Does post-editing increase usability? A study with Brazilian Portuguese as Target Language

Sheila Castilho
CNGL/SALIS
Dublin City University
Ireland
castils3@mail.dcu.ie

Sharon O'Brien
CNGL/SALIS
Dublin City University
Ireland
sharon.obrien@dcu.ie

Fabio Alves
Federal University of Minas Gerais
(UFMG)
Brazil
fabio-alves@ufmg.br

Morgan O'Brien
McAfee
Mahon, Cork
Ireland
morgan_o'brien@mcafee.com

Abstract

It is often assumed that raw MT output requires post-editing if it is to be used for more than gisting purposes. However, we know little about how end users engage with raw machine translated text or post-edited text, or how usable this text is, in particular if users have to follow instructions and act on them. The research project described here measures the usability of raw machine translated text for Brazilian Portuguese as a target language and compares that with a post-edited version of the text. Two groups of 9 users each used either the raw MT or the post-edited version and carried out tasks using a PC-based security product. Usability was measured using an eye tracker and cognitive, temporal and pragmatic measures of usability, and satisfaction was measured using a post-task questionnaire. Results indicate that post-editing significantly increases the usability of machine translated text.

1 Introduction

This paper discusses the measurement of usability for raw machine translated output and post-edited output for instructional text relating to a commercial PC security product machine translated from English into Brazilian Portuguese.

Authentic English source text relating to the software product (anonymised for confidentiality reasons) was identified and machine translated into Brazilian Portuguese using the freely available MT engine, Microsoft Bing.

Eighteen users were recruited to read the instructions and carry out tasks by creating files and folders, changing settings within the product etc. The participants were divided equally into two groups; one group used the raw machine translated instructions and the other used the post-edited instructions. The usability of both sets of instructions was investigated using screen recording, eye tracking and a post-task questionnaire. The main objective of this project was to investigate the extent to which human post-editing of machine translation impacted on the usability of instructional content.[1]

The paper is structured as follows: Section 2 discussed related research, Section 3 explains the methods used, Section 4 provides results and Section 5 the conclusions.

2 Related Work

The task and process of post-editing has received significant attention in the past few years (e.g. Guerberof (2014), De Almeida and O'Brien (2010), Depraetere (2010), Plitt and Masselot (2010), Sousa et al. (2011), Koponen (2012), O'Brien et al. (2012), O'Brien et al. (2013), Specia (2011)). While MT technology has made sig-

[1] This research is supported by the International Strategic Cooperation Award through Science Foundation Ireland and Dublin City University.

nificant strides in the last decade, it is accepted that post-editing is needed in cases where the content is required for more than gisting purposes. Empirical research has demonstrated that post-editing can lead to higher productivity, without having negative effects on quality (e.g. Guerberof, forthcoming), though it might have an impact on perceptions of stylistic quality (Fiederer and O'Brien 2009). Yet little empirical research has focused on the value of post-editing or on its return on investment (ROI). It is generally assumed that post-editing is required to bring content to a publication-ready level, but we know very little about the impact that post-editing has on the usability and, by extension, acceptability of machine translated content.

Related work is at this stage still somewhat limited. Jones et al (2005) present a usability test where participants answer questions from a machine translated version of an Arabic language test. Their results suggest that MT may enable an ILR level 2 (limited working proficiency) but it is not suitable for level 3 (general professional proficiency).

Stymne et al (2012) use eye tracking as a complement to MT error analysis. They found that MT errors have longer gaze time and more fixations than correct passages of text and the average gaze time is dependent on error types, which could indicate that some error types require more cognitive effort than others.

In 2010, Doherty, O'Brien and Carl tested the use of eye tracking as a machine translation evaluation technique, concluding that eye tracking was a reliable method for evaluating the quality of machine translated output. Building on this, Doherty and O'Brien (2014) conducted a study to compare the usability of raw machine translated output for four target languages against the usability of the source content (English). The conclusion of that study was that, although the raw MT output scored lower for usability measurements when compared with the source language content, the raw MT output was deemed to be usable, especially for Spanish as a target language. The target language Japanese, unsurprisingly, scored lowest in terms of usability.

The study by Doherty and O'Brien (2014) used both questionnaires and eye-tracking measurements to record levels of usability. The current study builds on that, but is different in several respects: (1) the content translated differs; (2) the target language in this case is Brazilian Portuguese, which was not included in the 2014 study; (3) the MT system differs and, most importantly, (4) the current study compares the usability of raw MT output against post-edited content, not against the usability of the source language content, which was the case for the previous study.

3 Methods

In this section we discuss the methods deployed to measure usability and the experiment design.

3.1 Measuring Usability

We adopt the ISO/TR 16982 definition for usability: "the extent to which a product can be used by specified users to achieve specified goals with **effectiveness**, **efficiency**, and **satisfaction** in a specified content of use" (ISO 2002).

When this definition is divided into its component parts (in bold above), it allows us to measure different aspects of usability using a variety of methods.

Effectiveness is measured through goal completion, that is, how successful the users were at accomplishing tasks documented in the instructions measured by observing the user interactions as recorded by a Tobii T60XL eye tracker.

Efficiency is measured as the number of successful tasks completed (out of all possible tasks) when total task time is taken into account. A second measure of efficiency is cognitive effort, i.e. how much cognitive effort is evident when users are reading the instructions and trying to complete their tasks? Cognitive effort is measured using typical indicators recorded via the eye tracking apparatus, i.e. mean total fixation time, mean fixation duration, total fixation count, average visit duration and visit count. Such fixation data are well established as indicators of cognitive effort (Rayner 1998, Rayner and Sereno 1994, Radach et al. 2004). For example, the more fixations there are on a set of instructions, the more probable it is that the reader is having difficulties in processing the instructions.

Satisfaction is a measure of user satisfaction with the translated content and, by extension, the product itself. As satisfaction is a multi-faceted concept, we measure it using a questionnaire with a Likert scale ranging from 1 (Strongly Disagree) to 5 (Strongly Agree). In our questionnaire, "satisfaction" is addressed using a number of statements (see section 4.8).

3.2 Content

In collaboration with an industry partner, we selected a security software product that controlled for viruses, allowed for the setting of parental controls and so on and instructional content in English on how to configure features of this product. The total number of words in the source content amounted to 594. This content was machine translated into Brazilian Portuguese using Microsoft's Bing engine.[2] Brazilian Portuguese was selected for this study as it was part of a Brazil/Ireland research collaboration project. The raw machine translated output was post-edited by a native speaker of Brazilian Portuguese who has an undergraduate degree in linguistics and literature and a Master's degree in natural language processing and human language technology. The post-editor had also conducted research previously on post-editing. The guidelines adhered to during post-editing were those of TAUS for the level "fit-for-purpose" (TAUS: online). From a practical perspective, this meant that edits were carried out when terminology did not conform to the client-specific glossary and grammatical errors were fixed. No edits were implemented for purely stylistic reasons and the focus was on accuracy and comprehensibility.

To measure how much post-editing was performed we conducted an automatic evaluation comparing the post-edited version against the MT output. We observed an average HTER score of 0.20 which indicates that post-editing was of a light nature.

3.3 Participants

18 native speakers of Brazilian Portuguese where recruited from the student body of the Federal University of Minas Gerais, Belo Horizonte, Brazil.[3] It was ensured that participants had no previous experience of this particular security product so that previous knowledge could not be used to compensate for poor quality machine translation output (Moravcsik and Kintsch 1995, Kaakinen et al. 2003).

The participants were randomly assigned to one of two groups: Group 1 used the raw machine translated output and were asked to follow the instructions while Group 2 read and followed the post-edited instructions. Neither group knew that the texts they were reading had been translated. Both groups were given a warm up task where they were asked to read a text in Brazilian Portuguese for comprehension; the text came from Wikipedia and explained the concept of virus checking. Fixation data gathered during this reading exercise were used as a baseline measurement for 'reading for comprehension' in Brazilian Portuguese among participants. Two participants (one from each group) appeared to be outliers in terms of several of the fixation measurements and were removed from each group.

Participants were seated at the eye tracker and were informed that they would be presented with some instructions on the left-hand side of the screen and a software product on the right hand side in which they had to perform five tasks as per the instructions (see Figure 1 for layout – for confidentiality reasons, company-specific information has been removed).

The tasks involved setting up an automatic cleaning schedule, setting parental controls, creating a vault, shredding files and deleting a vault. Participants were instructed not to reposition any of the windows relating to the software product or the instructions, so as to facilitate eye-tracking analysis. Once they had completed their tasks they responded the questionnaire.

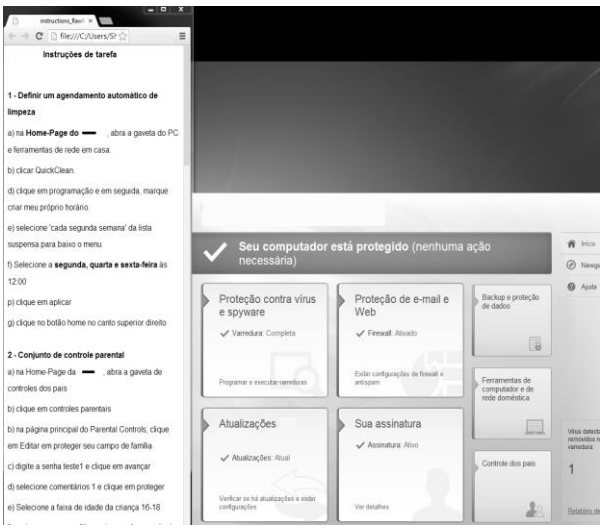

Figure 1- Set up screenshot

[2] Our intention had been to use the company-specific MT engine trained using the Microsoft Translator Hub. However, at the time of the experiment, technical difficulties prevented this and the company suggested the use of the generic Bing engine as an alternative.

[3] Ethics approval was granted by the relevant university research ethics committee.

4 Results

We first present the results from the eye tracking data, which, as discussed above, we treat as measures of efficiency. For all results "Baseline" refers to the baseline reading task of the Wikipedia text, "Instructions" refers to the eye tracking data for the area of the screen in which the instructions were displayed (the AOI, or Area of Interest) and "Interface" refers to the area on the screen in which the product itself was displayed and where users had to carry out the tasks required. For the eye tracking data, "MT" refers to the raw MT instructions and PE refers to the post-edited version. We first present cognitive indicators of efficiency (fixation measures: 4.1-4.5), then goal completion as a measure of effectiveness (4.6), followed by goal completion as a factor of time (also a measure of efficiency – 4.7) and finally satisfaction measures (4.8).

4.1 Mean Total Time in Fixation

The mean total time in fixation is the time spent in fixations combined for each group within an AOI (in seconds).

Figure 2 shows the mean across both groups for the baseline, MT and PE texts. Data for the baseline text is much shorter, as would be expected, because this was just one short text that had to be read and there was no other task associated with it. The mean total fixation time is higher for the MT group for both the Instructions AOI and the Interface AOI.

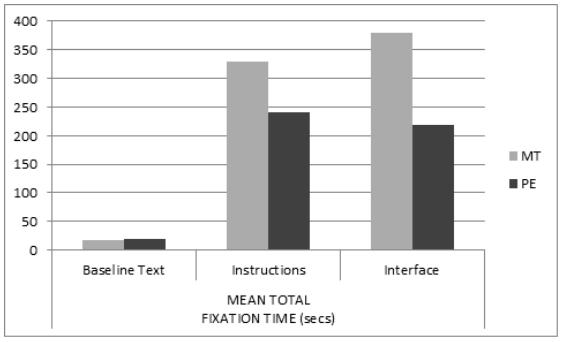

Figure 2- Mean Total Time in Fixation

An independent-samples t-test was conducted to compare both conditions. There was a significant difference in the scores for total fixation time for the Instructions AOI: *t-value* (14) = 2.83, *p-value* = .013 and for the Interface AOI: *t* (14) = 4.58, *p* = .001. There was no significant difference between groups for the Baseline (*p* = .65), which indicates that there was no difference in the baseline reading activity between the two groups. (All significance levels at p > 0.05.)

4.2 Mean Fixation Duration

Mean FD (in seconds) is the average length of fixations for all participants in both groups (Figure 3).

For both groups, the mean value is 0.33 for the baseline, again indicating that there was no difference across both groups for the baseline task. For the Instructions AOI, the mean fixation duration for the PE group is (0.45) and for the MT group (0.43). Both are greater than the baseline, suggesting that reading of the MT output (either in raw or post-edited form) required greater effort than reading the baseline text. Although the value for the MT text is slightly higher than that of the PE text (0.45 vs. 0.43), these are not statistically different. This is also the case for the Interface AOI.

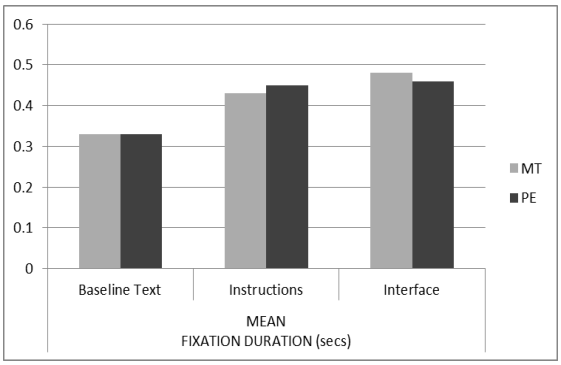

Figure 3- Mean Fixation Duration

4.3 Fixation Count

Fixation count (FC) is the total number of fixations within an AOI. The more there are, the higher the cognitive effort is deemed to be. As can be seen (Figure 4), the total FC is higher for the MT group for both the Instructions and Interface AOIs. Table 1 also shows the mean, median and standard deviations values for the Fixation Count measure. (Note: We do not report data for the baseline reading task here as comparisons of fixation count would be meaningless, given that the task and text differ substantially from the task and text used in the actual experiment. Comparisons for mean total fixation time (Fig. 2), on the other hand, are meaningful as they demonstrate that the groups did not differ radically in their baseline reading activity.)

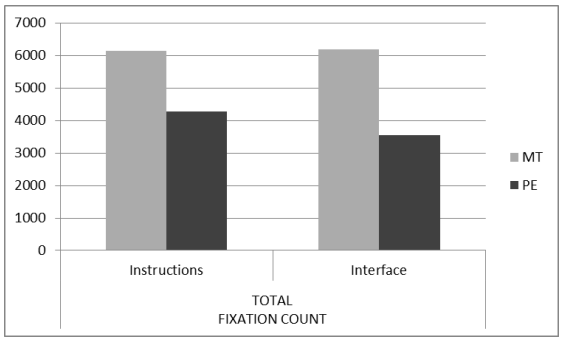

Figure 4- Total Fixation Count

A significant difference was found for Fixation Count on the AOI Instructions: t (14) = 4.43, p = .001 as well as for the Fixation Count on the AOI Interface, t (14) = 4.69, p <.001.

	Instructions			Interface		
	Median	Mean	St. Dev.	Median	Mean	St. Dev.
MT	808.50	767.00	88.93	731	773.88	158.99
PE	535.00	534.25	118.97	454.5	443.5	119.72

Table 1 – Fixation Count & St. Dev

4.4 Visit Duration

Visit duration (VD) is the total time (in seconds) spent looking at an AOI, starting with a fixation within the AOI and ending with a fixation outside this AOI, that is, saccades (or rapid eye movements between fixations) are also counted. Table 2 presents the values for the baseline, instructions and interface for both groups.

		Total	Median	Mean	St. Dev.
Baseline	MT	234.76	27.91	27.3	8.20
	PE	252.61	32.17	31.58	7.44
Instructions	MT	3085.00	392.01	2.91	50.69
	PE	2266.30	292.19	2.86	74.49
Interface	MT	3654.48	429.08	2.62	104.43
	PE	2098.77	276.4	2.13	70.59

Table 2 - Visit Duration (secs)

As Table 2 demonstrates, the mean VD is higher for the machine translation group for both the Instructions and Interface. A t-test found a significant difference between both conditions, where t(14) = 3.212, p = .006 for the AOI Instructions and t(14) = 4.363, p = .001 for the AOI Interface. For the baseline, t(14) = -.578, p = .578, again suggesting that there is no statistically significant difference between the two groups in the baseline task and so the effects we see between the two conditions MT and PE are likely to have been produced by the texts themselves and not by variances in the groups.

4.5 Visit Count

Visit Count is the number of visits (using eye movements as evidence) to an AOI. Table 3 shows the number for VC for both MT and PE groups:

		Total	Median	Mean	St. Dev.
Instructions	MT	1093	128	136	25.24
	PE	799	99.5	99.8	20.3
Interface	MT	1205	151	150.63	12.54
	PE	907	113.5	113.38	22.77

Table 3 - Visit Count

Note that the baseline is not shown here as the number of visits in a static text presented for reading would always be 1. The total VC is higher for the MT group for both AOIs. A t-test found a significant difference between both conditions, where t(14)= 3.209, p = .006 for the AOI Instructions and t(14)= 4.052, p = .001 for the AOI Interface.

4.6 Goal Completion - Effectiveness

All participants in the PE group were able to complete all the tasks, with the exception that one participant in the MT group skipped task 1 (Set an Automatic Cleaning Schedule). This demonstrates that, regardless of the type of instructions, participants were still able to complete their tasks. At the same time, it is worth pointing out some confusion among those who read the raw MT instructions: For Task 2 (Set Parental Controls) one of the options to be blocked by the participants had a different translation from the interface. As a result, some participants were not able to select that option and skipped it, but the task as a whole was completed. Also, Tasks 3 and 5 for the MT group resulted in participants erasing and moving incorrect files but, in the end, the task of creating and deleting the vault was completed. Table 4 gives the total task times for both groups.

	Total Time	Median	Mean	St. Dev.
MT	6885	828.6	860.63	139.99
PE	4540.2	582.3	567.52	138.43

Table 4 – Total Task Time (secs)

An independent-samples t-test was conducted to compare both conditions. There was a significant difference between the MT and PE groups; *t-value* (14) = 4.21, *p-value* = .001.

4.7 Efficiency

Efficiency is also measured as the number of successful tasks completed divided by the total task time (Table 5). The PE group were found to be more efficient (t (14) = 3.75, p = .002).

	Median	Mean	St. Dev.
MT	165.72	178.58	40.42
PE	116.46	113.5	27.68

Table 5 – Efficiency Scores (secs)

4.8 Satisfaction

As mentioned in Section 3, the participants responded to a post-task questionnaire that measured their level of satisfaction with the instructions through a range of questions. None of the participants knew that the instructions had been machine translated.

As a reminder, the statements they had to respond to were as follows:

1. The instructions were usable.
2. The instructions were comprehensible.
3. The instructions allowed me to complete all of the necessary tasks.
4. I was satisfied with the instructions provided.[4]
5. The instructions could be improved upon.
6. I would be able to use the software again in the future without re-reading the instructions.
7. I would recommend the software to a friend or a colleague.
8. I would consider buying this product after participating in this experiment.

Table 6 presents the results for each statement and each group. For all statements, except number 5, the higher score (5) indicates higher satisfaction (the opposite is true for statement 5). As can be seen, levels of satisfaction are generally higher for the post-edited instructions. Exceptions include statements 2, 6 and 5. In the case of 5, the lower score means higher satisfaction for the post-edited text. The considerable difference in scores for statements 7 and 8 are worth noting due to the potential commercial implications. Those who read the post-edited text would seem more inclined to recommend or purchase the product.

5 Conclusions and Future Work

We set about measuring and comparing the usability of instructions for a software product that had been machine translated and machine translated and lightly post-edited.

		S1	S2	S3	S4	S5	S6	S7	S8
MT	1	0	0	12.50%	12.50%	0	12.50%	12.50%	25.00%
	2	12.50%	37.50%	12.50%	62.50%	0	50.00%	25.00%	12.50%
	3	0	0	0	0	0	0	37.50%	37.50%
	4	75.00%	62.50%	37.50%	25.00%	0	12.50%	25.00%	25.00%
	5	12.50%	0	37.50%	0	100%	25.00%	0	0
	Median	**4**	**4**	**4**	**2**	**5**	**2**	**3**	**3**
PE	1	0	0	0	0	0	37.50%	0	0
	2	0	0	0	0	12.50%	25.00%	12.50%	0
	3	0	0	0	0	12.50%	0	12.50%	25.00%
	4	25.00%	62.50%	25.00%	50.00%	25.00%	37.50%	12.50%	37.50%
	5	75.00%	37.50%	75.00%	50.00%	50.00%	0	62.50%	37.50%
	Median	**5**	**4**	**5**	**4.5**	**4.5**	**2**	**5**	**4**

Table 6 – Post-Task Questionnaire

[4] We made sure the participants understood that by 'instructions' we meant the written task instructions provided to perform the tasks, not the verbal instructions given by the researcher on how the experiment would be carried out.

Our objective was to see whether the post-edited version was more usable than the raw MT output. The natural hypothesis is to assume that post-editing improves the quality and usability of a text, but this is usually measured using quality evaluation and not via end user eye tracking-based measurements. The empirical investigation we have carried out here is a validation of this hypothesis.Using the ISO/TR 16902 definition of usability, we undertook a suite of measurements to assess different parts of this definition. Measures of effectiveness included the cognitive measurements of mean total fixation time, mean fixation duration, fixation count, visit duration, and visit count. For all of these measures except mean fixation duration a statistically significant difference was found between the MT and PE groups implying that those who read the PE instructions were more effective and that therefore those instructions had a higher level of usability.

The measurement of goal achievement demonstrated that regardless of the type of instructions, both groups were successful in achieving their goals. We put this down to the use of human intelligence and experience in making sense of content that is not optimal. Moreover, a higher level of confusion was evident among the MT group, as discussed above.

Additional measures of effectiveness and efficiency also demonstrated that the PE instructions were more usable. Finally, the responses to a post-task questionnaire on satisfaction indicated a higher level of satisfaction among those who used the post-edited instructions. Noteworthy in particular are the responses regarding recommendation to a friend or the purchase of the product; for both statements those who read the post-edited instructions were more likely to do so, which has important implications for commercial users of MT.

We have shown that post-editing – even to the level of 'fit-for-purpose' – adds value to machine translated content because it increases usability and satisfaction levels. While this is perhaps an unsurprising result, the important aspect of this study is the number of measures of usability and the inclusion of end users actually performing tasks with the instructions and a software product. This lends a higher level of credibility to the claim of increased usability.

Obviously the sample size is small and we have included only one language pair so future work could build on the number of participants and language pairs. Another focus in the future will be comparisons between human translation and raw and post-edited MT as well as a focus on different kinds of content.

References

De Almeida, Gisele and Sharon O'Brien. 2010. Analysing Post-Editing Performance: Correlations with Years of Translation Experience. *Proceedings of the 14th Annual Conference of the EAMT*. St. Raphael, May 27-28.

Depraetere, Ilse. 2010. What Counts as Useful Advice in a University Postediting Training Context? Report on a case study. *Proceedings of the 14th Annual EAMT Conference*. St. Raphael, May 27-28.

Doherty, Stephen and Sharon O'Brien. 2013. Assessing the Usability of Raw Machine Translation Output: A User-Centered Study using Eye Tracking. *International Journal of Human-Computer Interaction*, 30: 40-51.

Doherty, Stephen, Sharon O'Brien and Michael Carl. 2010. Eye Tracking as an MT Evaluation Technique. *Machine Translation*, 24(1): 1-13.

Fiederer, Rebecca, and Sharon O'Brien. 2009. Quality and Machine Translation: A Realistic Objective? *Journal of Specialised Translation [online]*.

Guerberof, Ana. Forthcoming. The Role of Professional Experience in Post-Editing from a Quality and Productivity Perspective. In O'Brien, Sharon, Laura, Winther-Balling, Michael Carl, Michel Simard and Lucia Specia (eds) *Post-Editing of Machine Translation: Processes and Applications.* Cambridge Scholars Publishing, 51-76.

International Organization for Standardization. 2002. *ISO/TR 16982: Ergonomics of human-system interaction – Usability methods supporting human centered design*. Available from: http://www.iso.org/iso/catalogue_detail?csnumber=31176.

Jones, Douglas, Wade Shen, Neil Granoien, Martha Herzog and Clifford Weinstein. 2005. Measuring Translation Quality by Testing English Speakers with a New Defense Language Proficiency Test for Arabic. *Proceeding of the International Conference on Intelligence Analysis*. McLean, VA.

Kaakinen, Johann, Jukka Hyönä and Janice Keenan. 2003. How prior knowledge, WMC, and relevance of information affect eye fixations in expository text. *Journal of Experimental Psychology: Learning, Memory and Cognition*, 29(3): 447-457.

Koponen, Maarit. 2012. Comparing Human Perceptions of Post-editing Effort with Post-editing Operations. *Proceedings of the 7th Workshop on Statistical Machine Translation*. Association for Computational Linguistics, Montreal, June 7-8.

Moravcsik, Julia and Walter Kintsch. 1995. Writing quality, reading skills, and domain knowledge as factors in text comprehension. In John Henderson, M. Singer, and F. Ferreira (eds.) *Reading and Language Processing*. New York, London: Psychology Press, 232-246.

O'Brien, Sharon, Michel Simard and Lucia Specia (Eds.) 2012. Workshop on Post-editing Technology and Practice (WPTP 2012). *Conference of the Association for Machine Translation in the Americas (AMTA 2012).* San Diego, October 28.

O'Brien, Sharon, Michel Simard and Lucia Specia (Eds.) 2013. Workshop on Post-editing Technology and Practice (WPTP 2013). *Machine Translation Summit XIV*. Nice, September 2-6.

Plitt, Mirko, and Françoise Masselot. 2010. A Productivity Test of Statistical Machine Translation Post-editing in a Typical Localisation Context. *The Prague Bulletin of Mathematical Linguistics*. Prague: 7-16.

Radach, Ralph, Alan Kennedy and Keith Rayner. 2004. *Eye Movements and Information Processing during Reading*. Hove: Psychology Press.

Rayner, Keith. 1998. Eye Movements in Reading and Information Processing: 20 Years of Research. *Psychological Bulletin*, 124:372-422.

Rayner, Keith and Sarah Sereno. 1994. Eye Movements in Reading: Psycholinguistic Studies. In Gernsbacher M.A. (ed.), *Handbook of Psycholinguistics*. New York: Academic Press, 57-81.

Sousa, Sheila C.M., Wilker Aziz, and Lucia Specia. 2011. Assessing the post-editing effort for automatic and semiautomatic translations of DVD subtitles. *Proceedings of the Recent Advances in Natural Language Processing Conference*. Hissar, Bulgaria.

Specia, Lucia. 2011. Exploiting Objective Annotations for Measuring Translation Post-editing Effort. *Proceedings of the 15th Annual EAMT Conference*. Leuven, May 30-31.

Stymne, Sara, Henrik Danielsson, Sofia Bremin, Hongzhan Hu, Johanna Karlsson, Anna Prytz Lillkull and Martin Wester. 2012. Eye Tracking as a Tool for Machine Translation Error Analysis. *Proceedings of the Language Resources and Evaluation Conference*. 1121-1126. Istanbul. May 21-27.

Translation Automation User Society (TAUS). Online. *Machine Translation Post-Editing Guidelines*. Available at: https://evaluation.taus.net/resources/guidelines/post-editing/machine-translation-post-editing-guidelines