the attention follows eye movements. Empirical translation studies have shown that the average fixation duration and the number of fixations per word correlate with other effort indicators such as the pause-to-word ratio and production duration per word (Vieira 2016a). Also, the first fixation duration and the total fixation duration are used as indicators of effort (Schaeffer et al. 2019). Research show that differences are observed in post-editors' fixation behavior, with TT usually attracting more attention than the ST (Sanchis-Trilles et al. 2014, Vieira 2014).

## *2.2 Product Indicators*

Lexical variations and syntactic distortions (reordering) in the TT are indicators of post-editing effort, as they reflect activities of text modification or structure adjustment during post-editing (see, e.g., Vanroy et al. (this volume, Chap. 1); Lacruz et al. (this volume, Chap. 11); Ogawa et al. (this volume, Chap. 6)). The larger the number of alternative translations, the more effort is expected for post-editors to make the modification. Similar logic applies to the syntactic variations, which are measured by the vectorized word sequence distortion from the ST to the TT sentence. The concept of entropy (Carl this volume, Chap. 5), borrowed from information theory, is used to show the degree of translation variations of each ST word (Schaeffer and Carl 2014, Schaeffer et al. 2016). The features are generated by the analysis toolkit integrated into the CRITT TPR-DB, a large repository of translation process data (Carl et al. 2016). Overall, the indicators of lexical and syntactic modifications of the MT output give us a glimpse of post-editors' effort for each ST word, which is facilitated by the word-level analysis.

AltT, ITra, and HTra are measures of lexical translation choices. AltT is the number of alternative translations for an ST word in a given context across different participants and sessions. Higher AltT values indicate a wider range of TT words corresponding to the ST word. ITra is the self-information of a translation, computed as ITra $= \log_2(1/\text{ProbT})$, where ProbT is the probability of the translation, as provided in the CRITT TPR-DB table. Higher ITra values correspond to higher information content indicating that the current TT word is less frequently used. These two indicators can be calculated after the alignment of ST and TT words (or phrases) to display the translation choices of each participant for each ST word. HTra is the word translation entropy that multiplies the sum of ITra with their expectation (Schaeffer et al. 2016, Carl this volume, Chap. 5). The three measures differ in the sense that AltT and HTra values are identical among participants on the same ST token, while ITra is a participant-specific metric that relates to a particular translation. Previous studies have found that the HTra value correlates with process measures such as the duration of production, the number of insertions, and the number of fixations on ST (Bangalore et al. 2016, Vanroy et al. 2019, Wei this volume, Chap. 7; Lacruz et al. this volume, Chap. 11). It shows that the measures of lexical variations are robust and reliable to estimate post-editing effort.

Cross and HCross are measures for syntactic distortions. Cross is a vector of relative cross-linguistic distortion of word position between ST and TT. Its absolute value thus indicates the degree of adjustment made on the sentence syntactic structure. HCross is based on the Cross values of all participants on the same ST word and is calculated as the word order entropy (Carl and Schaeffer 2017). If the HCross value of a given ST word is 0, it means that the same relative translation re-ordering is chosen by all participants in the TT (Nitzke 2019, Carl this volume, Chap. 5). On the contrary, a high HCross value implies a larger variance among participants in choosing different relative TT word position. Therefore, Cross is a participant-unique and session-unique metric, and HCross is unique among ST words only. The two metrics are both used to measure the effort of adjusting sentence syntactic structures during post-editing.

## 3 Method

### *3.1 Material*

Twenty-one participants post-edited extracts of two audiovisual texts on the topic of law. One text was a documentary film, and the other was an episode of TV drama. In each text, two extracts of comparable length and durations were selected and allocated to participants in a random sequence. Although legal texts may be more specialized than general texts, the pieces did not contain difficult terminologies. For the material to be as authentic as possible for the sake of ecological validity (Orero et al. 2018), all selected scenes were self-contained clips with subtitles referring to a complete scenario. The documentary extracts have a mean duration of 25 s (sd = 3) and a mean length of 53 words (sd = 1). The drama extracts have a mean duration of 36.5 s (sd = 2.5) and a mean length of 96.5 words (sd = 2.5). As both subtitles and videos of source material are required for the professional working process of AVT (Díaz Cintas and Remael 2014), the corresponding video clip with subtitles in the SL was played twice before participants started the task. The video-watching process was not eye-tracked or key-logged. No time constraint was set on participants' post-editing process.

### *3.2 Participants*

All 21 participants were Masters students of translation. All of them were Chinese native speakers with English as the second language, between 22 and 32 years old. They had a similar level of English proficiency, with 81% having passed the China Accreditation Test for Translators and Interpreters (CATTI) between the English and Chinese language pair and all the rest scoring over 7 in International English

Language Testing System (IELTS).[1] None of them had professional experience or training on AVT or post-editing. Therefore, they were targeted as novice translators without expertise in AVT or post-editing. Our sampling method considers that post-editing audiovisual texts is currently not a common practice for professionals, so we only selected novice translators in this post-editing experiment. To imitate the actual working scenario, we provided participants with a brief guideline of quality and technical requirements in subtitle translation, i.e., the Code of Good Subtitling Practice (Ivarsson and Carroll 1998). They were asked to read and understand the guideline carefully before they start.

### *3.3 Apparatus*

This experiment used a portable Tobii X2–60 eye-tracker (60 Hz) and Translog-II software for the recording of both eye-tracking and keystroke-logging data. The eye-tracker was placed on the bottom of a 24″ monitor as shown in Fig. 1.

After watching the video and calibrating the eye-tracker, participants post-edited a transcript of the video text in Translog-II with the task setup as follows: as shown in Fig. 2, the English source texts (i.e., the subtitles) were displayed on the left


**Fig. 1** Eye-tracker setup of the experiment

[1]For more information, see http://www.catticenter.com/wzjs/452 (CATTI) and https://www.ielts.org/ (IELTS).

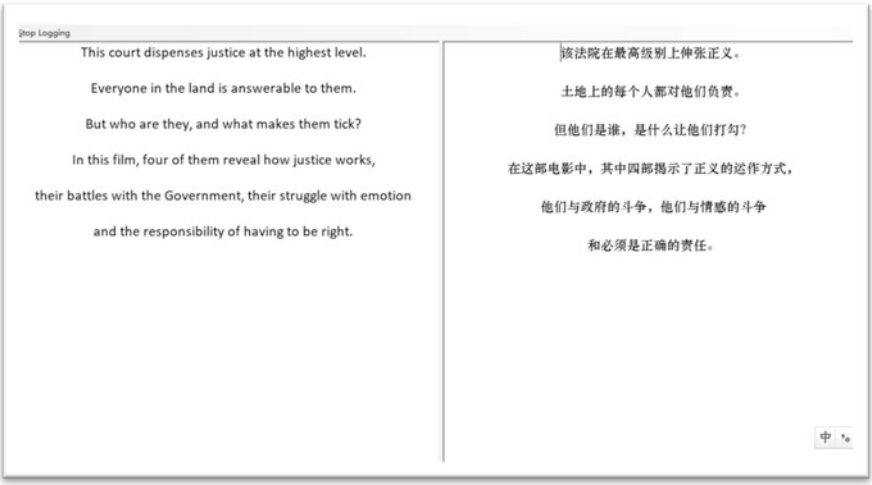


**Fig. 2** Translog-II task setup for participants

window of Translog-II in Calibri, 20-point size, and 1.5 line spacing, while the Chinese target texts were positioned on the right side with similar settings except that the font was changed to STZhongsong. The line breaks of both English and Chinese texts were displayed in the same way as in the original videos, in which a short sentence occupies a line and a long sentence is broken into two lines. After collecting the XML files produced by Translog-II, we added the data to the CRITT TPR-DB for further alignment and analysis.

## *3.4 Data Alignment*

The ST and TT were first sentence-segmented and tokenized[2] and aligned using the YAWAT tool (Germann 2008). To ensure consistency, a single researcher aligned each meaning unit in ST (a word or phrase) to its corresponding TT unit in all sessions by participants. Figure 3 shows a tokenized segment from the interface of the manual alignment tool with the ST on the left and the TT on the right. Successively, we aligned the ST and TT tokens on a level of minimal translation equivalence. In Fig. 3, the ST-TT alignment groups are as follows: *"Everyone"*—*"每个/人"*, *"in the land"*—*"国 土/上"*, *"is answerable to"*—*"回/应/. . . 的/诉求"*. The ST tokens *them* and punctuation "." are not translated and have no TT equivalents. As shown in this example, all of the ST and TT units represent the minimal meaning group in each language, including basic collocations. Phrase-level

[2]For Chinese, we used the Stanford Tokenizer (https://nlp.stanford.edu/software/tokenizer.shtml).

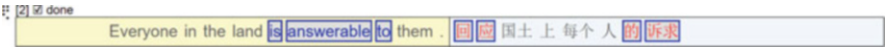


**Fig. 3** Alignment of ST and TT units in Yawat

units avoid the arbitrary separation of meanings and allow us to have consistent alignment in both ST and TT.

## *3.5 Computation of WHER*

The TER script computes an edit distance at the segment-level between a translation hypothesis (usually the MT output) and a translation reference (usually a human translation). It produces an *adjusted hypothesis*, which is tagged with edit operations on specific word positions. The adjusted hypothesis encodes edits and shifts, which represent the minimum edit distance between hypothesis and reference. The TER script also computes a cumulative TER score where each of the edits has a cost of 1, thus representing the distance between the hypothesis and the reference. Snover et al. also introduce the HTER score in which the reference is the post-edited MT hypothesis.[3] Given that TER is designed to be a measure of MT quality and there are many ways to translate a sentence, HTER (i.e., edit distance between MT output and its post-edited version) is believed to give more prominence to the "real" quality of the MT output, as compared to a perhaps very different human translation. However, in order to assess the amount of human editing (as compared to MT errors), do Carmo suggests reversing the reference and the hypothesis, so that the post-edited version becomes the hypothesis and the MT output the reference. do Carmo calls his reversed edit distance metric the HER which is presumably better suited to assess the post-editing effort. We extend the HTER score into a word-level HER (WHER), which (1) reverses the reference and the hypothesis in the HTER calculation following do Carmo's suggestion, (2) breaks down the segment-level HER operations to the word level, and (3) maps the operations on TT words to the aligned ST equivalents.

The edit distance between the hypothesis and the reference text consists of shifts of word positions (H), insertion (I), deletion (D), and substitutions (S). We take the hypothesis text to be the MT output and the reference text to be the final post-edited result. The minimum edit operations between the reference text and the adjusted MT are calculated to represent the minimum effort of post-editing from the product-wise perspective. A sample sentence containing all of the four types of WHER operations is presented in Fig. 4 illustrating the mappings between WHER operations and the exact TT tokens.

[3]But see do Carmo (this volume, Chap. 1) for a discussion on the terminological confusion of different definitions of HTER.

| ST word. | Everybody | | in the land | | | | is answerable to | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **WHER** | | | SH | | | | ISDDDSS | | | | |
| Keystrokes | 都对他们负责。 | | 国家的的家土上 | | | | 回应的诉求 | | | | |
| TT (Hyp.) | 回 | 应 | | 国土 | 上 | 每个 | 人 | | | | 的 | 诉求 |
| Adjusted Hyp. | I | S | | S | H | | | D | D | D | S | S |
| MT (Ref.) | | 土地 | 上 | 的 | | 每个 | 人 | 都 | 对 | 他们 | 负责 | 。 |

**Fig. 4** Example for calculating the minimum edit operations in WHER. Identical background colors indicate translation equivalence. Font colors green (I), insertion; blue (D), deletion; red (S), substitution; yellow (H), shift

Figure 4 reproduces the alignment from Fig. 3. It shows the ST segment with the WHER operations, the Chinese post-edited translation, an adjusted hypothesis which consists of the string of edit operations that are projected to the ST words, and the raw MT output. In addition, we add a row of the real keystroke activities below WHER to make comparisons. The adjusted MT is an intermediate construct that illustrates the assumed operations which have presumably occurred in the mapping of the post-edited hypothesis on the MT reference. Figure 4 also shows a sequence of deletions, substitutions, and insertion operations and a shift between the hypothesis and the MT reference. In addition, it shows the three aligned chunks between the ST segment and the TT hypothesis (i.e., the post-edited version). Notice that there is a syntactic inversion between "*Everybody*"—"每个/人" and *"in the land"*—"国土/上" and a discontinuous translation, *"is answerable to"*—"回/应/. 的/诉求".

Let's illustrate the calculation of WHER by taking an example of the ST phrase "is answerable to" which is aligned with two discontinuous Chinese segments, "回/应/" and "的/诉求" consisting of a total of four TT tokens.

Step 1: The edit operations are mapped to Chinese TT words. The shift operation (H) only happens in the position change of the hypothesis, so it is found here below "上" of the post-edited TT. Substitutions are identified as tokens occupying the same positions but with different contents between TT and MT. In this case, four substitutions are detected for "土地", "的", "负责", and "。". Deletions and insertions happen in places where tokens are missing or added from the MT to the TT. The vertical display of token positions in Fig. 4 shows that "回" is an insertion and "都", "对", and "他们" are deletions.

Step 2: The edit operations associated with the TT tokens are then mapped onto the ST words through their alignment links. The WHER operations for "is answerable to" is, consequently, "ISDDDSS."

Step 3: The WHER score for each ST token in an alignment unit is calculated as the sum of WHER operations because each operation type costs 1 in the same way. Then, the WHER score for "is answerable to" is 7, as a sum of all edit operations.

Next, we can compare WHER operations to the actual modifications[4] that were produced during post-editing session. The row of actual modifications as produced in the post-editing process is shown below the three ST chunks. Overall, there are 19 modifications of Chinese characters in the segment (which required 38 keystrokes with the Chinese input tool). Seven modifications (deletions), which make up five Chinese words, are assigned to "Everybody." These operations represent the deletion of the erroneous MT output "都/对/他们/负责/。/", at the end of the MT output. WHER assigns the modifications of these five words as three deletions and two substitutions to *is answerable to* (DDDSS). There is thus a discrepancy in the allocation of edit operations between WHER and the TPR-DB analysis. There are also seven modifications assigned to *in the land*. The translation of *in the land* actually consists of two tokens and three Chinese characters "国土/上/". However, the post-editor seems to have revised his/her own editing activities and inserted and deleted the two characters "家的". Finally, there are five insertions "回应的诉求" for *is answerable to* which corresponds to the discontinuous chunk "回应" and "的诉求". Notice that the first part (回应) consists of one insertion and one substitution in terms of WHER operations, but were produced as insertions in the post-editing process.

Comparing WHER to HER, do Carmo says "HER is to move the focus from errors to edits" and "an edit rate presents an improved perspective on actual editing" (see do Carmo this volume, Chap. 1). We pursue the same aim with the WHER score but on a word level. With WHER, we are looking at the human editing effort instead of MT errors on the word level. However, the only thing that changes from errors to edits is an inversion of labels for insertions and deletions: what is a deletion error in the MT output in TER becomes an insertion operation in the editing pattern for HER, and what is an insertion error in the MT output for TER becomes a deletion operation in HER. Shifts (H) and substitutions (S) are independent of the error/edit view as they are symmetrical. Figure 4 shows this reversed relation in the errors vs. edits view: the "回" in the post-edited translation of *is answerable to* was an omission error in the MT output and appears as insertion (I) in the string of edit operations. Similarly, "都对他们" was an erroneous MT insertion but appears as deletion (D) in the edit string. Shifts (H) and substitutions (S) also change directionality depending on whether we look from MT hypothesis to TT reference or the other way around.

[4]Note that there is a difference in produced keystrokes and modifications for Chinese: with an input method editor (IME), there are usually more (i.e., 2–3) keystrokes required to produce one Chinese character. In this analysis, we count the number of character modifications in the text, as opposed to keystrokes.

### *3.6 Features*

We correlate effort indicators from both the process and product data with WHER score. As mentioned above, keystroke, fixation, and duration information are indicators of post-editing effort during the process. The details of the process and product features are displayed in the following two tables.

As shown in Table 1, ten features from the three groups of process data are used in our experiment. For keystroke, Key_ins and Key_del measure the insertion and deletion activities recorded on the TT window by Translog-II, and Key_all is the total number of keystrokes that reflect the overall typing effort. Fixation features including Fix_S, Fix_T, Trt_S, and Trt_T cover both temporal and count data with a separation on ST and TT windows. The total values Fix_all and Trt_all are also included to provide a holistic view of the reading effort. Measuring the time spent from the first keystroke to the last keystroke, Dur is a temporal record of the technical effort.

Table 2 shows five features used for measuring lexical and syntactic variations of post-edited texts. AltT counts the number of alternatives for each ST token; therefore, it is not a participant-specific value. ProbT is the probability of the current translation choice relating to the ST token, which is sensitive to participants' individual differences. Based on the two values, HTra calculates the information entropy of each ST word across the whole data set and indicates the variations of translations for each ST token regardless of the participants' differences (Carl and

**Table 1** Process features

| Category | Feature name | Description |
|---|---|---|
| Keystroke | Key_ins | Number of keystroke insertions |
| | Key_del | Number of keystroke deletions |
| | Key_all | Total number of keystroke insertions and deletions |
| Fixation | Fix_S | Number of fixations on ST |
| | Fix_T | Number of fixations on TT |
| | Fix_all | Total number of fixations on both ST and TT |
| | Trt_S | Fixation duration on ST |
| | Trt_T | Fixation duration on TT |
| | Trt_all | Total fixation duration on both ST and TT |
| Duration | Dur | Production duration from the first keystroke to the last keystroke |

**Table 2** Product features

| Category | Feature name | Description |
|---|---|---|
| Lexical variation | AltT | Number of alternatives for ST tokens |
| | ITra | Self-information of current translation choice |
| | HTra | Word translation entropy |
| Syntactic variation | Cross | Cross value for ST tokens |
| | HCross | Word distortion entropy |

Schaeffer 2014, Schaeffer et al. 2016). Similarly, Cross and HCross are included to indicate the variation related to syntactic adjustments.

## 4 Results

This section discusses the results of the correlation tests. The collected data of the above features are not normally distributed except HTra, ITra, and HCross. The distributions of WHER, all of the process features, and the remaining product features are skewed to the right. As most of these data have 0 s, we decide to transform the above right-skewed data by adding a constant 1 to all values and taking their log transformation (Hancock et al. 2018). In this way, the features are more comparable, and we can use Pearson's $r$ as the correlation metric in all tests. Among all of the correlation results below, the asterisks indicate the significance levels of the correlations, where one asterisk (*) refers to a significant effect ($p$-value <0.05), two asterisks (**) for a highly significant effect ($p$-value <0.01), and three asterisks (***) for a very highly significant effect ($p$-value <0.001).

### *4.1 Process Features*

We collect process data from keystroke, duration, and fixation as indicators of technical, temporal, and cognitive effort during post-editing. With log transformations on both WHER and the other features, Table 3 shows the correlation results between them.

Table 3 shows that LogWHER strongly correlates with the insertion activities (LogKey_ins) and the number of overall keystrokes (LogKey_all). The correlation of LogKey_ins ($r = 0.76$) is higher than that of LogKey_all ($r = 0.68$), indicating a major contribution from the insertion activities. The number of deletions (LogKey_del) is only weakly but significantly correlated with LogWHER

**Table 3** Correlations between WHER score and the process features (log-transformed)

| Category | Feature name | Pearson's $r$ with LogWHER |
|---|---|---|
| Keystroke | LogKey_ins | 0.76*** |
| | LogKey_del | 0.13*** |
| | LogKey_all | 0.68*** |
| Fixation | LogFix_S | 0.01 |
| | LogFix_T | 0.15*** |
| | LogFix_all | 0.13*** |
| | LogTrt_S | 0.01 |
| | LogTrt_T | 0.12*** |
| | LogTrt_all | 0.11*** |
| Duration | LogDur | 0.35*** |

($r = 0.13$). The strong correlation suggests that WHER score is a good indicator of performed keystroke operations.

However, the correlation with fixation data is relatively weak. Fixations are counted separately for the ST and TT texts; we find higher scores in TT than ST where LogFix_T ($r = 0.15$) and LogTrt_T ($r = 0.12$) have significant correlations with LogWHER. The results corroborate the findings of previous studies that TT is likely to attract more attention than ST (Daems et al. 2017, Schaeffer et al. 2019). Overall, the correlations between fixation data and WHER are not as strong as those in keystroke data. One possible explanation might be the noise of fixation data and the gaze-to-word mapping errors. WHER is thus not predictive of the reading effort as indicated by the number and duration of fixations in this study.

We used the production duration of an ST token (Dur) as an indicator of temporal effort. The correlation result between LogDur and LogWHER is moderate ($r = 0.33$), which is higher than the scores for fixations and lower than those for keystrokes. It is also indicated that the WHER feature reflects, to some extent, the temporal effort during post-editing.

Overall, the keystroke and duration features are found to be at least moderately correlated with the WHER score, but the fixation data are only weakly correlated. This suggests that WHER is more indicative of the technical and temporal effort than for cognitive effort, as gathered from reading activities.

## *4.2 Product Features*

As mentioned above, AltT and Cross are distributed with skews to the right, so we added a constant 1 to all values and used their log-transformed data to correlate with LogWHER. In particular, Cross has both positive and negative numbers because it is a vector between the relative word positions in ST and TT. Its absolute value is thus indicative of the number of word distortions between ST and TT. Therefore, we take the absolute value of Cross before the log transformation to enable the correlation tests.

Table 4 shows relatively strong correlations between WHER and the three features of lexical variation. Although with slight differences, HTra, LogAltT, and ITra show a consistent moderate correlation with LogWHER, with the highest

**Table 4** Correlations between WHER and the indicators of product features (log-transformed)

| Category | Feature name | Pearson's *r* with LogWHER |
|---|---|---|
| Lexical variation | LogAltT | 0.51*** |
| | ITra | 0.70*** |
| | HTra | 0.54*** |
| Syntactic variation | LogCross | 0.30*** |
| | HCross | 0.36*** |

absolute correlation score being 0.61 for ITra and the lowest value being 0.54 for HTra. This means that less common translations require more operations than more frequent ones. For post-editing, it implies that rare translation alternatives are likely to relate with more edit operations. In general, the correlation results of the three indicators suggest that the effort of editing MT output can be well reflected by WHER.

For the features of syntactic variation, LogWHER correlates moderately but significantly with both HCross (0.36) and LogCross (0.30). The two correlation values are both lower compared with the above indicators of lexical variation. As the results of lexical and syntactic variation reflect the editing effort, the above results imply that lexical modifications on the MT output are more prominently reflected in the WHER score than syntactic modification. In other words, the overall editing effort indicated by WHER operations only overlaps partly with the effort of making syntactic choices.

## 5 Discussion and Conclusion

From the results presented above, we can see that the WHER correlates with both process and product measures of post-editing effort. Moderate to strong correlations are found between WHER and several features in keystroke, duration, and lexical variation. The correlation results support our assumption that the number of WHER operations reflects the post-editing effort.

Keystroke activities have stronger correlations with WHER than gaze data. The number of insertions has the strongest correlation with WHER, while the number of deletions is not correlated at all. The production duration indicating typing and pause effort has a moderate correlation with WHER. However, the weak correlation of the number of fixations and total fixation durations shows that the reading effort might not be well represented by the WHER operations. Effort indicators of making lexical modifications in the MT output have stronger correlations with WHER than the indicators of adjusting sentence structures. While all three lexical indicators are moderately correlated with WHER, ITra has the highest correlation, while HCross and Cross only have weak correlations with WHER. The correlation results are in line with the previous findings that a single measure is not robust enough to estimate the post-editing effort (Koponen 2012, Guerberof 2014). Our new WHER metric correlates better with keystroke activities (Key_ins) and the self-information of translation (ITra).

To sum up, this chapter introduces WHER as a measure to quantify the minimum per word edit operations. The study aims at finding out whether WHER is a reliable indicator of post-editing effort. By experimenting with audiovisual texts, we recorded participants' post-editing activities with keystroke-logging and eye-tracking devices. We collected the MT output and the final post-edited results and aligned the ST and TT. We computed the WHER score and correlated it with the process and product data on the word level. Our main contribution is the

development of the WHER score which extends the HTER score by Snover et al. (2006) and the HER score, as suggested by do Carmo (this volume, Chap. 1).

We have shown that WHER correlates with both process and product measures which indicate multiple aspects of post-editing effort. Measuring the edit operations mapped to each ST word, WHER provides a new perspective of looking at post-editors' effort. We find that it is possible to use WHER to estimate post-editors' typing activities, lexical modifications, and, to a lesser extent, word order changes in the target texts. However, as our data show, reading activities indicated by fixations are not associated with WHER. For future research, we would like to include more translation process recordings, language pairs, and text types to corroborate the usability of WHER. If WHER is proven to be stable in correlating with multiple process and product indicators of post-editing effort, it might be used as an indicator to estimate word post-editing difficulty and/or generate indications which ST fragments might be troublesome in MT and post-editing. Besides, as the sentence-based HTER score has been used in the translation industry (see Cumbreño and Aranberri this volume, Chap. 3, for a sentence-level HTER assessment) to predict post-editing difficulty and decide human pay rates, the WHER metric can help specify words that are harder to translate and thus encourage language service providers to make flexible strategies on translation and pricing.

## References

Armstrong S, Way A, Caffrey C et al (2006) Improving the quality of automated DVD subtitles via example-based machine translation. In: Proceedings of translating and the computer. Aslib, London. http://www.mt-archive.info/Aslib-2006-Armstrong.ppdf

Bangalore S, Behrens B, Carl M et al (2016) Syntactic variance and priming effects in translation. In: Carl M, Bangalore S, Schaeffer M (eds) New directions in empirical translation process research exploring the CRITT TPR-DB. Springer, Cham, pp 211–238

Basu P, Pal S, Naskar SK (2018) Keep it or not: word level quality estimation for post-editing. In: Proceedings of the third conference on machine translation (WMT). Association for Computational Linguistics, Brussels, pp 772–777

Burchardt A, Lommel A, Bywood L et al (2016) Machine translation quality in an audiovisual context. Target 28(2):206–221

Bywood L, Volk M, Fishel M, Georgakopoulou P (2013) Parallel subtitle corpora and their applications in machine translation and translatology. Perspectives 21(4):595–610

Campbell S (2017) Choice network analysis in translation research. In: Olohan M (ed) Intercultural faultlines: research models in translation studies I. Textual and cognitive aspects. Routledge, London, pp 29–42

Carl M (this volume) Information and entropy measures of rendered literal translation. In: Carl M (ed) Explorations in empirical translation process research. Springer, Cham

Carl M, Schaeffer M (2014) Word transition entropy as an indicator for expected machine translation quality. In: Miller KJ, Specia L, Harris K, Bailey S (eds) Proceedings of the workshop on automatic and manual metrics for operational translation evaluation. European Language Resources Association, Paris, pp 45–50

Carl M, Schaeffer MJ (2017) Why translation is difficult: a corpus-based study of non-literality in post-editing and from-scratch translation. J Lang Commun Bus 56:43–57

Carl M, Schaeffer M, Bangalore S (2016) The CRITT translation process research database. In: Carl M, Bangalore S, Schaeffer M (eds) New directions in empirical translation process research exploring the CRITT TPR-DB. Springer, Cham, pp 13–54

Cumbreño C, Aranberri N (this volume) What do you say? Comparison of metrics for post-editing effort. In: Carl M (ed) Explorations in empirical translation process research. Springer, Cham

Daems J, Vandepitte S, Hartsuiker R, Macken L (2017) Translation methods and experience: a comparative analysis of human translation and post-editing with students and professional translators. Meta 2:245–270

de Sousa SCM, Aziz W, Specia L (2011) Assessing the post-editing effort for automatic and semi-automatic translations of DVD subtitles. In: Proceedings of recent advances in natural language processing. Association for Computational Linguistics, Hissar, pp 97–103

DePalma DA, Pielmeier H, O'Mara P (2019) Who's who in language services and technology: 2019 rankings. Common Sense Advisory, Boston, MA. https://insights.csa-research.com/reportaction/305013039/Marketing

Díaz-Cintas J, Remael A (2014) Audiovisual translation subtitling. Routledge, New York, NY

do Carmo F (this volume) Editing actions: a missing link between translation process research and machine translation research. In: Carl M (ed) Explorations in empirical translation process research. Springer, Cham

Dragsted B (2012) Indicators of difficulty in translation—correlating product and process data. Across Lang Cult 13(1):81–98

Germann U (2008) Yawat: yet another word alignment tool. In: Proceedings of the 46th annual meeting of the association for computational linguistics. Association for Computational Linguistics, Columbus, pp 20–23

Guerberof A (2014) The role of professional experience in post-editing from a quality and productivity perspective. In: O'Brien S, Winther Balling L, Carl M et al (eds) Post-editing of machine translation: processes and applications. Cambridge Scholars, Newcastle Upon Tyne, pp 51–76

Hancock GR, Stapleton LM, Mueller RO (2018) The Reviewer's guide to quantitative methods in the social sciences, 2nd edn. Routledge, London

Ivarsson J, Carroll M (1998) Code of good subtitling practice. European Association for Studies in screen translation, Berlin. https://www.esist.org/wp-content/uploads/2016/06/code-of-good-subtitling-practice.pdf.pdf

Jia Y, Carl M, Wang X (2019) How does the post-editing of neural machine translation compare with from-scratch translation? A product and process study. J Spec Transl 31:60–86

Just MA, Carpenter PA (1980) A theory of reading: from eye fixations to comprehension. Psychol Rev 87(4):329–354

Koglin A (2015) An empirical investigation of cognitive effort required to post-edit machine translated metaphors compared to the translation of metaphors. Transl Interpret 7(1):126–141

Koponen M (2012) Comparing human perceptions of post-editing effort with post-editing operations. In: Proceedings of the seventh workshop on statistical machine translation. Association for Computational Linguistics, Montreal, QC, pp 181–190

Krings HP (2001) Repairing texts: empirical investigations of machine translation post-editing processes, Kent State University Press, Kent

Lacruz I, Ogawa H, Yoshida R, Yamada M, Martinez DR (this volume) Using a product metric to identify differential cognitive effort in translation from Japanese to English and Spanish. In: Carl M (ed) Explorations in empirical translation process research. Springer, Cham

Leijten M, Van Waes L (2013) Keystroke logging in writing research: using inputlog to analyze and visualize writing processes. Writ Commun 30(3):358–392

Martins AFT, Junczys-Dowmunt M, Kepler FN et al (2017) Pushing the limits of translation quality estimation. Trans Assoc Comput Ling 5:205–218

Melero M, Oliver A, Badia T (2006) Automatic multilingual subtitling in the eTITLE project. In: Proceedings of translating and the computer. Aslib, London. http://www.mt-archive.info/Aslib-2006-Melero.pdf

Moorkens J, O'Brien S, da Silva IAL et al (2015) Correlations of perceived post-editing effort with measurements of actual effort. Mach Transl 29:267–284

Mossop B (2007) Empirical studies of revision: what we know and need to know. J Spec Transl 8:5–20

Nitzke J (2019) Problem solving activities in post-editing and translation from scratch: a multi-method study. Language Science Press, Berlin

Ogawa H, Gilbert D, Almazroei S (this volume) redBird: rendering entropy data and ST-based information into a rich discourse on translation. In: Carl M (ed) Explorations in empirical translation process research. Springer, Cham

Orero P, Doherty S, Kruger JL et al (2018) Conducting experimental research in audiovisual translation (AVT): a position paper. J Spec Transl 30:105–126

Ortiz-boix C, Matamala A (2016) Post-editing wildlife documentary films: a new possible scenario? J Spec Transl 26:187–210

Sanchis-Trilles G, Alabau V, Buck C et al (2014) Interactive translation prediction versus conventional post-editing in practice: a study with the CasMaCat workbench. Mach Transl 28:217–235

Schaeffer M, Carl M (2014) Measuring the cognitive effort of literal translation processes. In: Proceedings of the workshop on humans and computer-assisted translation (HaCaT). Association for Computational Linguistics, Gothenburg, pp 29–37

Schaeffer M, Dragsted B, Hvelplund KT et al (2016) Word translation entropy: evidence of early target language activation during reading for translation. In: Carl M, Bangalore S, Schaeffer M (eds) New directions in empirical translation process research: exploring the CRITT TPR-DB. Springer, Cham, pp 183–210

Schaeffer M, Nitzke J, Tardel A et al (2019) Eye-tracking revision processes of translation students and professional translators. Perspectives 27(4):589–603

Snover M, Dorr B, Schwartz R et al (2006) A study of translation edit rate with targeted human annotation. In: Proceedings of the 7th conference of the Association for Machine Translation in the Americas. Association for Machine Translation in the Americas, Cambridge, pp 223–231

Vanroy B, De Clercq O, Macken L (2019) Correlating process and product data to get an insight into translation difficulty. Perspectives 27(6):924–941

Vanroy B, De Clercq O, Tezcan A, Daems J, Macken L (this volume) Metrics of syntactic equivalence to assess translation difficulty. In: Carl M (ed) Explorations in empirical translation process research. Springer, Cham

Vieira LN (2014) Indices of cognitive effort in machine translation post-editing. Mach Transl 28:187–216

Vieira LN (2016a) How do measures of cognitive effort relate to each other? A multivariate analysis of post-editing process data. Mach Transl 30:41–62

Vieira LN (2016b) Cognitive effort in post-editing of machine translation: evidence from eye movements, subjective ratings, and think-aloud protocols. Dissertation. Newcastle University, Newcastle Upon Tyne

Volk M (2008) The automatic translation of film subtitles: a machine translation success story? In: Nivre J, Dahllöf M, Megyesi B (eds) Resourceful language technology: Festschrift in honor of Anna Sågvall Hein. Uppsala University, Uppsala

Wei Y (this volume) Entropy and eye movement: a micro analysis of information processing in activity units during the translation process. In: Carl M (ed) Explorations in empirical translation process research. Springer, Cham

Xenouleas S, Malakasiotis P, Apidianaki M, Androutsopoulos I (2019) SUM-QE: a BERT-based summary quality estimation model. In: Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP). Association for Computational Linguistics, Hong Kong, pp 6005–6011