D6.3: Analysis of the third field trial

Vicent Alabau, Michael Carl, Mercedes García Martínez,
Jesús González-Rubio, Bartolomé Mesa-Lao, Daniel Ortiz-Martínez,
Sofia Rodrigues, Moritz Schaeffer

Distribution: Public

CasMaCat
Cognitive Analysis and Statistical Methods
for Advanced Computer Aided Translation

ICT Project 287576 Deliverable D6.3

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	January 7, 2015
Date of last update	January 7, 2015
Deliverable number	D6.3
Deliverable title	Analysis of the third field trial
Type	Report
Status & version	Final
Number of pages	21
Contributing WP(s)	WP7
WP / Task responsible	CS, CBS
Other contributors	
Internal reviewer	Daniel Ortiz
Author(s)	Vicent Alabau, Michael Carl, Mercedes García Martínez, Jesús González-Rubio, Bartolomé Mesa-Lao, Daniel Ortiz-Martínez, Sofia Rodrigues, Moritz Schaeffer
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

In this work package, we evaluate the CasMaCat workbench in field trials to study the use of the workbench in a real-world environment. We have also integrated the workbench into community translation platforms and collected user activity data from both field trials and volunteer translators interacting with the workbench.

This Deliverable covers Task 6.1 and 6.2.

- Task 6.1: Third field trial at a translation agency (Celer Soluciones SL in Madrid) to evaluate the CasMaCat workbench in a real-world professional translation environment.
- Task 6.2: Analysis of translator feedback and activity data. Collection of feedback of translators' self-estimation through questionnaires and retrospective interviews.

In addition to the originally planned third field trial for 2014, we have also conducted an additional longitudinal study between April and May 2014 (as discussed in the last review meeting – December 2013).

Contents

1 Introduction

This deliverable reports on two different field trials: the first conducted in April/May 2014 in the form of a longitudinal study (LS14 study) and another one conducted in June 2014 (CFT14 study). Both field trials were conducted at Celer Soluciones SL in Madrid with a sample of professional translators recruited by the company.

Results from the second field trial (June 2013) revealed that post-editors may need more time to fully grasp the benefits of ITP for post-editing purposes, so as discussed on the second review meeting (December 2013) and on the third internal CasMaCat meeting (March 2014), the consortium decided to run an additional longitudinal test with the aim of investigating whether and to what extent post-editors improve performance using the ITP (Interactive Translation Prediction) feature in CasMaCat over an extended period of time.

The longitudinal study (LS14) involved five post-editors working alternatively with traditional post-editing (baseline) and ITP over a period of six weeks. The aim was to test whether they become faster when working with ITP as they become more acquainted with this type of assistive technology. Results show that participants became indeed faster over the period of six weeks in the ITP condition and, according to the projection of the data collected, they could have been even more productive after 8 to 12 weeks of regular exposure to this new technology.

An extension of the LS14 study is the third CasMaCat field trial 2014 (CFT14). The CFT14 study was also conducted by Celer Soluciones SL aiming at assessing whether post-editors profit from ITP online learning as compared to traditional post-editing. A sample of seven post-editors participated in the CFT14 study and four of them had also taken part in the previous longitudinal study (LS14). The CFT14 study differs from the LS14 study in these two respects: the text type involved was general news in the case of LS14, while the text type in CFT14 was a more specialized one extracted from the EMEA corpus (medical domain). The number of source text words was also quite different between these two studies: LS14 involved 24 source texts of 1,000 words each, while CFT14 involved only two source text with 4,500 each (texts were much longer in CFT14, so as to test the online learning effect with tokens that occurred several times within each text).

Both studies combined involved around 33,000 source text words (205,000 target text tokens) and they have been included in the publicly available TPR-DB[1].

2 Longitudinal study (LS14)

This section presents an additional study, previous to the third year trial, investigating post-editors' performance over a period of six weeks (April-May 2014). The aim of this study was primarily to find out whether professional post-editors improved performance over time while interacting with the CasMaCat workbench featuring ITP. The findings are reported in section 2.4.2. We were also interested in uncovering any specific profiles of translators who behaved differently while post-editing with ITP depending on personal factors such as previous experience in post-editing and typing skills (see section 2.4.1). Finally, it was the aim to collect feedback from the post-editors in order to know more about their views regarding this type of technology. This is reported in section 5.

[1]The CRITT Translation Process Research Database. Available online at: `http://bridge.cbs.dk/platform/?q=CRITT_TPR-db`

2.1 Participant profiles

Five professional translators were recruited by Celer Soluciones SL to take part in the study. Participants were 33 years old on average (range 26-42) and all of them were regular users of computer-aided translation tools (mainly SDL Trados and WordBee) in their daily work as professional translators. All participants but one had previous experience in post-editing MT as a professional service. For three of the four participants with post-editing experience, their workload involving post-editing services did not exceeded 10% of their projects as reported in an introductory questionnaire. The fourth participant with post-editing experience reported that 75% of their workload as a professional translator involved post-editing projects.

More specific data on the participants' age, level of experience, professional education, etc., is available in the CRITT TPR Database (metadata folder).

2.2 Text type

The source texts involved in this longitudinal study were pieces of general new extracted from the WMT 2014 corpus. Each source text contained 1,000 words on average distributed over 48 segments on average (range 39-61).

2.3 Experimental design

The experimental design involved 24 different source texts which were post-edited from English into Spanish over a period of six weeks (four texts per week). MT was provided by the CasMaCat server and the participants were asked to work under the following conditions:

- *Condition 1*: Traditional post-editing (P), i.e. no interaction is provided during the post-editing process.
- *Condition 2*: Interactive post-editing (PI), i.e. interaction is provided during the post-editing process in the form of ITP.

Every week, post-editors worked in parallel on the same 4 source texts counterbalancing texts/conditions among participants in order to avoid any possible text/tool-order effect (two texts in condition 1 and two texts in condition 2). During the first and the last week of the study, post-editors worked from Celer Soluciones SL while their eye movements were recorded using an eye-tracker. From week 2 to week 4, post-editors worked from home as they usually do when completing jobs for the company. Meeting the participants at the company the first week was useful to make sure they understood the assignment before starting to post-edit (specific post-editing guidelines were given) as well as to offer them a hands-on tutorial on how ITP works from the user perspective (condition 2). During the last week of the experiment, participants returned to Celer Soluciones SL so that a second sample of their eye movements could be recorded and so that we could gather their feedback and their comments on the technology they had been using.

Each post-editor post-edited 1,154 segments in total with 146,358 source text words (half of them in each condition).

	Week 01*	Week 02	Week 03	Week 04	Week 05	Week 06*
P01	T00,T02 (P)	T04,T06 (P)	T08,T10 (P)	T12,T14 (P)	T16,T18 (P)	T20,T22 (P)
	T01,T03 (PI)	T05,T07 (PI)	T09,T11 (PI)	T13,T15 (PI)	T17,T19 (PI)	T21,T23 (PI)
P02	T01,T03 (P)	T05,T07 (P)	T09,T11 (P)	T13,T15 (P)	T17,T19 (P)	T21,T23 (P)
	T00,T02 (PI)	T04,T06 (PI)	T08,T10 (PI)	T12,T14 (PI)	T16,T18 (PI)	T20,T22 (PI)
P03	T01,T03 (PI)	T05,T07 (PI)	T09,T11 (PI)	T13,T15 (PI)	T17,T19 (PI)	T21,T23 (PI)
	T00,T02 (P)	T04,T06 (P)	T08,T10 (P)	T12,T14 (P)	T16,T18 (P)	T20,T22 (P)
P04	T00,T02 (PI)	T04,T06 (PI)	T08,T10 (PI)	T12,T14 (PI)	T16,T18 (PI)	T20,T22 (PI)
	T01,T03 (P)	T05,T07 (P)	T09,T11 (P)	T13,T15 (P)	T17,T19 (P)	T21,T23 (P)
P05	T00,T02 (P)	T04,T06 (P)	T08,T10 (P)	T12,T14 (P)	T16,T18 (P)	T20,T22 (P)
	T01,T03 (PI)	T05,T07 (PI)	T09,T11 (PI)	T13,T15 (PI)	T17,T19 (PI)	T21,T23 (PI)

Table 1: Experimental design for the longitudinal study(LS14) covering 6 weeks.

2.4 Results

2.4.1 Post-editing behaviour

The evaluation of the LS14 data is based on three different parameters computed at the segment level[2]:

1. *Fdur*: production time per segment, excluding pauses > 200 seconds, normalised by the number of characters in the source segment.
2. *Kdur*: duration of coherent keyboard activity per segment excluding keystroke pauses > 5 seconds, normalised by the number of characters in the source segment.
3. *Pdur*: duration of coherent keyboard activity per segment excluding keystroke pauses > 1 seconds, normalised by the number of characters in the source segment.

Participant	Cond	Fdur	Kdur	Pdur
P01	PI	563.64	254.30	113.33
P01	P	529.71	215.86	88.51
P02	PI	456.53	173.06	68.24
P02	P	439.87	157.46	68.51
P03	PI	623.81	223.79	85.26
P03	P	573.68	167.51	63.77
P04	PI	684.30	230.22	130.28
P04	P	701.46	161.53	88.32
P05	PI	320.72	158.18	69.99
P05	P	284.43	138.20	54.25

Table 2: Overall typing activity (insertions + deletions) and production times in the LS14 data.

Table 2 gives an overview of average post-editing durations and typing activities per source text character for all five post-editors on the two conditions during the six weeks. The data show that post-editors need more keystrokes in the PI condition than under the P condition, thus becoming slower in this condition. On average, there are fewer manual insertions in the P condition (52.5 per segment) than there are in PI (68.7 per segment) but there are more manual deletions in P (46.8 per segment) than in PI (32.5). Both these differences are significant ($p < .001$). As shown by *Kdur* values, post-editors needed between 138.2 and 215.86 ms per character for post-editing (P) while it took them on average between 158.18 and 254.30 ms in the PI mode. Duration values for *Fdur* and *Pdur* show a similar pattern.

[2]Further insights on this data are reported in deliverable D1.3

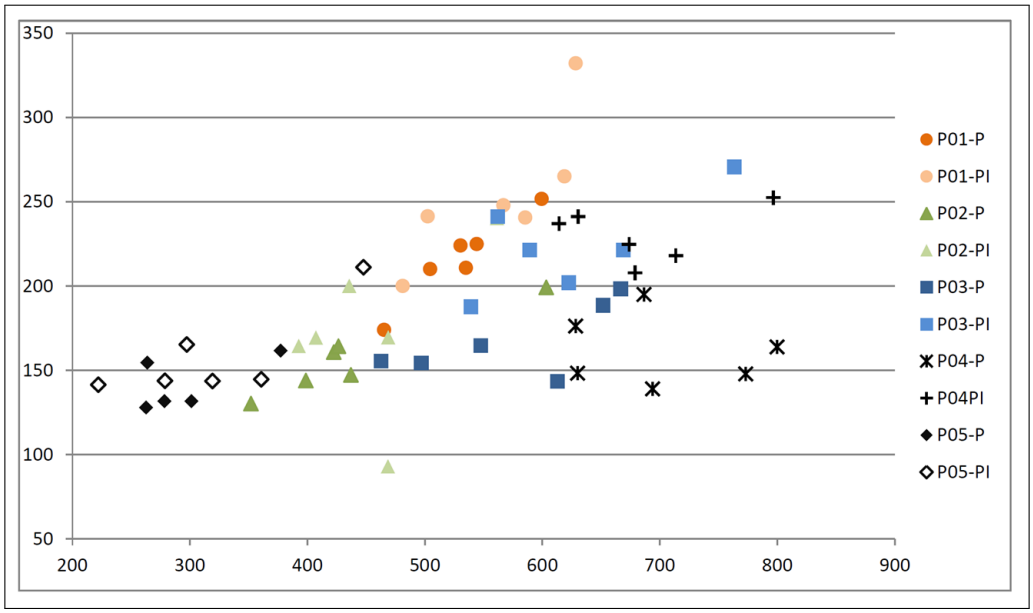

Figure 1: LS14 Study - Fdur (horizontal) vs. Kdur (vertical) for all 5 participants.

Figure 1 plots the relationship between *Fdur* and *Kdur* for the 5 participants in the longitudinal study. Each point in the graph shows the average $Kdur/Fdur$ ratio per source text character over one week of post-editing activity under both conditions. Since two texts of each approximately 1,000 words were edited in either of the two post-editing modes (P or PI), each dot represents the average per-character post-editing duration of 2,000 source text words.

Different post-editors show different $Kdur/Fdur$ correlations, comparing total post-editing duration (excluding pauses longer than 200 seconds) and typing activity duration (including insertions and deletions). Only the data for post-editor P04 (the only one who reported no previous post-editing experience despite being a professional translator) showed a much weaker correlation between *Kdur* and *Fdur* (0.40). P03 (the only one without formal training despite working as a freelance translator for Celer Soluciones SL) showed a slightly weaker correlation between these two measures (0.74). However, the other participants show a strong correlation between these two durations ({P01 = 0.78, P02 = 0.78, P05 = 0.82}). This suggests that professional translators whith training and experience in post-editing ({P01,P02,P05}) use their time more efficiently while they work on a segment. *Kdur* seems to be a better indicator for overall translation duration as reflected in active typing activity for more experienced post-editors than it seems to be for less experienced post-editors ({P03,P04}). Deliverable 1.3 discusses in more detail the user profiles of the post-editors as identified in the logged data, which could explain this difference in post-editing behaviour as reflected in typing activity.

2.4.2 Learning effects

With respect to the main aim of the LS14 study, when investigating the production times over a period of six weeks interacting with ITP in the CasMaCat workbench, it can be observed how post-editors become substanially quicker in the PI condition over time, while in the P condition (baseline) no significant change in *Kdur* effort can be observed. Figure 2 plots the effect of regular usage of the two CasMaCat settings on post-editing duration measured in terms of *Kdur* per source text character. For this analysis, skipped segments with either zero tokens in the final target text and/or with zero total duration and segments with more than one edit were excluded. Segments with more than one edit were segements which had been opened, corrected and closed once, before they were opened again for a revision. This was done, because participants complained that often when a segment was re-visited, the initial MT ouput rather

than the already corrected text appeared, which meant that translators had to edit text which they had already corrected. In total, 12% of the data was excluded. The two regression lines (based on simple linear models) show the projection of the average post-editing time under the PI and the P conditions over a hypothetical timeframe of 12 weeks, twice as much as the actual LS14 study duration. The grey areas around the linear regression lines represent the 95% confidence region for each regression. According to this projection, it is between weeks 9 and 10 that post-editors would become more efficient under the PI condition than under the P condition. While this is a hypothetical assumption, assuming a linear relationship between time spent working on the CasMaCat workbench and *Kdur*, this projection clearly shows a learning effect for the PI condition, which is absent in the P.

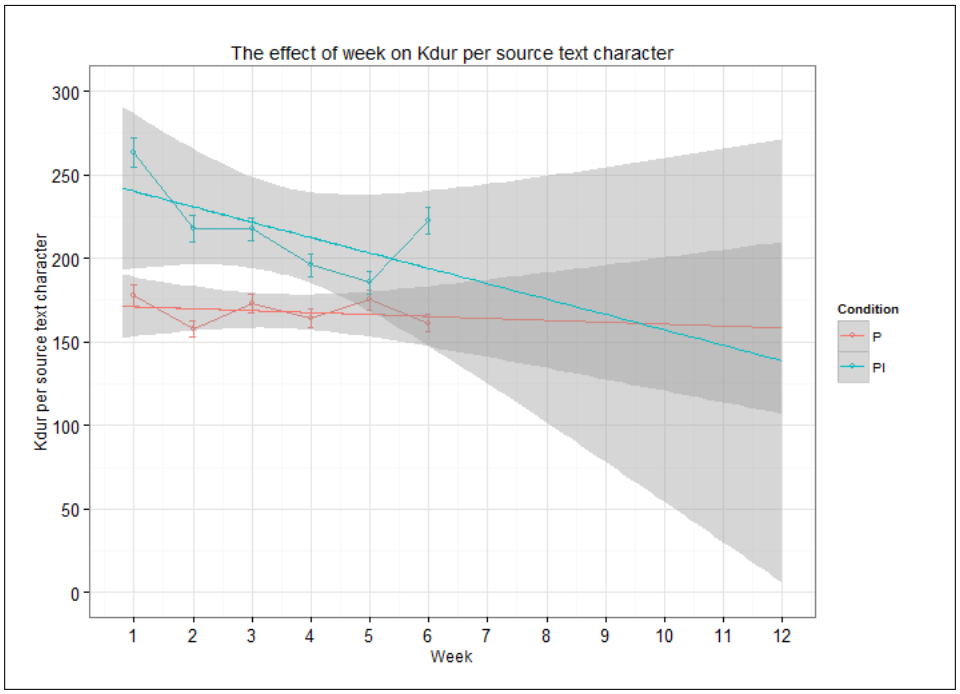

Figure 2: LS14 Study - Productivity projection as reflected in Kdur taking into account six weeks.

Despite the general downwards trend in PI over time, Figure 2 shows a difference in efficiency in week 1 and 6 as compared to the other weeks. The reason for these peaks in time for weeks 1 and 6 might be the experimental setup itself, since these two weeks involved eye-tracking apparatus and the request to post-edit from the company. The fact of having to work from Celer Soluciones SL, instead of from home, seems to have had a negative impact on post-editor's performance. During weeks 2 to 5, post-editors worked from home, which is what they are used to since all of them work as freelancers. In addition to this, using an eye-tracker involved limited head movement and sometimes recalibration during the process of post-editing was necessary. Together, these aspects may have had a negative effect on participants' productivity, in other words, the data might show a lab effect.

In addition to the above mentioned reasons, the productivity drop for week 6 under PI can also be found in the texts themselves that were post-edited under this condition that particular week. In an attempt to find out whether the texts involved in week 6 were somehow more challenging to post-edit under PI, TER values were computed for all the texts in LS14. The aim was to discover if such TER values were particularly higher for texts in week 6, which could be interpreted as a reason for the fact that we observe longer overall times since more edits were required during the post-editing process. After looking into TER values, we could indeed identify text 20 in week 6 (post-edited under PI by P01, P03 and P05) as one of the most difficult texts to post-edit as reflected in the number of edits recorded in TER values. Text 20 in LS14 was of a more specialized nature (legal text) since this piece of news was about the

code of conduct of US judges. This different degree in text specialization could be the reason for both lower MT quality and thus requiring more edits from the post-editors.

Assuming that working at home and working in the office are two different conditions, we re-calculated the learning projection shown in Figure 2 based only on the 4 weeks when post-editors worked in the office. Figure 3 plots the two conditions in LS14 showing that post-editing under the PI condition could have become –theoretically– more efficient already after 6 weeks. Grey areas in Figure 3 represent 95% confidence regions and regression lines are based on simple linear models.

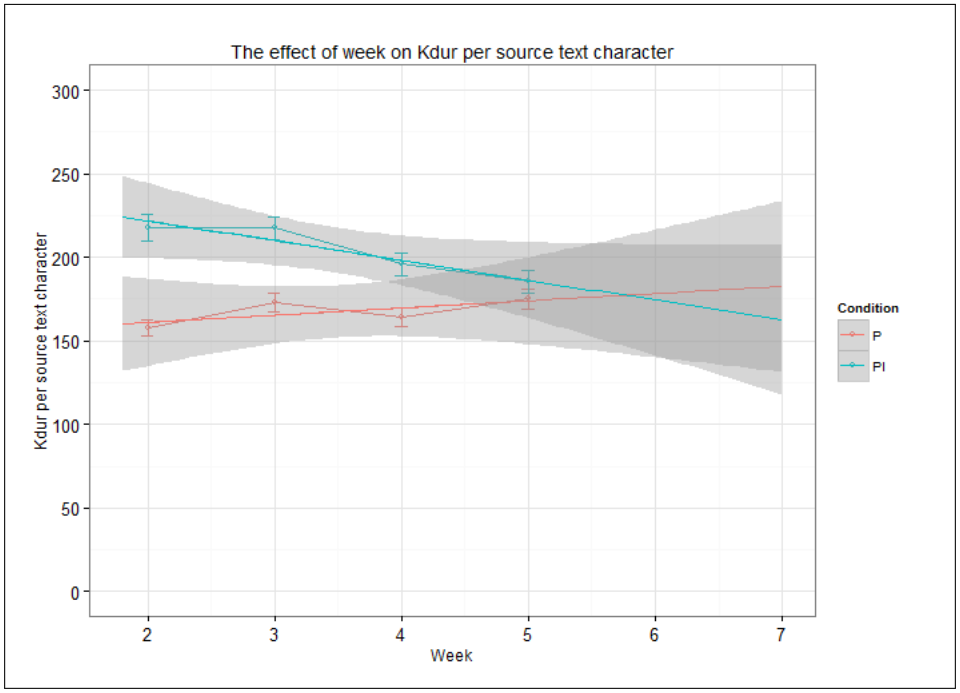

Figure 3: LS14 Study - Productivity projection as reflected in Kdur based only on the data from weeks 2-5 (working from home).

A closer look at the way post-editors became acquainted with ITP suggests that learning to interact with this interactive technology involves controlling typing speed in order to be able to fully benefit from the suggestions (i.e. autocompletions) provided by the system. Since all post-editors in the LS14 study were touch typists, they could only fully benefit from the ITP suggestions once they gradually learned to avoid overwriting suggestions and thus saving typing effort. Post-editor self-rated their typing skills as excellent in an introductory questionnaire and, indeed, their typing speed caused many cases of overwriting behaviour as they continued typing even though the right suggestions by the ITP system was already pasted in the target text. Learning to control this overwriting behaviour was also reported by the post-editors themselves when providing user feedback (see section 5).

3 Pre-field trial (PFT14)

This section presents the pre-field trial pilot study prior to CFT14 study (see section 4) conducted at the Copenhagen Business School for the language pair English to Danish. The main aim of this pilot study was to assess and compare online learning (OL) and active learning (AL) combined with ITP against conventional ITP (without OL/AL) with a view to deciding which of the two machine learning techniques should be further tested in the frame of the main field trial (CFT14 study). For a technical description on these two machine learning techniques see work package 4. A secondary aim of the PFT14 study was to collect user feedback while post-editing using machine learning techniques.

3.1 Experimental setup

As mentioned before, the three conditions in this pre-field trial pilot study were:

- ITP (i.e. post-editors were presented with alternative ways to complete words/phrases as they typed their changes).
- ITP with online learning (OL) techniques from segment to segment.
- ITP with active learning (AL) techniques from segment to segment.

The two parameters used to analyse the user activity data collected in this pre-field trial pilot were:

- *Speed*: total number of words translated divided by time in minutes.
- *Technical effort*: total number of edits done by the participant divided by the number of translated words.

A group of five participants volunteered to take part in this pre-field trial pilot post-editing from English into Danish. Table 3 summarizes the profile of the users.

	Native Danish speaker	Professional translator
P0	yes	no
P1	yes	yes
P2	no	yes
P3	yes	yes
P4	yes	yes

Table 3: Users' profile in the pre-field trial pilot study.

The type of text involved in this pilot was the same as the one used in the main field third field trial, i.e. specialized texts extracted from the EMEA corpus (domain: medical package leaflets - `http://opus.lingfil.uu.se/EMEA.php`).

This pre-field trial pilot involved two different experiments:

First experiment: P0 post-edited three comparable texts with 55 segments each (843 words, 803 words, and 1,005 words). This participant translated each text under a different condition: ITP, ITP with OL, and ITP with AL.

Second experiment: the rest of participants (P1 to P4) were asked to post-edit the same source text (the one with 1,005 words in the first experiment) under a different condition each. The aim was to compare results from different participants under different conditions.

3.2 Results

3.2.1 User activity data

First we will present the results comparing conventional ITP against ITP with OL. Table4 shows ITP and OL results for the first experiment in which P0 post-edited different texts under the three conditions. Table 5 shows the corresponding results for the second experiment, where the same text (1,005 words) was post-edited by the different participants.

Results show how OL significantly improved translation speed (about 2.5 more words translated per minute). Regarding the number of keystrokes, results are not consistent: no significant difference was found in the first experiment for the two conditions involving OL/AL while it was significantly better for OL in the second experiment involving different participants. The anomalous results for P2 could be explained by the different profile of this participant (i.e. P2 was not a native speaker of Danish despite being proficient in this language).

P0	ITP	ITP with OL
Words translated	843	803
Words/min.	14.1	16.4
Keystrokes/word	2.3	2.3

Table 4: First experiment: ITP vs. ITP with OL results.

	P1	**P2**	**P3**
Native	Yes	No	Yes
Condition	ITP	ITP with OL	ITP with OL
Words/minute	15.2	40.2	18.0
Keystrokes/word	2.9	0.6	1.8

Table 5: Second experiment: ITP vs. ITP with OL results.

Note: P4 is not included in Table 5 since she post-edited under the ITP with AL condition.

Regarding the results for post-editing through ITP with AL in the first experiment (performed by P0), post-editors were asked to post-edit first those segments for which the machine generated translations were considered to be worst (as judged by confidence measures). It is important to note that, since the participant did not post-edit all machine generated segments in this condition (just the ones with the lowest confidence scores), the final target text was a mixture of automatic and human post-edited translations.

With respect to this specific experiment involving only P0, the quality of the target text was computed using BLEU scores together with the technical effort invested (keystrokes per post-edited word) as a function of the number n of segments post-edited by the participant. Segments were ranged n between zero and 55, the number of segments in the text. Figure 4 shows the improvement in translation quality with respect to SMT output as a function of the technical effort invested by P0. Similar results were obtained when comparing P0 versus P4 in the second experiment. These results prove how, for the same amount of effort, ITP with AL provides a larger increase in translation quality as compared to post-editing just through conventional ITP.

3.2.2 User feedback

User feedback was collected after each post-editing session in the form of retrospective think-aloud protocols. The post-editing process was recorded using screen capture video and then replayed to the participants in order to elicit their actions and feelings as they went about with the post-editing tasks. Below, we include some comments and ideas provided by the participants.

P1 observations on post-editing through ITP *(professional translator)*

> "Compared with editing in a non-interactive setting, the interactive translation mode was generally quite a different experience from a users point of view. It was necessary to 'unlearn' some of the editing processes normally carried out during revision

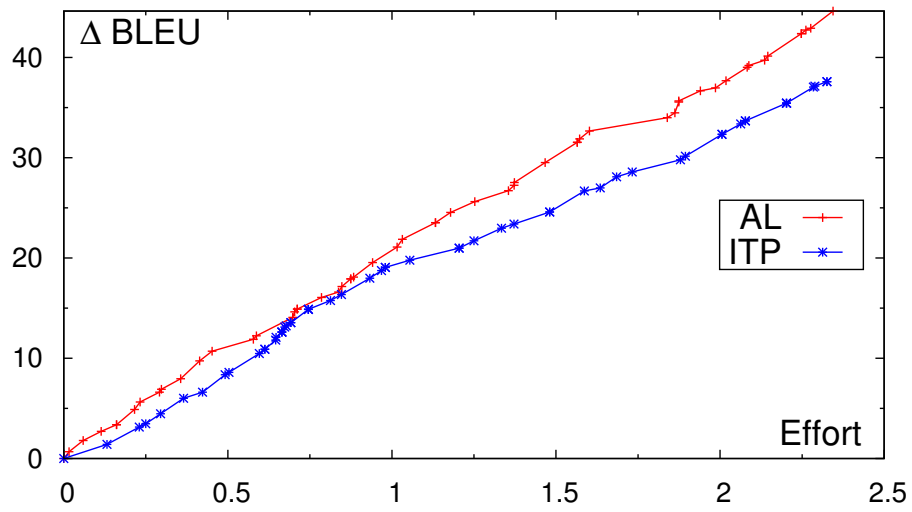

Figure 4: First experiment: improvement in translation quality with respect to SMT as a function of the technical effort (keystrokes/word) invested by P0.

of human or machine translation, such as highlighting words or segments and overwriting them with improved alternatives, and reading and planning a whole sentence before making corrections. This lead to a very different editing process, which required some getting used to and caused a good deal of frustration at first. However, after some time and practice, and 'unlearning' of old habits, efficiency improvements kicked in, but only to the extent that the dynamic changes were appropriate, which was not always the case. Thus, the problems experienced when working in the interactive mode were generally associated more with the quality of some of the dynamic corrections made by the system and less with the interactive mode as such.

On the positive side, the grammatical corrections generally worked well. For example, when the definite article (*det/den/de* in Danish) was inserted (by the user) before a pre-modifying adjective, the system automatically added the inflection -e to the adjective, which is the correct form in Danish. Also, when a noun was written as an alternative to the original MT solution, the original noun was automatically removed, which saved the user the delete action and thus improved efficiency.

On the negative side, dynamic corrections at the lexical level were not always appropriate. For example, when adding the morpheme 'op-' to the Danish noun 'løsning' to arrive at the Danish word for 'dissolution' ('opløsning'), rather than 'solution' ('løsning'), the system suggested 'opfølgning' ('follow-up'). This inappropriate dynamic correction then had to be revised by deleting 'følgning' and reinserting 'løsning', which lead to decreased efficiency in the post-editing process.

The gray/black distinction to differentiate between edited and non-edited text worked well for me. It was easy to keep track of already accepted text and output that was yet to be checked."

P0 observations on ITP with AL *(non-professional translator)*

"The use of AL features while post-editing helped me a lot especially when using a more technical vocabulary. The interactivity seems faster and easier to recall completely different words, but it is quite the opposite when it comes to introduce small grammatical chances, such as word endings in Danish. I think that I would need more hours interacting with the system to make the most of it, but it is a nice feature when the system is able to remember my word preferences to help me improving my productivity and consistency overall."

Based on the results of this pre-field trial pilot evaluating both OL and AL, the decision was made to further test the possibilities of ITP with OL in the context of third CasMaCat field trial. The next section will concentrate on the third field trial itself (included in the CRITT TPR-DB under the study name CFT14).

4 Third field trial (CFT14)

This section presents the third field trial with the latest prototype of the CasMaCat workbench conducted in June 2014. The main research aims of this field trial were:

- To measure the productivity benefits derived from introducing online learning techniques during the post-editing process.
- To investigate how post-editors use the biconcordancer tool integrated in the latest prototype of the CasMaCat workbench.
- To assess how professional reviewers use the e-pen functionalities while reviewing from CasMaCat.
- To collect feedback from reviewers using e-pen as an additional input method for revision.

4.1 Participant profiles

This third field trial involved seven post-editors and four reviewers. All post-editors and reviewers were freelancers recruited by Celer Soluciones SL. Participants were 35 years old on average (range 26-52) and all of them were regular users of computer-aided translation tools in their daily work. It is important to note that participants P01, P02, P03 and P04 also took part in the longitudinal study (LS14) described in section 2.

All participants but one had previous experience in post-editing MT as a professional service. More specific data on the participants' age, level of experience, professional education, etc., is available in the CRITT TPR Database (metadata folder).

4.2 Text type

Two texts were involved in this third field trial. As in the case of the pre-field trial pilot study (see section 3), the type of text involved in this third field trial was domain specific, i.e. medical specialised texts from the EMEA corpus (package leaflets for schizophrenic patients). They were approximately 4,500 words long comprising 131 and 141 segments respectively and they were pre-translated into Spanish by a SMT system and then loaded into the workbench for the participants to post-edit.

4.3 Experimental design

In order to assess and compare the effects of enabling interactivity and online learning techniques, each participant post-edited two texts each under one of the following conditions:

- *Condition 1*: Traditional post-editing with no assistance during the process (P).
- *Condition 2*: Post-editing through ITP featuring on-line learning (PIO).

	Condition	
Participant	P	PIO
P01	T1	T2
P02	T2	T1
P03	T1	T2
P04	T2	T1
P05	T1	T2
P06	T2	T1
P07	T1	T2

Table 6: Experimental design for third field trial (CFT14).

Each participant completed the two tasks in a single session (4 hours on average) from the Celer Soluciones' premises, where an eye-tracker was also used to record the gaze behaviour of the post-editors. In order to ensure an equal distribution of texts and conditions across the participants, both variables were counterbalanced from participant to participant (see table 6).

Before starting each task, participants were introduced again to the workbench and they were given time to familiarize themselves with the tool (specially participants P05, P06 and P07 who were post-editing using for the first time the CasMaCat workbench). Each of the post-edited texts were subsequently proofread by different reviewers (further details about the revision phase are provided in section 4.4.3).

4.4 Results

As in the case of the LS14 study, the productivity evaluation for CFT14 study is based on these three measure: *Fdur*, *KDur* and *PDur* (section 2.4 for a definition of these measures).

Participant	Cond	FDur	KDur	PDur
P01	P	468.90	290.38	138.29
P01	PIO	466.63	244.75	117.42
P02	P	417.87	265.46	128.63
P02	PIO	572.32	233.50	104.83
P03	P	420.29	226.68	71.36
P03	PIO	578.96	256.90	95.00
P04	P	656.72	216.61	111.54
P04	PIO	516.62	261.09	141.63
P05	P	330.74	261.63	132.10
P05	PIO	325.36	253.39	120.36
P06	P	704.48	229.87	84.19
P06	PIO	433.02	230.18	88.02
P07	P	529.58	196.90	63.31
P07	PIO	443.81	216.56	74.58

Table 7: Overall typing activity (insertions + deletions) and production times in the CFT14 data.

4.4.1 Productivity

To measure whether participants become faster when post-editing with interactivity and online learning techniques, we analyzed both time to complete the task and keystroke activity

as reflected in the log files computed in the CRITT TPR-DB. Time was measured using *Fdur* values (duration of segment production time excluding keystroke pauses > 200 seconds) and *Kdur* values (duration of coherent keyboard activity excluding keystroke pauses > 5 seconds).

In order to measure the productivity benefits derived from introducing online learning techniques during the post-editing process, the amount of technical effort (i.e. the number of insertions and deletions needed to correct the raw MT output) was calculated for the two conditions. Keystroke activity was measured by using *Mdel* values (number of manually generated deletions) and *Mins* values (number of manually generated insertions). It is important to make the distinction between manual and automatic insertions and deletions since ITP triggers a lot of automatic insertions/deletions which does not require any technical effort from the post-editor.

On the one hand, participants working under condition 1 (traditional post-editing) deleted 70.71 keystrokes and inserted 79.53 on average. On the other hand, post-editing with interactivity and on-line learning techniques made keyboard activity decrease as they inserted 68.73 keystrokes and deleted 36.94 on average. If we compare keyboard activity for both conditions, it can be claimed that there was a significant decrease in the number of insertions (Z=-3,677, p= .000) and deletions (Z=-13,156, p= .000) comparing both conditions. Since both texts were comparable in size and difficulty, this significant decrease in technical effort (typping) must be attributed to the benefits of online learning techniques during the post-editing process.

Apart from the productivity gains as reflected in less keyboard activity, the overall time spent on both conditions was also measured using *Fdur* and *Kdur* values. Contrary to what was expected, the decrease in time was not significant either for *Fdur* (Z=-1,745, p=.081) or *Kdur* (Z=-,524, p=.601). Considering the fact that there was a significant decrease in keyboard activity in the PIO condition but not a significant decrease in overall time, we decided to carry out further qualitative analysis in order to investigate the reasons for less keyboard activity without a subsequent decrease in overall time to complete the task. We hypothesized that this might be due to the time that post-editors spent outsite the CasMaCat workbench doing Internet searches in relation to the texts being post-edited.

This qualitative analysis consisted of visualizing the screen recordings of the post-editing process in order to select specific segments for subsequent analysis and comparison between the seven participants. Such segments were selected according to the following steps:

1. Selection of segments with higher *Fdur* values (6 segments)
2. Verification of the total number of words for those segments
3. Filtering of segments with less number of words and high duration on both conditions (P and PIO)

After watching the screen recordings for the selected segments, our initial hypothesis involving less/more time doing Internet searches while post-editing seemed to be true. The time spent outside the CasMaCat workbench during the post-editing task was computed again for all segments removing the time expend outside the workbench. The results showed then a significant decrease in time when the consultation of external resources is removed from the post-editing process both for *Fdur* values (Z= -3,148, p=.002) and *Kdur* values (Z= -2,524, p=.012) for PIO condition.

There is not a straightforward explanation for the fact that post-editors felt the need to make more Internet searches while post-editing in the PIO condition (which affected overall times), but watching the screen recordings it seems that they doubled checked many of their post-edited options even when seeing them populated from segment to segment thanks to machine learning technique implemented.

The next subsection still relates to the use of external resources while post-editing and it is linked to the second main aim of this third field trial, i.e. to investigate how post-editors use the biconcordancer tool integrated in the latest prototype of the CasMaCat workbench.

4.4.2 Use of external resources: biconcordancer

Research on translation technologies generally attempts to identify user needs, with a view to developing new resources or improving existing tools. In this section we will investigate translators interaction with the biconcordancer (BiConc) in the latest prototype of CasMaCat workbench. In addition to external online tools, participants in the CFT14 had the chance to use the BiConc feature while post-editing. Thanks to this BiConc tool, post-editors were able to retrieve relevant translations/collocations, sorted by their relative frequencies (i.e. the most probable translations are shown first), from the training data available in CasMaCat.

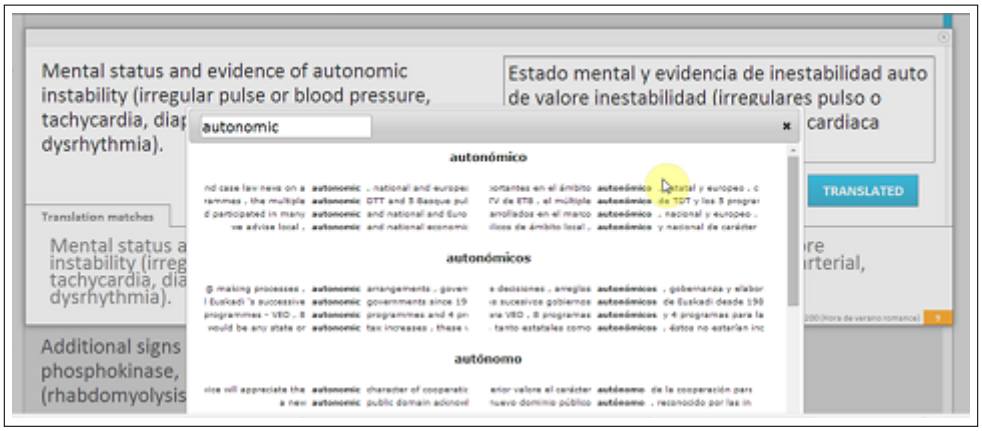

Figure 5: Biconcordancer tool in the third prototype of the CasMaCat workbench.

Using screen-capture recordings, we observed and analysed the way translators interact with the BiConc and other informational resources in order to solve particular translation problems while post-editing. Only three of the seven participants in the study made use of the BiConc tool, being participant P7 the one who logged more searches in the BiConc. Figure 6 shows the use count of the BiConc per conditions and participant.

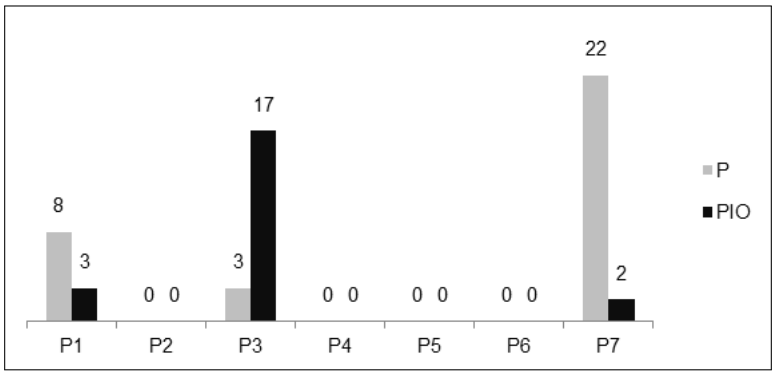

Figure 6: Use count of BiConc tool per participant and condition in CFT14.

The participants who never used the BiConc were also the ones who used less external resources while post-editing. Among the reasons for not using the BiConc, participants reported that they ignored/forgot they had this possibility and automatically used their well-known resources on the Internet.

When inspecting the usefulness of the BiConc tool for post-editing purposes, it can be observed the fact that participants who used the BiConc made it on both conditions but with a significant difference between them. P1 and P7 used the BiConc less under PIO (the condition involving online learning), which could be attributed to the fact that successful searches followed by edits in the text resulted in improved MT outputs where less specific domain information searches were needed. However, P3 shows the opposite search pattern having many more

searches in the PIO condition. In an attempt to find an explanation for this significant difference in the amount of searches between both condition for P3, the reason could be found in the experimental design and a close examination of the screen capture recording showing the post-editing process. P3 post-edited the first text in PIO making a regular use of the BiConc but with few cases of successful retrieval, which seem to have affected her trust in the BiConc in the second text post-edited under the P condition, where she still made many searches during the post-editing process but using other Internet resources instead of the BiConc. Specific insights on the use of concordance reported by the participants can be found in section 5.

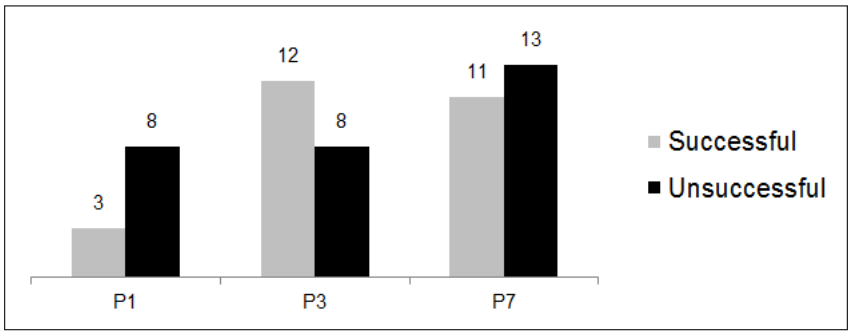

Figure 7: Success rate of retrieval using BiConc per participant.

In addition to the number of times post-editors actually used the BiConc, we were interested in the number of time such searches led to successful cases of information retrieval. A close examination of the searches they made shows how only 47,27% of the BiConc searches offered satisfying results that were subsequently used to post-edit the MT output. Such percentage cannot be considered high enough to develop the trust of the users and it must be related to the user experienced reported for participant P3. Having a close look at the screen capture recordings, it can also be observed how, due to the fairly low successful retrieval rate with the BiConc, participants often double-checked the proposal found in the BiConc against another Internet resource.

Interestingly enough, the most Internet-based resource used while post-editing was another biconcordance tool, i.e. Linguee[3], which was extensively used by all the participants in the field trial. These results show to what extend professional translators rate positively a biconcordance tool in their daily work, which points to the idea that it should be a feature included in the final prototype if the CasMaCat workbench after improving successful retrieval rates.

The following section reports on the third aim of this field trial, i.e. to assess how professional reviewers use the e-pen functionalities while reviewing from CasMaCat GUI

4.4.3 Revision with e-pen

Handwriting is known to be a slower method for text entry with respect to the keyboard and, more so, with respect to speech. However, handwriting can be useful in a reviewing scenario where the user is likely to introduce just a few changes and can benefit from some conventional proof-reading gestures. In such cases, writing with an epen as if it were a conventional proof-reading process on a paper could be faster and more ergonomic. Hence, one of the aims of the third field trial was to assess how professional reviewers use the e-pen functionalities implemented in the CasMaCat workbench. After the seven translators in the CFT14 post-edited the two texts, three different reviewers proof-read their final target texts under two different conditions:

- *Condition 1*: Traditional revision (R), i.e. using the keyboard as the only input method.

[3]http://www.linguee.es/

- *Condition 2*: Revision using an e-pen(RE), i.e. using an e-pen as an input method to enter corrections in the text.

The results show that condition was a bit slower than condition 1 on segment basis, but the differences could be acceptable. Unfortunately, e-pen recognition was perceived as not very accurate, causing many retries which made condition 2 less productive. On the positive side, e-pen gestures were appraised, in spite of minor problems related to font sizes. The evidence points out that, productivity differences could be reduced to the minimum if the UI was redesigned and implemented in a tablet with e-pen support, and the accuracy of the recogniser was improved. Unfortunately, that was out of the scope within the CasMaCat project.

A more detailed analysis of this evaluation can be found in D3.3 in section "Automatic Reviewing (Task 3.7)".

5 Eliciting user feedback

In this section we discuss the feedback provided by the users in the two studies reported in this deliverable. The main data collection tool used to gather user's feedback was a questionnaire that post-editors completed at the end of both studies and in which, apart from replying to the questions, participant could also make further comments.

Table 8 shows a general overview of the participants' profile involved in both studies. The most salient factors in the metadata collected for subject profiling are: i) P04 did not have previous post-editing experience, ii) P03 did not have formal translator training and was less experienced (despite being a regular freelance translator for Celer Soluciones SL), and iii) P05b had much more experience as a professional translator than the rest. Only the first factor seems to have played a role as shown in the collected feedback (this participant was extremely positive about ITP).

Participants	**P01**	**P02**	**P03**	**P04**	**P05a**	**P05b**	**P06**	**P07**
Gender	F	M	F	F	M	F	F	M
Years of translator training	4	4	0	3	14	5	4	4
Years of professional experience	8	8	1	3	14	27	3	11
Post-editing experience	Yes	Yes	Yes	No	Yes	Yes	Yes	Yes
Took part in LS14 study	Yes	Yes	Yes	Yes	Yes	No	No	No
Took part in the CFT14 study	Yes	Yes	Yes	Yes	No	Yes	Yes	Yes

Table 8: Participants' profile in the LS14 and CFT14 studies.

Note: In order to maintain the name conventions of the participants as they are stored in the TPR-DB, we make a distinction between P05a and P05b in this table to be able to differentiate between two different post-editors who were not simultaneously in LS14 and CFT14 and had the same participant number.

The questionnaire used to collect the user feedback presented in this section is available at this *introductory questionnaire*.

5.1 LS14 study

The user feedback derived from the longitudinal study was collected in week six right after post-editing the last text in the study. The main five questions that post-editors had to answer were:

1. If Celer Soluciones SL (or any other LSP) ever gave you the chance to post-edit with or without interactivity, what would you prefer?

2. In your daily work as a professional translator, do you prefer to translate from scratch instead of post-editing machine translation?

3. Would you use CasMaCat as a post-editing tool for your future projects?

4. According to your own personal opinion, what are the advantages of using interactivity while post-editing MT?

5. According to your own personal opinion, what are the disadvantages of using interactivity while post-editing MT?

The aim of the first question was to know if, after having post-edited using interactivity over an extended period of time, they would choose this form of post-editing over the traditional one. All participants, except one, stated that they would still prefer to post-edit without interactivity. Interestingly enough, the only post-editor who responded that she would prefer ITP over traditional post-editing was P03 (the only one without formal translator training and the only one with less than 2 years of experience).

When trying to find out more about their resistance to adopt ITP for post-editing purposes, in the open section of the questionnaire both P01 and P02 provided feedback along these lines: "having to post-edited with interactivity demands a controlled typing speed and this is difficult to achieve when you are an experienced touch typist". Advanced touch typists need to be aware of the fact that they will only benefit from ITP when they stop overwriting most of the suggestions offered by the system. As shown by the data collected, P01 and P02 are the two participants with more cases of overwriting behaviour due to their fast typing speed. In the case of P03, the only one who preferred ITP, she suggested that ITP becomes an effective way to retrieve equivalents as you type ("ITP helped me to find equivalents").

With respect to the second question, four out of the five post-editors in LS14 answered "It depends (on the text type, quality of the machine translation, etc.)". P02 was the only one who would prefer to translate instead of post-edit any under circumstances.

The third question in the final questionnaire wanted to explore to what extend translators would adopt the CasMaCat workbench as a professional tool. P02 and P05 were the only ones who would not use the workbench for further post-editing projects claiming that existing commercial CAT tools already serve this purpose. P01, P03 and P04 stated that they would adopt this workbench for post-editing purposes in the future.

When asked about the benefits of ITP, the responses collected were diverse: P05 stated that he was not able to mention any advantages and P02 argued that he rarely benefited from the suggestions provided by the system. The rest of the participants offered a more positive view of post-editing through ITP acknowledging, for instance, that the idea behind ITP certainly helps to decrease the technical effort (typing) from the post-editors. However, they would have to invest more time to make this principle productive in their personal experience learning not to overwrite many of the ITP suggestions. "I have to retrain myself on typing for ITP purposes", mentioned P01.

With relation to the disadvantages of ITP, all participants (except P03) mentioned that it is difficult to become familiar with the fact that the target text is constantly changing. It is difficult to pay attention to the source text, the target text and, in addition, all the suggestions triggered by the ITP as you type. In addition, P02 suggested that another area of the screen could be used to show these predictions (as it is the case with translation memory matches shown in a separate window).

The feedback collected seemed to offer a clear cut difference between the extremely positive attitude towards ITP shown by P03 (the only one without translator training and less years of experience) and the negative views offered by P05 (the participant with more years of formal training and more experience in this study). These two extremes in terms of experience and formal training certainly played a decisive role for ITP acceptance.

5.2 CFT14 study

The collection of user feedback for CFT14 involved two different types of users: post-editors working with online learning techinques (see section 4) and reviewers using an e-pen as an input method to enter edits (see section 4.4.3).

5.2.1 Online learning techniques

In the case of the CFT14, user feedback from the seven post-editors involved in the study was collected through a single question after completing both tasks in the experiments. Participant's were asked to formulate in which way they thought that the online learning techniques implemented in the CasMaCat workbench made the post-editing process easier for them. Their responses were collected though e-mail once they returned home after completing the experimental task from the company.

All the participants agreed that the benefits of online learning techniques were particularly noticeable at the terminology level, since many of their searches for terms in the text (e.g. *orodispensable, abnormalities, seizure, neuroleptic malignant syndrome*, etc.) were automatically populated from segment to segment. None of the participants had ever worked with a translator's workbench featuring machine learning techniques; however, P01 and P02 mentioned its similarities (implementation differences aside) to the "autopropagate" function that some CAT tools apply for full matches between identical segments. Indeed all participants would very much like to see this feature implemented in a post-editing workbench in order to avoid many of the repetitive tasks that post-editing involves when fixing many times the same MT error along the same text. This positive response with regard to online learning has been considered in the different configurations of the CasMaCat workbench presented in deliverable 1.3.

6 Discussion

This deliverable has reported on two different field trials involving professional translators recruited by Celer Soluciones SL. The first study was an additional longitudinal study (LS14) aiming at evaluation the learning effects while using ITP (PI condition) over a period of six weeks. The second study was the third field trial (CFT14) with the CasMaCat workbench featuring online learning techniques (PIO). The aim of this study was to explore the benefits of working with interactive machine translation combined with online learning techniques for post-editing purposes.

Results from the LS14 study showed how professional translators needed an average of six weeks (see Figure 3) to become familiar with interactivity features for post-editing purposes. The crucial factor in order to obtain a successful interaction between the post-editor and the ITP featured in CasMaCat is directly related to their typing behaviour. Only after post-editors stop overwriting most of the suggestions provided by the system is when productivity gains can be derived from using ITP. Touch typist post-editors find this trade-off between typing speed and the suggestions provided by the system somehow difficult to achieve, but this study have showed that after weeks of interaction a successful interaction can be achieved. It would be interesting to conduct further studies to explore if non-touch typists or non professional translators, with a slower keyboard activity, become acquainted with this technology within a shorter timespan.

With respect to CFT14, results have shown that working with online learning techniques made the post-editing process faster, but only when the time used by the post-editors to make Internet searches is not taken into account. Our analyses made clear that productivity metrics

in terms of overall time to complete the task might not be a good indicator of performance when the post-editor needs to conduct Internet searches in order to verify the quality of the MT provided. As shown in this CFT14 analyses, productivity studies should take into account the translation/post-editing process in order to control for any possible confounding variables that may affect the results. Counter to expectations, post-editors did not seem to be faster under the PIO condition. However, a more detailed qualitative analysis of the process data collected showed that the reason was the search behaviour of the participants while post-editing. They spent a considerable amount of time researching on the internet regardless of the productivity gains that could be derived from working under the PIO condition. Even though participants did not become faster in terms of overall time, their keyboard activity as reflected in $Kdur/Fdur$ values shows that post-editors had to type less when post-editing with interactivity and online learning techniques (condition 2) as opposed to doing traditional post-editing (condition 1). That means that basically these features help post-editors to save efforts during their work.

Based on the feedback collected from participants, most of the participants reported that they would prefer to work without interactivity but with online learning techniques (a hybrid of condition 2 - PO). These results show that further experiments could be run using a third conditions where the CasMaCat workbench features online learning but without interactive translation prediction.