# Chapter 5
# Learning Advanced Post-editing

**Vicent Alabau, Michael Carl, Francisco Casacuberta, Mercedes García Martínez, Jesús González-Rubio, Bartolomé Mesa-Lao, Daniel Ortiz-Martínez, Moritz Schaeffer, and Germán Sanchis-Trilles**

**Abstract** This chapter reports the results of a longitudinal study (LS14) in which the CASMACAT post-editing workbench was tested with interactive translation prediction (ITP). Whereas previous studies with the CASMACAT workbench (Sanchis-Trilles et al., *Machine Translation*, 2014) or similar systems (Langlais et al., *Machine Translation, 15*, 77–98, 2004) tested user interaction only for a few days, the aim of this study was primarily to find out whether and how the performance of professional post-editors improved over time when working with the CASMACAT ITP feature. We were also interested in uncovering any specific profiles of translators depending on personal factors such as previous experience in

V. Alabau (✉) • G. Sanchis-Trilles
Pattern Recognition and Human Language Technology Research Center, Universitat Politècnica de València, Camino de Vera s/n, 46021 Valencia, Spain

Sciling S.L., Valencia, Spain
e-mail: valabau@sciling.com; gsanchis@sciling.com

M. Carl • B. Mesa-Lao
Center for Research and Innovation in Translation and Translation Technology, Department of International Business Communication, Copenhagen Business School, Frederiksberg, Denmark
e-mail: mc.ibc@cbs.dk

D. Ortiz-Martínez • F. Casacuberta
Pattern Recognition and Human Language Technology Research Center, Universitat Politècnica de València, Camino de Vera s/n, 46021 Valencia, Spain
e-mail: dortiz@prhlt.upv.es; fcn@prhlt.upv.es

M. García-Martínez
Computer Laboratory, University of Maine, Le Mans, France
e-mail: mercedes.garcia_martinez@univ-lemans.fr

J. González-Rubio
Unbabel Lda., 1000-201 Lisboa, Portugal
e-mail: jesus@unbabel.com

M. Schaeffer
Center for Research and Innovation in Translation and Translation Technology, Department of International Business Communication, Copenhagen Business School, Frederiksberg, Denmark

Institute for Language, Cognition and Computation University of Edinburgh, Edinburgh, UK
e-mail: mschaeff@inf.ed.ac.uk



M. Carl et al. (eds.), *New Directions in Empirical Translation Process Research*,
New Frontiers in Translation Studies, DOI 10.1007/978-3-319-20358-4_5

post-editing and typing skills. Finally, the aim was also to collect feedback from the post-editors in order to know more about their views regarding this type of technology.

**Keywords** CASMACAT workbench • Interactive post-editing • Interactive translation prediction • Learning behavior in interactive post-editing • Production time • Typing time

## 5.1 Introduction

The way texts are produced changes with every technological invention. From paper and pencil to type-writers and computers, each new technology gives rise to new types of texts, new styles of authoring, and new ways of how texts are generated and perceived. Today we are experiencing increased automation of text production, in particular through the Internet and through novel forms of editing, authoring and translating digital content.

Within EU CASMACAT project (see Sanchis-Trilles et al., *Machine Translation*, 2014 and also Chap. 3 in this volume), we have developed an advanced post-editing platform with an interactive translation prediction mode, context dependent completions during the translation process (Langlais et al. 2004). Even though this feature was designed to help translators in their translation production, within a 3-days field study in a professional translation agency[1] Carl et al. (2013) it semed to hamper translators rather than help them to produce faster translations. Investigating some of the screen recordings, we hypothesized that post-editors might need to get more extended exposure to the CASMACAT workbench as its novel editing features might require completely different translation styles and translation procedures, which first would have to be learned (Sanchis-Trilles et al. 2014). This assumption is in line with experiences gained in a similar translation prediction system, TRANSTYPE (Langlais et al. 2004), where it was suggested that "over a longer period [the system] is expected to give a much better picture of its possibilities".

Accordingly, we conducted a longitudinal study (LS14) which involved five post-editors working alternatively with CASMACAT's traditional post-editing mode and the Interactive Translation Prediction (ITP) mode over a period of 6 weeks. The aim was to test whether post-editors become faster when working with ITP as they become more acquainted with this type of assistive technology, and to investigate whether exposure to this workbench over a longer period of time has an effect on editing behaviour.

[1] Field trials of the CASMACAT workbench were carried out at Celer Soluciones SL, Madrid, who were partner in the CASMACAT consortium

The LS14 study took place in May and June 2014. It was followed in July 2014 by the third CASMACAT field trial (CFT14), for which a more detailed description is contained in Chap. 7 of this volume. The CFT14 study was conducted at the same translation agency, aiming at assessing whether post-editors profit from ITP online learning as compared to traditional post-editing.[2] Seven post-editors participated in the CFT14 study from which four had also taken part in the previous longitudinal study (LS14). As a side effect, we can thus investigate what the four post-editors who participated in both studies have learned, compared to those three post-editors who only participated in the CFT13 study.

The CFT14 study differs from the LS14 study with respect to:

- the text type in LS14 was general news, while CFT14 was a specialized text from the medical domain extracted from the EMEA corpus.[3]
- The number of source text words was also quite different in these two studies: LS14 involved 24 source texts of 1000 words each, while CFT14 involved only two source text with 4500 words 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 add up to around 225,000 source text words which were translated into 249,000 target text words. The studies are included in the publicly available TPR-DB.[4]

Results show that LS14 participants became indeed faster over the period of 6 weeks working with the ITP system and, according to the projection of the data collected, they could have been even more productive after 6–7 weeks of regular exposure to this new technology.

A closer look at the way post-editors became acquainted with ITP suggests that learning to work with this interactive technology requires a different way of controlling the typing speed. In order to be able to fully benefit from the ITP suggestions (i.e. the translation autocompletions) provided by the system, post-editors need to check more frequently the proposals of the ITP 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 new suggestions and thus saving typing effort.

Section 5.2 introduces the LS14 study. It gives background on the participants, the experimental design and the results of the study. Section 5.3 compares behavioral patterns of LS14 participants with CFT14, and tries to describe what exactly is being learned over time. Section 5.4 corroborates these findings with the feedback from participants, as acquired on the basis of questionnaires.

[2] See also Chap. 3 for a comparison of online learning and active learning in the CASMACAT tool.

[3] http://opus.lingfil.uu.se/EMEA.php.

[4] The TPR-DB is available online free of charge from: http://sourceforge.net/projects/tprdb/. The TPR-DB website is at: https://sites.google.com/site/centretranslationinnovation/tpr-db.

## 5.2 A Longitudinal Study with Interactive Translation Prediction (LS14)

### *5.2.1 Participant Profiles*

Five professional translators {P01, P02, P03, P04, P05a} 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 (P04) had previous experience in post-editing MT as a professional service, and all post-editors considered themselves to have excellent typing skills. For three of the four participants with post-editing experience {P01, P02, P05a}, their workload involving post-editing services did not exceeded 10 % of their projects as reported in an introductory questionnaire. The fourth participant (P03) with post-editing experience reported that 75 % of their workload as a professional translator involved post-editing projects. The five post-editors can be grouped in two groups, $L_1$ and $L_2$, as follows[5]:

- $L_1$: {P01, P02, P05a} are the more experienced translators/post-editors
- $L_2$: {P03, P04} where:
  - P03: has no formal translator training and only 1 year experience
  - P04: has 3 years formal translator training and experience, but no post-editing experience

### *5.2.2 Text Type*

The source texts involved in this longitudinal study were pieces of general news extracted from the WMT 2014 corpus. Each source text contained 1000 words on average distributed over 48 segments on average (range 39–61).

### *5.2.3 Experimental Design*

The experimental design involved 24 different source texts which were post-edited from English into Spanish over a period of 6 weeks (four texts per week). MT was

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

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, all post-editors worked on the same four 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. Post-editing guidelines were given, similar to those discussed in Chap. 3, as well as 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 1154 segments, i.e., in total more than 140,000 source text words (half of them in each condition, as shown in Chap. 2, Appendix A). Presentation of texts and task order were counterbalanced, such that participants post-edited in the PI condition first and post-edited in the P condition afterwards half the time. In addition, texts were grouped in two lists: two participants post-edited list A (during their weekly assignments) in condition P and post-edited list B in condition PI, while the remaining three participants post-edited list A in condition PI and post-edited list B in condition P.

### *5.2.4 Results*

In Sect. 5.2.4.1 we provide an overall comparison of the translation durations, in terms of *FdurN*, *KdurN* and *PdurN*, which show that on average all translators slow down in the PI mode. Section 5.2.4.2 shows individual differences in post-editing behaviour: for some of the post-editors total post-editing time can be predicted by typing durations, while for other types of post-editors typing duration is less indicative of the total post-editing time (see Sect. 5.3).

#### 5.2.4.1 Overall Post-editing Durations

The evaluation of the LS14 data is based on three different parameters computed at the segment level[6]:

1. *FdurN*: production time per segment, excluding pauses >200 s, normalised by the number of characters in the source segment.
2. *KdurN*: duration of coherent keyboard activity per segment excluding keystroke pauses >5 s, normalised by the number of characters in the source segment.
3. *PdurN*: duration of coherent keyboard activity per segment excluding keystroke pauses >1 s, normalised by the number of characters in the source segment.
4. *Mins*: number of average manual insertions per source text character
5. *Mdel*: number of average manual deletions per source text character

Table 5.1 gives an overview of average post-editing durations (in ms) and typing activities per source text character for all five post-editors in the two conditions during the 6 weeks. The data show that post-editors needed more insertions but less deletion keystrokes in the PI condition than in the P condition. On average, there are 0.416 manual insertions per source text character in the P condition and 0.538 per source text character in PI, but there are more manual deletions in P (0.371 per source text character) than in PI (0.254). A Wilcoxon-Mann-Whitney test for categorical data revealed that there were more manual insertions in PI than in P ($W = 3{,}410{,}539, p < 2.2e-16$) and more manual deletions in P than in PI ($W = 5{,}617{,}674, p < 2.2e-16$).

**Table 5.1** Overall typing activity (insertions + deletions) and production times in the LS14 data

| Participant | Cond | *Mins* | *Mdel* | *FdurN* | *KdurN* | *PdurN* |
|---|---|---|---|---|---|---|
| P01 | PI | 0.744 | 0.399 | 563.64 | 254.3 | 113.33 |
| P01 | P | 0.595 | 0.545 | 529.71 | 215.86 | 88.51 |
| P02 | PI | 0.5 | 0.223 | 456.53 | 173.06 | 68.24 |
| P02 | P | 0.416 | 0.346 | 439.87 | 157.46 | 68.51 |
| P03 | PI | 0.429 | 0.187 | 623.81 | 223.79 | 85.26 |
| P03 | P | 0.353 | 0.319 | 573.68 | 167.51 | 63.77 |
| P04 | PI | 0.569 | 0.280 | 684.30 | 230.22 | 130.28 |
| P04 | P | 0.362 | 0.329 | 701.46 | 161.53 | 88.32 |
| P05a | PI | 0.447 | 0.181 | 320.72 | 158.18 | 69.99 |
| P05a | P | 0.354 | 0.314 | 284.43 | 138.20 | 54.25 |
| AV. | PI | 0.538 | 0.254 | 529.80 | 207.91 | 93.42 |
| AV. | P | 0.416 | 0.371 | 505.83 | 168.11 | 72.67 |
| AV. | Total | 0.477 | 0.312 | 517.82 | 188.01 | 83.05 |

[6]Further insights on this metrics are reported in Chap. 2

As shown by *KdurN* 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 *FdurN* and *PdurN* show a similar pattern.

#### 5.2.4.2 Individual Differences in P and PI

Figure 5.1 plots the relationship between *FdurN* and *KdurN* for the five participants in the LS14 study. Each point in the graph shows the average *KdurN*/*FdurN* ratio per source text and post-editing condition over 1 week of post-editing activity. Each dot represents the average per-character post-editing duration of approximately 2000 source text words per week and per condition in either of the two post-editing modes (P or PI). That is, each post-editor is represented with six dots representing 6 weeks for the P task and six dots for the performance of the 6 weeks in the PI task.

Different post-editors show different *KdurN*/*FdurN* relations: Experienced post-editors form group $L_1$ show a strong correlation between these two durations ({P01: R = 0.78, P02: R = 0.78, P05a: R = 0.82}), which is not the case for the less experienced translators in $L_2$: {P03: R = 0.74, P04: R = 0.40}). This suggests that experienced post-editors use their time more efficiently and predicably while they work on a segment. Despite being a professional translator, post-editor P04 showed a much weaker correlation between *KdurN* and *FdurN* (R = 0.40) than the other post-editors, probably related to the fact that he had no previous post-editing experience. 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 (R = 0.74).

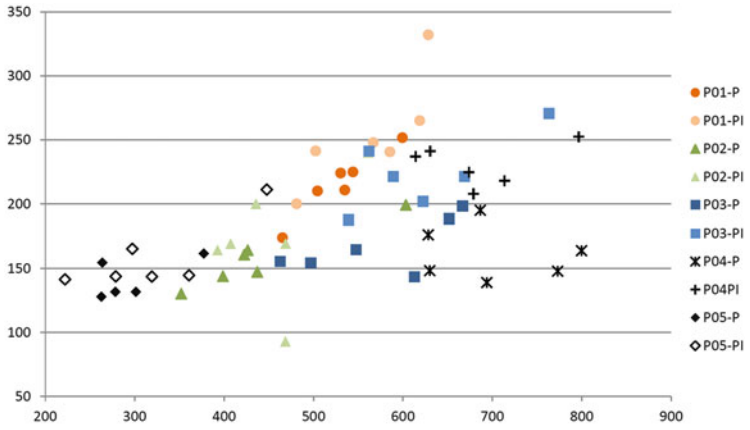


**Fig. 5.1** LS14 study—*FdurN* (*horizontal*) vs. *KdurN* (*vertical*) for all five participants

#### 5.2.4.3 Projecting Learning in P and PI

While the five post-editors show different behaviour, they become substantially quicker in the PI condition over time. However, in the baseline condition (P), there was no improvement over time. Figure 5.2 plots the effect of time on post-editing durations measured in terms of *Kdur* per source text character (i.e. *KdurN*) for the two CASMACAT settings. For this analysis, skipped segments with either zero tokens in the final target text and/or with zero total editing duration and segments with more than one revision were excluded. Segments with more than one revision were excluded, because participants complained that often when a segment was revisited, the initial MT output 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.

Despite the general downwards trend in PI over time, Fig. 5.2 shows a difference in efficiency in week 1 and 6 as compared to the other weeks. The reason for these peaks in production time for weeks 1 and 6 might be the experimental setup itself, since these 2 weeks involved eye-tracking apparatus and the request to post-edit from the company:

1. Having to work from the company office, rather than from home, seems to have had a negative impact on post-editor's performance. During weeks 2–5, post-editors worked from home, which is what they are used to, since all of them work as freelancers.

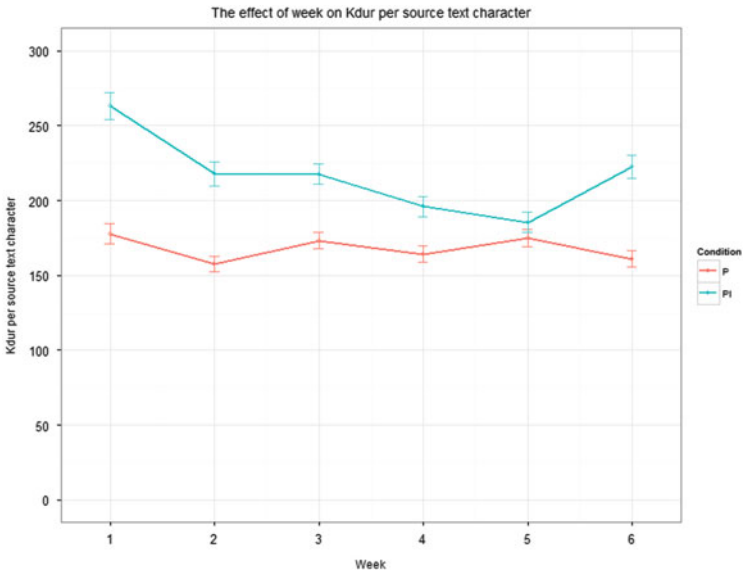


**Fig. 5.2** LS14 study—productivity as reflected in *KdurN* taking into account 6 weeks

2. The ITP mode involves a great amount of internet traffic: a new translation prediction is sent over the internet for (almost) every keystroke. This adds to the traffic from the gaze samples (at a rate of approximately 300 Hz), which are also sent over the internet to a remote server, so that a delay in response was frequently observed in the office of the translation agency when using CASMACAT in the PI setting

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 weeks 1 and 6, or—in other words—the data might show a lab effect.

The productivity drop for week 6 under PI can also be found in the difficulty of the texts themselves: TER values were computed for all the texts in LS14, and values were particularly higher for texts in week 6. We could identify text 20 in week 6 (post-edited under PI by participants P01, P03 and P05) as one of the most difficult texts to post-edit. Text 20 in LS14 was of a more specialized nature of legal text. This different degree in text specialization could be the reason for both lower MT quality and thus requiring more edits from the post-editors, as reflected in the higher number of edits recorded in TER values.

Assuming that working at home and working in the office are two different conditions, we calculated a learning projection based only on the 4 weeks when post-editors worked in the office. Figure 5.3 plots the two conditions in

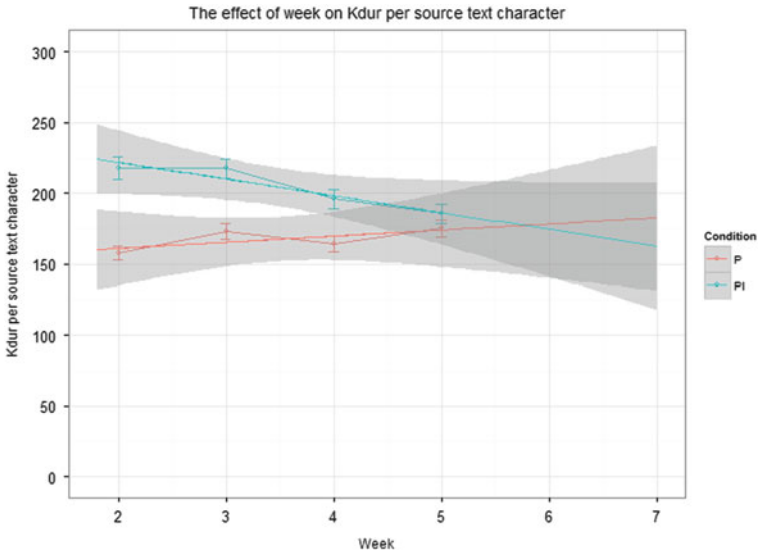


**Fig. 5.3** LS14 study—productivity projection as reflected in Kdur based only on the data from weeks 2–5 (working from home)

LS14 showing that post-editing under the PI condition could have become—theoretically—more efficient already after 6 weeks.

The two regression lines in Fig. 5.3 are based on simple linear models and show the projection of the average post-editing time under the PI and the P conditions over a hypothetical timeframe of 7 weeks. The grey areas around the linear regression lines represent the 95 % confidence region for each regression. According to this projection, it is between weeks 6 and 7 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 condition.

## 5.3 What is Learned During ITP

Singla et al. (2013) have shown that post-editor profiles can be detected automatically and single post-editors can be identified with a certain degree of accuracy on the basis of process data. They create n-gram models based on activity microunits,[7] as well as part-of-speech sequences to automatically cluster post-editors. Discriminative classifier models are used to characterize post-editors based on a diverse range of translation process features. Singla et al. (2013) conclude that classification and clustering of participants could be used to develop translation tool features which are customized to best fit an individual's behaviour.

However, as shown above, when working with the ITP system over a longer period of time, post-editors seem to change and adapt their behaviour, which indicates that translator profiles do not only refer to a static set of properties but a translator's profile can tell us also something about how the individual learns and adapts to new situations.

In this section we assess what it is that post-editors have learned in the 6 weeks during which they were working with the CASMACAT ITP mode. We compare the behaviour of the post-editors involved in the LS14 study with that in the subsequent CFT14 field trial.[8] We briefly introduce the participants in the CFT14 field trial, outline the differences of the texts used in LS14 and CFT14, and highlight the learning effects by comparing the two studies.

[7] Activity units are presented and discussed in Chap. 2. For an alternative approach to define activity microunits, see also Chaps. 8 and 14 in this volume.

[8] For more detailed information on the CFT14 data see Chap. 7.

### 5.3.1 Participants in LS14 and CFT14

Seven post-editors contributed to the CFT14 field trial. These can be separated into two groups: $C_1$:{P01,...,P04} are the four post-editors which previously participated in the LS14 study. In addition there was a group of three new post-editors $C_2$:{P05, P06, P07}, which had no experiences with the CASMACAT PE and ITP modes. This makes it interesting to investigate how the behaviour of the four $C_1$ post-editors who worked in both studies is different from the new $C_2$ post-editors.

Table 5.2 shows a general overview of the participants' profiles involved in both studies. The most salient factors in the metadata collected for subject profiling are:

1. P04 did not have previous post-editing experience
2. P03 did not have formal translator training and was less experienced, despite being a regular freelance translator for Celer Soluciones SL
3. P05 had much more experience as a professional translator than the rest
4. P05a did not participate in the CFT14 field trial

Note that we make a distinction between P05a and P05 in this table to differentiate between two different post-editors who were not simultaneously in LS14 and CFT14 and had the same participant number.

### 5.3.2 Texts in LS14 and CFT14

There were a few differences in the LS14 and the CFT14 studies.

1. For LS14, the goal was to compare the CASMACAT ITP and post-editing (P) modes, while CFT14 aimed at comparing post-editing (P) and ITP with online learning (PIO). A detailed description of the differences is contained in Chap. 3.

**Table 5.2** Information about participants in the LS14 and CFT14 studies

| Participants | P01 | P02 | P03 | P04 | P05a | P05 | 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 5.3** Comparing properties of EMEA corpus translations and the news translations

| Study | TType | *HTra* | *CrossS* | *CrossT* | *SLen* | *TLen* |
|---|---|---|---|---|---|---|
| LS14 | News | 0.612 | 1.60 | 1.29 | 25.0 | 27.85 |
| CFT14 | EMEA | 0.445 | 1.44 | 1.23 | 21.0 | 22.93 |

2. In order to appreciate online learning capacities, texts were much longer in CFT14 than in LS14, but there were only two texts for each translator, one to be post-edited in the P mode and the other in the PIO mode.[9]
3. Whereas the LS14 data is based on an English-to-Spanish news text, the CFT14 study used a medical text extracted from the EMEA corpus.

As shown in Table 5.3, segments on the source side (*SLen*) as well as on the target side (*TLen*) are on average shorter in the medical text than in the news text. The medical text has also a lower translation ambiguity, as indicated by the lower average word translation entropy *HTra*. This is likely due to dense terminology in the medical text, and the reduced choices for medical and chemical term translations, as compared to expressions in the news text. EMEA translations are also syntactically closer to the source text: lower *CrossS* and *CrossT* values indicate greater syntactic similarity between the source and the target language.[10] In summary, translations of the medical text tend to be more literal than news text translation.[11]

Despite the different nature of these texts, it can be expected that the experience with the ITP post-editing mode that $C_1$ translators obtained during the 6 weeks of the LS14 experiment would also carry over to the CFT14 study, while it is likely that the fresh translators in the group $C_2$ who do not have this experience thus show different behaviour.

### *5.3.3 Typing and Translation Times*

The aim of ITP is to reduce the relative time spent on mechanical translation production (i.e. typing). Taking into account individual differences in typing speed, orientation and translation times, we measure the desirable learning effects as the ratio of coherent keystroke activities (*Kdur*) and the filtered total production duration (*Fdur*): *KdurN* indicates the amount of coherent typing activity, while *Fdur* is the overall translation time, so that the ratio *KdurN*/*FdurN* indicates the relative proportion used for typing, the amount of which we want to reduce with the ITP mode.

[9] A comparison of the PIO mode and active learning is discussed in Chap. 3, this volume.

[10] See Chap. 2 in this volume for more details on these metrics.

[11] See also Chaps. 9 and 13 in this volume for a discussion on word translation literality, and how the *Cross* and the *HTra* features are indicators for this end.

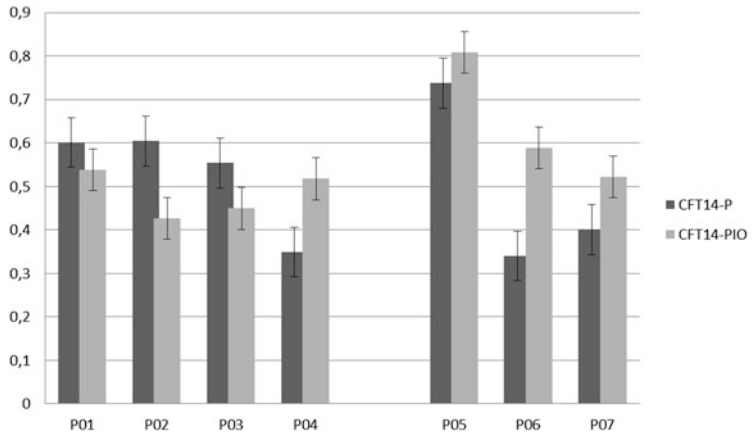


**Fig. 5.4** Ratio of typing time (*Kdur*) and production time (*Fdur*) for $C_1$ and $C_2$ translators

We take the traditional post-editing mode (P) as a baseline and compare the differences in relative typing effort for the ITP mode across different post-editors, and the two groups $C_1$ and $C_2$ with and without extended exposure to CASMACAT. We find that this measure provide another suited indicator for translator profiles, and to capture the learning IPT effects.

Figure 5.4 shows that most of the post-editors in the $C_1$ group ({P01, P02, P03}) have a lower proportion of coherent keystroke activities (*Kdur/Fdur*) in the PIO mode than in the P mode. That is, in the interactive ITP mode these post-editors seem to have learned to accept interactive suggestions which reduces the amount of their coherent typing time, which is not the case for the translators in the $C_2$ group. $C_1$ translators seemed to accept the interactive translation suggestions more often than the new $C_2$ translators by less frequently overwriting the ITP proposals.

Post-editor P04 is an exception in the $C_1$ group, which might be explained by the fact that she did not have any prior experience with post-editing MT output and performed already in the most unpredictable way during the LS14 study (see Fig. 5.1). P05 has the highest *KdurN*/*FdurN* ratio, indicating her ability to make use of her time in the most productive way. Comparing the performance patterns in Fig. 5.4 and taking into account that P04 is (one of) the least experienced translators, while P05 is the most experienced one suggest that the *KdurN*/*FdurN* measure captures some important features.

All post-editors 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 were already pasted in the target text. Learning to control this overwriting behaviour was also reported by the post-editors themselves when providing user feedback, as reported in the next section.

## 5.4 Eliciting User Feedback

User's feedback was collected with a questionnaire that post-editors completed at the end of both studies and in which, apart from answering the questions, participants could also make further comments.[12]

The user feedback derived from the longitudinal study was collected in week 6 right after post-editing the last text in the study. Post-editors had to answer the following five questions:

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, participants would choose ITP over a "traditional" form of post-editing. All participants, except P03, stated that they would still prefer to post-edit without interactivity. Interestingly, P03 was the only one without formal translator training and with less than 2 years of translation experience. She suggested that ITP becomes an effective way to retrieve equivalents as you type ("ITP helped me to find equivalents").

When trying to find out more about the 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 was also visible in the collected screen recordings, P01 and P02 are the two participants with more cases of overwriting behaviour due to their fast typing speed.

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 always prefer to translate instead of post-edit.

The third question in the questionnaire wanted to explore how likely it was that translators would adopt the CASMACAT workbench as a professional tool. P02 and P05a were the only ones who would not use the workbench for further post-editing

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

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: P05a 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 on ITP, acknowledging, for instance, that the idea behind ITP certainly helps to decrease the technical effort (typing). However, they would have to invest more time in order to increase productivity using this novel workbench by learning not to overwrite many of the ITP suggestions. In line with this finding, P01 mentioned "I have to retrain myself on typing for ITP purposes".

With respect 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, to all the suggestions triggered by the ITP. In addition, P02 suggested that another area of the screen could be used to show these predictions—similar to how translation memory matches are 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 experience) and the negative views offered by P05a (the participant with most years of formal training and many years of experience). These two extremes in terms of experience and formal training certainly played a decisive role for ITP acceptance.

## 5.5 Discussion

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 6 weeks (see Fig. 5.3) 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 can productivity gains be reached by using ITP. Touch typists find this trade-off between typing speed and the suggestions provided by the system somehow difficult to achieve. This study shows 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 more easily acquainted with this technology within a shorter timespan.

Most of the participants reported that they would prefer to work without interactivity but with online learning, a technique which is described in more detail in Chaps. 3 and 7 in this volume.

**Acknowledgements** The work described in this chapter was carried out under the auspices of the EU project CASMACAT: Cognitive Analysis and Statistical Methods for Advanced Computer Aided Translation, supported by the European Union 7th Framework Programme Project 287576 (ICT-2011.4.2). Website: http://www.casmacat.eu.

## References

Carl, M., Martinez, M.G., Mesa-Lao, B., Underwood, N., Keller, F., & Hill, R. (2013). *CASMACAT project tech report: D1.2: Progress report on user interface studies, cognitive and user modeling*. European Commission.

Langlais, P., Lapalme, G., & Loranger, M. (2004). Transtype: Development evaluation cycles to boost translators productivity. *Machine Translation, 15*, 77–98.

Sanchis-Trilles, G., Alabau, V., Buck, C., Carl, M., Casacuberta, F., García-Martínez, M., et al. (2014). Interactive translation prediction versus conventional post-editing in practice: A study with the CasMaCat workbench. *Machine Translation, 28*(3–4), 217–235.

Singla, K., Carmona, David O., Gonzales, A., Carl, M., & Bangalore, S. (2013). Predicting post-editor profiles from the translation process. In *Proceedings of the Workshop on Interactive and Adaptive Machine Translation, AMTA Workshop*, Vancouver, Canada (pp. 51–60).