Post-editing time as a measure of cognitive effort

Maarit Koponen
Dept of Modern Languages
University of Helsinki
maarit.koponen@helsinki.fi

Wilker Aziz
RIILP
University of Wolverhampton
w.aziz@wlv.ac.uk

Luciana Ramos
Scientific Translator and Interpreter
and Professional Trainer
translationandtraining@gmail.com

Lucia Specia
Dept of Computer Science
University of Sheffield
l.specia@sheffield.ac.uk

Abstract

Post-editing machine translations has been attracting increasing attention both as a common practice within the translation industry and as a way to evaluate Machine Translation (MT) quality via edit distance metrics between the MT and its post-edited version. Commonly used metrics such as HTER are limited in that they cannot fully capture the effort required for post-editing. Particularly, the cognitive effort required may vary for different types of errors and may also depend on the context. We suggest post-editing time as a way to assess some of the cognitive effort involved in post-editing. This paper presents two experiments investigating the connection between post-editing time and cognitive effort. First, we examine whether sentences with long and short post-editing times involve edits of different levels of difficulty. Second, we study the variability in post-editing time and other statistics among editors.

1 Introduction

As Machine Translation (MT) becomes widely available for a large number of language pairs and the demand for faster and cheaper translations increases, its adoption is becoming more popular in the translation industry. However, it is well known that except in very narrow domains with dedicated MT systems, automatic translations are far from perfect. A common practice is thus to have human translators performing post-editing of such translations. Post-editing has also been attracting increasing attention from researchers in MT as a way of evaluating the quality of machine translations, particularly for the purpose of comparing various MT systems. Effective ways of measuring post-editing effort – and thus MT quality – in both scenarios is a very relevant but open problem.

Standard MT evaluation metrics have proved to correlate significantly better with human assessments of quality when computed having a post-edited version of the automatic translation as reference as opposed to translations created independently from automatic translations. One of these metrics is HTER – the "Human-targeted Translation Edit Rate" – (Snover et al., 2006), which was used as the official metric in the DARPA GALE program (Olive et al., 2011).

HTER is an edit distance metric that computes the minimum number of edits between the system output and its (often minimally) post-edited version. It is a simple metric which has nevertheless shown to be very effective. However, this and other metrics that estimate the similarity or distance between a system translation and its post-edited version have a crucial limitation: they cannot fully capture the effort resulting from post-editing such a translation. Certain operations can be more difficult than others, based not only on the type of edit (deletion, insertion, substitution), but also on the words being edited. Edits due to incorrect morphological variants or function words are generally treated the same way as more complex edits such as fixing an untranslated content word. While variants of such metric assigning weights for specific edits or classes of words can be implemented (Snover et al., 2010; Blain et al., 2011), defining classes of complex words to post-

edit requires a lexicalised, linguistically- motivated and thus language-dependent approach. In addition, the complexity of a correction cannot always be characterized based only on a local edit, as it may depend on the neighbourhood of that edit.

Recently, Koponen (2012) conducted an error analysis on post-edited translations with HTER and 1-5 scores assigned by humans for post-editing effort. A number of cases were found where post-edited translations with low HTER (few edits) were assigned low quality scores (high post-editing effort), and vice-versa. This seems to indicate that certain edits require more cognitive effort than others, which is not captured by HTER.

Post-editing effort consists of different aspects: temporal, technical and cognitive (Krings, 2001). However, these aspects are highly interconnected. The temporal effort (time spent on post-editing) is the most easily measurable. Post-editing time reflects not only the technical effort needed to perform the editing, but also the cognitive effort required to detect errors and plan the necessary corrections.

We believe that measuring *post-editing time* is the most cost effective and straightforward way of quantifying at least some of the cognitive effort involved in post-editing. In order to verify this hypothesis, in this paper we study measurements of post-editing time on a number of English-Spanish translations produced by eight MT systems and revised by eight translators. We follow a similar methodology as Koponen (2012), but focus on discrepancies between post-editing *time* and HTER. The main purpose of this experiment is to identify different groups of errors in MT and correlate them to different levels of difficulty that may be involved in fixing them, where difficulty is defined in terms of post-editing time. We are particularly interested in sentences that take a long time to edit but involve relatively few edit operations (low HTER) and are not excessively long. In addition, we use time and other detailed post-editing effort indicators, such as number of keystrokes, to analyse the variance between different translators post-editing the same translations.

The remainder of this paper is organized as follows. Section 2 presents previous attempts to measure post-editing effort. Section 3 describes the dataset and method used for our analysis and Section 4 shows the results of this analysis.

2 Related work

The most commonly used approach to quantifying post-editing effort (and in general translation quality) has been the use of semi-automatic MT evaluation metrics such as HTER that measure the similarity or distance between the MT system output and its human post-edited version. However, while such metrics provide a good indication of the technical effort involved in post-editing, they do not capture its cognitive effort. Koponen (2012) shows that translator's perception of post-editing effort, as indicated by scores in 1-5, does not always correlate well with edit distance metrics such as HTER. In other words, sentences scored as requiring significant post-editing sometimes involve very few edits, and vice-versa. What this suggests is that technical and cognitive effort are not always equal: certain types of errors require considerable cognitive effort although the number of technical operations is low, while others may involve a large number of technical operations but may be cognitively easier.

Blain et al. (2011) introduce the Post-Editing Action (PEA), a new unit of PE effort which is a more linguistically founded way of measuring a traditional edit distance. In their approach, rather than treating each edited word as a separate action, PEAs incorporate several interrelated edit operations. For example, changing a noun propagates changes to its attributes (number, gender) which are then treated as one action. This approach has the disadvantages that it is hardly generalizable across languages, and it requires annotated corpus to train a model to classify PEAs for new texts.

A practical alternative, measuring *time* as a way of assessing post-editing effort, has only recently started to be used by researchers, although we believe this may be a more common practice in the translation industry.

Tatsumi (2009) examines the correlation between post-editing time and certain automatic metrics measuring textual differences. They find that the relationship between these two measures is not always linear, and offer some variables such as source sentence length and structure as well as specific types of dependency errors as possible explanations.

(Temnikova and Orasan, 2009; Temnikova, 2010) contrast the time translators spent fixing transla-

tions for texts produced according to a controlled language, versus translations produced using non-controlled language.

Sousa et al. (2011) compare the time spent on post-editing translations from different MT systems and on translating from scratch. The study has shown that sentences requiring less time to post-edit are more often tagged by humans as demanding low effort. It has also shown that post-editing time has good correlation with HTER for ranking both systems and segments.

Specia (2011) uses post-editing time as a way of evaluating quality estimation systems. A comparison is made between the post-editing of sentences predicted to be good and average quality sentences, showing that sentences in the first batch can be post-edited much faster.

Focusing on sub-segments, Doherty and O'Brien (2009) use an eye-tracker tool to log the fixation and gaze counts and time of translators while reading the output of an MT system. Overall translation quality was quantified on the basis of the number and the duration of fixations. Results show that fixation counts correlate well with human judgements of quality.

Following a similar approach, O'Brien (2011) measures correlations between MT automatic metrics and post-editing productivity, where productivity is measured using an eye tracker. Processing speed, average fixation time and count are found to correlate well with automatic scores for groups of segments.

Except the two latter approaches – which require eye-trackers –, to the best of our knowledge, no previous work focuses on using post-editing time as a measure of cognitive effort, and on how it correlates with technical effort. Using post-editing time for that has a number of open issues, such as the fact that it can vary significantly for different translators. In this paper we present some initial work in these two directions.

3 Materials and method

The data we have used for the experiments consists of English sentences machine translated into Spanish using eight MT systems, randomly selected from the WMT11 workshop dataset (Callison-Burch et al., 2011). The dataset includes 299 source sentences translated by two or more systems, resulting in 1484 translations. These systems were chosen based on the overall system ranking reported by WMT11: a manual evaluation had ranked the 15 participating systems in eight groups, where within each group the difference in performance was not found to be statistically significant. Within each group, we randomly picked a system: *cu-zeman*, *koc*, *online-A*, *rbmt-2*, *rbmt-4*, *rbmt-5*, *uedin*, *uow*.

	# main	# common	# all
SRC	279	20	299
SYS	8	8	8
MT	1464	20	1484
PE	1464	160	1624
MT/SRC	5.24	1	5.43

Table 1: Characteristics of the datasets: number of source sentences (SRC), systems being compared (SYS), machine translations (MT), post-edited translations (PE) and translations per source sentence (MT/SRC).

The machine translations were then edited by eight native Spanish speaking post-editors, who either were professional translators (six cases) or had some experience with post-editing (two cases). The 1484 translations were split to form two disjoint datasets (Table 1): i) a small dataset of 20 translations (one from each of 20 different sources) from randomly selected systems, and ii) a dataset made of the other 1464 translations (outputs of different systems to the remaining 279 sources - just over 5 translations per source). The first dataset (common) was edited by all the eight post-editors, that is, all of them post-edited the same 20 machine translations. The machine translations in the second dataset (main) were randomly distributed amongst the post-editors so that each of them only edited one translation for a given source sentence, and all of them edited a similar number of translations from each MT system (on average 23 per system). In sum, each post-editor edited 203 sentences (20 in common and 183 in main).

For each translation, post-editing effort indicators were logged using PET,[1] a freely available post-editing tool (Aziz et al., 2012). Among these indicators, of particular interest to our study are:

[1] http://pers-www.wlv.ac.uk/~in1676/pet/

- **TIME** the post-editing time of a sentence;
- **SPW** seconds per word, that is, the **TIME** the translator spent to post-edit the sentence divided by the length (in tokens) of the post-edited translation;
- **KEYS** the number of keystrokes pressed during the post-editing of the sentence (**PKEYS** is the subset of printable keys, that is, those that imply a visible change in the text); and
- **HTER** the standard edit distance between the original machine translation and its post-edited version (Snover et al., 2006).

Keystrokes and edit distance are natural candidates for measuring post-editing effort. To understand the usefulness of post-editing time (and its normalized version **SPW**) for this purpose, we first observed the performance of these time-based indicators at discriminating MT systems for quality. For that, we compare the system-level ranking reported by WMT11 with the rankings suggested by these indicators via Spearman's rank correlation coefficient ρ. In Table 2, the first column shows WMT11's official ranking - the numeric value is the percentage of times that the given system is better than any other system(s). The following columns show the rankings obtained by other indicators - the numeric value is the average score of each system in the `main` dataset according to that indicator. The last row shows the Spearman's rank correlation between the ranking of the gold standard (WMT11) and the ranking of each given metric. The time-based indicators, specially **TIME**, achieved a much stronger correlation with the gold standard ranking.

This initial analysis indicated that time can be a good metric to understand post-editing effort and translation quality. We then moved on to studying this metric in more detail at sentence and subsentence levels. More specifically, we analyse the `main` and `common` datasets in order to answer the following research questions, respectively:

- Can we characterise edits that require more cognitive effort from post-editors based on post-editing time?
- How do post-editors differ in terms of the time they spend, final translations they produce and strategies they use when post-editing?

The details on the methods used to address these two questions are given in the following sections.

3.1 Cognitive effort in post-editing

Our focus was on finding sentences that required a long time to edit and could therefore be expected to contain errors that are particularly difficult for the editor to correct. One relatively simple explanation for long editing time is sentence length (Tatsumi, 2009; Koponen, 2012). In order to target sentences where long editing time cannot be explained by sentence length alone, we chose to focus on post-editing time normalized by number of tokens in the translation. Long editing times can also be explained by the amount of editing needed in the sentence: low quality translations will naturally require more editing, but this does not necessarily mean that the edits are difficult. We thus decided to exclude cases where the sentence had undergone significant rewriting. For that, we used HTER and the observed edit operations performed as logged by PET to target sentences where relatively few changes had been made. These two indicators are different: while HTER only counts the operations that resulted in the changing of the translation, PET counts operations that were performed without necessarily changing translations, e.g, if a word is deleted and reinserted in its original form, one replacement operation is still counted.

The selection of potentially interesting examples of post-edited translations for this analysis was done with the aid of plots to visualise cases of high **SPW** and low **HTER** for each post-editor separately, to avoid any bias due to the variance in post-editing time across post-editors. For each post-editor, the four cases with the combination of longest **SPW** and lowest **HTER** were selected. A comparison set from the same post-editor with similar sentence length and similar **HTER** but short-to-average **SPW** was also selected. This was done for all post-editors, resulting in 32 cases of each type.

We then manually analysed these sentences to check if the types of errors edited differ in the two groups. Our hypothesis is that sentences with short editing times should contain more of the easy to fix errors and sentences with long edit times more of the difficult to fix errors. For error types and their level of cognitive effort, we used the 10 classes proposed in (Temnikova, 2010) with some modifica-

WMT11 ↑		KEYS (↓)		HTER (↓)		SPW (↓)		TIME (↓)	
online-A	0.72	uedin	56.29	online-A	0.229	online-A	3.06	online-A	64.48
uedin	0.64	online-A	57.04	uedin	0.242	rbmt-2	3.32	uedin	71.49
rbmt-4	0.61	rbmt-2	71.09	rbmt-2	0.281	uedin	3.33	uow	77.69
uow	0.59	rbmt-5	73.44	rbmt-5	0.291	rbmt-4	3.48	rbmt-4	78.07
rbmt-2	0.57	rbmt-4	73.81	rbmt-4	0.304	uow	3.58	rbmt-2	81.76
koc	0.56	uow	89.08	uow	0.306	rbmt-5	3.69	rbmt-5	85.20
rbmt-5	0.54	cu-zeman	94.36	koc	0.325	koc	3.84	koc	86.42
cu-zeman	0.49	koc	94.52	cu-zeman	0.331	cu-zeman	4.26	cu-zeman	100.32
Spearman's ρ		0.667		0.738		0.833		0.952	

Table 2: System-level rank correlation of each metric and WMT11's official ranking.

tions. Temnikova (2010) enriches a standard error classification (Vilar et al., 2006) for MT by ranking the error categories according to how cognitively costly she expects them to be. In addition, we hypothesise that longer edit times may involve more content words, e.g. verbs, nouns; while shorter times may involve more function words, e.g. determiners. We therefore further hypothesise that part-of-speech (POS) errors may be linked to longer edit times. Our adaptation of the classification in (Temnikova, 2010) resulted in the following error categories, from the easiest to the most difficult to fix:

0 Typographical: upper/lower case or similar orthographical edits

1 Incorrect word form

2 Incorrect style synonym: word substitutions that do not change the meaning

3 Incorrect word: divided into three cases

- **3a** Different word but same POS
- **3b** Different POS
- **3c** Untranslated source word in MT

4 Extra word

5 Missing word

6 Idiomatic expression missed

7 Wrong punctuation

8 Missing punctuation

9 Word order, word level

10 Word order, phrase level

This adaptation involved the addition of a category for orthographical edits, which is here assumed to be the easiest type. Category 3, "Incorrect word", was found to consist of different types of cases which might have cognitively different effects on the reader: an incorrect word that is the same POS as the correct one may not interfere with understanding of the sentence structure in the same way as a word that is also incorrect POS (e.g. noun instead of a verb) or an untranslated source language word. For this reason, we divided the category into three subcategories: different word but same POS, different POS, and untranslated word.

The sentences selected were lemmatised and POS tagged using the FreeLing software (Padró et al., 2010). The operation logs created by PET were used to track the changes made by the editors and then insertions, deletions and substitutions were labelled according to the error classification discussed above. Cases where the operations logged did not correspond to any changes visible in the final post-edited sentence, meaning typos and corrections made by an editor or cases where the editor revised their correction of some word or phrase several times, were not included in any error category.

3.2 Human variability in post-editing

The goal of this experiment is to analyse some aspects of the human variability in post-editing to understand whether any findings obtained using indicators from the post-editing process generalise across translators. A significant variance in segment-level post-editing time is not surprising: it is expected that different translators spend more or less time to edit the same translation, depending on their experience with the task, language-pair, text domain, etc. A variance in the final revised translations is also expected in some cases, as there is generally more than one way of expressing the source segment meaning. We were thus more interested in studying variations

in the strategies used by post-editors.

We used the 20 cases from the `common` dataset that had been edited by all eight translators. These were the last translations done by all editors. We analysed the operation history logs stored by PET to observe the changes made by the editors, post-editing time, HTER and keystroke counts, including not only the overall keystroke count, but also counts on groups of specific keys pressed:

- White keystrokes: space, tab and enter
- Alphanumeric: letters (including diacritical marks) and digits
- Control: delete, backspace, combinations such as ctrl+c etc.

We hypothesise that there may be differences in the amount of "visible" typing (alphanumeric and white keys), which would reflect the individual editors' choices of how much they chose to change the translations, but also in the use of control keys, for example some editors use the arrow keys to move around in the sentence while reading and editing.

4 Results

Figure 1 shows the correlations between **TIME**, **SPW**, **HTER**, and **PKEYS** and sentence length (LEN) in the `main` data. While, as expected, absolute post-editing time grows with sentence length (5th row, 2nd col.) and number of printable keystrokes (5th row, 3rd col.), **SPW** remains fairly constant (4th row, 2nd col.). Focusing on **HTER** vs. **TIME** (5th row, 1st col.) and **HTER** vs. **PKEYS** (3rd row, 1st col.), we can see that these have most of their points concentrated at around HTER=0.5. Those regions not only contain the majority of the points (which ultimately characterises the average **TIME** and **PKEYS**), but also the highest figures for both indicators, suggesting that although HTER reflects what the final translation looks like compared to the MT, it does not reveal much about the effort required to produce that final result in terms of time and keystrokes.

4.1 Cognitive effort in post-editing

The distribution of errors in the classes adapted from (Temnikova, 2010) is shown in Figure 2. The overall pattern observed with the error distribution is that

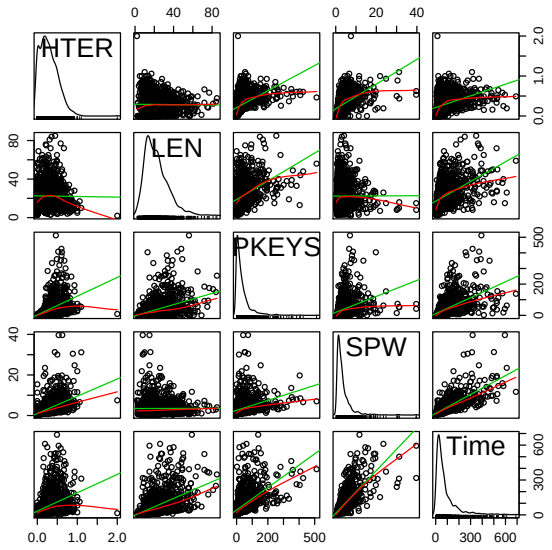


Figure 1: Scatter plot of the effort indicators: each cell in the main diagonal represents the distribution of the variable in that cell (**HTER**, **LEN**, etc.); the remaining cells correlate the variable in its column (projected on the x-axis) to the variable in its row (projected on the y-axis), and the lines are the best linear and the best polynomial fit.

errors assumed to be most cognitively difficult (idioms, punctuation and word order errors) are indeed more common in sentences with longer editing time.

For both sentences with short and long editing times, the most common errors involve category 3: "Incorrect word", with 29% in sentences with long editing time and 27% in those with short editing time. However, the distribution within the three subcategories differs: sentences with long editing time have larger proportion of the cases assumed to be more difficult, where the incorrect word has also wrong part of speech (3b) or is an untranslated source word (3c). Most of the cases in all subcategories involve nouns or verbs. Sentences with long editing times also include some 3b cases where a noun or a verb was mistranslated as an adverb. Such cases were not found in the set with short editing times.

For the sentences with short editing time, the second most common type of errors is incorrect form of a correct word (1). This is a less common type of errors in sentences with long editing times. Most

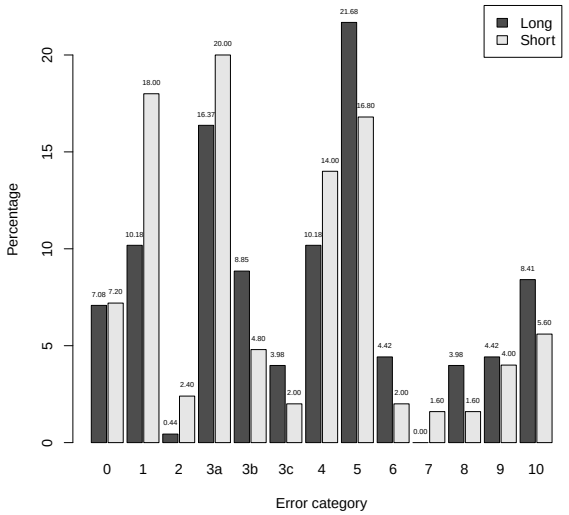


Figure 2: Comparison of error types between sentences with long and short editing times

word form errors involve verbs or determiners, but sentences with short editing times have a higher proportion (18%) of nouns with incorrect forms than those with long editing times (4%). The relative differences in proportions of word form errors between sentences with short or long editing times appears to support the ranking of these errors as "easy to fix".

Missing words are more common in sentences with long editing times, but extra words are more common in those with short editing times. In both cases, these are mostly function words, determiners and prepositions being the largest groups. The proportion of content words is larger in sentences with long editing times with one exception: sentences with short editing times contain a few more cases of extra verbs.

Errors related to mistranslated idioms, punctuation and word order are not very common overall in either set. Mistranslated idioms involve cases where an idiomatic expression has been translated literally word for word, often changing or obscuring the meaning of the original (e.g. *(to be) at odds* translated as *en probabilidades*, literally 'at probabilities'). They are slightly more common in sentences with long editing times. Another type of literal translation can be seen where a proper noun has been erroneously translated with a common noun (e.g. as *Marca de Stanley* 'the brand of Stanley' a for a person's name *Stanley Brand*) or an adjective (*mala Homburg* = 'bad Homburg' for the German place name *Bad Homburg*). Such errors were only found in the sentences with long editing time.

Cases of missing punctuation are more common in sentences with long editing times, and involve mainly missing commas. Cases of wrong punctuation (extra or replaced with other punctuation), on the other hand, were only found in the sentences with short editing times. However, at least on the surface, these few cases do not appear to be particularly critical to the understanding of the sentence: for example, substituting a comma for a semicolon or deleting an extra quotation mark. Although certain types of punctuation errors can have an effect on the meaning of a sentence by changing or obscuring the parsing of phrases, punctuation errors as a whole may not be cognitively as difficult as assumed in (Temnikova, 2010)'s classification.

Word order errors on the word level (e.g. transposition of nouns and adjectives) are about equally common in both types of sentences, but the need for reordering on the phrasal level is more common in sentences with long editing times. Furthermore, for sentences with long editing times this generally involves cases where individual words need to be reordered sometimes by long distances (and affecting the parsing of the sentence). In contrast, in sentences with short editing times about half the cases in this category involve moving groups of words into different location as whole phrases.

4.2 Human variability in post-editing

The comparison of cases where all editors post-edited the exact same machine translations show that even with the same sentence and same instructions, different editors approach the task in different ways.

Figure 3 shows the Pearson correlation coefficient for pairs of post-editors from the perspective of two different scores, namely **HTER** (bottom-left half of the figure) and **TIME** (top-right half of the figure), where darker cells indicate stronger correlation. We note that the **HTER** half is on average darker than the **TIME** half. This contributes to the hypothesis that although editors may apply a similar number of edits to the machine translation, the time they take to do it varies.

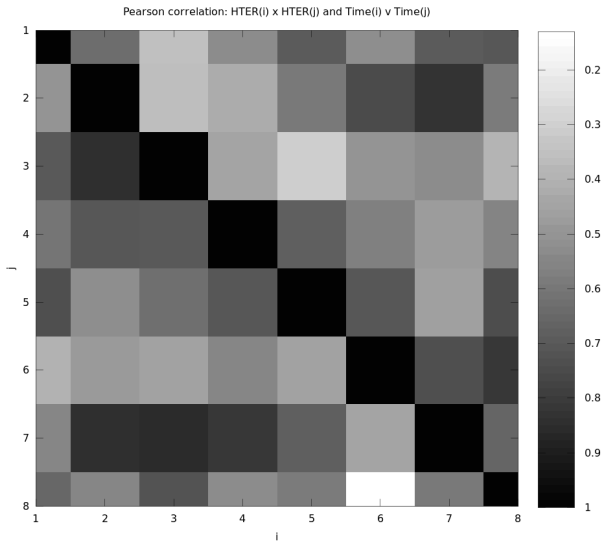


Figure 3: Each cell represents the Pearson correlation coefficient for a pair of post-editors according to **HTER** (bottom-left half) or **TIME** (bottom-right half) in the `common` dataset; darker cells indicate stronger correlation. The comparison of diagonally symmetric cells shows whether a pair of post-editors "agrees" more in terms of **HTER** or **TIME** (e.g. $\text{HTER}(3,5) > \text{TIME}(5,3)$; and $\text{HTER}(6,8) < \text{TIME}(8,6)$)

These variations may be explained by different post-editing strategies, which can be observed by comparing the different metrics: **SPW**, **HTER** and **KEYS**. Box plots for these metrics by post-editor are shown in Figure 4.

Two editors, A6 and A7, are the fastest, with the shortest editing time in 14 out of 19 sentences (in 1 case, none of the editors made any changes). On the other hand, the two slowest editors (A5 and A8) took the longest time in 11 out of the 19 cases.

In some cases, time relative to others does seem to reflect the amount of editing: the editor with the overall shortest editing times (A6) also has the lowest average **HTER**, and the two slowest editors (A5 and A8) have the highest **HTER** scores. Some differences do, however, appear: editor A4, whose editing times are third slowest of the eight editors, has in fact the second lowest **HTER**. In contrast, the second fastest editor, A7, has a considerably higher average HTER. Similarly for keystroke counts, some combine short/long editing times with low/high keystroke counts as might be expected, but despite relatively long editing times, A4 in fact uses less keystrokes on average than the two fastest editors.

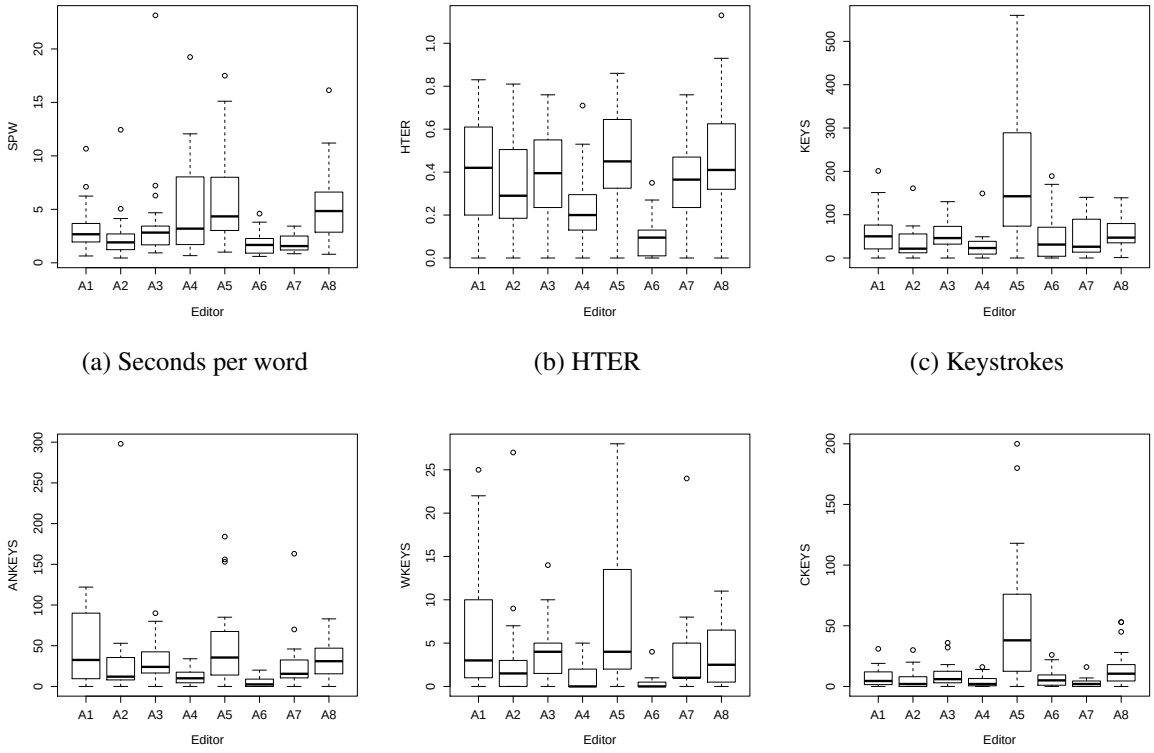
In addition to choices on how much to edit the MT sentence, some differences in post-editing times and keystrokes can also be explained by *how* the editor carries out the edits. Some editors appear make more use of the words already present in the MT as well as using cut-and-paste operations whereas others apparently prefer to type out their own version and deleting the MT words even if some parts are identical. Examples might be A7 (low **KEYS** but relatively high **HTER**) versus A1, A5 and A8 (relatively high **KEYS** and high **HTER**).

Some editors also seem to plan their edits beforehand, and edit what needs correcting only once, while others change the same passage several times or start and then delete several edits before settling on a final version. Examples may be displayed by A4 (relatively long **SPW** despite low **HTER** and low **KEYS**) versus A5 and A8 (long **SPW** combined with high **HTER** and high **KEYS**). Different editors also make the edits within the sentence in a different order, some proceeding from left to right while others move around between different parts of the sentence. Moving around inside the sentence with arrow keys may be one explanation for the very high keystroke count, and particularly high control key count by A5.

5 Conclusions

The goal of this study was to examine two questions: (i) can we characterise edits that require more cognitive effort from post-editors based on post-editing time? (ii) how do post-editors differ in terms of the time they spend, final translations they produce and strategies they use when post-editing?

The first experiment compared post-edited sentences with a long editing time to sentences with similar length and edit distance but short editing times. The errors that post-editors had corrected in these sentences were analysed according to a cognitively motivated error difficulty ranking (Temnikova, 2010), and the results suggest that the type of errors affects post-editing time. Shorter editing times seem to be associated with errors ranked cog-



(a) Seconds per word (b) HTER (c) Keystrokes

(d) Alphanumeric keys (e) White keys (f) Control keys

Figure 4: Post-editors' effort indicators in the `common` dataset.

nitively easiest, which include word form errors, synonym substitutions, and "simple" incorrect word substitutions where changing the part-of-speech is not necessary. On the other hand, substitutions involving an incorrect part-of-speech or an untranslated word, errors related to idiomatic expressions and word order, especially when reordering crosses phrase boundaries, seem to be connected with longer edit times.

These results may suggest some revisions to the assumed difficulty ranking. Sentences with short editing times in fact contained more errors labelled as extra words than sentences with long editing times. As the majority of extra/missing cases involved function words, this may indicate that extra words are not as cognitively challenging as assumed at least when they involve function words. Similarly, punctuation errors, which in turn were relatively rare in both types of sentences, showed little difference between the sentence types, and incorrect (as opposed to missing) punctuation was only found in the sentences with short editing times. Although there are certain situations where missing or incorrect punctuation could change or obscure the meaning of a sentence, perhaps not all punctuation errors need to be ranked as equally difficult.

In the second experiment, we examined post-editing effort indicators from different editors revising the same translations. Studying their variation in terms of time, edit distance and keystrokes suggests certain different editing strategies. Firstly, even with the same instructions to minimally change the machine translations, different editors make different choices about what constitutes minimal. Secondly, some editors maximize the use of MT words and cut-paste operations for reordering, while others appear to prefer writing out the whole corrected passage and then deleting MT words even when they are the same. Thirdly, some editors spend their time planning the corrections first and proceeding in order while others revise their own corrections and move around in the sentence. This could be an in-

dication that keystrokes, while very useful as a way to understand how translators work, may not be an appropriate measure to estimate cognitive effort.

Further work is needed for truly identifying cognitively difficult errors, including analyses with larger sets, as well as different language pairs, but we believe post-editing time is a variable that should certainly be considered in analyses of this type. In addition to sentence-level post-editing time, investigating editing times related to specific operations within the sentences could provide useful information on where editors spend their time. A revised set of error categories with more detailed error types (e.g "incorrect main verb", "incorrect prepositional attachment") is also an interesting direction to help understand the cognitive load in post-editing.

Studying the strategies of different post-editors can be potentially very useful for post-editing practice. Larger scale tests with editors editing the same translations, particularly where their backgrounds and levels of experience are similar would help understand whether the variances are systematic or very specific to individual translators.

References

Wilker Aziz, Sheila C. M. Sousa, and Lucia Specia. 2012. PET: a Tool for Post-editing and Assessing Machine Translation. In *8th International Conference on Language Resources and Evaluation*, Istanbul, Turkey.

Frédéric Blain, Jean Senellart, Holger Schwenk, Mirko Plitt, and Johann Roturier. 2011. Qualitative analysis of post-editing for high quality machine translation. In *MT Summit XIII*, Xiamen, China.

C. Callison-Burch, P. Koehn, C. Monz, and O. Zaidan. 2011. Findings of the 2011 Workshop on Statistical Machine Translation. In *6th Workshop on Statistical Machine Translation*, pages 22–64, Edinburgh, Scotland.

Stephen Doherty and Sharon O'Brien. 2009. Can MT Output be Evaluated through Eye Tracking? In *MT Summit XII*, pages 214–221, Ottawa, Canada.

Maarit Koponen. 2012. Comparing human perceptions of post-editing effort with post-editing operations. In *7th Workshop on Statistical Machine Translation*, pages 181–190, Montréal, Canada.

Hans P. Krings. 2001. *Repairing texts: Empirical investigations of machine translation post-editing process*. The Kent State University Press, Kent, OH.

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

Joseph Olive, Caitlin Christianson, and John McCary. 2011. *Handbook of Natural Language Processing and Machine Translation: DARPA Global Autonomous Language Exploitation*. Springer.

Lluís Padró, Miquel Collado, Samuel Reese, Marina Lloberes, and Irene Castellón. 2010. FreeLing 2.1: Five Years of Open-Source Language Processing Tools. In *7th International Conference on Language Resources and Evaluation*, pages 3485–3490.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A Study of Translation Edit Rate with Targeted Human Annotation. In *7th Conference of the Association for Machine Translation in the Americas*, pages 223–231, Cambridge, Massachusetts.

Matthew Snover, Nitin Madnani, Bonnie Dorr, and Richard Schwartz. 2010. TER-Plus: Paraphrase, Semantic, and Alignment Enhancements to Translation Edit Rate. *Machine Translation*, 22(2-3):117–127.

Sheila C. M. Sousa, Wilker Aziz, and Lucia Specia. 2011. Assessing the post-editing effort for automatic and semi-automatic translations of DVD subtitles. In *Recent Advances in Natural Language Processing Conference*, Hissar, Bulgaria.

Lucia Specia. 2011. Exploiting objective annotations for measuring translation post-editing effort. In *15th Conference of the European Association for Machine Translation*, pages 73–80, Leuven, Belgium.

Midori Tatsumi. 2009. Correlation between Automatic Evaluation Metric Scores, Post-Editing Speed, and Some Other Factors. In *MT Summit XII*, pages 332–33.

Irina Temnikova and Constantin Orasan. 2009. Post-editing Experiments with MT for a Controlled Language. In *International Symposium on Data and Sense Mining, Machine Translation and Controlled Languages*, Besançon, France.

Irina Temnikova. 2010. A Cognitive Evaluation Approach for a Controlled Language Post-Editing Experiment. In *7th International Conference on Language Resources and Evaluation*, Valletta, Malta.

David Vilar, Jia Xu, Luis Fernando D'Haro, and Hermann Ney. 2006. Error Analysis of Machine Translation Output. In *5th International Conference on Language Resources and Evaluation*, pages 697–702.