Noname manuscript No.
(will be inserted by the editor)

Interactive Translation Prediction vs. Conventional Post-editing in Practice: A Study with the CasMaCat Workbench

Germán Sanchis-Trilles · Vicent Alabau · Christian Buck · Michael Carl · Francisco Casacuberta · Mercedes García-Martínez · Ulrich Germann · Jesús González-Rubio · Robin L. Hill · Philipp Koehn · Luis A. Leiva · Bartolomé Mesa-Lao · Daniel Ortiz-Martínez · Herve Saint-Amand · Chara Tsoukala

Received: date / Accepted: date

Abstract We conducted a field trial in computer-assisted professional translation to compare Interactive Translation Prediction (ITP) against conventional post-editing (PE) of machine translation (MT) output. In contrast to the conventional PE set-up, where an MT system first produces a static translation hypothesis that is then edited by a professional translator (hence "post-editing"), ITP constantly updates the translation hypothesis in real time in response to user edits. Our study involved nine professional translators and four reviewers working with the web-based CasMaCat workbench. Various new interactive features aiming to assist the post-editor were also tested in this trial. Our results show that even with little training, ITP can be as productive as conventional PE in terms of the total time required to produce the final translation. Moreover, in the ITP setting translators require fewer key strokes to arrive at the final version of their translation.

Keywords CAT, SMT, interactive translation prediction, post-editing, field trial, user studies

G. Sanchis-Trilles · V. Alabau · F. Casacuberta · J. González-Rubio
L. A. Leiva · D. Ortiz-Martínez

Pattern Recognition and Human Language Technologies Center
Universitat Politècnica de València
Valencia, Spain
Tel.: +34 963 878 172
E-mail: gsanchis@dsic.upv.es

C. Buck and U. Germann · R. Hill · P. Koehn · H. Saint-Amand · C. Tsoukala

School of Informatics
University of Edinburgh
Edinburgh, United Kingdom

M. Carl · M. García-Martínez · B. Mesa-Lao

Department of International Business Communication
Copenhaguen Business School
Copenhaguen, Denmark

1 Introduction

Contemporary professional translators rarely produce translations entirely from scratch. Instead, they increasingly rely on Translation Memories (TM), that is, data bases of texts that have already been translated, and their translations. At translation time, translations of text fragments similar to the actual source text are retrieved from the data base and edited by the translator to bridge the mismatch between retrieved text fragments and an actual correct translation of the current source text. As the quality of the raw output of fully automatic machine translation (MT) systems is on the rise, so is the commercial interest in integrating MT as an alternative or supplement to traditional TMs into the professional translation workflow. Recent studies (Koehn 2009a; Plitt and Masselot 2010; Federico et al 2012; Flournoy and Duran 2009; Green et al 2013) have concluded that post-editing is, on average, more efficient than translating from scratch. However, the optimal form of human-machine interaction in the context of translation is still an open research question.

The open-source project CasMaCat addresses two needs in this area: first, it provides a new post-editing workbench for professional translators that is unobtrusive, yet provides support to the translator when it is relevant to do so; and second, it is able to log user activity in detail and thus record research data that can shed light on the mental processes underlying human translation in a computer-assisted translation (CAT) setting.

CasMaCat builds on the open-source, web-based post-editing tool MateCat[1] and adds several major capabilities to the framework:

1. It offers *interactive translation prediction* (Barrachina et al 2009) (ITP) as an alternative to classical post-editing. The ITP functionality used in this study has been implemented by means of the Thot toolkit for statistical MT (Ortiz-Martínez and Casacuberta 2014). Various auxiliary features and customizations have been implemented to help tailor the MateCat tool to the individual translator's preferences. They are described in Section 2.
2. CasMaCat can log user activity in detail and with precise timing information: key strokes, mouse activity, and translator's gaze (if used in combination with an eye tracker). Without eye tracking, the tool can be easily deployed in a web browser, eliminating the need for specialized hardware or software to run experiments. The logs from the user study discussed in this paper are available online for further analysis at `http://bridge.cbs.dk/platform/?q=CRITT_TPR-db`.
3. CasMaCat can be used with an e-pen as an alternative input device (Alabau et al 2014). There are a number of situations where such an interface is comfortable and effective. First, it is suited for post-editing sentences with only few errors, as it is often the case for sentences with strong fuzzy matches in translation memories, or during revision of human post-edited sentences. Second, it allows to perform such tasks while commuting, travelling or away from the desk for other reasons. The epen interface is also able to recognize gestures for interactive text editing, using a highly accurate, high-performance gesture recognizer (Leiva et al 2013).

[1] `www.matecat.com`

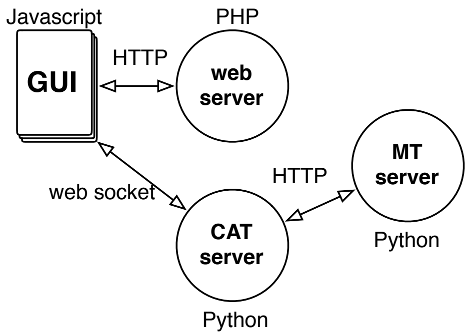

Figure 1 Components of the workbench

In the following, we present the results of a focused controlled user study of the CasMaCat workbench with professional translators that addressed the following questions:

- Does interactive translation prediction (ITP) boost or hinder overall translation productivity, especially when compared to conventional post-editing?
- What effect do different ITP visualization options have on the interactive translation process?
- How satisfied are users with regard to the produced translations?

2 The CasMaCat workbench

The CasMaCat workbench consists of several components (Figure 1).

1. a graphical user interface (GUI) implemented as a web browser plugin in JavaScript;
2. a web server backend implemented in PHP that retrieves translation jobs from a MySQL database;
3. a CAT server that manages interactive translation prediction and event logging during an edit session; and
4. an MT server that provides raw translations as well as the underlying search graphs (compact representations of all translation options considered) to the CAT server.

The latter two components are implemented in Python but interface and interact with additional third-party components written in a variety of programming languages.

The browser-based GUI and the CAT server communicate via web sockets for speed; the other communication pathways are handled over HTTP for maximum compatibility with other software components. For example, the communication between CAT server and MT server relies on an extension of the Google Translate API, so that other MT engines compliant with the Google Translate API can easily be swapped in if desired. The web back-end accepts translation job uploads and offers file downloads in standard XML Localization Interchange File Format (XLIFF).

Figure 2 Screenshot of CasMaCat with optional visualization features disabled

The CasMaCat workbench offers numerous user customization options. In its most basic form (Figure 2), the tool is reminiscent of standard CAT tools. The source text is partitioned into a series of translation segments (typically individual sentences), with the source text shown on the left and an edit window on the right that allows editing of the translation of the "current" segment.

This basic interface is augmented by additional functionalities and display customization options:

– **Intelligent autocompletion**: This is the fundamental interactive prediction feature of the CasMaCat workbench. Every time a keystroke is detected, the system produces a translation prediction for the entire sentence in accordance with the text that the user is writing or editing. The text to the left of the cursor is assumed to be approved by the human translator and serves as a prefix to identify the highest-scoring automatic translation that overlaps in this prefix. The remainder of the current transation prediction (to the right of the cursor) is then replaced with the updated prediction. The basic ITP feature is always enabled in ITP mode. ITP mode can be engaged by pressing the button labeled ITP and disengaged by pressing the button labeled PE (post-editing) below the text edit box. Post-editing mode and ITP mode are mutually exclusive.
– **Prediction rejection**: The current CasMaCat prototype also allows the translator to scroll through translation options by use of the mouse wheel (Sanchis-Trilles et al 2008). When the mouse wheel is turned over a word, the system invalidates the current prediction and provides the user with an alternate translation option in which the first new word is different from the one at the current mouse position. This option is one of the advanced ITP features.
– **Search and replace** (even in future predictions): the workbench extends standard search-and-replace functionality to future translation predictions. Whenever a new replacement rule is created, it is automatically propagated to the forthcoming predictions made by the system, so that the user only needs to specify them once. This specific function was implemented in response to user feedback in the first field trial of the tool. Note that this option implements a collection of replace rules, but does not resort to a fully-fledged SMT system for doing so as in (Simard and Foster 2013).

The user can also choose a number of advanced visualization options (Figure 3):

– **Visualization of MT system confidence**. Automatic estimation of the reliability of the MT system output, also known as *confidence estimation* for MT,

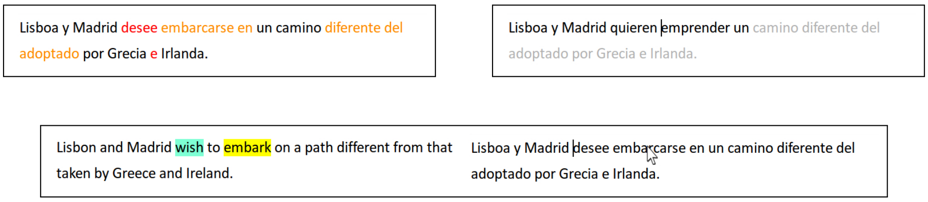

Figure 3 Advanced display options in CasMaCat: color-coding of confidence estimates (top left), limited prediction horizon (top right), and word alignment visualization.

is currently an active area of research. The CasMaCat workbench is able to visually mark up such confidence estimates in the prediction. MT output identified as probably incorrect is marked in red while MT output of questionable reliability in orange.

- **A limited prediction horizon**. Providing the user with a new prediction whenever a key is pressed has been shown to be cognitively demanding (Alabau et al 2012). In the current prototype, when this option is active, predictions are shown only up to the first word of low confidence according to the confidence estimates associated with the prediction. Pressing the TAB key allows the user to ask the system for the next set of predicted words, displaying the remaining words in the suggested translation in grey.
- **Word alignment information**. Alignment of source and target information is an important part of the translation process (Brown et al 1993). In order to display the correspondences between both the source and target words, this feature was implemented so that every time the user places the mouse (yellow) or the text cursor (cyan) on a word, the alignments made by the system are highlighted. The user can enable this visualization option by activating *displayCaretAlign* for the alignments with the cursor and *displayMouseAlign* for the alignments with the mouse.
- **Visualization of user edits** (not shown in Figure 3). This visualization option comes in three variants, all of them implemented with the purpose of helping the user locate which changes were introduced by him, or what was produced by the system without interaction.
  - *changed words only*: the system highlights in green the words that the user has modified.
  - *entire prefix*: the system highlights the prefix, i.e. the first part of the segment that the user has validated.
  - *last edit only*: the system highlights the last word that the user has modified.

3 Translation process data

Another important feature of the CasMaCat workbench is its ability to record user activity in fine detail for analysing human and computer-assisted translation processes scientifically. That is, the tool not only stores translation product information (the source, raw MT output and final translation), but can also provide detailed translation process data with precise timing information, including eye

Table 1 Translation process data logged and stored by CasMaCat. For further details about these features, see (Carl 2012b), (Carl 2014) and (Carl 2011).

Keystrokes (KD) : basic text modification operations (insertions or deletions), together with time of stroke, and the word in the final text to which the keystroke contributes.
Fixations (FD) : basic gaze data of text fixations on the source or target text, defined by the starting time, end time and duration of fixation, as well as the offset of the fixated character and word in the source or target window.
Production units (PU) : coherent sequence of typing, defined by starting time, end time and duration, percentage of parallel reading activity during unit production, duration of production pause before typing onset, as well as number of insertions and deletions.
Fixation units (FU) : coherent sequences of reading activity, including two or more subsequent fixations, characterized by starting time, end time and duration, as well as scan path indexes to the fixated words.
Activity Units (CU) : exhaustive segmentation of the session recordings into activities of typing, reading of the source or reading of the target text.
Source tokens (ST) : as produced by a tokenizer, together with TT correspondence, number, and time of keystrokes (insertions and deletions) to produce the translation and micro unit information (see below).
Target tokens (TT) : as produced by a tokenizer, together with ST correspondence, number, and time of keystrokes (insertions and deletions) to produce the token, micro unit information, amount of parallel reading activity during.
Alignment units (AU) : transitive closure of ST-TT token correspondences, together with the number of keystrokes (insertions and deletions) needed to produce the translation, micro unit information, amount of parallel reading activity during AU production, etc.
Segments (SG) : aligned sequences of source and target text segments, including duration of segment production, number of insertions and deletions, number and duration of fixations, etc.
Session (SS) : is a table which describes some properties of the sessions, such as source and target languages, total duration of session, beginning and end of drafting, etc.

tracking data if used in combination with an eye tracker.[2] A gaze-to-word mapping algorithm runs in real time, and maps gaze samples and fixation points to the nearest letter on the screen; the character offset is then logged together with the gaze data. The tool also keeps a record of the different translation options that were presented to the post-editor at the time. At storage time, CasMaCat aggregates and stores information about phases of coherent writing (production units; PU) and reading (fixation units; FU) from the raw user activity data (UAD). Table 1 summarizes the information stored during interactive translation and post-editing sessions. During analysis, we derived further aggregate information from the stored UAD. These derived measures are described in Section 5.

4 Field trial

In June 2013, we conducted the second CasMaCat field trial (CFT2) in cooperation with Celer Soluciones SL, a language service provider (LSP) based in Madrid. This trial involved nine freelance translators and four reviewers, all native speakers of Spanish offering translation and post-editing services on a regular basis for this LSP. Detailed information about participants' age (46 years old on average),

[2] In our experiments, we use an EyeLink1000 eye tracker.

Table 2 Task assignments in the field trial

	text								
	dataset 1			dataset 2			dataset 3		
	T1.1	T1.2	T1.3	T2.1	T2.2	T2.3	T3.1	T3.2	T3.3
segments	49	30	45	63	55	51	59	61	47
source words	952	861	1121	1182	1216	1056	1396	1427	1258
Part. 1	PE	ITP	AITP	ITP	AITP	PE	AITP	PE	ITP
Part. 2	AITP	PE	ITP	PE	ITP	AITP	ITP	AITP	AITP
Part. 3	ITP	AITP	PE	AITP	PE	ITP	PE	ITP	PE
Part. 4	ITP	AITP	PE	AITP	PE	ITP	PE	ITP	AITP
Part. 5	PE	ITP	AITP	ITP	AITP	PE	AITP	PE	ITP
Part. 6	AITP	PE	ITP	PE	ITP	AITP	ITP	AITP	PE
Part. 7	AITP	PE	ITP	PE	ITP	AITP	ITP	AITP	PE
Part. 8	ITP	AITP	PE	AITP	PE	ITP	PE	ITP	AITP
Part. 9	PE	ITP	AITP	ITP	AITP	PE	AITP	PE	ITP

years of experience in translation (23 years on average), education (Translation, Philology or another degree), etc., can be found in the CRITT Translation Process Research (TPR)-database under the metadata folder.[3]

The text type involved in this trial was general news from the WMT-2012 *news-commentary* corpus (Callison-Burch et al 2012). They consisted of approximately 1,000 words, distributed in 30 to 63 segments, as shown in Table 2. Each English source text was automatically translated into Spanish by a statistical MT system and then automatically loaded into the CasMaCat workbench for the participants to post-edit.

In an attempt to unify post-editing criteria among participants, all of them were instructed to follow the same post-editing guidelines aiming at a final high-quality target text (publishable quality). The post-editing guidelines distributed in hard copy were: i) Retain as much raw MT as possible; ii) Do not introduce stylistic changes; iii) Make corrections only where absolutely necessary, i.e. correct words and phrases that are clearly wrong, inadequate or ambiguous according to Spanish grammar; iv) Make sure there are no mistranslations with regard to the English source text; v) Publishable quality is expected. The work done by the four reviewers aimed at proofreading the final publishable quality of the translations produced by the post-editors.

4.1 Experimental design

Three system setups were evaluated in CFT2: conventional post-editing (PE), basic interactive translation prediction (ITP), and interactive translation prediction with advanced features (AITP). In each of the three conditions, the same set of nine different texts (approx. 1,000 words each), divided into three sets of three texts each, was translated three times by three different translators under each of the three conditions. Table 2 gives an overview of the task assignments. In each instance, keyboard and mouse activity was logged. Dataset 1 was processed under laboratory conditions recording additional eye-tracking activity from Celer Solu-

[3] These data are available on-line: CRITT Translation Process Research (TPR)database. URL: http://bridge.cbs.dk/platform/?q=CRITT_TPR-db

ciones SL. Datasets 2 and 3 were delivered over the Internet and processed at home by the nine post-editors.

For the conventional post-editing setup (PE) the highest-scoring translation hypothesis was used; ITP and AITP relied on a translation search graph delivered by the MT system. In the AITP condition, study participants were given access to all the advanced ITP features described in Section 2 and could freely choose which ones to enable and use.

The final translations of dataset 1 were subsequently proofread at Celer Soluciones SL, where each of the reviewers was assigned to review the work done by a maximum of three post-editors. Gaze and keyboard activity for reviewers was also logged.

Before starting their tasks, participants were introduced to the CasMaCat workbench and the three different conditions under consideration during the trial. They were given time to familiarise themselves with the tool and try out the different visualization options, and to decide which options they would enable when post-editing using AITP. After each session, participants were asked to complete an online questionnaire (see section 5.3). When all sessions at Celer Soluciones SL were completed, an additional in-depth interview was conducted with each of the translators. Table 3 summarizes the data collected during the trial.

5 System evaluation and results

User performance and evaluation is a central part of the CasMaCat project, and a rich dataset for analysis was collected during the field trial. This section provides several kinds of evaluation:

- Section 5.1 looks at the collected activity data, i.e. keystrokes and gaze data. In Section 5.1.1 we look at the amount of coherent typing activity needed to perform the post-editing task. Section 5.1.2 analyses the effort made by the post-editors in terms of the number of insertions and deletions, and Section 5.1.3 the gazing behavior.
- Section 5.2 describes several paths to assess the linguistic quality of the final post-edited text. Section 5.2.1 computes the edit distance between post-edited and reviewed versions of the text, and section 5.2.2 correlates post-editing time, number of text modifications, and edit distance between post-edited and reviewed texts.
- Section 5.3 presents the feedback provided by the translators in the form of questionnaires after completing each task.

5.1 Evaluation of activity data

Table 3 summarizes the user activity data that were collected during the field trial. For Dataset 1, gaze data were collected from all translators and reviewers. We analyzed the processing logs with respect to overall translation times, user effort in terms of edit operations, and gaze behavior.

[4] due to technical failure

[5] from logged segment translation pairs

Table 3 Data collected during the field trial. 460 distinct source segments were translated by 9 translators.

# of segment processing logs collected	total	with gaze data
PE condition	1,345	372
ITP condition	1,368	372
AITP conditon	1,373	372
lost data[4]	54	—
total	4,086	1,116

	English	Spanish
total tokens[5]	94,865	101,671
mean segment length	23	25

5.1.1 Overall translation time

In principle, the total processing time for a segment is the time lapsed between the moment the translator enters the edit box for a segment and the time he or she proceeds to the next one. However, in some of the logs from the sessions conducted from home, we observed very long pauses (up to several hours) suggesting that the respective participant interrupted these sessions and then returned to them later. By analysing the intervals between recorded edit events (recall that gaze data was not recorded for Datasets 2 and 3), we can make inferences about the underlying translation activity.

In our data, the vast majority of the pauses had a duration of a few seconds. Figure 4 shows the pause duration by means of a box plot. Box plots visualise data by means of a box that includes the first and third quartiles of the distribution as well as two arms or whiskers containing the extreme values. Box plots can also represent outliers[6] as isolated points at the left or at the right of the whiskers. Our data contain outliers so extreme that they could not be represented in the box plot without negatively affecting its legibility. Because of this, they have not been included in the diagram. Excluding outliers, all inter-keystroke intervals had a duration of 0.8 seconds or less. However, this does not mean that all of the outliers corresponded to noisy observations. Therefore, it is necessary to analyse the pauses more carefully. Here, we present two techniques to filter pause data in a meaningful way.

The first technique assumes that processing consists of alternating periods of typing and processing activity. Based on cognitive language processing and production theory (Alves and Vale 2009; Lacruz et al 2012; Carl 2012a), pauses between 0 and 5 seconds are used to segment the text production rhythm into "typing" and "processing" units.

In spite of the fact that the vast majority of the pauses had a duration of a few seconds, Figure 4 does not reflect their relative contribution to the total post-editing time. It is possible that there exist longer pauses that account for a substantial part of the segment post-editing time, even if they appear in a very small number (e.g. only one pause of 100 seconds accounts for the same time as one hundred pauses of 1 second). To clarify this, we generated a plot for different intervals of pause durations, summing their contributions to the total translation time. The result is represented as a weighted Pareto chart in Figure 5. Pareto

[6] Outliers are defined here as those points that exceed Q3+1.5 times the inter-quartile range, see (Montgomery 2004).

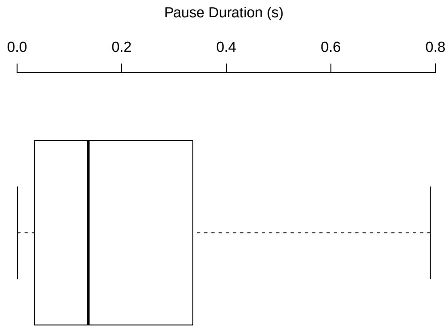

Figure 4 Boxplot for inter-keystroke pause duration in seconds (outliers not shown).

charts are used to highlight the most important factor among a typically large set of them. For this purpose, bars and a line graph are used, where the frequencies of individual values are represented in descending order by bars, and the cumulative total is represented by the line. Specifically, in a Pareto chart the left vertical axis represents the frequency of occurrence, while the right vertical axis is the cumulative percentage of the total number of occurrences. In weighted Pareto charts, frequencies are multiplied by specific magnitudes such as cost or loss associated with particular events so as to better analyse their importance (see for example (Montgomery 2004) for more details). The black line in the plot marks a relative frequency equal to 95%.

The plot given in Figure 5 provides valuable information about the effects of filtering pauses of a specific duration. The frequency of pauses belonging to a specific duration interval is weighted by such duration. For instance, the pauses with a duration between 0 and 10 seconds (0-10), consumed 58% of the post-editing time. Thus, according to the plot, filtering pauses of 10 seconds or more would remove the pauses that account for 42% of the total post-editing time. We think that such a filtering would alter the distribution of the post-editing times, resulting in average post-editing times that may not reflect correctly the real performance of each system. One alternative to filter the noisy inter-keystroke times mentioned at the beginning of this section would be to remove all pauses of 200 hundred seconds or more, since they roughly account for 95% percent of the post-editing time, as it can be seen in the plot.

Given these considerations, we executed two kinds of filtering over the set of inter-keystroke pauses, obtaining two new post-editing time measures (see also Section 3):

– **Kdur**: the total durations of *coherent typing* activity, excluding pauses where no keyboard activity was recorded lasting more than five seconds.
– **Fdur**: total durations of post-editing excluding pauses of 200 seconds or more.

Table 4 shows the average segment post-editing times in seconds for PE, ITP, and AITP systems for three different time measurements, namely Tdur (total duration without excluding any pauses), Kdur and Fdur. PE allowed for shorter post-editing times according to Kdur and Fdur measures. However, the differences between PE and ITP were very small when considering Fdur (the ITP system

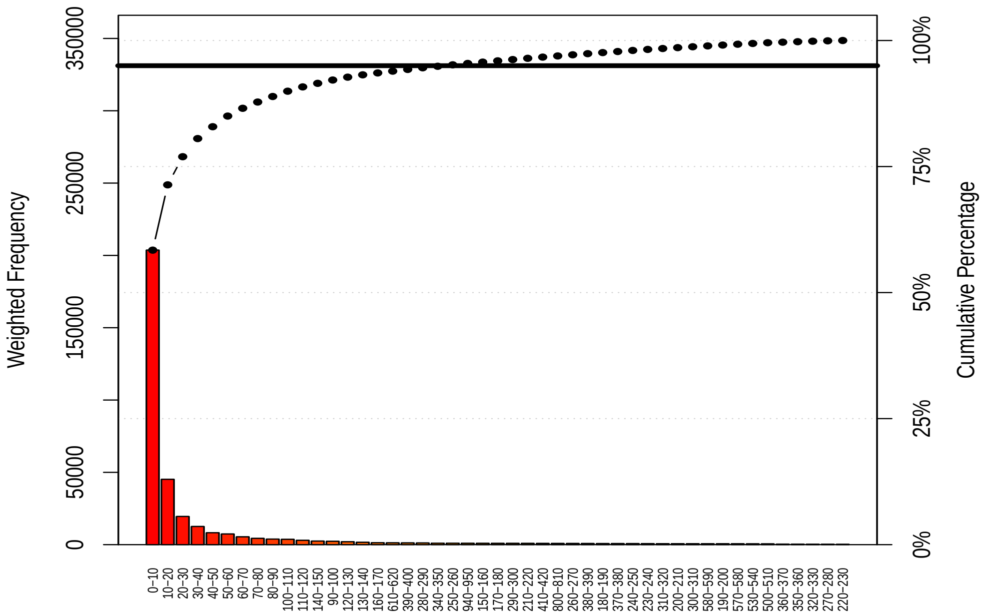

Figure 5 Weighted Pareto chart for inter-keystroke pause duration in seconds.

was 5% slower). One possible explanation for the greater differences between PE and ITP when considering Kdur may be due to the fact that ITP system users execute a higher number of short post-edit operations. Finally, Tdur values were different from the other two measures due to the noisy observations that have been mentioned above.

Table 4 Average post-editing times in terms of Tdur, Kdur and Fdur when using PE, ITP, and AITP systems.

System	Tdur	Kdur	Fdur
PE	104.0	21.7	73.0
ITP	80.7	27.0	77.0
AITP	117.1	29.6	92.4

One important thing to take into account when analysing the average post-editing times is the learning curves of each system. While PE systems are typically well-known by translators, this is not the case for ITP systems. For this reason, it is also interesting to compare the post-editing times that were required when translating from Celer Soluciones SL (dataset 1) with those obtained when translating from home (datasets 2 and 3). Since in our evaluation the translations were first generated from the office, it can be expected that users performed better with the ITP system from home after more hours of interaction (each dataset needed an average of 3.5 hours to be post-edited).

Table 5 shows a comparison of average post-editing time measured in terms of Kdur and Fdur for the PE, ITP and AITP systems, when translating both at the office or at home. As can be seen in the table, the average post-editing time was lower for all systems when the translations were generated at home. In addition to this, the post-editing time reduction for the ITP and AITP systems was greater than that for the PE system.

Table 5 Average post-editing time in terms of Kdur and Fdur when translating from the office or from home using PE, ITP and AITP systems.

	Kdur		**Fdur**	
System	**Office**	**Home**	**Office**	**Home**
PE	27.7	19.6	88.0	67.3
ITP	35.1	24.8	94.7	71.9
AITP	37.5	27.2	111.9	87.8

5.1.2 Typing activity

Enabling interactivity has also an effect on the number of insertions and deletions which the post-editor makes. Table 6 shows the average number of manual insertions and deletions per segment using the three systems at the office, at home or for all the sessions. According to the results, the ITP system required fewer operations than the rest of the systems.

Table 6 Number of insertion and deletions operations for translations generated at the office, at home or both using PE, ITP and AITP systems.

System	**Office**	**Home**	**All**
PE	114.9	134.6	131.3
ITP	109.6	127.2	123.6
AITP	143.2	137.0	132.6

It is important to note that these results must be interpreted in the light of the quality of the final output produced by the post-editors (see Section 5.2).

5.1.3 Gaze data

Drawing on the seminal work of (Just and Carpenter 1980), analyses based on the eye-mind hypothesis suggest that eye fixations can be used as a window into instances of effortful cognitive processing. Following this hypothesis, one could assume that eye-movement recordings can provide a dynamic trace of where a person's attention is being directed. This assumption is often taken for granted by eye-tracking researchers.

The average duration of gaze fixations in the source and target windows were calculated for each of the three systems in the field trial. Table 7 shows how participants exhibited a marked difference in the amount of time for which they gazed at the source and target windows. The use of interactivity features both in ITP and AITP triggered longer gaze fixations in the target window.

Under all three system configurations users exhibit on average more gaze fixations on the target rather than the source window. Unlike when translating from

Table 7 Average gaze fixations on source and target window per system.

System	**PE**		**ITP**		**AITP**	
	Nr.	%	Nr.	%	Nr.	%
Source window	18037	33.3	14422	26.0	16569	26.5
Target window	36193	66.7	41052	74.0	45999	73.5
Total	54230	100.0	55474	100.0	62568	100.0

scratch, the post-editor's task is to edit the MT output presented in the target window and thus it is not surprising that the primary focus is on that window. Enabling interactivity (ITP) and visualization (AITP), however, causes a decrease in the fixations on the source window and a corresponding increase in the target window.

5.2 Quality of post-edited data

This section evaluates the quality of dataset 1 in the trial. In section 5.2.1 we compare the post-edited version and the corresponding reviewed version using edit distance to assess post-editing quality. In section 5.2.2 we correlate edit distance with text modifications and revision time.

5.2.1 Edit distance in dataset 1

A quantitative analysis of the post-edited text has been carried out, based on the differences between the original post-edited version and the reviewed final texts.

Edit distances at word level have been used for this analysis. Words have been chosen as units because a word difference has typically much closer relation with both semantic quality and style than individual character differences. Moreover, rather than counting the absolute number of edit operations needed to transform the original text into the revised one, a relative figure (in %) is needed. This is important because the overall number of words is not the same for texts produced with the PE, ITP, and AITP systems and, without proper normalization, differences could be due to variations in text sizes, rather than to possible quality differences. Finally, in order to ensure the estimates are true percentages, one needs to normalize by the total number of edit operations, N, including non-error matches (i.e., $N = ins{+}del{+}sub{+}corr$, ins is for the number of inserted words, del is the number of deleted words, sub is the number of replaced words -substitutions- and $corr$ is the number of correct words). That is, the normalized edit distance is $(ins + del + sub)/N$. Such a normalization makes the product of the different systems fully and accurately comparable, regardless of the origin/reviewed sizes of each text.

The results of this analysis are presented in Table 8. Taking into account the 95% confidence intervals of these estimates ($\sim$1%), the conclusion is that the estimated quality of the translations — as assessed by the number of modifications introduced through the reviewer — is practically the same for the three assistance systems.

In this table it should be taken into consideration that only dataset 1 was analysed here. This means that the results are deduced from the translations generated while the post-editors were still getting used to the different systems.

Table 8 Quantitative analysis of the changes introduced by the reviewers.

Assistance system	**PE**	**ITP**	**AITP**
$ins + del + sub$	286	314	307
$ins + del + sub + corr$ $(=N)$	3082	2926	3050
Overall word changes (%)	9.3	10.7	10.1
Estimated quality (%)	90.7	89.3	89.9

5.2.2 Correlation of edit distance, revision time and text modifications

For this analysis, we counted the number of manual insertions and deletions for each of the four reviewers. Table 9 shows the average text modifications per system and reviewer R10 to R13. The table presents the average number of text modifications per segment divided by the length in characters of the segment for each of the three systems. In line with the results of (Guerberof 2012), reviewers seem to follow very different reviewing styles: reviewer R10 produces the least number of text modifications, while reviewer R13 is the most eager corrector. On average reviewers insert more modifications when the post-edited text was produced with system ITP.

Table 9 Average count, in percentage, of modifications (insertions and deletions) per character, reviewer and system in which the post-edited text was produced.

	PE	**ITP**	**AITP**	**total**
R10	8.9	0.8	4.8	4.8
R11	8.0	15.3	12.5	11.9
R12	9.4	9.8	8.8	9.3
R13	13.6	11.7	12.5	12.6
Total	10.0	9.4	9.7	9.7

We also computed the average revision time, edit distance and number of text modifications per reviewing session, which resulted in 12 data points for each of the variables (three systems × four reviewers). Unfortunately it was not possible to obtain reliable revision time on a segment level (which would have given many more data points) due to the fact that in the revision mode it was possible for the reviewer to read the segments, without loading them in the edit area of the workbench. As a consequence, we had to average over the entire revision session to get comparable numbers for average revision time, edit distance and number of text modifications.

Table 10 Correlations between keystrokes, edit distance and time in revision.

Assistance system	**PE**	**ITP**	**AITP**
Keystrokes vs. Time	$R^2 = .910$ $p > .081$	$R^2 = .998$ $p < .002$	$R^2 = .924$ $p > .076$
Edit dist. vs. Time	$R^2 = .740$ $p > .260$	$R^2 = .998$ $p < .002$	$R^2 = .946$ $p < .054$
Edit dist. vs. Keystrokes	$R^2 = .680$ $p > .320$	$R^2 = .999$ $p < .001$	$R^2 = .868$ $p > .132$

Table 10 summarises correlation and significance values, and shows that there is a strong correlation between these variables, but due to the small number of data points significance is not very high.

Figure 6 shows the correlations between text modifications and revision time. The highest correlation for all three variables can be observed in the ITP system and for the correlations between text modifications and revision time (Elming et al 2014).

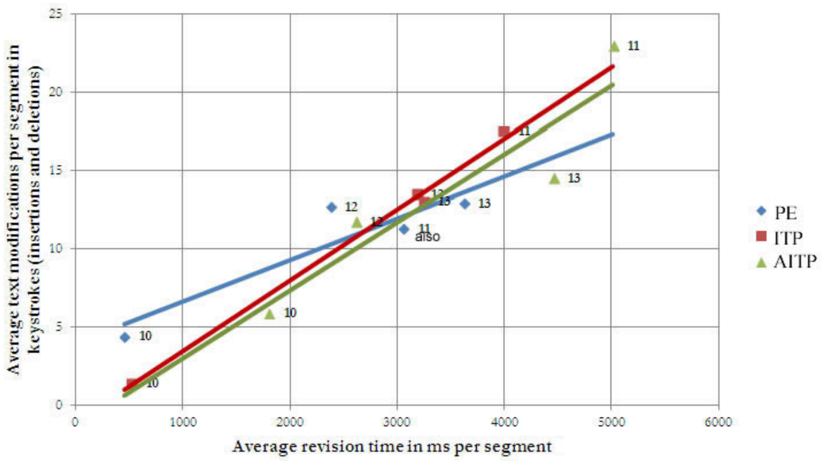

Figure 6 Correlation between: keystrokes (insertions and deletions) vs. time in milliseconds

5.3 User feedback

User feedback was elicited from the post-editors in the form of questionnaires. After each session, they were asked to rate their level of overall satisfaction on a 1-5 Likert scale, where 5 corresponded the highest positive reply and 1 the lowest.

User feedback was collected regarding the following questions:

– How satisfied are you with the translations you have produced ? (Satisfaction)
– How would you rate the workbench you have just used in terms of usefulness/aids to perform a post-editing task? (Tool)
– Would you have preferred to work on your translation from scratch? (From scratch)
– Would you have preferred to work on the machine translation output without the interactivity provided by the system? (No ITP)

Table 11 summarises the feedback provided by the post-editors after working with each of the three systems.

These results show different levels of satisfaction for the different systems. Some participants (i.e. 1, 3 and 4) seem to be more satisfied with the translations produced using interactive systems. Regarding the tool, interactive systems also are rated with a higher level of satisfaction overall, even though 7 out of 9 translators stated that they would have preferred not working with the interactivity provided by the system when using the ITP system. Their views are quite different when using AITP, since only two translators (6 and 8) continued thinking that they would have preferred to work without interactive features.

Table 11 Satisfaction ratings while using PE, ITP and AITP systems

	Satisfaction			Tool			From Scratch			No ITP	
	PE	ITP	AITP	PE	ITP	AITP	PE	ITP	AITP	ITP	AITP
Part.1	3	4	4	3	4	4	No	No	No	Yes	No
Part.2	4	4	4	3	2	4	Yes	Yes	Yes	No	No
Part.3	3	3	4	3	3	4	Yes	No	No	Yes	No
Part.4	4	4	5	3	4	4	No	No	No	No	No
Part.5	4	3	4	4	4	3	No	No	No	Yes	No
Part.6	5	5	5	3	3	2	No	No	No	Yes	Yes
Part.7	3	4	3	2	1	2	Yes	Yes	Yes	Yes	No
Part.8	4	4	3	2	2	3	Yes	No	No	Yes	Yes
Part.9	4	4	4	1	4	3	Yes	Yes	Yes	Yes	No

6 Related work

Improving the productivity of the translators is and has been a major driver of MT research. The hope is that, in many cases, post-editing MT output will help translators to perform their work faster. Several studies were performed to evaluate the potential benefit with generally positive results. Measured reductions in translation time typically range somewhere between 18% and 34% (Flournoy and Duran 2009; Guerberof 2009; Federico et al 2012) sometimes even reaching as high as 43% (Plitt and Masselot 2010).

Studies of translation vary in many dimensions which makes direct comparisons hard:

– Translators level of experience (volunteer, student (Koehn 2009a), professional (Plitt and Masselot 2010; Guerberof 2009))
– Suitability of the MT system, especially when comparing older (Krings 2001) and more recent (Plitt and Masselot 2010) studies
– PE software and subjects' familiarity with it
– Language pair and domain
– Data collection and filtering

Another question is whether post-editing leads to output of lower quality. Koehn (2009a) found that at least non-professional post-editors are generally both faster and produce better translations, a result that is consistent with later work (Plitt and Masselot 2010; Green et al 2013) investigating the same question with professional translators where a strong reduction in time and a reduced number of errors was found. Interestingly, Plitt and Masselot (2010) also find that the difference between individual translators is much stronger than between language pairs and MT systems of varying quality. Following this work, Skadiņš et al (2011) measure (slight) negative effects of the post-editing setting for both productivity and quality for some translators but still affirm the overall helpfulness of MT suggestions.

Along with the MT systems, PE environments have developed over time, recently converging towards web-based setups (Koehn 2009b; Green et al 2013) which integrate several aids in a single interface. Despite extensive research on Confidence Estimation for Machine Translation, such annotation has yet to be integrated. Bach et al (2011), for example, suggested visualizing word-level confidences by type size.

Besides quantity and quality, the translation process itself has been studied for many years, starting with explicit collection of translators' thoughts using Think Aloud Protocols (Krings 2001). Possible interference with the translation process quickly led to passive/indirect collection of user activity such as the logging of keystrokes and mouse movement (Langlais et al 2000) and, more recently, even gaze data (O'Brien 2009; Doherty et al 2010; Carl 2012b). By presenting multiple languages simultaneously in an ecologically valid environment, the combination of workbench and logging functions also offers a unique opportunity to investigate broader issues of applied bilingual cognitive processing.

7 Conclusions and future work

We have presented evaluation results that compare the performance of ITP versus conventional post-editing. More specifically, we defined two different kinds of ITP systems: a simple ITP system (referred to as ITP system) and an ITP system with advanced features (referred to as AITP system) and compared them with the post-editing system. Empirical results show that the ITP system accomplishes what it was designed to do, i.e., ITP minimises the number of key strokes that are required to generate the translations. In spite of this, the translation time per segment was a little bit higher for ITP system users than that required by the users of classical post-editing systems. Nevertheless, no substantial differences were found depending on how the translation times per segment were measured (using the Fdur measure, the ITP system required only a 5% more of translation time with respect to classical post-editing). In addition to this, results show that certain user profiles may benefit from interactivity when their experience with this kind of systems is increased. By contrast, the time results were worse for the AITP system, suggesting that some of the advanced features that were incorporated might not be useful to increase user productivity. However, we should take into account that the more complex the system, the steeper the learning curve. Considering that translators were already experienced post-editors, it seems logical to think that ITP and AITP systems had an initial disadvantage. In consequence, a longitudinal study would be necessary to shed more light on the effects of ITP and AITP systems.

On the whole, the analysis presented here includes results for the different system configurations calculated across users and text segments as a whole. A logical next step is to look in detail at the different post-editors and texts in order to see whether post-editing performance shows differences to identify user types who could most benefit from a post-editing workbench featuring interactivity.

References

Alabau V, Leiva LA, Ortiz-Martínez D, Casacuberta F (2012) User evaluation of interactive machine translation systems. In: Proceedings of the 16th Annual Conference of the European Association for Machine Translation (EAMT), pp 20–23

Alabau V, Buck C, Carl M, Casacuberta F, Garca-Martnez M, Germann U, Gonzlez-Rubio J, Hill R, Koehn P, Leiva L, Mesa-Lao B, Ortiz-Martnez D, Saint-Amand H, Sanchis-Trilles G, Tsoukala C (2014) Casmacat: A computer-assisted translation workbench. In: Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics

Alves F, Vale D (2009) Probing the unit of translation in time: aspects of the design and development of a web application for storing, annotating, and querying translation process data. Across Languages and Cultures 10(2):251–273

Bach N, Huang F, Al-Onaizan Y (2011) Goodness: A method for measuring machine translation confidence. In: Proceedings of the Annual Meeting of the Association for Computational Linguistics, pp 211–219

Barrachina S, Bender O, Casacuberta F, Civera J, Cubel E, Khadivi S, Lagarda AL, Ney H, Toms J, Vidal E, Vilar JM (2009) Statistical approaches to computer-assisted translation. Computational Linguistics 35(1):3–28

Brown PF, Della Pietra SA, Della Pietra VJ, Mercer RL (1993) The mathematics of statistical machine translation: Parameter estimation. Computational Linguistics 19(2):263–311

Callison-Burch C, Koehn P, Monz C, Post M, Soricut R, Specia L (2012) Findings of the 2012 workshop on statistical machine translation. In: Proceedings of the Seventh Workshop on Statistical Machine Translation, pp 10–51

Carl M (2012a) The CRITT TPR-DB 1.0: A Database for Empirical Human Translation Process Research. In: Proceedings of the AMTA 2012 Workshop on Post-Editing Technology and Practice (WPTP 2012), pp 1–10

Carl M (2012b) Translog-ii: a program for recording user activity data for empirical reading and writing research. In: Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)

Carl M (2014) Produkt- und prozesseinheiten in der critt translation process research database. In: Ahrens B (ed) Translationswissenschaftliches Kolloquium III: Beitrge zur bersetzungs- und Dolmetschwissenschaft (Kln/Germersheim), Peter Lang, Frankfurt am Main

Carl M Michael; Kay (2011) Gazing and typing activities during translation : A comparative study of translation units of professional and student translators. Meta 56(4):952–975

Doherty S, OBrien S, Carl M (2010) Eye tracking as an mt evaluation technique. Machine Translation 24(1):1–13

Elming J, Carl M, Balling LW (2014) Investigating user behaviour in post-editing and translation using the casmacat workbench. Cambridge Scholars Publishing, editors Sharon OBrien, Michael Simard, Lucia Specia, Michael Carl and Laura Winther Balling, chap 7, pp 147–169

Federico M, Cattelan A, Trombetti M (2012) Measuring user productivity in machine translation enhanced computer assisted translation. In: Proceedings of the Tenth Biennial Conference of the Association for Machine Translation in the Americas

Flournoy R, Duran C (2009) Machine translation and document localization at adobe: From pilot to production. In: MT Summit XII: Proceedings of the Twelfth Machine Translation Summit

Green S, Heer J, Manning CD (2013) The efficacy of human post-editing for language translation. In: Proceedings of SIGCHI Conference on Human Factors in Computing Systems, pp 439–448

Guerberof A (2009) Productivity and quality in mt post-editing. In: Proceedings of MT Summit XII-Workshop: Beyond Translation Memories: New Tools for

Translators MT
Guerberof A (2012) Productivity and quality in the post-editing of outputs from translation memories and machine translatio. Ph.D. Thesis
Just MA, Carpenter PA (1980) A theory of reading: from eye fixations to comprehension. Psychological review 87(4):329
Koehn P (2009a) A process study of computer-aided translation. Machine translation 23(4):241–263
Koehn P (2009b) A web-based interactive computer aided translation tool. In: Proceedings of ACL-IJCNLP 2009 Software Demonstrations, pp 17–20
Krings HP (2001) Repairing texts: empirical investigations of machine translation post-editing processes, vol 5. Kent State University Press
Lacruz I, Shreve GM, Angelone E (2012) Average Pause Ratio as an Indicator of Cognitive Effort in Post-Editing: A Case Study. In: Proceedings of the AMTA 2012 Workshop on Post-Editing Technology and Practice (WPTP 2012), pp 21–30
Langlais P, Foster G, Lapalme G (2000) Transtype: A computer-aided translation typing system. In: Proceedings of the 2000 NAACL-ANLP Workshop on Embedded Machine Translation Systems, pp 46–51
Leiva LA, Alabau V, Vidal E (2013) Error-proof, high-performance, and context-aware gestures for interactive text edition. In: Proceedings of the 2013 annual conference extended abstracts on Human factors in computing systems (CHI EA), pp 1227–1232
Montgomery D (2004) Introduction to Statistical Quality Control. Wiley
O'Brien S (2009) Eye tracking in translation process research: methodological challenges and solutions, Copenhagen studies in language, vol 38, Samfundslitteratur, Copenhagen, pp 251–266
Ortiz-Martínez D, Casacuberta F (2014) The new Thot toolkit for fully automatic and interactive statistical machine translation. In: Proceedings of the 14th Annual Meeting of the European Association for Computational Linguistics: System Demonstrations, pp 45–48
Plitt M, Masselot F (2010) A productivity test of statistical machine translation post-editing in a typical localisation context. The Prague Bulletin of Mathematical Linguistics 93(1):7–16
Sanchis-Trilles G, Ortiz-Martínez D, Civera J, Casacuberta F, Vidal E, Hoang H (2008) Improving interactive machine translation via mouse actions. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)
Simard M, Foster G (2013) Pepr: Post-edit propagation using phrase-based statistical machine translation. In: Machine Translation Summit, pp 191–198
Skadiņš R, Puriņš M, Skadiņa I, Vasiļjevs A (2011) Evaluation of smt in localization to under-resourced inflected language. In: Proceedings of the 15th International Conference of the European Association for Machine Translation EAMT, pp 35–40