Comparing Post-editing based on Four Editing Actions against Translating with an Auto-Complete Feature

Félix do Carmo
Centre for Translation Studies
University of Surrey
f.docarmo@surrey.ac.uk

Abstract

This article describes the results of a workshop in which 50 translators tested two experimental translation interfaces, as part of a project which aimed at studying the details of editing work. In this work, editing is defined as a selection of four actions: deleting, inserting, moving and replacing words. Four texts, machine-translated from English into European Portuguese, were post-edited in four different sessions in which each translator swapped between texts and two work modes. One of the work modes involved a typical auto-complete feature, and the other was based on the four actions. The participants answered surveys before, during and after the workshop. A descriptive analysis of the answers to the surveys and of the logs recorded during the experiments was performed. The four editing actions mode is shown to be more intrusive, but to allow for more planned decisions: although they take more time in this mode, translators hesitate less and make fewer edits. The article shows the usefulness of the approach for research on the editing task.

1 Introduction

1.1 Purpose

This article describes an experiment that is based on a theoretical framework in which editing is defined as being composed of four actions (delete, insert, move and replace). This framework also includes the definition of an editing threshold, which is a rate above which one may consider that the translator is no longer editing but translating the segment. The editing threshold was experimentally set at 25% for the project, as an inversion of the 75% fuzzy match initial band used in the translation industry. (do Carmo 2017)

The motivation for the experiment was to investigate how translators edited machine translation (MT) output, with and without consideration for the four editing actions.

At the beginning of the project, there was the expectation that this could contribute to the development of smart editing tools, which could learn patterns of editing based on these four actions, and then use this learned knowledge to support translators' editing work. If such systems employed features like Online Learning (Ortiz-Martínez et al. 2016) to record and reuse, for example, the word substitutions that are required, each edit could be made more efficiently. Such a system would show the translator good candidates for deletion, suggest words that might be missing from the MT output and indicate possible new positions for words being moved. These features are particularly useful in texts with high internal repetition, and when the output only requires minor editing.

After an analysis of the scope of the project, it was decided to focus on testing forms of interface for supporting editing work. The practical part was outlined as the comparison of an experimental interface based on the four editing actions, against an interface based on an auto-complete feature. This comparison of a novel interface against the main form of support offered by interactive translation tools to help translators while they edit (see below section 1.2) would create the opportunity to study in detail effects of different modes of work.

The specific research objectives for the experiment were two: (i) to collect opinions and effects of this description of editing, in a qualitative and quantitative study with professional translators and (ii) to compare two modes of editing, and measure the effects on editing practices of these two modes. The research questions explored during the workshop are presented in section 3.4, together with the variables that were measured.

1.2 Related research

In the localisation industry, professional translators not only translate texts from one language to another and revise translated texts, but they also regularly post-edit MT content. Several studies have highlighted the differences between the processes of translation, revision and post-editing (PE) (Alves and Vale 2011; Carl, Dragsted, and Jakobsen 2011; Carl et al. 2016).

PE is a demanding process, which requires that the translator pays a lot of attention to details, while using diverse resources and repeating very minute technical actions. Krings (2001) first described the three dimensions for which researchers must collect data to study the effort required by PE: temporal, technical and cognitive. While the most important dimension for the improvement of the processes is the cognitive effort, this is the hardest dimension to collect data from.

Despite a reasonable body of research, and research on the development of tools that integrate MT features to minimise translation effort (Green et al. 2014; Forcada and Sánchez-Martínez 2015), the computer aided-translation (CAT) tools used by professional translators are mostly the same for the three processes, and are oriented towards the use of translation memory (TM) features rather than for MT (Moorkens and O'Brien 2017).

In research on interactive translation, there are two main paradigms to feed MT content into translation tools: by presenting a full MT suggestion, which must then be edited by the translators, or by presenting suggestions to complete the translation, as it is typed. The first model is known as typical PE, and the second as Interactive Machine Translation (IMT) (Green et al. 2014; Sanchis-Trilles et al. 2014; Forcada and Sánchez-Martínez 2015; Ortiz-Martínez 2016).

Current interactive tools, like CasMaCat (Alabau et al. 2013) and Lilt (Green et al. 2014) offer a type of support which is based on IMT: translators type their translation from beginning to end and there is an auto-complete feature that suggests how to finish the word or sentence the translator is typing. (The distinction between PE and IMT concerns the features that form the interaction and interface with the user, at the segment level. Other adaptive features are not used for this distinction.) It is not clear yet whether this type of IMT feature will be fully adopted by translators, as it has been shown that it implies a big cognitive interference with the translation effort (Alabau et al. 2016).

Translation process research collects data from keylogs at the character-level, producing a detailed output that is hard to interpret. Levels of recursiveness, non-linearity in the writing actions, the fact that several edits may not be linked to specific words, cutting and pasting or moving words, all these factors increase the difficulty (Carl and Jakobsen 2009; Alves and Hurtado Albir 2010).

Extensive research has been done using translation edit rate (TER) (Snover et al. 2006) and other edit distances as identifiers of the editing operations that are required to transform one version of a text into another. Implementations of TER start by aligning words in source and target segments, and then estimate the least number of insertions, deletions and replacements that are necessary for the second version generation. Word movements are calculated afterwards. This means that TER is calculated from finished products, not during the process. As demonstrated in do Carmo (2017), TER is quite accurate at identifying insertions, deletions and replacements when only one word is edited in a segment, but this metric is not so accurate when editing involves more actions. Movements of words are very difficult to estimate accurately, especially at the end of a segment.

2 The experiment setting

This section describes the set-up of a three-hour workshop with 50 translators, which was performed in January 2017. This included a two-hour presentation and discussion on the conceptual framework of this project, followed by a one-hour hands-on experiment. This article summarises the experiment, which is described in detail in do Carmo (2017).

The setup of the experiment was exploratory. Our main purpose was to test a novel interface in as many scenarios as possible in a short time, and to collect translators' impressions on it. In this sense, this could be framed as an extensive pilot test, even though there were no plans for a proper test to the novel interface.

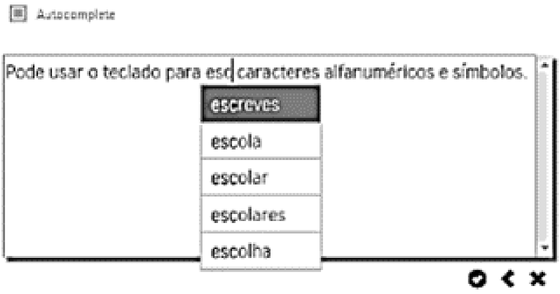
The use of edit distances to estimate operations performed by translators was one of the dimensions being tested during the experiments. As we wanted to collect data on the actual actions (the words that were deleted, inserted and replaced, besides those that were moved, together with a record of all the positions these edits were made on), we would need an interface that recorded those actions while they were being performed, not postprocess estimations as edit distances are. It was also clear that this new interface would ask for conscious decisions from translators on using only the four actions.

The tool developed during the study had usability issues, but it allowed us to explicitly ask translators to consider which words to delete, insert, move or replace, before performing any of those actions. This allowed us to gather information which would not be possible from an unconstrained work environment, in which translators could, for example, delete a whole MT suggestion, and write the whole translation, eventually retyping words already in the suggestion.

2.1 Translation tool and interfaces

The translation tool used in the experiment was HandyCAT (Hokamp and Liu 2015), in a specific implementation that Chris Hokamp agreed to create for this workshop.[2] This version included two modes of interaction:

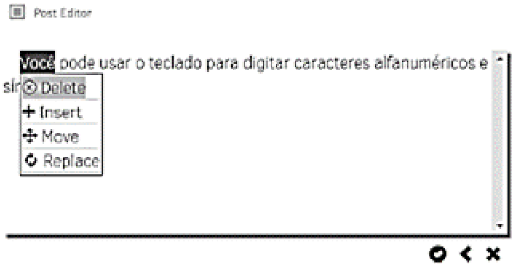
Auto-complete (AC mode): Figure 1 - this is an implementation of the technical principles of predictive writing, used in IMT. Users type their translations and get suggestions on how to complete each word.



Pode usar o teclado para esc caracteres alfanuméricos e símbolos.

Figure 1 – Auto-complete mode (AC mode).

Post-editor (PE mode): Figure 2 - this mode constrained the users to select a token and then choose from a contextual menu one of the four editing actions to apply to that token.



Você pode usar o teclado para digitar caracteres alfanuméricos e si

Figure 2 – Post-editor mode (PE mode).

2.2 Workshop and data collection

After a presentation and a discussion on the theoretical foundations of the project, translators received an explanation of the work sessions that they were going to perform and their purpose. In this explanation, it was mentioned that there would be no evaluations of the quality of the translations they produced, because we were only interested in collecting data on *how* they used the two interfaces.

During the hands-on sessions, translators edited four texts, from English into European Portuguese. Each text was extracted from a different technical document, aiming at a diversified experience from the participants. One of the texts was extracted from an electronic device instruction guide (text A), another from a marketing questionnaire (text B), the third text (C) was part of a product catalogue, and the last one (text D) was the initial paralegal text of a technical manual. The texts had been pre-translated using MateCAT (Federico et al. 2014).

Text A was used for familiarisation with both working modes (AC and PE), and there was no data collection from this stage. Then, users edited the MT outputs of the other 3 texts, in a random distribution, in four different sessions. In the first session, each translator edited one of the three available texts (B, C or D) in AC mode for 10 minutes. Next, they edited one of the other texts in PE mode for 15 minutes, so to allow for a richer experience with a method that was new for all users. After these two sessions, they performed two 5-minute sessions with the third text, first in AC mode and then in PE mode. All 50 translators edited each of the 3 texts, but not in the same mode, nor in the same sequence.

The length of each text varied, between 500 and 970 words, but there was no requirement to finish editing any of the texts, since all sessions were time-limited. Thus, we released each translator from concerns about speed.

We analysed data that corresponded to a total of 26 hours of work, during which 8565 segments were edited by the 50 translators.

Before, during, and after the workshop, participants filled in questionnaires, to identify their opinions on the conceptual structure of the project, and on the use of the two modes of interaction in a translation tool (the results of these 150 questionnaires are presented in section 3). Besides,

[2] This version of HandyCat is available from https://github.com/chrishokamp/handycat/releases/tag/porto_v0.1.

HandyCAT collected activity logs, which were also used in the data analysis (section 4). To collect edit scores, we used an add-on of SDL Trados Studio called Post-Edit Compare, or PEC (Hartnett and SDL Community 2014).

3 Data analysis

3.1 Characteristics of the participants

The selection of participants in the workshop was non-probabilistic and purposive (Trochim 2006), as the aim was to get specialised feedback. Most participants in the workshop were freelancers (58%) or they worked in translation agencies (36%). There was a fair distribution between experience ranges, of 1 year to 20 or more years. Most users were very comfortable with technology, and a fair majority (68%) preferred to type over the source text than to write the translation in an empty window. 90% of the users worked with suggestions from CAT tools, and most also used predictive writing features and/or support by MT systems. 75% of the users had some or a lot of experience in doing PE. Finally, a big proportion of users (66%) considered that the new technologies will not have a negative impact on the profession.

3.2 Receptivity to the proposal

Although the workshop was a pilot test on concepts that were still in development, most participants (80%) accepted that the concepts discussed in the workshop were useful for clarifying the tasks that they perform, and a large majority (90%) said at the end that the workshop had allowed them to rethink what they do when they post-edit.

3.3 Specialised answers

Data collected from activity logs confirmed most of the intuitive answers these specialised users gave in the questionnaires. For example, the texts that, after a quick reading, were classified as more complex and which showed a lower MT quality were the ones that later required the highest TER scores. Users also identified the mode in which they edited more (AC mode), and the actions they used the most (replace, followed by delete) – see section 4.5.

They classified AC mode as faster, easier to work with and more adjusted to translation work; they classified PE mode as slower, more intrusive, and more adapted to PE work, namely when only small changes were necessary. Another positive outcome was visible after the qualitative answers were codified and analysed. Translators considered, for example, that PE mode forced them to think and plan the changes that were required, and to focus on priority issues. They also said that their view on PE had changed, as an effect of the work they had done at the workshop.

3.4 Data variables in the activity logs

The input variables, the main questions related to them, and the sections in which these are presented are shown below:

- **Texts:** Did text features influence the editing practices? (Section 4.1)
- **Segments:** Did longer or shorter segments affect the results? (Section 4.2)
- **Users:** Is it possible to distinguish users according to numbers of edits and speed? (Section 4.3)
- **Modes:** Did work modes affect how translators edited? (Section 4.4)
- **Actions:** Which differences can be observed in the way users applied the different editing actions? (Section 4.5).

The main dependent variables in the study were:

- **Edit scores:** number of edits per segment and TER scores (number of edits divided by segment length).
- **Speed:** number of seconds per edit.

Section 4.6 focuses on the results obtained by looking for the most influential factors for the results shown by the two dependent variables.

4 Discussion of results

4.1 Text variable

With only 3 texts being used, and for a short time, it would not be possible to extract strong evidences that different texts were associated with different editing practices. However, interesting patterns started emerging at this level. The main two results at the level of texts were the edit scores, and the related editing threshold.

	AC	*PE*	*Total*
B - Questionnaire	17%	18%	**18%**
C - Catalogue	34%	29%	**31%**
D - Manual	29%	24%	**26%**
Total average	**26%**	**24%**	**25%**

Table 1 – Average TER per text and mode.

Text B (a marketing questionnaire) had the lowest TER score, Text C (a catalogue of office supplies) the highest TER score, and Text D (the initial instructions of a manual) presented an intermediate edit score. Across all texts, the average editing score is 25% (at our proposed editing threshold). This result might be read as a global indicator of the high quality of the MT output, since it only required that circa 25% of the words were edited. However, this hid a more complex situation, since text C in AC mode required editing to more than 33% of its words. The features of this text (a list of short descriptions of products) might justify this, but the study was not planned to collect enough contrastive data on text types to make such a claim.

4.2 Segment variable

Segments allowed for a more detailed analysis of the values collected at text level, but this presented its own challenges. The main research question at this level was whether longer or shorter segments had affected the results, but it was not possible to identify a meaningful correlation between segment length and editing scores or speed. This is related to the dependence of segment level to text type. In fact, the text with the longest segments was also the text with the shortest segments (B – Questionnaire). This was also the text in which more segments were edited, almost double than those for text C, the catalogue. Apart from this, the fact that users could spend a long time in a short segment and make only one edit, skewed the influence of those data points in all estimates related to segments.

At segment level, it was interesting to analyse edit scores, especially at the extremes (segments with zero edits and high number of edits in specific segments). Globally, 34% of the segments required editing above the 25% editing threshold, but in text C – Catalogue 53% of the segments were above the threshold. Figure 3 shows the spread of scores in all segments, with the percentage of edited segments per text and mode in the vertical axis, and the distribution in TER ranges in the horizontal axis.

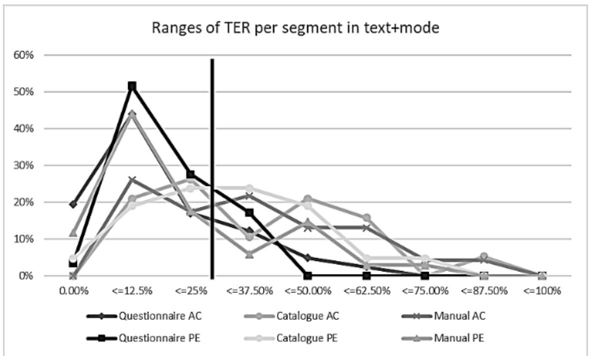


Figure 3 – Segment distribution in TER ranges.

Although most segments, in both modes and all texts, have edit scores below or equal to the 25% threshold (as described by the highest points in all curves at the second and third ranges at the left of the chart), there is a fair number of segments that show higher edit scores, up to 87.5%, as we move to the right of the chart. Catalogue shows the highest number of segments above the threshold, as is visible in the data points of the two curves (one for each mode) that describe this text, especially at and above the 50% editing range.

Next, we looked at extreme editing. We measured numbers of zero edited segments (see the first column of data points in Figure 3), but the results in the opposite extreme were more interesting. In all texts, there were five segments that presented editing scores above 50% on average, among all users. At this level of intense editing, there are not only long segments, as might be expected, if we consider the strong correlation between editing effort and segment length (Popovic et al. 2014). In fact, only two of these five segments have more than six words. This shows that short segments, like the ones one finds in software localisation, in the translation of lists of technical terminology, and in other types of length-restricted projects, may imply an editing effort which is not proportional to their size.

4.3 Users variable

The analysis of the different behaviours of users showed a few outliers, which were essentially users who had had technical problems and only reported a few of the work sessions. Besides this, there were users who had left segments open for a long time, and others who came back and reopened segments. This behaviour had not been anticipated, but it was possible to reclassify the data and get more accurate records of the number of times users opened and closed segments, with or without editing them, or to re-edit them. We could then see that this behaviour was more frequent in AC mode than in PE mode.

Users' editing behaviour tends to be more consistent in PE mode than in AC mode. For example, they tend to edit each segment in PE mode in a time range from 01:15 to 02:15, but in AC mode it ranges from 01:00 to 03:45. This wider variation of values in AC mode was visible in other scores.

Figure 5 shows the distribution of users in each mode according to TER scores, in ranges of 5%. The averages in PE mode are more concentrated in central cases, and they go up to users with an average of 46-50% edit scores. In AC mode, there are users with very small edit score averages and with very high edit scores. The number of users with average editing scores above the 25% threshold is the same for both modes (56%) and is higher than those below the threshold.

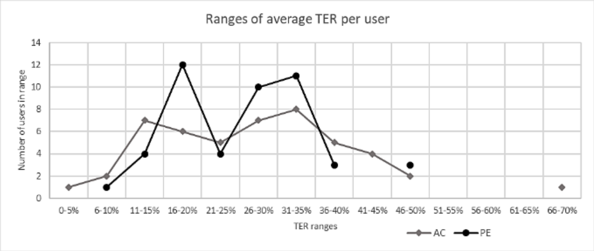


Figure 5 – Distribution of TER in both modes.

Another interesting result of the analyses based on users was a matrix of sorted results in terms of speed and TER, which showed four clear clusters of users, as described in Figure 6.

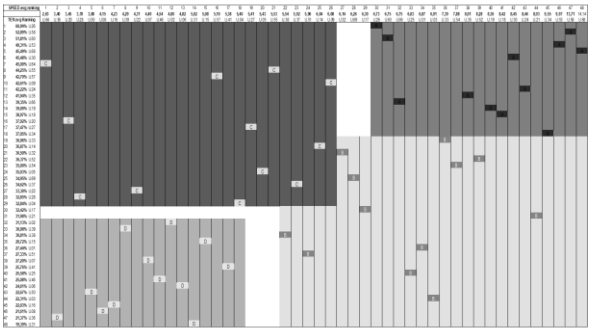


Figure 6 – Groups of users per TER and speed.

This figure was created by placing each user in a matrix which ranked them from the highest TER score to the lowest (vertical axis), and the highest speed to the lowest (horizontal axis). Separating lines around each area were then drawn, at the frontiers that divided each variable by the same number of users: the 12 faster users separated from the 12 slowest, and the 12 users who edited more separated from the 12 who edited less. These groups could be further explored, in terms of their characteristics, but none of the analysis employed showed a strong consistency. We found this clustering of users an interesting outcome of our experiment, but our position is that this should not be used in any form of profiling translators, since many other factors, not measured in our experiment, have to be taken into account when quality is discussed in professional uses.

4.4 Modes variable

The general perception that PE mode created more difficulties for the users is confirmed by the scores obtained. However, PE mode may have helped users achieve higher efficiency. For example, assuming that users are trying to produce a similar edited result in both modes, they tend to reopen and re-edit segments less often when they are in PE mode. They also make less edits (2.85 edits per segment in PE mode, against 3.21 in AC mode).

The total average TER scores are very similar for both modes, but PE mode has a higher score, slightly above the threshold (26% against 24% in AC mode). The fact that these global averages are so close to the suggested threshold does not recommend immediately that the threshold should be revised and repositioned. If more experiments identify a very frequent number of cases around this threshold (i.e. if most segments and texts require that around a quarter of the words are edited), we suggest that the threshold should become an object of study in itself. Arguments for studies focusing on the threshold, besides its statistical relevance, would include quality and effort, since, as Krings (2001) mentions, medium quality segments seem to be the ones that require the highest cognitive effort.

Mode is the variable in which considerations on speed are more relevant, related to users' efficiency. In fact, speed reveals a clearer separation between the two work modes than edit intensity.

In the four work sessions (two in each work mode), each edit took 16 seconds on average in PE mode, but only 10 seconds in AC mode. This difference can be attributed to the intrusiveness of the PE mode, but there are other factors to take into account, like the fact that AC is a method that some users already knew and used regularly, whereas none of the users had ever used PE mode. The two final sessions, in which the same text was edited in the two different modes (first AC and then PE), confirmed that users were still faster applying each edit in AC mode (10 seconds on average, against 12 seconds in PE mode).

The two sessions (second and fourth) with PE mode also show an interesting result: users improved their times when they applied the PE method for the second time. The average speed for each edit in the first session in PE mode was 19 seconds, but this improved to 12 seconds in the second session, when translators edit for the second time the same content they had edited in AC mode. This might simply arise from repetition, but it shows that the constraints of the PE mode can be overcome with that repetition.

We stress again that we make no claims on the usability of the method — this only shows that repetition (of method and text) may lead to a high increase in speed (37%), which may be considered relevant for a feature that had implementation issues. We should also note that questions on the quality and reasonability of the edits made by the translators at any of these sessions cannot be answered in this experimental setup and are not considered in these observations.

4.5 Actions variable

The last variable that was studied involved the four editing actions. For this variable, we chose a more descriptive research question: which differences can be observed in the way users applied the different editing actions?

First, we wanted to know how accurate the estimations of edit distances are, in relation to real actions performed by users. The analysis of this data requires a detailed description of the methods of collection and treatment of data by the tools that we used.

Every time a user selected an action (delete, insert, move or replace) in PE mode from HandyCAT's contextual menu, this event was recorded in a separate row in the log. So, we might have only one row or many rows describing each segment, according to the number of events for that segment: the edits each translator did in that segment. This data was then analysed in terms of TER edits measured by PEC, and a comparison was made between them. This confirmed the initial analysis of a disconnection between the users' actual actions and the description of TER.

Table 2 summarises these results. The first columns represent the events selected in HandyCAT; the four last columns, under 'Edits by PEC', summarise the total numbers of each type of edit that PEC estimated. In each row in the table, we show the number and types of edits PEC estimated for each type of action selected by the participants.

		Edits by PEC			
Actions	**No. events**	**Delete**	**Insert**	**Move**	**Replace**
PE.delete	1122	**1280**	0	0	273
PE.insert	570	22	**650**	1	174
PE.move	265	107	46	**184**	107
PE.replace	1470	118	342	1	**1197**
Total	**3427**	**1527**	**1038**	**186**	**1751**

Table 2–Relation between actions/events and TER.

If we look to the first row of table 2, we can see that users chose the 'delete' action (PE.delete) from the contextual menu in HandyCAT, 1122 times in total. When PEC compared the effect of each of these deletions in each segment, it identified 1280 deletions, more than those actually performed. Furthermore, it also identified in the same deletion events 273 substitutions. This illustrates the lack of reliability in descriptions of editing behaviour based on the estimations made by edit distances like TER.

Replace was the action most often well identified by PEC (of the 1470 times replace was selected, 1197 times a replacement was identified by PEC, an accuracy of 81%). Move was the action most often wrongly identified (the accuracy was 69%). In total, TER estimated 31% more actions than those chosen in HandyCAT (a total 4502 against 3427).

Then, we looked at the distribution of the editing actions chosen by the participants. Replace was the action most frequently chosen by participants (43% of the events), followed by delete (33%), then insert (17%), and finally move (only 7%). This is in line with the results obtained by Krings (2001) – in very different technical circumstances – and Snover et al (2006).

Part of the disconnection between the actions chosen in a menu and the estimates by PEC may be due to technical issues, such as the fact that edits to spaces and capitalisation are registered by HandyCAT, but not counted by PEC. There may also be errors in user's selection of actions, which we analysed carefully in our data. However, the differences in these numbers are mainly related to the way PEC estimates the edits, using an estimation optimised for efficiency, which is not necessarily the method used by translators.

We also investigated how edit distances considered contiguous edits: when the user applied the same action (e.g. delete) to two consecutive words, it could be expected that PEC would identify this as just one edit (one deletion of two words). However, edits are estimated per word, which means that the average number of edited words for most estimated actions is one. Movement is an exception to this: TER implementations, as the one PEC applies, maximise efficiency for movement by starting to estimate movements of phrases. This is the reason why there was an exceptional average of 1.68 words being edited in each row where one move action was selected.

The values for speed of each editing action are easy to understand. Delete is the fastest action to apply (an average 8 seconds), then move (11 seconds), while insert and replace take, with a slight rounding, the same time (14 seconds). The delay in both these actions is associated with the time required to type the new words. Choosing move is faster than deleting and inserting a word.

4.6 Factors for editing scores and speed

After all the results were compiled, we did an analysis per user, combining answers from questionnaires and activity logs, by applying different statistical tests and tools in R (R Development Core Team 2008). The objective was to identify the factors that most contributed to the values obtained in each dependent variable (edit scores and speed).

One of the tools that presented the clearest results was a decision tree approach, a tool which, although not having a very strong explanatory power, shows interesting trends in the data. We fed a total of 27 factors to the decision tree, some of which were codified qualitative answers to the questionnaires (like experience with PE, or opinion on the impact of the workshop) and others performance indicators (like the number of segments with zero edits, ranking according to speed, average and total edit scores and speeds, etc.). Then, we estimated which of these factors had a stronger relation with the distribution of results for TER and speed.

For TER scores, this tool showed that the ‘text’ was the most influential factor, as we can see in Figure 6. The second factor, some way behind it, was ‘experience with PE’. However, the influence of text was so determining that if we removed the data for ‘text’ from the list of factors, the program could not calculate and produce a decision tree.

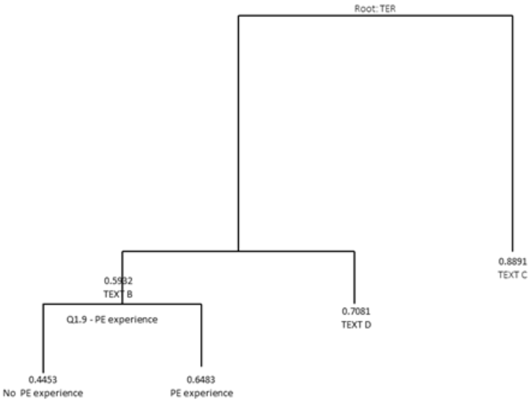


Figure 6 – Decision tree for TER scores.

The results for speed were different. Speed seems to be a measure that depends more on the work method used. That was an impression collected from users, and it was confirmed in the global analysis of the data collected. The second most influential factor was the previous use or not of predictive writing features by translators, and finally their experience with MT or PE. The text, for example, does not appear in the sequence of influencing factors for speed of editing, as shown in Figure 7.

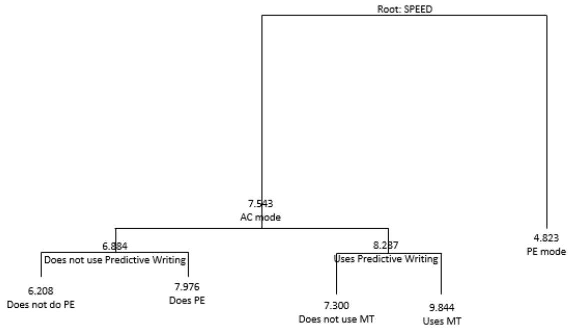


Figure 7 – Decision tree for speed factors.

5 Discussion and conclusions

Following the aim of the project, we could confirm in the workshop that there is a research interest, from different points of view, in the approach of editing as being formed by four actions.

First, it was confirmed by the participants in the surveys conducted during the workshop. Although it was related to a disruption in work habits, translators admitted that this perspective made sense as a description of what they do when they edit MT text, at least from a technical perspective. As commented in the final surveys, the main advantage of this method, which is related to its intrusiveness, is the demand for translators to plan their editing strategy before applying it, thus bringing to light an often-unconscious decision process, forcing them to make more efficient actions. This efficiency, measured only as technical action data, is, naturally, meaningless if we aim at relating it to effectiveness, but it is relevant for the development of better forms of support in translation tools.

As admitted from the outset, the four editing actions interface tested during the workshop was not appropriate for commercial settings, because of its lack of usability development. There are, however, potential research and pedagogical implications in this mode of work. The work mode based on the four editing actions gives visibility to the decision process in time-constrained and efficiency demanding work contexts, as is the case of PE.

This study, namely in its analysis of the lack of correspondence between the estimations of edit distances and the actions employed by translators, also calls the attention to the complexity of editing work. This calls for a cautious use of metrics like TER as descriptors of processes, and to the need to study in closer detail the different levels of complexity in PE. Our suggestion of an editing threshold that sets a boundary between levels in which editing can be easily described and others in which it cannot is a contribution to this type of study.

The methodology applied in this study was adjusted to the objectives and constraints of the experiment. The amount of questions this exploratory approach raised meant that a choice had to be made between extensiveness or depth of statistical analysis. This approach enabled a rich discussion with the participants, on the two work modes and around concepts and practices that were novel, and to collect data for the comparison between the two modes, thus fulfilling the experiment objectives.

The analysis of the results of our experiment exploited the weight of different dimensions of PE. The suggested editing threshold of 25% was the most frequent global average in the different levels of analysis, but relevant differences in editing intensity in different texts, by different users and using different work modes were also observed. We also saw how the number of edits and speed may be associated with different factors, text and work mode, respectively, but also how these two dependent variables helped us group users in four different clusters. This shows how the threshold may act as an important instrument to highlight varied degrees of complex editing in MT content.

The experiment also showed that users were more consistent, and, in a certain sense, more efficient in PE mode, since they made fewer edits and returned fewer times to the same segment. This efficiency is contradicted by the fact that PE mode requires more time per edit, but we also observed improvements in speed of use of this work method. This highlights the usefulness of making the decision process more conscious, something which may be explored in pedagogic contexts.

Finally, this work showed how the four editing actions are useful features to describe and research editing work. The most frequent action employed by translators was replace, followed by delete, then insert and finally move, a result confirmed in similar experiments. Insert and replace take longer to perform because they require typing. A tool that aims at supporting the actions performed by translators needs to act on all of these actions. This presents implementation challenges, not only at the data collection stage, but also regarding the usefulness and usability of the interfaces. For example, actions that require typing (like insert and replace) may be supported by auto-complete features, but delete and move only have an added value if they can appear previously as suggestions to the users, highlighting words which may have to be deleted and positions where words may be moved into. So, a combination of methods, flexibly adjusted to different degrees of editing, seems to be the most reasonable approach.

Although it focused on micro actions, this project allowed us to learn many things about the editing task. It has raised relevant questions related to technical effort, productivity, usability and usefulness of action support features, and on how to support specialised users. We suggest that more research around the editing actions and the editing threshold, besides new applications to support editing actions, should be pursued by translation technology researchers and developers.

Acknowledgements

This Project has received funding from the European Union's Horizon 2020 research and innovation programme under the EDGE COFUND Marie Skłodowska-Curie Grant Agreement no. 713567. This publication has emanated from research supported in part by a research grant from Science Foundation Ireland (SFI) under Grant Number 13/RC/2077.

References

Alabau, Vicent, Ragnar Bonk, Christian Buck, Michael Carl, Francisco Casacuberta, Mercedes García-Martínez, Jesús González, et al. 2013. "CASMACAT: An Open Source Workbench for Advanced Computer Aided Translation." *The Prague Bulletin of Mathematical Linguistics* 100 (100): 101–12. https://doi.org/10.2478/pralin-2013-0016.

Alabau, Vicent, Michael Carl, Francisco Casacuberta, Mercedes García Martínez, Jesús González-Rubio, Bartolomé Mesa-Lao, Daniel Ortiz-Martínez, Moritz Schaeffer, and Germán Sanchis-Trilles. 2016. "Learning Advanced Post-Editing." In *New Directions in Empirical Translation Process Research*, edited by Michael Carl, Srinivas Bangalore, and Moritz Schaeffer, 95–110. Heidelberg: Springer International Publishing Switzerland. https://doi.org/10.1007/978-3-319-20358-4_5.

Alves, Fábio, and Amparo Hurtado Albir. 2010. "Cognitive Approaches." In *Handbook of Translation Studies - Volume I*, edited by Yves Gambier and L. van Doorslaer. Amsterdam/Philadelphia: John Benjamins Publishing Company.

Alves, Fábio, and Daniel Couto Vale. 2011. "On Drafting and Revision in Translation: A Corpus Linguistics Oriented Analysis of Translation Process Data." *Translation: Computation, Corpora, Cognition* 1 (1): 105–22. http://www.mt-archive.info/TC3-2011-Alves.pdf.

Carl, Michael, Barbara Dragsted, and Arnt Lykke Jakobsen. 2011. "On the Systematicity of

Human Translation Processes." In *Tralogy 2011. Translation Careers and Technologies: Convergence Points for the Future*. Paris, France. http://research.cbs.dk/en/publications/on-the-systematicity-of-human-translation-processes(0225f368-7eea-4ed2-8a35-9a7d9efbce6c).html.

Carl, Michael, and Arnt Lykke Jakobsen. 2009. "Towards Statistical Modelling of Translators' Activity Data." *International Journal of Speech Technology* 12 (4): 125–38. https://doi.org/10.1007/s10772-009-9044-6.

Carl, Michael, Isabel Lacruz, Masaru Yamada, and Akiko Aizawa. 2016. "Measuring the Translation Process." In *The 22nd Annual Meeting of the Association for Natural Language Processing, NLP 2016*, 0–3. Japan. http://openarchive.cbs.dk/bitstream/handle/10398/9280/Michael Cral_2016_02.pdf.

do Carmo, Félix. 2017. "Post-Editing: A Theoretical and Practical Challenge for Translation Studies and Machine Learning." Universidade do Porto. https://repositorio-aberto.up.pt/handle/10216/107518.

Federico, Marcello, Nicola Bertoldi, Mauro Cettolo, Matteo Negri, Marco Turchi, Marco Trombetti, Alessandro Cattelan, et al. 2014. "The Matecat Tool." *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: System Demonstrations* 7 (287688): 129–32. http://aclweb.org/anthology/C14-2028.

Forcada, Mikel L, and Felipe Sánchez-Martínez. 2015. "A General Framework for Minimizing Translation Effort: Towards a Principled Combination of Translation Technologies in Computer-Aided Translation." *Proceedings of the 18th Annual Conference of the European Association for Machine Translation*, 27–34. http://aclweb.org/anthology/W15-4904.

Green, Spence, Jason Chuang, Jeffrey Heer, and Christopher D Manning. 2014. "Predictive Translation Memory: A Mixed-Initiative System for Human Language Translation." In *Proceedings of the 27th Annual ACM Symposium on User Interface Software and Technology (UIST '14)*. New York: ACM. https://doi.org/http://dx.doi.org/10.1145/2642918.2647408.

Hartnett, Patrick, and SDL Community. 2014. "Post-Edit Compare." 2014. http://posteditcompare.wiki-site.com/index.php/Main_Page.

Hokamp, Chris, and Qun Liu. 2015. "Handycat: The Flexible CAT Tool for Translation Research." In *EAMT 2015*. Istanbul.

Krings, Hans P. 2001. *Repairing Texts: Empirical Investigations of Machine Translation Post-Editing Processes*. Edited by Geoffrey S. Koby. Kent, Ohio & London: The Kent State University Press.

Moorkens, Joss, and Sharon O'Brien. 2017. "Assessing User Interface Needs of Post-Editors of Machine Translation Interfaces for Editing Human Translation and Post-Editing Machine." In *Human Issues in Translation Technology*, edited by Dorothy Kenny and Jenny Williams. London: Routledge.

Ortiz-Martínez, Daniel. 2016. "Online Learning for Statistical Machine Translation." *Computational Linguistics* 42 (1 (March 2016)): 121–61. https://doi.org/http://dx.doi.org/10.1162/COLI_a_00244.

Ortiz-Martínez, Daniel, Jesús González-Rubio, Vicent Alabau, German Sanchis-Trilles, and Francisco Casacuberta. 2016. "Integrating Online and Active Learning in a Computer-Assisted Translation Workbench." In *New Directions in Empirical Translation Process Research*, edited by Michael Carl, Srinivas Bangalore, and Moritz Schaeffer, 57–76. Heidelberg: Springer International Publishing Switzerland. https://doi.org/10.1007/978-3-319-20358-4-3.

Popovic, Maja, Arle Richard Lommel, Aljoscha Burchardt, Eleftherios Avramidis, and Hans Uszkoreit. 2014. "Relations between Different Types of Post-Editing Operations, Cognitive Effort and Temporal Effort." In *The Seventeenth Annual Conference of the European Association for Machine Translation (EAMT 14)*, 191–98.

R Development Core Team. 2008. "R: A Language and Environment for Statistical Computing." Vienna, Austria. http://www.r-project.org.

Sanchis-Trilles, Germán, Vicent Alabau, Christian Buck, Michael Carl, Francisco Casacuberta, Mercedes García-Martínez, Ulrich Germann, et al. 2014. "Interactive Translation Prediction versus Conventional Post-Editing in Practice: A Study with the CasMaCat Workbench." *Machine Translation* 28 (3–4): 217–35. https://doi.org/10.1007/s10590-014-9157-9.

Snover, Matthew, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. "A Study of Translation Edit Rate with Targeted Human Annotation." *Proceedings of AMTA 2006*, no. August: 223–31. https://doi.org/10.1.1.129.4369.

Trochim, Joachim. 2006. "The Research Methods Knowledge Base." 2006. http://www.socialresearchmethods.net/kb/index.php.