Measuring Post-editing Time and Effort for Different Types of Machine Translation Errors

Anna Zaretskaya
University of Málaga, SPAIN

Mihaela Vela
University of Saarland, GERMANY

Gloria Corpas Pastor
University of Málaga, SPAIN

Miriam Seghiri
University of Málaga, SPAIN

ABSTRACT

Post-editing (PE) of machine translation (MT) is becoming more and more common in the professional translation setting. However, many users refuse to employ MT due to bad quality of the output it provides and even reject post-editing job offers. This can change by improving MT quality from the point of view of the PE process. This article investigates different types of MT errors and the difficulties they pose for PE in terms of post-editing time and technical effort. For the experiment we used English to German translations performed by MT engines. The errors were previously annotated using the MQM scheme for error annotation. The sentences were post-edited by students in translation. The experiment allowed us to make observations about the relation between technical and temporal PE effort, as well as to discover the types of errors that are more challenging for PE.

KEYWORDS: post-editing, machine translation, error typology, post-editing effort

1. Introduction

Post-editing (PE) of machine translation (MT) is usually understood as "a human being (normally a translator) comparing a source text with the machine translation and making changes to it to make it acceptable for its intended purpose" (Koby 2001:1). In recent years PE has become a common

practice in the translation industry. One of the leading language service providers SDL[1] reported post-editing 25% of all the company's translation orders and was expecting this figure to increase, including incorporating more languages (Stevens and Filipello 2014). This is not surprising as a number of studies have shown that post-editing of machine translation can boost productivity and save costs (Laübli et al. 2013, Plitt and Masselot 2010, Zampieri and Vela 2014, Zhechev 2014).

However, it is still recognized that the use of MT followed by post-editing is not suitable in every translation scenario. Many translators remain reluctant for reasons relating to their working languages, subject domain, working practices and, most importantly, the quality of translation offered by MT engines. Evaluation of machine translation output in terms of its usability for post-editing can help to overcome this problem. One possible approach to this task consists in investigating which MT errors have more influence on the post-editing process and introduce more challenges for post-editors. This knowledge can then be used for MT evaluation by, for instance, assigning weights to more "difficult" errors. It can also help improve MT systems and increase usability of post-editing software by taking into account the difficulty of different errors. Finally, this knowledge can be used for post-editing training.

In this article we present the results of an experiment that aims at investigating how different types of MT errors influence the post-editing process. To this end, we consider different aspects of post-editing effort. According to Krings (2001), there are three types of post-editing effort: temporal, cognitive and technical. In our research we only consider the temporal and the technical types. A study on the cognitive effort with the same post-edited data is planned within further research, as measuring cognitive effort requires either special software or a time-consuming annotation process. Temporal effort refers to the time needed to post-edit a segment, to which we will henceforward refer as post-editing time, or *PE time*. Technical effort consists of the changes made in the MT output to obtain the final version. It can be represented by the number of keystrokes or the number of deletions, insertions and substitutions made. In this research we will use a quantitative measure of technical effort called *post-editing effort (PEE)*, which reflects the number of editing operations

[1] http://www.sdl.com/ (accessed 27 January 2016).

necessary to transform the MT output into the final version.[2] This measure, along with others, will be further used to describe the experiment results as a numerical indicator of the amount of technical effort made during post-editing and will be explained in greater detail in Section 3.2.

Participants in the experiment were students (German native speakers) enrolled in a translation studies programme and familiar with MT concepts and post-editing. The students were given a post-editing task, where the errors made by an MT engine were previously annotated by language professionals. This would allow us to observe the effect of different errors on the PE process. Thus, the students' task consisted in correcting these specific errors. The PE time and PEE were then compared among different error types, which allowed us to draw conclusions about their post-editing difficulty.

The remainder of this article is structured as follows: Section 2 presents relevant work on the subject of MT errors and post-editing; Section 3 presents our methodology, including the data we used (3.1) and the experiment setup (3.2); Section 4 presents the results of this study; and, finally, Section 5 discusses the findings and outlines some ideas for future research.

2. Methodological Background and Related Research

In recent years several studies have been carried out which investigate different aspects of the post-editing process, including different post-editing effort indicators and their dependency on MT (Gupta et al. 2015, Moorkens et al. 2015, Lacruz and Shreve 2014, Scarton et al. 2015). In this section, we draw particular attention to works that, similarly to our own study, investigate different types of MT error in relation with post-editing process. Some of the methodological aspects described here will be taken up in the subsequent sections of the paper. By way of example, Koponen (2012) considers differences in cognitive and technical effort in order to identify types of errors that require more cognitive effort and are therefore harder to correct. The data used for the study consists of a corpus of sentences that were manually annotated for cognitive post-editing effort. Then, the actual Translation Edit Rate (TER) score (Snover et al. 2006) for these sentences

[2] Unless specified otherwise, we will further use the term *post-editing effort (PEE)* to refer to this specific measure of technical post-editing effort as opposed to the broad sense defined by Krings (2001).

was calculated. The TER metric compares the machine translation of a sentence with its edited version and counts the minimum word-level changes that are needed to transform the MT output into the final version. The number of changes is then divided by the number of words in the edited sentence to obtain the TER score, which is attributed values between 0 (when no changes are made) and 1 (when the entire sentence is changed). The aim of this study was to investigate what factors caused the difference between the perceived cognitive effort (represented by the manual score) and the technical effort represented by the TER score.

The author discovered that one of the reasons is sentence length: long sentences tend to obtain low human scores even if there are few changes to be made. Reordering operations are also related to low human scores, which means that word order error types seem difficult to correct. Examination of errors in different parts of speech showed that errors in open class words (i.e. parts of speech that constantly acquire new members, like nouns, verbs, adjectives and adverbs) tend to correlate more with human effort scores than in closed class words (prepositions, determiners, conjunctions, etc), which suggests that open class words are more cognitively difficult to correct. Word form errors, on the contrary, turned out to be the easiest to rectify, as well as determiner errors (Koponen 2012).

In a related study, Koponen et al. (2012) consider how different types of errors influence post-editing time. The study is based on the assumption that there is a ranking of cognitive difficulty of errors, namely the one proposed by Temnikova (2010), and that the PE time depends on the error type and its position in the ranking. Indeed, their results showed that shorter editing times were related to cognitively easier errors, namely word form errors, synonym substitutions, and incorrect word substitutions within the same part-of-speech. On the other hand, substitutions involving an incorrect part-of-speech or an untranslated word, errors related to idiomatic expressions and word order errors seemed to render longer editing time.

Some of the abovementioned results are in line with a more recent study by Popović et al. (2014), who explore the relations of five different types of edit operations with cognitive and temporal PE effort. Cognitive effort is again represented by manually assigned difficulty scores. The operation types are based on *edit distance*, which is a metric used to measure the difference between two sentences and which reflects the minimal number of deletions, insertions and substitutions needed

to transform one sentence into another. The operations include correcting word form, correcting word order, adding omission, deleting addition, and correcting lexical choice. Similarly to Koponen (2012), the analysis revealed that word order errors correlate most strongly with the perceived cognitive effort, together with mistranslations, and that lexical errors require the longest post-editing time. Another interesting finding is the strong correlation between edit rates and cognitive effort regardless of the PE time, which contradicts the results discussed above. The fact that edit rate was more closely related to cognitive effort than to PE time highlights the need to investigate how PE time and technical PE effort are related, which we plan to do in this study. PE time, in its turn, proved to be strongly dependent on sentence length.

In a somewhat similar study, Lacruz et al. (2014) consider different types of MT errors while comparing cognitive demand and cognitive effort. Cognitive demand is closely related with MT utility and expresses the usefulness of MT output for producing a correct translation. Cognitive effort is the actual effort needed to identify the errors and plan the necessary corrections during post-editing. They found that mistranslations, omissions, additions and syntax errors have stronger correlation with cognitive demand and cognitive effort (for instance, as indicated by pause to word ratio) than "less cognitively challenging errors" such as punctuation and word form errors. The error classification in this study was based on the ATA grading rubric.[3]

Another study conducted by Daems et al. (2015) measured the effect of different types of MT errors on different indicators of PE effort: average number of production units, average duration per word, average fixation duration, average number of fixations, pause ratio and average pause ratio. In order to calculate the effect coefficient, they used linear mixed models. Both the classification of errors and the annotation was performed by the authors. In this study, they discovered that the most common error types affecting PE effort indicators are mistranslations, structural issues and word order issues. But more importantly, they found that different error types affect different PE effort indicators. Thus, errors such as mistranslation, grammar, structure, and word order affect the pause ratio, the average pause ratio and the average number of production units, which are therefore related to each other. The average duration per word is affected the most by coherence and structure

[3] http://www.atanet.org/certification/aboutexams_error.php (accessed 27 January 2016).

issues. Fixation duration is strongly related to mistranslation errors. The average number of fixations is predicted by coherence issues, which is a more cognitively demanding error type. It is not always clear, however, how the effects of different error types interact with each other, i.e. when there are several error types in one segment it cannot be determined in which exact place the fixations or the pauses took place. Although we do not consider the same PEE indicators in our work, it will be interesting to compare these results with our own in terms of their impact on overall PE outcomes.

Temnikova (2010) investigated how pre-editing source text using a given controlled language influences the cognitive effort involved in correcting different error types. The error classification was based on the one presented by Vilar et al. (2006) and included four main categories: (1) missing words, (2) word order, (3) incorrect words and (4) punctuation error. The errors were ranked according to their cognitive difficulty. The results showed that applying controlled language rules leads to a lower rate of cognitively difficult errors and increases the number of easy errors, while the original texts preserve a high number of difficult errors while having a lower number of easy errors.

Finally, Kirchhoff et al. (2012) explored user preferences regarding different types of machine translation errors, without any consideration of the post-editing process, however. Word-order errors were the least preferred type, followed by word-sense errors. Morphology errors were ranked third, and function-word errors were the most preferable compared to the other error types under consideration.

One of the major differences of the approach presented in this article is the way it deals with the problem of identification and classification of MT errors. In some of the studies presented above, it is the authors themselves who perform the classification and annotation of the errors (Daems et al. 2015), other studies use the ATA grading classification (Lacruz et al. 2014), discover the errors by analysing the qualitative data from the experiments (Koponen 2012) or base their classification on the editing operations (Popović et al. 2014). Our approach pays special attention to the difficulty and complexity of the MT error annotation task. In many cases annotators disagree on what an error is, and which type it belongs to (Lommel 2013, Burchardt et al. 2013). That is why this study is

based on a robust classification of errors that was elaborated in various stages. The data used for the experiments was annotated for errors by several language professionals, who followed the same annotation scheme. Some of the error types in our classification overlap with the ones considered in previous works, but this classification definitely represents a more complete range of errors. Another significant difference of our methodology is that it allows all error types to be considered separately as well as the PE difficulty indicators for each type. In other words, because each sentence used in the experiment contains only one error type, the sentences used can be grouped according to the error type they contain. Observations that are then made relate to each error group. This could be achieved because all sentences contained only one error.

Finally, the studies overviewed in this section included interesting observations on the relation among different indicators of PE difficulty. Thus, technical effort is more related to cognitive effort than to PE time, and PE time is strongly dependent on sentence length (Popović et al. 2014). Current research further investigates how sentence length influences temporal and technical PE effort and how the two PE difficulty indicators correlate.

3. Methodology

In our experiments we used a corpus of English sentences translated into German with different MT engines and subsequently annotated for translation errors. This point of the methodology is similar to the studies reported by Koponen et al. (2012), Lacruz et al. (2014) and Daems et al. (2015), who also used a corpus of annotated MT errors. Our corpus is described in greater detail in Section 3.1. It contained sentences that contained only one error, which allowed us to separate the effects of each error type on the post-editing process. As it has already been pointed out in the previous section, one of the main differences that sets apart our experiment was that the location of the errors was indicated using braces to ensure that the students post-edited only the same strings that were annotated as erroneous. In other words, we wanted the students to exclusively correct the annotated errors.

Identification of the MT errors is an important step in the post-editing process. For instance, Valotkaite and Asadullah (2012) conducted a post-editing experiment which proves that by previously highlighting errors, post-editors become more efficient, i.e. they miss fewer errors.

Nevertheless, we had to discard the error identification step of the post-editing process in our experiment in order to limit the editing area of the segment to be corrected. The post-editors did not have to spend time and effort on searching for the error, and could concentrate exclusively on correcting the errors indicated for them. This way, our experiment is more controlled, but, on the other hand, it is more distant from a regular post-editing setting that professional translators work in, which naturally includes error identification. Despite the limitations, we considered these measures necessary for comparing different error types in terms of post-editing difficulty. It was only possible to draw conclusions on specific error types and compare them, because we had previously ensured that the participants interpreted the error and corrected only the erroneous part of the sentence.

Another difference between our experiment and common professional post-editing practices was that segments, or sentences in the corpus, were not related to each other, which made it more difficult for the students. In fact, in the informal feedback we received from them, students commented that some segments were hard to translate without knowing the context. This limitation is due to the corpus, as well as the condition of having only one error per segment, which is impossible in an entire text.

3.1. Data

The data used for carrying out the experiments was selected from the MQM error annotation corpora (Burchardt et al. 2013). The MQM corpora contain English source sentences and their translation into German produced by statistical, rule-based and hybrid engines. The corpora were designed to contain sentences that exhibited only few errors. One part of the corpus originated from previously existing publicly available corpora of machine-translated texts drawn either from the technical domain or the news (Avramidis et al. 2012, Specia 2011). The second part was taken from the TSNLP Grammar Test Suite[4] for English. All the data is publicly available online in XML format.[5]

[4] http://www.delph-in.net/tsnlp/ftp/tsdb/ (accessed 27 January 2016).
[5] http://www.qt21.eu/deliverables/annotations/index.html, http://www.qt21.eu/deliverables/test- suite/ (accessed 27 January 2016).

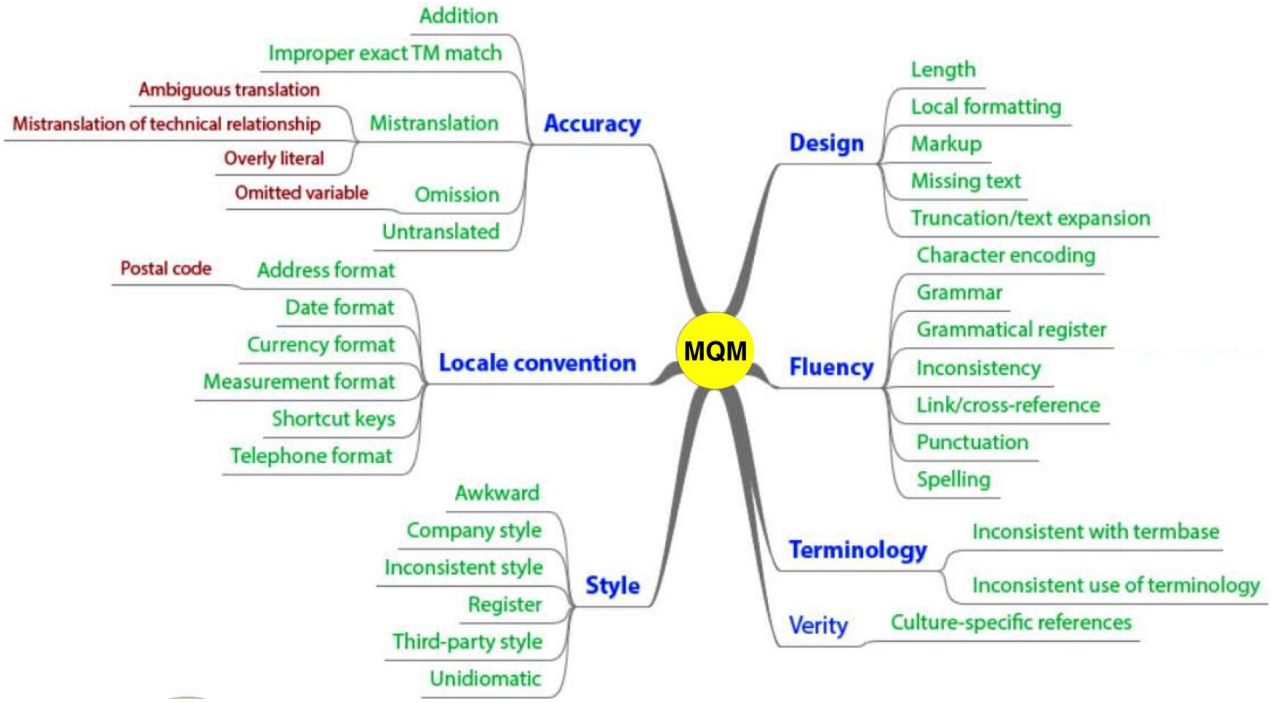
The translations in the corpora were annotated for errors by translation professionals. Each automatically translated sentence was annotated by up to five translators. The annotation was performed according to the Multidimensional Quality Metric (MQM), developed within the QTLaunchPad project.[6] The metric was designed to provide a method for translation error annotation for various purposes and with various degrees of granularity (Lommel 2013) and further contains an error taxonomy as well as guidelines for annotation.

Consider Figure 1 as an example of MQM error taxonomy. It should be mentioned that, since the metric was designed to be adaptable in order to serve different error annotation purposes, this figure illustrates only one of the many possible configurations of the taxonomy. It is formed by several bigger branches, which represent the first-level categories: Accuracy, Locale convention, Style, etc. These branches are further divided into more specific categories. For example, within the Accuracy category we find Addition, Omission, Mistranslation, and others, which all are subcategories of Accuracy errors. Finally, some of these subcategories are again divided into narrower ones, producing the third level of taxonomy.

In order to adapt the MQM taxonomy to the task of creating corpora of MT errors, the creators of the corpora developed annotation guidelines and a special taxonomy configuration, which is provided below.

[6] http://www.qt21.eu/launchpad/ (accessed 27 January 2016).

Figure 1: MQM Taxonomy.



- Accuracy
    - Mistranslation
        - Terminology
        - Unit conversion
        - Overly literal
        - Number
        - False friend
        - Entity
        - Should not have been translated
    - Locale convention
    - Inconsistency
    - Omission
    - Addition
    - Untranslated
- Fluency
    - Style/register

- Unidiomatic
- Spelling
  - Capitalization
- Typography
  - Punctuation
- Grammar
  - Word form
    - Part of speech
    - Agreement
  - Word order
  - Function words
    - Missing function word
    - Erroneous function word
- Unintelligible

The annotation procedure was supported by a set of clear guidelines, which helped annotators resolve problematic cases. For example, 'Locale convention' errors can sometimes be confused with 'Spelling' or 'Punctuation' errors, as they are issues that concern locale-specific spelling and punctuation (such as quotation mark format, which is different depending on the country). The guidelines helped annotators to choose the correct error type.

In our experiments, we wanted to avoid very narrow and rare error types, so that we have enough data for all types to reach valid conclusions. Thus, to reduce the number of types, some categories were joined together. The hierarchical principles of the MQM taxonomy allowed us to join together categories that belong to the same common node, therefore there were no risks of merging error types that are not related to each other. More specifically, we merged some of the subcategories that fall within the same larger category. For example, 'Part of speech' and 'Agreement' errors were joined together under the umbrella category 'Word form'. Similarly, 'Capitalization' was joined with the more general category 'Spelling', as well as 'Typography' (issues related to the mechanical presentation of text in general) and 'Punctuation' (as a specific type of typography issue). The

different types of 'Function words' errors were discarded, so we consider 'Function words' errors as one type.

At this point, it should be clarified that the annotators could choose categories of any level of the taxonomy when classifying errors. Let us consider the example of 'Mistranslation' errors, which occur when "the target content does not accurately represent the source content." (Burchardt et al. 2013: 23) In the taxonomy provided above, it has seven subcategories: 'Terminology', 'Unit conversion' (e.g. pounds into kilograms), 'Overly literal' (an expression which is translated literally), 'Number', 'False friend', 'Entity' (such as a company name, brand name, book title, etc.), and 'Should not have been translated' (cases like song titles, which should stay unchanged in the target language). When encountering a case of mistranslation, the annotators could choose one of the subcategories or, when it was not possible, the general 'Mistranslation' category. The hierarchical structure of the taxonomy allowed a reduction in subjectivity on the part of annotators.

In total, we obtained 23 final error types. Table 1 shows the distribution of error types. 'Mistranslation' was the most frequent one, followed by 'Word form' and 'Function words'. These three types were significantly more frequent than others, while some occurred only once, e.g. 'Character encoding', 'Date/time', 'False friend', among others. In order to compare error types, it was necessary to select sentences that contained only one error. Thus, we selected the sentences where only one error was found by all or the majority of annotators. In sentences where the error types were different among annotators, we chose the more frequent annotation. Thus, 200 sentences were selected. The total number of source words amounted to 1941 (approximately 10 words per sentence on average). The selected sentences were then extracted from the XML files.

Table 1: Distribution of errors.

Error type	Frequency
Mistranslation	41
Word form	34
Function words	30
Overly literal	18
Locale convention	14
Omission	12
Spelling	9
Word order	9
Typography	7
Entity	6
Terminology	3
Addition	3
Grammar	2
Unintelligible	2
Untranslated	2
Character encoding	1
Date/time	1
False friend	1
Inconsistency	1
Number	1
Should not have been translated	1
Style/register	1
Unidiomatic	1

3.2. Experimental Setup

The experiment participants were 19 native German speakers enrolled in the translation program, 12 of whom were on their last year of a bachelor course, and 7 were master students. All of them were familiar with MT concepts and post-editing.

Prior to the actual experiments, the students were given written instructions explaining that they should correct, if possible, only the segment parts within the braces. They were also instructed to limit themselves to corrections considered necessary to achieve a grammatically and semantically correct translation. In addition, they completed a prior short exercise, during which they could practise and prepare for the task.

The CAT (Computer-assisted Translation) tool used for the experiment was Matecat.[7] It was given preference for the editing log feature it provides and its user-friendliness. In addition, it is web-based, which allowed the students to access their translation tasks without any prior software installation, and at the same time it gave us control over the post-editing process. And finally, it is free of charge. The editing log gives an overview of all the corrections made and registers different statistical information about the translation process, including the following:

1) edit time for each segment;
2) post-editing effort (PEE) for each segment, which is a numerical indicator of the amount of editing performed within the segment and is based on fuzzy match algorithms used in CAT tools, as well as the TER metric (Snover et al., 2006). In general, it is calculated as a number of changed words in the sentence divided by the total number of words.
3) total time for the job;
4) average edit time and PEE for the job.

In order to distribute the sentences among the participants we, first of all, randomly selected ten sentences that were to be included in every student's translation job. This was done in order to measure the annotators' agreement. The rest of the selected sentences from the corpora were randomly divided into four sets, each set had 47 or 48 sentences. Each student received a translation job with one of these sets with the ten common sentences added at the end. This way, each sentence except for the ten common sentences was post-edited by four or five students. Each student received a URL with their personal translation job in Matecat, which included a document to translate and a TM (Translation Memory) containing the annotated translations. All translation units in the TM had a 100% match. During the post-editing process, the students were allowed to consult external resources, as they would do during a regular translation assignment. The experiment organisers controlled the students' activity to ensure their full involvement in the process. In order to assure that participants only edit the errors annotated in the corpus, they appeared in braces. In the example below taken from the test exercise, the article *der* is marked with braces, indicating that it is not correct (the correct article would be *das*), thus being an example of a'Function words' error.

(1) Fill the glass almost to the top with ice.

[7] https://www.matecat.com/ (accessed 27 January 2016).

{Der} Glas bis fast zum Rand mit Eis auffüllen.

4. Results and Discussion

In this section we present the results of the experiment, including general statistics followed by a comparison of different measures among different error types.

4.1. General Results

In total, we collected 19 translation sessions. Due to a technical problem in Matecat, some target segments were not retrieved from the translation memory files, so the students could not edit them and, therefore, they were missing during the analysis stage. Furthermore, one segment was eliminated as the students commented during the experiment that it was too short and too difficult to understand without any context. Thus, only 196 source sentences were valid for analysis, with the total edited 861 target segments where post-editing time was not equal to zero (these are the segments that the students started to work on but did not necessarily finish).

We measured inter-annotator agreement for post-editing time and PEE. Inter-annotator agreement shows how similarly the participants behaved in terms of a chosen measure. In our particular case, it indicates the proximity in the length of time the students took to edit the same sentences (inter-annotator agreement for PE time), and whether they made approximately the same number of changes (agreement in PEE). Inter-annotator agreement was calculated using the Intraclass Correlation Coefficient (ICC) (Shrout and Fleiss 1979), which is attributed values between -1 and 1, where -1 suggests strong disagreement and 1 suggests absolute agreement. We observed the ICC in post-editing effort at a value of 0.36 ($ICC_{PEE} = 0.36$) and the ICC in post-editing time at a value of 0.16 ($ICC_{time} = 0.16$). These values are quite low, especially in the case of PEE: since the errors were marked, we expected students to make approximately the same number of corrections, which would imply high agreement in PEE. On the other hand, the fact that ICC_{PEE} is higher than ICC_{time} was quite predictable, for the same reason of marked errors. The assumption was that the error marking would restrict participants to correcting the same parts of text and thus to making approximately the same number of corrections, while the time taken to do it could vary more, since it depends on the participants' individual characteristics, like typing speed and time needed to think

of the right translation. One of the reasons why agreement in PEE was lower than expected was found in the data retrieved from the editing logs. It turned out that the variety of different edited versions of one sentence is quite high even in spite of the error marking. In other words, students tend to choose different final translations for the same error. Thus, about 13% of the sentences were corrected differently by all students, compared to 17% that had an identical final version among all students. Consequently, as the final versions are very different among post-editors, the number of changes made is also different.

It is worth mentioning that similar results were obtained by Koponen (2013), who investigated how differently post-editors performed in terms of PEE and PE time. The data was generated using a controlled language – i.e. prior to the translation, the source segments were modified according to certain rules to make them easier to translate automatically. Less variation was observed in PEE than in PE time. On the other hand, differently from our case, most of the sentences only had one or two PE versions, which means that there was high agreement on what the right translation was, which was probably due to the controlled language. Thus, one can make an assumption that controlled language is more efficient in restricting subjectivity on the part of post-editors rather than error marking.

On average, the time students needed to correct one sentence was 40.88 seconds, and the average PEE value was 0.24. The plotted distributions of the values are shown in Figures 2 and 3, respectively. Note that we previously removed the PE time values greater than 100 to avoid too many outliers: if a student took more than 100 seconds to edit one segment, it is probably due to some distraction and does not reflect the actual editing time. Thus, 78 data points were removed, which corresponded to 9% of the total number of 861 edited target segments with non-zero edit time.

The graphs show how the values of PE time (Figure 2) and PEE (Figure 3) are distributed among the experiment participants. The boxes in the graphs represent 50% of the data. The horizontal line inside each box shows the median, or the mid-point of the data. This means that the numbers of data points below and above the horizontal line are equal. The lines outside the boxes, or the *whiskers*, each represent 25% of the data. In addition, the circles above the upper whiskers are the *outliers*;

these are the data points that were not considered when computing the shape of the box plot. In general, box plots show how the data are distributed: if the boxes are short with short whiskers it means that the data in the sample is very similar, while tall boxes with long whiskers mean that the distribution is spread out and that there are significant differences between the data points.

Figure 2: Distribution of time.

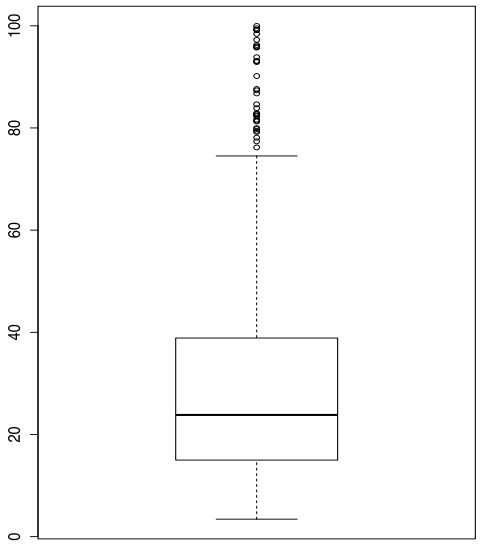
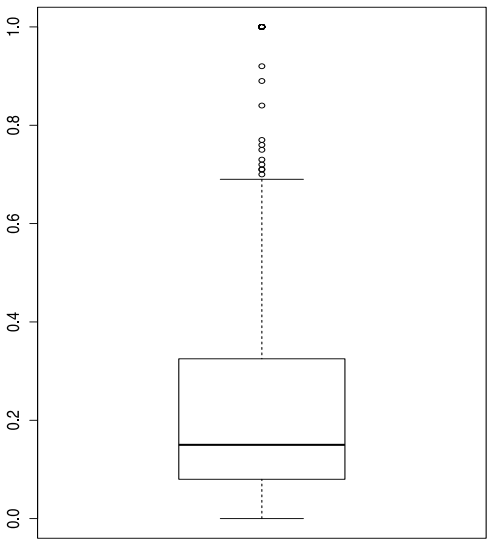


Figure 3: Distribution of PEE.

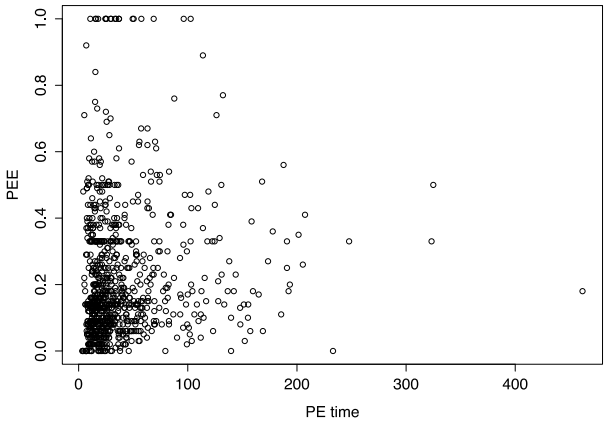


As we can see, the general shape of the box plots, as well as their positions are quite similar. However, the box plot is a little flatter in case of PEE, while the box plot for PE time is taller: the values are more spread, there are many more outliers, and the whiskers are longer. This means that there is more variation in PE time than in the PEE, in other words, the students took a different amount of time to edit the sentences, but the PEE values are somewhat more similar among different students. This is also in line with the obtained inter-annotator agreement scores.

In order to investigate whether there is a dependency between the time students took and the number of changes they made, we calculated the correlation between PE time and PEE using Pearson's correlation coefficient, which is a statistical measure that takes values from -1 to 1, where -1 indicates strong negative correlation (the two variables are inversely related so that when one decreases the other one increases), 1 indicates strong positive correlation (the two variables are

related so that when one increases the other one does too), and 0 indicates no correlation. The edit time and PEE variables showed very weak correlation with the Pearson's correlation coefficient equal to r(859)=0.10 (with the p-value=0.0026),[8] which is illustrated in Figure 4. In this scatter plot, the dots represent the edited sentences, and there are multiple dots for the same source sentence corresponding to the edits by various participants. The axes represent the values of PE time and PEE that each edited sentence rendered. In the case of strong positive correlation, the dots would form a diagonal line, as with increasing PE time the PEE would also increase. In our case, it did not happen and the dots are very spread out around the surface of the plot. This is an interesting result from the point of view of post-editing process: it appears that, even when the error requires a large number of editing operations, it does not necessarily mean that it requires much time. And, vice versa, when there are only few corrections to be made, the editor might still spend a long time finding the right translation. Thus, there is no direct dependency between the temporal and the technical PE effort.

Figure 4: Correlation between time and PEE.



In addition, we discovered that there is only a weak correlation of r(195)=0.26 (with p-value=0.0002) between the post-edited segment length (in number of tokens) and post-editing time. It is interesting to compare these results with those presented by Koponen (2012), who found that the human perception of post-editing effort correlates strongly with the sentence length. Our results

[8] The Pearson's correlation coefficient was calculated for all data points, where a data point corresponds to one sentence edited by one participant.

show that sentence length affects post-editing time only slightly. This can act as proof that there is a difference between perceived PE effort and the actual time it takes to post-edit. Even so, when comparing the two findings, one has to take into account the influence of the error marking, which also affects the editing time. There is a stronger negative correlation between sentence length and PEE: r(195)=-0.37 (with p-value=0.00001). This is due to the way PEE is calculated. In fact, it is similar to the proportion of editing operations on tokens to the total number of tokens in the sentence, so the PEE score decreases when sentence length increases.

4.2. Comparison of Error Types

In this section we consider differences in post-editing time and effort between the error types. Table 2 shows the average values of PE time and PE effort for all the error types, as well as the error frequency. Especially high values of average time and PEE can be observed in the 'Overly Literal' category, with an average time of 57.6 seconds and average PEE of 0.43. In other categories, only one of the two indicators can stand out, such as 'Entity' with a high average time of 55.6 seconds and a relatively regular average PEE of 0.3 (consider similar picture in 'Should not have been translated', where average time is high, but average PEE is neither specifically high or low). In one case, we observed a particularly high PE time combined with a very low PEE, which is the 'Style/register' error. Finally, 'Inconsistency' and 'Number' errors had a low average PEE value and a regular PE time.

Table 2: Average time and post-editing effort in different error types.

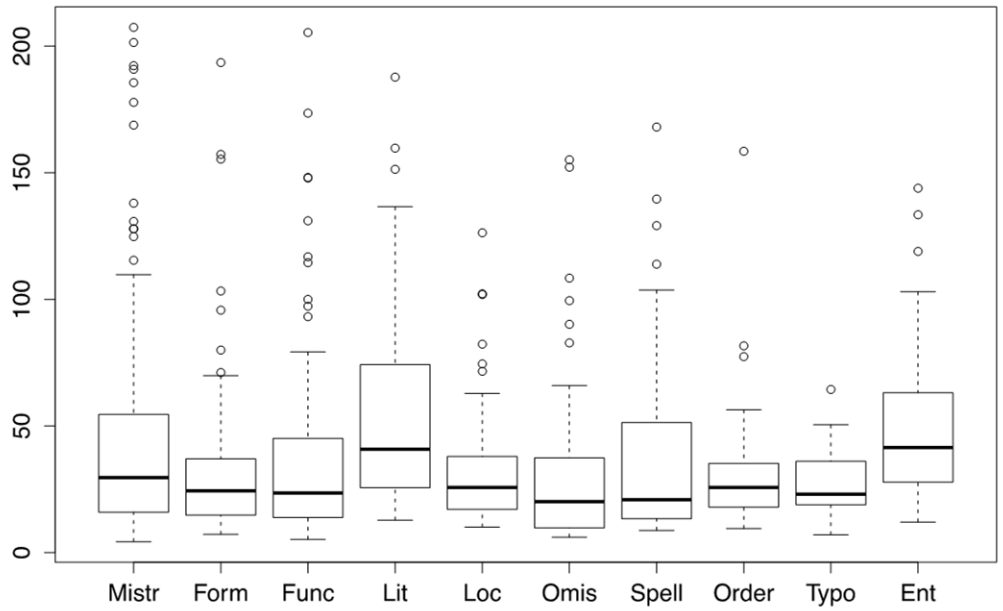
Error type	Frequency	Avg. time	Avg. PEE
Mistranslation	41	46.81	0.26
Word form	34	32.78	0.18
Function words	30	37.19	0.18
Overly literal	18	57.58	0.43
Locale convention	14	35.25	0.22
Omission	12	33.02	0.17
Spelling	9	39.92	0.16
Word order	9	36.99	0.16
Typography	7	27.08	0.36
Entity	6	55.63	0.3
Terminology	3	46.8	0.23
Addition	3	28.68	0.15
Grammar	2	49.3	0.21

Unintelligible	2	30.2	0.24
Untranslated	2	34.8	0.17
Character encoding	1	11.7	0.19
Date/time	1	29.3	1
False friend	1	39.5	0.19
Inconsistency	1	25.3	0.12
Number	1	26.4	0.12
Should not be translated	1	93.3	0.3
Style/register	1	58.9	0.09
Unidiomatic	1	52.5	0.34

We looked at the distribution of the time and PEE values in different errors. Figure 5 shows the time distribution for the ten most frequent error types. We can see that the box plot for the 'Overly literal' type ('Lit') is the tallest and the highest at the same time. This means that generally this error type takes more time, and that the time varies significantly between the post-editors. 'Word form' ('Form') and 'Function words' ('Func'), on the contrary, have comparatively flat boxes and their upper whiskers are shorter (with some outliers), so there is less variation in post-editors' speed for this type of error. The same is true for 'Locale convention' ('Loc'), 'Omission' ('Omis'), 'Word order' ('Order') and 'Typography' ('Typo'). 'Spelling' errors had a generally low median but a tall box, indicating high variation. The box representing 'Mistranslation' error ('Mistr') is among the highest, thus being associated with long PE time. It is also one of the tallest boxes, so time varied significantly for this type of error.

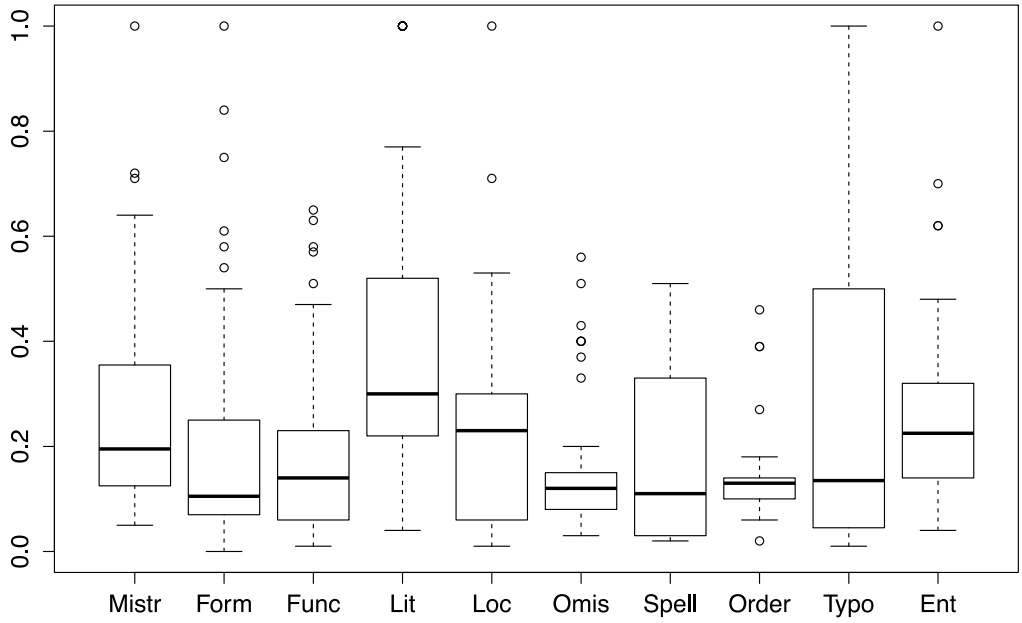
To summarize, it is easily seen from Figure 5 that 'Mistranslation', 'Overly literal', 'Spelling' and 'Entity' errors show more variation in edit time and are on average slower to edit. At the same time, 'Word form', 'Function words', 'Locale convention', 'Omission', 'Word order', and 'Typography' errors generally take less time to correct, and the variation between post-editors is lower. This confirms the results presented by Popović et al. (2014), who found that lexical errors take longer to correct. The fact that 'Word order' errors took less time was quite surprising, as both Koponen (2012) and Popović et al. (2014) report these errors to be cognitively difficult based on translators' manual scores. This can indicate that these errors are perceived as difficult, but in practice do not require much time.

Figure 5: Distribution of PE time by error type.



Similarly, we plotted the distribution of the PEE values among error types (Figure 6). This time, the error types are much more different from each other and the whole picture is also quite different from Figure 5. The 'Overly literal' box plot is again the highest, with the median of about 0.3, however this time it is not the tallest; in other words, correcting this type of error required many editing operations for most students. 'Mistranslation' errors generally have a lower PEE. Curiously, PEE value in 'Typography' errors was very spread out, despite the fact that typography corrections do not require many editing operations. Qualitative analysis of the editing log data revealed that this is mostly due to different strategies post-editors employ when correcting the same errors. For instance, when a quotation mark is missing, some editors simply insert one, while others also delete the preceding word and rewrite it together with the missing quotation mark.

Figure 6: Distribution of post-editing effort by error type.



Another two types of error where post-editors had less agreement in PEE were 'Locale convention' and 'Spelling'. However, their medians were not high (generally not higher than 0.5), which was quite expected. Thus, we can conclude that these errors do not require many operations to correct, but editors tend to have different strategies while post-editing them, such as replacing just one symbol, or the whole word or phrase. 'Word form' and 'Function words' box plots are very similar: they are shorter and lower than the errors mentioned above, so they required less editing and exhibited less variety in PEE among students. 'Omission' and 'Word order' have a similar median under 0.2, but the boxes are much shorter, so for these errors the students showed similar PEE scores with less variation.

If we compare the two figures we can see that 'Word form' errors generally require a short time to correct (the median is low while the box and the upper whisker is short), while the PEE scores are more spread. The same tendency is observed in other errors as well, in particular 'Locale convention' and 'Typography'. This might be due to the differences in the way translators correct the errors, as was observed earlier, but can also depend on sentence length. In general, we observed

more variation in PEE scores among different error types than in PE time both in terms of the median and in terms of variation between translators within each of the types.

One of the aims of this study was to find out whether the students' corrections of the errors resulted in different translation versions, i.e. if there was an agreement between the participants on what the correct translation was. In order to do that we looked at the number of different solutions students proposed for the same sentence. As we commented in Section 4.1, it turned out that, even though the errors were highlighted, the final versions were identical only in 17% of sentences. For each sentence we calculated the number of post-edited translations divided by the number of different versions. This number represents the average number of identical final versions for one sentence which will be called *identical version score*. For instance, if four students post-edited one sentence, and three of the edited versions were identical while one was different, the number would be equal to 4/2=2, where 4 is the number of post-edited target sentences and 2 is the number of different versions. Thus, the higher the score, the more similar the post-edited versions for this sentence were. Then, we calculated the mean of these values for each error type (Table 3).

Table 3: Average identical version score for each error type.

Character encoding	4	Unintelligible	2.09
Spelling	3.21	False friend	2
Function words	2.95	Number	2
Addition	2.83	Mistranslation	1.79
Typography	2.72	Inconsistency	1.67
Omission	2.69	Style/register	1.67
Word order	2.53	Unidiomatic	1.66
Locale convention	2.35	Grammar	1.61
Word form	2.3	Terminology	1.5
Untranslated	2.25	Overly literal	1.31
Entity	2.17	Should not be translated	1.25

We can see that the errors with the highest scores are 'Character encoding', 'Spelling', 'Function words', 'Addition', 'Typography', and 'Omission'. They all have a score higher than 2.5. This means that the students' translations were very similar and that there was high agreement on what the correct translation was for these errors. 'Terminology', 'Overly literal' and 'Should not have been translated' received the lowest scores (1.5 and less), suggesting that there is more variety of

possible translations for these errors. To summarize, most of the errors that concern punctuation, spelling and grammar showed higher scores. At the same time the errors that concern lexical choice, idiomaticity and style received lower scores, which demonstrates that these issues can have a broader range of possible solutions.

5. Conclusions and Future Work

In this article we presented the results of a post-editing experiment that aimed at investigating the post editing process from the point of view of PE time and technical PE effort.[9] We compared how these indicators of PE difficulty change with different types of MT errors. The sentences to post-edit contained one error each. The errors were highlighted for post-editors, who were instructed only to correct the highlighted part of the sentence if possible. The sentences were taken from a corpus that was previously annotated for MT errors.

Due to a technical problem some of the data items were missing, which made data scarcer than expected. In addition, participants mentioned that in some cases it was not clear how to proceed with the task due to the lack of context. And indeed, we could see that some segments were not edited at all, which we interpreted as difficulties in understanding sentence meaning. Nevertheless, the data analysis revealed some interesting findings.

Despite the instructions provided, there was only low agreement between annotators. The agreement was higher for PEE than for PE time, which can be partially explained by the error marking. There was also a significant variety in correct solutions provided by participants. All these results show that the PE process is individual and different for each translator and that the choice of the final versions is also subjective, as in most cases there is more than one possible solution for most MT errors.

The observed correlation between PE time and PEE was only weak. This means that different indicators of post-editing effort are not necessarily related: some errors require more time to find the right solution but do not necessarily involve many editing operations. A practical implication of

[9] The latter represented by the Matecat PEE score.

this finding is that, when assessing the difficulty of a given MT output for post-editing, one has to consider these two indicators separately, or decide which is more important in the specific setting – i.e. whether time or technical effort should be given priority.

We observed more variety in PE time between translators in errors that require lexical choice ('Mistranslation') and errors related to multiword expressions, such as 'Overly literal' and 'Entity'. These errors also exhibited longer average time to correct. This was not always true in the case of PEE, where, apart from 'Overly literal' and 'Mistranslation', high variation and average value were also observed in 'Typography'. We assumed that this is due to differences in the operations translators perform while correcting the same error. The only category of error that had high indicators both in temporal and in technical PEE was 'Overly literal'. Thus, multi-word expressions being a well-known problem for many types of MT systems, are also one of the main difficulties for human translators. Even though a number of researchers reported that word order and structural errors are perceived as difficult for post-editing, our experiment did not yield significantly high PE time or PEE values for this type of error. This could suggest that some error types exhibit differences between the perceived cognitive effort and the temporal and technical effort actually exerted during the PE process. We also compared the diversity of final versions provided by translators for the same error. The most versions were found for 'Overly literal', 'Entity' and terminological errors, and, surprisingly, for the general 'Grammar' category.

One can envisage a number of ways in which these findings can be used with the aim of improving the post-editing workflow. Being a well-known challenge for automatic translation, multi-word expressions and named entities should be given special attention when developing an MT engine for PE purposes, as these are the types of errors that were also shown to be specifically difficult for post-editing. When evaluating MT output for usefulness in post-editing, these errors, along with mistranslations, should be given more weight than, for instance, word order, word form, and function word errors, as those proved to take less time and effort. Finally, the differences between different error types have to be taken into account when training post-editors. At the same time, such training should also consider that there are significant differences among post-editors regarding their speed, the editing strategies they use, as well as the agreement on the final correct translation.

Considering the results presented in this article, our plans for future work include conducting a similar experiment with an MQM-annotated corpus available for the English-Spanish language pair and comparing the findings in order to investigate possible differences in the PE process between German and Spanish as target languages. In addition, it would be interesting to see whether the PEE indicators for the same error types are different if the errors were performed by different types of MT systems: statistical, hybrid and rule-based. Finally, annotating more data would help obtain a more representative corpus with a more balanced distribution of errors, which would allow us to obtain more reliable results.

Acknowledgements

Anna Zaretskaya is supported by the People Program (Marie Curie Actions) of the European Union's Framework Program (FP7/2007-2013) under REA grant agreement no 317471. The research reported in this article has been partially carried out in the framework of the research group Lexytrad. The authors would like to thank Prof. Josef van Genabith and Anne Schumann for their useful comments and suggestions.

Anna Zaretskaya
annazar@uma.es
Mihaela Vela
m.vela@mx.uni-saarland.de
Gloria Corpas Pastor
gcorpas@uma.es
Miriam Seghiri
seghiri@uma.es

References

Avramidis, Eleftherios, Aljoscha Burchardt, Christian Federmann, Maja Popović, Cindy Tscherwinka and David Torres Vilar (2012) 'Involving Language Professionals in the Evaluation of Machine Translation', *Proceedings of the 8th ELRA Conference on Language Resources and Evaluation*, Istanbul, Turkey, May 2012, 1127-1130.

Burchardt, Aljoscha, Arle Lommel and Maja Popović (2013) *Deliverable 1.2.1. TQ Error Corpus,* 31 July 2013.

Daems, Joke, Sonia Vandepitte, Robert Hartsuiker and Lieve Macken (2015) 'The Impact of Machine Translation Error Types on Post-Editing Effort Indicators', *Proceedings of the 4th Workshop on Post-Editing Technology and Practice (WPTP4)*, Miami (Florida), October 30-November 3 2015, 31-45.

Gupta, Rohit, Constatin Orasan, Marcos Zampieri, Mihaela Vela and Josef van Genabith (2015) 'Can Translation Memories afford not to use Paraphrasing', *Proceedings of the European Association of Machine Translation (EAMT-2015),* Antalya, Turkey, 2015.

Kirchhoff, Katrin, Daniel Capurro and Anne Turner (2012) 'Evaluating user preferences in machine translation using conjoint analysis', *Proceedings of the 16th Annual Conference of the European Association for Machine Translation (EAMT 12)*, Trento, Italy, May 2012, 119-126.

Koby, Geoffrey S. (2001) 'Editor's Introduction', in Hans P. Krings Repairing Texts: Empirical Investigations of Machine Translation Post-editing Processes, 1-23.

Koponen, Maarit (2012) 'Comparing human perceptions of post-editing effort with post-editing operations', *Proceedings of the 7th Workshop on Statistical Machine Translation*, Montréal, Canada, Association for Computational Linguistics, June 2012, 181-190.

Koponen, Maarit, Wilker Aziz, Luciana Ramos and Lucia Specia (2012) 'Post-Editing Time as a Measure of Cognitive Effort', *Proceedings of the AMTA 2012 Workshop on Post-Editing Technology and Practice (WPTP 2012),* 28 October 2012, San Diego, USA, edited by Sharon O'Brien, Michel Simard and Lucia Specia, 11-20.

Koponen, Maarit (2013) 'This translation is not too bad: An analysis of post-editor choices in a machine translation post-editing task', *Proceedings to the MT Summit XIV Workshop on Post-editing Technology and Practice*, Nice, France, September 2013, 1-9.

Krings, Hans P. (2001) Repairing Texts: Empirical Investigations of Machine Translation Post-editing Processes, Kent State University Press.

Lacruz, Isabel, Michael Denkowski and Alon Lavie (2014) 'Cognitive Demand and Cognitive Effort in Post-Editing', *Proceedings of the Third Workshop on Post-editing Technology and Practice*, Vancouver (Canada), 26 October 2014, 73-84.

Lacruz, Isabel and Gregory M. Shreve (2014) 'Pauses and cognitive effort in post-editing', *Post-editing of Machine Translation: Processes and Applications*, Newcastle upon Tyne: Cambridge Scholars Publishing, 246-273.

Laübli, Samuel, Mark Fishel, Gary Massey, Maureen Ehrensberger-Dow and Martin Volk (2013) 'Assessing Post-Editing Efficiency in a Realistic Translation Environment', *Proceedings of MT Summit XIV Workshop on Post-Editing Technology and Practice*, 83-91.

Lommel, Arle (2013) Deliverable D 1.1.2. Multidimensional Quality Metrics. 28 June 2013.

Moorkens Joss, Sharon O'Brien, Igor da Silva, Norma Fonseca and Fabio Alves (2015) 'Correlations of perceived post-editing effort with measurements of actual effort', *Proceedings of the Translating and the Computer Conference 27*, London, UK, 26-27.

Plitt, Mirko and François Masselot (2010) 'A Productivity Test of Statistical Machine Translation Post-Editing in a Typical Localisation Context', *The Prague Bulletin of Mathematical Linguistics*, January 2010, 7-16.

Popović, Maja, Arle Richard Lommel, Aljoscha Burchardt, Eleftherios Avramidis and Hans Uszkoreit (2014) 'Relations between different types of post-editing operations, cognitive effort and temporal effort', *The Seventeenth Annual Conference of the European Association for Machine Translation (EAMT 14)*, Dubrovnik, Croatia, 191-198.

Scarton, Carolina, Marcos Zampieri, Mihaela Vela, Josef van Genabith and Lucia Specia (2015) 'Searching for Context: a Study on Document-Level Labels for Translation Quality Estimation', *Proceedings of the 18th Annual Conference of the European Association for Machine Translation (EAMT 2015)*, Antalya, Turkey, 121-128.

Shrout, Patrick E. and Joseph L. Fleiss (1979) 'Intraclass Correlations: Uses in Assessing Rater Reliability', *Psychological Bulletin*, 86(2), 420-428.

Snover, Matthew, Bonnie Dorr, Richard Schwartz, Linnea Micciulla and John Makhoul (2006) 'A Study of Translation Edit Rate with Targeted Human Annotation', *Proceedings of Association*

for Machine Translation in the Americas, Cambridge, Massachusetts, USA, August 8-12, 2006, 223-231.

Specia, Lucia (2011) 'Exploiting Objective Annotations for Measuring Translation Post-editing Effort', *Proceedings of the 15th Annual Conference of the European Association for Machine Translation (EAMT 11),* Leuven (Belgium), May 2011. 73-80.

Stevens, Andrea and Valeria Filipello (2014) 'MT in Practice of Language Service Provider', *MT@Work 2015. Machine Translation in Translation Practice. Annual User Conference*. Brussels, Belgium, 4 December 2014.

Temnikova, Irina (2010) 'Cognitive evaluation approach for a controlled language post-editing experiment', *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10),* Valletta, Malta, May 2010, 3485-3490.

Valotkaite, Justina and Munshi Asadullah (2012) 'Error Detection for Post-editing Rule-based Machine Translation', *Proceedings of the AMTA 2012 Workshop on Post-editing Technology and Practice (WPTP 2012),* San Diego, California, October 2012.

Vilar, David, Jia Xu, Luis Fernando D'Haro and Hermann Ney (2006) 'Error Analysis of Statistical Machine Translation Output', *Proceedings of LREC 2006*, 697-702.

Zampieri, Marcos and Mihaela Vela (2014) 'Quantifying the Influence of MT Output in the Translators' Performance: A Case Study in Technical Translation', *Proceedings of the EACL Workshop on Humans and Computer-assisted Translation (HaCat)*, May 2014, 93-98.

Zhechev, Ventsislav (2014) 'Analyzing the post-editing of machine translation at Autodesk', *Post-editing of Machine Translation: Processes and Applications*, Newcastle upon Tyne: Cambridge Scholars Publishing, 2-24.