frontiers

Post-editing Effort of a Novel with Statistical and Neural Machine Translation

Antonio Toral,[1,*] **Martijn Wieling,**[1] **and Andy Way** [2]

[1] *Center for Language and Cognition, Faculty of Arts, University of Groningen, The Netherlands*
[2] *ADAPT Centre, School of Computing, Dublin City University, Ireland*

Correspondence*:
Corresponding Author
a.toral.ruiz@rug.nl

ABSTRACT

We conduct the first experiment in the literature in which a novel is translated automatically and then post-edited by professional literary translators. Our case study is *Warbreaker*, a popular fantasy novel originally written in English, which we translate into Catalan. We translated one chapter of the novel (over 3,700 words, 330 sentences) with two data-driven approaches to Machine Translation (MT): phrase-based statistical MT (PBMT) and neural MT (NMT). Both systems are tailored to novels; they are trained on over 100 million words of fiction. In the post-editing experiment, six professional translators with previous experience in literary translation translate subsets of this chapter under three alternating conditions: from scratch (the norm in the novel translation industry), post-editing PBMT, and post-editing NMT. We record all the keystrokes, the time taken to translate each sentence, as well as the number of pauses and their duration. Based on these measurements, and using mixed-effects models, we study post-editing effort across its three commonly studied dimensions: temporal, technical and cognitive. We observe that both MT approaches result in increases in translation productivity: PBMT by 18%, and NMT by 36%. Post-editing also leads to reductions in the number of keystrokes: by 9% with PBMT, and by 23% with NMT. Finally, regarding cognitive effort, post-editing results in fewer (29% and 42% less with PBMT and NMT respectively) but longer pauses (14% and 25%).

Keywords: literary translation, post-editing, neural machine translation, statistical machine translation, foreign literature, foreign fiction

1 INTRODUCTION

Machine Translation (MT) is widely used in the translation industry today to assist professional human translators, as using MT results in notable increases in translator productivity compared to translation from scratch. This has been empirically shown in many use-cases over the last decade that rely on the phrase- and rule-based paradigms to MT (PBMT and RBMT), for several text types, including technical documents (Plitt and Masselot, 2010) and news (Martín and Serra, 2014), to mention just two.

The most common workflow employed is post-editing, a sequential pipeline in which the source document is first translated with MT, and subsequently, a translator edits the MT translation (e.g. fixing errors) to produce the final translation. In most of the use-cases explored in the literature the translation aim is

dissemination, and the translations obtained via post-editing have been found to be of equivalent or higher quality (Plitt and Masselot, 2010; Green et al., 2013) to those produced from scratch.

Nonetheless, post-editing has been found to prime the translator, thus resulting in a final translation that is similar to that initially suggested by the MT system (Green et al., 2013). Because the MT approaches most widely used to date in post-editing translation workflows – RBMT and, above all, PBMT – are known to lead to literal translations, post-edited translations are also perceived as being more literal than translations from scratch (Martín and Serra, 2014).

While this is acceptable for text types such as technical documents, as the main objective of the translation for these types of texts is to preserve the meaning of the original, it might not be the case for other text types of a more creative nature, such as literary texts, because in this case the objective of the translation is twofold: not only the meaning of the source text needs to be preserved but also its reading experience (Jones and Irvine, 2013).

Recently, neural machine translation (NMT) has emerged as a new paradigm in MT, and has been shown to considerably improve the translation quality achieved, regardless of the language pair (Toral and Sánchez-Cartagena, 2017). In addition, the translations produced by NMT are much more fluent (Bentivogli et al., 2016) than those derived by PBMT, until recently by far the most dominant paradigm in the field. In addition, relevant to this work, it has been claimed that NMT does not lead to literal translations,[1] as is the case with PBMT and RBMT.

At this point, because of (i) the maturity of post-editing in industry, and (ii) the rise of a new MT paradigm (NMT) that results in more fluent and less literal translations than previous models (PBMT and RBMT), it is timely to study the extent to which current MT technology can be useful in assisting with professional translations of literary text. In this work we take the first steps in this direction by assessing the effort involved in the post-editing of a novel, along the three dimensions commonly studied in the literature (Krings and Koby, 2001), which constitute the research questions (RQs) underpinning this work:

- RQ1 (temporal effort). Does post-editing an MT output (using the NMT or PBMT paradigm) result in shorter translation time compared to post-editing of outputs from the other type of MT system and/or to translation from scratch?
- RQ2 (technical effort). Does post-editing on one of the two MT paradigms result in a lower number of keystrokes than the other MT paradigm and/or than translation from scratch?
- RQ3 (cognitive effort). Does post-editing on one of the MT paradigms result in changes in cognitive effort?

In this work we translate a fragment of a novel with NMT and PBMT. Subsequently, six professional translators with previous experience in literary translation translate subsets thereof under three different conditions: from scratch (the norm in the novel translation industry), post-editing the translation produced by the PBMT system and post-editing that generated by the NMT system. For each sentence translated, we record (i) the time spent to translate it, (ii) the number of keystrokes used, and (iii) the number of pauses and time devoted to them. We then use these three measurements to attempt to provide answers to questions RQ1, RQ2 and RQ3, respectively.

[1] "Neural network-based MT can, rather than do a literal translation, find the cultural equivalent in another language", according to Alan Packer, Engineering Director at Facebook, in 2016, cf. https://slator.com/technology/facebook-says-statistical-machine-translation-has-reached-end-of-life

The rest of the paper is organised as follows. Section 2 outlines the state-of-the-art in MT of novels. Next, Section 3 presents the MT systems (Section 3.1) and the novel (Section 3.2) used in our experiment, followed by the experimental set -up (Section 3.3). Section 4 presents and discusses the results. Finally, in Section 5, we draw our conclusions and propose lines of future work.

2 STATE-OF-THE-ART IN LITERARY TRANSLATION USING MT

Voigt and Jurafsky (2012) studied how referential cohesion is expressed in literary (short stories) and non-literary (news stories) texts and how this cohesion affects translation. They found that literary texts use more dense reference chains to express greater referential cohesion than news. They then compared the referential cohesion of human versus machine translations of short stories from Chinese-to-English. MT systems had difficulty in conveying the cohesion, which is attributed to the fact that they translate each sentence in isolation while human translators can rely on information beyond the sentence level.

Jones and Irvine (2013) used generic PBMT systems to translate samples of French literature (prose and poetry) including a fragment of Camus' *L'Étranger* into English. They analysed the translations from a qualitative perspective to address what makes literary translation hard and to discover what the potential role of MT could be.

Besacier and Schwartz (2015) presented a pilot study where a generic PBMT system followed by post-editing was applied to translate a short story from English into French. Post-editing was performed by non-professional translators, and the authors concluded that such a workflow can be a useful low-cost alternative for translating literary works, albeit at the expense of lower translation quality.

Simultaneously to the previous work, Toral and Way (2015) built a PBMT system tailored to a contemporary best-selling author (Ruiz Zafón) and then applied it to translate one of his novels, *El prisionero del cielo*, between two closely-related languages (Spanish-to-Catalan). For 20% of the sentences, the translations produced by the MT system and the professional translator (i.e. taken from the published novel in the target language) were exactly the same. In addition, a human evaluation revealed that for over 60% of the sentences, Catalan native speakers judged the translations produced by MT and by the professional translator to be of the same quality.

Murchú (2017) machine translated the sci-fi novel *Air Cuan Dubh Drilseach* from Scottish Gaelic to Irish, a pair of closely related languages, using the hybrid MT system Intergaelic and subsequently post-edited the resulting MT output. Post-editing was 31% faster than translating from scratch and less than 50% of the tokens in the MT output were corrected by the translator.

Toral and Way (2018) built PBMT and NMT systems tailored to novels for the English–Catalan language pair. These were evaluated on a set of twelve widely known novels from the 20^{th} and 21^{st} centuries by authors such as Joyce, Orwell, Rowling and Salinger, to name but a few. Overall, NMT resulted in an 11% relative improvement (3 points absolute) over PBMT according to the BLEU evaluation metric (Papineni et al., 2002). In a human evaluation conducted on the books by Orwell, Rowling and Salinger, the translations generated by the NMT system were perceived by Catalan native speakers to be of equivalent quality to the professional human translations for 14%, 29% and 32% of the sentences, respectively, compared to 5%, 14% and 20% respectively, with the PBMT system. These findings have encouraged us to expand this study to the post-editing of a novel, which we detail below.

3 MATERIAL AND METHODS

3.1 MT Systems

We trained two MT systems belonging to two different paradigms: PBMT and NMT. Both are tailored to novels and a brief description of them follows. For a more detailed account, the reader is referred to Toral and Way (2018).

The PBMT system is trained on a linear interpolation of in-domain (133 parallel novels from different genres amounting to over 1 million sentence pairs) and out-of-domain (around 400,000 sentence pairs of subtitles)[2] parallel data, with version 3 of the Moses toolkit (Koehn et al., 2007). The n-gram-based language model, in addition to the training parallel data, uses monolingual in-domain (around 1,000 books written in Catalan amounting to over 5 million sentences) and out-of-domain (around 16 million Catalan sentences crawled from the web (Ljubešić and Toral, ????)) data. The system uses 3 reordering models (lexical- and phrase-based, and hierarchical), an operation sequence model (Durrani et al., 2011) and an additional language model based on continuous space n-grams (Vaswani et al., 2013). The last two models are trained on the in-domain parallel data.

The NMT system follows the encoder-decoder approach and is built with Nematus (Sennrich et al., 2017).[3] This system is trained on the concatenation of the parallel in-domain training data (133 parallel novels) and a synthetic corpus obtained by machine-translating the Catalan in-domain monolingual training data (1,000 books) into English. The system uses sub-words as the basic translation unit; we segmented the training data into characters and performed 90,000 operations jointly on both the source and target languages (Sennrich et al., 2016). Finally, we generate an n-best list with the NMT system and rerank it with a left-to-right NMT system.[4]

3.2 Novel

The novel used in this experiment is Sanderson's *Warbreaker*.[5] This book fulfills our two requirements, namely (i) literary quality, to make sure that the task is indeed challenging, and (ii) being freely redistributable, to guarantee the reproducibility of our experiment. The first criterion is attested by its reviews by critics, while the second is met as the book was published under a Creative Commons License (CC-NC-ND specifically).

Warbreaker is pre-processed in the same way as the training data, namely it is sentence-split with NLTK (Bird, 2006) and subsequently tokenised, truecased and normalised (in terms of punctuation) with the corresponding Moses scripts.

In order to have an estimate of the difficulty posed by the translation of Warbreaker, we use two automatic metrics. The first, type-token ratio (TTR), provides an indication of the richness of the vocabulary used in the book. The second, n-gram overlap, corresponds to the percentage of n-grams in the novel that are also found in the training data used to build the MT system. This measure thus provides an indication of the degree of lexical divergence (or 'novelty') of the book that is to be translated with respect to the training data.

[2] `http://opus.lingfil.uu.se/OpenSubtitles.php`

[3] `https://github.com/rsennrich/nematus`

[4] This system has the same settings as the regular NMT system, the only difference being that the target sentences of the training data are reversed at the word level.

[5] `https://brandonsanderson.com/books/warbreaker/warbreaker/`

Table 1 shows the TTR and n-gram overlap (for $n = \{2, 3, 4\}$) of Warbreaker (both for the whole book and for some individual chapters)[6] as well as for the 12 books previously translated with our MT systems (Toral and Way, 2018). For the latter we show the mean value for the 12 books as well as the 95% confidence interval. In addition, we calculate the average sentence length (average number of words per sentence) as previous research has shown that the performance of current NMT systems degrades with increasing sentence length (Toral and Sánchez-Cartagena, 2017).

Comparing the scores of *Warbreaker* to those of the twelve well-known books we have previously translated allows us to have an approximation as to how challenging translating *Warbreaker* is going to be. The scores for *Warbreaker* (full book) fall inside the confidence intervals obtained for the twelve books for two measures (2-gram overlap and TTR), they are slightly higher for another two (3- and 4-gram overlap) and slightly lower for the remaining one (sentence length). According to these results we expect the novel chosen to be slightly easier to translate than the average of the twelve novels we translated previously.

As for Warbreaker's individual chapters, we select one for our experiment that has similar values to the whole book, as that would make it (to some extent) representative of the book as a whole. We show the values for the first three (prologue and Chapters 1 and 2) in Table 1 and pick Chapter 1 as it is the one whose results are closest to the whole book for all the metrics considered (except sentence length, whose value is longer than the average).

3.3 Experimental Setup

The professional translators performed the translation using `PET` v2.0 (Aziz et al., 2012),[7] a computer-assisted translation tool that supports both translating from scratch and post-editing. `PET` is used with its default settings. A snapshot of the tool, as used in our experiment, is shown in Figure 1.

The source text translated in the experiment (Warbreaker's Chapter 1) is made up of 3,743 words distributed in 330 sentences. We divided it into 33 translation jobs, each of which is made of 10 consecutive sentences (translation segments). There are three types of translation jobs (translation conditions): translation from scratch (HT), and post-editing the translation produced by the PBMT (MT1) and NMT systems (MT2).[8]

Six translators (henceforth T1 to T6) took part in the study. They saw all factors but not all combinations, since they translated each job in one translation condition. The type of translation to be carried out for each job by each translator is chosen randomly, with the following three constraints:

1. The first job is set to translation condition HT for translators T1 and T2, to MT1 for T3 and T4 and to MT2 for T5 and T6.
2. Two consecutive jobs by a translator cannot follow the same translation condition.
3. For each translator the number of jobs under each translation condition is equal, i.e. each translator translates 11 jobs under translation condition HT, 11 under MT1 and 11 under MT2.

We provided the translators with comprehensive translation guidelines,[9] where it is stated that the aim is to achieve publishable professional quality translations, both for translations from scratch and for post-editing. With respect to post-editing, the guidelines encourage the translator to try to fix the translation provided

[6] TTR scores are not shown for chapters as it is computed on 20,000 words, a much bigger amount of text than what makes up a chapter.

[7] `http://rgcl.wlv.ac.uk/projects/PET/resources/PET-v2.0.tgz`

[8] We referred to the two MT systems as MT1 and MT2 throughout the experiments so that the translators could not know anything about the MT paradigm into which they fell.

[9] The manual is available at TODO

by the MT system. Only if this is deemed too time-consuming to fix (e.g. because the quality of the MT output is too low) were the translators instructed to delete it and carry out the translation from scratch.

As in other computer-assisted tools, translations in PET are related to source sentences on a one-to-one basis. In other words, each source sentence corresponds to one target sentence (see Figure 3.3). However, in the translation of novels it is not that uncommon to have some cases of many-to-one (more than one source sentence translated as one target sentence) or one-to-many (one source sentence translated as more than one target sentence) translations. Due to this characteristic of literary translation, translators were told that they could, in addition to one-to-one translations, perform one-to-many and/or many-to-one translations. Details on how they could go about this are provided in the translator's manual.

For each research question (temporal, technical and cognitive effort), we first report the (descriptive) results for the samples. For example, for temporal effort, the relative change in translation productivity with post-editing versus translating from scratch is provided. Subsequently, we aim to generalise from the samples (the translators that participate in the study and the sentences they translate) to populations (any translator and any similar text) by using mixed-effects regression models (Baayen, 2008).[10] Mixed-effects regression models distinguish both fixed effects (i.e. the effects we are usually interested in) from random effects (i.e. the factors we would like to generalise over). With respect to random effects, a distinction can be made between random intercepts (i.e. the value of the dependent variable varies on the basis of the level of the random-effect factor), and random slopes (i.e. the strength of the effect of a predictor varies on the basis of the level of the random-effect factor). Specifically, we will build models where we contrast the three translation conditions by including them as fixed effects, while including the translators (6 levels) and translation segments, or sentences (330 levels) as random effects.

Previous studies in post-editing have shown that results vary considerably between translators and segments. By taking a mixed-effects regression approach, we are able to include the by-translator and by-segment random intercepts and slopes to model the variability associated with translator and segment. For example, one individual translator may tend to take longer, or rewrite a larger part of a sentence than another, which is modelled by a by-translator random intercept. Similarly, one sentence (due to its structure) may be more likely to be rewritten than another, or may take more cognitive effort to translate, which is modelled by a by-segment random intercept. In addition, one translator might show a greater difference between the three conditions than another, which is modeled by by-translator random slope for translation condition. Similarly, a by-segment random slope for translation condition is able to model that translation condition may show a greater difference for one sentence than for another.

We conduct exploratory analyses, in which we first include random intercepts for translator and segment, and subsequently add fixed-effect predictors one by one. For each predictor, we check whether its addition results in a significantly better statistical model by comparing the model that adds that predictor to a simpler model without that predictor. Any pair of models is subsequently compared in terms of Akaike's Information Criterion (AIC; Akaike, 1973). If the AIC of the model that includes the predictor is at least 2 points lower than the model without the predictor then we consider the first model significantly better. The evidence ratio can be calculated on the basis of the AIC difference[11] and represents the relative probability that the model with the lowest AIC is more likely to provide a more precise model of the data. By using a threshold of 2 (see also Groenewold et al. (2014)), we only select a more complex model if it is 2.7 times more likely than the simpler model. After including the fixed effect predictors separately, we evaluate

[10] For our analysis we use the lme4 R package, for a normal linear regression model or a Poisson generalised linear regression model, but for a ratio as the dependent variable, we use the package mgcv, as beta regression is not implemented in the lme4 package.

[11] Evidence ratio: $e^{\frac{\delta AIC}{2}}$

(using AIC comparisons) if interactions between the fixed-effect predictors are necessary. After obtaining the best fixed-effects structure, we evaluate the optimal random-effects structure (i.e. by including random effects, and evaluating their inclusion again using AIC comparison) and retain all fixed-effect factors which are significant when the appropriate random effects structure is included. This approach is similar to that used by Wieling et al. (2011).

An ethics approval for this study was obtained from The Research Ethics Committee of the Faculty of Arts, University of Groningen. The professional translators involved in the study gave written informed consent in accordance with the Declaration of Helsinki.[12]

4 RESULTS AND DISCUSSION

As was previously mentioned in Section 1, this work has three research questions, concerning temporal (RQ1), technical (RQ2) and cognitive effort (RQ3). Next we detail the pre-processing of the data. The subsequent three subsections attempt to provide answers to these three questions, based on the experimental data collected.

4.1 Pre-processing

For each translated sentence by each translator, we extract the following elements from the `PET` logs: length of the source and target text (in words and characters), translation condition (HT, MT1 or MT2), translation time, number of keystrokes (total as well as belonging to different categories: letters, digits, whitespace, symbols, navigation, deletion, copy, cut and paste), and number of pauses and their duration. Following the findings by Lacruz et al. (2014), we include only pauses longer than 300 milliseconds.

We also pre-processed those translations without a 1-to-1 sentence equivalence. None of the translators produced any 1-to-many translations, and only three out of the six translators generated many-to-1 translations. Moreover, these translators performed such translations in very few cases: from 6 to 10 sentences, i.e. from 1.8% to 3% of the translation units. The reason given by the translators as to why some of them produced many-to-1 translations but no 1-to-many was due to the fact that sentences in novels in Catalan tend to be longer than in English. Accordingly, conflating more than one English sentence into a single Catalan translation equivalent made sense, albeit on rare occasions. The fact that the vast majority of translations were 1-to-1 could be attributed to either of the two following reasons (or a combination of both):

- While the instructions allowed for translations beyond the 1-to-1 sentence equivalence, the computer-assisted tool used expects 1-to-1 sentence equivalence, so translators may feel discouraged to do otherwise;
- While there may be the perception that in original novels and their translations, sentences do not tend to correspond 1-to-1, this is actually the most frequent case, at least for the language pair we cover in this study. In Toral and Way (2018), we sentence-aligned over 100 novels in English and their translations in Catalan. Overall, 77% of the sentences were successfully aligned 1-to-1. The remaining 23% is made up of alignments that are not 1-to-1 but also of 1-to-1 alignments that the alignment tool could not align confidently.

[12] `https://web.archive.org/web/20091015082020/http://www.wma.net/en/30publications/10policies/b3/index.html`

4.2 Temporal Effort

First, we report on translation productivity (measured as words per hour) per translation condition, as this is a metric commonly used in related work, e.g. Plitt and Masselot (2010). Overall, translators produce 503 words per hour when translating from scratch (condition HT). Compared to this, post-editing the translation produced by the PBMT results in 594 words per hour, an 18% increase in productivity, while post-editing the NMT output leads to double that figure: 36% (685 words per hour). This is clearly indicative of the fact that NMT outputs were superior to those from PBMT.

We now zoom in and look at each translator individually. Results are shown in Figure 2. We can observe a large variability in translation speed, from the lowest value of 402 words/hour (translator T3, condition HT) to the highest of 1,140 (T2, MT2). Despite this variability, we can observe clear trends when comparing translation conditions: all translators are faster in condition MT1 compared to HT (relative increases range from 1% for T6 to 46% for T2), and all are faster with MT2 than with MT1 (increases range from to 0.001% for T1 to 37% for T3).

Next, in order to generalise from samples to populations and to find out whether differences are statistically significant, we build a linear mixed-effects regression model in which we predict translation time[13] given two (fixed-effect) numerical predictors (length of the segment in characters and trial number), one fixed-effect factorial predictor (translation condition) and two random-effect factors (translators and segments). Numerical predictors are centred and scaled. After fitting the final model, we conduct model criticism by excluding data points which have an observed value deviating more than 2.5 standard deviations from the predicted value by the model[14] and refit the model. In this way, we prevent that potentially significant effects are "carried" by these outliers (which are not well represented by the model; Baayen, 2008). We assessed that the residuals of our final model approximately followed a normal distribution and were homoscedastic.

In the best model, the two numerical fixed predictors are significant: translators take longer time the longer the input text and shorter time as they advance through the experiment (trial number). The effect of translation condition is also significant: compared to HT, translation time in condition MT1 is significantly reduced, and so this is also the case for MT2 compared to MT1.

We find a significant interaction between input length and translation condition. Figure 3 shows that the longer the input sentence the lower the advantage of MT2 over HT. There is no such effect for MT1 though. The fact that post-editing NMT is not advantageous over translating from scratch for long sentences corroborates the finding that the translation quality provided by NMT degrades with sentence length (Toral and Sánchez-Cartagena, 2017). Table 3 shows the significance level for each predictor and interaction between predictors, not only for the model built for temporal effort but also for those used for technical and cognitive effort (see Sections 4.3 and 4.4, respectively).

In terms of the random-effects structure, the final model included both random intercepts (by segment and translator), and a by-item (segment) random slope for translation condition. The random slope reflects that the difference in temporal effort between the three conditions varies per segment.

[13] We transform it logarithmically, since its distribution is heavily skewed to the right.

[14] A total of 56 out of 1,980 data points (2.8%) were removed.

4.3 Technical effort

We measure the technical effort by means of the number of keystrokes used to produce the final translation. Similarly to what was done for temporal effort (cf. Section 4.2), we calculate the number of keystrokes per character in the source sentence and per translation condition (HT, MT1 and MT2), i.e. the number of keystrokes that it takes to translate one character with each translation method. Overall, it takes 1.94 keystrokes to translate each character when translating from scratch (condition HT). Compared to this, post-editing PBMT (condition MT1) results in a 9% reduction (1.76 keystrokes per character), while NMT leads to more than double that reduction, 23% (1.49 keystrokes per character).

We now zoom in and look at each translator individually. Results are shown in Figure 4. As with temporal effort, there is large variability across translators and conditions, the lowest value being 0.8 keystrokes per second (translator T2, condition MT2) and the highest 2.9 (translator T5, condition MT1). Some trends arise but they are not as clear as was the case with temporal effort. Compared to HT, the number of keystrokes is reduced with MT1 for three translators (maximum reduction: 45%, T2) and is increased with the other three (maximum increase: 13%, T5). Compared to HT, MT2 results in a reduced number of keystrokes for all translators except T6, for whom it increases slightly (2%). The maximum reduction is, as in the case of MT1, for T2 (59%).

Next, as for temporal effort, we build a statistical mixed model to predict technical effort, for which we consider the same set of predictors. Our dependent variable is the total number of keystrokes. As this dependent variable reflects count data, we use Poisson generalised linear mixed-effects regression. As in temporal effort, all the fixed predictors are significant. The longer the input, the more keystrokes are used and the further a translator advances in the experiment, the fewer keystrokes s/he uses. The effect of post-editing is significant, fewer keystrokes are required with MT1 compared to HT, and the same occurs when we compare MT2 to MT1.

The interaction between input length and translation condition, which was significant for temporal effort, is significant here too, but again only shows a difference between HT and MT2. The interaction is shown in Figure 5. The longer the input sentence, the smaller the difference becomes between the number of keystrokes used in conditions HT and MT2.

The optimal random-effects structure, in this case, consists of both a by-translator and a by-segment random slope for translation condition, and a by-translator random slope for trial (reflecting that the trial, i.e. learning effects, are different per translator).

In the experiment we not only logged the number of keystrokes used but also their type. We now delve deeper into the keystroke results by differentiating the keystrokes into three groups: content (digits, letters, white space and symbols), navigation keys and erase keys.[15] Figure 4 showed the average number of keys per source character for each translator under each of the translation conditions. Now, we show a different perspective in Table 2, where we break up the average number of total keys into three groups of keys and we aggregate the data for all the translators.

It has been previously shown that post-editing leads to a very different usage of the keyboard compared to translation from scratch (Carl et al., 2011). Our results corroborate this: while post-editing reduces considerably the number of content keywords used (-55% with PBMT and -63% with NMT), that translation

[15] Other types of keystrokes were logged too, for the operations copy, cut, paste and undo. However, their usage was negligible in the experiment; they account for just 0.1% of the total number of keystrokes used, so have not been included in our analysis.

pipeline results in a massive increase in the use of navigation keys (228% with PBMT and 195% with NMT) and, to a lesser extent, erase keys (105% for PBMT and 72% with NMT).

Figure 7 shows a complementary view of this data. For each translation condition, we depict the proportion of keys that belong to each of the three groups considered (content, navigation and erase). In translation from scratch content keystrokes make up 79% of the total, navigation 9% and erase the remaining 12%. Post-editing leads to roughly equal percentages for each keystroke category: 39% content, 34% for navigation and 27% for erase with PBMT and 38%, 36% and 27%, respectively with NMT.

Finally, we show the complete piecture with three variables at once (translators, translation condition and keystroke type) in Figure **??**. The trend is similar across translators for content keys; all of them use substantially more keystrokes when translating from scratch than when post-editing. Navigation is the type of keystrokes for which we observe the highest variation across translators; on one extreme two translators (T2 and T6) use very few navigation keys, regardless of the translation condition. On the other, one translator (T5) uses more than double the number of navigation keys than the second translator in number of navigation keys (T3). For erase keys, we see similar trends across translators; all of them except T2 use more erase keys when post-editing than when translating from scratch.

4.4 Cognitive Effort

We use pauses as a proxy to measure cognitive effort (Schilperoord, 1996; O'Brien, 2006). We consider three different ways of expressing the dependent variable (Green et al., 2013):

- Count: the number of pauses.
- Mean duration: how long pauses take on average.
- Ratio: the amount of time devoted to pauses divided by the total translation time.

The number of pauses correlates strongly with the number of keystrokes ($R = 0.87$). Due to this and because number of pauses is a count-dependent variable, we fit number of pauses as the dependent variable with the Poisson regression model previously built for technical effort (see Section 4.3). According to the model, there are 15.3 pauses per sentence when translating from scratch. Condition MT1 significantly reduces this by 29% (10.9) and MT2 by 42% (8.8).

The mean duration of pauses correlates weakly with translation time ($R = 0.25$) and has no correlation with number of keystrokes ($R = -0.02$). We fit the mean duration of pauses[16] with the model previously built to predict translation time (see Section 4.2). Pauses have a mean duration of 2,243 milliseconds in the translating-from-scratch condition. In condition MT1 this significantly increases by 14% (2,559 milliseconds), while in MT2 this increases further, by 25% (2,810 milliseconds).

The ratio of pauses is a proportion, and thus we use beta regression. Pause ratio correlates with translation time ($R = 0.57$) and hence we will use the same predictors, interactions and slopes as in the model previously built to predict time (see Section 4.2). According to the model, pauses take 63% of the translation time in condition HT. Post-editing, be it with MT1 or MT2, leads to significant increments of around 2.5 percentage points (65.6% and 65.3%, respectively) of the time devoted to pauses. The difference between MT1 and MT2 is not significant.

[16] We transform it logarithmically, since its distribution is heavily skewed to the right.

5 CONCLUSIONS AND FUTURE WORK

We have conducted the first experiment in the literature in which a fragment of a novel is translated automatically and then post-edited by professional translators. Specifically, we have translated one chapter of *Warbreaker* (over 3,700 words) from English into Catalan with domain-specific PBMT and NMT systems.

The experiment has been conducted by six professional translators, who translated consecutive fragments of 10 sentences each in three alternating conditions: from scratch, post-editing PBMT, and post-editing NMT. The time taken for each segment as well as the keystrokes used, the number of pauses and the duration of pauses were recorded, which has allowed us to analyse the translation logs and study how post-editing with PBMT and NMT affects temporal, technical and cognitive effort.

Regarding temporal effort, compared to translation from scratch, both PBMT and NMT lead to substantial increases in translation productivity (measured as word per hour), of 18% and 36%, respectively. This demonstrates convincingly that post-editing MT output – whatever the system – makes translators faster than when they translate from scratch. Furthermore, it indicates that translations output by NMT engines were better than those from the corresponding PBMT systems. In addition, we found that the gain with PBMT remains constant regardless of the length of the input sentence, while the gain with NMT decreases with long sentences.

With respect to the number of keystrokes used (the measure used for technical effort), NMT again resulted in a more substantial reduction (23%) than PBMT (9%). As with temporal effort, the reduction in the number of keystrokes for PBMT remains constant across input sentences of different length, while the reduction with NMT decreases for long sentences. Finally, we have observed that the distribution of types of keystrokes is very different in post-editing compared to translation from scratch. While the first results in considerably fewer content keywords, it notably increases the number of navigation and erase keystrokes.

As for cognitive effort, which we measured using pauses as proxies, we found that NMT – and to a lesser extent PBMT – significantly reduce the number of pauses (42% and 29%, respectively). Pauses are considerably longer when post-editing (14% with PBMT and 25% with NMT) than when translating from scratch. Finally, we observed that pauses take a longer fraction of the total translation time when post-editing, and that the difference between PBMT and NMT is not significant.

In this study we have looked at post-editing effort, covering its three dimensions: temporal, technical and cognitive. In the next phase of this work, we will explore translators' perceptions, which we recorded during the experiments by means of pre- and post-experiment questionnaires and a debriefing session, and compare these perceptions to the results and conclusions from the current study.

Finally, we will assess the quality of the resulting post-edited translations. In previous post-editing studies this is commonly measured by assessing the translations in terms of adequacy and fluency. For literary texts, however, there is an additional requirement, namely that the translation should preserve the reading experience of the source text. Accordingly, we aim to measure this in our future work.

CONFLICT OF INTEREST STATEMENT

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

AUTHOR CONTRIBUTIONS

AT conceptualised the research, co-designed and conducted the experiments and wrote the manuscript. MW directed the statistical analyses and reviewed/edited the manuscript. AW co-designed the experiments and reviewed/edited the manuscript. All authors listed have made a substantial, direct, and intellectual contribution to the work and approved it for publication.

FUNDING

The research leading to these results has received funding from the European Association for Machine Translation through its 2015 sponsorship of activities programme, proposal named "Pilot on Post-editing Novels (PiPeNovel)". The ADAPT Centre for Digital Content Technology at Dublin City University is funded under the Science Foundation Ireland Research Centres Programme (Grant 13/RC/2106) and is cofunded under the European Regional Development Fund.

ACKNOWLEDGMENTS

We would like to thank the six professional translators that took part in this study, in alphabetical order: X, Y, Z, and N translators that preferred to remain anonymous. In addition, we would like to thank Sheila Castilho and Joss Moorkens for their feedback on the experiment set up and the translation guidelines and Wilker Aziz for his help on processing the PET log files.

SUPPLEMENTAL DATA

The translation guidelines provided to translators, the raw logs from PET and an R notebook (source code and HTML report) with all the statistical analyses conducted are provided with this manuscript as supplementary data.

REFERENCES

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In *Second International Symposium on Information Theory* (Budapest, Hungary), 267–281

Aziz, W., Castilho, S., and Specia, L. (2012). PET: a Tool for Post-editing and Assessing Machine Translation. In *Proceedings of the 8th International Conference on Language Resources and Evaluation* (Istanbul, Turkey), 3982–3987

Baayen, R. H. (2008). *Analyzing linguistic data: A practical introduction to statistics using R* (Cambridge, UK: Cambridge University Press)

Bentivogli, L., Bisazza, A., Cettolo, M., and Federico, M. (2016). Neural versus phrase-based machine translation quality: a case study. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing* (Austin, Texas), 257–267

Besacier, L. and Schwartz, L. (2015). Automated translation of a literary work : a pilot study. In *Proceedings of the Fourth Workshop on Computational Linguistics for Literature* (Denver, Colorado, USA), 114–122

Bird, S. (2006). NLTK: the natural language toolkit. In *Proceedings of the COLING/ACL on Interactive presentation sessions* (Sydney, Australia), 69–72

Carl, M., Dragsted, B., Elming, J., Hardt, D., and Jakobsen, A. L. (2011). The Process of Post-Editing : a Pilot Study. In *Proceedings of the 8th international NLPSC workshop. Special theme: Human-machine interaction in translation* (Copenhagen, Denmark), 131–142

Durrani, N., Schmid, H., and Fraser, A. (2011). A joint sequence translation model with integrated reordering. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1* (Portland, Oregon, USA), 1045–1054

Green, S., Heer, J., and Manning, C. D. (2013). The efficacy of human post-editing for language translation. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Paris, France), 439–448

Groenewold, R., Bastiaanse, R., Nickels, L., Wieling, M., and Huiskes, M. (2014). The effects of direct and indirect speech on discourse comprehension in dutch listeners with and without aphasia. *Aphasiology* 28, 862–884

Jones, R. and Irvine, A. (2013). The (Un)faithful Machine Translator. In *Proceedings of the 7th Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities* (Sofia, Bulgaria), 96–101

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., et al. (2007). Moses: open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions* (Prague, Czech Republic), 177–180

Krings, H. and Koby, G. (2001). *Repairing Texts: Empirical Investigations of Machine Translation Post-editing Processes*. Translation studies (Kent State University Press)

Lacruz, I., Denkowski, M., and Lavie, A. (2014). Cognitive demand and cognitive effort in post-editing. In *AMTA 2014: proceedings of the eleventh conference of the Association for Machine Translation in the Americas, Workshop on Post-editing Technology and Practice (WPTP-3)* (Vancouver, BC), 73–84

Ljubešić, N. and Toral, A. (????). In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14), title = caWaC - a Web Corpus of Catalan and its Application to Language Modeling and Machine Translation, year = 2014, pages = 1728–1732* (Reykjavik, Iceland)

Martín, J. A. A. and Serra, A. C. (2014). Integration of a machine translation system into the editorial process flow of a daily newspaper. *Procesamiento del Lenguaje Natural* 53, 193–196

Murchú, E. Ó. (2017). Bearnaí i litríocht na gaeilge a líonadh: Réiteach úr? (filling gaps in irish-language literature: A novel approach). Presented at AR AN IMEALL I LÁR AN DOMHAIN: An tairseachúlacht i litríocht agus i gcultúr na hÉireann agus na hEorpa, Prague, 14 Sept 2017

O'Brien, S. (2006). Pauses as indicators of cognitive effort in post-editing machine translation output. *Across Languages and Cultures* 7, 1–21

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). BLEU: A Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics* (Philadelphia, PA, USA), 311–318

Plitt, M. and Masselot, F. (2010). A productivity test of statistical machine translation post-editing in a typical localisation context. *Prague Bull. Math. Linguistics* 93, 7–16

Schilperoord, J. (1996). *It's about Time: Temporal Aspects of Cognitive Processes in Text Production*. Utrecht studies in language and communication (Rodopi)

Sennrich, R., Firat, O., Cho, K., Birch, A., Haddow, B., Hitschler, J., et al. (2017). Nematus: a toolkit for neural machine translation. In *Proceedings of the Software Demonstrations of the 15th Conference of the European Chapter of the Association for Computational Linguistics* (Valencia, Spain), 65–68

Sennrich, R., Haddow, B., and Birch, A. (2016). Neural Machine Translation of Rare Words with Subword Units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (Berlin, Germany), 1715–1725

Toral, A. and Sánchez-Cartagena, V. M. (2017). A multifaceted evaluation of neural versus phrase-based machine translation for 9 language directions. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers* (Valencia, Spain), 1063–1073

Toral, A. and Way, A. (2015). Translating Literary Text between Related Languages using SMT. In *Proceedings of the Fourth Workshop on Computational Linguistics for Literature* (Denver, Colorado, USA), 123–132

Toral, A. and Way, A. (2018). What level of quality can neural machine translation attain on literary text? In *Translation Quality Assessment: From Principles to Practice*, eds. J. Moorkens, S. Castilho, F. Gaspari, and S. Doherty (Berlin/Heidelberg: Springer)

Vaswani, A., Zhao, Y., Fossum, V., and Chiang, D. (2013). Decoding with large-scale neural language models improves translation. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing* (Seattle, Washington, USA), 1387–1392

Voigt, R. and Jurafsky, D. (2012). Towards a Literary Machine Translation: The Role of Referential Cohesion. In *NAACL-HLT Workshop on Computational Linguistics for Literature* (Montréal, Canada), 18–25

Wieling, M., Nerbonne, J., and Baayen, R. H. (2011). Quantitative social dialectology: Explaining linguistic variation geographically and socially. *PloS one* 6(9). 14pp.

TABLES

Document	***N*-gram overlap**			**TTR**	**Sentence length**
	2	3	4		
Warbreaker	0.86	0.67	0.41	0.15	12.54
Prologue	0.86	0.63	0.38		13.14
Chapter 1	0.87	0.66	0.41		13.81
Chapter 2	0.89	0.67	0.42		13.08
12 books	0.86±0.03	0.63±0.03	0.37±0.03	0.17±0.03	16.78±3.03

Table 1. N-gram overlap ($n = \{2, 3, 4\}$), TTR and sentence length for *Warbreaker* and the means and 95% confidence intervals of those measures for the 12 books previously translated by Toral and Way (2018).

Keystroke type	**Task Type**				
	ht	**mt1**	Δ%	**mt2**	Δ%
Total	1.94	1.76	-9%	1.49	-23%
Content	1.52	0.69	-55%	0.56	-63%
Navigation	0.18	0.59	228%	0.53	195%
Erase	0.23	0.47	105%	0.40	72%

Table 2. Average number of different types of keystrokes used to translate each source character in each translation condition. For conditions MT1 and MT2, the relative changes with respect to translation from scratch (HT) are shown alongside the absolute values.

FIGURE CAPTIONS

Predictor	**Temporal** (time)	**Technical** (keystrokes)	**Cognitive** (pauses) number	mean duration	ratio
Source length	↑***	↑***	↑***	↑***	↑*
Trial	↓***	↓*	↓·	↓**	-
Condition (MT1 vs HT)	↓***	↓**	↓***	↑***	↑**
Condition (MT2 vs HT)	↓***	↓**	↓***	↑***	↑*
Condition (MT2 vs MT1)	↓*	↓**	↓***	↑**	-
Length:MT1	-	-	-	-	-
Length:MT2	↑**	↑***	-	-	-

Table 3. Significance of predictors in the mixed models built for each effort dimension. Significance levels: - ($p > 0.1$), · ($p \leq 0.1$), * ($p \leq 0.05$), ** ($p \leq 0.01$), *** ($p \leq 0.001$). Direction: ↑ (significantly higher), ↓ (significantly lower). Two comparisons are carried out for level MT2 of the predictor condition (i.e. against levels HT and MT1), hence we correct these p-values with Holm-Bonferroni.

Figure 1. Snapshot from the translation environment showing the third task, in which the translator is to post-edit the translations produced by the NMT system for sentences 21–30 of the chapter.

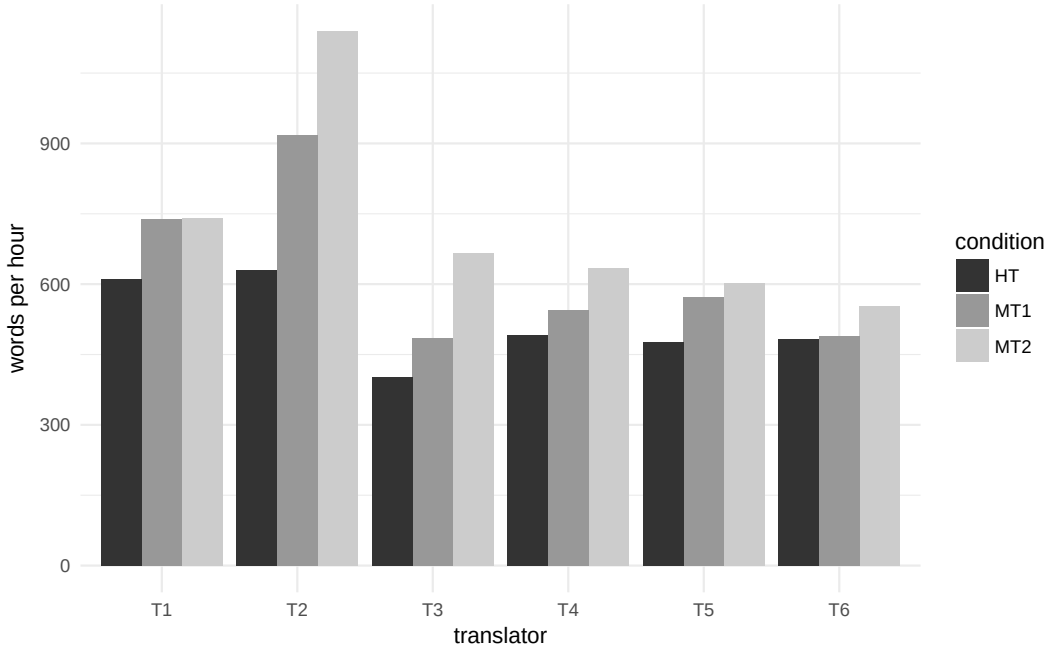

Figure 2. Translation productivity measured as words per hour for each of the translators in each of the translation conditions.

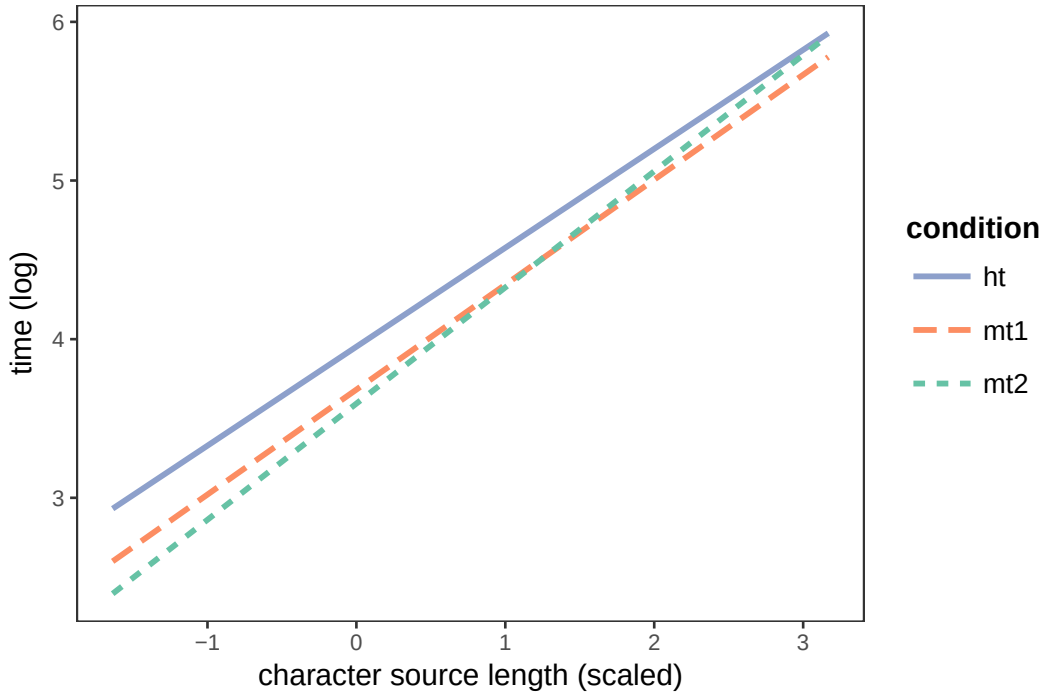

Figure 3. Interaction between the length of the source sentence and the translation condition on temporal effort.

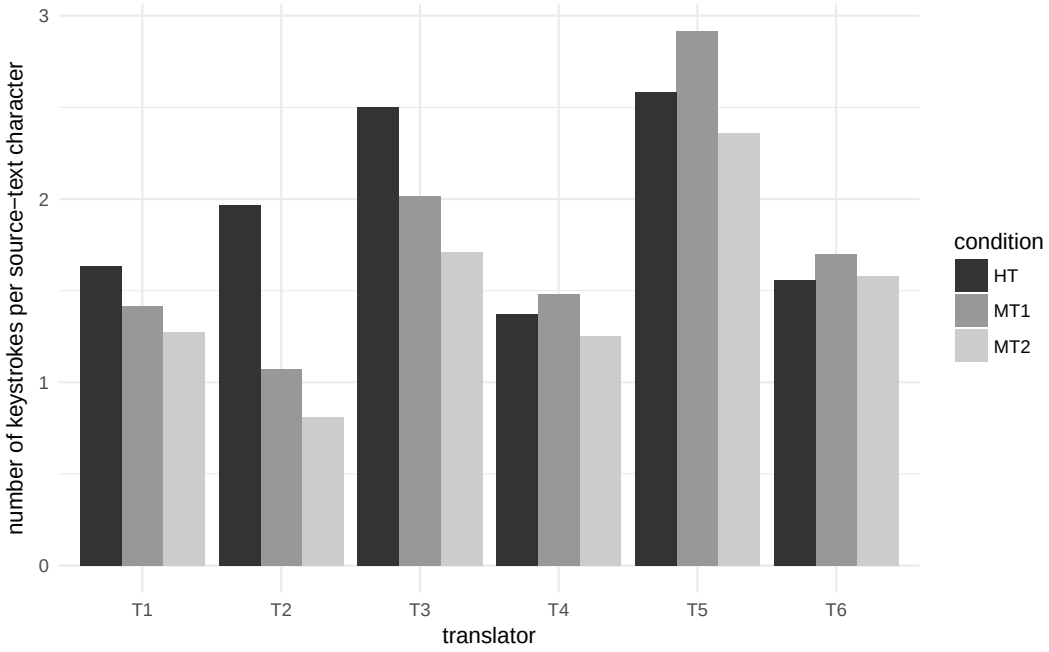

Figure 4. Technical effort measured as number of keystrokes per source character for each of the translators under each of the translation conditions.

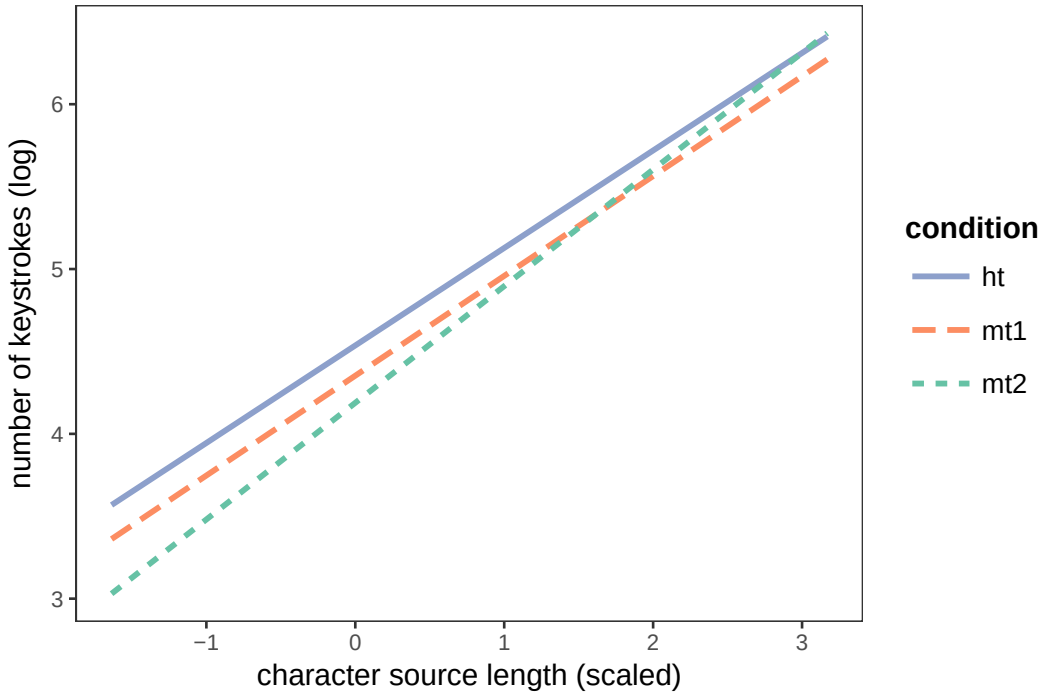

Figure 5. Interaction between the length of the source sentence and the translation condition on technical effort.

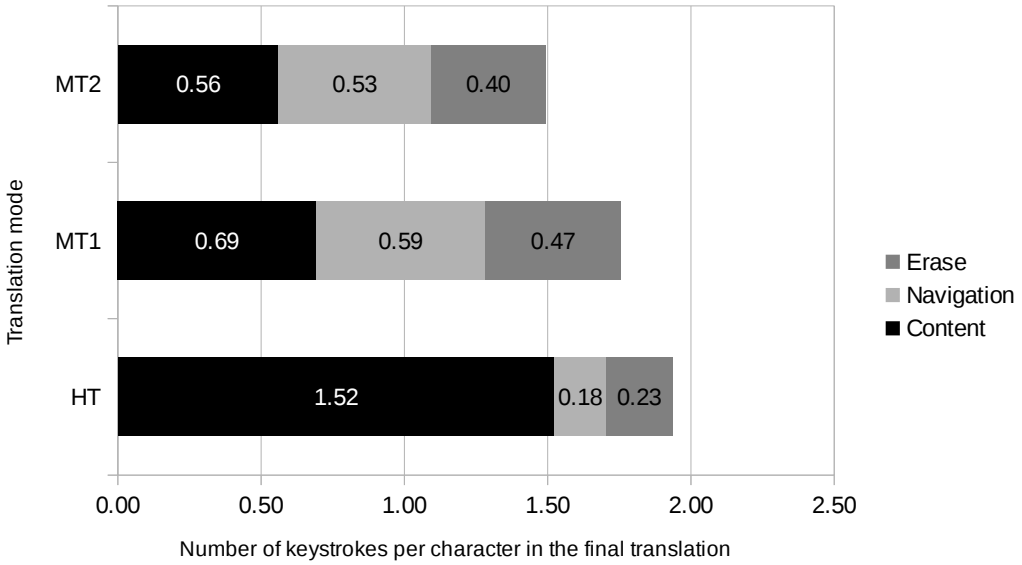

Figure 6. Proportion of each keystroke type (content, navigation and erase) in each translation condition (HT, MT1 and MT2) aggregating all the translators.

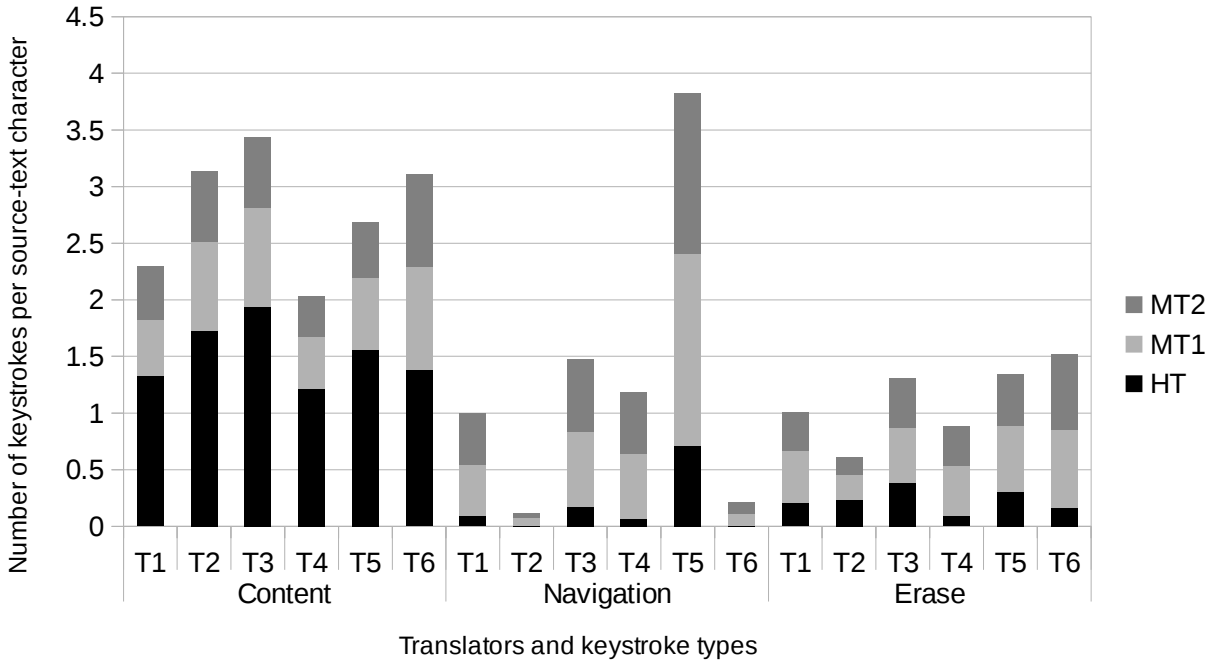

Figure 7. Proportion of each keystroke type (content, navigation and erase) in each translation condition (HT, MT1 and MT2) and for each translator.