



D1.2: Progress report on user interface studies, cognitive and user modelling

Michael Carl, Mercedes García Martínez,
Bartolomé Mesa-Lao, Nancy Underwood,
Frank Keller, Robin Hill

Distribution: Public

CasMaCat

Cognitive Analysis and Statistical Methods
for Advanced Computer Aided Translation

ICT Project 287576 Deliverable D1.2



Project funded by the European Community
under the Seventh Framework Programme for
Research and Technological Development.



Project ref no.	ICT-287576
Project acronym	CASMACAT
Project full title	Cognitive Analysis and Statistical Methods for Advanced Computer Aided Translation
Instrument	STREP
Thematic Priority	ICT-2011.4.2 Language Technologies
Start date / duration	01 November 2011 / 36 Months

Distribution	Public
Contractual date of delivery	October 31, 2013
Actual date of delivery	November 8, 2013
Date of last update	November 8, 2013
Deliverable number	D1.2
Deliverable title	Progress report on user interface studies, cognitive and user modelling
Type	Report
Status & version	Draft
Number of pages	69
Contributing WP(s)	WP7
WP / Task responsible	CBS, UEDIN
Other contributors	
Internal reviewer	
Author(s)	Michael Carl, Mercedes García Martínez, Bartolomé Mesa-Lao, Nancy Underwood, Frank Keller, Robin Hill
EC project officer	Kimmo Rossi
Keywords	

The partners in CASMACAT are:

University of Edinburgh (UEDIN)
Copenhagen Business School (CBS)
Universitat Politècnica de València (UPVLC)
Celer Soluciones (CS)

For copies of reports, updates on project activities and other CASMACAT related information, contact:

The CASMACAT Project Co-ordinator
Philipp Koehn, University of Edinburgh
10 Crichton Street, Edinburgh, EH8 9AB, United Kingdom
pkoehn@inf.ed.ac.uk
Phone +44 (131) 650-8287 - Fax +44 (131) 650-6626

Copies of reports and other material can also be accessed via the project's homepage:
<http://www.casmacat.eu/>

© 2013, The Individual Authors

No part of this document may be reproduced or transmitted in any form, or by any means, electronic or mechanical, including photocopy, recording, or any information storage and retrieval system, without permission from the copyright owner.

Executive Summary

This WP presents the empirical foundations for the development of the CASMACAT workbench. A series of experiments are being run to establish basic facts about translator behaviour in computer-aided translation, focusing on the use of visualization options and input modalities while post-editing machine translation (sections 1 and 2). Another series of studies deals with cognitive modelling and individual differences in translation production, in particular translator types and translation/post-editing styles (sections 3 and 4).

This deliverable, D1.2, is a progress report on user interface studies, cognitive and user modelling. It reports on post-editing and interactive translation experiments, as well as cognitive modelling covering Tasks 1.1, 1.2, 1.3 and 1.5. It also addresses the issues that were raised in the last review report for the project period M1 to M12, in particular:

- the basic facts about the translator behaviour in CAT (sections 1 and 4) highlighting usage of visualization and input modalities (see also D5.3).
- the individual differences in translator types and translation styles, (section 3, see also terminology, section A.1)
- the results and conclusions of preliminary studies conducted to investigate post-editing and translation styles (section 2 and 5)

From the experiments and analyses so far, it is clear that the data collected in the CRITT TPR-DB (Translation Process Research database) is an essential resource to achieve the CASMACAT project goals. It allows for large-scale in depth studies of human translation processes and thus serves as a basis of information to empirically grounded future development of the CASMACAT workbench. It attracts an international research community to investigate human translation processes under various conditions and to arrive at a more advanced level of understanding. Additional language pairs and more data increase the chances to better underpin the conclusions needed, as will be shown in this report, and as concluded in section 5.

Contents

1	Post-editing (Task 1.1 - completed)	5
1.1	Post-editing and translator behaviour using CASMACAT prototype-I	5
1.2	User feedback using CASMACAT prototype-I	5
1.3	A large scale analysis of post-editing behaviour	5
2	Interactive translation (Task 1.2 - ongoing)	6
2.1	Post-editing in CASMACAT prototype-II	6
2.2	Pilot experiments	9
2.2.1	Pre-pilot experiments	10
2.2.2	Pilot experiment	10
2.3	The second CASMACAT field trial	11
2.3.1	Participant profiles	12
2.3.2	Experimental design	12
2.4	Translation process data	13
2.4.1	The CRITT Translation Process Research (TPR) database	13
2.4.2	CASMACAT field trial data	14
2.5	Findings of the second CASMACAT field trial	16
2.5.1	Post-editing time	16
2.5.2	Typing activity	18

2.5.3	Gaze data	19
2.5.4	Post-editing quality	20
2.6	Review of post-edited data	20
2.6.1	Manual scoring	20
2.6.2	Edit distance	21
2.6.3	Edit-distance, revision time, text modifications	22
2.6.4	Final remarks and future work	24
3	Translator types and post-editing styles (Task 1.3 - completed)	24
3.1	Post-editing styles	24
3.2	Backtracking moves	25
3.2.1	Local backtracking	26
3.2.2	Long-distance backtracking	26
3.2.3	Post-editing strategies	27
3.2.4	Conclusions	27
4	Cognitive modelling (Task 1.5 - ongoing)	28
4.1	Tracing literal translation alignment in the translator’s mind	28
4.2	Psycholinguistic understanding of translation error detection	28
4.2.1	Detection by Monolinguals	32
4.2.2	Detection by Multilinguals	35
4.2.3	Comparison between Mono and Multilinguals	38
4.3	Modelling post-editing behaviour: an analysis of post-editing changes	39
4.3.1	Typology for the classification of post-editing changes	40
4.3.2	Results: Post-editing changes made in dataset 3	41
5	Connections with the rest of the project and further CasMaCat development	42
6	References	44
A	Appendix	46
A.1	Terminology	46
A.2	Second field trial description	56
A.3	Post-editing times	60
A.4	Total revisions per reviewer	62
A.5	Tracing literal translation alignment in the translator’s mind	64

1 Post-editing (Task 1.1 - completed)

This section briefly discuss results of the first CASMACAT field trial (sections 1.2 and 1.1) and a large scale analysis of post-editing behaviour in section 1.3.

1.1 Post-editing and translator behaviour using CasMaCat prototype-I

The analysis of the first field trial was based on more than 90 hours of English to Spanish post-editing and translation sessions performed by professional translators working for Celer Soluciones SL. The findings support an average time saving of 25% for post-editing machine translation compared to translation from scratch. The time saving correlates to a large degree with the number of keystrokes that a post-editor performs. It is interesting that this is a much better predictor than edit distance between the machine translation output and the final translation product, which is often used in the literature. The post-editor has to perform a large number of keystrokes before post-editing stops paying off compared to translation from scratch. It is also observed that translators had more gaze activity on the source segment when translating than when post-editing. This can be explained by the fact that a translation suggestion is already presented for post-editing, so less inspiration from looking at the source is needed. A detailed analysis of this investigation is presented in (Elming, Jakob, Michael Carl, and Laura Winther Balling, Forthcoming).

1.2 User feedback using CasMaCat prototype-I

Retrospective interviews were held with the five professional post-editors who were involved in the first CASMACAT field trial. The post-editors gave feedback on their experience working with the CASMACAT prototype-I and suggested new functionalities. A key finding of that investigation was the positive attitude of translation professionals towards greater automation during post-editing.

Interviewees asked for autowrite/autocomplete functionalities, search and replace functions, and quality control checks. These features have been implemented in CASMACAT prototype-II and tested in the second CASMACAT field trial. The most frequently discussed feature that post-editors believed needed greater automation while post-editing was the autopropagation of already fixed segments and sub-segments. This feature is anticipated in the CASMACAT prototype-III (Task 4.3) as part of the on-line learning functionalities.

A detailed evaluation of the interviews and the list of desirable features of a post-editing workbench are discussed in deliverable D6.2 on user evaluation.

1.3 A large scale analysis of post-editing behaviour

Using the data collected in the CRITT TPR-DB, a large-scale multi-lingual comparison of translation and post-editing behaviour was conducted to investigate and compare the behaviour of 68 different translators when translating and post-editing six English texts into four different languages: German, Spanish, Hindi and Chinese.

Through the analysis of key-logging and eye-tracking data, the main aim of this research was to evaluate human translators, performance with a view to assess different assistance possibilities for automated translation support. More specifically, this analysis aimed at explaining differences in the production time of Alignment Units (AUs) i.e. sequences of source-target correspondences. The main findings of this research in regard to AU can be summarised as follows:

- **Translation task:** from-scratch translation always takes longer than post-editing.
- **Inefficiency:** the more keystrokes are produced the longer it takes to produce the translation.
- **Parallel processing:** shifting attention frequently between different areas (TT, ST and keyboard) is time-consuming.
- **Average word frequency:** lower word frequency results in slower production time; this tendency is more pronounced for student translators.
- **Number of different possible translations:** high translation ambiguity only has a slow-down effect in post-editing.
- **Edit distance:** large edit distance between the source and target sides has a slow-down effect particularly for German.
- **Alignments:** alignment crossing distance only has significant effects for post-editing German and Spanish.
- **Target language:** with respect to overall translation time Hindi is slowest, Spanish is quickest, with no significant difference between German and Chinese.

A detailed description of the analysis in (Balling Winther, Laura; Carl, Michael, 2013).

2 Interactive translation (Task 1.2 - ongoing)

This section reports on a set of post-editing experiments to collect process data when using the second prototype of the CASMACAT workbench. One of the basic and most novel features that has been implemented in this CASMACAT prototype-II is Interactive Machine Translation (IMT), also known as Interactive Translation Prediction (ITP). This section provides a brief introduction to IMT and then reports on a series of experiments to compare IMT against more traditional ways of post-editing.

For a detailed technical and implementation description of IMT, please refer to deliverable D2.2.

2.1 Post-editing in CasMaCat prototype-II

An alternative to the traditional post-editing workflow is represented by the interactive machine translation (IMT) approach (Langlais and Lapalme, 2002; Casacuberta et al., 2009; Barrachina et al., 2009). In the IMT approach, a fully-fledged MT engine is embedded into a post-editing workbench allowing the system to look for alternative translations whenever the human translator corrects the MT output. MT technology is used to produce full target sentences (hypotheses), or portions thereof, which can be interactively accepted or edited by a human translator. The system continues searching for alternative renditions as the translator edits the text. The MT engine then exploits the changes made by the translator to produce improved outputs, and provides the user with fine-tuned completions of the sentence being translated.

CASMACAT prototype-I was redesigned in order to include IMT in the new CASMACAT prototype-II. First the MATECAT post-editing interface (Bertoldi et al., 2012) was leveraged and, secondly, the visualization of the advanced features of IMT were implemented on top of the interface. Figure 1 shows the implemented CASMACAT interface with all the advanced features enabled.

The main new features implemented in prototype-II are:

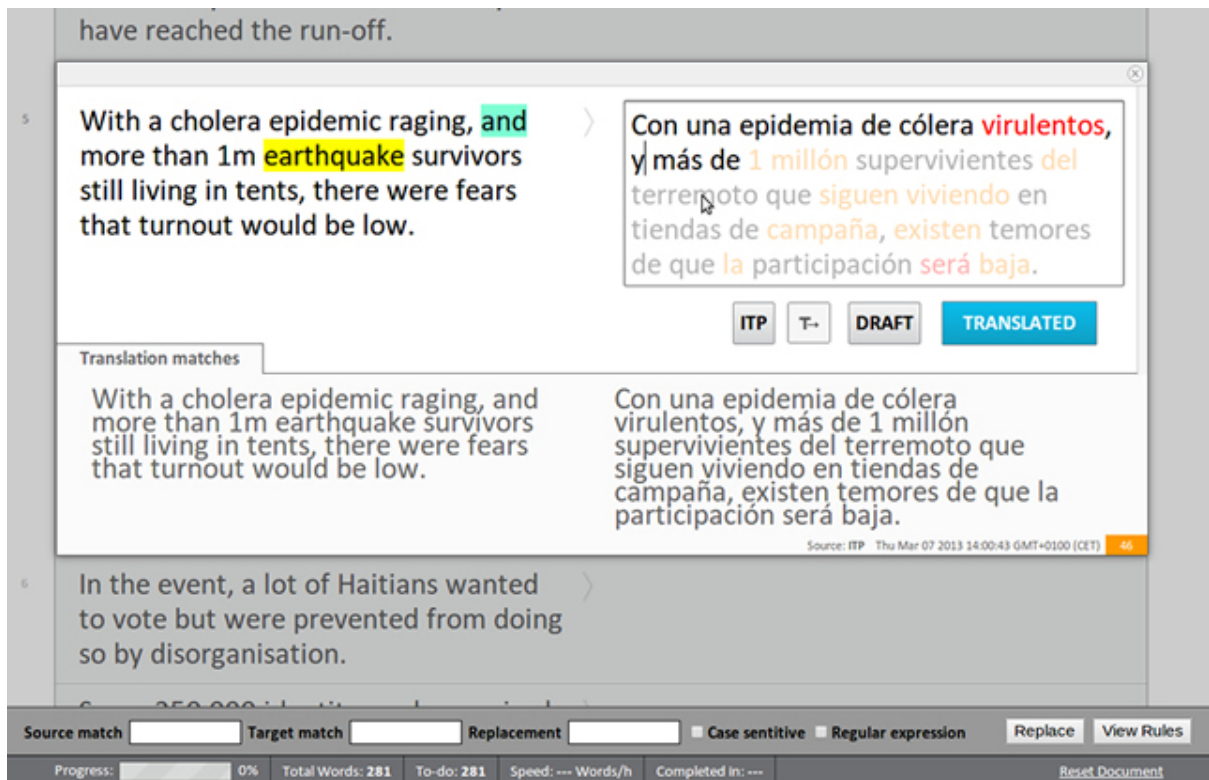


Figure 1: Screenshot of CSMACAT prototype-II with all new features enabled



Figure 2: Screenshot of CSMACAT prototype-II with optional visualization features disabled

- **Intelligent autocompletion** - basic IMT feature: Intelligent autocompletion takes place every time a manual keystroke is detected by the system. In such an event, the system produces a (full) suitable prediction according to the text that the user is writing. This new prediction replaces the remaining words of the original sentence to the right of the text cursor. This basic IMT feature is always enabled when the ITP mode is activated. This feature can be enabled or disabled by pressing the buttons ITP (Interactive Translation Prediction) and PE (post-editing). Whenever the user presses the PE button this intelligent autocompleting is disabled.
- **Prediction rejection:** This prototype also supports a mouse wheel rejection feature (Sanchis-Trilles, 2008) with the purpose of easing user interaction. By scrolling the mouse wheel over a word, the system invalidates the current prediction and provides the user with an alternate translation in which the first new word is different from the previous one. This option is one of the advanced IMT features.
- **Search and replace:** The CASMACAT workbench features a straightforward function to run search and replacement rules on the fly. Whenever a new replacement rule is created, it is automatically populated to the forthcoming predictions made by the system, so that the user only needs to specify them once. This specific function was implemented after the users asked for it in the first CASMACAT field trial (cf. deliverable D6.1). This option is always located at the bottom on the interface and it is also considered an advanced feature in the workbench. Figure 3 shows this feature:

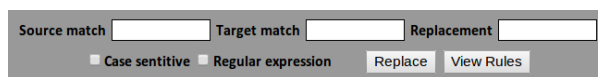


Figure 3: Search and replace bar at the bottom of the CASMACAT GUI

The user can also choose a number of advanced visualization options activating them as shown in Figure 2. These visualization options are:

- **Confidence measures:** The CASMACAT prototype-II workbench features confidence measures to inform post-editors about the reliability of translations under two different criteria. On the one hand, those machine translation outputs that are likely to be incorrect are highlighted in red. On the other hand, those machine-translated words that are considered dubious are highlighted in orange. This option is one of the advanced IMT features and it can be activated by selecting the visualization option *displayConfidences*. Figure 4 presents an example of this feature:

Lisboa y Madrid **desee** embarcarse en un camino **diferente** del
adoptado por Grecia **e** Irlanda.

Figure 4: Example of translated segment featuring confidence measures

- **Limit suffix length:** The number of predicted words that are shown to the user is limited to only predicting up to the first word with a low confidence measure according to the system. 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. This option is one of the advanced IMT features and it can be activated by selecting the visualization option *limitSuffixLength*. Figure 5 presents an example of this feature:

Lisboa y Madrid quieren emprender un camino diferente del
adoptado por Grecia e Irlanda.

Figure 5: Example of translated segment featuring limit suffix length

- **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. Figure 6 presents an example of this feature:

Lisbon and Madrid wish to embark on a path different from that
taken by Greece and Ireland.
Lisboa y Madrid desee embarcarse en un camino diferente del
adoptado por Grecia e Irlanda.

Figure 6: Example of word alignment information between source and target segments

In addition to these advanced features, three extra highlighting options are also available in CASMACAT prototype-II:

- *highlightValidated*: the system highlights in green the words that the user has modified.
- *highlightPrefix*: the system highlights the prefix. The prefix is defined as the first part of the segment that the user has validated.
- *highlightLastValidated*: the system highlights the last word that the user has modified.

2.2 Pilot experiments

This section details the research and experiments in a CASMACAT pre-field trial (PFT) study, performed before the second CASMACAT field trial. It enabled us to decide on the elements and configuration of the second prototype of the CASMACAT workbench before running the second CASMACAT field trial.

For the purpose of evaluating the visualization of advanced IMT features, four different configurations of the workbench were tested (see Table 1). Each of them differs in the set of features that are included (see section 2.1). System PFT1 was a baseline system for IMT including only basic intelligent autocompletion. Systems PFT2 to PFT4 included the intelligent autocompletion feature (IMT) together with some of the advanced features described above.

The main goal of this research was to measure user satisfaction when performing post-editing tasks using different workbench features (see table 1). In this context, we were interested in knowing whether translators find such features useful while post-editing MT outputs. All the logging files of these experiments are available in the CRITT TPR database¹.

¹ This data is available on-line: CRITT Translation Process Research (TPR) database. URL: http://bridge.cbs.dk/platform/?q=CRITT_TPR-db. See also section 2.4.2

Workbench features	PFT1	PFT2	PFT3	PFT4
basic intelligent autocompletion (IMT)	*			
IMT + confidence measures		*		
IMT + limit suffix length			*	
IMT + search and replace				*
IMT + word alignment information				*
IMT + prediction rejection				*

Table 1: List of the workbench features included in each of the four evaluated systems (PFT1 to PFT4)

2.2.1 Pre-pilot experiments

We carried two pre-pilot experiments in March 2013 from the CRITT premises at the Copenhagen Business School (CBS). One for the language pair English into Danish (4 hours) and another one for the language pair English into Spanish (10 hours). The main aim of these pre-pilot experiments was to test the eye-tracker plug-in implemented for the CASMACAT workbench (a brand-new plug-in for prototype II) as well as to debug any possible problems in the logging of IMT features while post-editing for the replay mode.

Due to the quality of the logged data, only the pre-pilot experiments for the language pair English to Spanish will be reported here. Five participants were each asked to interact during two hours with the CASMACAT prototype across the four configurations described in table 1 post-editing machine translation outputs from English into Spanish. At the end of each session, they answered some questions about their satisfaction with regard to the advanced features of IMT described in section 2.1.

Two major findings emerged from these pre-pilot experiments:

- The workbench should be able to identify when the translator edits back from right to left in order not to trigger further IMT suggestions that could change some of the editing work already done by the post-editor.
- All of the five post-editors in the pre-pilot test agreed that it would be useful to be able to enable or disable IMT at their convenience. Based on this request, the button ITP (Interactive Translation Prediction) was added to the GUI in order to be able to work with or without IMT at the segment level.

Figure 7 shows the ranking of most valued IMT features according to answers and comments from the five participants in these pre-pilot experiments. The feature most valued among the users was *Limit suffix length* and the least popular was *confidence measures*.

2.2.2 Pilot experiment

In preparation for the second CASMACAT field trial, in April 2013 a group of 16 participants volunteered to perform an evaluation of the CASMACAT workbench as described in Table 1. All participants in this pilot experiment had a degree in Translation Studies and were regular users of computer-aided translation tools (i.e., SDL Trados, MemoQ, etc.), but they had never used IMT technology to post-edit.

The aims of this pilot experiment were:

- From the technical point of view: To test the logging functions of CASMACAT prototype II under long-lasting experiments (1 hour of logging).

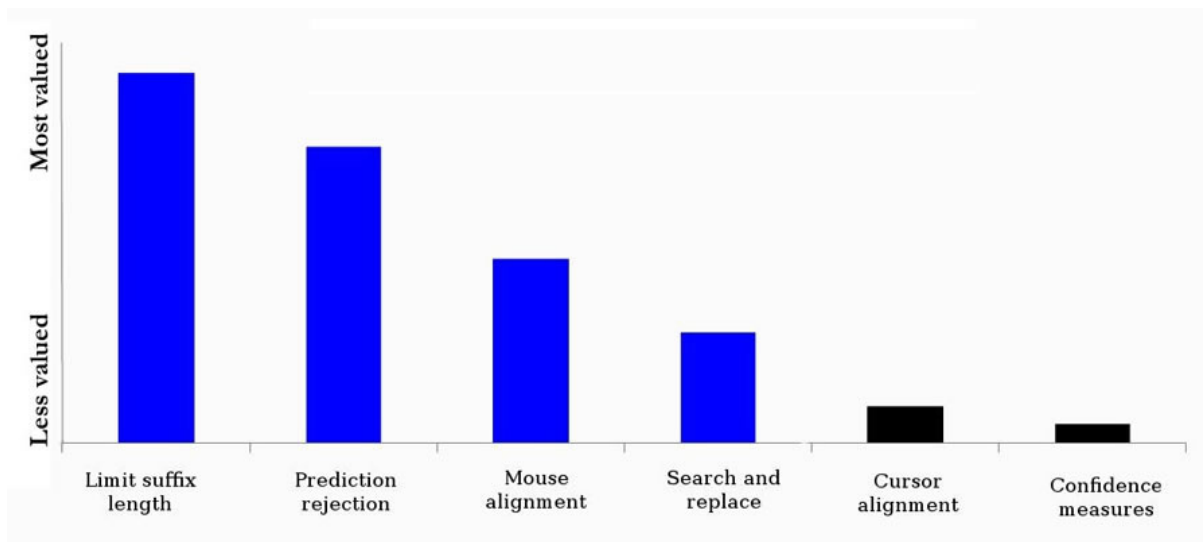


Figure 7: Ranking of advanced IMT features in CASMACAT prototype-II, in the pre-pilot study

- From the user point of view: To collect feedback on user satisfaction while using the prototype featuring IMT.

For this reason, a system usability scale (SUS) questionnaire was used to collect quantitative data on user satisfaction. Users had to assess each system on a typical five-level Likert scale, with 5 denoting the highest satisfaction, right after performing a post-editing task using each of the four different systems described in Table 1. In addition to the Likert scale, each questionnaire also included a text area for users to submit additional comments and feedback on the features being tested. A final overall questionnaire was also filled out in order to know which of the four configurations of the workbench was most preferred.

In preparation for the second field trial in Celer Soluciones SL, the language pair involved in this pilot experiment was English to Spanish. Each system was tested using a different data set consisting of 20 segments each; two pieces of news per system extracted from the News Commentary corpus². No time constraints were imposed on the participants involved in the evaluation.

Figure 8 shows the results of this user evaluation. For each of the evaluated systems, we display the average of the satisfaction scores given by the users (blue box), the 95% confidence interval for the average satisfaction score (black whisker), and the actual distribution of user satisfaction scores (gray pattern). The baseline system (system PFT1) was given an average satisfaction score of 2.4. In comparison, system PFT2 was given a slightly worse satisfaction score (2.1) while both system PFT3 (3.3) and system PFT4 (2.9) scored clearly above the baseline.

Overall, the most popular workbench configuration among participants was system CFT3 as was the case in the pre-pilot experiment. This was the reason why it was decided to include the feature *Limit suffix length* by default as an advanced IMT feature in the second CASMACAT field trial. A detailed description of this user evaluation study is provided in (Alabau, V., Mesa-Lao, B., et al. 2013) and reproduced in Appendix A.1 .

2.3 The second CasMaCat field trial

This section describes the second CASMACAT field trial (CFT) held at Celer Soluciones SL in June 2013. Following this description a series of findings are presented regarding time, effort and quality in post-editing with and without IMT.

² Training corpus for the sixth workshop on SMT 2011. URL: <http://www.statmt.org/wmt11>

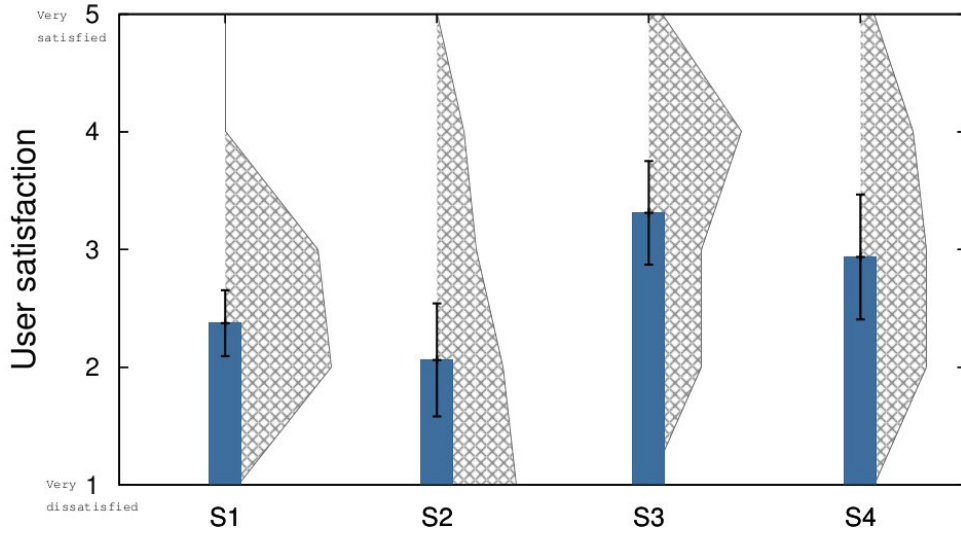


Figure 8: Results of user satisfaction in the pilot experiment prior to the second field trial

2.3.1 Participant profiles

The field trial involved nine post-editors and four reviewers. The post-editors were all freelance translators employed by Celer Soluciones SL and all but one (Participant 04) had previous experience of post-editing MT as a professional service. They all have Spanish as their first language and seven had English as their L2. Two had English as their L3. Of the reviewers who took part in the quality evaluation study, there were two in-house and one freelance reviewer and a freelance translator.

Details about participants' age, level of experience, professional education, etc., is available in the TPR database under the metadata folder³.

2.3.2 Experimental design

Based on the experiences gained in the pre-field trial experiments, we reduced the number of systems in the CFT study from four to three. In order to assess and compare the effects of post-editing using the IMT features described in section 2.1, each of the nine participants in the field trial was required to work with three different systems (all of them different configurations of the CASMACAT prototype-II):

Workbench features	CFT1 (P)	CFT2 (PI)	CFT3 (PIA)
traditional post-editing (no IMT)	*		
basic IMT (intelligent autocompletion)		*	
advanced IMT (intell. autocompletion + a choice of all other features)			*

Table 2: Workbench features included in each of the three evaluated systems (CFT1 to CFT3) in the second field trial

In the case of system CFT3 (the one featuring advanced IMT), participants were presented with all the advanced IMT features described in section 1 and they could choose which ones they would use while working in this system.

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

Every participant post-edited the same set of nine different texts, which makes a total of 81 post-editing sessions in all. The nine texts were organised in three datasets (dataset 1 to 3, see Table 3), so that for each dataset, each of the three systems under evaluation were used. All of these systems logged all keyboard and mouse activity while post-editing.

To ensure an equal distribution of texts and systems across the participants each one was assigned tasks specifying which system (CFT1:P, CFT2:PI or CFT3:PIA) they should use for each text so that all texts were post-edited by all participants using all different systems. The assignment of texts and systems was randomized in order to avoid any ordering effects. In addition to the post-editing task, the post-edited texts of dataset 1 were also reviewed in the offices of Celer Soluciones SL. Each of the reviewers was assigned to review the work of two or more of the post-editors. An overview of the systems and texts for each participant in the second field trial is provided in Appendix A.2.

Each source text consisted of approximately 1,000 words (distributed in 30 to 63 segments, see table 3) and they were all short news items in English (from the corpus news-commentary 2012, Callison-Burch et al., 2008). Each source text was machine-translated into Spanish by the statistical MT system developed by the UEDIN partners and then loaded into the CASMACAT workbench for the participants to post-edit.

Dataset	dataset 1			dataset 2			dataset 3		
Texts	11	21	31	12	22	32	31	32	33
#Segments	49	30	45	63	55	51	59	61	47

Table 3: Number of segments per text used in the CFT13 study - second CASMACAT field trial

In order to collect the gaze data described in section 3 and 4, each of the participants had to carry out the three post-editing sessions for dataset 1 (Texts 11,12 and 13, one with each of the three systems P, PI and PIA) at the offices of Celer Soluciones SL, where an eye-tracker (EyeLink 1000) was used to record gaze behaviour. The other six sessions of datasets 2 and 3 were performed at home where no eye-tracking was available. Revisions made by the reviewers at the office were also monitored with an eye-tracker.

Before starting their tasks, participants were introduced to the CASMACAT workbench (Prototype-II) and the three different systems under evaluation during the second field trial. They were given time to familiarise themselves with the tool and try out the different visualization options. The participants themselves then chose which options they would enable when post-editing using S3:PIA. After each session participants were asked to fill out an on-line questionnaire and were then interviewed in depth after all sessions at Celer Soluciones were completed. Details about these questionnaires and interviews are included in deliverable D6.2.

2.4 Translation process data

This section presents the collected data of the field trial and how it was processed in the CRITT TPR database. The raw logging data of the CASMACAT workbench was post-processed and compiled into the CRITT TPR database format so it can be easily accessed, analyzed and compared with other translation studies by anyone interested in post-editing.

2.4.1 The CRITT Translation Process Research (TPR) database

The CRITT TPR database contains user activity data (UAD), such as keystrokes and gaze data of reading and writing activities recorded both with Translog-II and the CASMACAT workbench (prototypes I and II). The data is available as raw logging data and has been converted into data tables, annotated with metadata (Hvelplund and Carl 2012).

A first version of the CRITT TPR database was released in May 2012 with a total of ten studies amounting for a total of 456 (translation, post-editing, editing or text copying) sessions (Carl 2012). In its present state, the CRITT TPR database contains a total of 17 studies. The most recent contribution to it has been the logging files resulting from this second CASMACAT field trial. The raw data logging and alignments can be checked out via:

```
svn co https://130.226.34.13/svn/tpr-db/
```

This svn repository includes Translog-II data and the logging data of all CASMACAT field trials and pre-field trials. There are compiled versions of the TPR-DB ⁴ which contain a number of post-processed more comprehensible data tables, as described in section 2.4.2. The database comprises:

- CFT13 - data from the second CASMACAT field trial, June 2013. This contains logging data of the 81 post-editing and review sessions. More than 120 hours of UAD:
http://bridge.cbs.dk/platform/?q=CRITT_TPR-db#CFT13
- PFT13 - data from the CASMACAT pre-field trial experiments 2013 can be downloaded from: http://bridge.cbs.dk/platform/?q=CRITT_TPR-db#PFT13

Outside the CASMACAT consortium, there are - to our knowledge - a number of research groups in Wolverhampton, Sheffield, Leicester, Durham (all UK), Germersheim (Germany), Kent (USA), Mumbai (India), Macao (China) and Tokyo (Japan), working with the TPR-DB and we expect the number of interested researchers to grow. The results produced by these external researchers complement our own research and has resulted in a number of forthcoming publications.

2.4.2 CasMaCat field trial data

As outlined above, a total of 460 source text segments distributed over nine texts were post-edited by nine post-editors. From the total of 4,140 segments, 54 segments were lost due to logging problems so that 4,086 segments were kept. The logging data of these segments is included in the CRITT TPR database. The total amount of logged segments account for 94,865 English source tokens (average 23 source text words/segment) translated into 101,671 Spanish words (average 25 words/segment). The three texts of dataset 1 (Texts 11, 12 and 13) with a total number of 124 source text segments were post-edited by 9 translators and subsequently reviewed by four different reviewers. Gaze data was collected with an eyekink 1000. These sessions amount to a total of 1,116 post-edited and reviewed segments. Table 4 gives an overview of the collected data available in CFT13 study, indicating that almost all the 460 different segments have been translated by three translators with all three systems.

System	#Segments	Segments containing gaze data	Segments reviewed
CFT1: P	1345	372	372
CFT2: PI	1368	372	372
CFT3: PIA	1373	372	372
Total	4086	1116	1116

Table 4: Data overview of the CFT13 study - second CASMACAT field trial

The available data in the TPR database consists basically of two types of process information that is most interesting for further investigation: (i) keystroke data, i.e. information about the time and kind of text insertions and deletions performed by the post-editors, and (ii) information

⁴The TBR-DB website is http://bridge.cbs.dk/platform/?q=CRITT_TPR-db

on gaze fixations on the source or the target text, and the different translation options that were presented to the post-editor via the interactivity provided by the system.

The data made available in the TPR-DB CFT13 study contains complex units derived from the process information: text production units (PU) and fixation units (FU), which represent sequences of coherent writing and reading respectively. From the final translation product are extracted: source text tokens (ST), target text tokens (TT) and alignment units (AU) of source and target tokens. The TPR-DB post-processes the raw logging data and extracts the following information:

- **Keystrokes:** basic text modification operations (insertions or deletions), together with time of stroke, and the word in the final target text to which the keystroke contributes.
- **Fixations:** 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 character offset and word index of fixated symbol in the source or target window.
- **Production units:** coherent sequence of typing (cf. Carl and Kay, 2011), 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 insertion, deletions.
- **Fixation units:** 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.
- **Source tokens:** as produced by a tokenizer, together with TT correspondence, number, and time of keystrokes (insertions and deletions) to produce the translation, micro unit information.
- **Target tokens:** as produced by a tokenizer, together with ST correspondence, number, and time of keystrokes (insertions and deletions) to produce the token, micro unit information, and the amount of parallel reading activity.
- **Alignment units:** 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:** describe the source and target segments, annotated with the number of keystrokes, insertions, deletions, fixations on the source and the target side.

In addition, a table contains for each post-edited segment the following information:

- **Nedit:** number of times the segment was opened.
- **Tdur:** cumulative duration in which the segment was opened.
- **Kdur:** cumulative duration with keystroke activity (excluding pauses of 5 seconds or longer).
- **Fix1:** fixation duration on source window.
- **Fix2:** fixation duration on target window.
- **Mins:** manual insertions.
- **Ains:** automatic insertions.

- **Adel**: automatic deletions.
- **TokS**: number of tokens in the source segment.
- **LenS**: number characters in the source segment.
- **TokT**: number tokens in the target segment.
- **LenT**: number characters in the target segment.

In summary, the UAD available from CFT13 study can be analyzed from many different angles in a more comprehensive manner than the original logging data and may serve as a basis for further investigation for anyone wanting to use this resource.

2.5 Findings of the second CasMaCat field trial

This section presents preliminary investigations into the field trial data described in section 2.3 and in section 2.4. The post-edited data was analyzed by looking at: (i) the time needed to perform the post-editing task (i.e. productivity), (ii) the effort made by the post-editors in terms of the number of insertions and deletions, (iii) the gazing behavior, and (iii) the linguistic quality of the final post-edited text. In section 2.6, we report analyses of the editing behavior of the reviewers in terms of (i) manual assessment of reviewer activities, and (ii) time needed for reviewing, number of text modifications, and edit distance between post-edited and reviewed texts.

Due to logging problems in some sessions performed at home, a number of segments were not logged correctly. In particular, certain participants had difficulties in saving either the first or last segment of a text. Therefore, in order to ensure valid comparisons, if a participant was unable to save a particular segment, the corresponding segment in all other translations of the same text were ignored in these analyses of the data. In the case of text 1.2 one user was unable to save 22 segments and so for this preliminary analysis all the translations of this text were also excluded. Similarly the eye-tracking for two participants (participant 01 and 06) was found to be unreliable and excluded from the analyses. In the case of participant 01, she suffered from nystagmus which distorts the eye-tracking data. In the case of participant 06, his typing skills were so poor that he spent most of the sessions looking at the keyboard.

2.5.1 Post-editing time

In the sessions carried out at home some participants registered very long pauses (up to several hours) which would seem to indicate that the participant interrupted these sessions and then returned to them later. Therefore for the purposes of comparison we calculated the durations of user activity excluding pauses lasting over five seconds.

Pauses of a length between one and five seconds (inter-keystroke time) have been used in previous studies (Alves et al 2009, Lacruz 2012, Carl 2012) to fragment the text production rhythm into typing or processing units. With a pause of 5 seconds we are thus on the upper limit of what researchers consider a boundary for coherent typing. In our data, 95% of all successive keystrokes are produced within 5 seconds. Figure 9 shows the distribution of pauses between keystrokes according to their duration. The 5-seconds bar indicates the 95% threshold. Note, however, that there is a greater contribution of short pauses to the total translation time for PI and PIA (CFT2 and CFT3) systems as compared to the P (CFT1) system. As a result, removing pauses of more than 5 seconds increases the relative differences between traditional and interactive post-editing.

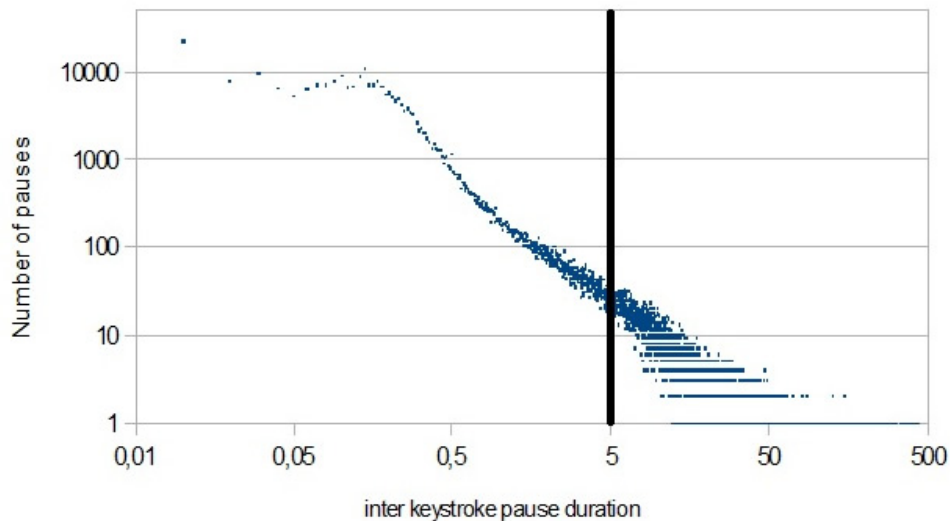


Figure 9: Distribution of inter-keystroke pauses in seconds

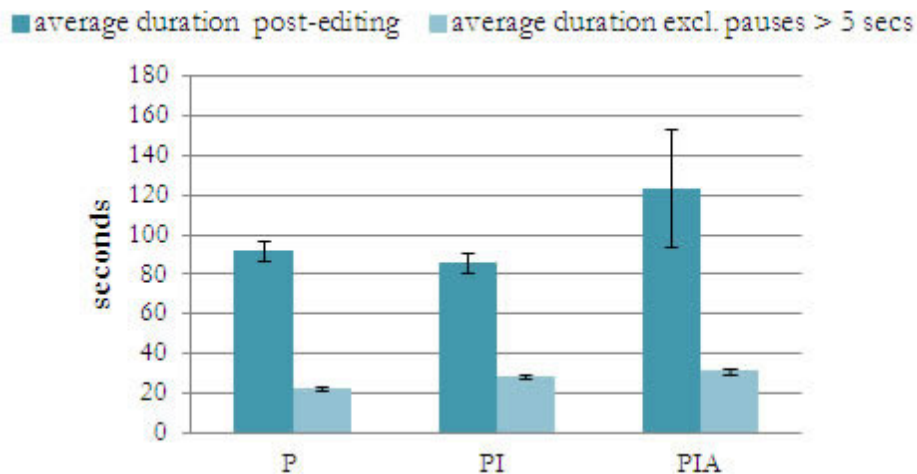


Figure 10: Average post-editing time per segment - Tdur vs Kdur

Figure 10 shows a comparison between the total duration of the post-editing sessions (Tdur) with the duration excluding pauses over five seconds (Kdur).

When comparing the time spent post-editing using the three different systems included in the field trial, as can be seen in Figure 11, the use of interactivity seems to increase post-editing time. However, comparing the data collected from the initial sessions at the office and the data collected from subsequent sessions at home, there is a reduction in the average post-editing time per segment (Kdur) as shown in Figure 12.

Although there is still a marked effect when using interactivity during post-editing tasks, the average time for post-editing with interactivity enabled (PI) decreased more than for the other two configurations (P and PIA). The reasons for this time decrease can be interpreted in different ways, but it would be interesting to investigate whether this is due to some sort of learning effect (i.e. the more you interact with the system the less time you need to perform a post-editing task) or simply the effect of being in a less formal experimental setting when working with the workbench from home.

It will also be worthwhile to drill down to consider the figures according to the post-editor and text to get a more accurate view since, certain post-editors are inherently slower than

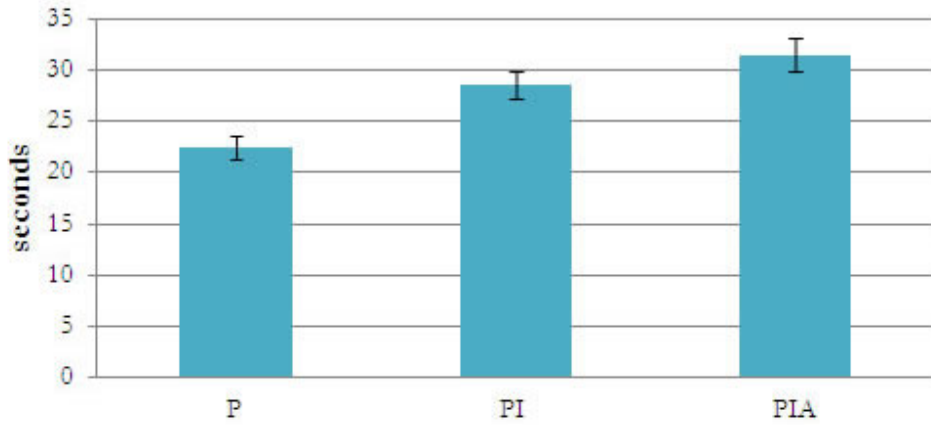


Figure 11: Activity based post-editing times (Kdur)

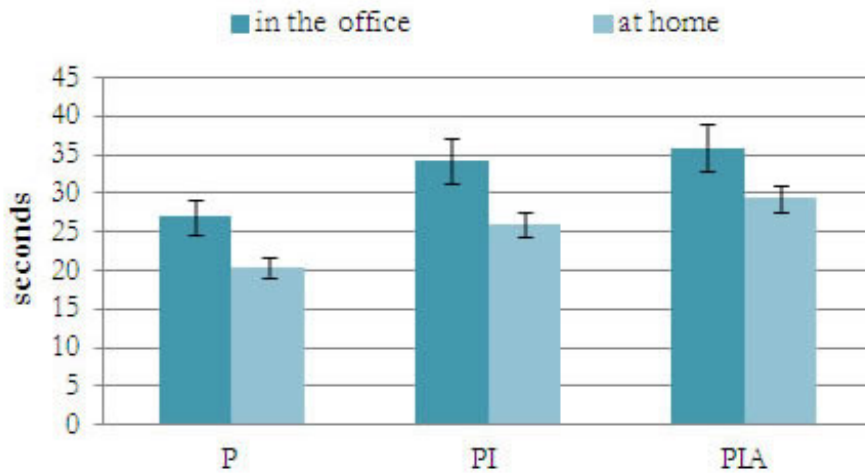


Figure 12: Average post-editing times per segment (Kdur)

others, whilst different texts are more or less easy to post-edit, depending on the quality of the MT output.

Appendix A.3 presents some preliminary data about the average time per token for each participant using the different texts and systems configurations.

2.5.2 Typing activity

Enabling interactivity also has an effect on the number of insertions and deletions which the post-editor makes. Figure 13 shows the average number of manual insertions and deletions per segment for the three systems in all the sessions.

This effect seems more pronounced in the sessions which were carried out from Celer Soluciones SL (Figure 14). Maybe the reason again is that it was their first experience working with the workbench under these three configurations.

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 (section 2.6.1). Comparing Figure 14 with Figure 16 we see that the number of keystrokes by the user is in inverse proportion to the number of errors still present in the final text.

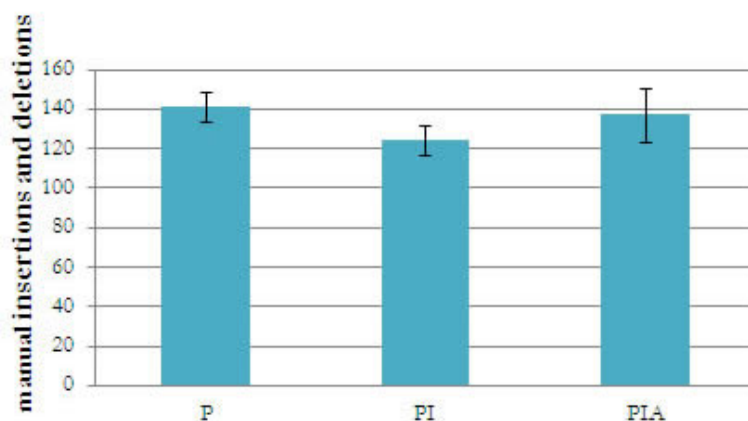


Figure 13: Average manual insertions and deletions per system in all sessions

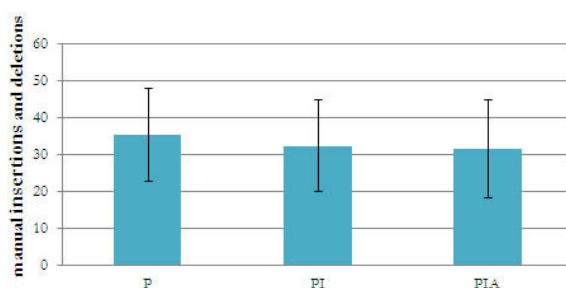


Figure 14: Average manual insertions and deletions per system in sessions performed at Celer offices

2.5.3 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 today taken for granted by eye tracker researchers. In the eye-tracking literature the number of fixations recorded is assumed to correlate with cognitive load.

The average duration of gaze fixations in the source and target windows were calculated for each of the three system in the field trial. Figure 15 shows how participants exhibited a marked difference in the amount of time which they gazed at the source and target windows. The use of interactivity features both in PI and PIA 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 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 (PI) and visualization (PIA), however, causes a decrease in the fixations on the source window and a corresponding increase in the target window. The next logical step in this analysis is to drill down and consider the behaviour of individual post-editors and correlate this with specific texts and the visualization options which were chosen. Further details on gaze behaviour are presented in section 3.



Figure 15: Average gaze fixations on source and target window per system

2.5.4 Post-editing quality

The errors in the post-edited dataset 3 (approx. 3,000 words) were manually counted, and classified into different groups (for detailed analysis see Section 4.3.1). This analysis suggested that enabling IMT (system PI) produced texts with less errors overall. Although the differences in quality associated with the different systems are not significant, the manual analysis of the residual errors (i.e., the errors that the post-editors overlooked (*essential changes not implemented*) or errors introduced) revealed that among all three systems (P, PI and PIA), PI was the one with the least amount of residual errors.

This contrasts with the conclusions drawn from the reviewers’ work (see next section 2.6), where all systems are deemed indistinguishable in terms of edit-distance, with PI presenting a slight increase in the number of edits required with respect to P. The reason for this different assessment of the systems may be due to the fact that the reviewers worked on the datasets 1 (which was the first to be translated), whereas the manual analysis was performed on dataset 3 (the last of the datasets translated). This suggest that the post-editors need some time to get used to IMT and to learn how to take advantage of its strengths, but that it might actually result in better quality translations once translators are used to the system. A longitudinal study, where linguists would work for several weeks with the systems, would shed more light on the effects of IMT.

2.6 Review of post-edited data

The dataset 1, i.e. the 27 texts that were produced at the office in Celer Soluciones SL from nine post-editors (three texts of each post-editor) were also revised by four reviewers. In this section we discuss an analysis of the reviewers’ behaviour: (i) a manual assessment of the text modifications, and (ii) the edit-distance between the post-edited texts and the reviewed version and (iii) the correlation between text modifications, edit-distance and revision time.

2.6.1 Manual scoring

A qualitative analysis of the revised translations produced by the P, PI, and PIA systems shows that the reviewers did make a very good and careful work (Appendix A.4 provides details about the work carried out by the reviewers). Only about 10-15% of the modifications introduced by

the reviewers can be considered style or presentation minded (such as changing the character used for quotes, etc.) The remaining 85-90% modifications were really needed in order to render the result semantically equivalent to the source. While there are a few discrepancies among reviewers (and even for the same reviewer) about what needs to be revised and what is not needed, these discrepancies only affect the style changes just mentioned.

2.6.2 Edit distance

A quantitative analysis of the changes introduced by the reviewers has been carried out on the differences between original translations and the resulting, revised 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 P, PI, and PIA 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 plotted in Table 5

Assistance system	P	PI	PIA
$ins + del + sub$	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

Table 5: Quantitative analysis of the changes introduced by the reviewers

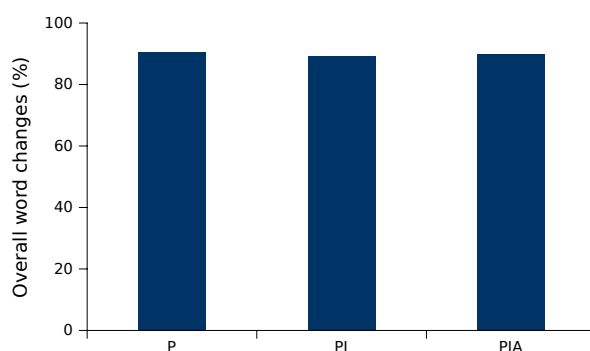


Figure 16: Revisions made by the reviewers per system for the sessions performed at the office

A graphical summary of this table is presented in Figure 16). 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 these graphs it should be taken into consideration that dataset 1 was here analyzed. This means that the results are deduced from the translations generated while the post-editors were

still getting used to the different systems. An analysis of the dataset 3 and their discrepancies was already discussed above.

2.6.3 Edit-distance, revision time, text modifications

We counted the number of manual insertions and deletions for each of the four reviewers. Table 6 shows the average text modifications per system and reviewer R10 to R13, while the graph 17 represents the average for each system visually. The table plots the average number of text modifications per segment divided by the length in characters of the segment for each of the three systems. Reviewers seem to follow different reviewing styles: reviewer R10 produces the least number of text modification, while reviewer R13 is the most eager corrector. Figure 17 shows that on average reviewers produce most relative text modifications when the post-edited text was produced with system PI.

	P	PI	PIA
R10	0.0891	0.0077	0.0483
R11	0.0795	0.1531	0.1253
R12	0.0941	0.0977	0.0880
R13	0.1357	0.1172	0.1248

Table 6: Average count of modifications (insertions and deletions) per reviewer and system

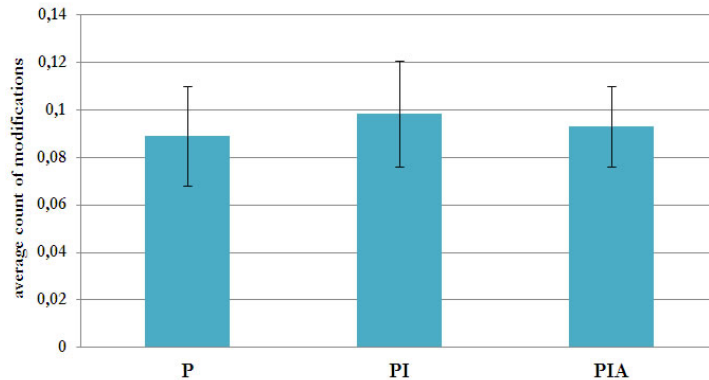


Figure 17: Average count of modifications (insertions and deletions) per system

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 \times 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. However, only when loading a segment into the edit area would also revision time be allocated for that segment. If no changes were required in a segment, it was usually not loaded into the edit area. 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.

Assistance system	P	PI	PIA
Keystrokes vs. Time	$R^2 = .910$ $p > .081$	$R^2 = .998$ $p < .002$	$R^2 = .924$ $p > .076$
Edit distance vs. Time	$R^2 = .740$ $p > .260$	$R^2 = .998$ $p < .002$	$R^2 = .946$ $p < .054$
Edit distance vs. Keystrokes	$R^2 = .680$ $p > .320$	$R^2 = .999$ $p < .001$	$R^2 = .868$ $p > .132$

Table 7: Correlations between keystrokes, edit distance and time in revision

Table 7 summarizes 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.

Figures 18, 19, and 20 show respectively the correlations between text modifications and revision time, edit-distance and revision time, and edit-distance and text modifications. The highest correlation for all three variables can be observed in the PI system. In a previous study (see section 1.1) we had observed that post-editing time and text modifications are better correlated than post-editing time and edit-distance. For the reviewing sessions of the CASMACAT field trial we could not confirm these findings based on our few data points.

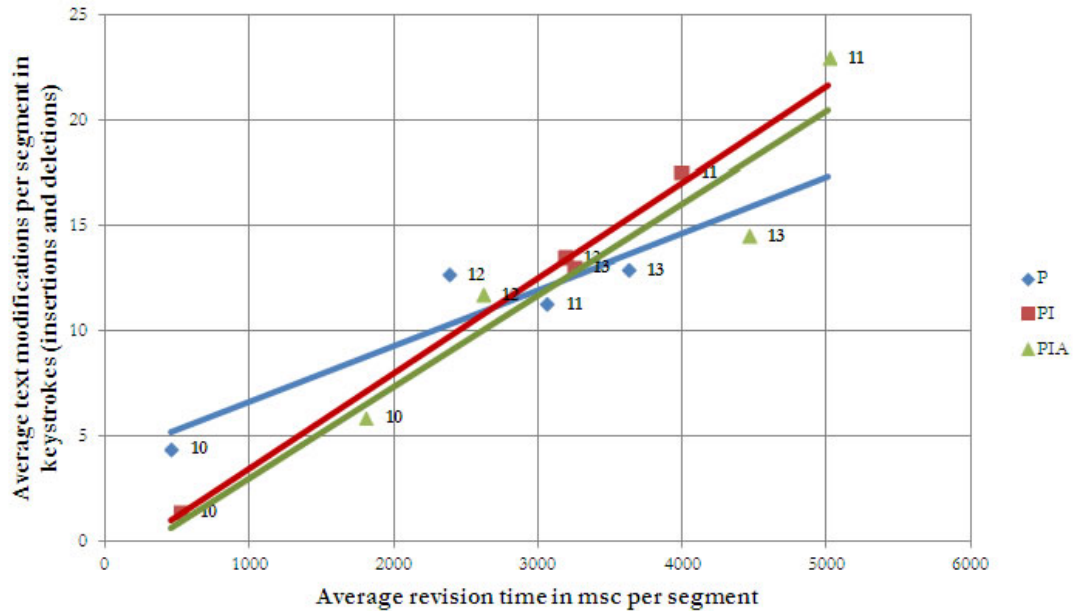


Figure 18: Correlation between: keystrokes (insertions and deletions) vs. time

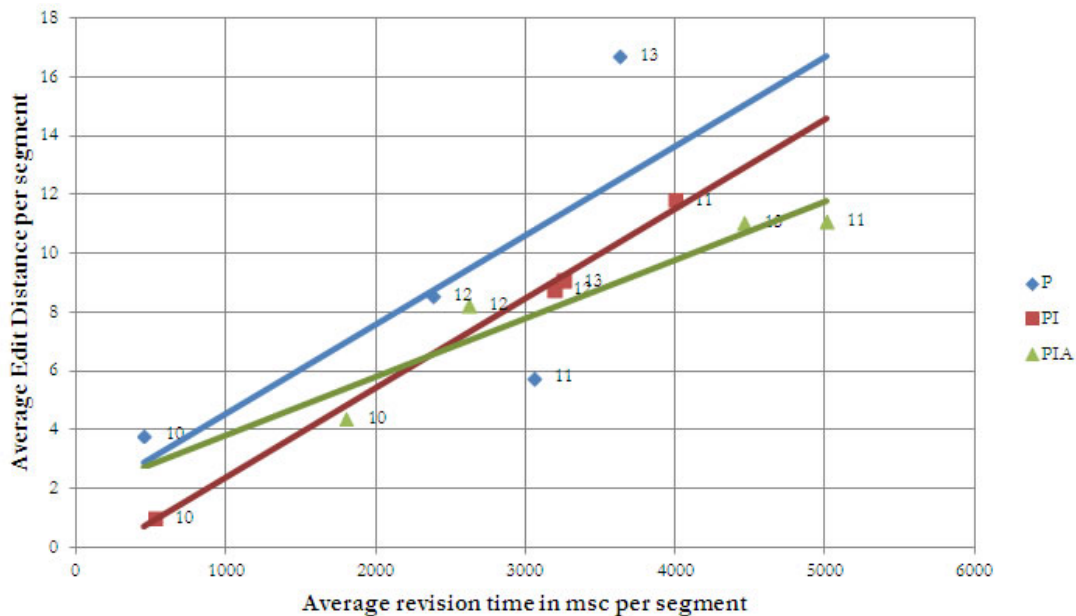


Figure 19: Correlation between: edit distance vs. time

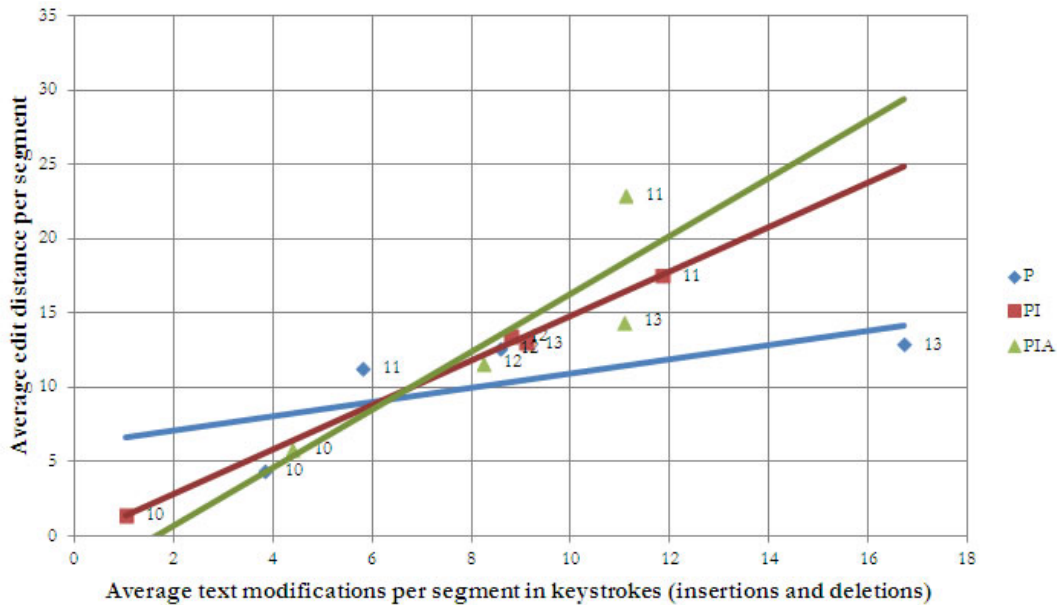


Figure 20: Correlation between: edit distance vs. keystrokes

2.6.4 Final remarks and future work

The overall results of this preliminary analysis are not immediately positive for interactive features in the current version of the CASMACAT workbench. Indeed they tend to confirm the results of a similar user evaluation of such features (see Appendix A.1). Nevertheless they are not entirely negative for all participants and it seems that certain user profiles may benefit from interactivity after more interaction with the system featuring interactivity (PI and PIA).

On the whole the preliminary analysis presented here involves results for the different system configurations calculated across all users and text segments. To get a fuller picture we are performing further analyses based on these two variables (participant and texts). A logical next step is to look in detail at the results for different post-editors and see whether and how these correlate with their professional profiles to identify user types who could benefit most from the interactive features.

For the purposes of this second CASMACAT field trial PI and PIA were enabled during an entire session. However, it is in fact possible for users to switch on and off interactivity (ITP/PE button) as well as some of the visualization options at their convenience to make the most of IMT.

3 Translator types and post-editing styles (Task 1.3 - completed)

This section describes the post-editing styles as identified in the data collected during the second CASMACAT field trial. Section 3.1 gives an account of post-editing styles based on gaze behaviour. Section 3.2 presents post-editors behaviour in terms of backtracking moves among the segments of the text being post-edited.

3.1 Post-editing styles

Figures 21, 22, 23 and 24 show the four different styles which were identified in the user activity data of dataset 1, collected using an eye-tracker. Each of these four styles differ in the reading

pattern followed by the post-editor when processing the MT output and they must be interpreted at the segment level:

- In style 1 the post-editor first reads the target text (raw MT output) and then refers to the source text before making changes in the target text.
- In style 2 the post-editor first reads the source text and then proceeds to read the target text looking for changes needed in the MT output.
- In style 3 the post-editor works on the target text without referring to the source text at any time (monolingual post-editing).
- In style 4 the post-editor reproduces style 1, but adding local backtracking to previous or subsequent segments before actually making any changes in the target text.

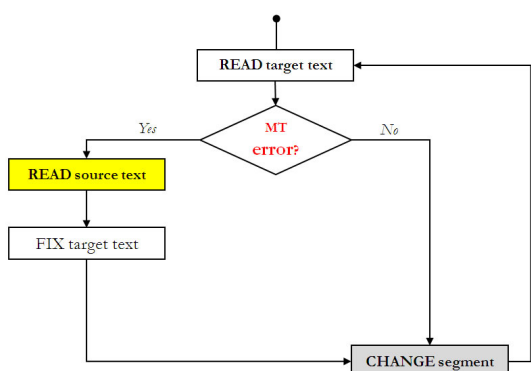


Figure 21: Post-editing style 1

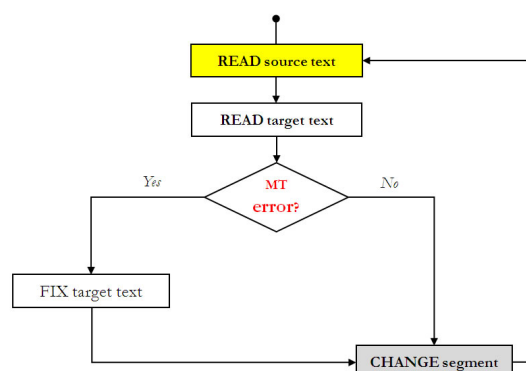


Figure 22: Post-editing style 2

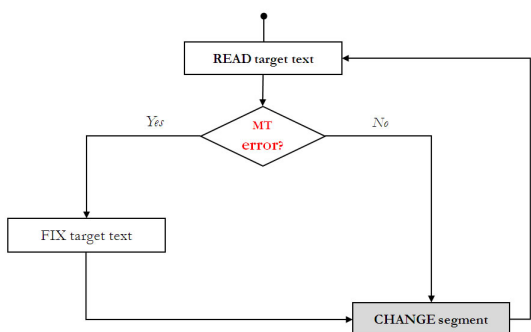


Figure 23: Post-editing style 3

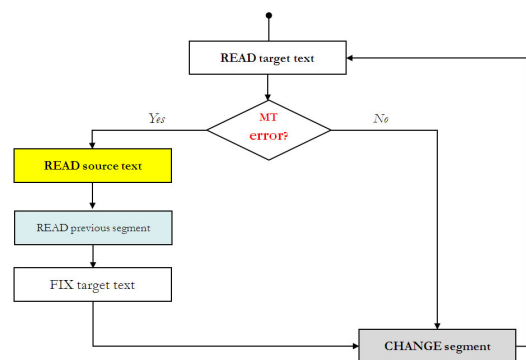


Figure 24: Post-editing style 4

Table 8 presents the distribution of these four styles across the participants and the three system configurations in the second CASMACAT field trial. No information is provided for participants 1, 3 and 4 due to the poor quality of the gaze data collected.

*: predominant style, •: style also present

Either style 1 or style 3 (monolingual post-editing) seem to be predominant regardless of the system involved in the post-editing process.

3.2 Backtracking moves

The data collected during the second CASMACAT field trial also reveals interesting differences on how different post-editors move between the different segments to post-edit a whole text. The

	Style 1			Style 2			Style 3			Style 4		
	P	PI	PIA	P	PI	PIA	P	PI	PIA	P	PI	PIA
P01												
P02	*	*	*	•	•	•	•			•	•	•
P03												
P04	•	*	*				*	•	•	•	•	•
P05	•	•	•				*	*	*	•	•	•
P06												
P07	*	*	*				•	•	•	•	•	•
P08	*	*	*	•	•	•				•	•	•
P09	•	•	•				*	*	*	•	•	•

Table 8: Post-editing styles identified in the gaze data

workflow in the CASMACAT workbench (prototype-II) is segment-based (the user is presented with segments to be post-edited one at a time). The user is thus able to backtrack and re-open previously edited/saved segments in order to review and/or revise them in the light of their work on later segments.

A number of different types of backtracking have been identified. Broadly these can be divided into local backtracking (where the user backtracks to within up to 4 segments from the current segment) and long-distance backtracking (where the user hops to segment(s) further away from the current segment). These two types of backtracking can be further refined as follows (letters of the alphabet are used here to indicate the patterns of backtracking):

3.2.1 Local backtracking

- **Immediate repetition:** the user immediately returns to the same segment. In our data this can occur up to 4 times (e.g. AAAA).
- **Local alternation:** the user switches back and forth between adjacent segments, often singly (e.g. ABAB) but also for longer stretches (e.g. ABC-ABC).
- **Local orientation:** this is characterized by very brief reading of a number of segments in order before returning to each one and editing them (e.g. ABCDE-ABCDE).

3.2.2 Long-distance backtracking

- **Long-distance alternation:** the user switches between the current segment and different previous segments (e.g. JCJDJFJG)
- **Text final backtracking:** the user backtracks to specific segments after having edited all the segments at least once.
- **In-text long distance backtracking:** instances of long distance backtracking as the user proceeds in order through the text.

The reasons for backtracking can either be to check previous segments to help the user understand and/or edit the current segment or, alternatively, to revise a previously edited segment in the light of edits made in the current section.

Table 9 shows the total number of the different types of local backtracking per participant in the second CASMACAT field trial.

The long-distance backtracking is by its nature more complex than local backtracking and deserves more detailed analysis into the numbers and length of distance between the segments. Nevertheless from the current data we are able to identify the following different post-editing styles or strategies.

Participant	Immediate repetitions	Local alternation	Local orientation
01	7	13	2
02	9	17	1
03	4	11	-
04	4	33	-
05	10	49	-
06	2	7	-
07	6	34	-
08	2	14	1
09	3	28	1

Table 9: Local backtracking moves per participant

3.2.3 Post-editing strategies

Unsurprisingly the overwhelming tendency is for post-editors to begin with the first segment and proceed in order to the last segment. As can be seen from Table 9, all post-editors engaged in some backtracking and, clearly, the amount and type of backtracking depends on issues arising in the particular text being post-edited. However, within these constraints it is possible to identify the following post-editing styles:

- **Only local backtracking:** The post-editor concentrates on the local context only and performs only local backtracking proceeding from the first to the last segment and only rarely, if ever, backtracking over a long distance (due to a particular issue in the text). For example in eight out of the nine sessions participant 01 only performed local backtracking. Participants 02 and 04 exhibited the same tendency whilst participants 05 and 03 did so in the sessions carried out at the Celer Soluciones SL offices although in sessions at home some long distance backtracking occurred.
- **Text final long distance backtracking:** The post-editor goes through all segments to the end performing only local backtracking. Long-distance backtracking then occurs after all the segments have been edited at least once. Participants 06, 08 and 09 employed this strategy
- **Mixed in-text backtracking:** Although the overall progress of the task is from the first to last segment, the post-editor performs both local and long-distance backtracking throughout the editing process including long-distance backtracking after the final segment is edited. This was the strategy adopted by participant 07 who also tended to go through all segments again in order at least once and make final corrections.

3.2.4 Conclusions

The main difference in post-editing styles in terms of backtracking revolves around whether the post-editors concentrate only on local context and eschew a final review/revisions phase once the whole text has been edited or not. However a number of factors may affect this choice. First of all, it should be noted that all the texts are short (with a maximum of 63 segments in the longest text) and thus at least some post-editors may be expected to be able to retain information about previous segments without needing to backtrack. It would be interesting to investigate the effect of longer texts on the strategies employed by the post-editors. The nature of the text and any linguistic problems will also have an effect on the necessity for the post-editor to go back and revise previous segments. This requires a detailed investigation of the linguistic features of the texts where backtracking occurs. For example one segment which caused a post-editor (who normally did not backtrack) to backtrack several times included a metaphor which appeared to cause problems.

4 Cognitive modelling (Task 1.5 - ongoing)

This section presents three different perspectives aiming at developing a model of the post-editing process and the individual differences between translators. Modelling translators behaviour can inform further implementation of the CASMACAT workbench in order to make it as adaptable as possible as well as tailor the type of assistance needed.

4.1 Tracing literal translation alignment in the translator’s mind

A previous deliverable (D5.3: *Representation in the Translation Process Research DB*) describes the *cross value* as a feature of the TPR database which represents word alignment information in a procedural manner. By following the ST-TT alignment links, the cross value indicates the minimum number of successive words towards the left or the right that need to be scanned to find the translation of the next target word in the source text. Languages with similar word order will have low average cross values. In a monotonous 1-to-1 translation all cross values are 1. The more syntactic re-ordering between source and target text take place, the higher the average of cross value will be.

Appendix A.5 describes the correlation of Cross values with gazing times. It was found that translators produce a cognitive alignment between source and target text representations during the translation process. Translators apparently cope with differences in word order by mentally aligning word meanings of the two languages following alignment links.

This process is more effortful for higher Cross values than in a monotonous word-to-word translation situation as indicated by longer gaze durations. The insights gained from this investigation lead to a revision of the monitor model and the literal default translation hypothesis (Tirkkonen-Condit 2005). The literal translation hypothesis assumes that an automatic translation procedure produces default translations in the translator’s mind, on the basis of shared representations (horizontal translation processes). Vertical processes interrupt this automatic text generation if the resulting target text is not acceptable.

The revised monitor model consists of a recursive cycle which integrates horizontal and vertical translation processes from source to target as well as from target to source: the monitor assesses whether the source text corresponds to the target text, but it is equally important to make sure that the target text is equivalent to the source text. Vertical processes access the output from the automatic default procedure recursively in both the source and the target language and monitor consistency as the context during translation production increases. This process can be observed during translation and during post-editing as the investigation of the Cross feature suggests. See (Schaeffer and Carl, in print) for a more detailed description of this study.

4.2 Psycholinguistic understanding of translation error detection

The initial stage of the post-editing process involves translators reading the machine translated output and evaluating its veracity. However, remarkably little is known about the cognitive processes involved in checking for lexical, syntactic and semantic violations in this type of task. The literature that does exist is almost exclusively in proofreading by native language speakers and limited to simple typographical errors arising from mistyping. For a complete understanding of the processing conducted by translators it is therefore necessary to establish clear baselines by examining how monolinguals perform in this reading environment. Replication of the findings for the simple typographical cases in monolingual English speakers will help validate the methodology employed and enable "levels of difficulty" to be determined between different classes of errors as well as differences between native and non-native language processing (L1 vs. L2).

Classifying types of errors

For simplicity and to maintain strict experimental control, there was a maximum of one error in each sentence. However, there are many different possible types of errors that can arise in written translations and five classes were defined to cover this range. Some are clearly objectively determinable, such as spelling mistakes or grammatical errors, while others are more subjective and may depend upon individual differences and style.

TE) Transposition (easy). This was considered the easiest error to spot and was essentially to provide a baseline for comparing the other error types against.

- (1) Picasso said that good artists ocpy [copy], great artists steal.
- (2) [Picasso sagte, dass gute Künstler kopieren, großartige Künstler klauen.]

TD) Transposition (difficult). While also a letter transposition, these examples involved two internal letters being switched (harder to spot than incorrect characters at the beginning or end of words) that produced an incorrect but legitimate word (and would therefore pass a spell-check or cursory examination).

- (3) I have decided to write all my deepest thoughts in a dairy [diary] again.

WO) Word order. Again, this was a transposition error, but at the word level rather than letter level. To minimise deviation in distance or alignment order, it was always the case that two adjacent words were switched rather than a random order reassignment (word salad).

- (4) Mostly were affected [affected were] the vegetable, corn and chickpea crops.
- (5) [Betroffen waren vor allem der Gemüse-, Mais- und Kichererbsenanbau.]

MT) Mistranslation of tense or agreement. These sentences contain a within-sentence violation in verb tense or a mismatch in gender or number agreement.

- (6) Many of our friend [friends] are surfers and I have a great friend who lives in Tamarindo.
- (7) [Viele unserer Freund sind Surfer und ich habe einen groartigen Freund, der in Tamarindo lebt.]
- (8) The cuts were [would] ultimately hit the combat troops.
- (9) [Die Kürzungen würden letztendlich die Kampftruppen treffen.]

ML) Mistranslated lexical item. In these sentences, the critical words were related or semantically connected to the correct word, but contextually odd or inappropriate.

- (10) Judge Torkjel Nesheim canceled [interrupted] Breivik during his monologue.
- (11) [Richter Torkjel Nesheim unterbrach Breivik während diesem Monolog.]

Experiment materials and design

The original materials for four of the error conditions were drawn from the German-to-English Machine Translation Marathon 2012 (MTM12) competition dataset [http://matrix.statmt.org/matrix/systems_list/1692]. This dataset comprised 3000 German input sentences with corresponding English reference translations, and machine translation output for each of the 19 systems that entered the competition. Emphasis was placed on using the Edinburgh submission [http://matrix.statmt.org/matrix/output/1692?run_id=2517], thereby utilising a project-related corpora and ensuring authentic stimuli.

24 sentence frames were constructed for each of the five error conditions. Each item had two variants: a correct version and a version where one word was the primary source of an error.

Two item lists were created. For each error condition in List A, the first half of the sentences were displayed correctly (i.e. no error) while the second half contained one appropriate error; List B was the matched reciprocal set. There were therefore 12 examples containing errors in each of the 5 conditions and 12 examples that did not contain a mistake, totalling 60 sentences with error and 60 without. Word frequency, error location and word length of the target error words were balanced across the two versions of the materials. Both lists included the same 60 filler sentences extracted from native-English corpora (i.e. fluent, error-free sentences), making 180 sentences in total, and 4 practice items. Participants were presented with this mixture of error-containing and error-free sentences in random order.

Sentences were displayed on the central line of a 22-inch widescreen monitor. Eye movements were recorded using an SR Research EyeLink 2K (binocular recording; 1KHz sample rate per eye).

A pre-experiment questionnaire (see Appendix) covered demographic details and linguistic background, as well as familiarity with machine translation (e.g. Google Translate) and personal usage.

Procedure for participants

After completing their background questionnaire, participants were handed an instruction sheet and told that they would be shown a series of sentences that had been automatically translated by a computer from German into English. The task was divided into two stages. The first was to read the sentence in its entirety and then to click the left mouse button if there was something incorrect about the translation or the right button if there was nothing wrong. If the participant decided there was a mistake then they next had to click on the error word to identify the location, but did not have to rectify the mistake. Following the set-up and calibration of the eye-tracker, participants were given four practice items with full instructions displayed between each practice trial. Each of the 180 experimental trials would only commence after fixating a contingent trigger that corresponded to the position of the first word in the sentence, ensuring accurate and reliable eye-movement data. This also means that participants controlled the display of each new item and could rest when required. Recalibrations were performed if targeting the contingent marker failed.

Data and statistics

Trial-by-trial data was collected, including yes/no participant decisions, error word identification and sentence reading times. Additionally, high resolution eye-movement data was recorded to enable a fine-grained investigation of the dynamic and online human processing involved in performing the task (spatial accuracy: 0.25° to 0.5° visual angle; sampling rate: 1 millisecond).

For the purposes of the analyses reported here, the data used are contingent on whether participants made a correct detection response (error or no error). Other datasets are also available: all trials irrespective of response, trials where the error was missed, and false positive trials (error detected when there was none). Mixed-design Analyses of Variances were the primary tests for statistical significance. Sentence List (A or B) was included in these analyses as a dummy between-subjects variable to verify that the two lists were equivalent. Newman-Keuls post-hoc comparisons were used in the breakdown of significant multi-level ANOVA main effects or interactions. On graphs, the vertical bars denote 0.95 confidence intervals. Global analyses refer to sentence-wide effects while localised analyses involve effects at the sub-sentence level. Three sets of localised analyses were conducted for each of the eye-movement dependent variables: measures on the critical target word N alone (i.e. error location); the subsequent "spillover" word, N+1; and a combined two-word region of interest, N & N+1 (particularly relevant for the word order errors where the two words swapped position). Different patterns in these are indicative of whether shallow processing is sufficient to identify a problem (immediate disruption upon fixating the word), whether further context is required or integration issues arise (processing difficulty occurs further downstream) or whether wider re-reading and regressive movements are required.

Eye-Movement Dependent Variables In addition to the total reading times for each sentence, mean fixation durations and mean fixation counts were calculated. Saccadic amplitude was also recorded: short, regular movements are indicative of problem-free serial processing of a sentence; while more variable, larger movements are symptomatic of the need to access parts of the sentence out of order and/or re-reading. Potential blink-related and pupillometric indicators of processing load were also examined: mean pupil dilation, maximum dilation, number of blinks and blink rate (blinks per minute).

A range of further reading measures were utilised for determining localised effects on or around the target word:

- **First Fixation Duration.** The duration of the first fixation made on a target region. This commonly reflects fundamental aspects of reading such as lexical access as this is a measure of immediate processing.
- **Gaze Duration.** The sum of all fixations made on a region before the eyes moves onto another region (either progressive or regressive).
- **Right-Bounded Duration.** The sum of all fixations made in a region until the eyes move rightwards out of it.
- **Regression Path Duration.** The total duration of all fixations that occurred from the first fixation on a target region until the target region was exited in a progressive manner. This is similar to Right-Bounded Duration but includes any time spent re-reading earlier text in the sentence (not just the fixations on the region itself but also regressions up to that point). This tends to reveal the total integration cost at the critical point of a sentence.
- **Re-reading Duration.** The difference between Regression Path Duration and Gaze Duration. Essentially the amount of time spent re-reading earlier words as a consequence of reading the target region.
- **Total Duration.** The total duration of all fixations in a target region, irrespective of the order they occurred in.

As well as these reading time variables, there were three fixation ratio measures used to examine gaze behaviour. These are spatial descriptions of reading patterns rather than temporal ones:

Error Type	Mean	Percent
TE	11.85	98.75
TD	8.90	74.17
WO	10.65	88.75
MT	10.50	87.50
ML	8.25	68.75
False Positives	16.85	14.04

Table 10: Detection rate by error type for monolinguals. Maximum score of 12 for each error class and 120 for false positives.

- First-Pass Regression. Whether the first saccade out of the target region was leftwards (regressive) rather than progressive.
- First-Pass Fixation. Whether the target region was fixated before any progressive region of text was fixated (i.e. it was not skipped during first-pass reading).
- First-Pass Multi-Fixation. Whether multiple fixations were made on the target region before the eyes move away from it.

4.2.1 Detection by Monolinguals

Participants 20 monolingual native English speakers were recruited through the University of Edinburgh careers service (11 Male, 9 female; mean age 23.05). All participants gave written consent and received £ 10 compensation for taking part. Half were tested with the List A sentences while the other half were given the reciprocal List B.

Global Trial Effects Response decisions are listed in Table 4.2.1. Each participant read 12 examples of each error type and could therefore score a maximum of 12 in each category. The number of false positives made is also listed. This refers to when participants decided there was something wrong with a sentence when it was in fact correct. There was a maximum of 120 instances of this. The scores indicate a generally good detection rate, but there were significant differences across the different error types [$F(4,72) = 26.026$, $p < 0.001$]. A post-hoc analysis of all the pair-wise comparisons showed reliable differences between them all, with only two exceptions: TD & ML (the two lowest scoring sets) and WO & MT.

The global trial reading measures for each of the error types is given in Table 4.2.1, along with a summary of the ANOVA results in Table 4.2.1. The presence of an error was strong enough to induce a greater average fixation duration even at the sentence level (239ms vs 208ms). This was reliable for every error type, averaging a 15% increase. However, this simple pattern is not directly reflected in the total trial duration or the total number of fixations. An error resulted in longer times (with more fixations) in the ML condition (10329ms vs 7135ms), shorter times (and fewer fixations) in the TE condition (6160ms vs 7041ms) but no significant difference between the other three. This pattern was also present in the blink count, but it disappears when duration is taken into account (i.e. there was a correlation between the number of blinks made and the length of time taken to read a sentence rather than anything meaningful). There are also shorter saccades made for every error type except ML (presence of an error makes no reliable difference in this case). No evidence of any pupillometry effects was found.

So, detecting an error reliably slows down eye movements with the eyes holding longer at the points they are fixating. But this remains independent of other processing consequences and does not mean that sentences containing an error necessarily take longer to read than the same version without an error. In other words, longer fixations do not automatically lead to an

	Error Type				
	TE	TD	WO	MT	ML
Fixation Duration	229	225	225	220	219
Saccadic Amplitude	2.87	3.17	3.02	3.27	3.17
Blink Count	2.35	2.74	3.15	3.01	3.22
Trial Duration	6600	7934	8163	8075	8732
Fixation Count	24.97	29.81	30.64	30.73	33.04
Mean Pupil Size	725	823	598	862	878
Max Pupil Size	651	640	630	649	647
Blink Rate	23.27	23.86	25.45	24.58	25.41

Table 11: Monolingual global reading measures for each error condition.

	Error Type		Error Present		Interaction	
	F(4, 72)	P-value	F(1, 18)	P-value	F(4, 72)	P-value
Fixation Duration	5.7602	0.000	98.587	0.000	3.384	0.014
Saccadic Amplitude	25.228	0.000	21.718	0.000	9.900	0.000
Blink Count	6.495	0.000	2.477	0.133	12.251	0.000
Trial Duration	20.967	0.000	4.205	0.055	14.257	0.000
Fixation Count	29.321	0.000	0.010	0.923	13.896	0.000
Mean Pupil Size	1.184	0.325	1.013	0.328	0.723	0.579
Max Pupil Size	0.456	0.768	0.136	0.717	1.850	0.129
Blink Rate	2.400	0.058	1.888	0.186	2.790	0.033

Table 12: ANOVA summary of monolingual global reading measures. Main effects of Error Type (TE, TD, WO, MT, ML), Error Present (Yes, No) and the interaction of these two variables.

increase in overall reading time. Errors are disrupting usual reading patterns, however, not only in terms of fixation duration but also in the distance between saccades in the majority of cases. Deciding that a semantically related but inappropriate word is present does increase the reading time for a sentence whereas spotting a simple typographic error actually reduces the total time spent reading the sentence. It is likely that there is more re-reading of the entire sentence in the ML condition (second pass reading) hence longer times, more fixations, but a similar scanning pattern (saccade amplitude); i.e. repeated "normal" reading of the same sentence for comprehension. This seems a reasonable hypothesis for higher-level error detection/evaluation. In contrast, TE errors are easy to spot immediately. These may result in faster (skimming) of subsequent text and no need to re-read sentence. Testing for localised effects may help provide a clearer picture of the changes in reading behaviour.

Local Effects The reading measures for each of the error types on the target word N, subsequent word N+1, and the conjoined region N & N+1, is given in Table 4.2.1, along with a summary of the ANOVA results in Table 4.2.1. The First Fixation times reveal a significant immediate impact for the errors MT (191ms vs 154ms), TD (230ms vs 196ms), and especially TE (268ms vs 198ms), but not so for WO (186ms vs 161ms) or ML (190ms vs 183ms). It appears that tense or agreement integration violations are immediately obvious while problems with plausible semantic integration are not. There were only longer first fixations on the second word (N+1) for the Word Order condition. Initially, this makes sense due to the nature of the word transposition, but there is in fact no interaction with the presence of an error.

Both letter transpositions induce much longer gaze durations on the target word itself (TE: 639ms vs 262; TD: 376ms vs 247ms). This is strong enough to persist across the two-word region and extends here to include the WO cases, this time conditional on the presence of an error (632ms vs 480ms). The pattern for the spillover-only region (N+1) is identical to the first fixation results: nothing significant except for an overall WO effect. The Right-Bound Duration on the critical word (the total time spent on the word until the eyes move onto new text) shows longer times when there is an error for every condition except ML. Again, letter transposition produces the strongest effect, both Easy (762ms vs 307ms) and Difficult (523ms vs 315ms).

	Word N					Word N+1					Words N & N+1				
	TE	TD	WO	MT	ML	TE	TD	WO	MT	ML	TE	TD	WO	MT	ML
First Fixation	233	213	174	173	186	135	128	189	134	134	245	225	210	206	216
Gaze Duration	447	329	254	212	233	164	145	250	162	184	646	561	556	401	451
Right-Bounded	534	419	284	244	273	196	185	320	192	240	886	755	710	495	643
Regression Path	642	510	345	315	392	506	469	502	294	577	1148	979	847	609	968
Re-reading Duration	210	192	108	121	184	549	580	304	215	582	506	420	297	214	527
Total Duration	1343	1658	1328	1237	1332	361	371	920	556	644	1703	2029	2247	1793	1976
First-Pass Regression Probability	0.254	0.259	0.225	0.185	0.211	0.387	0.412	0.318	0.315	0.316	0.350	0.292	0.302	0.236	0.267
First-Pass Fixation Probability	0.931	0.927	0.801	0.802	0.847	0.633	0.595	0.873	0.641	0.649	0.991	0.991	0.977	0.967	0.985
First-Pass Multi-Fixation Probability	0.594	0.472	0.457	0.367	0.377	0.275	0.310	0.480	0.380	0.413	0.819	0.726	0.876	0.726	0.717

Table 13: Monolingual reading measures for the target word (N), subsequent word (N+1) and the combined two-word region. Times are in milliseconds.

	Word N						Word N+1						Words N & N+1					
	Error Type		Error Present		Interaction		Error Type		Error Present		Interaction		Error Type		Error Present		Interaction	
	F(4, 72)	P	F(1, 18)	P	F(4, 72)	P	F(4, 72)	P	F(1, 18)	P	F(4, 72)	P	F(4, 72)	P	F(1, 18)	P	F(4, 72)	P
First Fixation	24.055	0.000	42.209	0.000	3.383	0.014	14.253	0.000	0.057	0.814	0.837	0.506	12.409	0.000	17.873	0.001	2.231	0.074
Gaze Duration	45.857	0.000	95.490	0.000	32.549	0.000	18.177	0.000	0.346	0.564	0.955	0.438	17.783	0.000	53.509	0.000	11.830	0.000
Right-Bounded	47.346	0.000	124.529	0.000	28.824	0.000	15.945	0.000	1.507	0.235	3.073	0.021	11.368	0.000	107.167	0.000	0.465	0.761
Regression Path	21.228	0.000	69.065	0.000	6.735	0.000	4.968	0.001	30.316	0.000	3.123	0.020	11.368	0.000	107.167	0.000	0.465	0.761
Re-reading Duration	2.694	0.038	2.072	0.167	0.987	0.420	6.198	0.000	26.637	0.000	2.322	0.065	6.352	0.000	20.363	0.000	4.488	0.003
Total Duration	9.859	0.000	165.200	0.000	1.805	0.137	31.634	0.000	32.634	0.000	28.093	0.000	9.509	0.000	152.864	0.000	6.649	0.000
First-Pass Regression Probability	1.724	0.154	4.085	0.058	3.425	0.013	3.003	0.024	51.190	0.000	1.678	0.165	3.208	0.018	15.482	0.001	3.361	0.014
First-Pass Fixation Probability	13.271	0.000	16.049	0.001	1.648	0.172	22.805	0.000	0.066	0.800	0.419	0.795	3.195	0.018	3.204	0.090	0.989	0.419
First-Pass Multi-Fixation Probability	13.236	0.000	34.799	0.000	3.890	0.006	7.611	0.000	107.167	0.000	0.465	0.761	16.736	0.000	41.792	0.000	2.535	0.047

Table 14: Summary of ANOVA results for monolingual localised effects.

There were clear effects for all five error types over the larger two-word region. However, there was a lack of any effect for four of the five error types (same pattern as First Fixation and Gaze Duration) on Word N+1 alone, with a difference between the error and no-error cases only evident in the Word Order condition (354ms vs 286ms).

Regression Path (includes any leftwards fixations into previous text before the eye moves into new text to the right of the region) for the target word only demonstrates the same pattern of results as the Right-Bound Duration. Again, there were clear effects for all 5 error types over the two-word region, although the easy transposition is no longer as dramatically different from the other cases. On the spillover word itself, there is only a difference in the ML case (794ms vs 360ms).

Re-reading Duration (leftward regression times) indicates that no significant time was spent making regressions as a consequence of reading the target word itself. There are clear regression effects only for the ML conditions in the two-word and spillover-only analyses, which corroborates the Regression Path finding. The lexical mistranslation is therefore being detected primarily on the subsequent N+1 word (spillover region) resulting in immediate regressions to previous text but not any hesitation. It seems that this type of problem is detected late, having already moved on to the next word, but once identified then the re-reading of the sentence begins immediately.

Total Duration indicates a big overall effect arising on an error word: the total time spent on the target word is about 4 times as long, on average, as when there is no error (2107ms vs 652ms). The effect remains highly significant in the two-word analysis, although there is only an increase in the size of the effect for WO; given that both words can be considered critical in this case, this makes sense. The pattern is more complicated for the spillover word. An error results in longer times spent on Word N+1 for ML, MT and especially WO; there is no difference for TD; but there are actually shorter times for TE condition. The long instant impact of encountering an easy transposition error may allow some attention to shift and parafoveal processing of the next word to occur, thus reducing the time required to directly fixate it. This

Error Type	Mean	Percent
TE	11.50	95.83
TD	9.40	78.33
WO	10.85	90.42
MT	9.95	82.92
ML	7.65	63.75
False Positives	20.85	17.38

Table 15: Detection rate by error type for multilinguals. Maximum score of 12 for each error class and 120 for false positives.

is a good example of how specific localised fluctuations can average out over the course of a sentence and may never be identified if only global measures of reading are studied.

First-pass Regression Probability: only in the MT condition is there a higher probability of making a regression after fixating the error word (25.8% vs 11.1%). This is also the case for the two-word region, although WO almost achieves significance ($p=0.054$). There is also a greater likelihood of making a regression from the spillover word for these types of error, and it is almost significant for TD ($p=0.061$). Participants were more likely to fixate the target word during first-pass reading when it was an error (89.1% vs 83.2%) but this did not interact with error type. The two-word region was almost always fixated and there were no differences here (ceiling effect). The presence of an error did not affect the fixation rate of the spillover word although it was much more likely to be fixated in the WO condition even when it is the correct order (i.e. no error).

The First Pass Multi-fixation Probability did not quite exhibit the same as pattern as Gaze Duration. Here, only TE produced a higher likelihood of multiple fixations (75.2% vs 43.5%), not TD. Typically, multiple fixations and gaze duration are highly correlated. There was a greater probability of multiple fixations across the two-word region when there is an error in all cases except for WO. The presence of an error did increase the likelihood of making multiple fixations on the spillover word, but this difference was consistent across all types of error.

4.2.2 Detection by Multilinguals

Participants 20 multilingual non-native English speakers took part (6 Male, 14 female; mean age 30.2). Every participant had a European first language (L1) and English as their second language (L2), having spoken English for an average of 20.6 years. Seven were bilingual while 13 were fluent in three or more languages. Seven had experience or training in professional translation. All had used machine translation before with nine having experience of post-editing the output. All participants gave written consent and received 10 compensation for taking part. As before, half were tested with the List A sentences while the other half were given the reciprocal List B.

Global Trial Effects Response decisions are listed in Table 4.2.2. Again there was an overall high detection rate with clear differences between the different types of error [$F(4,72)=22.611$, $p<0.001$]. A post-hoc breakdown of this effect showed reliable differences between them all with two exceptions: TE & WO and TD & MT.

Table 4.2.2 lists the global trial reading measures for each of the error types, with a summary of the ANOVA results contained in Table 4.2.2. The presence of an error somewhere in the sentence is enough to significantly increase the average fixation duration across the entire sentence (10% increase: 233ms vs 212ms). This was consistent across all five types of error. There was also a main effect of Error on mean saccade amplitude, with eye movements being

	Error Type				
	TE	TD	WO	MT	ML
Fixation Duration	229	222	223	218	220
Saccadic Amplitude	2.91	3.13	3.10	3.28	3.33
Blink Count	3.00	4.43	3.78	3.86	4.68
Trial Duration	7472	9409	9118	8970	10301
Fixation Count	28.17	35.81	34.43	33.61	38.77
Mean Pupil Size	660	666	661	666	662
Max Pupil Size	737	787	558	977	662
Blink Rate	26.74	32.37	27.20	28.44	30.46

Table 16: Multilingual global reading measures for each error condition.

	Error Type		Error Present		Interaction	
	F(4, 72)	P-value	F(1, 18)	P-value	F(4, 72)	P-value
Fixation Duration	6.942	0.000	41.650	0.000	2.876	0.029
Saccadic Amplitude	24.516	0.000	14.462	0.001	9.508	0.000
Blink Count	5.680	0.000	0.140	0.713	4.601	0.002
Trial Duration	10.548	0.000	0.007	0.935	7.928	0.000
Fixation Count	17.341	0.000	1.713	0.207	11.274	0.000
Mean Pupil Size	0.630	0.643	31.935	0.000	0.772	0.547
Max Pupil Size	1.749	0.149	0.117	0.736	1.467	0.221
Blink Rate	2.364	0.061	0.785	0.387	1.153	0.339

Table 17: ANOVA summary of multilingual global reading measures. Main effects of Error Type (TE, TD, WO, MT, ML), Error Present (Yes, No) and the interaction of these two variables.

6.2% shorter on sentences with an error compared to those without one. However, a breakdown of the interaction indicates that this effect only holds for the TE, TD and WO conditions, with the mistranslated cases (MT and ML) not exhibiting any difference in this type of reading pattern.

The presence of an error did not lead to longer reading times overall (9039ms vs 9069ms), although there was an interaction with sentence type. This arose from two balanced but reliable differences: ML sentences with errors were slower to read than any other type of sentence while TE sentences with errors were faster. An identical pattern of results was found for the mean number of fixations made. Similarly, there was an interaction without a main effect of Error on the number of blinks, but in this case it was driven entirely by a reduction in the number for sentences containing an incorrect initial bigram (TE). And even this disappeared when the total number was normalised by reading time.

A small, but significant pupillometry result was found. Sentences containing errors stimulated a 2% increase in the maximum dilation of the eyes. This effect was transitory as it did not translate into an increase in average pupil size across the entire sentence. Errors therefore appear to lead to longer individual fixations but not necessarily longer overall reading times for sentences. They also altered the pattern of eye movements in three of the sentence frames tested but not the two mistranslation structures. It is most likely that the MT and ML errors result in rereading rather than more immediate first-pass disruption. There is also some evidence to suggest that they can lead to greater pupil dilation in multilinguals, often an indicator of increased cognitive demand.

Local Effects Table 4.2.2 summarises the reading measures for each of the five error conditions on the target word N, subsequent word N+1, and the combined two-word region N & N+1. Table 4.2.2 summarises the ANOVA output for each Dependent Variable. There was a main effect of both Error Type and Error Presence on First Fixation times. However, the interaction between these variables revealed that only the most blatant error, the initial letter transposition, produced a first fixation effect in the bilingual participants (TE with error: 272ms; TE without: 190ms). Although the average first fixation was slightly longer due to an

	Word N					Word N+1					Words N & N+1				
	TE	TD	WO	MT	ML	TE	TD	WO	MT	ML	TE	TD	WO	MT	ML
First Fixation	231	218	191	183	203	152	132	189	143	224	243	228	222	207	274
Gaze Duration	474	380	284	259	308	181	162	261	185	282	736	640	592	473	628
Right-Bounded	571	434	318	288	359	219	204	321	216	336	992	776	728	559	841
Regression Path	709	520	374	334	452	681	446	460	307	899	1390	966	834	640	1351
Re-reading Duration	249	152	109	89	161	675	441	238	185	762	662	329	246	173	734
Total Duration	1403	1810	1401	1200	1421	393	500	1115	624	850	1796	2310	2516	1824	2271
First-Pass Regression Probability	0.258	0.183	0.194	0.154	0.178	0.361	0.389	0.281	0.220	0.258	0.359	0.225	0.233	0.194	0.231
First-Pass Fixation Probability	0.935	0.941	0.855	0.861	0.895	0.665	0.649	0.849	0.666	0.746	0.989	0.986	0.989	0.975	0.986
First-Pass Multi-Fixation Probability	0.691	0.550	0.497	0.456	0.471	0.324	0.375	0.542	0.443	0.388	0.889	0.818	0.871	0.799	0.842

Table 18: Multilingual reading measures for the target word (N), subsequent word (N+1) and the combined two-word region. Times are in milliseconds.

	Word N						Word N+1						Words N & N+1					
	Error Type		Error Present		Interaction		Error Type		Error Present		Interaction		Error Type		Error Present		Interaction	
	F(4, 72)	P	F(1, 18)	P	F(4, 72)	P	F(4, 72)	P	F(1, 18)	P	F(4, 72)	P	F(4, 72)	P	F(1, 18)	P	F(4, 72)	P
First Fixation	7.179	0.000	28.440	0.000	3.349	0.014	2.008	0.103	2.087	0.166	0.787	0.537	0.826	0.513	3.407	0.081	1.153	0.339
Gaze Duration	35.992	0.000	65.504	0.000	28.080	0.000	3.271	0.016	2.734	0.116	0.706	0.590	5.174	0.001	62.115	0.000	4.943	0.001
Right-Bounded	38.374	0.000	87.180	0.000	23.840	0.000	4.438	0.003	4.751	0.043	0.999	0.414	10.355	0.000	57.581	0.000	3.178	0.018
Regression Path	31.100	0.000	48.601	0.000	8.411	0.000	8.589	0.000	2.808	0.111	1.981	0.107	15.918	0.000	16.512	0.001	1.271	0.289
Re-reading Duration	5.350	0.001	3.519	0.077	1.918	0.117	10.329	0.000	1.675	0.213	0.986	0.421	14.438	0.000	0.059	0.811	3.355	0.014
Total Duration	14.384	0.000	173.210	0.000	0.808	0.524	27.999	0.000	25.197	0.000	10.886	0.000	12.696	0.000	114.410	0.000	5.746	0.000
First-Pass Regression Probability	2.395	0.058	1.874	0.188	1.958	0.110	6.707	0.000	47.962	0.000	1.850	0.129	7.602	0.000	5.771	0.027	2.555	0.046
First-Pass Fixation Probability	10.533	0.000	7.305	0.015	0.813	0.521	15.152	0.000	0.234	0.634	0.644	0.633	0.914	0.461	0.102	0.753	0.400	0.808
First-Pass Multi-Fixation Probability	15.630	0.000	11.876	0.003	5.343	0.001	8.581	0.000	1.982	0.177	1.496	0.213	4.663	0.002	6.922	0.017	2.286	0.068

Table 19: Summary of ANOVA results for multilingual localised effects.

error for all the other classes of sentence (in the order of 10 to 15 milliseconds for each) this was non-significant. Any sign of an immediate delay on processing was absent on the spill-over word or the combined two-word region.

There was a slightly stronger effect in Gaze Duration (this includes multiple fixations if they occurred, rather than just the first) on the critical word, with both types of letter transposition leading to significantly longer times compared to any other sentence structure (TE: 675ms vs 274ms; TD: 441ms vs 319ms). This effect remained over the combined two-word analysis region but did not persist in the spill-over word alone. This was matched by the Right-Bound Duration results with the addition of a processing delay for the Mistranslated Lexical sentences in the combined two-word region (1028ms vs 654ms).

For the Regression Path, which includes time on the word plus any regressions until a progressive eye movement is made onto new text, the results are much more similar to the initial First Fixation findings. On the target word itself there is a very strong effect of encountering an incorrect initial bigram (TE: 935ms vs 483ms). Again, as with First Fixation results, the other errors did produce slightly longer reading times, but not significantly longer. The TE error effect was weak ($p=0.095$) when the subsequent word was included with the critical one, and disappeared altogether when the spill-over word was analysed on its own. However, there was evidence of increased times on the spill-over word (N+1) when there were errors in the ML sentences. So, this kind of semantic integration difficulty may arise just after the problem word itself is encountered, delaying the arrival of new progressive information. There was no evidence of an increase in the time spent immediately re-reading previous text after encountering an error.

There was remarkably similar increase in the total time spent reading an incorrect word of just over a second for all five error types (differences TE: 1111ms; TD: 1261ms; WO: 1359ms; MT: 1146ms; ML: 1218ms). Unsurprisingly, total times were also longer in the two-word region, although there was more variance between the types of errors. On the spill-over word alone, an error increased times for the WO (1491ms vs 739ms) and ML (1079ms vs 620ms) cases.

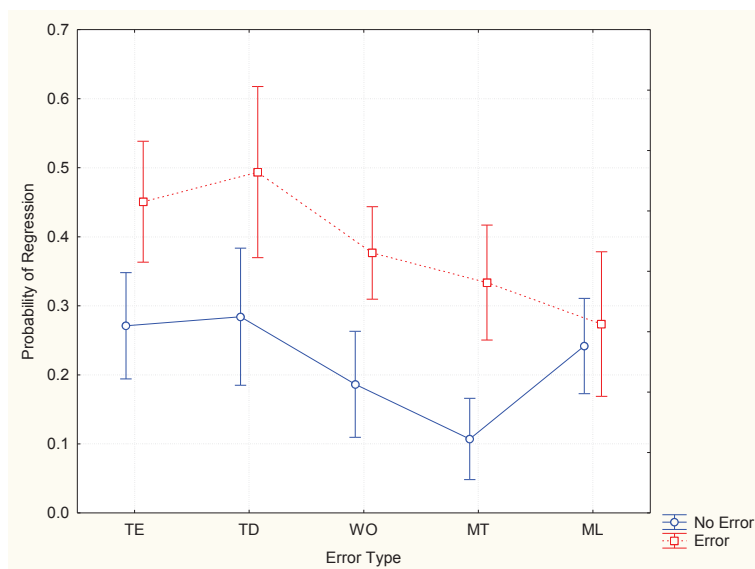


Figure 25: Probability of a multilingual participant making a regressive eye movement into previous text after fixating the spill-over word (target +1).

In parallel with the lack of any re-reading time increase, there was no evidence of an error triggering an immediate leftwards saccade into the previously read text (20.3% chance of a regression vs 18.4%). There is, however, an increase in regression probability after the spill-over word (N+1) is fixated, almost doubling the likelihood, for all but the ML sentences (see Figure 4.2.2). There was a small, significant overall increase in an error being fixated during first pass reading, but fixation probabilities were high anyway (main effect: 91.3% vs 88.2%). The probability of making multiple fixations on the first reading of an error word was only significantly increased for the TE sentences (82.8% vs 55.4%) which corresponds to the increase found in Gaze Duration.

4.2.3 Comparison between Mono and Multilinguals

Decision responses made by participants indicate remarkably little difference in spotting anomalies between the two linguistic populations overall [$F(1,36)=0.274$, $p=0.604$], displaying similar patterns (see Figure 4.2.3) across all five classes of error [$F(4,144)=1.3351$, $p=0.260$]. Multilinguals are therefore as good at detecting the full range of problems encountered in the English materials as native English speakers. While multilinguals did produce a slightly higher number of false-positive responses (mean 20.85 vs 16.85 out of 120), perhaps indicative of being over-cautious, this was not significant [$F(1, 36)=1.0786$, $p=0.30593$].

In terms of ranking the classes of error by successful detection, $TE > WO \geq MT > TD > ML$. (The difference between WO and MT was only marginally significant, $p=0.078$.) Overall reading times did not differ reliably between the two linguistic groups either (see Figure 4.2.3; $F(1,36)=1.666$, $p=0.205$). This was also the case for the set of 120 "filler" sentences [$F(1,36)=2.168$, $p=0.150$]. Other than a pupillometric response to an error for the multilinguals, both groups exhibited very similar patterns across the global trial analyses. Therefore, as far as end performance is concerned, participants scored consistently well and took a similar length of time, irrespective of whether they were native or non-native speakers of English. Ranking the errors in terms of reading and decision times showed TE to be the fastest, ML to be the slowest, but little difference between the other three.

At the local level, concentrating on gaze behaviour around the target word itself, more subtle differences between the linguistic groups emerge. Problems seem to emerge faster for the monolinguals compared to the multilinguals, with only one type of error producing longer

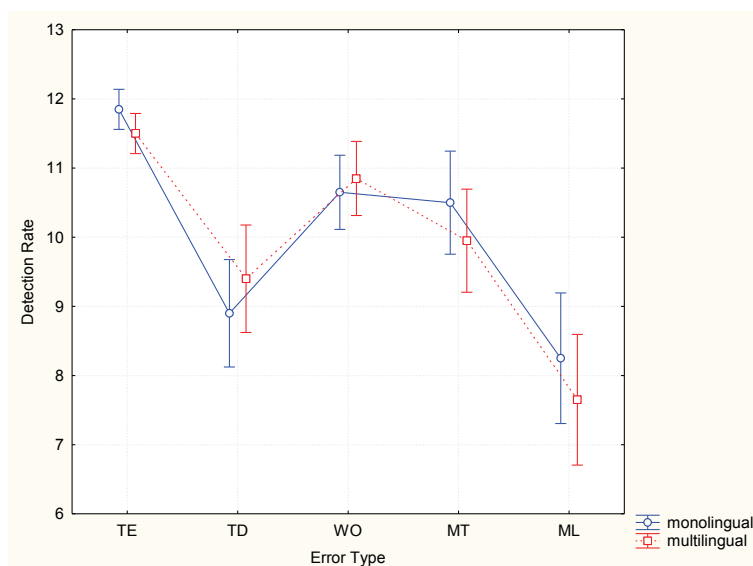


Figure 26: Detection rates for mono- and multilinguals across the five types of errors.

first fixation durations for the multilinguals rather than three. Neither was there any indication of another common sign of immediate processing difficulty, in the form of triggering an instant regression and re-reading. The single multilingual result was for the simplest baseline condition, TE, which would arise at the early lexical access stage of language processing, before there is any need for integration into a sentence. Additionally, the initial bigrams may also be rare or illegal in their own native language (L1) as well as in English (their L2), and so faster to respond to. However, there was evidence of stronger effects for the multilinguals when the word after the error was included in the analysis, suggesting that some stages of sentence processing may be slightly delayed (integration rather than word identification), producing a small disassociation between eye-movement control and sentence processing. There was, for example, a greater likelihood of making a leftwards regressive movement, but only two or more fixations after initially encountering the error.

It therefore appears that the task of detecting a range of different mistakes in text can be performed equally well by proficient, but non-native language speakers, as by educated native language speakers. While the overall results were similar there is some evidence to suggest that the point of detection was slightly later in L2, but that by the end of the sentence the consequences of this were minimal.

4.3 Modelling post-editing behaviour: an analysis of post-editing changes

In this section we include descriptive results for modelling post-editor profiles taking into account the type of changes and edits introduced by the nine participants in the second CASMACAT field trial.

The dataset selected for this analysis was dataset 3 (3,000 words), the one that all participants post-edited at the end of this field trial after having post-edited 6,000 words using the three systems (P, PI, and PIA). Investigating this particular dataset provides us with a clearer picture of the type of changes that post-editors introduce in the text once they are familiar enough with the three systems as well as to what extent each system modifies, if that is the case, the type of changes they introduce. Ultimately we wanted to know (i) whether it is possible to model post-editing behaviour based on the types of changes that post-editors introduce and (ii) whether the three systems involved in this second CASMACAT trial have an impact on the types of changes they introduce.

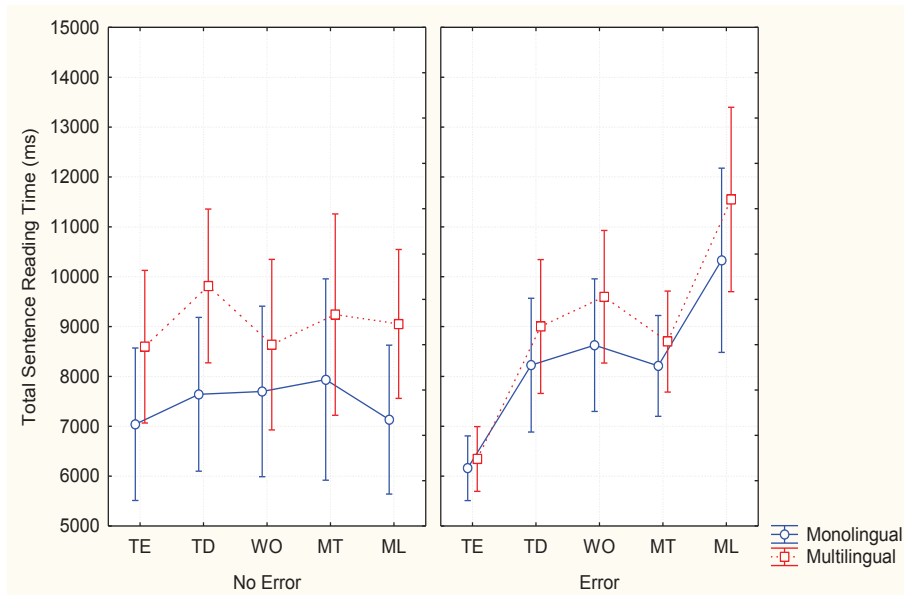


Figure 27: Overall mean reading times for the five sentence constructions, with and without errors, for both linguistic groups.

4.3.1 Typology for the classification of post-editing changes

A typology for the classification of post-editing data was devised for the present research. This typology derived from three main sources: de Almeida (2013), the LISA QA Model and the GALE Post-editing Guidelines. This typology is based in four master categories: (i) **essential changes**, (ii) **preferential changes**, (iii) **essential changes not implemented**, and (iv) **errors introduced**.

A change was considered as *essential* when, if the change is not implemented, the sentence (or part of it) is either:

1. Grammatically incorrect (i.e. it obviously breaches a grammatical rule), or
2. Grammatically correct, but not accurate in comparison to the source text (i.e. it does not contain all the information that is present in the source text, or it contains extra information that is not present in the source text).

Conversely, a change was considered *preferential* if the sentence from the raw MT output would still be grammatically correct, intelligible and accurate in relation to the source text, even if the change in question was not implemented. In order to differentiate essential and preferential changes, these two definitions were strictly followed in the analysis.

As well as accounting for the corrections made, it was also important to keep track of any essential changes not implemented by the post-editors. For this reason issues in the raw MT output that were not corrected by were also identified. When an essential correction was not implemented by a given participant in this field trial, it was counted as *Essential change not implemented*. Whenever several essential changes in the same segment were not implemented, they were also all counted as discrete occurrences.

Finally, a post-editing change was considered under the category *error introduced* if:

1. The error was not present in the raw MT output, and it was introduced by the post-editor while editing a sentence;

2. Because of it, the sentence (or part of it) is grammatically incorrect and/or inaccurate.

The category *introduced errors* caters for errors introduced by the post-editors (as opposed to errors that were present in the raw MT output). Examples might include (but are not limited to) typos and misspellings.

Table 20 presents details of the main categories and subcategories used to count the post-editing changes made for each of the four master categories in this study. The main categories in bold are from the LISA QA Model. The subcategories come from the GALE PE Guidelines are marked with the symbol *, while the subcategories devised by de Almeida (2013) are marked with the \$ symbol.

Main categories	Subcategories (<i>if applicable</i>)
Accuracy (completeness)	Extra information in MT output* Information missing from MT output* Untranslated text \$
Consistency	N/A*
Country	Decimal points* Quotation marks* Currency symbol* Date/time format \$
Format	N/A*
Language	Adjectives* Adverbs* Capitalisation* Conjunctions \$ Determiners* Gender \$ Nouns \$ Number \$ Phrasal ordering* Prepositions* Pronouns* Punctuation* Spelling* Verb tense*
Mistranslation	N/A*
Style	N/A*
Lexical Choice	N/A*

Table 20: Typology for the classification of post-editing changes borrowed from de Almeida (2013)

4.3.2 Results: Post-editing changes made in dataset 3

Table 21 summarises the number of post-editing changes introduced by the nine post-editors who took part in the second CSMACAT field trial.

Looking at the differences between the four master categories in the three different systems (P, PI, PIA), it can be seen that the number of errors introduced in dataset 3 overall were less in the case of one of the system featuring interactivity (PI). This is not the case, however, in the case of PIA, where the number of error introduced by the post-editors are twice as much as those introduced in the two other systems. Taking a closer look to the errors introduced in

Configuration	Essential changes			Preferential changes			Essential changes not implemented			Errors introduced		
	P	PI	PIA	P	PI	PIA	P	PI	PIA	P	PI	PIA
<i>Lexical choice</i>				123	144	107						
<i>Style</i>												
<i>Format</i>				3		1						
<i>Accuracy</i>	25	30	4									
<i>Mistranslation</i>	133	99	88				9	10	7		4	
<i>Language</i>	202	240	250	3		1	42	29	47	27	14	51

Table 21: Types of errors per category and system in dataset 3

PIA, these errors could easily have been avoided with a final revision using a spell checker since many of these errors are typos and missing punctuation marks (i.e. Example 1:[...] *violada por un sompañero* [...] - the post-editor deleted *soldado* to replace it for *compañero*, but finally types a hybrid non-existent word: *sompañer*; example 2: [...] *recuerdo decirle a mi mando* *No le* [...], after editing the text, the post-editor omits any punctuation mark).

Figure 28 show the distribution of post-editing changes made across participants in the second CASMACAT field trial.

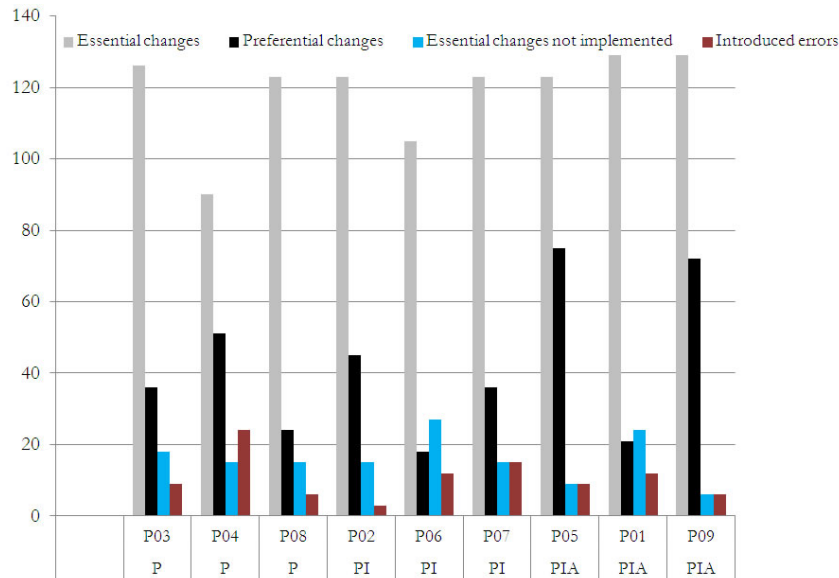


Figure 28: Overall count of post-editing changes per system and participant in dataset 3

5 Connections with the rest of the project and further CasMaCat development

As recommended in the first review report, this section addresses the connection of the findings of this workpackage with both the remaining tasks in WP1 and the rest of the project. It highlights how the results of WP1 contribute to the development and improvement of the CASMACAT workbench.

The analyses carried out on the data and user feedback collected from the field trials produce a snapshot of the current acceptability and performance of the CASMACAT workbench and directly feed into its further development. For example two new functionalities identified as

necessary and desirable in the first field trial were subsequently introduced in prototype-II (i.e. systematic search and replace, and copying source text to target window (WP2)).

In the context of this second CASMACAT field trial, there has been a close collaboration between WP2 (UPVLC) and WP1 (CBS). UPVLC have made direct use of the logging data made available by CBS and the two groups carried out analyses of the data from different perspectives resulting in the complementary analyses reported in section 2.5.

The translation styles reported in section 3.1 can also feed further CASMACAT development. In particular, style 3 showing monolingual reading could ground the implementation of a monolingual GUI in CASMACAT for both revision and monolingual post-editing tasks where the user is only presented with the text in the target language (WP5). It is now common practice for some language service providers to commission monolingual post-editing at a lower rate.

During the second field trial, where PI and PIA have shown different user satisfaction and productivity gains, interactivity was activated during all the post-editing session whenever the user was requested to work in system CFT2 or CFT3. Based on the findings of this experiment, the next logical step will be to give users the opportunity to turn IMT on and off as they choose. We envisage making the existing short-cut *ESC* or GUI button ITP/PE more flexible so that it applies both at the segment and text level (WP5). New experiments are being planned with IMT into Danish and German giving the user the chance to work with this feature as they choose in order to see how and when they make use of it.

Those users who reported lower satisfaction when using IMT in the second field trial (see deliverable D6.2) might benefit from a different way to show and rank translation options. Therefore further exploitation of IMT with different visualization options are being implemented to expand the functionalities pioneered by Caitra (WP3 - Task 3.5).

Taking into account the feedback provided by Celer post-editors indicating a desire for a concordancer and the work done in WP3 to develop such a tool (task 3.5), future user experiments will include investigations the use of this tool.

In the pilot experiment reported in section 2.2.2, confidence measures achieved very poor user satisfaction due to the lack of accuracy in most of the scores for proper nouns and acronyms. This issue is already being addressed as part of the WP2 (Task 2.3). Since proper nouns and acronyms are often translated correctly by the MT system, but show a low confidence score because of low representativeness in the corpus, the possibility of using a named entities recognizer is being explored to classify such words as correct translations.

The introduction of reviewers into the second field trial provided useful information on the revision process. This will be continued in the next field trial. It will be interesting to see how reviewers, as opposed to post-editors, can benefit from e-pen integration (WP5 - Task 5.3) whenever they are presented with a revision task.

The eye-tracking study reported in section 4.2 also offers interesting conclusions to be considered in further implementation work:

- Monolinguals could be used to do a cheap first pass, correcting obvious mistakes and flagging ones that may need the original source text to rectify or check, as detection rates were fairly similar between mono- and multi-linguals. In other words, an initial clean-up of the MT output does not require knowledge beyond the target language, meaning that the skills of professional translators can be further optimized.
- For the display of text in the CASMACAT interface, it is best to avoid sentence and clause breaks over lines as there is some evidence that the eyes of multilingual readers are more likely to have moved onto the next word in the MT output before a mistake is identified, increasing regressions. If a return sweep has already been made to the beginning of the next line, regressions to earlier text are more costly and disruptive. This is not so important for monolinguals as more of the error types seemed to be identified before their eyes fixate new information.

- The separation into global and localised effects complements the Quality Estimation work of WP3. WP3 examines both sentence-level post-editing effort (global processing) and word-level confidence measures (local effects), as well as making a similar distinction in paraphrasing granularity (sentential or clausal paraphrasing versus lexical or phrasal paraphrasing).

6 References

- Alabau, V., Mesa-Lao, B., et al. 2013. "User Evaluation of Advanced Interaction Features for a Computer-Assisted Translation Workbench". In Depraetere, H., Forcada, M.L., Grasmick, D., Sima'an, K., Way, A., (eds). *Proceedings of the XIV Machine Translation Summit - Nice*, September 2-6, 2013. pp. 361 - 368.
- Alves, F. and 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
- Balling Winther, L.; Carl, M. (forthcoming - 2013). "Production time across languages and tasks: a large-scale analysis using the CRITT translation process database", in Aline Ferreira and John Schwieter (eds). *The development of translation competence: Theories and methodologies from psycholinguistics and cognitive science*. Cambridge Scholars Publishing.
- Barrachina, S.; Bender, O.; Casacuberta, F.; et al. 2009. "Statistical approaches to computer-assisted translation." *Computational Linguistics*, 35(1): 3 - 28.
- Bertoldi, N.; Cattelan, A.; Federico, M. 2012. "Machine translation enhanced computer assisted translation. First report on lab and field tests". Available from: <http://www.matecat.com/wp-content/uploads/2013/01/MateCat-D5.3-V1.2-1.pdf>.
- Brown, P.F. et al. 1993. "The mathematics of statistical machine translation: Parameter estimation." *Computational Linguistics*, 19(2): 263 - 311.
- Callison-Burch, C., Fordyce, C., Koehn, P., Monz, C., and Schroeder, J. (2008). "Further meta-evaluation of machine translation." In *Proceedings of the Third Workshop on Statistical Machine Translation*, StatMT '08, pages 70-106, Stroudsburg, PA, USA. Association for Computational Linguistics.
- Carl, M.; Kay, M. 2011. "Gazing and Typing Activities during Translation: A Comparative Study of Translation Units of Professional and Student Translators". *Meta* 56:4, 952 - 975
- Carl, M. 2012. "The CRITT TPR-DB 1.0: A Database for Empirical Human Translation Process Research". *Proceedings of the AMTA 2012 Workshop on Post-Editing Technology and Practice*. Sharon O'Brien; Michel Simard; Lucia Specia (eds.). Stroudsburg, PA : Association for Machine Translation in the Americas (AMTA), pp. 9 - 18.
- Casacuberta, Francisco, Jorge Civera, Elsa Cubel, Antonio Luis Lagarda, Guy Lapalme, E. Macklovitch, and Enrique Vidal. 2009. "Human interaction for high quality machine translation". *Communications of the ACM*, 52(10): 135 - 138.
- Elming, Jakob, Michael Carl, and Laura Winther Balling. (Forthcoming). "Investigating User Behaviour in Post-editing and Translation Using the CASMACAT Workbench." *Expertise in Post-editing: Processes, Technology and Applications*, edited by Sharon O'Brien, Michael Simard, Lucia Specia, Michael Carl and Laura Winther Balling. Cambridge Scholars Publishing.

- Flanagan, M. A. 1994. "Error classification for MT evaluation". *Proceedings of the 1st conference of the Association for Machine Translation in the Americas*, AMTA. Columbia, MD, USA. October 5 - 8, pp.65-72.
- Guerberof, Ana. 2012. *Productivity and quality in the post-editing of outputs from translation memories and machine translation*. Ph.D. thesis, Tarragona: Universitat Rovira i Virgili.
- González-Rubio, Jesús, Daniel Ortiz-Martínez, and Francisco Casacuberta. 2010. "On the use of confidence measures within an interactive-predictive machine translation system". In *Proceedings of the 14th annual conference of the European Association for Machine Translation (EAMT)*. St. Raphael, France, 27 - 28 May.
- Hansen-Schirra, S., Neumann, S., Steiner, E. 2007. "Cohesion and Explicitation in an English-German Translation Corpus". *Languages in Contrast* 7(2): 241 - 265.
- Horning, A. S. and A. Becker (eds.). 2006. *Revision: history, theory, and practice*. West Lafayette, IN, USA: Parlor Press LLC.
- Hvelplund, K. T., and Carl, M. (2012). User Activity Metadata for Reading, Writing and Translation Research. In Arranz, V., Broeder, D., Gaiffe, B., Gavrilidou, M., Monachini, M., and Trippel, T. (Eds.), *Proceedings of The Eighth International Conference on Language Resources and Evaluation. LREC 2012*. (pp. 55-59). Paris: ELRA.
- Just, M.A. and Carpenter, P.A. 1980. "A theory of reading: from eye fixations to comprehension". *Psychological Review* 87: 329 - 354.
- Koehn, Philipp. 2010b. *Statistical Machine Translation*. Cambridge University Press.
- Krings, H. P. 2001. *Repairing texts: empirical investigations of machine translation post-editing processes*. Kent, OH, USA: The Kent State University Press, edited/translated by G. S. Koby.
- Lacruz, I., Shreve, G., Angelone, E. "Average Pause Ratio as an Indicator of Cognitive Effort in Post-Editing: A Case Study". *Proceedings of the AMTA 2012 Workshop on Post-Editing Technology and Practice (WPTP 2012)*. ed. / Sharon O'Brien; Michel Simard; Lucia Specia. Stroudsburg, PA : Association for Machine Translation in the Americas (AMTA), 2012. pp. 1 - 10.
- Langlais, Philippe, George Foster, and Guy Lapalme. 2000. "TransType: unit completion for a computer-aided translation typing system, applied natural language processing". *Applied Natural Language Processing (ANLP)*, pp. 46 - 51.
- Loffler-Laurian, A.-M. 1984. "Machine translation: what type of post-editing on what type of documents for what type of users". *Proceedings of the 10th international conference on computational linguistics and 22nd annual meeting of the Association for Computational Linguistics*. Stanford, CA, USA, July 2 - 6, pp. 236-238.
- Loffler-Laurian, A.-M. 1996. *La traduction automatique*. Paris, France: Presses Universitaires du Septentrion.
- Marrafa, P. and A. Ribeiro. 2001. "Quantitative evaluation of machine translation systems: sentence level". *Proceedings of the 8th MT Summit*, Santiago de Compostela, Spain, September 18 - 22 (no page numbers).
- Martínez-Gómez, P., Sanchis-Trilles, G.; Casacuberta, F. 2012. "Online adaptation strategies for statistical machine translation in post-editing scenarios". *Pattern Recognition*, 45:9, pp. 3193 - 3203.

- Mesa-Lao, Bartolomé. 2012. "The next generation translator's workbench: post-editing in CASMACAT v.1.0." *Proceedings of the 34th Translating and the Computer Conference*, ASLIB, 29 & 30 November 2012.
- Pym, A. 1992. "Translation error analysis and the interface with language teaching". Dollerup, C. and A. Loddegaard (eds.), *Teaching translation and interpreting*. Amsterdam and Philadelphia: John Benjamins Publishing Company, pp. 279 - 288.
- Sanchis-Trilles, G., Daniel Ortiz-Martínez, Jorge Civera, Francisco Casacuberta, Enrique Vidal, and Hieu Hoang. 2008. "Improving interactive machine translation via mouse actions". *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP.
- Moritz Schaeffer and Michael Carl. (in print). In *Describing cognitive processes in translation: Acts in events* - Special issue of *Translation and Interpreting Studies* 8:2 (Benjamins Publishing Company).
- Tirkkonen-Condit, Sonja. 2005. "The Monitor Model Revisited: Evidence from Process Research." *Meta* 50: 405 - 414

A Appendix

A.1 Terminology

This section presents some basic concepts used throughout this report:

- **cognition**: the mental process of knowing, including aspects such as awareness and perception. In this context it refers to the reasoning made by the human translator while post-editing machine translation outputs.
- **cognitive analysis**: investigation of the mental processes involved during translation and post-editing by means of eye-movements and keyboard activity.
- **eye-tracking**: research method measuring either the point of gaze (where one is looking) or the motion of an eye over a given area of interest. An eye tracker is a device for measuring eye positions and eye movements.
- **human-computer interaction** (HCI) involves the study, planning, and design of the interaction between users (humans) and computers. It is often regarded as the intersection of computer science, behavioural sciences and design.
- **interactive machine translation** (IMT): sub-field of computer-aided translation. Under this translation paradigm, the computer software that assists the human translator attempts to predict the text the user is going to input by taking into account all the information it has available. Whenever such prediction is wrong and the user provides feedback to the system, a new prediction is performed considering the new information available. Such process is repeated until the translation provided matches the user's expectations. IMT can also be referred under the term *interactive translation prediction*, ITP.
- **interactive translation prediction** (ITP): see *interactive machine translation*.
- **key-logging**: data collection method recording the keys struck on a keyboard, typically in a covert manner so that the person using the keyboard is unaware that their actions are being monitored.

- **post-editing:** proofreading and correcting a pre-translated text generated by a machine translation system against an original source text in order to comply with a set quality criteria. We use this term also when the machine translated text is dynamically created in a session of *interactive machine translation*.
- **monolingual post-editing:** kind of *post-editing* where the user is only presented with the text in the target language.
- **translation style:** recurrent pattern of activity identified in the user activity data (collected through e.g. eye-tracking or key-logging) during the translation or post-editing process.
- **translator type:** translator realizing a set of particular *translation styles* in certain contexts such as: when translating certain types of texts, when using certain types of translation assistance (i.e. GUI), when working under a certain translation brief, when working in different environments (e.g. at home, in the office) etc.

User Evaluation of Advanced Interaction Features for a Computer-Assisted Translation Workbench

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

Departamento de Sistemas Informáticos y Computación
Universitat Politècnica de València

Camí de Vera s/n, 46021 Valencia (Spain)

{valabau, jegonzalez, luileito}@dsic.upv.es
{daormar, gsanchis, fcn}@dsic.upv.es

B. Mesa-Lao and R. Bonk and M. Carl and M. García-Martínez

Center for Research and Innovation in Translation and Translation Technology (CRITT)

Copenhagen Business School

Dalgas Have 15, 2000 Frederiksberg (Denmark)

{bm.abc, rbo.abc, mc.abc, mgm.abc}@cbs.dk

Abstract

This paper reports on the results of a user satisfaction survey carried out among 16 translators using a new computer-assisted translation workbench. Participants were asked to provide feedback after performing different post-editing tasks on different configurations of the workbench, using different features and tools. Resulting from the feedback provided, we report on the utility of each of the features, identifying new ways of implementing them according to the users' suggestions.

1 Introduction

Machine translation (MT) technology has been playing an increasingly important role within translation over the past six decades. Nowadays its impact is undisputedly extensive and has reached an unprecedented level that deserves careful consideration as a crucial factor which affects human translators.

The use of MT systems for the production of post-editing drafts has become a widespread practice among many Language Service Providers (LSPs). This is confirmed by an extensive market study (TAUS, 2009) in which industry practices were surveyed in regard to translation automation in 129 LSPs. 40% of the surveyed LSPs reported that they are already using MT, while 89% of the remaining 60% reported that they were planning to integrate MT in their translation processes within the following two years.

The reasons for this increase in the adoption of MT technology are diverse. Apart from the productivity gains in the translation industry reported by several studies (de Almeida and O'Brien, 2010; Plitt and Masselot, 2010; Guerberof, 2012), there are many other reasons behind such a recent MT adoption. Some of these reasons could be a greater availability of resources and tools for the development of MT systems, a change in the expectations of MT users, as well as a successful integration of MT systems in already well-established computer-assisted translation (CAT) workbenches.

Traditionally post-editing workflows only take into account the human component in a serial process (Isabelle and Church, 1998). First the MT system provides complete translations which are then proofread by a human translator. In such a serial scenario, there is no actual interaction between the MT system and the human translator, making it impossible for the MT system to benefit from overall human translation skills and preventing the human translator from making the most out of the adaptive ability of some MT systems.

An alternative to this traditional workflow is represented by the interactive machine translation (IMT) approach (Langlais and Lapalme, 2002; Casacuberta et al., 2009; Barrachina et al., 2009). In the IMT approach, a fully-fledged MT engine is embedded into a post-editing workbench allowing the system to look for alternative translations whenever the human translator corrects the MT output. MT technology is used to produce full target sentences (hypotheses), or portions thereof, which can be interactively accepted or edited by a human translator. The system continues search-

ing for alternative renditions as the translator edits the text. The MT engine then exploits the changes made by the translator to produce improved outputs, and provides the user with fine-tuned completions of the sentence being translated.

IMT can be seen as an evolution of the statistical MT (SMT) framework (Koehn, 2010b). Within the IMT framework, a state-of-the-art SMT system is used in the following way. For a given source sentence, the SMT system automatically generates an initial translation. A human translator checks this machine translation, correcting the first error. The SMT system then proposes a new completion or suffix, taking the correction into account. These steps are repeated until the whole input sentence has been correctly translated.

The present study reports on a user evaluation of an IMT workbench being implemented as part of the CASMACAT project¹. Research was devised so as to investigate user satisfaction while post-editing MT outputs using a translation workbench featuring different tools and resources. The ultimate aim of testing these different configurations was to assess their potential and decide which of them can be successfully integrated into the second prototype of the CASMACAT workbench for the benefit of the human translator. This study also aimed at fine-tuning some of the IMT features tested in light of the feedback provided by the users.

Improving and maximizing the potential of a post-editing workbench is one of the priorities set by both the industry and researchers when addressing the technological challenges faced by human translators. The motivation behind this research ultimately comes from a desire to know how such tools can be of greater support to translation professionals, and how technology can even empower them to make an unrestricted choice of the translation methods, strategies and tools they feel comfortable with and which bring out the best of their skills (Mesa-Lao, 2012).

2 Background research

Human translator interaction with MT technology harks back to the emergence of the first effective MT systems (Vasconcellos and León, 1985). Traditionally this human-computer interaction involves the human translator as a post-editor (proof-

reader) of MT outputs, but rarely involves the human translator guiding the decisions of an MT system. Recent seminal efforts on building interactive MT systems include Langlais et al. (2000) and Barrachina et al. (2009). Both studies develop research systems looking into a tighter integration of human translators in MT processes by developing a prediction model that interactively suggests translations to the human translator as she types. Similar work was carried out by Koehn (2010a), displaying different translations to human translators and letting them choose the one that better suited their needs for post-editing.

An important contribution to IMT technology was pioneered by the TRANSTYPE project, where data driven MT techniques were adapted for their use in an interactive translation environment. Langlais et al. (2002) performed a human evaluation on their interactive prototype emulating a realistic working environment in which the users could obtain alternative renditions as they were typing to fix MT outputs. In this study, post-editors' productivity decreased by 17%, but they appreciated such an interactive system and declared that it could help them to improve their productivity after proper training.

In line with the aims of the TRANSTYPE project, Barrachina et al. (2009) also worked with the IMT approach by using fully-fledged MT systems to produce MT hypotheses. Translators could choose new suggestions from the SMT system as they were correcting MT outputs. Each corrected output was used by the system as additional information to achieve future improved suggestions. Further research has also been carried out as part of the TRANSTYPE2 project (Casacuberta et al., 2009). In this project, post-editors' performance tended to increase as they became acquainted with the system over a 18-month period.

A slightly different approach was studied in Koehn (2010a), where monolingual users evaluated a translation interface supporting predictions and the so-called "translation options". On Arabic-English and Chinese-English, using standard test data and current SMT systems, 10 monolingual users were able to translate 35% of Arabic and 28% of Chinese sentences correctly on average, with some of the participants coming close to professional bilingual performance on some of the texts.

¹CASMACAT: *Cognitive Analysis and Statistical Methods for Advanced Computer Aided Translation*. Project co-funded by the European Union under the Seventh Framework Programme Project 287576 (ICT-2011.4.2).

3 Workbench Features

For the purpose of the evaluation we decided to implement a web-based prototype supporting IMT features. Web-based applications present several advantages. Firstly, they provide a powerful and mature environment to implement dynamic interfaces with advanced visual features. Secondly, they can be easily deployed worldwide reaching virtually anyone. For this purpose, we leveraged the MATECAT post-editing interface (Bertoldi et al., 2012), which is an open source web application. On top of their interface, we implemented the visualization of the advanced features, connected to our IMT servers. Figure 1 shows the implemented CASMACAT interface with some features that we believe are desirable in any IMT-based workbench.

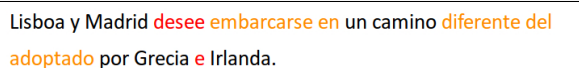
In the following subsections we present a short description of the main features that were implemented in the prototype. Such features are different in nature, but all of them aimed at facilitating the post-editing process.

3.1 Intelligent Autocompletion

IMT with intelligent autocompletion takes place every time a keystroke is detected by the system (Barrachina et al., 2009). In such an event, the system produces a (full) suitable prediction according to the text that the user is writing. This new prediction replaces the remaining words of the original sentence at the right of the text cursor.

3.2 Confidence Measures

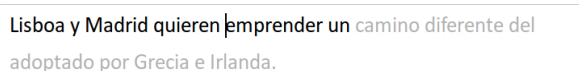
Current MT systems are still far from perfect. It would be thus desirable to improve their use by adding information on the reliability of the output produced. A way to do so would be by highlighting chunks of translated text that, according to the system knowledge, are not reliable enough (González-Rubio et al., 2010). In the CASMACAT workbench, we use confidence measures to inform post-editors about the reliability of translations under two different criteria. On the one hand, we highlight in red those translated words that are likely to be incorrect. We use a threshold that maximizes precision in detecting incorrect words. On the other hand, we highlight in orange those translated words that are dubious for the system. In this case, we use a threshold that maximizes recall.



Lisboa y Madrid desee embarcarse en un camino diferente del adoptado por Grecia e Irlanda.

3.3 Prediction Length

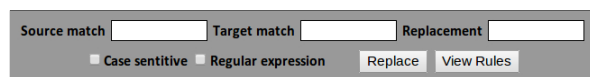
Providing the user with a new prediction whenever a key is pressed has been proved to be cognitively demanding (Alabau et al., 2012). For this reason it was decided to limit the number of predicted words that are shown to the user by only predicting up to the first word with a low confidence measure according to the system. In our implementation, pressing the Tab key allows the user to ask the system for the next set of predicted words, painting in gray the remaining words in the suggested translation.



Lisboa y Madrid quieren emprender un camino diferente del adoptado por Grecia e Irlanda.

3.4 Search and Replace

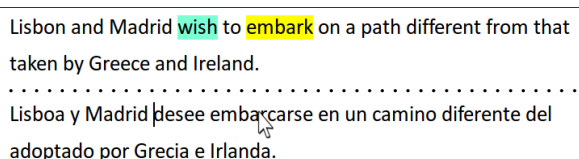
Most of the computer-assisted translation tools provide the user with intelligent search and replace functions for fast text revision. The CASMACAT workbench also features a straightforward function to run search and replacement rules on the fly. Whenever a new replacement rule is created, it is automatically populated to the forthcoming predictions made by the system, so that the user only needs to specify them once.



Source match Target match Replacement
 Case sensitive Regular expression

3.5 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 in a way 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.



Lisbon and Madrid wish to embark on a path different from that taken by Greece and Ireland.
.....
Lisboa y Madrid desee embarcarse en un camino diferente del adoptado por Grecia e Irlanda.

3.6 Prediction Rejection

With the purpose of easing user interaction, our prototype also supports a mouse wheel rejection

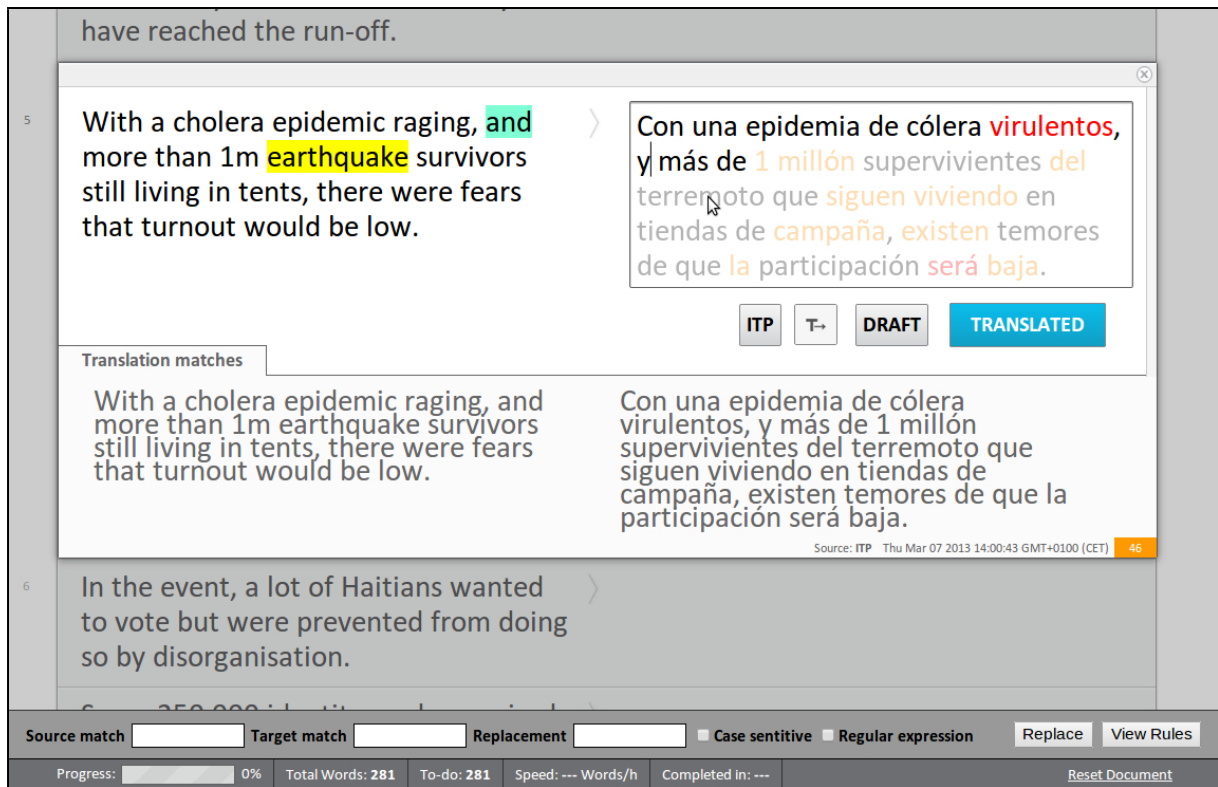


Figure 1: Screenshot of our workbench with all its features enabled.

feature (Sanchis-Trilles et al., 2008). By scrolling the mouse wheel over a word, the system invalidates the current prediction and provides the user with an alternate translation in which the first new word is different from the previous one.

4 User Evaluation

The main goal of this research was to measure user satisfaction when performing post-editing tasks using different workbench features (see Table 1). In this context, we were interested in knowing whether translators find the use of such features useful while post-editing MT outputs.

4.1 Workbench Configurations

For this purpose, we defined four different configurations of the workbench (see Table 1). Each of them differs in the set of features that are included (see section 3). System 1 (S1) was a baseline system for IMT including only basic intelligent auto-completion. Systems 2 to 4 (S2–S4) included intelligent auto-completion together with some of the advanced features described above.

4.2 Participants Profile

A group of 16 users (10 females and 6 males) aged between 21 and 34 volunteered to perform

the evaluation of the different systems. All participants had a degree in translation studies and were regular users of computer-aided translation tools (i.e., SDL Trados and MemoQ), but they had never used IMT technology to post-edit. When asked about previous experience in post-editing of MT outputs, 55% of claimed to have previous experience in post-editing assignments. This difference in post-editing experience was not considered a bias in the sample of the study, since the aim was not to measure productivity but user satisfaction.

4.3 Questionnaires

A system usability scale (SUS) questionnaire was used to collect quantitative data on user satisfaction. Users had to assess each system in a typical five-level Likert scale, with five denoting the highest satisfaction, right after performing a post-editing task in each of the four different systems. In addition to the Likert scale, each questionnaire also included a text area for users to submit additional comments and feedback on the feature being tested. A final overall questionnaire was also filled out in order to know which of the four configurations of the workbench was most preferred.

Workbench features	Systems			
	S1	S2	S3	S4
basic intelligent autocompletion (IMT)	*			
IMT + confidence measures		*		
IMT + prediction length control			*	
IMT + search and replace function				*
IMT + word alignments				*
IMT + prediction rejection				*

Table 1: List of the workbench features included in each of the four evaluated systems (S1 to S4).

4.4 Source Texts

The source texts compiled for this user evaluation were short pieces of news that are likely to appear in any general scope newspaper, extracted from the News Commentary corpus². No expert knowledge was thus required in order to successfully perform the post-editing task. The language pair involved was English to Spanish.

4.5 Procedure

Each system was tested using a different data set consisting of 20 segments each; two pieces of news per system. Before performing the evaluation, participants were asked to fill out an introductory questionnaire in order to collect data about their profile as professional translators, as well as their previous experience in post-editing. The evaluation always involved System 1 in the first place, since it was considered as a baseline prior to testing the advanced IMT features implemented in the other systems. The evaluation of Systems 2, 3, and 4 was done in a randomized order in order to minimize the effect of any ordering on user satisfaction (i.e., due to learning or fatigue effects). The presentation of the different source texts was also randomized across the different systems so as to avoid the potential effect of text difficulty on the evaluation of the system. No time constraints were imposed on the participants involved in the evaluation.

5 Results

From the submitted questionnaires, an overview of user satisfaction for the different systems is shown in Figure 2 following the above described five-level Likert scale, where 5 denotes the high-

est satisfaction. For each of the evaluated systems, we display the average of the satisfaction scores given by the users (blue box), the 95% confidence interval for the average satisfaction score (black whisker), and the actual distribution of user satisfaction scores (gray pattern). The baseline system (System 1) was given an average satisfaction score of 2.4. In comparison, System 2 was given a slightly worse satisfaction score (2.1) while both System 3 (3.3) and System 4 (2.9) scored clearly above the baseline. Moreover, the confidence intervals for System 1 and System 3 do not overlap.

Overall, the most popular workbench among participants was System 3 (the one featuring prediction length control). Participants seemed to favor the idea of editing chunks of information while having such a visual aid; i.e., showing in black the text that has already been post-edited and showing in gray the text that still needs revision. As stated by one participant, “[...] *This feature guided me in the post-editing process, having a greater control of what I had actually edited in the text. I didn’t have the feeling that the system was making too many changes at a time and I felt more in control of the editing process*”. System 2, featuring confidence measures (red for wrong and orange for dubious translations), recorded the lowest user satisfaction scores. However, some participants reported in the open-ended questionnaire that this feature seems to be very promising if a more reliable implementation was deployed. “*I could definitely benefit from this type of visual aid, but the system still needs to make better predictions. Many times the words marked by the system as wrong were actually correct, while wrong translations remained in black. In the end I had to double-check most of the sentences to make sure that words marked in black were actually acceptable translations*”, stated one participant.

²Training corpus for the sixth workshop on SMT 2011 (see <http://www.statmt.org/wmt11>)

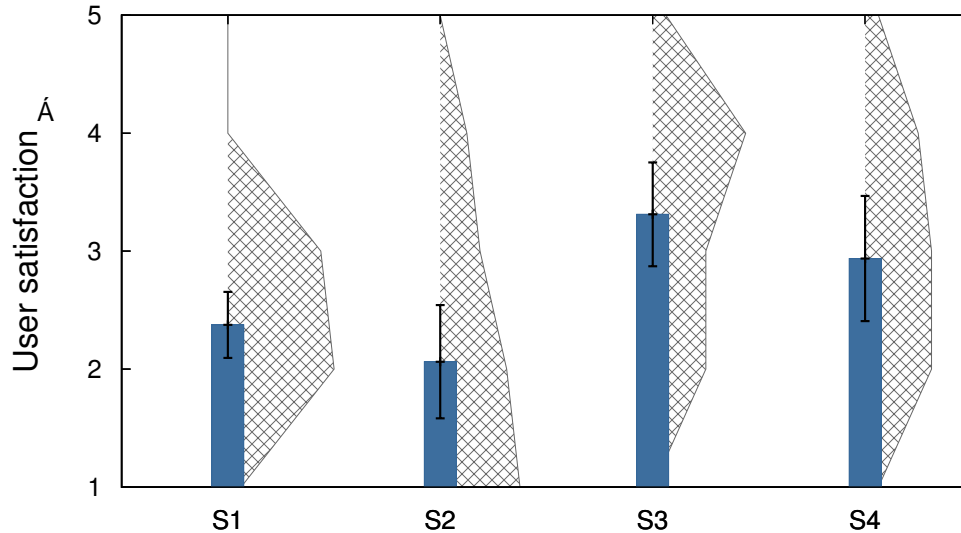


Figure 2: Average user satisfaction reported by the users for each system (S1–S4). We additionally display in black the 95% confidence interval for the average satisfaction score, and in gray the actual distribution of the satisfaction values given to each system.

None of the participants rated the baseline system above 3 and actually 50% of them were dissatisfied with the translations produced. These poor results could be attributed to the fact that System 1 was used as a baseline (featuring basic intelligent autocompletion with no advanced features), and therefore it was always evaluated first. Users would have certainly benefited from a warm-up session to become acquainted with IMT before the formal evaluation.

In line with previous findings by Barrachina et al. (2009) and Casacuberta et al. (2009), the more the participants became familiar with the system, the less the system was perceived as being cumbersome. Feedback recorded in the open-ended questionnaire showed that on-demand word alignment, implemented in the System 4, was very positively perceived by the users as a real aid to spot sources of mistranslations.

Another important finding was the fact that, as most of the participants were experienced touch-type translators, some of them reported that it would have been faster for them to type longer strings of text instead of having to interact with the IMT system. In this regard, some of them suggested an extra feature for enabling and disabling IMT depending on the segment that is being post-edited.

In addition to the general findings described above, the feedback provided by the users contains valuable information that can be used to guide

the future development of the CASMACAT workbench. The next sections describe the lessons learned about each of the features and tools included in the workbench.

5.1 Confidence Measures

The clarifications made by the users revealed that the main problem of this feature stems in the tendency of the system to classify as incorrect words that, from the translator’s point of view, are clearly correct. For example, proper names are usually classified as incorrect since they tend to appear few times, if any, in the training data. Such errors are infrequent, so they do not penalize much the performance of the confidence measure as evaluated in most automatic measures. However, these errors are quite annoying for the users who then distrust the confidence information provided by the system.

Users also provided us with feedback on how to display the confidence measures computed by the system. All participants agreed that the color selection was adequate, allowing for an easy identification of potential wrong translations in red and dubious in orange. However, they had mixed opinions regarding the usefulness of showing both wrong and dubious equivalents. Five users considered confidence measures for dubious equivalents (words in orange) a source of visual noise. They pointed that it is only useful to highlight confidence measures for clearly wrong equivalents

(words in red). The rest of the participants preferred both thresholds (wrong and dubious equivalents) to be displayed. As a consensus, it seems that translators should be provided with both options and let them decide which of these options, if not both, they want to use.

5.2 Prediction Length

In contrast to the criticism received by the system including confidence measures, the system featuring the prediction length did yield positive satisfaction ratios, even though the length of the prediction is set according to the same confidence measures. Users stated that this feature eased their interaction with the system, by reducing the stress involved in deciding upon the acceptability/correctness of the (sometimes quite different) completions provided by the system.

Some users commented that the limitation imposed by this feature to the autocompletions was a good indicator of what had actually been edited in the text. Nevertheless, this was not the intended purpose of this feature, but this seems to suggest that users would find useful a specific feature targeted to identifying already edited words. For instance, already edited words could be highlighted in green or a special symbol could be used to display the last position of the caret.

5.3 Search and Replace

Although the evaluation did not present enough sentences to the users so that the search and replace feature could be actually assessed, it was perceived positively. Translators agreed in that it is indeed a must in any professional workbench. So far, our search and replace module operates on the autocompletions provided by the system by dynamically applying replace rules. However, since the traditional search and replace feature is perceived as so valuable, future work will be addressed to find different ways of integrating it into the CAS-MACAT workbench.

5.4 Word Alignment Information

Word alignment information was considered to be quite useful. However, user opinions were mixed regarding the utility of the different visualization options. One frequent comment was that the alignment information triggered by the cursor position can be considered a source of distraction during the translation process as aligned words kept changing as the user edited the MT output. Therefore we

conclude that word alignment information should only be displayed on user demand. For instance, it could be shown only when the user presses a given keyboard shortcut.

5.5 Prediction Rejection

This feature also received positive reviews by most of the participants on this user evaluation. Nonetheless, some users reported that the implemented interaction mechanism was somehow unexpected. They would have expected the rejection operation to affect only the word under the cursor, instead of operating on the whole of the remaining sentence to the right of the cursor. Users suggested that this prediction rejection feature should be limited to single words (i.e. looking for alternative equivalents) instead of triggering further changes at the sentence level. Some users also commented that, instead of having to jump from prediction to prediction before finding the right one, a drop-down list would be preferable. Such an implementation of this feature could show several predictions at a time, making the interaction with the system faster. This suggestion, however, challenges the TRANSTYPE2 findings (Langlais and Lapalme, 2002), where drop-down lists were perceived as too overwhelming by the participants in the study. Further research is still needed on how best we can present predictions to the user.

6 Summary and Conclusions

This user evaluation of features for an IMT-based workbench has proved to be successful in addressing the actual benefits of automating interactivity between the MT system and the human translators. In this sense, the surveyed translators provided us with valuable feedback from real users in order to fine tune some of the tested features. One of the key findings of this user satisfaction study is the lack of agreement from most of the translators about which features they want to see implemented in a workbench to make post-editing a more rewarding task. This is certainly a crucial issue that needs further consideration by both human translators and tool developers. Overall, the workbench configuration that translators seem to be more satisfied with is the one featured in System 3 (with prediction length control). Further research is still needed with different user profiles as well as with more hours of interaction with the different features of the workbench.

Acknowledgments

Work supported by the European Union 7th Framework Program (FP7/2007-2013) under the CASMACAT project (grants agreement n° 287576), by Spanish MICINN under grants TIN2009-14205-C04-02 and TIN2012-31723, and by the Generalitat Valenciana under grant ALMPR (Prometeo/2009/014).

References

- Alabau, Vicent, Luis A. Leiva, Daniel Ortiz-Martínez, and Francisco Casacuberta. 2012. User evaluation of interactive machine translation systems. In *Proceedings of the 16th Annual Conference of the European Association for Machine Translation (EAMT)*, pages 20–23.
- Barrachina, Sergio, Oliver Bender, Francisco Casacuberta, Jorge Civera, Elsa Cubel, Shahram Khadivi, Antonio Lagarda, Hermann Ney, Jesús Tomás, Enrique Vidal, and Juan-Miguel Vilar. 2009. Statistical approaches to computer-assisted translation. *Computational Linguistics*, 35(1):3–28.
- Bertoldi, Nicola, Alessandro Cattelan, and Marcello Federico. 2012. Machine translation enhanced computer assisted translation. First report on lab and field tests. Available from: <http://www.matecat.com/wp-content/uploads/2013/01/MateCat-D5.3-V1.2-1.pdf>.
- Brown, Peter F, Vincent J Della Pietra, Stephen A Della Pietra, and Robert L Mercer. 1993. The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics*, 19(2):263–311.
- Casacuberta, Francisco, Jorge Civera, Elsa Cubel, Antonio Luis Lagarda, Guy Lapalme, E. Macklovitch, and Enrique Vidal. 2009. Human interaction for high quality machine translation. *Communications of the ACM*, 52(10):135–138.
- de Almeida, G. and Sharon O’Brien. 2010. Analysing post-editing performance: correlations with years of translation experience. In *Proceedings of the 14th annual conference of the European Association for Machine Translation (EAMT)*. St. Raphael, France, 27-28 May.
- González-Rubio, Jesús, Daniel Ortiz-Martínez, and Francisco Casacuberta. 2010. On the use of confidence measures within an interactive-predictive machine translation system. In *Proceedings of the 14th annual conference of the European Association for Machine Translation (EAMT)*. St. Raphael, France, 27-28 May.
- Guerberof, Ana. 2012. *Productivity and quality in the post-editing of outputs from translation memories and machine translation*. Ph.D. thesis, Tarragona: Universitat Rovira i Virgili.
- Isabelle, Pierre and Ken Church. 1998. Special issue on: New tools for human translators. *Machine Translation*, 12(1/2).
- Koehn, Philipp. 2010a. Enabling monolingual translators: post-editing vs. options. In *NAACL HLT 2010 - Human Language Technologies: Proceedings of the 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 537–545.
- Koehn, Philipp. 2010b. *Statistical Machine Translation*. Cambridge University Press.
- Langlais, Philippe and Guy Lapalme. 2002. TransType: development-evaluation cycles to boost translator’s productivity. *Machine Translation*, 17(2):77–98.
- Langlais, Philippe, George Foster, and Guy Lapalme. 2000. TransType: unit completion for a computer-aided translation typing system, applied natural language processing. In *Applied Natural Language Processing (ANLP)*, pages 46–51.
- Mesa-Lao, Bartolomé. 2012. The next generation translator’s workbench: post-editing in CasMaCat v.1.0. In *Translating and the Computer Conference Proceedings*, 34.
- Plitt, Mirko and François Masselot. 2010. A productivity test of statistical machine translation post-editing in a typical localisation context. *Prague Bulletin of Mathematical Linguistics*, 93:7–16.
- Sanchis-Trilles, G., Daniel Ortiz-Martínez, Jorge Civera, Francisco Casacuberta, Enrique Vidal, and Hieu Hoang. 2008. Improving interactive machine translation via mouse actions. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing, EMNLP*.
- TAUS. 2009. LSPs in the MT loop: current practices, future requirements [report]. Available from: <http://www.translationautomation.com/reports/lsp-in-the-mt-loop-current-practice-future-requirements>.
- Vasconcellos, Muriel and Marjorie León. 1985. Spanam and engspan: machine translation at the pan american health organization. *Computational Linguistics*, 11(2-3):122–136.

A.2 Second field trial description

This appendix offers an overview of the texts and systems involved in the second field trial. This information was used to schedule each of the post-editing sessions from Celer Soluciones SL.

Appendix 7.8 - Second CASMACAT field trial at CELER Soluciones (June 2013)

Week 1: Post-editing sessions from CELER using the eye-tracker (EyeLink 1000)

Monday June 10		Tuesday June 11		Wednesday June 12		Thursday June 13		Friday June 14	
TRANSLATOR 01	<i>Task:</i> PE of Dataset 1 subset 1.1 (1,000 words) in prototype 1 Code: P	TRANSLATOR 03	<i>Task:</i> PE of Dataset 3 subset 3.1 (1,000 words) in prototype 1 Code: P	TRANSLATOR 04	<i>Task:</i> PE of Dataset 1 subset 1.1 (1,000 words) in prototype 2 Code: PI	TRANSLATOR 05	<i>Task:</i> PE of Dataset 1 subset 2.1 (1,000 words) in prototype 2 Code: PI	TRANSLATOR 06	<i>Task:</i> PE of Dataset 1 subset 3.1 (1,000 words) in prototype 2 Code: PI
TRANSLATOR 02	<i>Task:</i> PE of Dataset 2 subset 2.1 (1,000 words) in prototype 1 Code: P	TRANSLATOR 03	<i>Task:</i> Revision of Dataset 1 subset 1.1 (1,000 words) by Translator 01	TRANSLATOR 04	<i>Task:</i> Revision of Dataset 2 subset 2.1 (1,000 words) in prototype 3 Code: PIA	TRANSLATOR 05	<i>Task:</i> Revision of Dataset 1 subset 3.1 (1,000 words) in prototype 3 Code: PIA	TRANSLATOR 06	<i>Task:</i> Revision of Dataset 1 subset 1.1 (1,000 words) in prototype 3 Code: PIA
TRANSLATOR 01	<i>Task:</i> PE of Dataset 2 subset 2.1 (1,000 words) in prototype 2 Code: PI	TRANSLATOR 03	<i>Task:</i> Revision of Dataset 2 subset 2.1 (1,000 words) by Translator 01	TRANSLATOR 04	<i>Task:</i> Revision of Dataset 3 subset 3.1 (1,000 words) by Translator 02	TRANSLATOR 05	<i>Task:</i> Revision of Dataset 1 subset 1.1 (1,000 words) by Translator 02	TRANSLATOR 06	<i>Task:</i> Revision of Dataset 2 subset 2.1 (1,000 words) by Translator 04
TRANSLATOR 02	<i>Task:</i> PE of Dataset 3 subset 3.1 (1,000 words) in prototype 2 Code: PI	TRANSLATOR 03	<i>Task:</i> Revision of Dataset 3 subset 3.1 (1,000 words) by Translator 01	TRANSLATOR 04	<i>Task:</i> Revision of Dataset 1 subset 1.1 (1,000 words) by Translator 02	TRANSLATOR 05	<i>Task:</i> Revision of Dataset 3 subset 3.1 (1,000 words) by Translator 02	TRANSLATOR 06	<i>Task:</i> Revision of Dataset 2 subset 2.1 (1,000 words) by Translator 04
TRANSLATOR 01	<i>Task:</i> PE of Dataset 3 subset 3.1 (1,000 words) in prototype 3 Code: PIA	TRANSLATOR 03	<i>Task:</i> Revision of Dataset 3 subset 3.1 (1,000 words) by Translator 01	TRANSLATOR 04	<i>Task:</i> Revision of Dataset 1 subset 1.1 (1,000 words) by Translator 01	TRANSLATOR 05	<i>Task:</i> Revision of Dataset 1 subset 1.1 (1,000 words) by Translator 01	TRANSLATOR 06	<i>Task:</i> Revision of Dataset 3 subset 3.1 (1,000 words) by Translator 04
<i>interview</i>	<i>interview</i>	<i>interview</i>	<i>interview</i>	<i>interview</i>	<i>interview</i>	<i>interview</i>	<i>interview</i>	<i>interview</i>	<i>interview</i>

00:51 - 00:60

PROTOTYPES	LANGUAGE PAIRS	TRANSLATORS	REVIEWERS
Prototype 1: traditional PE Prototype 2: basic ITP Prototype 3: advanced ITP	English to Spanish	9	4

Appendix 7.8 - Second CASMACAT field trial at CELER Soluciones (June 2013)

Week 2: Post-editing sessions from CELER using the eye-tracker (EyeLink 1000)

	Monday June 17	Tuesday June 18	Wednesday June 19	Thursday June 20	Friday June 21
	TRANSLATOR 07	TRANSLATOR 08	TRANSLATOR 09	REVIEWER 03	REVIEWER 03
	Task: PE of Dataset 3 subset 3.1 (1,000 words) in prototype 3 Code: PIA	Task: PE of Dataset 1 subset 2.1 (1,000 words) in prototype 3 Code: PIA	Task: PE of Dataset 3 subset 3.1 (1,000 words) in prototype 3 Code: PIA	Task: Revision of Dataset 3 subset 2.1 (1,000 words) by Translator 05	Task: Revision of Dataset 3 subset 3.1 (1,000 words) by Translator 09
	TRANSLATOR 07	TRANSLATOR 08	TRANSLATOR 09	REVIEWER 03	REVIEWER 03
	Task: PE of Dataset 1 subset 2.1 (1,000 words) in prototype 1 Code: P	Task: PE of Dataset 2 subset 3.1 (1,000 words) in prototype 1 Code: P	Task: PE of Dataset 2 subset 1.1 (1,000 words) in prototype 1 Code: P	Task: Revision of Dataset 3 subset 3.1 (1,000 words) by Translator 05	Task: Revision of Dataset 3 subset 1.1 (1,000 words) by Translator 09
	TRANSLATOR 07	TRANSLATOR 08	TRANSLATOR 09	REVIEWER 03	REVIEWER 03
	Task: PE of Dataset 3 subset 3.1 (1,000 words) in prototype 2 Code: PI	Task: PE of Dataset 1 subset 1.1 (1,000 words) in prototype 2 Code: PI	Task: PE of Dataset 1 subset 2.1 (1,000 words) in prototype 2 Code: PI	Task: Revision of Dataset 3 subset 1.1 (1,000 words) by Translator 05	Task: Revision of Dataset 3 subset 2.1 (1,000 words) by Translator 09
	interview	interview	interview		
08:00 - 15:00					

Appendix 7.8 - Second CASMACAT field trial (June 2013)

Weeks 1, 2, 3: Post-editing tasks to be performed by the translators from home without eye-tracker after they have completed the first 3,000 words from CELER.

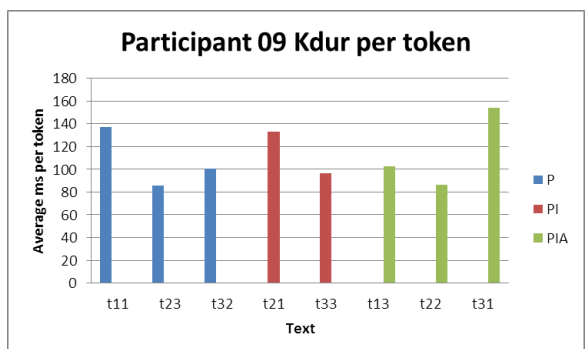
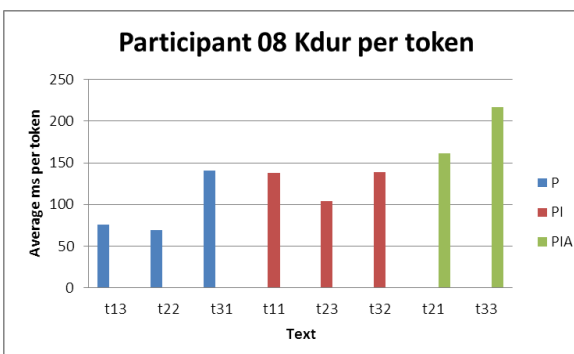
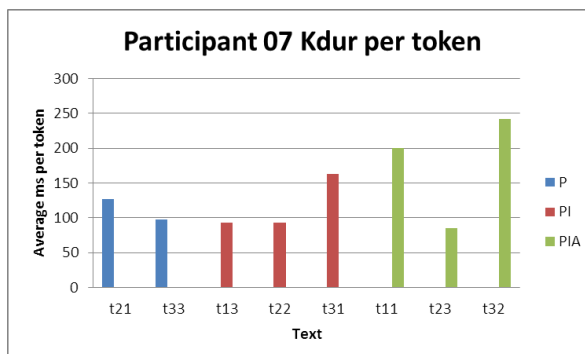
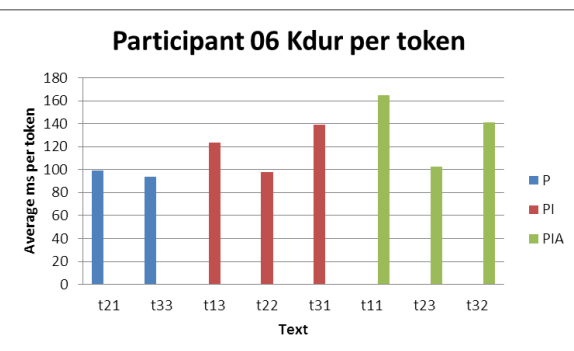
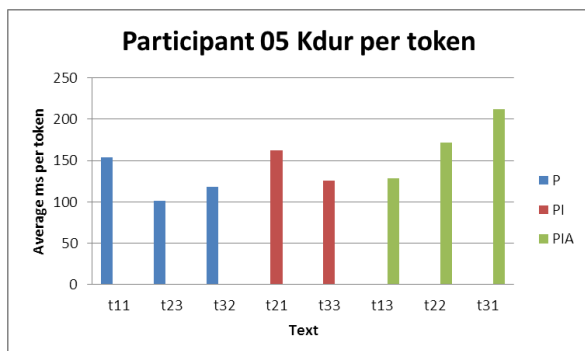
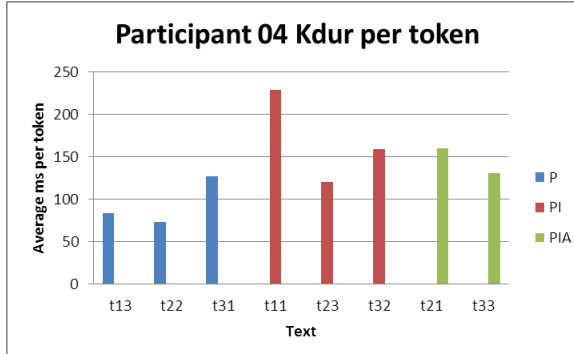
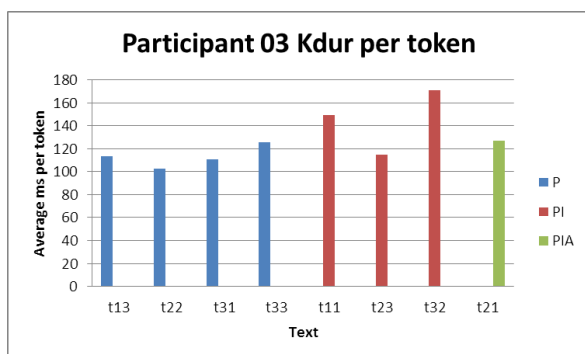
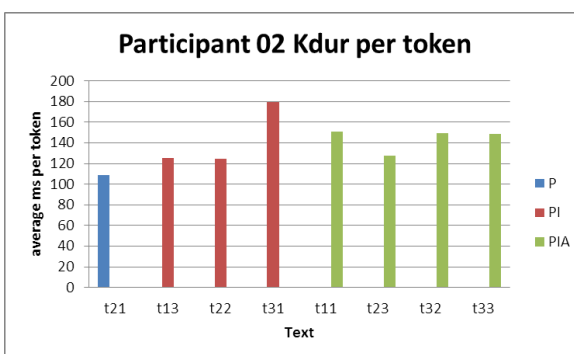
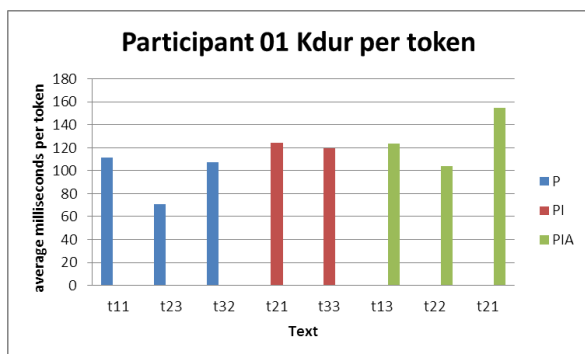
<p>TRANSLATOR 01:</p> <ul style="list-style-type: none"> Task: PE of Dataset 1 subset 1.2 (1,000 words) in prototype 2 (PI) Task: PE of Dataset 2 subset 2.2 (1,000 words) in prototype 3 (PIA) Task: PE of Dataset 3 subset 3.2 (1,000 words) in prototype 1 (P) Task: PE of Dataset 1 subset 1.3 (1,000 words) in prototype 3 (PIA) Task: PE of Dataset 2 subset 2.3 (1,000 words) in prototype 1 (P) Task: PE of Dataset 3 subset 3.3 (1,000 words) in prototype 2 (PI) 	<p>TRANSLATOR 02:</p> <ul style="list-style-type: none"> Task: PE of Dataset 2 subset 2.2 (1,000 words) in prototype 2 (PI) Task: PE of Dataset 3 subset 3.2 (1,000 words) in prototype 3 (PIA) Task: PE of Dataset 1 subset 1.2 (1,000 words) in prototype 1 (P) Task: PE of Dataset 2 subset 2.3 (1,000 words) in prototype 3 (PIA) Task: PE of Dataset 3 subset 3.3 (1,000 words) in prototype 1 (P) Task: PE of Dataset 1 subset 1.3 (1,000 words) in prototype 2 (PI)
<p>TRANSLATOR 03:</p> <ul style="list-style-type: none"> Task: PE of Dataset 3 subset 3.2 (1,000 words) in prototype 2 (PI) Task: PE of Dataset 1 subset 1.2 (1,000 words) in prototype 3 (PIA) Task: PE of Dataset 2 subset 2.2 (1,000 words) in prototype 1 (P) Task: PE of Dataset 3 subset 3.3 (1,000 words) in prototype 3 (PIA) Task: PE of Dataset 1 subset 1.3 (1,000 words) in prototype 1 (P) Task: PE of Dataset 2 subset 2.3 (1,000 words) in prototype 2 (PI) 	<p>TRANSLATOR 04:</p> <ul style="list-style-type: none"> Task: PE of Dataset 1 subset 1.2 (1,000 words) in prototype 3 (PIA) Task: PE of Dataset 2 subset 2.2 (1,000 words) in prototype 1 (P) Task: PE of Dataset 3 subset 3.2 (1,000 words) in prototype 2 (PI) Task: PE of Dataset 1 subset 1.3 (1,000 words) in prototype 1 (P) Task: PE of Dataset 2 subset 2.3 (1,000 words) in prototype 2 (PI) Task: PE of Dataset 3 subset 3.3 (1,000 words) in prototype 3 (PIA)
<p>TRANSLATOR 05:</p> <ul style="list-style-type: none"> Task: PE of Dataset 2 subset 2.2 (1,000 words) in prototype 3 (PIA) Task: PE of Dataset 3 subset 3.2 (1,000 words) in prototype 1 (P) Task: PE of Dataset 1 subset 1.2 (1,000 words) in prototype 2 (PI) Task: PE of Dataset 2 subset 2.3 (1,000 words) in prototype 1 (P) Task: PE of Dataset 3 subset 3.3 (1,000 words) in prototype 2 (PI) Task: PE of Dataset 1 subset 1.3 (1,000 words) in prototype 3 (PIA) 	<p>TRANSLATOR 06:</p> <ul style="list-style-type: none"> Task: PE of Dataset 3 subset 3.2 (1,000 words) in prototype 3 (PIA) Task: PE of Dataset 1 subset 1.2 (1,000 words) in prototype 1 (P) Task: PE of Dataset 2 subset 2.2 (1,000 words) in prototype 2 (PI) Task: PE of Dataset 3 subset 3.3 (1,000 words) in prototype 1 (P) Task: PE of Dataset 1 subset 1.3 (1,000 words) in prototype 2 (PI) Task: PE of Dataset 2 subset 2.3 (1,000 words) in prototype 3 (PIA)
<p>TRANSLATOR 07:</p> <ul style="list-style-type: none"> Task: PE of Dataset 1 subset 1.2 (1,000 words) in prototype 1 (P) Task: PE of Dataset 2 subset 2.2 (1,000 words) in prototype 2 (PI) Task: PE of Dataset 3 subset 3.2 (1,000 words) in prototype 3 (PIA) Task: PE of Dataset 1 subset 1.3 (1,000 words) in prototype 2 (PI) Task: PE of Dataset 2 subset 2.3 (1,000 words) in prototype 3 (PIA) Task: PE of Dataset 3 subset 3.3 (1,000 words) in prototype 1 (P) 	<p>TRANSLATOR 08:</p> <ul style="list-style-type: none"> Task: PE of Dataset 2 subset 2.2 (1,000 words) in prototype 1 (P) Task: PE of Dataset 3 subset 3.2 (1,000 words) in prototype 2 (PI) Task: PE of Dataset 1 subset 1.2 (1,000 words) in prototype 3 (PIA) Task: PE of Dataset 2 subset 2.3 (1,000 words) in prototype 2 (PI) Task: PE of Dataset 3 subset 3.3 (1,000 words) in prototype 3 (PIA) Task: PE of Dataset 1 subset 1.3 (1,000 words) in prototype 1 (P)
<p>TRANSLATOR 09:</p> <ul style="list-style-type: none"> Task: PE of Dataset 3 subset 3.2 (1,000 words) in prototype 1 (P) Task: PE of Dataset 1 subset 1.2 (1,000 words) in prototype 2 (PI) Task: PE of Dataset 2 subset 2.2 (1,000 words) in prototype 3 (PIA) Task: PE of Dataset 3 subset 3.3 (1,000 words) in prototype 2 (PI) Task: PE of Dataset 1 subset 1.3 (1,000 words) in prototype 3 (PIA) Task: PE of Dataset 2 subset 2.3 (1,000 words) in prototype 1 (P) 	



A.3 Post-editing times

This appendix provides details on post-editing times across participants for the different texts and systems (P, PI, and PIA) involved in the second CASMACAT field trial.

Appendix 7.9 Post-editing times of individual participants wrt. texts and GUIs



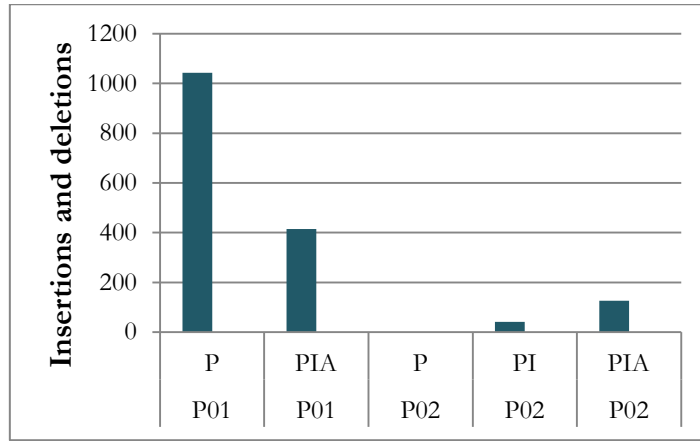
A.4 Total revisions per reviewer

This appendix provides details on the number of revisions made by the four reviewers involved in the second CASMACAT field trial.

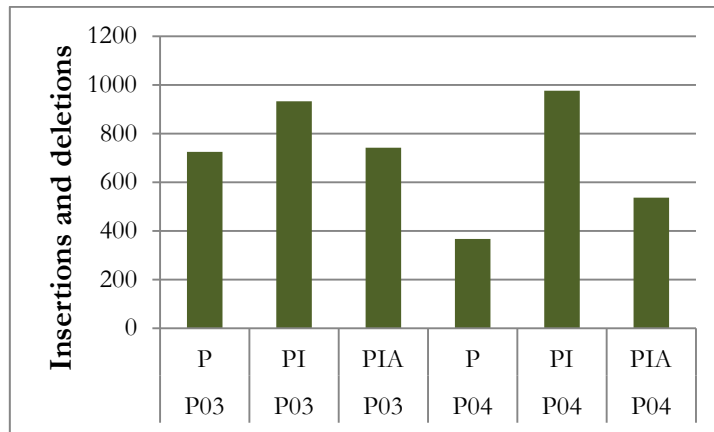
The charts also show the number of the participant reviewed by each reviewer and the system which each post-editors used.

All revisions were made from system CFT1 (the one featuring no interactivity - P).

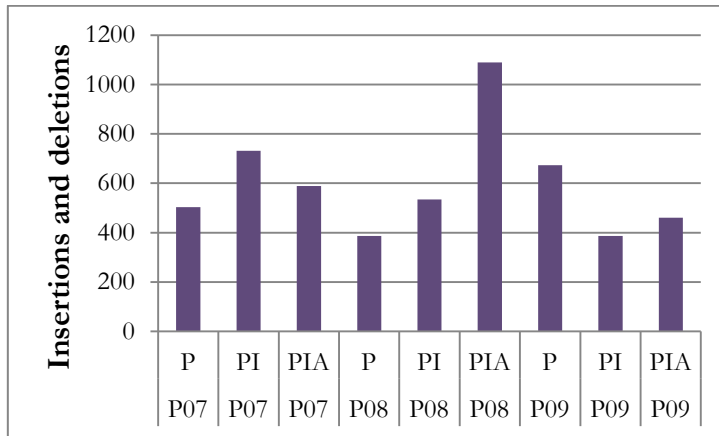
Total revisions per reviewer



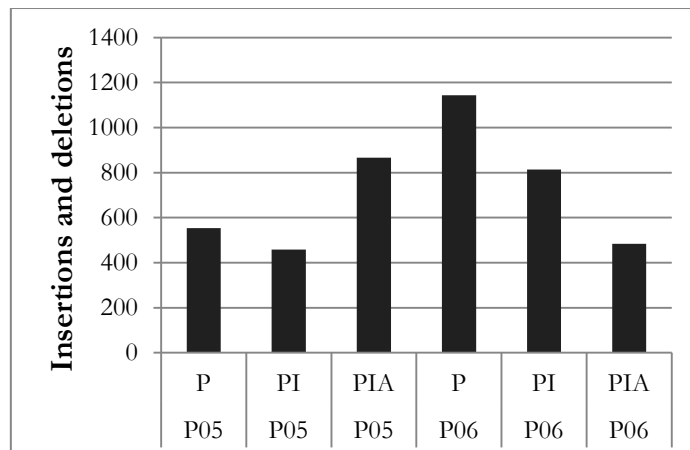
Reviewer 01



Reviewer 02



Reviewer 03



Reviewer 04

A.5 Tracing literal translation alignment in the translator's mind

The cross value feature:

The Cross feature

The Cross feature represents alignment information in a procedural manner. It indicates how many words need to be consumed in the (source) text to produce the next word in the translation. The assumption is that the target text is produced word by word from left to right, while words in the source text are successively consumed to find the one(s) that produce the translation. By following the ST-TT alignment links, the minimum number of words that need to be moved in the ST to produce the successive TT words represents the Cross value. Figure 9 gives an example from an

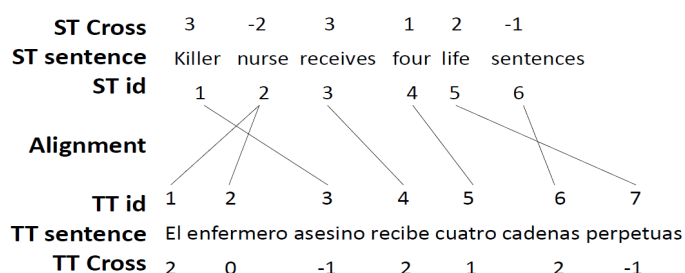


Figure 1 Mapping Alignment information into Cross features

English → Spanish translation. The figure shows for each language side the sentence, the Cross value, the actual sentence and the enumeration of the words in the sentence, in addition to the actual ST-TT links.

A declarative representation of the alignment would relate pairs of ST / TT word ids into alignment sets for instance as follows $\{\{1,3\}, \{2,1\}, \{2,2\}, \{3,4\} \dots\}$. A procedural alignment indicates word reordering relative to the previous alignment link counting the number of tokens that have to be processed to produce the next output: in order to

produce the first Spanish TT word (“El”), two English words (“Killer” and “nurse”) have to be consumed, which results in a Cross value of 2. Since the second source word (“nurse”) emits two adjacent TT words, no further ST word has to be consumed to produce “enfermero”, which results in a Cross value is 0. To produce the third Spanish word, “asesino”, one ST word to the left of ”nurse” has to be processed, leading to the Cross value -1. The next Spanish word ”recibe” is the translation of two words to the right of the current ST cursor position; ”cuatro” one ST word ahead etc. with their respective Cross values of 2 and 1.

A Cross value is also computed for the source language. The ST Cross values provide the alignment values under the assumption that the translation was produced from the target into the source text. Languages with similar word order will have low average Cross values. In a monotonous 1-to-1 translation all Cross values are 1. The more syntactic reordering between source and target text take place the higher the average Cross value will be.

The Cross feature across Texts and Languages

This section describes an analysis of 328 translations of 18 different source texts into six different languages as a subset from the TPR-DB v1.2. It shows that the Cross values correlate with the total reading time per word, as depicted in Figures 6 A, B, C and D.

Only Cross values from -8 to 8 are reported because items with higher Cross values are very rare, resulting in vastly unequal numbers of items. A reference line at Cross value 1 has been inserted in all graphs, given that CrossS (Cross value for the source text) or CrossT (Cross value for the target text) 1 represents the ideal literal translation and the least effortful items to process. Graphs 1A and C depict eye movements on the source text while graphs B and D depict eye movements on the target text. All graphs show that higher CrossS and CrossT values, both positive and negative are more effortful to process than lower CrossS and CrossT values, as indexed by higher total reading times.

A simple linear regression was carried out for all Cross values for total reading time and fixation count to ascertain the extent to which Cross values can predict total reading time and fixation count. The correlation was calculated from Cross value 1 to the peak in each distribution in both directions (negative and positive).

Correlation between CrossS values and Total Reading Time on Source Text

A strong positive correlation was found between negative CrossS values and total reading time on the source text ($r = .63$), but the regression model only predicted 37% of the variance and the model was not a good fit for the data ($F = 3.60$, $p < .11$), suggesting a nonlinear relationship. A strong positive correlation was found between positive CrossS values and total reading time on the source text ($r = .89$) and the regression model predicted 79% of the variance. The model was a good fit for the data ($F = 23.04$, $p < .003$). Note, though, that for every single increase in the positive CrossS value, the total reading time on the source text only increased by 85ms: while source text words with high positive CrossS values are more effortful to process than source text words with lower CrossS values, the increases are modest in comparison to all other Cross values, as shown below.

Correlation between CrossS values and Total Reading Time on Target Text

A strong positive correlation was found between negative CrossS values and total reading time on the *target text* ($r = .92$) and the regression model predicted 84% of the variance. The model was a good fit for the data ($F = 36.97$, $p < .001$). For every single increase in the negative CrossS value, the total reading time on the target text increased by 389ms.

A strong positive correlation was also found between positive CrossS values and total reading time on the target text ($r = .93$) and the regression model predicted 86% of the variance. The model was a good fit for the data ($F = 30.69$, $p < .003$). For every single increase in the positive CrossS value, the total reading time on the target text increased by 301ms.

Correlation between CrossT values and Total Reading Time on Source Text

For negative CrossT values a strong positive correlation was found between negative CrossT values and total reading time on the source text ($r = .97$) and the regression model predicted 97% of the variance. The model was a good fit for the data ($F = 205.7$, $p < .0005$). For every single increase in the negative CrossT value, the total reading time on the source text increased by 516ms.

A strong positive correlation was found between positive CrossT values and total reading time on the source text ($r = .91$) and the regression model predicted 82% of the variance. The model was a good fit for the data ($F = 22.89$, $p < .005$). For every single increase in the positive CrossT value, the total reading time on the source text increased by 347ms.

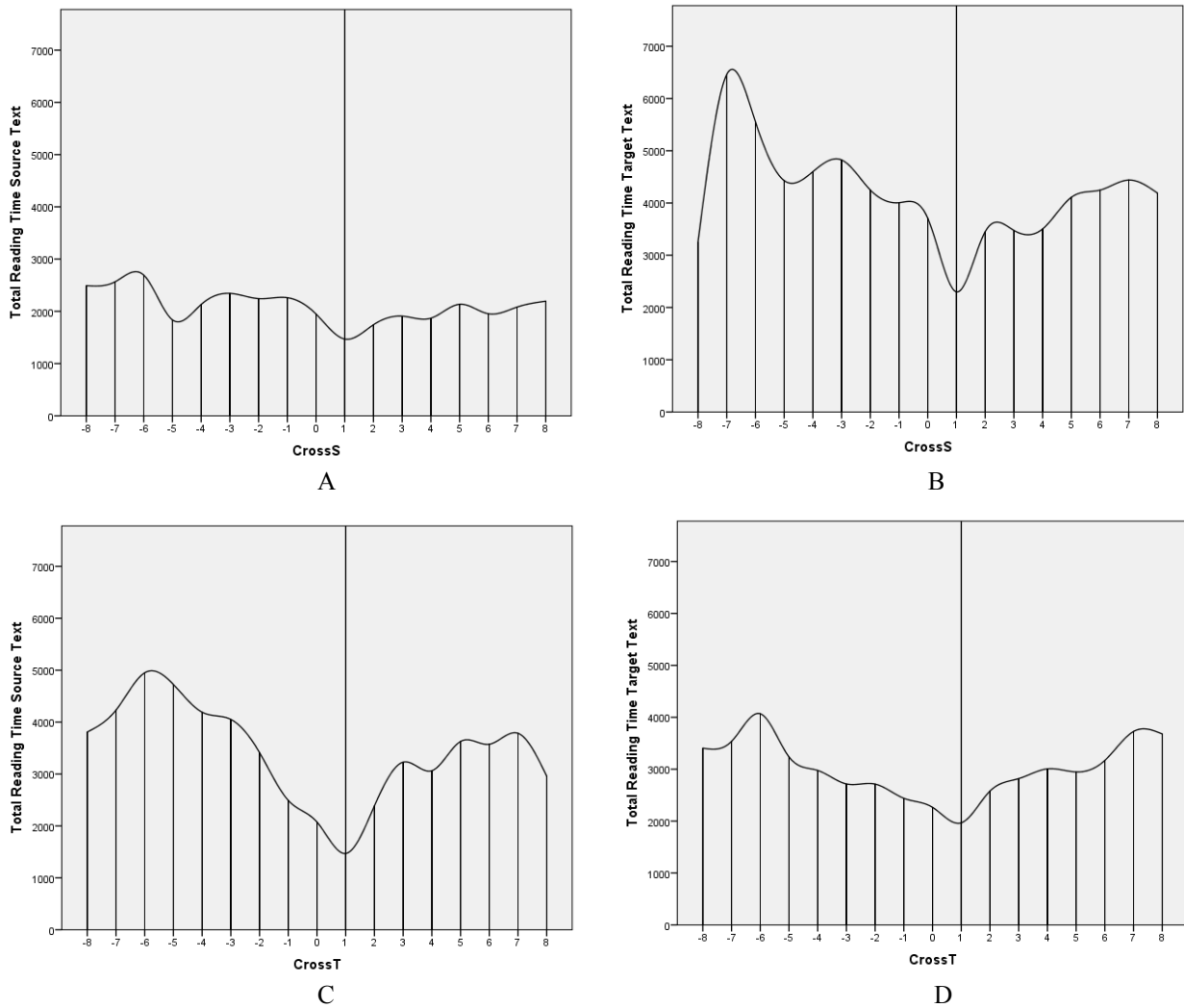


Figure 6 A, B, C and D: Total reading time for CrossS and CrossT on source and target window. Graph 1A plots the total reading time on the source text for CrossS values from -8 to 8, while graph 1B depicts total reading time on the target window for CrossS values from -8 to 8. Graph 1C plots the total reading time on the source text for CrossT values from -8 to 8, while graph 1D depicts total reading time on the target window for CrossT values from -8 to 8.

Correlation between CrossT values and Total Reading Time on Target Text

A strong positive correlation was found between negative CrossT values and total reading time on the target text ($r = .94$) and the regression model predicted 89% of the variance. The model was a good fit for the data ($F = 56.07$, $p < .0005$). For every single increase in the negative CrossT value, the total reading time on the target text increased by 226ms.

A strong positive correlation was found between positive CrossT values and total reading time on the target text ($r = .94$) and the regression model predicted 89% of the variance. The model was a good fit for the data ($F = 39.02$, $p < .002$). For every single increase in the positive CrossT value, the total reading time on the target text increased by 235ms.

The increases in the fixation count on source text words for high CrossT values are particularly large (graph 2C) and the increases for high CrossT values on target words are less pronounced (graph 2D).

A simple linear regression was carried out for all Cross values for fixation count to ascertain the extent to which Cross values can predict fixation count. Essentially the same pattern as for total reading time emerges.

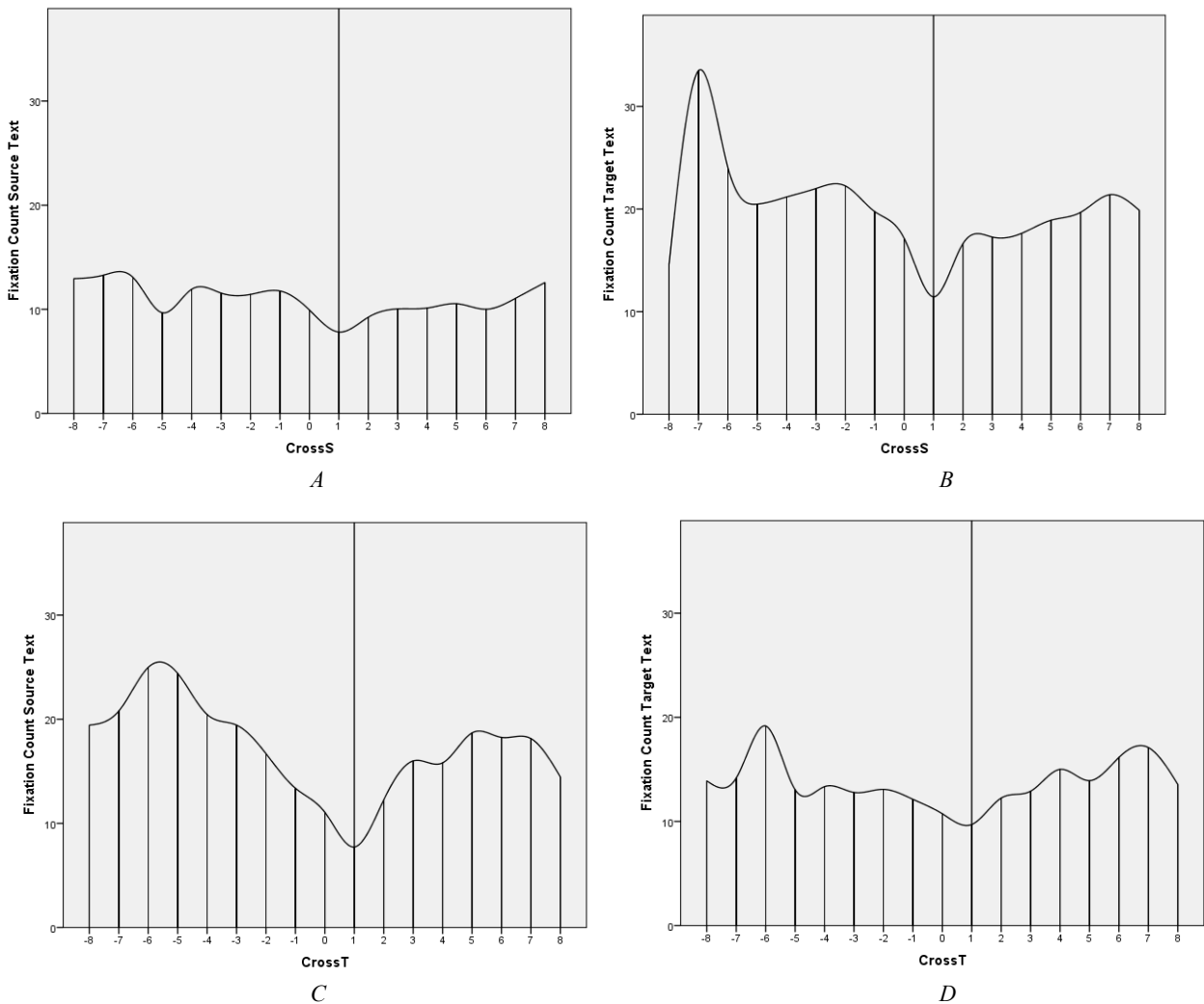


Figure 7 A, B, C and D: Fixation count for CrossS and CrossT on source and target window. Graph 2A plots the number of fixations on the source text for CrossS values from -8 to 8, while graph 2B depicts number of fixations on the target window for CrossS values from -8 to 8. Graph 2C plots the number of fixations on the source text for CrossT values from -8 to 8, while graph 2D depicts number of fixations on the target window for CrossT values from -8 to 8.

Correlation between CrossS values and Fixation Count on Source Text

A strong positive correlation was found between negative CrossS values and fixation count on the source text ($r = .76$) and the regression model predicted 57% of the variance. The model was a good fit for the data ($F = 10.79$, $p < .01$). For every single increase in the negative CrossS value, the fixation count on the source text increased by .412 fixations. For positive CrossS values, on the other hand, a strong positive correlation was found between positive CrossS values and fixation count on the source text ($r = .90$) and the regression model predicted 81% of the variance. The model was a

good fit for the data ($F = 26.02$, $p < .002$). For every single increase in the positive CrossS value, the fixation count on the source text increased by .548 fixations.

Correlation between CrossS values and Fixation Count on Target Text

For negative CrossS values a strong positive correlation was found between negative CrossS values and fixation count on the target text ($r = .84$) and the regression model predicted 71% of the variance. The model was a good fit for the data ($F = 17.34$, $p < .004$). For every single increase in the negative CrossS value, the fixation count on the target text increased by 1.8 fixations.

A strong positive correlation was found between positive CrossS values and fixation count on the target text ($r = .90$) and the regression model predicted 81% of the variance. The model was a good fit for the data ($F = 21.23$, $p < .006$). For every single increase in the positive CrossS value, the fixation count on the target text increased by 1.36 fixations.

Correlation between CrossT values and Total Reading Time on Source Text

For negative CrossT values a strong positive correlation was found between negative CrossT values and fixation count on the source text ($r = .99$) and the regression model predicted 98% of the variance. The model was a good fit for the data ($F = 400.46$, $p < .0005$). For every single increase in the negative CrossT value, the fixation count on the source text increased by 2.46 fixations.

A strong positive correlation was found between positive CrossT values and fixation count on the source text ($r = .96$) and the regression model predicted 92% of the variance. The model was a good fit for the data ($F = 39.00$, $p < .008$). For every single increase in the positive CrossT value, the fixation count on the source text increased by 2.6 fixations.

Correlation between CrossT values and Fixation Count on Target Text

For negative CrossT values a strong positive correlation was found between negative CrossT values and fixation count on the target text ($r = .83$) and the regression model predicted 69% of the variance. The model was a good fit for the data ($F = 13.21$, $p < .01$). For every single increase in the negative CrossT value, the fixation count on the target text increased by .76 fixations.

A strong positive correlation was found between positive CrossT values and fixation count on the target text ($r = .96$) and the regression model predicted 92% of the variance. The model was a good fit for the data ($F = 59.21$, $p < .001$). For every single increase in the positive CrossT value, the fixation count on the target text increased by 1.07 fixations.

It is interesting to note that both total reading time and fixation count peak at Cross values around +/-7. For a Cross value of +/-7, a translator has processed about 7 words once the current token can be mapped onto its equivalent. It has been found that about seven items is a limit beyond which the information stored in memory needs to be either recoded into higher units or it decays (Miller, 1956). The fact that total reading time and fixation count peaks at Cross values around +/-7 may indicate peak performance in translation production without the need to recode or higher level representation. While the Cross value does not necessarily represent the actual reading or writing path that a translator carries out, it nevertheless supports a literal translation hypothesis, which we will discuss in the next section.

Ideal literal default translation

While the Cross values are inferred from the alignment of the translation product and represent the syntactic similarity of the source and target languages, they also seem to carry a psychological dimension. In the translation product, the Cross feature describes an idealized literal translation, because it represents the distance, in number of words from a source target alignment with exactly the same word order, i.e. a Cross value of 6 for a word describes a situation in which this particular word is 5 words removed from an alignment where the source and target sentences have the same word order. The large increases in total reading time for higher CrossT values during source text reading (Figure 6C and 7C) and the large increases in total reading time for higher CrossS values during target reading (Figure 6B and 7D) are best explained in terms of effort related to maintaining and processing an increasing number of items in working memory. The Cross feature thus describes the translation process in the form of ideal literal translations: the ideal situation in which every word of a sentence has a Cross value of 1 serves as a reference with which divergences from this ideal situation are compared. And, as the distance between source and target words increase, also the effort to integrate larger units of both source and target text representations grows, as measured by the increases in number of fixations and reading time.

This suggests that, whenever possible, translators produce translations in as small pieces as possible, keeping one minimal chunk at a time in mind while translating it into the target language, thereby reducing memory load and the

need to anticipate future translations. Similarly, while checking the accuracy of the produced target language only a minimal chunk is re-scanned and mapped on the appropriate piece of source text. In extreme cases, when stretches of Cross value are 1, this translation behaviour may turn into a word-for-word translation, which reduces the need to read ahead in the text and thus decreases the number of items kept the memory buffer to a minimum. Words with a CrossS or CrossT value of 1 are the least effortful to process while higher CrossS or CrossT values require more effort, because more items need to be maintained in working memory and need to be integrated into a larger representations.

These findings corroborate the “literal default translation hypothesis” by which a translator starts out with a literal default rendering, “which goes on until it is interrupted by a monitor that alerts about a problem in the outcome. The monitor’s function is to trigger off conscious decision-making to solve the problem” (Tirkkonen-Condit, 2005)

Schaeffer and Carl (2013) extend the monitor model by a “Recursive model of translation” which re-conciliates horizontal (i.e. literal default rendering) and vertical (i.e. control and monitoring) translation processes in one cognitive architecture. While horizontal translation processes are triggered early through various levels of linguistic priming, the vertical monitor processes emerge as the context becomes available during the translation task. The extended model consists of a recursive cycle where the vertical processes acts as a monitor to assess whether the source text encodes the same meaning as the target, and to make sure that the target is the same as the source. Vertical processes access the output from a horizontal, automatic default procedure, recursively in both the source and the target language and monitor consistency as the context during translation production increases:

During decoding, both horizontal and vertical processes are always active at the same time. We assume that the horizontal process is an early process while the vertical processes depend on context which becomes available later, as processing advances in the chunk or text. Early during source text reading, shared representations are activated which then serve as a basis for regeneration in the target language. As long as the target text being produced conforms to the target norms and contextual considerations of the vertical processes, regeneration on the basis of shared representations is not interrupted. But when the target text is not acceptable, the interim translation, either kept in working memory or already partially produced as target text, is adapted to target norms by vertical encoding processes (Schaeffer and Carl, 2013)

The results of the Cross analysis indicate that translators produce a cognitive alignment between source and target text representations during the translation process. Translators circumvent differences in word order and number by mentally forcing one language to adopt the other language’s word order and number. This process is more effortful for higher Cross values than in a word-to-word translation alignment situation (for Cross values of 1). Such an ideal literal translation is rarely found in actual real texts, and requires more gaze and cognitive activity as the word order in both languages differ. Our findings indicate that this literal default translation procedure may buffer up to 7 items, before higher order monitor processes intervene. It is likely that this process is not only useful as a “check on meaning” (Ivir 1981: 58), but that it also facilitates the process of translation: it is easier to process source items which have the same order as the target text and when there is no such sameness, the translator mentally forces it, postponing the re-ordering to a later stage of processing, i.e. to target text production.

The strong correlation between Cross values and reading time suggest that translators across languages and texts operate on the basis of an ideal literal translation, i.e. an ideal alignment in which the number and order of source and target text items are identical. The correlation for both, the CrossS and the CrossT values, further suggest that translators mentally map the source language on the target language word order when producing the translation. During translation, a translator mentally re-maps the source text word order so that it coincides with the ideal literal translation according to the demands of the target language. Similarly, during translation revision or checking for accuracy the translator mentally re-orders the target language to adopt the source language word order, as evidenced by the large increases in eye movements on the target text for higher CrossS values in the corpus analysis above. Ivir (1981: 58) describes the ideal literal translation thus: even if the translator departs from formal correspondence, he or she “makes use of formal correspondence as a check on meaning - to know what he is doing, so to speak.” Mapping one language to another language’s word order can therefore act as one literal mental representation of two different texts, which may oscillate between the source or the target language as reference, depending on whether the translation is checked against the source text or vice versa.

References

- Ivir, V. (1981). Formal Correspondence vs. Translation Equivalence Revisited. *Poetics Today*, 2(4), 51–59.
- Schaeffer, Moritz and Michael Carl. 2013. Shared Representations and the Translation Process: A Recursive Model. *Translation and Interpreting Studies*, in press
- Tirkkonen-Condit, Sonja. 2005. “The Monitor Model Revisited: Evidence from Process Research.” *Meta* 50: 405-414