The Process of Post-Editing: a Pilot Study

Michael Carl, Barbara Dragsted, Jakob Elming,
Daniel Hardt, and Arnt Lykke Jakobsen*

Department of International Language Studies and Computational Linguistics
Copenhagen Business School, Dalgas Have 15, DK2000 Frederiksberg, Denmark

Abstract. We report on experiments in which manual translation is compared with a process in which automatically translated texts are post-edited. The translations were performed using Translog, a tool for monitoring and collecting keystroke and gaze data. The results indicate that the post-editing process resulted in a modest improvement in quality, as compared to the manual translations. Translation times were lower for the post-editing. It was also found that post-editing involved notable differences in gaze behavior.

Key words: Translation Process, Post-editing, Machine Translation

1 Introduction

The results of empirical research on translators' productivity as they post-edit machine-translated text in comparison with their productivity when they translate text more traditionally, either manually without any special technological support or with the support of a translation memory system, have mostly been inconclusive [10, 9]. There may be many reasons why no very conclusive results have been produced. A major factor may be that translators often object to being asked to improve on a machine's inferior text.

This situation strikes us as necessarily transitional. Not very long ago there was similar resistance among translators to using translation memory (TM) systems, but that has been almost universally overcome, and everywhere the professional translation environment now includes a TM system.

Perhaps a TM system has a more human appearance than an MT system. Both the fact that it is conceptualized as a 'memory' and the fact that its database is a record of human translations, and perhaps also the fact that the human translator has full control of how the translation is constructed, contribute to making this kind of man-machine interaction acceptable and indeed meaningful to the human user.

However, the most recent TM systems now include an MT component so that users of TM systems have the opportunity to interact with the machine

* Thanks to Kristian T. H. Jensen for help with information on the results from his experiments on manual translations of these texts. Also, thanks to our colleagues at Copenhagen Business School for serving as subjects of the experiments.

in a different mode, namely by post-editing text generated not by a human translator but by the machine. This addition of MT to successful TM solutions reflects the widespread view of MT developers that MT, especially statistical MT (SMT), has improved quite radically in recent years and deserves to be more widely used and accepted. In their view, considerable productivity gain could be obtained (a) if post-editing was accepted as a meaningful method of producing a translation and (b) if acceptance was followed up by post-editing training.

In order to properly test such assumptions, we plan to conduct a longitudinal study in order to trace the effect of training on positively motivated translators. The pilot study reported in the present paper lacks this longitudinal dimension, but was undertaken in order to find out how translators with no post-editing training at all would perform when asked to post-edit MT-produced text in comparison with the performance of a group of translators who had translated the same texts manually, without any dictionary or technical assistance. We chose three English news texts which were to be translated into Danish. We specifically wanted to see how post-editing Google Translate versions of the three texts would compare with translating the three texts manually, in terms of the quality of the translations produced, and the time it took to produce them. We also investigated various features of keyboard and gaze activity.

2 Experiments

The manual translation data was elicited in experiments conducted by K. T. H. Jensen in 2008-2009 [8]. In his PhD study of allocation of cognitive resources in translation, he had 24 participants translate a warm-up text and three British newspaper texts, A, B, and C, assumed to be at different levels of difficulty. 12 participants were MA students (one male), 12 were professional translators (three male), all with Danish L1 and English L2. The English source texts averaged about 850 characters and were translated under different time constraints.

In the current experiment, we chose 8 translations from each of the manually translated A, B and C texts which had no time constraints. In our post-editing experiment, we used the same three texts (A, B, and C), and asked 7 translators to post-edit English-to-Danish machine-translated versions produced by Google Translate.

All 7 translators were native Danish speakers. Three of them had professional translation experience, two post-editors had a formal translation background (but no extended professional experience), and one post-editor was a bilingual native Danish speaker with no formal translation education. None of them had significant experience in using CAT tools. Three of the translators had already manually translated the texts 2 years before in the context of the manual translation, but we think that this did not have a measurable impact on the translation performance, given the long lapse of time between these two events and also the different nature of the two tasks.

The post-editing was performed using Translog [3], a tool for monitoring and collecting keystroke and gaze data during translation. Translog consists of two windows: the source text is shown in the top window, and the translator types the translation into the bottom target window. At the beginning of the post-editing experiment, the Source Text (ST) was displayed in the top window, and the Google Translate output was pasted into the target window at the bottom of the screen. These translations were then post-edited by the translators. Table 1 gives an overview of the properties of the manual and the post-edited translations. On average, the post-edited translations were slightly shorter than the manually translated versions, there were many more deletions during post-editing than during manual translation, there are less insertions, and when post-editing, translators used navigation keystrokes and mouse clicks much more often.

Table 1. Averaged keyboard activity data over 7 versions of three post-edited and three manually translated texts from seven translators

	Post-editing						Manual Translation				
	Google	TT len	insert.	delet.	navi.	mouse	TT len	insert.	delet.	navi.	mouse
A text	834	853	221	112	491	12	884	945	61	35	5
B text	863	903	281	127	379	21	949	1089	127	183	6
C text	865	915	181	74	390	13	905	976	66	47	3

The Google translations of the three English texts consisted of A:834, B:863 and C:865 characters, whereas the average length of the post-edited translations was A:853, B:903 and C:915 characters, and the average length of the manual translations was A:884, B:949 and C:905 characters. It is interesting to note that almost all translations (the post-edited as well as the manual translations) were longer than the Google translations.

Note that the number of insertion keystrokes minus the number of deletion keystrokes does not equal the length of the final TT translations, since highlighting a word by using, e.g., the left or right arrow in combination with shift+control would count as one (navigation) keystroke, but the deletion of a highlighted sequence can be achieved by just hitting the delete (or backspace) key once, or by overwriting it with another (sequence of) character. The latter activity would then count as an insertion, rather than (or in addition to) a deletion, even though the highlighted sequence is deleted. The table shows that the usage of the keyboard for post-editing and manual translation is quite different.

2.1 Evaluation of Translation Quality

The quality of each translation was evaluated by seven native Danish speaker evaluators. Four of the evaluators were professional translators from the CBS

teaching staff and two evaluators had at least 3-5 years of translator training at CBS, and again one evaluator had no translator background, but was a Danish native and fluent English speaker. Each evaluator was presented with a source sentence together with four candidate translations. In each case two translations had been produced using manual translation and two had been produced using post-editing. The presentation order was randomized. Evaluators were instructed to order (rank) the candidate translations from best to worst quality, with ties permitted. This method is frequently used for the evaluation of MT system output [1, 2], but is less familiar in evaluating human-produced translations.

Each sentence was ranked by at least two evaluators. Also, each evaluator was presented with two repeats of a source sentence together with the same four proposed translations. This was done to permit calculation of inter-coder and intra-coder agreement.

Inter-coder agreement: agreement is defined with respect to a given pair of candidate translations for a given source sentence. That is, for two coders c_1 and c_2, we have a source sentence s, and two candidate translations t_1 and t_2, both of which received rankings from coders c_1 and c_2. We say that the two coders agree if their rankings for t_1 and t_2 stand in the same relation. In other words, there is agreement if one of the following three conditions holds:

1. $rank(c_1, t_1) > rank(c_1, t_2)$ AND $rank(c_2, t_1) > rank(c_2, t_2)$
2. $rank(c_1, t_1) < rank(c_1, t_2)$ AND $rank(c_2, t_1) < rank(c_2, t_2)$
3. $rank(c_1, t_1) == rank(c_1, t_2)$ AND $rank(c_2, t_1) == rank(c_2, t_2)$

There were a total of 125 pairs which were evaluated by two coders. Of these, there was agreement in 57 of them, or 46%. Assuming that chance agreement is 33%, we compute a Kappa of 0.188. While this is better than chance, it is considered Slight agreement [6]. This is consistent with the general feeling of evaluators that ordering candidate translations was a difficult task.

Intra-coder agreement: there were a small number (14) of repeat sentences, where the same coder was presented with identical pairs of candidate translations. Here six were in agreement (42.8%), for a Kappa of 0.147.

The data show that intra-coder agreement is even lower than inter-coder agreement. The fact that agreement is so low suggests to us that the assessment of translation quality was simply too difficult.

3 Analysis

3.1 Translation Quality

As mentioned above, evaluators ranked 4 translations of one source sentence at a time, where 2 translations were taken from the manual translations and 2 from the post-edited translations. Subsequently, each sentence was scored according to how often it was ranked better than the translations of the other mode. For instance, if a post-edited sentence was ranked better than one manual translation and worse than the other manual translation, it received a score of 1. If a manual

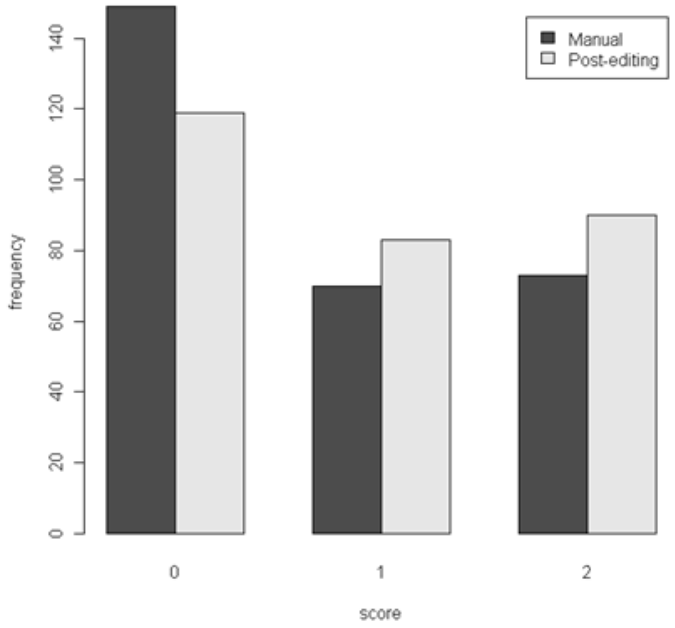

Fig. 1. A comparison of the frequency of evaluation scores of the manual translations and post-edited sentences. Higher scores are better.

Fig. 2. Average scores of post-edited and manually translated texts: A-texts (left) B-texts (middle) and C texts (right).

translation was ranked better than both post-edited translations, it received a score of 2, and if it was not better than any translations of the other mode, it received a score of 0. Accordingly a sentence can have one of 3 scores, where higher scores represent better rankings. The score of a manual or post-edited translation was as follows:

- 2: better than both of the other-mode translations
- 1: better than one of the other-mode translation
- 0: not better than any of the other-mode translations

The distribution of sentence evaluation scores is shown in Figure 1. The graph indicates that the post-edited translations are judged to be better than the equivalent manual translations. The difference is not quite significant, according to the Wilcoxon signed-rank test ($p = 0.05053$). It is however an interesting result that translation quality does not seem to be reduced by the integration of machine translation in the translation process.

The average scores over all the sentences in the post-edited and the manually translated A, B and C texts are shown in Figure 2 below.

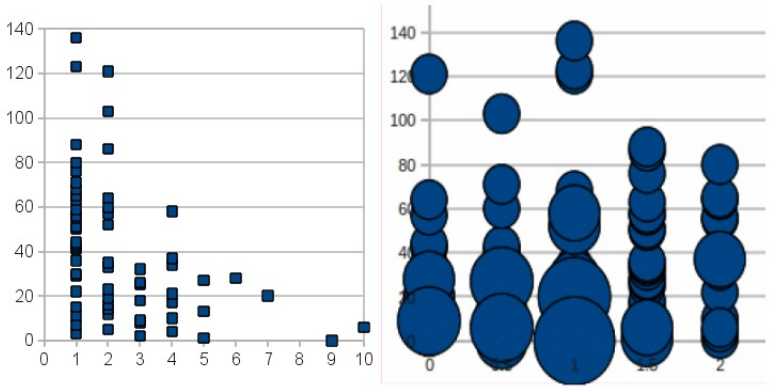

Fig. 3. Correlation between the number of edit operations (vertical axis) and correlation with translation score (right) and number of sentences (left).

3.2 Edit Distance and Translation Quality

The 21 post-edited texts consisted of 133 sentences (8 sentences in the A text, 6 sentences in the B text and 5 sentences in the C text). For each of the 133 post-edited sentences the edit distance was computed. The edit distance indicates the minimum number of changes between the Google translation and its post-editing version. Figure 3 shows that there are between 0 and up to 136 edit operations per sentence. The distribution of edit operations is shown in

figure 3 (left): As can be expected, there are only a few sentences with many operations, and there are more sentences with few operations, e.g. 1 sentence with 136 edit operations, but 10 sentences with 6 operations. We also computed the correlation between the average score of the post-edited sentences, as described in section 3.1, and the number of edit operations per sentence. Since there were two scores from two different evaluations per post-edited sentence, we computed 266 correlations between edit distance and translation score, which are shown in figure 3 (right). Bigger bubbles represent more occurrences of the operation/score relation. Surprisingly, there is no correlation between the score of the post-edited sentence and the number of edit operations, indicating that more post-editing does not necessarily lead to better translations.

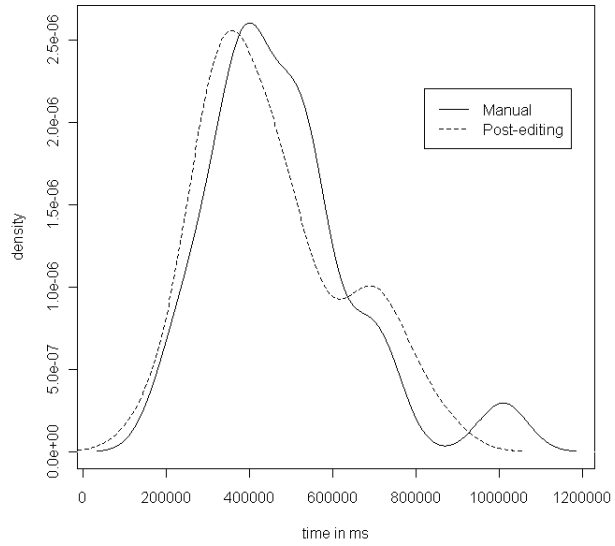

Fig. 4. The estimated distribution of the time spent manually translating and post-editing a text.

3.3 Time

One of the most obvious reasons for engaging in post-editing is the desire to save time. Figure 4 shows the estimated distribution of the time spent on manually translating a text compared to post-editing a text. The two distributions are quite similar, but there is an indication that post-editing may lead to some time saving, though not a significant difference ($p = 0.7118$). This may partly be due to the low number of participants in the tasks. On average a text was post-edited in 7 minutes 35 seconds, while a manual translation took 7 minutes 52

seconds. In this context it should be noted that while all manual translators had experience in translating, none of the post-editors had experience post-editing or using CAT tools. We expect that more post-editing experience will yield a margin of time saving.

3.4 Gaze

We recorded participants' gaze activity in the manual translation and the post-editing task. In the manual translation task, we used a Tobii 1750 eye tracker, which runs at a frame-rate of 50 Hz, and in the post-editing task, we used a Tobii eye tracker which runs at 60 Hz.[1] Both are remote eye-trackers which use binocular eye tracking. The texts were presented on a TFT display with a resolution of 1280x1024 pixels. Participants sat 60 cm from the screen, and were calibrated with a standard 9-point grid.

A basic assumption in eye movement research is that "the eye remains fixated on a word as long as the word is being processed" [5]. Gaze duration is thus taken to signal the cognitive effort associated with processing a particular item, and fixations in reading tend to be longer on items requiring effortful processing, for instance less frequent words, words containing spelling errors, ambiguous words and words which are inappropriate in a given context [7]. Evidence from reading studies suggest that the majority of words in a text are fixated during reading, but that some words are skipped and some words are fixated more than once [11]. The number of regressions has been found to increase as texts become more difficult. Conversely, the eyes are less likely to fixate highly predictable words [7]. In translation, fixation counts are generally higher than in reading, with regressions occurring more frequently [4], and the average fixation count per minute has been found to be significantly higher in complex texts than in simpler texts. Gaze times have similarly been used as indicators of particular items requiring larger cognitive effort [12].

Total gaze time on both areas of the screen (ST and TT) was approximately the same in the two tasks, 263,938ms in the manual translation task and 295,508ms in the post-editing task on average across the three texts. Since the average total task time was lower in the post-editing task (see section 3.3), a higher proportion of time was spent looking at the screen. The slightly larger gap between total task time and total gaze time in the manual translation task may indicate that more time was spent looking outside the screen, most likely at the keyboard, when the translation was produced manually. Another intuitive explanation may be that when producing a translation from scratch, translators may stare off into the space as they await inspiration - something they would not do in a more "mechanical" post-editing task. However, off-screen fixations were not recorded in any of the tasks and the distribution between gaze time and task time will need to be investigated further in future studies.

We analysed the distribution of gaze activity in Translog's ST window vs. its TT window in the two tasks, using the measures fixation count and total gaze time, to investigate which of the two areas attracted most visual attention. In the manual translations, the number of fixations was distributed more evenly

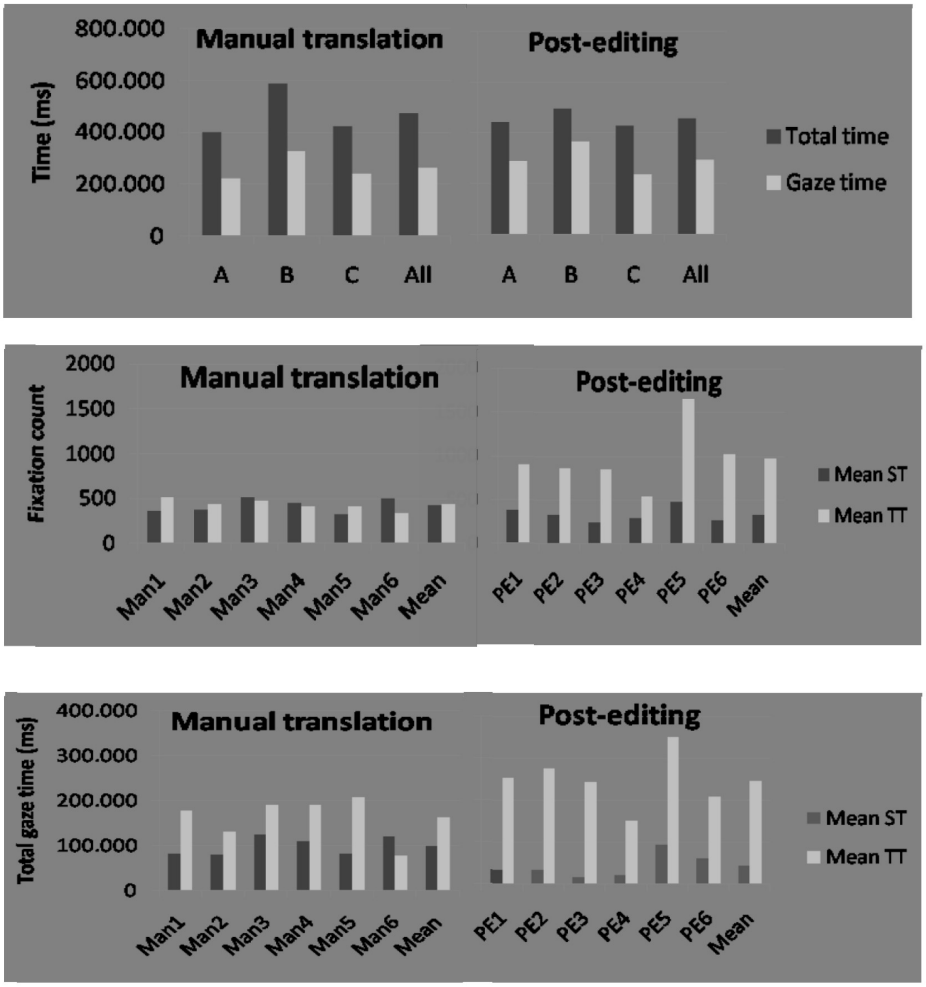

Fig. 5. Comparision of post-editing vs. manual translation behaviour with respect to 1. total translation time vs. total gaze time (top) 2. mean fixation counts on the source vs. target text (middle), 3. total gaze time on the source vs. target text (bottom)

on ST and TT than in the post-editing task. The average fixation count in the manual translation was 420 on the ST area of the screen and 434 on the TT. In the post-editing task, participants fixated the ST 334 times on average against 975 fixations on the TT. Means for six participants in each group are shown in Figure 5. The total gaze time was higher on the TT area than on the ST area in both tasks: 163,364ms on average on the TT against 100,575ms in the ST in the manual translation task, and 247,226ms on average on the TT against 43,055ms on the ST in the post-editing task.

Differences between fixation count and total gaze time in terms of ST/TT distributions show that participants had longer average fixation durations on the TT area in both tasks (Figure 5), but the tendency for most visual attention to be on the TT was most pronounced in the post-editing task, and both the fixation count and the total gaze time on the TT were significantly higher in post-editing than in manual translation according to an unpaired two-sample t-test ($p < 0.01$).

Editing SMT output thus apparently requires a higher TT reading and rereading effort than manual translation. The gaze activity in the post-editing task reflects a process, it may be assumed, of first reading a segment of raw SMT output, then comparing this against a segment in the ST that it is a translation of, and then possibly correcting the machine-translated output and reading the corrected version one or several times. In manual translation, TT gaze activity simply involves monitoring and possibly correcting one's own manual translation output, a process which, based on the eye movement data, requires less effort.

The ST was consulted more frequently (see Figure 5, middle, the dark bars in fixation count) and in particular attracted longer fixations (see Figure 5, bottom[1]) when participants produced a translation manually than when they post-edited SMT output. The number of fixations on the ST was not very different from the post-editing task (it was slightly higher, but the difference was not significant according to an unpaired two-sample t-test, $p = 0.07998$), but the duration of each fixation was longer on average, leading to significantly longer total gaze time on ST during manual translation ($p < 0.01$). This indicates that a different type of ST comprehension is involved in a post-editing task than in manual translation. Manual translation seems to imply a deeper understanding of the ST, requiring more effort and thus longer fixations, whereas in post-editing, the ST is consulted frequently but briefly in order to check that the SMT output is an accurate and/or adequate reproduction of the ST. Also, it may be assumed that in post-editing, the translator reads the SMT output in the TT window before consulting the ST, whereas in manual translation, the ST is naturally attended to first. Note, however that none of our translators had experience in post-editing. The observed behaviour might change dramatically as the translators become more acquainted with the task. This will have to be investigated further.

[1] The total gaze time is the product of fixation count and fixation duration.

4 Conclusion

MT technology has been developing rapidly in recent years, and many have suggested that it can have a major impact on productivity in the translation process, when followed by a post-editing process. However, there is a widespread belief among translators that MT has a negative effect on translation quality, and there is also skepticism that post-editing MT can be done as quickly as ordinary translation. The present study represents a preliminary attempt to address these issues. We found striking differences in both the keyboard and gaze activity of translators when doing post-editing as opposed to manual translation. Furthermore, we found that translation speeds were on average somewhat faster with post-editing, together with a modest increase in translation quality.

These results provide indications that post-editing MT may indeed be shown to have a positive effect on productivity. Given the small scale of the current study, however, no firm conclusions can yet be drawn. Furthermore, our results show that the evaluation of translation quality was extremely difficult. We believe that this difficulty derived in large part from the fact that evaluators were asked to perform relative evaluations, and nearly all the translations were of very high quality. In subsequent studies, we intend to address this issue by asking evaluators to perform more traditional categorical evaluations, in particular asking them to focus on clearly identifiable problems in translation quality. Our results, preliminary as they are, are consistent with a widespread belief that reductions in translation time are possible by doing post-editing. In subsequent work we will pose the question: under what conditions are such reductions possible without a negative effect on translation quality?

References

1. Callison-Burch, C., Fordyce, C., Koehn, P., Monz, C. & Schroeder, J. 2007. (meta-) evaluation of machine translation. In Proceedings of the Second Workshop on Statistical Machine Translation (pp. 136–158). Association for Computational Linguistics.
2. Elming, J. 2008. Syntactic Reordering in Statistical Machine Translation. Ph.D. Thesis. Copenhagen Business School, Denmark.
3. Jakobsen, A. L. 1999. Logging target text production with Translog. Copenhagen Studies in Language, 24. pp. 9–20.
4. Jakobsen, A. L. and Jensen, K. T. H. 2008. Eye movement behaviour across four reading tasks. In Göpferich, S., Jakobsen A. L. and Mees, I. M. (eds) Looking at eyes. Eye-tracking Studies of Reading and Tranlsation Processing. Copenhagen Studies in Language 36. 103-124.
5. Just, M.A. and Carpenter, P.A. 1980. A theory of reading: From eye fixations to comprehension. Psychological Review. Vol. 87. No. 4. 329-354.
6. Landis, J. Richard and Koch, Gary G. 1977. The measurement of observer agreement for categorical data. Biometrics, 33, 159–174.
7. McConkie, G.W. and Yang, S. 2003. How cognition affects eye-movements during reading. In Hyönä, J., Radach, R. and Deubel, H. (eds) The mind's eye: Cognitive and applied aspects of eye movement research. 413-427.

8. Jensen, K. T. H. (to appear in 2011) Allocation of cognitive resources in translation: an eye-tracking and key-logging study. Ph.D. Thesis. Copenhagen Business School.
9. Krings, H. P. 2001 (tr. and ed. by G. S. Koby et al.) Repairing Texts. Empirical Investigations of Machine Translation Post-Editing Processes. Kent, Ohio: Kent State UP.
10. O'Brien, S. 2007. Pauses as Indicators of Cognitive Effort in Post-Editing Machine Translation Output. Across Languages And Cultures, 7, 1, pp1–21 Rayner, K. 1998. Eye movements in reading and information processing: 20 years of research. Psychological Bulletin Vol. 124. No. 3. 372–422.
11. Rayner, K. 1998. Eye movements in reading and information processing: 20 years of research. Psychological Bulletin Vol. 124. No. 3. 372–422.
12. Sharmin, S. Špakov, O. Räihä, K. J. and Jakobsen, A. L. 2008. Where on the screen do translation students look while translating, and for how long? In Göpferich, S., Jakobsen A. L. and Mees, I. M. (eds) Looking at eyes. Eye-tracking Studies of Reading and Tranlsation Processing. Copenhagen Studies in Language 36. 31-51.