D5.4: Final release of the CASMACAT workbench

Mercedes García-Martínez, Dan Cheung Petersen, Chara Tsoukala, Vicent Alabau, Daniel Ortíz-Martínez, Philipp Koehn, Michael Carl

Distribution: Public

CasMaCat
Cognitive Analysis and Statistical Methods
for Advanced Computer Aided Translation

ICT Project 287576 Deliverable D5.4

This project has received funding from the European Union's Seventh Framework Programme for research, technological development and demonstration under grant agreement no 287576.

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	April 30, 2014
Actual date of delivery	April 30, 2014
Date of last update	April 30, 2014
Deliverable number	D5.4
Deliverable title	Final release of the CASMACAT workbench
Type	Report
Status & version	Draft
Number of pages	36
Contributing WP(s)	WP5
WP / Task responsible	UEDIN, CBS, UPV
Other contributors	
Internal reviewer	Philipp Koehn
Author(s)	Mercedes García-Martínez, Dan Cheung Petersen, Chara Tsoukala, Vicent Alabau, Daniel Ortíz-Martínez, Philipp Koehn, Michael Carl
EC project officer	Aleksandra Wesolowska
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/

Executive Summary

This document contains details about the implementation of the 3rd prototype of the CASMACAT workbench as well as the CRITT Translation Process Research Database (TPR-DB). It outlines the improvements of the workbench respect of the previous Deliverable 5.3. This deliverable will be updated in month 36 of the project with further improvements.

Sections 1, 2, 3, 4, 5 and 6 explain the tasks 2,3,4,5,7 and 8 respectively and section 7 addresses the comments of the last review report.

Contents

1 T5.2 Graphical Interface

This section describes the improvements made in the 3rd CasMaCat prototype compared with the 2nd CasMaCat prototype. A description of the 2nd CasMaCat prototype can be found in deliverable 5.3.

1.1 Advanced Interactive Translation Prediction (ITP)

Figure 1: Visualization of advanced features

A bar containing a number of visualization options of the advanced ITP features has been added to the top of the text editor for each translation segment. The user is able to activate the features he wants to use by clicking checkboxes, as shown in Figure 1. The description of the visualization options can be found in [3] and Section 8 (appendix). This bar can also be configured by a special file specified by URL. This file can setup the visualization options enabled by default but also disabled completely.

1.2 Improvements of the eye-tracking module

Two new buttons that allow the user to re-calibrate and to download logging EDF files from Eyelink1000 have been added to the GUI of the Translation View (Figure 2)

Figure 2: GUI heading buttons

These buttons accommodate new features implemented in the browser plug-in. The browser plug-in update includes a method for re-calibration of the eye-trackers, a method for downloading the data file directly from the eye-tracker, a reduction in the polling frequency is implemented to accommodate the need for more computing power other than JavaScript calls.

The eye-link calibration screen has been revamped and now also includes methods for validation of calibration, (Re)calibration, camera view and drift-correction. Pressing Re-calibration button in the GUI prompts the calibration screen. Download of the Eyelink1000 data files (*.edf) has been added, so it is possible to analyse gaze and fixation data with the Eyelink analysis software.

1.3 Floating prediction

Floating prediction (Figure 3) is an alternative way to present the completion suggestions to the user. The autocompletion is no longer inserted in the editor; instead, the next three predicted words are shown in a floating box next to the caret position. The user can then accept the word prediction that is in bold by pressing `TAB`, or ignore it and keep typing. If the user continues or starts typing, new suggestions are generated.

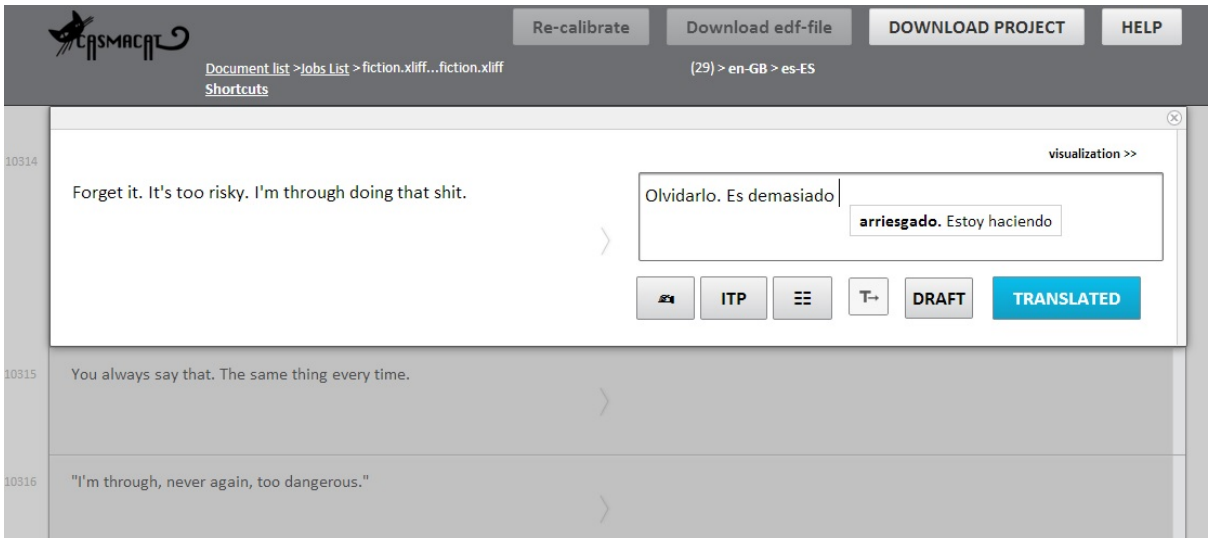

Figure 3: Floating prediction

1.4 Biconcordancer

The 3rd CasMaCat prototype also provides the biconcordancer functionality. The user can select a word from the source text and either press the biconcordancer button or type `CTRL+B`. Alternatively, the user can press the biconcordancer button and type a word in the textbox within the biconcordancer window that pops up. An example of the biconcordancer is shown in Figure 4; alternative translations (in bold) of the selected input words are presented in context, which helps the users to disambiguate the word.

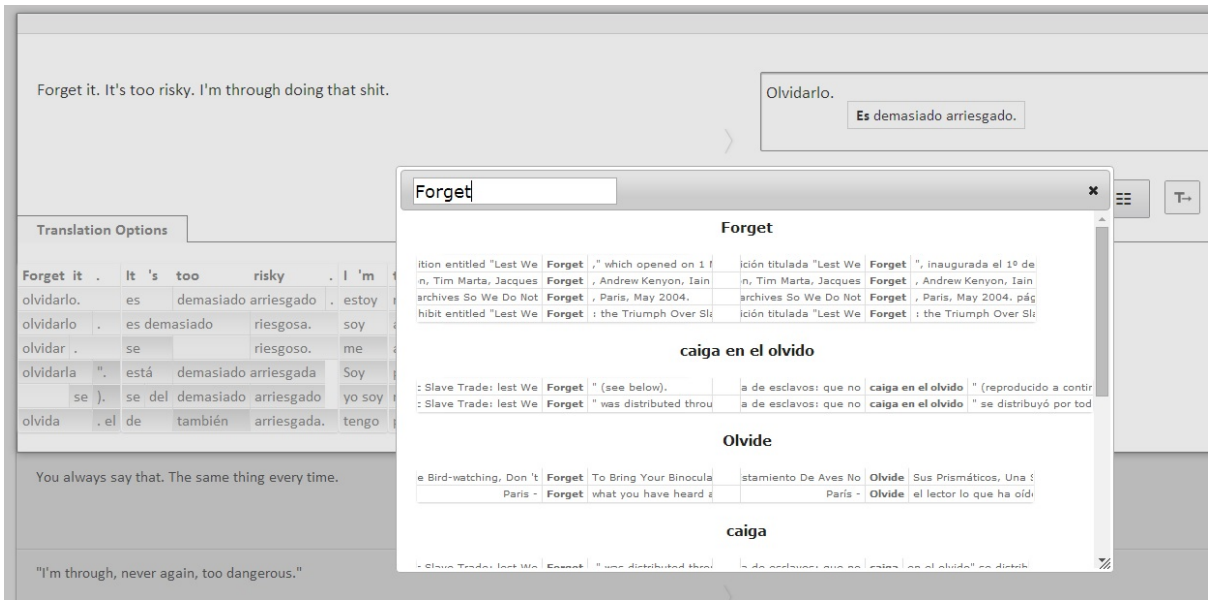

Figure 4: Biconcordancer

For further information see previous deliverable 3.2 section 5.

1.5 Translation options

Another feature which has been added to the 3rd CasMaCat prototype is information regarding alternative translation options for each segment of the source phrase, as shown in Figure 5. The user is able to see a table below the current segment, where alternative ways to translate a

covered source segment are displayed. The options are displayed in levels, according to their probability scores. The ones that are on top are more likely to be close to the translation the user is looking for, whereas the ones that are displayed at the bottom of the table are marked with a darker colour to indicate that they are less probable. If the user wants to add one of the options to the text editor, he can click on that option and it will be automatically pasted in the target window.

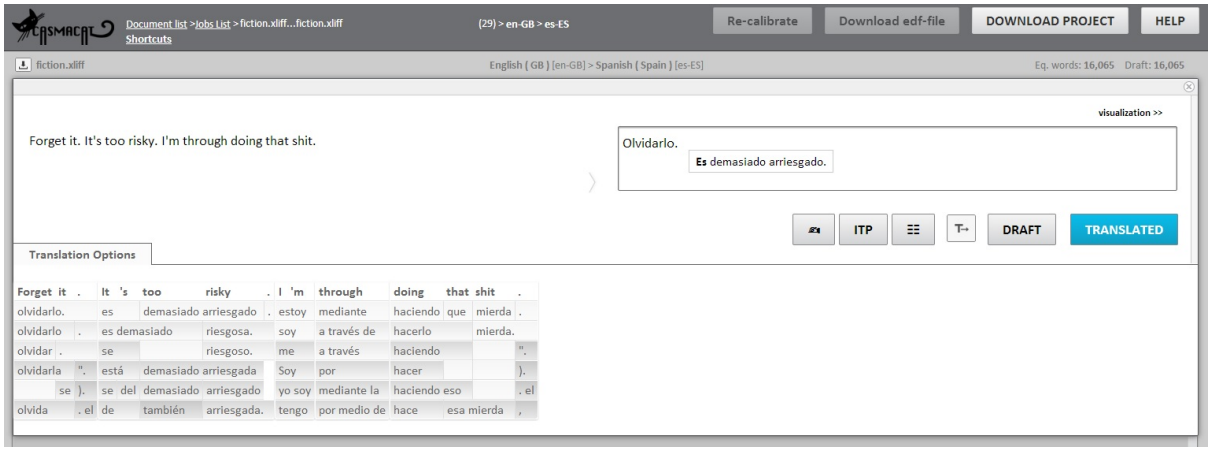

Figure 5: Translation options

For further information see previous deliverable 3.2 section 5.

1.6 Alternative ITP visualization

The second field trial (2013) and other lab tests showed that the way predictions were presented could be annoying/confusing for the user. Thus, we propose an alternative visualization to alleviate such problems. Similarly to the *floating prediction* feature presented above, in alternative ITP visualization the system suggestions are shown in a floating area so that the user is not bothered with constant changes in the suffix. However, in contrast to *floating prediction*, here a whole translation is shown at any moment. The floating predictions show the differences of the current translation and the system suggestion up to the first coincidence (Figure 6). In addition, the size of the floating area indicate which target words are going to replaced when the user presses `TAB`.

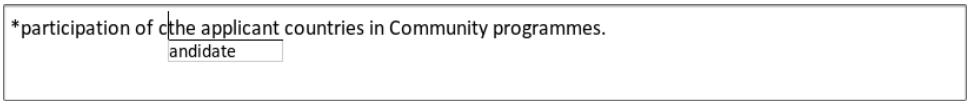

Figure 6: Alternative visualization of ITP shows only the changes up to the first coincidence between the text being edited and the new alternative hypothesis given by the system.

1.7 List of documents

A new function has been added to the list of documents view in the 3rd CasMaCat prototype to check the logging files quickly. When the user selects a documents and presses the button "check logging", the user gets a new file in the downloads folder (http://URL/downloads) with the main information about the logging of the selected file (configuration, number of fixations, translations and events). Figure 7 shows the new view of the list of documents.

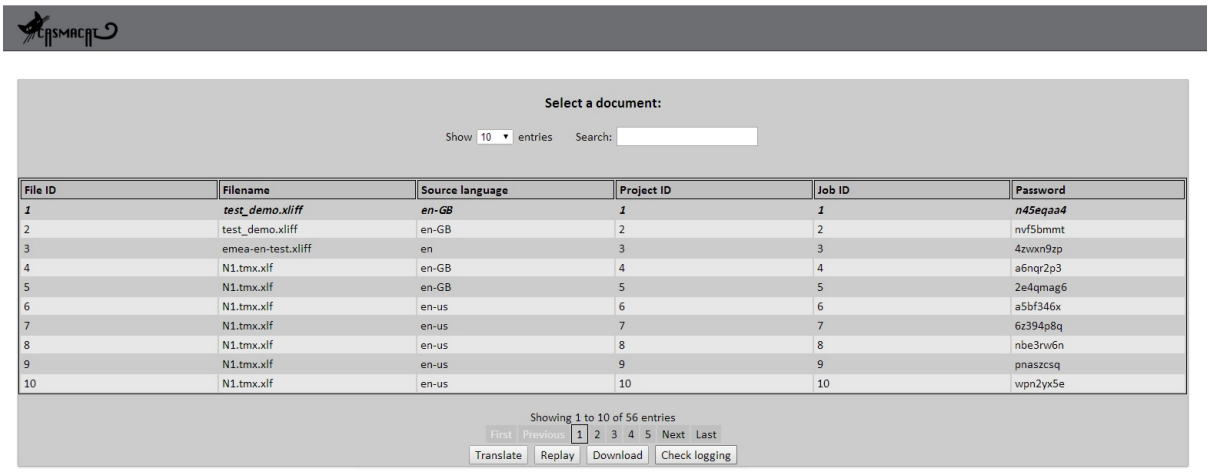

Figure 7: List of documents

1.8 Demo

The 3rd CasMaCat prototype can be tested by submiting a request in the *Workbench* section of the official CasMaCat website: `casmacat.eu`
`http://casmacat.eu/index.php?n=Workbench.Workbench`

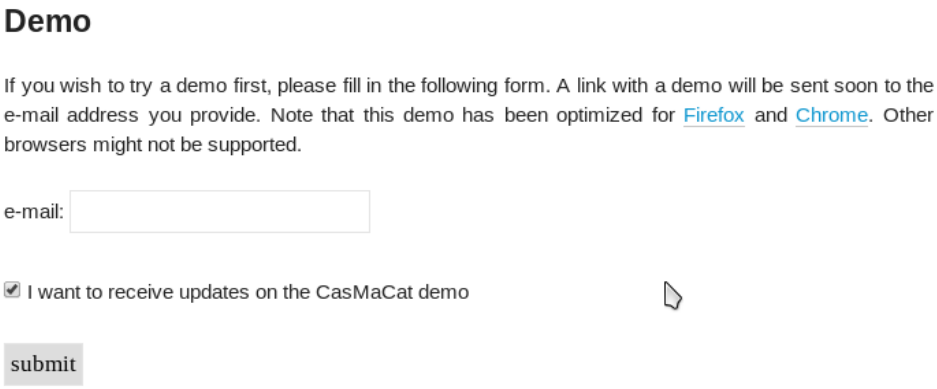

Figure 8: Demo request

The form will automatically upload a demo document to the server and create an URL that will be only accessible to the user. The URL will be sent to the user by e-mail. In case the user tries to re-submit a request, a page will appear with a message pointing to the assigned URL.

2 T5.3 E-pen Interaction

In the previous workbench we enabled e-pen interaction. It consisted of an online HTR recognizer that allowed to substitute words, and a MinGestures gesture recognizer that allowed insertions, deletions, undo and redo actions among other things. In this release of the CasMaCat workbench we have two new features:

2.1 N-best lists

The online HTR recognizer can retrieve an arbitrary number of alternative transcriptions to the user's handwriting, i.e. a list of n-best solutions. By default this number is set to 10 but

can be changed. The GUI has been modified so that, at the bottom of the drawing area, a grey box appears as a notification area (Figure 9). It gives feedback of which gestures are being recognized as well as the status of the HTR recognizer. When a word is recognized it also displays a list of the n-best options. Then, the user can be selected one of such words to replace the original recognition by clicking on them. Until the user produces another gesture or tries to substitute another word, the n-best list is mantained on screen.

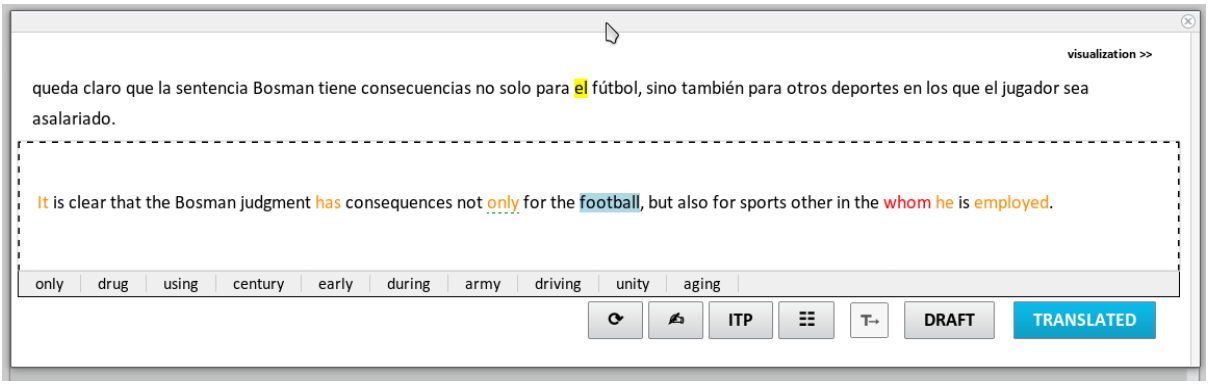

Figure 9: A list of the 10-best recognition results appear at the bottom of the e-pen drawing area. If the user clicks on any of the 10-bests the text in the edit area is automatically replaced.

2.2 Text selection

In order to allow for the gestures and handwriting to operate on a segment of words, the user has to select such segment first. Thus, by clicking on one word the selection mode is enabled (Figure 10). That word establishes the beginning of the selection. Then the user must click on other word to close the selection (Figure 11). Now, any gesture or handwriting issued will only affect the selected text.

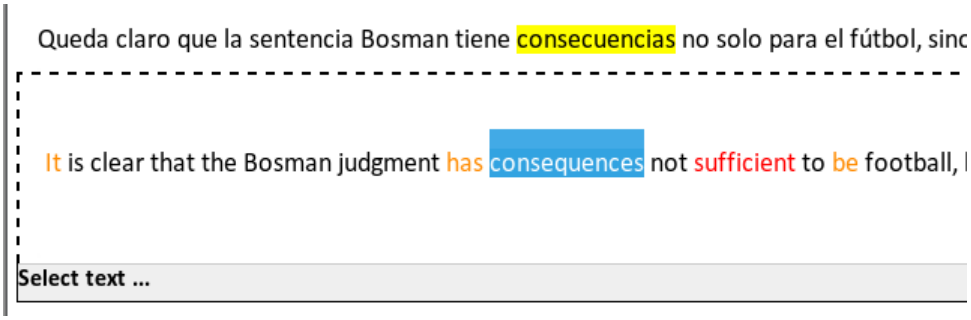

Figure 10: Double click enables selection mode by selecting the word under the e-pen.

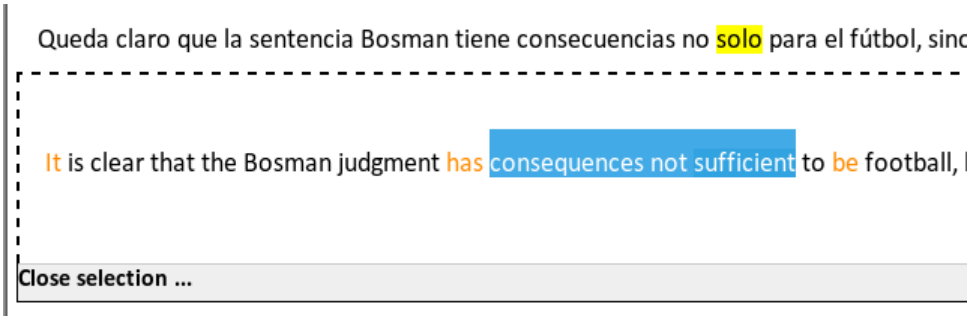

Figure 11: When in selection mode, a click indicates the last word to be selected.

3 T5.4 Logging Functions

When the user presses "Download" in the list of documents view (see Figure 1.7), a new XML file[1] is generated with the logging information for the selected file in the downloads folder. This XML file contains information about the following translation process research items:

- Source and initial and final translation of each segment.
- Configuration of the sessions.
- Start and stop times for the session.
- Segments opened and closed.
- Keystrokes.
- Fixations on source/target text.
- Scroll and resize activity.
- Search and replace used.
- Suggestions loaded and suggestion chosen.
- Mouse activity.
- ITP (Interactive Translation Prediction).

The logging functions have been extended in the 3rd CasMaCat prototype with the following items:

- Full target: The user gets the full target text each time the translation is changed.
- Previous target: The user also gets the full target text from the previous edition of the translation.
- Type of edit: The user gets the type of edit (manual, from ITP, copy and paste, search and replace, etc.).
- New CasMaCat features: The logging has been extended with the new features (floating prediction, biconcordancer, translation options and gaze to word).
- The scripts from CRITT TPR[2] database to check to log files have been modified to check these new features logged.

[1] See an example of XML log file in:
http://bridge.cbs.dk/prototype3/casmacat-el/downloads/log_id1_test_demo.xliff.xml

[2] This data is available on-line: http://bridge.cbs.dk/platform/?q=CRITT_TPR-db.
See also deliverable 1.2 section 2.4

4 T5.5 Machine Translation Server

During the second field trial, experiments with professional translators from CELER Soluciones SL were carried out to evaluate the performance of ITP compared with standard post-editing. In addition to this, for the third and last field trial, a translation system with online learning will be compared with a conventional system where the statistical models are not updated for newly available training pairs.

The above novel functionality, namely, ITP and online learning for statistical machine translation (SMT), have not previously received much attention from open-source software developers. The lack of previously existing public software poses important implementation challenges when developing the CasMaCat Workbench. In an effort to make ITP and online learning publicly available to the machine translation community, we have developed a new version of the Thot toolkit for SMT [1, 2].

4.1 The Thot Toolkit for SMT

The Thot toolkit was initially designed to train phrase-based models but its functionality has been greatly extended during the last years, providing a state-of-the-art SMT system as well as tools to estimate all of the models involved in the translation process.

This is a list of the different features included in the toolkit:

- Phrase-based statistical machine translation decoder.
- Computer-aided translation (post-editing and interactive machine translation).
- Incremental estimation of all of the models involved in the translation process.
- Robust generation of alignments at phrase-level.
- Client-server implementation of the translation functionality.
- Single word alignment model estimation using the incremental EM algorithm.
- Scalable implementation of the different estimation algorithms using Map-Reduce.

During the CasMaCat project, the work on the Thot toolkit has been strongly focused on developing scalable and incremental training techniques as well as on providing efficient ITP functionality following a well founded statistical formalism. Thot can be easily integrated into the CasMaCat workbench, providing the basic and advanced SMT features that are required within the different field trials.

Thot has been coded using C, C++ and shell-scripting. Thot is known to compile on Unix-like and Windows (using Cygwin) systems. It is released under the GNU Lesser General Public License (LGPL) and can be downloaded from GitHub (`https://github.com/daormar/thot/`).

4.2 Home Edition

While the CasMaCat workbench has been designed as a web-based platform, which enables its installation on a central server and accessed from anywhere over the web, it is also an objective of the project to give individual translators the ability to install the workbench on their own machine. Such a local installation guards intellectual property concerns about translation memories used for training and project files that are edited.

The CasMaCat workbench is a complex software product consisting of many different components and external tools. Many of these tools have been developed and tested on Unix-based operating systems (mostly Linux and MacOS) and it would be significant work and beyond the capabilities of this project to port them to other operating systems. However, most home users are using Microsoft Windows as operating system on their home computers.

We address this conundrum by providing a packaged solution that installs within a virtual machine running the Ubuntu flavour of Linux. Installation is straightforward:

- The user installs virtual machine software such as Oracle's Virtualbox, which is freely available.
- The user creates a virtual machine and installs Ubuntu Linux within it.
- The user installs the CasMaCat workbench on the virtual machine.

Installation of the CasMaCat workbench on the virtual machine is done by the download and execution of an install script. Detailed instruction for this installation process, including setting up a virtual machine, can be found on the web at

`http://www.casmacat.eu/index.php?n=Installation.HomePage`

See Figure 12 for a screenshot of the installation process. Upon installation, the CasMaCat workbench can be controlled over a web based interface.

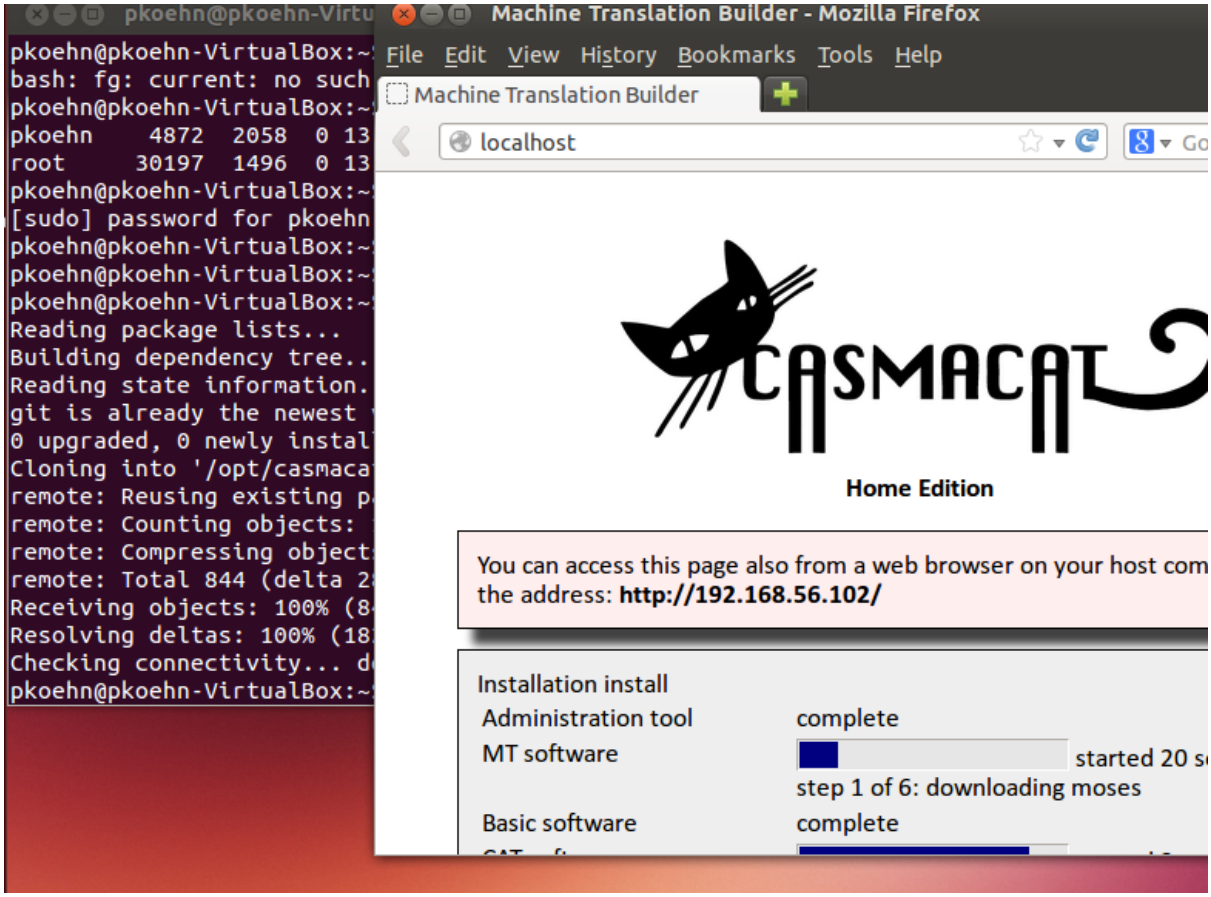

Figure 12: Installation of the CasMaCat Home Edition

The most common use case of the local installation of the CasMaCat workbench by an individual translator involves building a customized machine translation system on her private

translation memory. The CasMaCat Home Edition administrative user interface allows the configuration of the training pipeline of the Moses machine translation toolkit to create such a system.

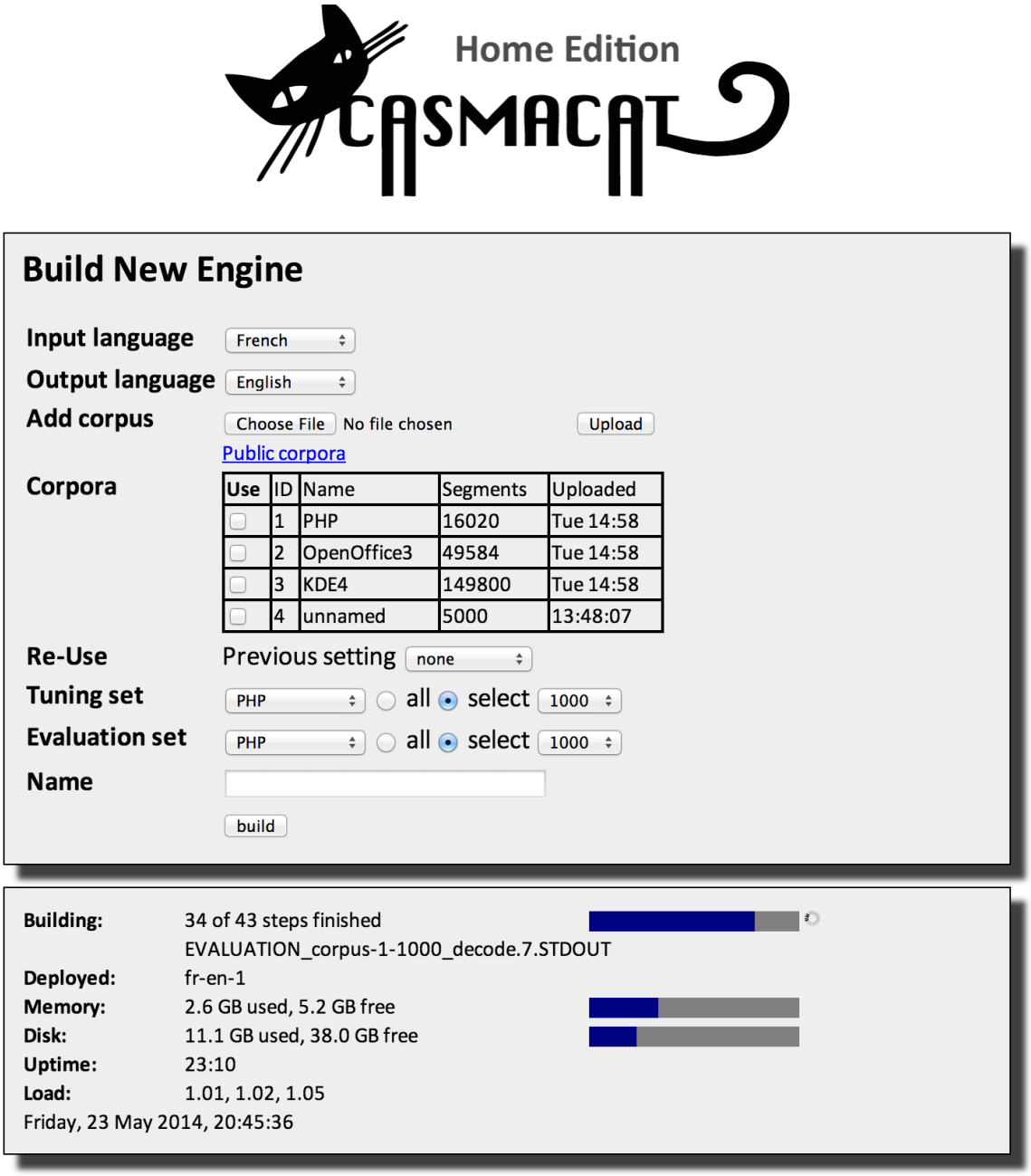

Figure 13: Setup building of a machine translation system in the CasMaCat Home Edition

Figure 13 shows a screenshot of the setup screen for building such a system. System building may also use publicly available corpora. The Home Edition links directly to corpora available on OPUS (`http://opus.lingfil.uu.se/`) and additional corpora made available by the CasMaCat project from public sources, such as training data created for the WMT and IWSLT machine translation evaluation campaigns. A user may build several systems for any language pair, and deploy the most appropriate machine translation engine for each translation project.

At the time of the writing of the deliverable, a fully functional version of the CasMaCat Home Edition has been developed and is being tested by early adopters. We expect to extend and improve it throughout the end of the project and beyond.

4.3 Moses MT Server

Using the same API as the Thot Toolkit, server infrastructure around the Moses toolkit has been extended to offer much of the same functionality. The Moses MT server has been described in previous deliverables.

Significant recent improvements are:

- Support for display of translation options
- Full support for word alignment visualization
- Incremental updating of the translation model

5 T5.7 Automatic Gaze-to-Word Alignment

The 3rd CasMaCat prototype includes a new function for automatic gaze-to-word alignment. The system implements an automatic gaze-to-word Alignment algorithm using gaze and fixation sent from an eye-tracker device via the browser plug-in to the JavaScript algorithm. This section only describes the JavaScript function.

Automation of the gaze-to-word algorithm is handled by JavaScript and the browser plug-in. The conversion of gaze data to screen position is handled by the browser plug-in and then sent through the browser to be caught by JavaScript, which is then able to decipher the screen position, converting it into a caret position (an indication of place in written text) to provide a single character. The function is generic, so it it can accommodate multiple characters or words if necessary.

In addition to the characters found from the gaze data, two additional possible candidates of Gaze-to-Word Alignment are also calculated by the JavaScript algorithm, these character candidates are found by adding or subtracting the text line height, returning the character just above and below the character. Such candidates are used for further disambiguation during the replay view.

All three gaze-mappings are saved in the log. Providing the possibility to alter the Gaze-to-Word gold standard.

For further information see section 5 in the previous deliverable 5.3.

6 T5.8 Replay Mode for User Activity Data

An additional option has been added to the CasMaCat GUI of the Replay View accommodating gaze-to-word alignment visualizations.

Gaze-to-word alignment visualizations for the Replay View display the information gained thanks to the automatic gaze-to-word alignment algorithm. This is visualized by three coloured boxes (magenta, yellow and cyan) surrounding the characters in focus (see Figure 14). These boxes are clickable, so that the user can choose to change the golden standard character if necessary (e.g. to fix driftings). The golden standard is displayed with a green border.

A toggle has been added to the header of the Replay view (see Figure 15) to enable or disable displaying the gaze-to-word Alignment visualizations.

For further information see section 6 in the previous deliverable 5.3.

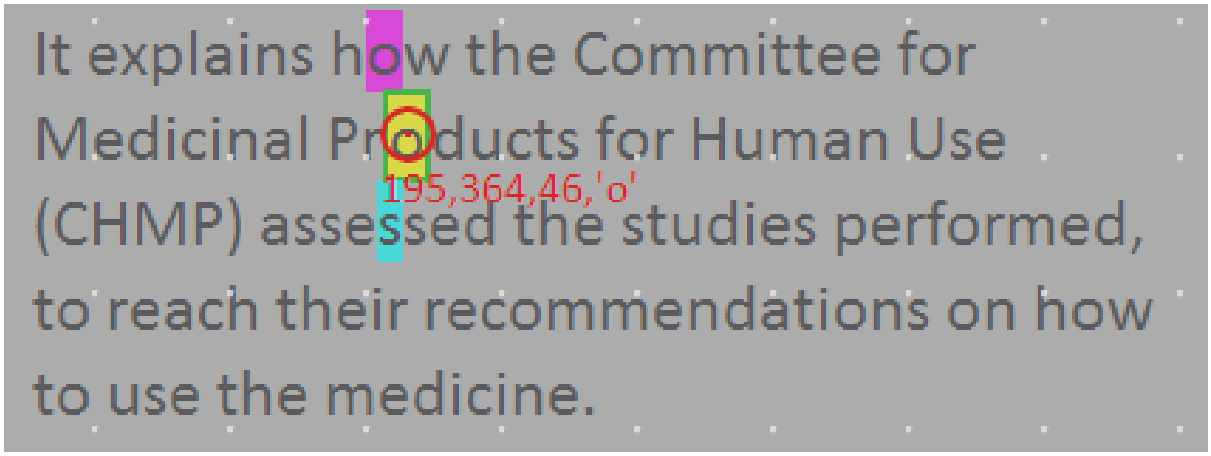

Figure 14: Gaze-to-Word Visualization in the Replay View

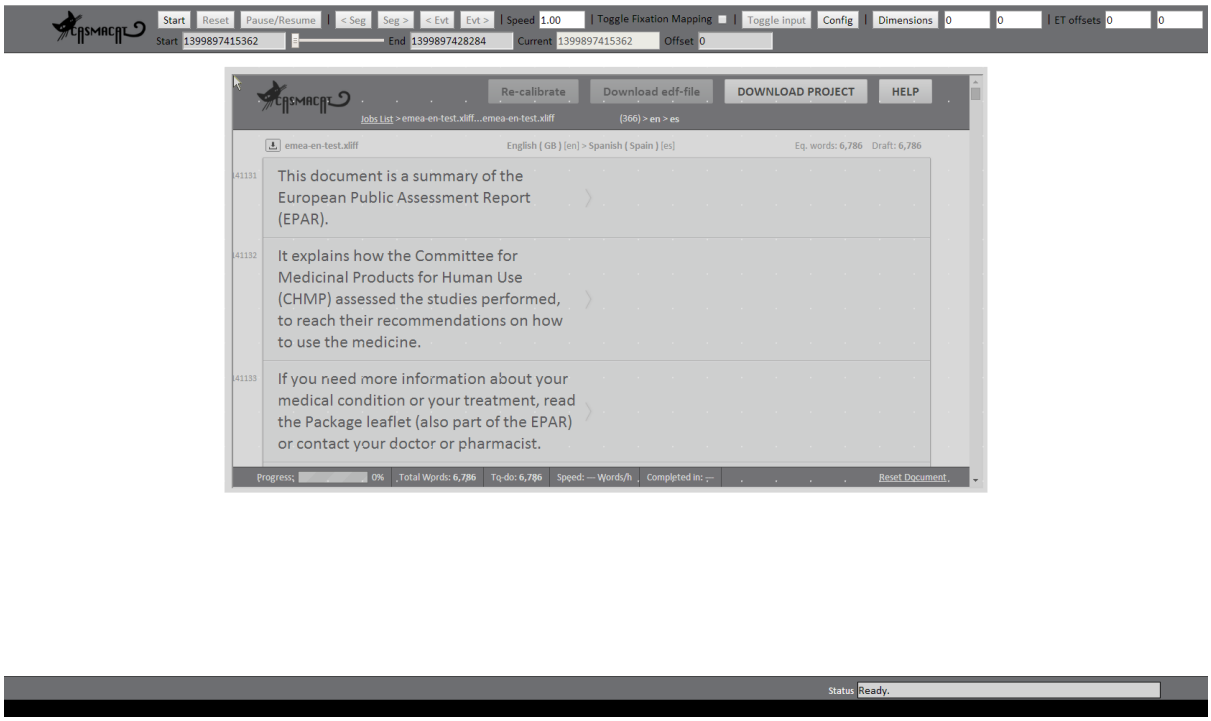

Figure 15: Replay view

7 Review report: addressing reviewers' comments

This section addresses the reviewers' comments for the 2nd review report (Project period M13 to M24) WP5 pages 11-13.

- **The workbench that included the novel functionalities was available for a Linux Workstation solely, and reviewers could not install (...) However, they tried the online version. (...) The system seemed "frozen" at the analysis step.**

  We have included in section 1.8 information about the demo with the ITP functionalities working and we also have included in section 4.2 information about the home edition to be able to run it in Windows as a virtual machine.

- **Furthermore, the Web-based interface should showcase the CasMaCat workbench's functionalities but not just to mirror the MateCat workbench if one does not have some prerequisite data and configuration (...).**

At the second year of CasMaCat project, it was decided to merge the interface of MateCat and CasMaCat due to the similarities between the two projects and in order to benefit from having a common interface. The functionalities can be switched on and off.

- **How do all novel functionalities developed by the project integrate into the final version of the CasMaCat workbench (both standalone and Web-based service)?**

  Based on the experiments run with the 2nd CasMaCat prototype, we have worked on the following items:

  1. Floating prediction: some translators complained about the translation changing automatically. So, we propose an alternative visualization option to improve user satisfaction (section 1.6).
  2. Confidence measures shown in colours: some translators suggested that less colours should be used to show confidence measures at the word level, so we have decided to use the confidence measures to trigger the amount of text automatically validated when the user presses `TAB` (the text is painted in black from that point to the next word with a low confidence measure). See [3] and (Section 8).
  3. Alignments: translators suggested in the field trials that they prefer to trigger alignments using the mouse instead of having them automatically shown by default using the cursor position.

- **How does the project solve the formatting and encoding issues within the shared workbench? (...).**

  It is not in the scope of the project to work with all possible file formats. The CasMaCat workbench only works with UTF-8, which it is the most common and complete format.

- **How does the project solve the issue with a multi-user capabilities of the shared workbench?**

  MateCat was thought to be able to have capability for multi-user, but when CasMaCat and MateCat merged this functionality was not implemented. The integration of multi-user capability could certainly be implemented in the future, because the tool is prepared to such an implementation. Currently, it is possible to distinguish different users by file and user IDs.

- **Task 5.2. During the review meeting, it was mentioned that there were some difficulties with the way the predictions were presented, and it was annoying/confusing for the tester.**

  As stated in the 3rd item of this section 7 is shown, an alternative visualization option has been implemented to address such comments from the users (section 1.6)

- **Task 5.3. What is the respond time for the e-pen functionality implemented via API?**

  The handwriting recognition system has been tuned for a compromise between performance and accuracy. Currently, the system responds with the 10-best list in less than 1.5 seconds, including network communication. It is not rare that, in some cases such as short texts or very clean handwriting, the response time can be less than one second. On the other hand, the gesture recognizer is executed in the client side and the algorithm is very fast. Results are obtained in a fraction of a millisecond.

- **Task 5.5. The translation systems Moses and Thot are called via API that extends the basic implementation of the Google Translation API by returning additional information, (...). The MT server also accepts new sentence pairs**

for incremental updating, tokenises monolingual sentences, and aligns words in sentence pairs. UPVLC integrated its ITP server (...) into the CasMaCat workbench. A new version of the Thot system was released. During the review, it was clear that upon editing a word, ITP would predict the whole of the rest of the sentence. It was discussed that the project should consider the functionality whereby single words or phrases could be edited in isolation without changing the whole of the rest of the sentence.

This was achieved in two different ways in the previous version of the workbench. When users changed the prefix (the segment of the translation from the first word to the last edited word) ITP was disabled by default. Also, users could disable ITP on demand by clicking on the ITP button or by pressing `ESC`. In both cases, the system would behave just like a traditional post-editing tool (without interactivity). In addition, in the current version we have developed two extra kinds of floating prediction. One of them, the one presented in Section 1.3, proposes segments and not full suffixes. A second one, the alternative visualization presented in Section 1.6, predicts whole suffixes but only displays the changes up to the first coincidence with the previous suffix, leaving the rest of the sentence unchanged. Thus, the user always sees a whole sentence but the prediction affects only a part of it.

- **Task 5.7: Two methods were implemented and reported in two publications (...). Some difficulties coordinating the gaze information with the word alignments were experienced during the work due to imperfect calibration of the eye trackers.**

  Further work has been carried out in regard with this issue. Section 5 describes it and it will be tested in June 2014. The next update of the deliverable will show the results.

- **Task 5.8. The replay mode functionality was further implemented, and the presentation of the log information was improved with extra information (...). Other functionalities are planned to be added in due course (e.g., search through the replay).**

  Searching in the the replay mode can be done by introducing the required time in the heading tool of the replay as well as going to the next or previous segment or event (see Figure 15).

- **Task 5.9. Fixations on the source text, insertions, (...) were visualised in TPR-DB (from the submitted paper "Feature Representation in the Translation Process Research DB", the author(s) are not mentioned in the deliverable). During the review, some of the functionalities were demonstrated (however, the paraphrasing tool was not shown that would be interesting to see). Using the link to the Web-based server from the deliverable, it is not possible to see the functionalities visualised, as one has to have some prerequisite data and configuration. Otherwise, the CasMaCat workbench available online is the mirror of the MateCat workbench. The list of languages in GUI is also shared between two projects (...).**

  The last version of a submitted paper "Feature Representation in the Translation Process Research DB" is available in: https://dl.dropboxusercontent.com/u/7757461/TPR-DB1.4.pdf.

  It must be taken in to account that the paraphrasing tool is not planning to be included in the CasMaCat GUI; for more information see task 3.6 of the project.

  The source and target languages in the workbench are shared between MateCat and CasMaCat. The future users of the workbench will be able to work with such a variety of languages, although we only focus on a small group for testing purposes (English, Spanish, German, Danish).

References

[1] D. Ortiz, I. García-Varea, and F. Casacuberta. Thot: a toolkit to train phrase-based statistical translation models. In *Machine Translation Summit*, pages 141–148, Phuket, Thailand, September 2005.

[2] D. Ortiz-Martínez and F. Casacuberta. The new thot toolkit for fully automatic and interactive statistical machine translation. In *14th Annual Meeting of the European Association for Computational Linguistics: System Demonstrations*, pages 45–48, Gothenburg, Sweden, April 2014.

[3] G. Sanchís-Trilles, Alabau V., C. Buck, M. Carl, F. Casacuberta, M. García-Martínez, U. Germann, J. González-Rubio, P. Hill, R. L.; Koehn, L. A. Leiva, B. Mesa-Lao, D. Ortíz-Martínez, H. Saint-Amand, and C. Tsoukala. Interactive translation prediction vs. conventional : Post-editing in practice - a study with the casmacat workbench. Forthcoming 2014.

8 Appendix

This appendix contains a forthcoming publication in a special issue on post-editing in the MT journal that gives more details to the advanced interactive translation prediction in Section 1.1.

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

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

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

Received: date / Accepted: date

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

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

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

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

School of Informatics
University of Edinburgh
Edinburgh, United Kingdom

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

Department of International Business Communication
Copenhaguen Business School
Copenhaguen, Denmark

required to produce the final translation. Moreover, in the ITP setting translators require fewer key strokes to arrive at the final version of their translation.

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

1 Introduction

Contemporary professional translators rarely produce translations entirely from scratch. Instead, they increasingly rely on Translation Memories (TM), that is, data bases of texts that have already been translated, and their translations. At translation time, translations of text fragments similar to the actual source text are retrieved from the data base and edited by the translator to bridge the mismatch between retrieved text fragments and an actual correct translation of the current source text. As the quality of the raw output of fully automatic machine translation (MT) systems is on the rise, so is the commercial interest in integrating MT as an alternative or supplement to traditional TMs into the professional translation workflow. Recent studies [21, 30, 15, 16, 17] have concluded that post-editing is, on average, more efficient than translating from scratch. However, the optimal form of interaction between man and machine in the context of translation is still an open research question.

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

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

1. It offers *interactive translation prediction* [6] (ITP) as an alternative to classical post-editing. The ITP functionality used in this study has been implemented by means of the Thot toolkit for statistical MT [29]. Various auxiliary features and customizations have been implemented to help tailor the MateCat tool to the individual translator's preferences. They are described in Sec. 2.
2. CasMaCat can log user activity in detail and with precise timing information: key strokes, mouse activity, and translator's gaze (if used in combination with an eye tracker). Without eye-tracking, the tool can be easily deployed in a web browser, eliminating the need for specialized hardware or software to run experiments. The logs from the user study discussed in this paper are available online for further analysis at `http://bridge.cbs.dk/platform/?q=CRITT_TPR-db`.
3. CasMaCat can be used with an e-pen as an alternative input device [2]. There are a number of situations where such an interface is comfortable and effective. First, it is suited for post-editing sentences with only few errors, as it is often the case for sentences with strong fuzzy matches in translation memories, or during revision of human post-edited sentences. Second, it allows to perform such tasks while commuting, travelling or away from the desk for other reasons.

[1] `www.matecat.com`

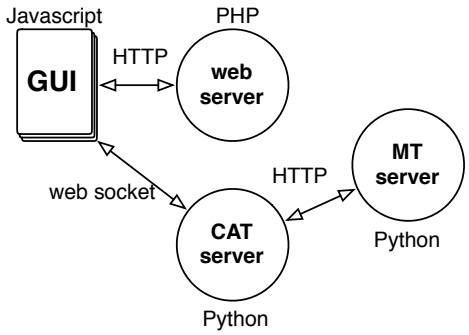

Fig. 1 Components of the workbench

The epen interface is also able to recognize gestures for interactive text editing, using a highly accurate, high-performance gesture recognizer [26].

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

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

2 The CasMaCat workbench

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

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

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

The browser-based GUI and the CAT server communicate via web sockets for speed; the other communication pathways are handled over HTTP for maximum compatibility with other software components. For example, the communication between CAT server and MT server relies on an extension of the Google Translate API, so that other MT engines compliant with the Google Translate API can

Fig. 2 Screenshot of CasMaCat with optional visualization features disabled

easily be swapped in if desired. The web back-end accepts translation job uploads and offers file downloads in standard XML Localization Interchange File Format (XLIFF).

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

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

- **Intelligent autocompletion**: This is the fundamental interactive prediction feature of the CasMaCat workbench. Every time a keystroke is detected, the system produces a translation prediction for the entire sentence in accordance with the text that the user is writing or editing. The text to the left of the cursor is assumed to be approved by the human translator and serves as a prefix to identify the highest-scoring automatic translation that overlaps in this prefix. The remainder of the current transation prediction (to the right of the cursor) is then replaced with the updated prediction. The basic ITP feature is always enabled in ITP mode. ITP mode can be engaged by pressing the button labeled ITP and disengaged by pressing the button labeled PE (post-editing) below the text edit box. Post-editing mode and ITP mode are mutually exclusive.
- **Prediction rejection**: The current CasMaCat prototype also allows the translator to scroll through translation options by use of the mouse wheel [31]. When the mouse wheel is turned over a word, the system invalidates the current prediction and provides the user with an alternate translation option in which the first new word is different from the one at the current mouse position. This option is one of the advanced ITP features.
- **Search and replace** (even in future predictions): the workbench extends standard search-and-replace functionality to future translation predictions. Whenever a new replacement rule is created, it is automatically propagated to the forthcoming predictions made by the system, so that the user only needs to specify them once. This specific function was implemented in response to user feedback in the first field trial of the tool. Note that this option implements a collection of replace rules, but does not resort to a fully-fledged SMT system for doing so as in [1].

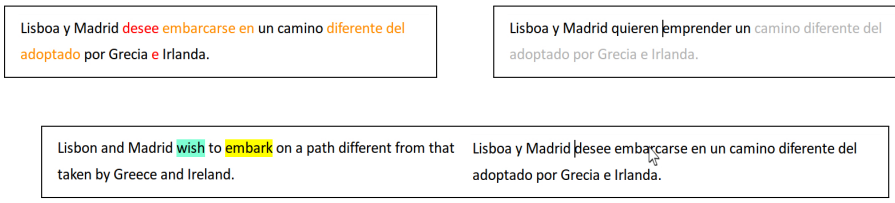

Fig. 3 Advanced display options in CasMaCat: color-coding of confidence estimates (top left), limited prediction horizon (top right), and word alignment visualization.

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

- **Visualization of MT system confidence**. Automatic estimation of the reliability of the MT system output, also known as *confidence estimation* for MT, is currently an active area of research. The CasMaCat workbench is able to visually mark up such confidence estimates in the prediction. MT output identified as probably incorrect is marked in red while MT output of questionable reliability in orange.
- **A limited prediction horizon**. Providing the user with a new prediction whenever a key is pressed has been shown to be cognitively demanding [3]. In the current prototype, when this option is active, predictions are shown only up to the first word of low confidence according to the confidence estimates associated with the prediction. Pressing the TAB key allows the user to ask the system for the next set of predicted words, displaying the remaining words in the suggested translation in grey.
- **Word alignment information**. Alignment of source and target information is an important part of the translation process [7]. In order to display the correspondences between both the source and target words, this feature was implemented so that every time the user places the mouse (yellow) or the text cursor (cyan) on a word, the alignments made by the system are highlighted. The user can enable this visualization option by activating *displayCaretAlign* for the alignments with the cursor and *displayMouseAlign* for the alignments with the mouse.
- **Visualization of user edits** (not shown in Figure 3). This visualization option comes in three variants, all of them implemented with the purpose of helping the user locate which changes were introduced by him, or what was produced by the system without interaction.
  - *changed words only*: the system highlights in green the words that the user has modified.
  - *entire prefix*: the system highlights the prefix, i.e. the first part of the segment that the user has validated.
  - *last edit only*: the system highlights the last word that the user has modified.

3 Translation process data

Another important feature of the CasMaCat workbench is its ability to record user activity in fine detail for analysing human and computer-assisted translation

Table 1 Translation process data logged and stored by CasMaCat. For further details about these features, see [10], [11] and [12].

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

processes scientifically. That is, the tool not only stores translation product information (the source, raw MT output and final translation), but can also provide detailed translation process data with precise timing information, including eye tracking data if used in combination with an eye tracker.[2] A gaze-to-word mapping algorithm runs in real time, and maps gaze samples and fixation points to the nearest letter on the screen; the character offset is then logged together with the gaze data. The tool also keeps a record of the different translation options that were presented to the post-editor at the time. At storage time, CasMaCat aggregates and stores information about phases of coherent writing (production units; PU) and reading (fixation units; FU) from the raw user activity data (UAD). Table 1 summarizes the information stored during interactive translation and post-editing sessions. During analysis, we derived further aggregate information from the stored UAD. These derived measures are described in Sec. 5.

4 Field trial

In June 2013, we conducted a field trial (AFT) in cooperation with Celer Soluciones SL, a language service provider (LSP) based in Madrid. This trial involved nine freelance translators and four reviewers, all native speakers of Spanish offering

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

Table 2 Task assignments in the field trial

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

translation and post-editing services on a regular basis for this LSP. Detailed information about participants' age, level of experience, professional education, etc., can be found in the CRITT Translation Process Research (TPR)-database under the metadata folder.[3]

The text type involved in this trial was general news from the WMT-2012 *news-commentary* corpus [8]. They consisted of approximately 1,000 words, distributed in 30 to 63 segments, as shown in Table 2. Each English source text was automatically translated into Spanish by a statistical MT system and then automatically loaded into the CasMaCat workbench for the participants to post-edit.

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

4.1 Experimental design

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

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

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

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

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

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

5 System evaluation and results

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

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

5.1 Evaluation of activity data

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

[4] due to technical failure

[5] from logged segment translation pairs

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

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

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

5.1.1 Overall translation time

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

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

The first technique assumes that processing consists of alternating periods of typing and processing activity. Based in cognitive language processing and production theory [4, 24, 9], pauses between 0 and 5 seconds are used to segment the text production rhythm into "typing" and "processing" units.

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

[6] Outliers are defined here as those points that exceed Q3+1.5 times the inter-quartile range, see [27].

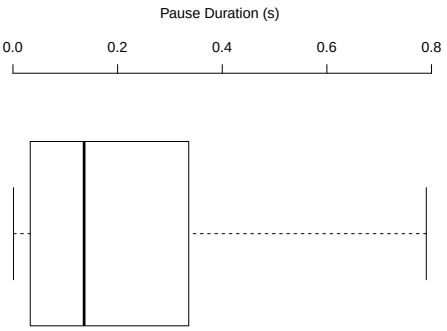

Fig. 4 Boxplot for inter-keystroke pause duration in seconds (outliers not shown).

large set of them. For this purpose, bars and a line graph are used, where the frequencies of individual values are represented in descending order by bars, and the cumulative total is represented by the line. Specifically, in a Pareto chart the left vertical axis represents the frequency of occurrence, while the right vertical axis is the cumulative percentage of the total number of occurrences. In weighted Pareto charts, frequencies are multiplied by specific magnitudes such as cost or loss associated to particular events so as to better analyse their importance (see for example [27] for more details). The black line in the plot marks a relative frequency equal to 95%.

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

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

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

Table 4 shows the average segment post-editing times in seconds for PE, ITP, and AITP systems for three different time measurements, namely Tdur (total duration without excluding any pauses), Kdur and Fdur. PE allowed to obtain shorter post-editing times according to Kdur and Fdur measures. However, the differences between PE and ITP were very small when considering Fdur (the ITP system was 5% slower). One possible explanation for the greater differences between PE and ITP when considering Kdur may be due to the fact that ITP system

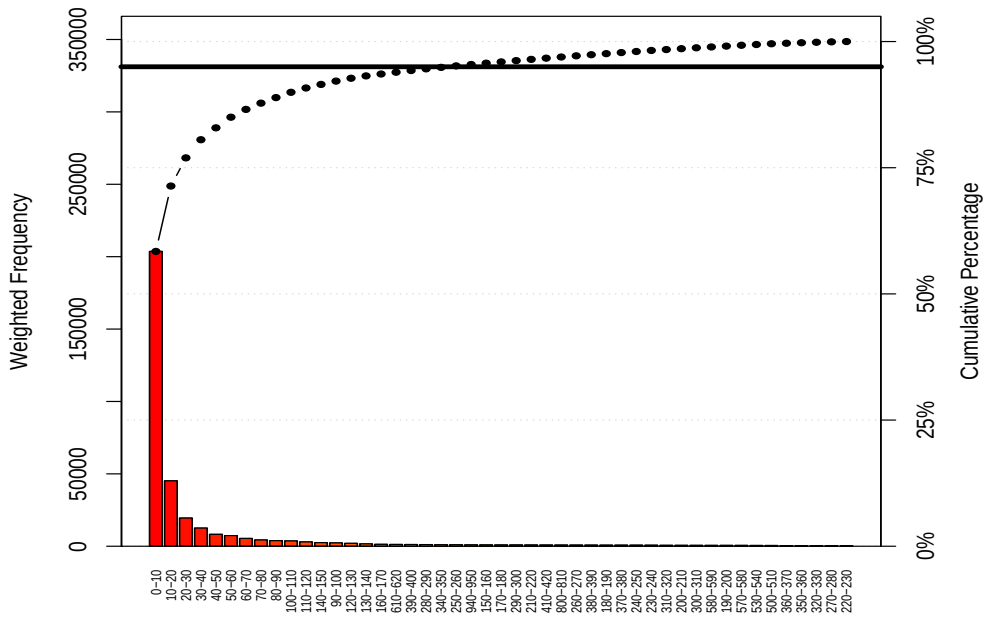

Fig. 5 Weighted Pareto chart for inter-keystroke pause duration in seconds.

users execute a higher number of short post-edit operations. Finally, Tdur values were different from the other two measures due to the noisy observations that have been mentioned above.

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

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

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

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

this, the post-editing time reduction for the ITP and AITP systems was greater than that for the PE system.

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

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

5.1.2 Typing activity

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

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

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

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

5.1.3 Gaze data

Drawing on the seminal work of [20], analyses based on the eye-mind hypothesis suggest that eye fixations can be used as a window into instances of effortful cognitive processing. Following this hypothesis, one could assume that eye-movement recordings can provide a dynamic trace of where a person's attention is being directed. This assumption is often taken for granted by eye-tracking researchers.

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

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

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

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

Enabling interactivity (ITP) and visualization (AITP), however, causes a decrease in the fixations on the source window and a corresponding increase in the target window.

5.2 Quality of post-edited data

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

5.2.1 Edit distance in dataset 1

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

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

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

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

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

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

5.2.2 Correlation of edit-distance, revision time and text modifications

For this analysis, we counted the number of manual insertions and deletions for each of the four reviewers. Table 9 shows the average text modifications per system and reviewer R10 to R13. The table presents the average number of text modifications per segment divided by the length in characters of the segment for each of the three systems. In line with the results of [19], reviewers seem to follow very different reviewing styles: reviewer R10 produces the least number of text modifications, while reviewer R13 is the most eager corrector. On average reviewers produce most relative text modifications when the post-edited text was produced with system ITP.

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

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

We also computed the average revision time, edit distance and number of text modifications per reviewing session, which resulted in 12 data points for each of the variables (three systems $\times$ 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 of the workbench. As a consequence, we had to average over the entire revision session to get comparable numbers for average revision time, edit distance and number of text modifications.

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

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

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

Figure 6 shows the correlations between text modifications and revision time. The highest correlation for all three variables can be observed in the ITP system and for the correlations between text modifications and revision time [14].

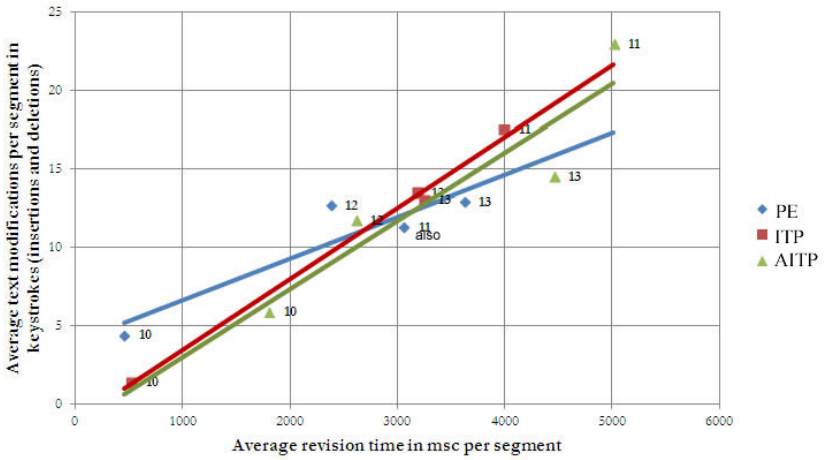

Fig. 6 Correlation between: keystrokes (insertions and deletions) vs. time

5.3 User feedback

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

User feedback was collected regarding the following questions:

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

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

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

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

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

6 Related work

Improving the productivity of the translators is and has been a major driver of MT research. The hope is that, in many cases, post-editing MT output will help translators to perform their work faster. Several studies were performed to evaluate the potential benefit with generally positive results. Measured reductions in translation time typically range somewhere between 18% and 34% [16, 18, 15] sometimes even reaching as high as 43% [30].

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

- Translators level of experience (volunteer, student [21], professional [30, 18])
- Suitability of the MT system, especially when comparing older [23] and more recent e.g. [30] studies
- PE software and subjects' familiarity with it
- Language pair and domain
- Data collection and filtering

Another question is whether post-editing leads to output of lower quality. Koehn [21] found that at least non-professional post-editors are generally both faster and produce better translations, a result that is consistent with later work [30, 17] investigating the same question with professional translators where a strong reduction in time and a reduced number of errors was found. Interestingly, Plitt and Masselot [30] also find that the difference between individual translators is much stronger than between language pairs and MT systems of varying quality. Following this work, Skadiņš et al. [32] measure (slight) negative effects of the post-editing setting for both productivity and quality for some translators but still affirm the overall helpfulness of MT suggestions.

Along with the MT systems, PE environments have developed over time, recently converging towards web-based setups [22, 17] which integrate several aids in a single interface. Despite extensive research on Confidence Estimation for Machine Translation, such annotation has yet to be integrated. Bach [5], for example, suggested visualizing word-level confidences by type size.

Besides quantity and quality, the translation process itself has been studied for many years, starting with explicit collection of the translators thoughts using Think Aloud Protocols [23]. Possible interference with the translation process quickly led to passive/indirect collection of user activity such as the logging of keystrokes and mouse movement [25] and, more recently, even gaze data [28, 13,

10]. By presenting multiple languages simultaneously in an ecologically valid environment, the combination of workbench and logging functions also offers a unique opportunity to investigate broader issues of applied bilingual cognitive processing.

7 Conclusions and future work

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

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

References

1. Pepr: Post-edit propagation using phrase-based statistical machine translation. In: Machine Translation Summit, pp. 191–198 (2013)
2. Alabau, V., Buck, C., Carl, M., Casacuberta, F., Garca-Martnez, M., Germann, U., Gonzlez-Rubio, J., Hill, R., Koehn, P., Leiva, L., Mesa-Lao, B., Ortiz-Martnez, D., Saint-Amand, H., Sanchis-Trilles, G., Tsoukala, C.: Casmacat: A computer-assisted translation workbench. In: Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics (EACL) (2014). Software demonstrations
3. Alabau, V., Leiva, L.A., Ortiz-Martínez, D., Casacuberta, F.: User evaluation of interactive machine translation systems. In: Proceedings of the 16th Annual Conference of the European Association for Machine Translation (EAMT), pp. 20–23 (2012)
4. Alves, F., Vale, D.: 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 (2009)
5. Bach, N., Huang, F., Al-Onaizan, Y.: Goodness: A method for measuring machine translation confidence. In: Annual Meeting of the Association for Computational Linguistics, pp. 211–219 (2011)

6. Barrachina, S., Bender, O., Casacuberta, F., Civera, J., Cubel, E., Khadivi, S., Lagarda, A.L., Ney, H., Toms, J., Vidal, E., Vilar, J.M.: Statistical approaches to computer-assisted translation. Computational Linguistics **35**(1), 3–28 (2009)
7. Brown, P.F., Della Pietra, S.A., Della Pietra, V.J., Mercer, R.L.: The mathematics of statistical machine translation: Parameter estimation. Computational Linguistics **19**(2), 263–311 (1993)
8. Callison-Burch, C., Koehn, P., Monz, C., Post, M., Soricut, R., Specia, L.: Findings of the 2012 workshop on statistical machine translation. In: Proceedings of the Seventh Workshop on Statistical Machine Translation, pp. 10–51. Association for Computational Linguistics, Montréal, Canada (2012). URL `http://www.aclweb.org/anthology/W12-3102`
9. Carl, M.: The CRITT TPR-DB 1.0: A Database for Empirical Human Translation Process Research. In: AMTA 2012 Workshop on Post-Editing Technology and Practice (WPTP 2012), pp. 1–10. Association for Machine Translation in the Americas (AMTA), San Diego, USA (2012)
10. Carl, M.: Translog-ii: a program for recording user activity data for empirical reading and writing research. In: Eight International Conference on Language Resources and Evaluation (LREC'12). European Language Resources Association (ELRA), Istanbul, Turkey (2012)
11. Carl, M.: Produkt- und prozesseinheiten in der critt translation process research database. In: B. Ahrens (ed.) Translationswissenschaftliches Kolloquium III: Beitrge zur bersetzungs- und Dolmetschwissenschaft (Kln/Germersheim). Peter Lang, Frankfurt am Main (2014)
12. Carl Michael; Kay, M.: Gazing and typing activities during translation : A comparative study of translation units of professional and student translators. Meta **56**(4), 952–975 (2011)
13. Doherty, S., OBrien, S., Carl, M.: Eye tracking as an mt evaluation technique. Machine Translation **24**(1), 1–13 (2010)
14. Elming, J., Carl, M., Balling, L.W.: Investigating user behaviour in post-editing and translation using the casmacat workbench. Expertise in Post-editing: Processes, Technology and Applications (2014)
15. Federico, M., Cattelan, A., Trombetti, M.: Measuring user productivity in machine translation enhanced computer assisted translation. In: Tenth Biennial Conference of the Association for Machine Translation in the Americas (2012)
16. Flournoy, R., Duran, C.: Machine translation and document localization at adobe: From pilot to production. MT Summit XII: Proceedings of the Twelfth Machine Translation Summit (2009)
17. Green, S., Heer, J., Manning, C.D.: The efficacy of human post-editing for language translation. In: SIGCHI Conference on Human Factors in Computing Systems, pp. 439–448. ACM (2013)
18. Guerberof, A.: Productivity and quality in mt post-editing. In: MT Summit XII-Workshop: Beyond Translation Memories: New Tools for Translators MT (2009)
19. Guerberof, A.: Productivity and quality in the post-editing of outputs from translation memories and machine translatio. Ph.D. Thesis (2012)
20. Just, M.A., Carpenter, P.A.: A theory of reading: from eye fixations to comprehension. Psychological review **87**(4), 329 (1980)
21. Koehn, P.: A process study of computer-aided translation. Machine translation **23**(4), 241–263 (2009)
22. Koehn, P.: A web-based interactive computer aided translation tool. In: ACL-IJCNLP 2009 Software Demonstrations, pp. 17–20. Association for Computational Linguistics, Suntec, Singapore (2009). URL `http://www.aclweb.org/anthology/P/P09/P09-4005`
23. Krings, H.P.: Repairing texts: empirical investigations of machine translation post-editing processes, vol. 5. Kent State University Press (2001)
24. Lacruz, I., Shreve, G.M., Angelone, E.: Average Pause Ratio as an Indicator of Cognitive Effort in Post-Editing: A Case Study. In: AMTA 2012 Workshop on Post-Editing Technology and Practice (WPTP 2012), pp. 21–30. Association for Machine Translation in the Americas (AMTA), San Diego, USA (2012)
25. Langlais, P., Foster, G., Lapalme, G.: Transtype: A computer-aided translation typing system. In: 2000 NAACL-ANLP Workshop on Embedded Machine Translation Systems, *NAACL-ANLP-EMTS '00*, vol. 5, pp. 46–51 (2000). DOI 10.3115/1117586.1117593. URL `http://dx.doi.org/10.3115/1117586.1117593`
26. Leiva, L.A., Alabau, V., Vidal, E.: Error-proof, high-performance, and context-aware gestures for interactive text edition. In: Proceedings of the 2013 annual conference extended abstracts on Human factors in computing systems (CHI EA), pp. 1227–1232 (2013)

27. Montgomery, D.: Introduction to Statistical Quality Control. Wiley (2004). URL `http://books.google.es/books?id=Qpt4QgAACAAJ`
28. O'Brien, S.: Eye tracking in translation process research: methodological challenges and solutions, *Copenhagen studies in language*, vol. 38, pp. 251–266. Samfundslitteratur, Copenhagen (2009)
29. Ortiz-Martínez, D., Casacuberta, F.: The new Thot toolkit for fully automatic and interactive statistical machine translation. In: 14th Annual Meeting of the European Association for Computational Linguistics: System Demonstrations, pp. 45–48. Gothenburg, Sweden (2014)
30. Plitt, M., Masselot, F.: A productivity test of statistical machine translation post-editing in a typical localisation context. The Prague Bulletin of Mathematical Linguistics **93**(1), 7–16 (2010)
31. Sanchis-Trilles, G., Ortiz-Martínez, D., Civera, J., Casacuberta, F., Vidal, E., Hoang, H.: Improving interactive machine translation via mouse actions. In: Conference on Empirical Methods in Natural Language Processing (EMNLP) (2008)
32. Skadiņš, R., Puriņš, M., Skadiņa, I., Vasiļjevs, A.: Evaluation of smt in localization to under-resourced inflected language. In: Proceedings of the 15th International Conference of the European Association for Machine Translation EAMT, pp. 35–40 (2011)