View metadata, citation and similar papers at core.ac.uk
brought to you by CORE
provided by DCU Online Research Access Service

Eye-Tracking as a Measure of Cognitive Effort for Post-Editing of Machine Translation

Abstract:

The three measurements for post-editing effort as proposed by Krings (2001) have been adopted by many researchers in subsequent studies and publications. These measurements comprise temporal effort (the speed or productivity rate of post-editing, often measured in words per second or per minute at the segment level), technical effort (the number of actual edits performed by the post-editor, sometimes approximated using the Translation Edit Rate metric (Snover et al. 2006), again usually at the segment level), and cognitive effort. Cognitive effort has been measured using Think-Aloud Protocols, pause measurement, and, increasingly, eye-tracking. This chapter provides a review of studies of post-editing effort using eye-tracking, noting the influence of publications by Danks et al. (1997), and O'Brien (2006, 2008), before describing a single study in detail.

The detailed study examines whether predicted effort indicators affect post-editing effort and results were previously published as Moorkens et al. (2015). Most of the eye-tracking data analysed were unused in the previous publication, and the small amount presented was not described in detail due to space constraints. This chapter focuses instead on methodology and the logistics of running an eye-tracking study recording over 70 sessions. We present results in which average fixation count per segment correlates strongly with temporal effort.

Introduction

The increase in machine translation (MT) quality for many language pairs since Statistical Machine Translation (SMT) became the dominant MT paradigm has resulted in an associated increase in the use of MT for industry production of translated texts. Although initial reports of the usefulness of post-editing (PE) were highly negative (Beyer 1965, ALPAC 1966), by the 1980s there was sporadic use of post-edited Rule-Based MT for production (Hutchins 1992; Vasconcellos and León 1985). The shift to SMT (and the associated improvement in quality) has subsequently made MTPE an appealing method of translating large volumes of text at reduced cost in localisation workflows (DePalma and Lommel 2016).

Initial PE research focussed on temporal effort and/or technical effort. Temporal effort may be defined as the speed or productivity rate of post-editing, often measured in words per second or per minute at the segment level. Technical effort is the number of actual edits performed by the post-editor, either measured using keylogging software or approximated using the hTER metric, developed by Snover et al. (2006), which calculates the fewest possible edits required from a pre- to post-edited segment. These measures of PE effort were often presented in comparison with translation from scratch or with the aid of translation memories (for example, Bruckner and Plitt 2001) or for MT system evaluation (Su et al. 1992). Krings (2001: 179) introduced the measurement of cognitive effort for post-editing, and he used Think-Aloud Protocol (TAP) to discover the "type and extent of cognitive processes" required to "remedy a given deficiency" in MT. Cognitive effort had been measured in Translation Process Research (TPR) since the early 1980s, but not previously in PE research, where the addition of raw MT output to the source texts or segments may be associated with additional cognitive load. Shreve and Diamond (1997: 243) highlighted the "reduction in efficiency" associated with TAP, which

is problematic when measuring cognitive and temporal effort concurrently using that method. In his study, Krings (2001) found that processing speed without TAP was roughly 30% faster, and considers that TAP can only possibly report conscious processes without explaining automatic processes. Nunes Vieira (2015), however, suggests that TAP is still useful for detailed relative measurements of cognitive effort within a dataset, and found that coded TAP ratings correlated strongly with other measures of cognitive effort in his study.[1]

Some alternative methods of measuring cognitive effort, such as keyboard logging (Jakobsen 1999; O'Brien 2005), pause measurements (O'Brien 2006; LaCruz et al. 2014), and more recently fMRI (Functional Magnetic Resource Imaging; Chang 2009), and EEG (Electroencephalography; Hansen-Schirra 2017) are sometimes used, either alone, or in combination with other methods (Dragsted 2010; Hvelplund 2011). However, eye-tracking has become a particularly popular method for measuring cognitive effort in translation studies and is used often for measuring post-editing effort, due to influential studies such as O'Brien's pilot study of fuzzy match editing and post-editing effort in 2006. Translation researchers, especially those with a psychological background, quickly saw the potential of eye-tracking as an non-intrusive and objective research tool, and adopted the process due to the possibility of collecting empirical cognitive data with relatively mature software packages available for its analysis, and for its relative affordability when compared with options such as EEG and fMRI.

This chapter introduces the task of post-editing and presents a review of post-editing research using eye-tracking, before looking in detail at the methodology and previously unpublished results of a single study that examines whether predicted effort indicators effect post-editing effort.

[1] Editors' note: readers can find observations on cognitive effort, including an early definition, in relation to eye-tracking research in the Introduction.

The task of post-editing

Post-editing is a task that "entails correction of a pre-translated text rather than translation 'from scratch'" (Wagner 1985: 1), with the task of the post-editor defined by Allen as to "edit, modify, and/or correct pre-translated text that has been translated by an MT system from a source language into (a) target language(s)" (2003: 297). Many translators dislike revision or editing tasks (Mossop 2007), but the task of PE differs from revision of human translations in that the types of errors that the post-editor is required to correct often "contain errors which no human, even a small child or a non-native speaker, would ever make", errors that post-editors may find "irritating and 'stupid'" (Wagner 1985: 2). PE is usually introduced in order to increase productivity in response to growing demands and to cut costs (Senez 1998), but has grown more popular in recent years due to incrementally increasing MT quality, ever-faster production cycles, and growing amounts of texts to translate amidst economic constraints (Moorkens 2017).

Initial industrial deployments of PE were often for assimilation or gisting purposes (such as Senez 1998 and 'rapid' or 'light' PE in Wagner 1985), but PE was also used for publication, in which case 'fully' post-edited text could be "indistinguishable from human translation" (Wagner 1985: 4). More recently, some companies have offered light, medium, and full PE, gradations that are difficult to precisely define, may be interpreted differently by the post-editor, and that make reliable and generalizable measurement of task effort problematic. New uses are continually being found for (to a greater or lesser extent) post-edited or even raw MT for publication or dissemination, based on two concepts as introduced by Way (2013): fitness for purpose (when the quality is 'good enough' or 'acceptable') and perishability of content to be translated. Way suggests that the use of MT should be in line with the "purpose, value and

shelf-life" of the text (2013: 2). Continuing incremental increases in MT quality should result in reduced PE effort, as noted by Wilms (1981). These increases in MT quality, added to economic pressures, have resulted in more pragmatic interpretations of acceptable quality, bringing new use cases for raw and PEMT (Schmidtke 2016). This trend is likely to continue, based on initial PE evaluations of neural MT (Bentivogli et al. 2016, Castilho et al. 2017). DePalma and Lommel (2016) report that over 80% of Language Service Providers surveyed in 2016 now offer a PEMT service. This means that more translators are being asked to post-edit MT, a task that they tend not to be fond of.

Wagner noted in 1985 (2) that "working by correction rather than creation" comes as a shock to translators, and there is still widespread user dissatisfaction reported in PE studies. Complaints include finding a limited opportunity to create quality, the perception of MT as a threat to the profession of translation, and the perception that MTPE is slower than translating from scratch. Studies of temporal PE effort have been particularly useful for testing the latter perception, finding that all (Plitt and Masselot 2010, Läubli et al. 2013) or some (Garcia 2011, Gaspari et al. 2014) participants studied were faster when post-editing than translating from scratch. Despite repeated findings of lower temporal effort when post-editing, many translators still prefer to translate from scratch, ignoring the potential productivity gains (Teixeira 2014). This contradictory but wide-spread preference for translation from scratch suggests that there may be a usability problem with the method of deployment of MT via PE, and/or that there may be increased cognitive effort associated with the addition of MT output to the source and target segments that the translator usually works with.

Krings (2001) wrote that "the availability of a machine translation often does not lead to the expected reduction in cognitive effort during post-editing" (320). In fact, he found cognitive

effort for PE generally to be higher than for translation from scratch, independent of varying MT quality, and only reported decreased cognitive effort for PE tasks performed without access to the source text. This was despite his finding that most cognitive processing effort is required for target text production in PE and physical writing. As mentioned previously, TAP was found to be an inefficient method of measuring cognitive effort in Krings' study. Since its introduction as a measure of cognitive effort in TPR, eye-tracking has been adopted as a more efficient way to measure cognitive effort for the task of post-editing, levels of cognitive effort associated with repairing different error types from the MT output, and for testing features and functionality that may mitigate that cognitive effort required for PE in order to make the task more acceptable to translators. Some of these studies are reported in the following section.

Eye-tracking measures used in studies of post-editing

Most eye-tracking studies of post-editing have measured fixations, although a smaller number have reported measurements for pupil dilation, and in one instance, saccades. O'Brien (2006) used pupil dilation as a measure of cognitive effort, but in a 2008 study found it inappropriate for translation tasks and instead focussed on fixation count and fixation duration based on Just and Carpenter's (1980: 330) theory that "the time it takes to process a newly fixated word is directly indicated by the gaze duration". Saldanha and O'Brien (2014) cited difficulties in controlling variables when measuring pupil dilation (see also Caffrey 2009), a factor which may threaten ecological validity, and noted the additional problem of allowing for latency or delays in changes to pupil size. Lacruz and Shreve (2014) suggested that it may be useful to triangulate pupil dilation data with keystroke logs, but that this may be so labour intensive as to be unfeasible. At the time of writing, there has been little focus on saccade measurements in post-editing studies. Gonçalves (2016) carried out a pilot study to assess whether saccade

direction and distance correlate with fixation measures of cognitive effort in reading, translating, and post-editing tasks. His findings were inconclusive, partially due to the frequency limit of 60Hz for the eye-tracker used for this research (see Duchowski 2003), but a follow-up study will employ a 300Hz eye-tracker. Many eye-tracking studies in TPR and post-editing have designated AOIs in source and target text sections of the user interface. In this way, Carl et al. (2011) compared source and target text editing behaviours among translators and post-editors, and found that both fixation count and total gaze time (per AOI) when post-editing appears to be heavily focussed on the target text, concurring with Krings' findings as reported in Section 2.

Several studies have used eye-tracking to measure cognitive effort when post-editing. O'Brien (2011) asked seven participants to post-edit 60 segments of English-French SMT output - 20 segments in three categories of GTM (General Text Matcher, Turian et al. 2003) score - within the Alchemy Catalyst editing environment. She found that average fixation duration per word and average fixation count per word correlated strongly with the GTM categories, suggesting that the GTM metric may be a useful predictor of cognitive PE effort. For the eye-tracking portion of his study, Nunes Vieira (2015) asked 19 participants to post-edit two texts (of roughly 400-word length) from a news article corpus[2] that had been translated from French to English using SMT. He found strong correlations between cognitive effort and METEOR metric (Denkowski and Lavie 2014) scores below 0.6; these findings led him to suggest that source text features, such as frequency of prepositional or verb phrases and type-token ratio, may be good predictors of cognitive PE effort. He also suggests that the mixed-methods

[2] These texts were extracted from the newstest2013 corpus, extracted articles from various online publications used at the WMT Shared Task events and available from http://www.statmt.org/wmt14/translation-task.html.

approach employed for this study enriches the findings by adding details of the quality or intensity of cognitive effort expended along with the amount.

Koglin (2015) had 14 translation students post-edit two texts about the Tea Party movement in the USA that had been translated from English to Portuguese using both Systran and Google Translate MT systems within the Translog-II environment, and found PE to require less cognitive effort than translating the texts from scratch. Carl, Gutermuth, and Hansen-Schirra (2015) compared translation from scratch with two Google MT post-editing tasks in which 24 translators translated six English texts into German using the Casmacat interface. They found that, despite a stated post-task preference for translation from scratch, and a lack of experience of PE, all participants were more efficient in terms of temporal, technical, and cognitive effort when post-editing. They also found that source text complexity had more of an impact on processing effort when translating from scratch than when post-editing.

Finally, many studies have used eye-tracking to test novel functionality or new ways to categorise user behaviour when post-editing. Alves et al. (2016), for example, used the Casmacat interface to carry out an A/B test, asking participants to post-edit with and without interactive machine translation (IMT) functionality, to investigate the impact of IMT on PE behaviour. When IMT is active, the MT suggestion it updated in real time based on the user's edits. The authors' hypothesised that technical and temporal effort would be less in the interactive PE mode, but made not predictions about cognitive effort. In fact, neither technical nor temporal effort was decreased as expected, but mean fixation duration was lower than with regular PE. Although fixation count was higher with IMT, the authors concluded that this was a promising study for improved PE usability due to the drop in mean fixation duration, notwithstanding the small sample size (10) and the single language pair tested (EN to PT-BR).

Läubli and Germann (2016) comment that, despite eye-tracking and key-logging becoming commonplace for TPR, data analysis is still "tedious and difficult" (160; see also the following section), and thus difficult to perform manually. They created a statistical model for annotating PE based on the number of keystrokes, mouse clicks, and eye fixations in a segment. In comparison with a gold standard sample annotation of 7 PE sessions, ten experienced annotators were more accurate than the statistical model, and two were less accurate. This is a promising result for automatic annotation, but suggests that, for now, data analysis for eye-tracking TPR data will remain a labour-intensive activity.

Nitzke and Oster (2016) introduced a novel annotation schema for PE, and compared general and domain-specific translation and PE data using this schema. The data are in the English to German language pair and are drawn from the large TPR database collected by the CRITT (Centre for Research and Innovation in Translation and Translation Technology) at the Copenhagen Business School. They subdivide the orientation phase, when a post-editor is fixating on the text on-screen before beginning translation or editing, depending on whether the focus is on the source or target text. The revision phase is annotated based on whether there is a single round of post-editing or the user jumps back through the text to make changes. For translation from scratch there is an additional drafting stage. Perhaps predictably, the authors found gaze behaviour to be similar for target texts when post-editing or translating, but that for PE the source text receives far less attention. This tendency was particularly notable for domain-specific texts. The authors suggest that application of the review categories could reveal differences between PE behaviours for texts from different domains.

A study by Moorkens et al. (2015) investigated whether human estimates of PE effort were accurate predictors of actual PE effort, and whether post-editor behaviour was different when PE effort estimation indicators (based on real user ratings) were displayed to participants. A moderate correlation was found between measurements of PE effort and mean user ratings (six participants rated the segments that has been machine translated from English to Portuguese), which lead to a conclusion that "human ratings of PE effort do not correlate strongly with the actual time required during post-editing" (Moorkens et al. 2015: 281). The moderate correlation meant that, as participants moved through the texts to be post-edited, there was some relationship between the three-category, 'traffic light' indicator colouring scheme, and the final measurements of temporal and technical effort, but user behaviour did not appear to change. In the following section, we provide some more detail about that study, and analyse some further eye-tracking data that may add further detail to the conclusions as originally published.

Post-editing study

This section describes a study of PE effort, measuring temporal, technical, and cognitive effort (using eye-tracking) that was carried out as part of a larger study in ADAPT Centre in Dublin in collaboration with Sharon O'Brien. The research questions for this study were:

1. Are human estimates of PE effort accurate predictors of actual post-editing effort?
2. Does the display of PE effort estimation indicators to post-editors influence post-editing behaviour?

Methodology

The study employed a test interface that, after some further development, became HandyCat (Hokamp and Liu 2015). This is a web-based, horizontally aligned translation editing tool, hosted on a server that saves User Activity Data (UAD) for analysis, including timings for editing actions, and pre- and post-edited texts. It enables researchers to note a session ID to attribute anonymously to an eye-tracking session, and features, such as the PE effort indicators added in this study, may be toggled on and off. Two Wikipedia source texts (about Paraguay and Bolivia) were chosen and machine translated into Portuguese using Microsoft Bing Translator, at that time an SMT system. Familiar topics were chosen so that participants would be unlikely to require consultation with external resources, as this would be problematic when using an Internet browser within a Tobii Studio environment. Post-editing was carried out at the Laboratory for Experimentation in Translation (LETRA) at the Federal University of Minas Gerais (UFMG) in Brazil, while using a Tobii T60 eye tracker.

The research was carried out in three stages. In the first stage, six members of staff at UFMG who have translation and PE experience each rated the machine translation quality of our two test sets containing 40 segments according to the following categorisation:

- Red: Requires complete retranslation
- Amber: Requires some editing, but PE still quicker than retranslation
- Green: Little or no PE needed

The second stage of the research began after a break of at least two weeks. Four of the same participants (two dropped out) were asked to post-edit the texts, to see whether their actual PE effort matched their predicted effort. Participants were introduced to the PEARL interface, requested not to answer phones (inevitably, two participants did) or leave the eye-tracking

room, not to leave the PEARL webpage, and provided with the following PE guidelines based on O'Brien (2010):

- The message transferred should be accurate
- Grammar should be accurate
- Ignore stylistic and textuality problems
- Ensure that key terminology is correctly translated
- Edit any offensive, inappropriate or culturally unacceptable information
- All basic rules regarding spelling, punctuation, and hyphenation still apply
- Quality expectations: medium

In the third stage, 33 undergraduate and Master's translation students with little PE experience were asked to post-edit the two texts as in Stage 2, however this time one of the tasks was carried out with colour-coded Post-Editing Effort Estimation Indicators (PEEIs) displayed for each segment based on the ratings from Stage 1. The order of the tasks was randomised, with eight participants each following one of the four conditions as shown in Table 1.

	Condition 1	**Condition 2**	**Condition 3**	**Condition 4**
Test & Feature Set	Paraguay/No PEEI	Paraguay/PEEI	Bolivia/PEEI	Bolivia/No PEEI
Test & Feature Set	Bolivia/PEEI	Bolivia/No PEEI	Paraguay/No PEEI	Paraguay/PEEI

Table 1. Randomly ordered tasks in Stage 3; PEEI= post-editing effort indicator

Each participant was scheduled a one-hour slot to complete the tasks between 9am and 9pm during an 8-day period. One participant only completed one of the PE tasks during her slot, another took far longer than other participants, so her data was discounted as an outlier for temporal effort, and 22 task recordings did not log properly, meaning that data for temporal and technical effort was lost for several participants.

Temporal effort was calculated from the first edit to the 'segment-finished' tag for the first segment, then between 'segment-finished' tags for each subsequent segment. Technical effort was estimated using the hTER metric (Translation Error Rate with human targeted references; Snover et al. 2006), which calculates the minimum number of edits to get from MT output to the post-edited segment. Cognitive effort was measured using the eye-tracking software package Tobii Studio (v.3.1 for Stage 2, v.3.4 for Stage 3) to calculate fixation count and total fixation duration for each segment within the areas of interest for source and target text areas of the screen. Each recording was manually marked when editing had been completed for each text segment. Tobii Studio segments of equavalent length to text segment editing time were created from these marks, and these Tobii segments were in turn added to a Segment Group, numbered from 1 to 40 (again to match the translation segments). The step of grouping recordings by all participants for each text segment allowed the statistics for each segment to be calculated within the Statistics view in Tobii Studio and exported.

Results: Stage 1

The results of the first and second stage will only be summarised here, as they may be read in detail in Moorkens et al. (2015). O'Brien (2011: 201) has commented on the subjectivity of human ratings, how they may be "influenced by the previous rating, and fatigue or boredom may influence the motivation of raters". For this reason, inter-rater reliability is often low. In this study, the correlation between predicted PE effort as judged by each participant and the mean rating of all participants is weak (r_s=0.373, $p<0.001$). Participant assessments were 100% equivalent for only 13 of the 80 segments presented in the test data. Nonetheless, a mean rating between 0 and 1 for each MT segment was calculated, and this was the basis for the colour-coding appended to each segment, and made visible to participants in Stage 3. A segment with

a mean score of ≤0.3 was marked 'green', suggesting that the segment would require little editing. A segment with a mean score of ≥0.7 was marked as 'red', suggesting that the segment would require heavy editing. The remaining segments were marked 'amber' (as previously: requires some editing, but PE still quicker than retranslation).

Results: Stage 2

Mean ratings from Stage 1 were found to only correlate moderately with the two eye-tracking measurements, despite the fact that both stages involved members of the same participant group. However, the ratings were found to have a strong correlation with technical effort, as may be seen in Table 2.

		Fixation Count	**Fixation Duration**	**Mean Rating**	**Mean Temporal Effort**
Total Fixation Duration	Correlation (r_s)	0.366	-	0.505	0.431
Mean PE Edit Rating	Correlation (r_s)	0.411	0.505	-	0.492
Mean Temporal Effort	Correlation (r_s)	**0.942**	0.431	0.492	-
Mean Technical Effort	Correlation (r_s)	0.432	**0.759**	**0.652**	0.524

Table 2. Spearman correlations (all p<0.001) between mean ratings (Stage 1) and measures of actual effort in Stage 2

Interestingly, the strongest correlations were between (the means calculated for) fixation count and temporal effort (very strong, where r_s=0.942, p<0.001), and between (the means calculated for) total fixation duration and technical effort (r_s=0.759, p<0.001). Mean temporal, technical, and cognitive effort appeared to fit roughly with categorisation, i.e. higher for red (poorly rated) segments, lower for green (positively rated) segments.

Results: Stage 3

The student participants in Stage 3 of this study took, on average, 9% longer to complete the tasks when compared with Stage 2 participants with professional experience. While this may be expected, the difference was not as pronounced as in Moorkens and O'Brien (2015), which used the same PEARL interface. For Stage 3, data from the ten participants with the highest gaze sample rate (of over 85%) was chosen for analysis. This time a poor correlation was found between temporal and technical PE effort, and a moderate correlation was found between mean ratings and temporal effort, as may be seen in Table 3.

		Fixation Count	**Fixation Duration**	**Mean Rating**	**Mean Temporal Effort**
Total Fixation Duration	Correlation (r_s)	**0.965**	-	0.356	0.298
Mean PE Edit Rating	Correlation (r_s)	0.319	0.356	-	0.484
Mean Temporal Effort	Correlation (r_s)	**0.669**	**0.639**	0.484	-
Mean Technical Effort	Correlation (r_s)	0.310	0.298	0.236	0.109

Table 3. Spearman correlations (all $p<0.035$) between mean ratings (Stage 1) and measures of actual effort in Stage 3

In Stage 3, the relationship between fixation count and fixation duration was found to be very strong, and strong correlation was found between mean temporal effort and fixation duration. Both Stage 2 and Stage 3 showed a strong relationship between fixation count and mean temporal effort. 79.25% of fixations were measured in the target text AOI, which is consistent with the findings of Carl et al. (2011).

The two research questions posed in this study were answered with the caveat that they are limited by the size and high variability of ratings, which we repeat here. The answer to Question 1 was that human ratings were not a good predictor of PE effort when using the same participants group. On analysing the data from a second participant group, this conclusion is unchanged. To answer Question 2, the PE effort indicators appeared not to change actual PE effort for both participant groups. For one user in Stage 3, there were fewer fixations when the indicators were on, and for some users we noticed that they did not look at the source text at all when the indicator was green. However, on average, there was no real difference. The does not necessarily mean that confidence indicators are not worth persevering with. Any new feature needs to show its usefulness and a fit with users' judgement and workflow in order to gain their trust. A model based on previous post edits, as suggested by Specia (2011), may be more useful here. In addition, if a feature increases usability despite not ameliorating PE effort, that is still worthwhile. A user experience focused or mixed methods study should make this benefit apparent.

Conclusion

The use of eye-tracking for PE studies has become commonplace in recent years, as evidenced by the number of studies reviewed in this chapter. These studies tend to base their measures of PE effort on those identified by Krings (2001), using eye-tracking to measure cognitive effort. Previous studies focused first on pupil dilation, then on fixation duration and fixation count, and current studies using eye-trackers with greater sampling frequency are being used to measure saccadic movement when post-editing. While there have been many PE studies using eye-tracking at this stage, the number of participants is usually small, and each study has a different focus, making results difficult to compare directly. Some elements have become

standard, such as the three-category rating system as used by Krings (2001), Specia et al. (2009, and Moorkens et al. (2015).

The final section of this chapter addresses some difficulties in running eye-tracking studies for PE, such as finding willing participants, scheduling sessions, and retention of data from the eye-tracker and the user interface. The findings from the eye-tracking study detailed are presented with the proviso that they are based on a single language pair, and a small number of participants. Evaluating the usefulness of a feature for PE (or otherwise) using only quantitative data is difficult. More broadly, this is a limitation with many empirical user studies. Nunes Vieira (2015) addressed this by using a mixed methods approach in his PE study. Mixed methods studies based on the pragmatic paradigm may be a worthwhile avenue to pursue to add new insights for future eye-tracking studies of user interaction with machine translation. PE remains a contentious activity for many people involved with translation, and finding a way to make the interaction with MT more acceptable will necessarily involve input from users.

References

Allen, Jeffrey. 2003. "Post-editing". In *Computers and Translation*, ed. by Harold Somers, 297-318. London: John Benjamins.

ALPAC. 1966. *Languages and machines: computers in translation and linguistics. A report by the Automatic Language Processing Advisory Committee.* Washington, DC.

Alves, Fabio, Arlene Koglin, Bartolomé Mesa-Lao, Mercedes García Martínez, Norma B. de Lima Fonseca, Arthur de Melo Sá, José Luiz Gonçalves, Karina Sarto Szpak, Kyoko Sekino, and Marceli Aquino. 2016. "Analysing the Impact of Interactive Machine Translation on Post-

editing Effort". In *New Directions in Empirical Translation Process Research*, ed. by Michael Carl, Srinivas Bangalore, Moritz Schaeffer, 77-94. Heidelberg: Springer.

Bentivogli, Luisa, Arianna Bisazza, Mauro Cettolo, Marcello Federico. 2016. "Neural versus Phrase-Based Machine Translation Quality: a Case Study". In *Proceedings of Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Austin, Texas.

Beyer, Robert T. 1965. "Hurdling the language barrier". *Physics Today* 18(1), 46–52.

Bruckner, Christine, Mirko Plitt. 2001. "Evaluating the Operational Benefit of Using Machine Translation Output as Translation Memory Input". In *Proceedings of MT Summit VIII*, 61–65.

Caffrey, Colm. 2009. *Relevant abuse? Investigating the effects of an abusive subtitling procedure on the perception of TV anime using eye tracker and questionnaire*. PhD diss. Dublin City University.

Carl, Michael, Barbara Dragsted, Jakob Elming, Daniel Hardt, and Arnt Lykke Jakobsen. 2011. "The Process of Post-Editing: a Pilot Study". In *Proceedings of NLPSC 2011*.

Carl, Michael, Silke Gutermuth and Silvia Hansen-Schirra. 2015. Post-editing machine translation: Efficiency, strategies, and revision processes in professional translation settings. Psycholinguistic and Cognitive Inquiries into Translation and Interpreting. In *Psycholinguistic and Cognitive Inquiries into Translation and Interpreting*, ed. by Aline Ferreira and John W. Schwieter, 145-174. London: John Benjamins.

Castilho, Sheila, Joss Moorkens, Federico Gaspari, Iacer Calixto, John Tinsley, Andy Way. 2017. “Is Neural Machine Translation the New State of the Art?” *Prague Bulletin of Mathematical Linguistics* 108.

Chang, Vincent Chieh-Ying. 2009. *Testing Applicability of Eye-tracking and fMRI to Translation and Interpreting Studies: An Investigation into Directionality*. PhD diss. Imperial College, London.

Danks, Joseph H., Gregory M. Shreve, Stephen B. Fountain, Michael K. McBeath. 1997. *Cognitive Processes in Translation and Interpreting*. Thousand Oaks, California: Sage Publications.

Denkowski, Michael, Alon Lavie. 2014 “Meteor Universal: Language Specific Translation Evaluation for Any Target Language”. In *Proceedings of the EACL 2014 Workshop on Statistical Machine Translation*.

Dragsted, Barbara. 2010. “Coordination of Reading and Writing Processes in Translation: An Eye on Uncharted Territory”. In *Translation and Cognition,* ed. by Gregory M. Schreve, Erik M. Angelone, 41-62, Amsterdam: John Benjamins.

Duchowski, Andrew. 2003. *Eye Tracking Methodology: Theory and Practice*. Heidelberg: Springer.

García, Ignacio. 2011. “Translating by post-editing: is it the way forward?” *Machine Translation* 25, 217-237.

Gaspari, Federico, Antonio Toral, Sudip Kumar Naskar, Declan Groves, Andy Way. 2014. “Perception vs Reality: Measuring Machine Translation Post-Editing Productivity”. In *Proceedings of AMTA 2014 Workshop on Post-editing Technology and Practice*, Vancouver, 60-72.

Gonçalves, José Luiz. 2016. “Investigating saccades as an index of cognitive effort in post-editing and translation”. In *Proceedings of EST Congress 2016*, Aarhus, Denmark.

Hansen-Schirra, Silvia. 2017. “EEG and Universal Language Processing in Translation”. In *The Handbook of Translation and Cognition*, ed. by John W. Schwieter and Aline Ferreira, 232-247. Hoboken: John Wiley & Sons.

Hokamp, Chris, Qun Liu. 2015. “HandyCAT”. In *Proceedings of European Association for Machine Translation (EAMT) 2015*, Antalya, 216.

Hutchins, W. John. 1992. “Météo”. In *An Introduction to Machine Translation*, ed. by W. John Hutchins and Harold L. Somers, 207–220. London: Academic Press.

Hvelplund, Kristian T. 2011. *Allocation of cognitive resources in translation: an eye-tracking and key-logging study*. PhD diss. Copenhagen Business School.

Jakobsen, Arnt Lykke. 1999. “Logging target text production with Translog”. *Copenhagen Studies in Language* 24, 9-20.

Koglin, Arlene. 2015. “An empirical investigation of cognitive effort required to post-edit machine translated metaphors compared to the translation of metaphors”. *Translation and Interpreting* 7(1): 126-141.

Krings, Hans P. 2001. *Repairing Texts: Empirical Investigations of Machine Translation Post-editing Processes*, translated by G. S. Koby. Kent, OH: The Kent State University Press.

Lacruz, Isabel, Gregory M. Shreve. 2014. “Pauses and Cognitive Effort in Post-editing”. In *Post-editing of Machine Translation: Processes and Applications*, ed. by Sharon O'Brien, Laura Winther Balling, Michael Carl, Michel Simard and Lucia Specia, 246–272. Newcastle-Upon-Tyne: Cambridge Scholars.

Läubli, Samuel, Mark Fishel, Gary Massey, Maureen Ehrensberger-Dow, Martin Volk. 2013. “Assessing Post-Editing Efficiency in a Realistic Translation Environment”. In *Proceedings of MT Summit XIV Workshop on Post-editing Technology and Practice*, Nice, 83-91.

Läubli, Samuel, Ulrich Germann. 2016. Statistical Modelling and Automatic Tagging of Human Translation Processes. In *New Directions in Empirical Translation Process Research*, ed. by Michael Carl, Srinivas Bangalore, Moritz Schaeffer, 77-94. Heidelberg: Springer.

Lommel, Arle R., Donald A. DePalma. 2016. *Europe’s Leading Role in Machine Translation: How Europe Is Driving the Shift to MT.* Boston: Common Sense Advisory Report.

Moorkens, Joss, Sharon O'Brien, Igor Antonio Lourenco Silva, Norma Fonseca, Fabio Alves. 2015. "Correlations of perceived post-editing effort with measurements of actual effort". *Machine Translation* 29 (3-4), 267-284. doi: 10.1007/s10590-015-9175-2

Moorkens, Joss. 2017. "Under pressure: Translation in times of austerity". *Perspectives: Studies in Translation Theory and Practice*, 25(3), doi: 10.1080/0907676X.2017.1285331

Mossop, Brian. 2007. *Revising and Editing for Translators*. Manchester: St. Jerome.

Nitzke, Jean, Katharina Oster. 2016. Comparing Translation and Post-editing: An Annotation Schema for Activity Units. In *New Directions in Empirical Translation Process Research*, ed. by Michael Carl, Srinivas Bangalore, Moritz Schaeffer, 77-94. Heidelberg: Springer.

Nunes Vieira, Lucas. 2015. *Cognitive Effort in Post-Editing of Machine Translation: Evidence from Eye Movements, Subjective Ratings, and Think-Aloud Protocols*. PhD diss. Newcastle University.

O'Brien, Sharon. 2005. "Methodologies for measuring the correlations between post-editing effort and machine translatability". *Machine Translation* 19(1): 37-58.

O'Brien, Sharon. 2006. "Eye Tracking and Translation Memory Matches." *Perspectives: Studies in Translatology* 14 (3): 185–205.

O'Brien, Sharon. 2008. "Processing Fuzzy Matches in Translation Memory Tools: An Eye-tracking Analysis." In *Looking at Eyes. Eye Tracking Studies of Reading and Translation Processing*, ed. by Susanne Göpferich, Arnt Lykke Jakobsen and Inger Mees, 79–102. Copenhagen: Samfundslitteratur. [Copenhagen Studies in Language, 36.]

O'Brien, Sharon. 2011. "Towards Predicting Post-Editing Productivity". *Machine Translation* 25, 197.

Plitt, Mirko, François Masselot. 2010. "A productivity test of statistical machine translation post-editing in a typical localization context". *Prague Bulletin of Mathematical Linguistics* 93, 7-16.

Saldanha, Gabriela, Sharon O'Brien. 2014. *Research Methodologies in Translation Studies*. London: Routledge.

Schmidtke, Dag. 2016. "Large scale Machine Translation publishing, with acceptable quality, for Microsoft Support content". In *Proceedings of AMTA 2016 Workshop on Interacting with Machine Translation (iMT 2016)*, Austin, Texas.

Senez, Dorothy. 1998. "Post-editing service for machine translation users at the European Commission". Translating and the Computer 20. Proceedings of the Twentieth International Conference. 12-13 November 1998 (London: Aslib, 1998); 6pp

Shreve, Gregory M., Bruce Diamond. 1997. "Cognitive Processes in Translation and Interpreting: Critical Issues". In *Cognitive Processes in Translation and Interpreting*, ed. by

Joseph H. Danks, Gregory M. Shreve, Stephen B. Fountain, Michael K. McBeath, 233-251. Thousand Oaks, California: Sage Publications.

Snover, Matthew, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, John Makhoul. 2006. "A Study of Translation Edit Rate with Targeted Human Annotation". In *Proceedings of Association for Machine Translation in the Americas*, pages 223-231. August 8-12, 2006, Cambridge, Massachusetts, USA.

Specia, Lucia, Nicola Cancedda, Marc Dymetman, Marco Turchi, Nello Cristianini. 2009. "Estimating the sentence-level quality of machine translation systems". In *Proceedings of the 13th Annual Conference of the EAMT*, Barcelona, 28–35.

Specia, Lucia. 2011. "Exploiting objective annotations for measuring translation post-editing effort". In *Proceedings of the 15th conference of EAMT*, Leuven, 73–80.

Su, Keh-Yih, Ming-Wen Wu, Jing-Shin Chang. 1992. "A New Quantitative Quality Measure for Machine Translation Systems". In *Proceedings of COLING-92*, Nantes, 433-439.

Teixeira, Carlos S. C. 2014. "Perceived vs. measured performance in the post-editing of suggestions from machine translation and translation memories". In *Proceedings of AMTA 2014 Workshop on Post-editing Technology and Practice*, Vancouver, 45-59.

Turian, Joseph P., Luke Shen, and I. Dan Melamed. 2003. "Evaluation of Machine Translation and Its Evaluation". In *Proceedings of MT Summit 2003*, 386- 393, New Orleans, Louisiana.

Vasconcellos, Muriel, Marjorie León. 1985. “SPANAM and ENGSPAN: machine translation at the Pan American Health Organization.” *Computational Linguistics* 11 (2-3): 122-136.

Wagner, Emma 1985. “Post-editing systran - a challenge for commission translators”. *Terminologie et Traduction*, 3: 1-7.

Way, Andy. 2013. “Traditional and emerging use-cases for machine Translation”. In *Proceedings of Translating and the Computer 35*, London.

Wilms, Franz-Josef M. 1981. “Von SUSY zu SUSY-BSA: Forderungen and ein anwenderbezogenes MU-System”. *Sprache und Datenverarbeitung* 5: 38-43.

Biography

Joss Moorkens is an Assistant Professor of Translation Studies in the School of Applied Language and Intercultural Studies at Dublin City University (DCU) and a researcher in the ADAPT Centre and the Centre for Translation and Textual Studies. Within ADAPT, he has contributed to the development of translation tools for both desktop and mobile, and he led development of a multimodal translation editing interface. He is co-editor of a book on human and machine translation quality and evaluation and has authored journal articles and book chapters on topics such as translation technology, post-editing of machine translation, human and automatic translation quality evaluation, and ethical issues in translation technology in relation to both machine learning and professional practice.