The Impact of Machine Translation Error Types on Post-Editing Effort Indicators

Joke Daems joke.daems@ugent.be
Department Translation, Interpreting and Communication, Ghent University, Ghent, 9000, Belgium
Sonia Vandepitte sonia.vandepitte@ugent.be
Department Translation, Interpreting and Communication, Ghent University, Ghent, 9000, Belgium
Robert Hartsuiker robert.hartsuiker@ugent.be
Department of Experimental Psychology, Ghent University, Ghent, 9000, Belgium
Lieve Macken lieve.macken@ugent.be
Department Translation, Interpreting and Communication, Ghent University, Ghent, 9000, Belgium

Abstract

In this paper, we report on a post-editing study for general text types from English into Dutch conducted with master's students of translation. We used a fine-grained machine translation (MT) quality assessment method with error weights that correspond to severity levels and are related to cognitive load. Linear mixed effects models are applied to analyze the impact of MT quality on potential post-editing effort indicators. The impact of MT quality is evaluated on three different levels, each with an increasing granularity. We find that MT quality is a significant predictor of all different types of post-editing effort indicators and that different types of MT errors predict different post-editing effort indicators.

1. Introduction

In recent years, machine translation (MT) and its subsequent post-editing have become more widely accepted in the translation industry. Especially when it comes to technical texts, machine translation has proven its worth, with companies like Autodesk reporting on productivity increases in comparison with human translation ranging from 20 to 131%, depending on the language combination and translator (Plitt & Masselot, 2010). The main goal of post-editing research is no longer finding out whether or not post-editing can be used, but rather finding out when it cannot be used, and how machine translation systems can be improved to better suit post-editors' needs.

While post-editing is generally assumed to be faster than human translation, speed is not the only factor that should be taken into account when assessing the post-editing process. More recent studies have looked at ways of determining post-editing effort. This knowledge can be used, on the one hand, to improve the quality of MT systems, and, on the other hand, to reduce post-editors' frustration by only presenting them with a segment containing MT output when the effort required to post-edit that segment is not too high.

Krings (2001) mentioned three levels of post-editing effort: temporal effort, or the time needed to post-edit a given text, cognitive effort, or the activation of cognitive processes during post-editing, and technical effort, or the technical operations such as insertions and deletions that are performed during post-editing. According to Krings (2001), post-editing research should concentrate on causes and manifestations of post-editing effort with a focus on cognitive effort: "The type and extent of cognitive processes triggered by the post-editing task must be defined qualitatively and quantitatively, and correlated to the corresponding deficiencies in machine translations as triggering factors" (p. 182).

In this paper, we will first discuss some previous work on the effort involved in post-editing and the problems that arise when trying to measure cognitive effort in isolation. We then present the results of our study, examining the impact of different types of machine translation errors on post-editing effort indicators with student translators post-editing from English into Dutch.

2. Related research

The ultimate goal of post-editing process research is predicting how much effort a post-editor will need to correct a segment before presenting the post-editor with that segment. Depending on the expected effort, a translator can then be given MT output to post-edit whenever the effort to post-edit would be lower than the effort needed when translating that segment from scratch. Two aspects need to be researched in order to reach that ultimate goal: firstly, we need to establish which types of effort we take into account and how we can objectively measure them, and secondly, we need to find ways of predicting effort on the basis of elements contained in either the source text or the MT output. Both aspects will be discussed in the following paragraphs.

A number of potential post-editing effort indicators have been introduced in previous research. The distinction between temporal, cognitive and technical effort as proposed by Krings (2001), however, does not seem to be a clear distinction.

While temporal effort seems the easiest to measure, as it is simply the time needed to translate a word, segment or text, Koponen et al. (2012) found evidence that post-editing time can also be an indication of cognitive effort. They use a cognitively motivated MT error classification created by Temnikova (2010), but finish their paper with a few remarks on the classification and a suggestion for future work: "A revised set of error categories with more detailed error types (...) is also an interesting direction to help understand the cognitive load in post-editing" (p. 20).

Koponen et al. (2012) also looked at a technical effort indicator - keystrokes - and its relationship to cognitive load. However, they found that keystrokes were influenced more by individual differences between participants than by cognitive load. We therefore decided not to include keystrokes as such in our analysis. Related to keystrokes are production units, or sequences of coherent typing activity. Although producing translation output in itself is clearly a technical activity, Lacruz et al. (2012) intuitively felt that an increase in the number of complete editing events (which correspond to the notion of production units) would lead to an increase in cognitive demand as well, making it a cognitive effort indicator in addition to a technical effort indicator. The question remains whether editing events really correspond to cognitive effort. For example, many spelling errors or adjective-noun agreement errors will require quite a few (local) editing events, but are not really difficult to solve.

Lacruz et al. (2012) further introduce the average pause ratio (the average time per pause in the segment divided by the average time per word in the segment) as an answer to O'Brien's pause ratio (2006) - the total time in pauses divided by the total editing time. O'Brien (2006) did not find conclusive evidence for a relationship between pauses and cognitive activity. Lacruz et al. (2012) argue that pause ratio is not sensitive enough as a measure for cognitive activity, as it does not take average pause length into account. We include both pause measures in our study, to establish whether or not they can both be used, and whether or not they are indicators for different causes of effort. Lacruz et al. (2012) found a relationship between average pause ratio and the number of production units. As production units are delimited by pauses, and the average pause ratio is influenced by the number of pauses, perhaps this finding is related more to intrinsic correlation than to actual impact of cognitive load on pause behavior, although the relationship is most likely more complex. We will look at production units and average pause ratio in isolation to better understand the differences and similarities between both variables.

Some of the few effort indicators that seem to be exclusively related to cognitive post-editing effort, are the average fixation duration and the number of fixations. Building on the eye-mind hypothesis from Just and Carpenter (1980), a person is cognitively processing what they are looking at. Longer fixations should thus be an indication of more cognitive processing. This assumption was confirmed by Jakobsen and Jensen (2008), who found longer average fixation durations and a higher number of fixations as the complexity of the task increased from reading to translation. Doherty and O'Brien (2009), however, found a higher number of fixations for bad MT output than for good MT output, but they did not find a significant difference between the average fixation durations for both types. We will include both average fixation duration and number of fixations as potential cognitive post-editing effort indicators.

From the abovementioned research, it becomes clear that the distinction between the different types of effort indicators is not always easily made. Correlations are identified between different indicators without really knowing whether or not they measure different things. To avoid this circular thinking, we need to find a way of studying the post-editing effort indicators in isolation, by linking them to source text and MT output characteristics rather than to other post-editing effort indicators. O'Brien (2004) has taken a step in this direction by looking at negative translatability indicators (NTIs) in the source texts, or elements that can reasonably be considered to be problematic for MT systems, for example, long noun phrases or gerunds. Although some NTIs indeed seem to have an impact on post-editing effort, there are some NTIs that have no effect, and O'Brien (2004) further found post-editing activity in segments that did not contain NTIs. From these findings, we can derive that NTIs do not conclusively predict post-editing effort, and perhaps another focus is needed.

In this paper, we take a look at a fine-grained MT quality assessment and whether or not the average MT error weight of a segment has an impact on the post-editing process. In line with previous research, we take a look at different types of post-editing effort indicators.

3. Methodology

3.1. Participants

Participants were ten master's students of translation. All of them were native speakers of Dutch. They had no previous experience post-editing and had passed their final English General Translation exam. They received two gift vouchers of 50 euros each. As we are working with students, it is of course hard to say whether our results will generalize to the professional translation process. However, we have repeated the experiment with professional translators (but the process data has not yet been analyzed), and we found no significant differences in proficiency or attitude towards post-editing between the two groups, so perhaps they are more comparable than often thought.

3.2. Text selection

The present study is a part of a larger study aimed at comparing the differences between the human translation process and the post-editing process for students and professional translators for general text types. In the present study, the focus will be on the post-editing process of the students only, but the texts have been selected with the larger study in mind.

Originally, fifteen different English newspaper articles were selected from newsela.com, a website providing newspaper articles at different levels of complexity, as indicated by a Lexile score. We selected articles with the same level of complexity, i.e., Lexile scores between 1160L and 1190L[1], to try to control for textual differences in our studies. Each article was reduced to its first 150-160 words, and then analyzed for additional readability measures and potential translation problems. Texts with on average less than fifteen or more than twenty words per sentence were discarded, as well as texts that contained too many or too few complex compounds, idiomatic expressions, infrequent words or polysemous words. Sentence length ranged from seven to thirty-five words, with an average of eighteen point three, and a median of eighteen words per sentence. The texts were then translated into Dutch by the statistical machine translation system Google Translate. We annotated the MT output for quality, as will be discussed in section 3.4. From the original fifteen texts, the eight texts that were most comparable in difficulty - based on the potential translation problems and MT output quality - were retained. Texts have different subjects and don't require specialist knowledge to be translated.

3.3. Experimental setup

Two sessions were planned for each participant. During the first session, students had to first fill out a survey, and take a LexTALE test (Lemhöfer & Broersma, 2012) to be able to take their English proficiency into account. This was followed by a copytask and a warmup task combining both post-editing and human translation, so the students could get used to the tools and different types of tasks. After the warmup, students post-edited two texts and translated two texts from scratch. During the second session, students again started with a warmup task, followed by two human translation tasks and two post-editing tasks. The order of texts and tasks was balanced in a Latin square design across all participants, to reduce task order effects. The second session ended with a retrospective part, during which students could highlight elements in the text that they found most difficult to translate or post-edit, and another survey to measure how students experienced the experiment and the different tasks.

To be able to look at different aspects of post-editing effort, we used a combination of keystroke logging tools and eye tracking. The process itself was registered by the CASMACAT translator's workbench (Alabau et al., 2013), which looks like an actual translation environment to improve ecological validity, yet contains keystroke logging and mouse tracking software for researchers to better be able to observe the translation and post-editing process in detail. The texts were presented to the students one by one, and each text was subdivided in translation segments, corresponding to sentences in the source text. The number of segments in each text ranges from seven to ten. A plugin connects Casmacat to the EyeLink 1000 eyetracker that was used to register the students' eye-movements while translating and post-editing. In addition to these tools, an extra keystroke logging tool, Inputlog (Leijten & Van Waes, 2013) was running in the background. While the CASMACAT software is capable of performing a detailed logging within the CASMACAT interface, it cannot log external applications. Inputlog registers when and which applications other than CASMACAT are

[1] The authors would like to thank MetaMetrics® for their permission to publish Lexile scores in the present paper. https://www.metametricsinc.com/lexile-framework-reading

opened, and which keystrokes are performed within those screens. Though not applicable for the present study, this information can lead to better insights regarding translators' usage of external resources.

In total, we collected forty student post-editing sessions and forty student human translation sessions. In this paper, we'll focus on the post-editing sessions only. Each of the eight texts was post-edited by five different students. For some segments, some of the data was missing, so these segments were left out of the analysis. The final dataset consisted of 317 post-edited segments.

3.4. MT quality annotation

MT quality can be measured in a myriad of ways, depending on the goal of the assessment and the means available. Automatic metrics like BLEU (Papineni, Roukos, Ward, & Zhu, 2002) are often used to evaluate the output of MT systems by comparing it to reference translations. While these metrics give an indication of an MT system's performance, they rely on the idea that MT quality in itself should approach human quality. Metrics like human-targeted translation error rate (HTER)(Snover et al., 2006), focus more on the perspective of post-editing: how much editing effort is needed to make the MT output match a reference translation? This is the difference between judging the quality of MT output as a final product, and judging the utility of MT output for subsequent post-editing, which is discussed in more detail by Denkowski and Lavie (2012). While both types of metric have proven their worth for different applications, they depend on the availability of human reference translations, which is not something that is always readily available. Ideally, MT evaluation can take place without resorting to reference translations, so it can be used on new texts as well. The translation quality assessment approach presented in this paper builds on that notion, while at the same time being flexible enough so it can be used to evaluate human translation quality and post-editing quality as well (while not relevant for this particular study, it is of importance to the larger study we are conducting).

The quality of MT output is judged from two different perspectives. On the one hand, there is the adherence to the target text and target language norms, also known as acceptability, and, on the other, the adherence to the source text norms and meaning, also known as adequacy (Toury, 1995). This distinction has been used in context of human translation evaluation by Koby and Champe (2013), with acceptability and adequacy issues being called mechanical and transfer errors, respectively. In a more recent study by Koponen and Salmi (2015), where participants had to correct MT output without the source text, the distinction was used successfully as well. Yet the researchers felt the need for a more fine-grained error analysis to better establish which MT errors are the most difficult to edit, and which MT errors lead to meaning loss in the final post-edited text. Koponen and Salmi use Temnikova's (2010) MT error classification and cognitive ranking, but identify a few shortcomings of the ranking, especially with regards to punctuation errors. Lacruz et al. (2014) propose another MT error classification. They use the ATA grading rubric (Koby & Champe, 2013) to distinguish between mechanical and transfer errors, and also create their own classification, which is a simplified version of the ATA's rubric. Lacruz et al. (2014) expect that "cognitive demand placed on post-editors by transfer errors is greater than the cognitive demand resulting from mechanical errors" (p. 77). Mechanical and transfer errors correspond roughly to acceptability and adequacy errors as discussed below, although this is somewhat of an oversimplification. Following Lacruz et al.'s definition (2014), mechanical errors are those errors that can be solved without looking at the source text, whereas our acceptability errors can be identified as errors without looking at the source text, but they cannot necessarily be solved without consulting the source text.

To be able to distinguish more clearly between the effects of acceptability and adequacy issues in MT, we suggest adopting a two-step translation quality assessment approach. In a first step, annotators only receive the target text (in this case, MT output) and they annotate all acceptability issues (grammar and syntax, coherence, lexicon, style and register, spelling). In a second step, annotators receive both the source and the target text, and they annotate all discrepancies in meaning between source and target text, i.e. adequacy issues. It is possible for issues to have an impact on adequacy as well as acceptability, in which case both errors will be annotated. For example, a word sense error is an adequacy issue that can lead to a logical problem in the target text, which in turn is an acceptability issue.

We have tested and fine-tuned our two-step translation quality assessment approach in two pilot studies with student translators, on two different text types (newspaper articles and technical texts), and have successfully applied the approach on MT output as well. To ensure as much objectivity and quality of the annotations as possible, two different people - both authors of this paper with a background in translation studies and evaluation - annotated all the texts. After the annotation process, the annotators discussed discrepancies in their annotations, and only the annotations that both annotators agreed on were kept for the final analysis.

The annotations were made with the brat rapid annotation tool (Stenetorp et al., 2012). To allow for a deeper analysis than hitherto possible, we created a very fine-grained analysis for acceptability and adequacy issues (for an overview of subcategories, see Daems, Macken, & Vandepitte, 2013). Though originally intended for translation evaluation of English to Dutch texts, the categorization builds on common evaluation metric categories and can easily be expanded to other languages. For example, when working with grammatically more complex languages, subcategories for cases could be added to the acceptability category 'grammar and syntax'.

In line with Temnikova (2010) and Lacruz et al. (2014), we believe that error categorizations need to incorporate some method of ranking the different errors according to severity. The errors in our categorization receive error weights ranging from 0 (no actual error, but can be interesting to annotate, such as explicitations), to 4 (critical problems, such as contradictions, that have a critical impact on the understandability of the text). Depending on the text type or task, the weights can be set differently. For example, in technical texts, terminology errors are critical errors, whereas they are not as dramatic in general texts. While we did not originally assign error weights with cognitivity in mind, but rather with the translation product in mind, we do see a close correspondence between the two aspects. For example, structural issues (error weight = 3) will be cognitively demanding to solve, but they also make the text as a product harder to understand. Likewise, capitalization errors (error weight = 1) are easy to solve, and they hardly impair the understanding of the text.

3.5. MT data

Of the 63 source text segments, only three segments contained no errors in the MT output. In total, 201 acceptability issues were identified, and 86 adequacy issues. Though the original error categorization is really fine-grained to allow for detailed analysis, the current dataset is a bit too small to be able to perform any statistical tests on such a detailed level. We therefore decided to group some of the categories together into higher-order categories, so that each category appeared at least ten times in the dataset. The final classification can be seen in Figure 1.

All subcategories for style and spelling have been grouped together into the main categories, since there were very few instances of these subcategories. For lexicon, the subcategory 'wrong collocation' appeared often enough to stand alone, the other subcategories (wrong preposition, named entity, and word non-existent) have been grouped into 'lexicon other'. For coherence issues, the category 'logical problem' occurred more than ten times, but the other categories together (conjunction, missing info, paragraph, and inconsistency) did not occur more than ten times, so all coherence categories were grouped together. The category 'adequacy' in Figure 1 contains all forms of adequacy issues. Mistranslations and word sense issues occurred frequently enough to be considered as separate categories, the other subcategories (additions, deletions, misplaced words, function words, part of speech, inconsistent terminology) were grouped together into 'adequacy other'. Within the grammar and syntax category (the most common error category for MT output), word order issues, structural issues and incorrect verb forms occurred more than ten times each. The different types of agreement issues (noun-adjective, article-noun, subject-verb, and reference) were grouped into a new 'agreement' category, and the other grammatical issues are contained in the 'grammar other' category (superfluous or missing elements).

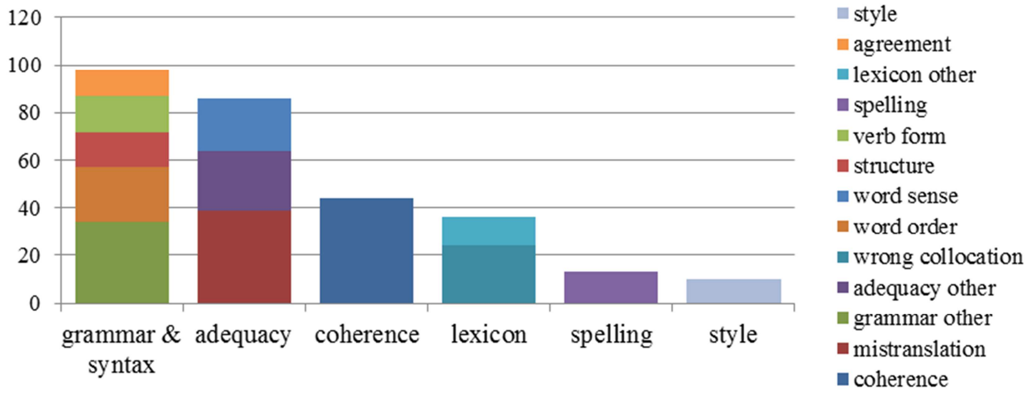

Figure 1. Overview of regrouping and number of occurrences of each error type in the MT output.

4. Analysis

The main goal is to analyze the effect of machine translation quality on different types of post-editing effort indicators. We looked at MT quality on three different levels. For the first level, we simply look at the effect of the average total error weight per word on the different effort indicators. As discussed above, some errors lead to adequacy problems as well as acceptability problems. In these cases, only the adequacy error weight was taken into account for the calculation of the total error weight, as the acceptability error was caused by the adequacy error. For the second level, we look at the impact of the average acceptability and adequacy error weight per word on the different post-editing effort indicators. For the third level, we go even more fine-grained, and we identify the different subcategories that are best suited to predict changes in the post-editing effort indicators.

Based on previous research, we look at the following post-editing effort indicators as dependent variables:

- **Average number of production units**: technical and/or cognitive effort (see discussion of Lacruz et al., 2012 in Section 2), calculated by dividing the number of production units of a segment by the number of source text words in that segment

- **Average time per word**: temporal and/or cognitive effort (see discussion of Koponen et al., 2012 in Section 2), calculated by dividing the total editing time (in ms) of a segment by the number of source text words in that segment
- **Average fixation duration**: cognitive effort (Just and Carpenter, 1980), calculated by dividing the total fixation duration (in ms) of a segment by the number of fixations within that segment
- **Average number of fixations**: cognitive effort (Doherty and O'Brien, 2009), calculated by dividing the number of fixations in a segment by the number of source text words in that segment
- **Pause Ratio**: technical and/or cognitive effort, as suggested by O'Brien (2006), calculated by dividing the total time in pauses (in ms) for a segment by the total editing time (in ms) for that segment
- **Average Pause Ratio:** technical and/or cognitive effort, as suggested by Lacruz et al. (2012), calculated by dividing the average time per pause in a segment by the average time per word in a segment[2]

The dependent variables were derived from the SG-data files obtained by processing the output from CASMACAT. Production units are sequences of coherent typing activity, separated from one another by pauses of at least 1000 ms. A segment in CASMACAT corresponds to a sentence in the source text. A pause in CASMACAT corresponds to any pause in typing activity lasting at least 1000 ms - the lowest possible pause threshold present in the CASMACAT output. The fixation information is added to the CASMACAT tables via the EyeLink plugin.

4.1. Level 1: Average total MT error weight

We used the R statistical software (R Core Team, 2014) to perform linear mixed effects analyses with the lme4 package (Bates et al., 2014) and the lmerTest package (Kuznetsova et al., 2014). For each of our independent variables, we built a null model without fixed effect, but with sentence code and participant as random factors, to account for between text and between participant variation. The only exception was the dependent variable 'average fixation duration', for which the output from the step-function from the lmerTest package showed that only participant was necessary as a random factor. We then tested this null model against a model with fixed effect: the average total MT error weight. As can be seen in Table 1, the model with fixed effect was always significantly different from the null model without fixed effect, with p ranging from < 0,001 (when predicting the average number of production units or the average pause ratio) to 0,017 (when predicting the average fixation duration). Likewise, the Akaike's Information Criterion (AIC) value is always lower for the model with predictor. AIC (Akaike, 1974) is a method designed for model selection, based on a comparison between models. According to Burnham and Anderson (2004), the best model is the model with the lowest AIC value. Their rule of thumb states that if the difference between models is less than 2, there is still substantial support for the weakest model. If the difference is between 4

[2] We would like to thank one of the anonymous reviewers for pointing out that the measure of average pause ratio in this paper is somewhat different from that presented in Lacruz et al. (2012), since the pause threshold is set at 1000ms, whereas Lacruz et al. included clusters of shorter pauses as well. This needs to be taken into account when studying these findings.

and 7, there is far less support for the weakest model, and if the difference is greater than 10, there is hardly any support for the weakest model. As can be seen in Table 1, the difference in AIC values ranges from 17 to 4, always in favor of the model with average total MT error weight as predictor variable. It must be noted that the AIC value in itself has no meaning. The values are used for comparison between models predicting the same dependent variable, but cannot be compared across models predicting different dependent variables.

Dependent variable	**Random factors**	**AIC without predictor**	**AIC with predictor**	**effect**	**p**
Average number of production units	sentence code, participant	-78	-95	0,3 (± 0,06)	< 0,001
Average duration per word (in ms)	sentence code, participant	5979	5974	3077 (± 1153)	0,01
Average fixation duration (in ms)	participant	2890	2886	12 (± 5)	0,017
Average number of fixations	sentence code, participant	2268	2262	8,6 (± 3)	0,005
Pause Ratio	sentence code, participant	-688	-698	-0,07 (± 0,02)	0,002
Average Pause Ratio	sentence code, participant	1596	1580	-3,86 (± 0,85)	< 0,001

Table 1. Summary of mixed models with average total MT error weight as fixed effect.

The impact of the predictor variable can be derived from the 'effect' column in Table 1. The column should be read as follows: for each increase of the average MT error weight per word by one unit, the corresponding dependent variable changes by the value in the 'effect' column. For example, for each unit increase in the average MT error weight per word, the average duration per word increases with 3 seconds. All of the models show the expected effects. A decrease in MT quality, i.e., an increase in average MT error weight, leads to an increase of the number of production units, the average duration per word, the average fixation duration, and the average number of fixations, and to a decrease of the pause ratio and average pause ratio. The latter is in line with findings by Lacruz et al. (2012) that high cognitive load is related to lower average pause ratios. It's remarkable as well that we find a small but statistically significant effect of MT quality on pause ratio. O'Brien (2006) looked at the effect of negative translatability indicators on pause ratio and did not find a statistically significant difference between sentences with NTIs and with few or no NTIs. We can assume that MT error weights provide a more accurate estimation of pause behavior than NTIs, although the direction of the effect is somewhat surprising. Following O'Brien (2006), higher cognitive load should be associated with a higher pause ratio, which is in contrast with the effect we see in Table 1. More detailed analysis is needed to further examine this effect. Our findings for the average fixation duration seem to support the findings by Jakobsen and Jensen (2008) that increased task complexity leads to longer fixations. It must be noted, however, that - though statistically significant - the observed change of 12 ms in our study is rather small and perhaps not practically significant. The change in average number of fixations seems to be a more convincing effort indicator, as was found by Doherty and O'Brien (2009).

4.2. Level 2: Average acceptability and adequacy MT error weight

In the second step, we used the same null models as in the step above, but we applied a more fine-grained approach. Both average acceptability and adequacy error weight were added as possible predictor variables to the model. The step-function from the lmerTest package was used to find out which elements needed to be retained in the final model. This step-function showed that adding adequacy as a predictor did not significantly improve the model, so the final model consisted of only average acceptability error weight as a predictor. The random effects were the same as in the previous analysis. A summary of the models can be found in Table 2.

Dependent variable	**Predictor variable(s)**	**AIC without predictor**	**AIC with predictor**	**effect**	**p**
Average number of production units		-78	-92		
	acceptability			0,32 (± 0,07)	< 0,001
Average duration per word (in ms)		5979	5973		
	acceptability			3347 (± 1312)	0,013
Average fixation duration (in ms)		2890	2884		
	acceptability			15 (± 5)	0,007
Average number of fixations		2268	2264		
	acceptability			8,5 (± 3,4)	0,015
Pause Ratio		-688	-697		
	acceptability			-0,08 (± 0,02)	< 0,002
Average Pause Ratio		1596	1580		
	acceptability			-4,35 (± 0,97)	< 0,001

Table 2. Summary of mixed models with average total adequacy and acceptability error weight as potential fixed effects.

The fact that the average adequacy error weight is not retained in the model is a bit counterintuitive, as Lacruz et al. (2014), for example, found that transfer errors (which roughly correspond to adequacy errors) were cognitively more demanding than mechanical errors. A possible explanation can be that transfer errors and mechanical errors are not entirely comparable to adequacy and acceptability errors, respectively. From Table 2, we can derive that average MT acceptability error weight is a significant predictor for all the different post-editing effort indicators, with p-values below the 0,01 level, with the exception of the p-values for dependent variables 'average duration per word' and 'average number of fixations', although the values are still well below the generally accepted 0,05 significance threshold.

The AIC values are somewhat different from the AIC values of the models with average total MT error weight as predictor variable, but the difference is never greater than three, so we can assume that both models can be supported. Again, we observe the same trend as with the average error weight per word. An increase in average acceptability error weight, leads to an increase of the number of production units, the average duration per word, the average fixation duration, and the average number of fixations, and to a decrease of the pause ratio and average pause ratio.

4.3. Level 3: Average MT error weight for all subcategories

In a final step, we wanted to get a better idea of exactly which types of machine translation errors best predict the different types of post-editing effort indicators. We again used the post-editing effort indicators as dependent variables, and sentence and participant as random factors. This time, however, we added the average MT error weight for all the different subcategories to the model as potential predictor variables: mistranslation, word sense, adequacy other, coherence, style, lexicon other, wrong collocation, spelling, grammar other, structure, verb form, word order, and agreement (see Figure 1). The step-function was used to identify the variables that significantly added to the model. Only these variables were added to the final model, of which the results can be seen in Table 3.

The column 'predictor variables' gives an overview of the different subcategories that predict a change in the dependent variable. Comparing the AIC values of the model with predictors as shown in Table 3 with those from Table 2, we can see that there is more support for the fine-grained model than for the model with the average total MT error weight as a predictor for average duration per word, average fixation duration, and pause ratio. The opposite is true for average number of production units, and average pause ratio. It must be noted that AIC penalizes models with more predictor variables, and seeing how both the model predicting production units and the model predicting average pause ratio contain three or four predictor variables (in comparison with only one or two for the other post-editing effort indicators), this is not such a surprising fact.

What is interesting, is how the different post-editing effort indicators are influenced by different MT error types. The pause ratio and average pause ratio seem to be predicted by a subset of the variables that are predictors for the average number of production units. This is in line with our hypothesis that production units and pauses are closely related to one-another (seeing how the boundaries of production units are defined by pauses). Our findings are comparable to those of Lacruz et al. (2014), who found a strong correlation between pause to word ratio (an alternative for average pause ratio) and mistranslations and structural issues. They also found a correlation with insertions and deletions that we did not find in our data. This can be explained by the fact that insertions and deletions rarely occurred in our data (three and six times respectively), and perhaps their effect is nullified by grouping them together with other categories. The surprisingly negative effect of error weigh on pause ratio as found in Table 1, might be explained by the types of errors found in Table 3: grammatical errors and word order errors are easily spotted, and also easily corrected. This would imply that sentences containing a lot of grammatical or word order errors require fewer time in pauses than sentences containing other types of errors, since these errors can be solved immediately. The average duration per word is predicted most by average MT error weight for coherence and structure issues, which indeed take a lot of time to solve: coherence issues require a post-editor to figure out how the text is semantically structured, whereas structural issues often contain a combination of grammatical structures, so that there are different ways of solving the problem, leading to a higher cognitive load and thus processing time. Fixation duration can be predicted by the average MT error weight for mistranslations, which can be explained by the fact that a mistranslation draws the attention and the problem is often harder to understand than is the case with, for example, grammatical errors. The average number of fixations can be predicted by the average MT error weight for coherence issues. Solving coherence issues requires a post-editor to look back and forth throughout the text to figure out how everything is connected, and so more fixations are needed.

Dependent variable	Predictor variables	AIC without predictors	AIC with predictors	effect	p
Average number of production units		-78	-91		
	mistranslation			0,32 (± 0,13)	0,014
	grammar			0,34 (± 0,41)	0,018
	structure			0,58 (± 0,22)	0,01
	word order			0,43 (± 0,19)	0,028
Average duration per word (in ms)		5979	5964		
	coherence			6365 (± 2464)	0,012
	structure			8020 (± 3912)	0,044
Average fixation duration (in ms)		2890	2882		
	mistranslation			30 (± 9)	0,002
Average number of fixations		2268	2264		
	coherence			16,4 (± 6,5)	0,015
Pause Ratio		-688	-700		
	grammar			-0,15 (± 0,04)	0,001
	word order			-0,13 (± 0,06)	0,036
Average Pause Ratio		1595	1587		
	mistranslation			-4,52 (± 1,74)	0,012
	structure			-6,27 (± 3,04)	0,043
	word order			-6,1 (± 2,66)	0,025

Table 3. Summary of mixed models with average MT error weight for the subcategories retained by step function as fixed effects and sentence code and participants as random factors.

5. Conclusion and discussion

To be able to predict post-editing effort, we need to look at source text and MT output features as possible influencing factors of the post-editing process. In this paper, we investigated how translation students' post-editing process was influenced by the average error weight of the MT output. We found that average MT error weight is a good predictor of six different post-editing effort indicators (average number of production units, average duration per word, average fixation duration, average number of fixations, average pause ratio, and pause ratio). The analysis was conducted on three levels of granularity, by means of linear mixed effects models. With regards to the more fine-grained level, we found that the different post-editing effort indicators are predicted by different MT error categories, with mistranslations, structural issues and word order issues being the most common categories. The average number of production units and the (average) pause ratio seem to be linked, as they are best predicted by comparable MT error categories, consisting of more technical fixes. Cognitively more demanding fixes (coherence and mistranslation) are better predictors for other types of post-editing effort indicators (average fixation duration, average number of fixations, and average duration per word).

We only looked at a few potential post-editing effort indicators and only at MT quality as a possible cause, but there are of course many more indicators and potential causes that can help us better understand the post-editing process. In the future, we would like to look at syntactical variety between source and target language and translation entropy (Carl & Schaeffer, 2014). Other directions for future research include a more fine-grained analysis. We now looked at the segment level, but it could be interesting to look at production units and pauses in isolation, as there is usually a very long pause at the beginning of a segment before the first edit that might influence the pause data. In addition, we want to compare fixations on the source and target text, and focus more on specific MT errors rather than on the entire dataset at once.

Though we only had time to analyze the students' data, we conducted the same experiment with professional translators, and it will be interesting to compare our current findings with the results from the professional translators' data.

References

Akaike, H. (1974), A new Look at the Statistical Model Identification. *IEEE Transactions on Automatic Control, 19*(6), 716–723.

Alabau, V., Bonk, R., Buck, C., Carl, M., Casacuberta, F., Martínez, M., González, J., Koehn, P., Leiva, L., Mesa-Lao, B., Ortiz, D., Saint-Amand, H., Sanchis, G., & Tsoukala, C. (2013). CASMACAT: An Open Source Workbench for Advanced Computer Aided Translation. *The Prague Bulletin of Mathematical Linguistics, 100,* 101–112.

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2014). lme4: Linear mixed-effects models using Eigen and S4. R package version 1.1-7, url: http://CRAN.R-project.org/package=lme4

Burnham, K., & Anderson, D. (2004). Multimodel inference: understanding AIC and BIC in Model Selection. *Sociological Methods & Research, 33*, 261–304.

Carl, M., & Schaeffer , M. (2014). Word Transition Entropy as an Indicator for Expected Machine Translation Quality. In K. J. Miller, L. Specia, K. Harris, & S. Bailey (Eds.), *Proceedings of the Workshop on Automatic and Manual Metrics for Operational Translation Evaluation. MTE 2014*. Paris: ELRA, 45-50).

Daems, J., Macken, L., & Vandepitte, S. (2013). Quality as the sum of its parts: A two-step approach for the identification of translation problems and translation quality assessment for HT and MT+PE. In *Proceedings of the MT Summit XIV Workshop on Post-editing Technology and Practice*, 63-71.

Denkowski, M., & Lavie, A. (2012). Challenges in Predicting Machine Translation Utility for Human Post-Editors. In *Proceedings of AMTA 2012.*

Doherty, S., & O'Brien, S. (2009). Can MT Output be Evaluated through Eye Tracking*? In MT Summit XII*, Ottawa, Canada, 214–221.

Jakobsen, A., & Jensen, K. (2008). Eye Movement Behaviour Across Four Different Types of Reading Task. *Copenhagen Studies in Language, 36,* 103-124.

Just, M., & Carpenter, P. (1980). A theory of Reading: From Eye Fixations to Comprehension. *Psychological Review 87*(4), 329-354.

Koby, G., & Champe, G. (2013). Welcome to the Real World: Professional-Level Translator Certification. *Translation & Interpreting, 5*(1), 156-173.

Koponen, M., & Salmi, L. (2015). On the Correctness of Machine Translation: A Machine Translation Post-Editing Task. *The Journal of Specialised Translation, 23,*118-136.

Koponen, M., Aziz, W., Ramos, S., & Specia, L. (2012). Post-editing Time as a Measure of Cognitive Effort. In *Proceedings of the AMTA 2012 Workshop on Post-editing Technology and Practice (WPTP 2012)*, San Diego, California, 11-20.

Krings, H. (2001). *Repairing Texts: Empirical Investigations of Machine Translation Post-Editing Processes*. Kent, Ohio and London: Kent State University Press. (Translated from 1994 German thesis and edited by G. S. Koby).

Kuznetsova, A., Brockhoff, P, & Christensen, R. (2014). lmerTest: Tests in Linear Mixed Effects Models. R package version 2.0-20. url: http://CRAN.R-project.org/package=lmerTest

Lacruz, I., Denkowski, M., & Lavie, A. (2014). Cognitive Demand and Cognitive Effort in Post-Editing. In S. O'Brien, M. Simard, and L. Specia (Eds.), *Third Workshop on Post-Editing Technology and Practice*. *AMTA 2014*, 73-84.

Lacruz, I., Shreve, G., & Angelone, E. (2012). Average Pause Ratio as an Indicator of Cognitive Effort in Post-Editing: A Case Study. In S. O'Brien, M. Simard, L. Specia (Eds.), *Proceedings of the AMTA 2012 Workshop on Post-Editing Technology and Practice*. Stroudsburg, PA: Association for Machine Translation in the Americas (AMTA).

Leijten, M., & Van Waes, L. (2013). Keystroke Logging in Writing Research: Using Inputlog to Analyze and Visualize Writing Processes. *Written Communication 30*(3), 358-392.

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

O'Brien, S. (2004). Machine Translatability and Post-Editing Effort: How do they Relate? *Translating and the Computer, 26*. London: Aslib.

O'Brien, S. (2006). Pauses as Indicators of Cognitive Effort in Post-Editing Machine Translation Output. *Across Languages and Cultures, 7*(1), 1-21.

Papineni, K., Roukos, S., Ward, T., & Zhu, W.-j. (2002). BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, 311-318.

Plitt, M., and Masselot, S. (2010). A Productivity Test of Statistical Machine Translation Post-Editing in a Typical Localisation Context. *The Prague Bulletin of Mathematical Linguistics, 93*, 7-16.

Snover, M., Dorr, B., Schwartz, R., Micciulla, L., & Makhoul, J. (2006). A Study of Translation Edit Rate with Targeted Human Annotation. In *Proceedings of AMTA 2006*

Stenetorp, P., Pyysalo, S., Topić, G., Ohta, T., Ananiadou, S., & Tsujii, J. (2012). brat: a Web-based Tool for NLP-Assisted Text Annotation. In *Proceedings of the Demonstrations Session at EACL 2012*.

R Core Team (2014). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. url http://www.R-project.org/.

Temnikova, I. (2010). Cognitive Evaluation Approach for a Controlled Language Post-Editing Experiment. In *Proceedings of the seventh international conference on Language Resources and Evaluation (LREC 2010)*, Valletta, Malta, 3485-3490.

Toury, G. (1995). The Nature and Role of Norms in Translation. *Descriptive Translation Studies and Beyond*. 53-69. Amsterdam-Philadelphia: John Benjamins.