Moorkens, J., O'Brien, S., Silva, I. A. L., Fonseca, N., Alves, F. 2015. Correlations of perceived post-editing effort with measurements of actual effort. Machine Translation 29(3).

Correlations of perceived post-editing effort with measurements of actual effort

Joss Moorkens, ADAPT Centre/School of Computing, Dublin City University, Ireland. E-mail: joss.moorkens@dcu.ie
Sharon O'Brien , ADAPT Centre/ SALIS/CTTS, Dublin City University, Ireland. E-mail: sharon.obrien@dcu.ie
Igor A. L. da Silva, Instituto de Letras e Linguística, Federal University of Uberlândia, Brazil. E-mail: ials@ileel.ufu.br
Norma B. de Lima Fonseca, Laboratory for Experimentation in Translation (LETRA), Federal University of Minas Gerais, Brazil. E-mail: normafonseca@gmail.com
Fabio Alves, Laboratory for Experimentation in Translation (LETRA), Federal University of Minas Gerais, Brazil. E-mail: fabio-alves@ufmg.br

Abstract

Human rating of predicted post-editing effort is a common activity and has been used to train confidence estimation models. However, the correlation between human ratings and actual post-editing effort is under-measured. Moreover, the impact of presenting effort indicators in a post-editing user interface on actual post-editing effort has hardly been researched. In this study, ratings of perceived post-editing effort are tested for correlations with actual temporal, technical and cognitive post-editing effort. In addition, the impact on post-editing effort of the presentation of post-editing effort indicators in the user interface is also tested. The language pair involved in this study is English-Brazilian Portuguese. Our findings, based on a small sample, suggest that there is little agreement between raters for predicted post-editing effort and that the correlations between actual post-editing effort and predicted effort are only moderate, and thus an inefficient basis for MT confidence estimation. Moreover, the presentation of post-editing effort indicators in the user interface appears not to impact on actual post-editing effort.

Keywords

Post-editing, post-editing effort, eye-tracking, confidence estimation, confidence indicators, machine translation user evaluation.

1 Introduction

As the quality of machine translation (MT) has incrementally improved in recent years, post-editing of MT has been increasingly implemented as a method of translating large volumes of text at relatively low cost in localization workflows (DePalma et al. 2013). Research has indicated that translation productivity may be improved by introducing post-editing (PE) of MT for certain domains, language pairs, and when MT quality is sufficient (Guerberof 2009; Plitt and Masselot 2010; de Almeida and O'Brien 2010).

Translators are, however, still faced with varied MT quality. To improve both productivity and to reduce cognitive friction (Cooper 2004), it would seem useful to only present translators with high-quality MT for post-editing, i.e. to implement a method for only displaying fair to excellent MT output, or at least to indicate the quality of the MT output to the translator, just as fuzzy match scores indicate similarity in Translation Memory (TM) systems. Post-editors themselves have previously requested that MT confidence estimations be displayed within their translation interface (Moorkens and O'Brien 2013) and, as TM and MT technologies become more integrated, it can be expected that metadata pertaining to the origin and quality of an MT suggestion within a TM interface will become increasingly important. Such confidence, or trust scores, need to be trustworthy and reliable and should reflect how much post-editing effort is really required. MT quality and confidence scores are often 'tuned' using human ratings of post-editing effort (Quirk 2004). The accuracy of these ratings is still open to question, however. Moreover, little is known about how much attention post-editors might pay to such scores and how they might impact on PE effort.

To at least start addressing this research gap, in this paper we report on a three-stage study whereby six experienced translators were first asked to rate segments of English-to-Brazilian Portuguese machine translated output from two texts for estimated post-editing effort (Stage 1). The effort categories

were simplified to fit with a three-colour 'traffic light' confidence indicator to be used at Stage 3 of the study. In the second stage, after a break of several weeks, the same raters were asked to post-edit the machine translated texts (using a beta PE environment) and their estimated and actual post-editing effort were compared. This allowed us to establish whether post-editors' predictions about PE effort at the segment level are borne out by correlations with multiple measurements of their actual effort.

PE effort for Stage 2 was measured in three ways, as specified in Krings (2001): cognitive, temporal, and technical effort. Eye-tracking techniques were used to capture fixation count and duration as a proxy for cognitive effort; temporal effort was captured using timestamps taken from user activity data (UAD), i.e. logs of PE activity recorded within the editing environment; and technical effort was measured using the TER (Translation Edit Rate; Snover et al. 2006) metric.

In Stage 3, a different set of participants (students with minimal post-editing experience) completed the same post-editing tasks, and their technical and temporal effort was recorded[1]. The first objective of this stage was to measure actual PE effort of one group against both the actual and predicted PE effort of a different group (the Stage 1 raters/post-editors). The second objective in this stage was to see what, if any, impact the display of effort indicators had on actual PE effort. To accomplish this second objective, one task was completed using what we term Post-Editing Effort Estimation Indicators (PEEIs). These were colour-coded indicators of estimated segment-level effort displayed in the user interface based on ratings from Stage 1.

It should be noted that our PEEIs are not the same as confidence scores automatically generated by an MT system because they were created using human ratings of predicted effort. However, measurements of predicted effort using similar scales to this study have been used to build sentence-level confidence estimation (CE) models, intended to give post-editors an indication of whether an unseen MT segment is worth post-editing (Specia et al. 2009). Estimating sentence-level ratings of predicted PE effort was also a shared task at a recent Workshop on Machine Translation[2]. Therefore, there is a relationship between our PEEIs and automatically generated confidence scores. Thus, the findings here should have implications for automatic generation of confidence scores, which we will return to in Section 5.2.

2 Related work

Post-editing effort has previously been researched using one or more of Krings' distinctions (2001). While Krings' own work employs all three proposed measurements for PE effort, others (O'Brien 2005; Carl et al. 2011) have focused on PE time, as productivity is regularly of interest in applied translation research. Technical effort has been measured using several methods, including automatic evaluation metrics such as TER (Tatsumi 2009; Temnikova 2010; Blain et al. 2011). Cognitive measurements of PE effort have been produced by measuring pause time using key- or input-logging (Krings 2001; Lacruz and Shreve 2014) and eye-tracking techniques (Carl et al. 2011). The use of eye-tracking to measure cognitive load is well established at this stage (Rayner 1998), and measurements of fixation duration and fixation count have been used by many researchers of translation and PE (O'Brien 2011; Doherty 2012).

Tatsumi and Roturier (2010) measured temporal and technical PE effort in order to compare with Systran and Acrolinx measurements for ambiguity, complexity and style compliance. Koponen (2012) compared perceived technical PE effort with actual technical effort (measured using the TER metric), counting the types of edits when the disparity between perceived and actual PE effort was large. Koponen (2012) posited that the tendency for participants to rate long segments poorly, even when they may not need so many edits, suggests that segment length affects the cognitive effort required to assess and correct MT errors. She also found that in her study, using English-to-Spanish SMT (Statistical Machine Translation) of news texts, word order changes and certain parts of speech that need correction may be associated with the "perception of more effort" (Koponen 2012). O'Brien (2011) found that, using English-to-French SMT, GTM (General Text Matcher; Turian et al. 2003) and TER metrics correlated well with PE productivity and cognitive effort.

[1] Cognitive effort was also recorded for this cohort using eye tracking, but the data will be analysed at a later stage.

[2] See http://www.statmt.org/wmt14/.

Gaspari et al. (2014) compared PE productivity with translators' post-task perception of their own PE effort and speed. This evaluation revealed participants' bias against PE in their perceptions of speed improvements and in their continued preference for translation from scratch. Teixeira (2014) compared PE performance (comprising temporal and technical effort along with the number of errors in the target text) in three settings, with participants' post-task ratings of their own performance in those settings, and found that the two did not correlate for all segments. He also compared a translation memory (TM) work environment (with translation metadata) with a PE-focused work environment (without metadata) and found, perhaps unsurprisingly, a correlation between familiarity with a task/setting and preferring to carry out the task. Although performance was sometimes better in the PE-focused environment, participants believed that they had been more productive in the familiar setting. Läubli et al. (2013) also suggested using a fully-featured TM environment for environmentally valid PE studies, although this is not always possible as a TM environment may not always have the requisite functionality (see Vieira and Specia 2011).

Krings' three measurements for PE effort are, at this stage, relatively well established and widely used. Both Koponen (2012) and Teixeira (2014) found an occasional disconnect between perception and reality for participants in PE studies, and some research has searched for a predictor, such as Working Memory Capacity, to explain participant variability (Vieira 2014). In this paper, we add to this work by first comparing perceived PE effort with each of Krings' three measurements of PE effort in order to see which, if any, correlates closely. We then investigate what effect the display of confidence indicators, based on the scores for perceived effort, may have on actual PE effort.

CE models can be created by feeding examples of source and translation features plus quality annotation to machine learning algorithms. These annotations can be from automatic evaluation metrics (such as BLEU or NIST: Blatz et al. 2004), human-annotated MT output (Quirk 2004; Specia et al. 2009), pseudo-references (data output from different MT systems; Soricut and Narsale 2012), data close to the training and test set (Biçici and Way 2014), or measurements of PE time (Specia 2011). Specia (2011) noted the expense of human annotation of MT and found that PE time may be a more productive basis for CE modelling. A comparison of perceived PE effort and three measures of actual PE effort, along with the follow-on study of the effectiveness of confidence indication based on perceived effort ratings, is also likely to benefit research on CE, and help with selecting which of several translation options to display to the user (Shah and Specia 2014) or whether to display MT output below a chosen quality threshold at all.

3 Methodology

3.1 Research Questions

This work follows on from an earlier study that sought to identify PE-specific features that could be incorporated into editing environments to make the task more efficient for post-editors as described in Moorkens and O'Brien (2013). The current study is intended to answer two research questions:

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?

In answering Question 1, we compare user PE effort prediction ratings with three measurements of actual PE effort by the same users and by a second independent group of users to search for correlations. We expect (or hope) that ratings of perceived PE effort are reflective of actual PE effort by the same cohort of people and by an independent cohort. That is, we are testing whether rating of perceived effort is a reliable method. For Question 2, our null hypothesis is that post-editors are not influenced by the confidence score displayed in the user interface (UI); despite what the score suggests, they will use their subjective judgement to decide whether the MT output needs little or extensive post-editing.

3.2 Research Design

Moorkens, J., O'Brien, S., Silva, I. A. L., Fonseca, N., Alves, F. 2015. Correlations of perceived post-editing effort with measurements of actual effort. Machine Translation 29(3).

The study used a beta PE environment (called "PEARL: Post-Editing Assistance and ReLearning" – see Figure 1) for the PE task[3]. There are several reasons for choosing to use the PEARL tool. Firstly, the tool is web-based and can be used anywhere without installation. The hosting server saves logs of UAD for each PE session, containing timestamps for temporal measurement and final post-edited texts for technical measurement. These logs were attributed to a random session ID created within the tool, to ensure anonymity of participants[4]. Secondly, as we have control over the UI, we can add PEE-Is onscreen for the Stage 3 in this study.

Figure 1. The PEARL post-editing interface

As the PEARL interface is self-contained and the Tobii Studio eye-tracking software ends the current session when the user closes the active browser window, it was decided to use general text for this study, which would limit and, hopefully, eliminate the need for translation-associated research by the participants. Two sets of data were created, each using 40 source text segments taken from the English-language Wikipedia pages describing Paraguay (Test Set 1) and Bolivia (Test Set 2). The source text data was machine translated into Portuguese using Microsoft's Bing Translator, pre-assessed by one of the research team (who made some manual edits to seven of the segments to increase the number that were likely to be rated poorly and so even out the quality of the segments), and populated into an XLIFF file to use in PEARL. The source and MT segments were also entered into an online survey[5] to be completed by raters, who could rate each MT segment using a simplified version of the categories from Specia et al. (2009). As mentioned, the categories were simplified to fit with a three-colour 'traffic light' confidence indicator to be used in Stage 3 of this research. Krings (2001) also categorised MT quality ratings into three groupings, which he called good, medium, and poor.

The three categories in this study were:

1: Requires complete retranslation (red)
2: Requires some editing, but PE still quicker than retranslation (amber)
3: Little or no PE needed (green)

User Group 1 (see Section 3.3) participated in the first two stages of this study. Beforehand, they signed a Portuguese language informed consent form (as required by the research ethics committees of both DCU and UFMG). In Stage 1 they were requested to evaluate every segment within both test sets, to enter their online ratings, and then given break of at least two weeks so that they would not

[3] This is a fork of HandyCAT (Hokamp and Liu 2015).
[4] Participant anonymity is a standard requirement of the university research ethics approval process.
[5] Using the esurv.org platform.

easily remember their ratings. In Stage 2 they were first informed of some PE guidelines (based on O'Brien 2010), and then asked to post-edit the same segments within the PEARL interface while using a Tobii T60 eye tracker in the Laboratory for Experimentation in Translation (LETRA) at the Federal University of Minas Gerais (UFMG) in Brazil. The PE guidelines were as follows:

- 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

Participants were given a brief introduction to the PEARL interface before beginning the PE task. The Stage 2 participants completed the PE tasks without seeing any confidence indicators in the user interface.

Participants in Stage 3 of the study were students from User Group 2. These users received the same tasks and PE guidelines as in Stage 2, however in the third stage one of the tasks was completed 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, and of which task contained the PEEIs, was randomised, with eight participants assigned each one of the four conditions shown in Table 1.

	Condition 1	**Condition 2**	**Condition 3**	**Condition 4**
Test & Feature Set A	TS1/No PEEI	TS1/PEEI	TS2/PEEI	TS2/No PEEI
Test & Feature Set B	TS2/PEEI	TS2/No PEEI	TS1/No PEEI	TS1/PEEI

Table 1. Randomised order of tasks in Stage 3; TS = test set; PEEI= post-editing effort estimate

3.3 Participant Profiles

The first group of participants (User Group 1) who participated in Stage 1 of this study were six members of staff, post-doctoral researchers, and PhD students at UFMG in Belo Horizonte in Brazil. Five are native speakers of Brazilian Portuguese (the remaining one is Argentinian but has lived in Brazil for many years), and all have extensive experience of translation and post-editing. All six rated the two sets of 40 segments according to the categories described in Section 3. In Stage 2, two participants did not complete the post-editing task. This was due to scheduling issues in the case of one participant and the withdrawal of the other participant. Hence, we had six raters, four of whom also did the post-editing task.

User Group 2 was comprised of 33 undergraduate and Masters' level translation students, with minimal PE experience. All but one participant completed 2 tasks each, for a total of 65 completed tasks. Due to logging problems on the remote server, UAD was not saved for 22 of these tasks. The results presented for technical effort are therefore based on 20 participant logs for Test Set 1, and 24 participant logs for Test Set 2. One further outlier result (P37, Test Set 2) was not considered for temporal effort, as the time spent per segment was up to ten times as long as the average.

3.4 Measurements

Following the manual rating task in Stage 1, temporal, technical, and cognitive effort was measured for four participants in Stage 2. Only temporal and technical effort was measured in Stage 3. Temporal effort was measured for each segment by converting the Unix timestamp information for the first edit on the opening segment to a human readable date, and then recording the timing for each

'segment-finished' tag in the logs, marking the completion of that segment. A calculation of seconds per segment could then be made for each participant. Technical effort was measured retrospectively using the TER COMpute Perl code[6] to get a TER score for each segment. This use of TER to calculate the minimum number of edits required to change the MT output to the post-edited segment has been used in previous studies (such as Alves et al. 2015; Da Silva et al. 2015), even though the TER score may differ slightly from the number of edits actually made by the post-editor. Cognitive effort was measured using the eye-tracking software analysis tool Tobii Studio (v.3.1) to analyse fixation data within two assigned areas of interest (source and target text) recorded during PE of each segment. For fixation data, we were interested primarily in the total number of fixations and mean fixation duration as indicators of cognitive effort.

4 Results

4.1 Stage 1

Once all six participants had finished the Stage 1 rating task, a mean rating for each MT segment was calculated to compare with PE effort. This was calculated by scoring (per participant) 1 point to each segment that was said to require complete retranslation, 0.5 points to each that required some retranslation (but PE still quicker than retranslation), and 0 to each segment that required little or no PE, then dividing the total score by the number of participants. A segment with a mean score of ≤0.30 was marked 'green', meaning that our raters believed that the segment would require little editing. A segment with a mean score of ≥0.70 was marked as 'red', and thus likely to require heavy editing. The remaining segments were marked as 'amber' - requiring some editing, but PE was still perceived to be quicker than retranslation.

The correlation between the expected post-editing effort determined by each participant and that determined by means of *all* participants scores is r_s=0.373 (p=0.000). This is indicative of a certain level of subjectivity in assessing post-editing effort, that is, the participants do not agree as to what level of post-editing effort each segment should be ascribed. In fact, their assessments were 100% equivalent for only eight of the 40 segments in Test Set 1, and five of the 40 segments in Test Set 2. Already this suggests that rating of perceived post-editing effort is not very reliable. The assessments were equivalent among at least three of the participants for 23 (57.5%) of the segments in Test Set 1 and 21 (52.5%) of the segments in Test Set 2.

The segments marked red, amber, and green based on the average rating were relatively homogeneous in terms of mean number of words (19.53, 22.93, and 18.72 respectively and 21.00 for all segments – well within the standard deviation of 9.20), and mean word length (5.07, 5.13, and 5.20 respectively, 5.14 for all segments). Based on the average ratings, 24 segments were categorised as 'green', 15 as 'red', and 41 as 'amber'.

4.2 Stage 2: Temporal effort

In this stage, the actual PE time of the raters was recorded and compared to their perceived effort ratings. Table 2 shows an example of segment rating alongside a temporal measurement of seconds spent on the segment by each participant, followed by the mean PE duration. For space reasons, we show just a sample of each category here in order of average participant rating.

PEEI Category	Segment	PE Effort Estimate	P1 Sec/Seg	P2 Sec/Seg	P3 Sec/Seg	P4 Sec/Seg	Mean Seconds/Segment
Green	P32	0.08	16	5	17	19	14.25
Green	B16	0.17	7	6	47	7	16.75
Green	B35	0.25	14	16	37	12	19.75
Amber	B04	0.33	13	11	83	41	37.00

[6] Available from www.cs.umd.edu/~snover/tercom.

Amber	P05	0.50	20	22	48	52	35.50
Amber	B34	0.58	64	66	120	60	77.50
Red	B01	0.75	51	39	146	44	70.00
Red	P39	0.83	43	45	92	32	53.00

Table 2. Results for temporal effort in Stage 2

Mean temporal effort was somewhat higher for poorly-rated segments (amber and red), as may be seen in Figures 2 and 3, and there is a trend towards higher perceived effort ratings being related to higher temporal effort. However, the relationship between average ratings of predicted PE effort and actual temporal effort was not strong - r_s=0.492 (p=0.000), which suggests no more than a moderate correlation. The relationship between individual ratings and temporal effort was also moderate (r_s=0.465 (p=0.000) for Participant 1, r_s=0.528 (p=0.000) for Participant 2, the correlation for Participant 3 was not considered significant, r_s=0.488 (p=0.000) for Participant 4).

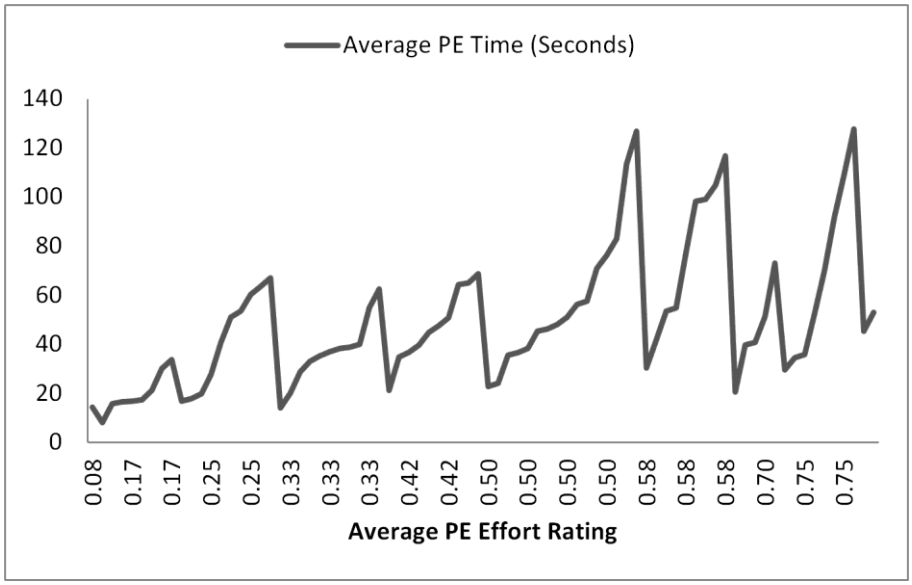

Figure 2. A chart of average Stage 1 rating vs. temporal effort measured in Stage 2

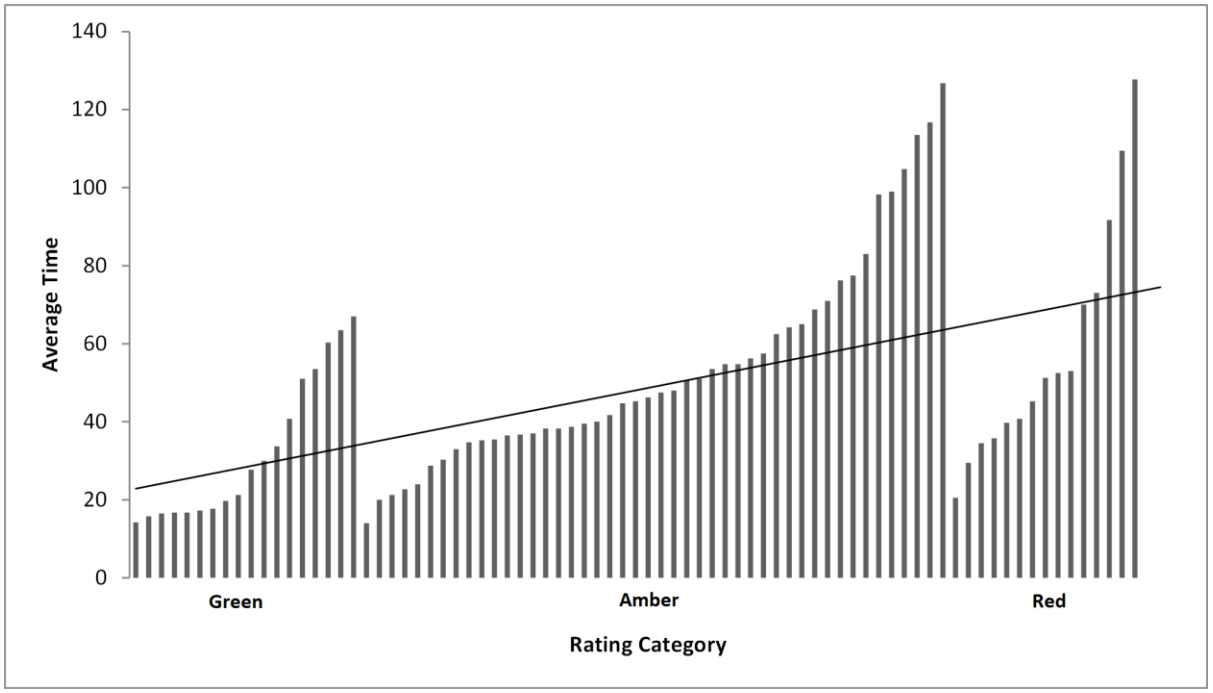

Figure 3. A chart of Stage 1 rating category vs. temporal effort measured in Stage 2

4.3 Stage 2: Cognitive effort

Color/ Measure		Fix. No.	Total Fix.	Mean Fix.	Fix No. ST	Total Fix. ST	Mean Fix. ST	Fix. No. TT	Mean Fix TT
Red	Mean	8.30	1984.11	241.22	2.78	555.74	195.10	1428.36	260.29
	n	60	60	60	60	60	60	60	60
	SD	5.55	1319.74	36.34	2.71	584.09	33.96	825.23	45.23
Amber	Mean	7.02	1668.44	237.11	2.35	450.67	187.69	1217.77	258.93
	n	164	164	164	164	164	164	164	164
	SD	4.38	1093.64	40.86	1.71	341.31	34.07	821.16	52.87
Green	Mean	5.37	1200.48	220.17	2.09	398.17	188.53	802.31	236.32
	n	96	96	96	96	96	96	96	96
	SD	3.82	903.98	40.62	1.90	384.59	27.20	593.32	52.63

Table 3. Mean and standard deviation for fixation measures

Table 3 shows the eye-tracking measures for the 80 segments classified into green, amber and red (respectively, 24, 41, and 15 segments). By examining the total number of fixations and mean fixation duration, we can clearly see that cognitive effort decreases as the post-editors move through the segments categorised as red, amber, and green. Table 4 shows that effort is significantly different between red and green segments and between amber and green segments, but not between red and amber segments. This may indicate that there is not a substantial enough differentiation between the quality of red and amber segments for this to be manifested in the cognitive effort measurements using eye tracking.

	Red vs. Amber	Red vs. Green	Amber vs. Green
Fix. No.	0.134	0	0
Total Fix.	0.071	0	0
Mean Fix.	0.47	0	0
Fix. No. ST	0.561	0.039	0.039
Total Fix. ST	0.475	0.033	0
Mean Fix. ST	0.227	0.895	0
Fix. No. TT	0.051	0	0
Total Fix. TT	0.048	0	0
Mean Fix. TT	0.716	0	0

Table 4. Significance comparing red, amber and green segments using Mann-Whitney U test

When checking Spearman's correlations between the predicted effort and the eye-tracking informed measures, most correlations are either very weak (0.00-0.19) or weak (0.20-0.39) as may be seen in Table 5. A moderate correlation (r_s=0.443) is found only for the mean score of total number of fixations on the target text. No strong correlations were found. This seems to indicate that humans' ratings for predicted PE effort are moderately, but not very strongly correlated to actual post-editing effort, when measured through fixation data. Total number of fixations on the TT would appear to be

the best indicator of actual PE effort, which stands to reason (the post-editor has to look at the TT longer if it requires substantial editing). Carl et al. (2011) found that SMT output requires more reading and rereading effort than manual translation, based on both fixation count and fixation duration.

Participant	r_s for Rating vs. Avg. Fixation Duration	r_s for Rating vs. Fixation Count
P1	0.383 (p=0.000)	0.439 (p=0.000)
P2	0.440 (p=0.000)	0.483 (p=0.000)
P3	0.276 (p=0.013)	0.136 (p=0.227)
P4	0.336 (p=0.002)	0.375 (p=0.000)

Table 5. Spearman's correlations for individual ratings vs. cognitive effort

4.4 Stage 2: Technical effort

PEEI Category	**Segment**	**PE Effort Estimate**	**P1 TER**	**P2 TER**	**P3 TER**	**P4 TER**	**Avg. TER**
Green	P32	0.08	0	0	0	44.44	11.11
Green	B16	0.17	0	0	0	0	0
Green	B35	0.25	0	5.56	5.56	5.56	4.17
Amber	B04	0.33	4.35	4.35	13.04	21.74	10.87
Amber	P05	0.50	10.53	5.26	10.53	42.11	17.10
Amber	B34	0.58	26.09	21.74	34.78	21.74	26.09
Red	B01	0.75	56.00	8.00	56.00	28.00	37.00
Red	P39	0.83	0	35.29	35.29	47.06	29.41

Table 6. Stage 2 technical effort data

Table 6 shows each Stage 2 participant's TER score and the average TER score for a sample of ten segments (again, for space reasons). Figure 4 shows the relationship between technical effort (measured using the TER metric) and the average predicted effort rating.

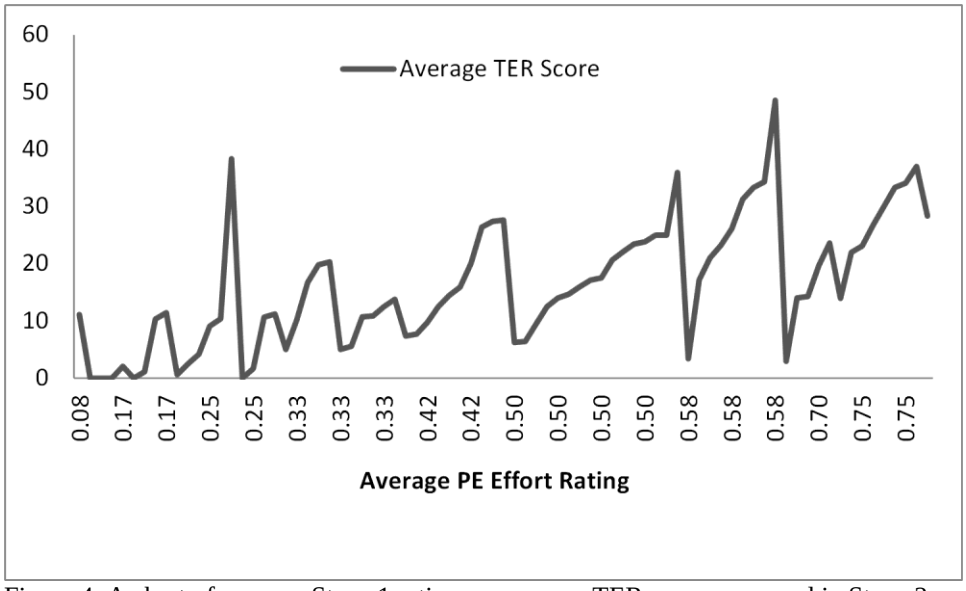

Figure 4. A chart of average Stage 1 rating vs. average TER score measured in Stage 2

As with temporal effort, there is a trend towards higher Stage 1 predicted effort ratings being related to higher technical effort. However, for this metric we found a strong correlation, with r_s=0.652 (p=0.000), although the relationship between individual ratings and technical effort was only moderate (r_s=0.431 (p=0.000) for Participant 1, r_s=0.552 (p=0.000) for Participant 2, r_s=0.326 (p=0.003) for Participant 3, and r_s=0.515 (p=0.000) for Participant 4). When compared with the moderate correlation between Stage 1 ratings and temporal PE effort, this suggests a disconnect between temporal and technical effort within this study. The results for technical and temporal effort show a moderate correlation, with r_s=0.524 (p=0.000).

Table 7 shows the relationships between predicted PE effort and the three measurements for actual PE effort used in Stage 2 of this study, with strong correlations in bold. The strongest correlation found was between PE time and mean fixation count, whereas the mean duration of fixations correlated strongly with TER scores. This suggests that some segments presented time-consuming PE problems that required a related measure of cognitive effort without requiring a related amount of edits to the text.

		Fix Count	**Fix Dur**	**Avg. Rating**	**Avg. Time**
Fixation Duration	Corr. (r_s)	0.366	-	0.505	0.431
	Sig.	0.001	-	0.000	0.000
Average Est. Edit Rating	Corr. (r_s)	0.411	0.505	-	0.492
	Sig.	0.000	0.000	-	0.000
Average Time	Corr. (r_s)	**0.942**	0.431	0.492	-
	Sig.	0.000	0.000	0.000	-
Average TER	Corr. (r_s)	0.432	**0.759**	**0.652**	0.524
	Sig.	0.000	0.000	0.000	0.000

Table 7. Spearman correlations between Stage 1 estimated PE effort and Stage 2 measures of actual effort

4.5 Stage 3: Temporal Effort

Participants in Stage 3 were students of translation and, as has been established previously (Moorkens and O'Brien 2015), could be expected to take more time to post-edit each segment than the more expert User Group 1 who participated in Stage 2, who have professional translation and post-editing experience (in this study the student group took 9% longer on average). The correlation between participants' average time spent post-editing each segment and raters' estimated post-editing effort is, as with Stage 2, only moderate, with r_s=0.487 with PEEI off and r_s=0.479 with PEEI on (p=0.000). This correlation was negatively affected by several segments where estimated and actual effort diverged, such as the 9th segment in Test Set 2[7]. This segment received an average rating of 0.25 and was therefore marked with a green PEEI, yet post-editing took on average 54.75 seconds with PEEI off and 71.27 seconds with PEEI on. The latter time may be due to the dissonance between the green rating

[7] "Its main economic activities include agriculture, forestry, fishing, mining, and manufacturing goods such as textiles, clothing, refined metals, and refined petroleum. Bolivia is very wealthy in minerals, especially tin." This was the only segment that contained two sentences due to a segmentation error.

when the PEEI was displayed and the actual post-editing effort required for this segment. In comparison, the average 'green' segment in this test set required 29 seconds to post-edit.

	PEEI Off	PEEI On
TS1. Average Time/All Segments	47.84	50.69
TS2. Average Time/All Segments	53.95	53.91
TS1. Average Red Segment	49.17	49.49
TS1. Average Amber Segment	56.29	62.86
TS1. Average Green Segment	37.60	38.04
TS2. Average Red Segment	73.68	82.47
TS2. Average Amber Segment	62.67	61.47
TS2. Average Green Segment	29.05	28.90

Table 8. Temporal effort (seconds per segment) from Stage 3

Temporal effort, measured for all segments and for segments in each colour category, was very similar with and without the PEEI. Table 8 shows the average times for each test set per colour category. These results fit with our null hypothesis with regard to the influence of PEEI. Figure 5 shows the relationship between PE effort estimation and actual temporal effort with PEEI off and on.

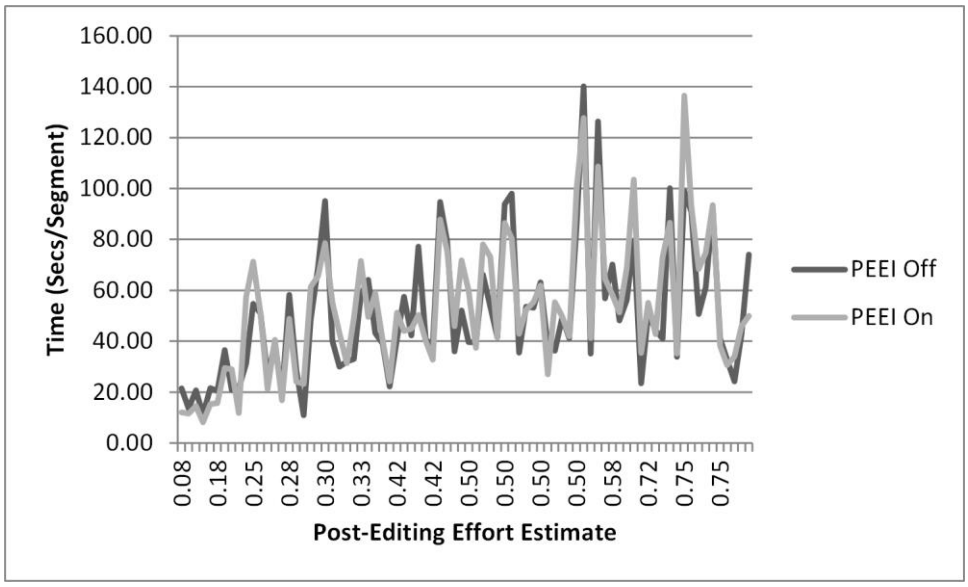

Figure 5. The relationship between estimated effort (Stage 1) and actual temporal effort in Stage 3

4.6 Stage 3: Technical Effort

Participants in Stage 3 tended to edit segments more heavily than Stage 1 participants, with an average TER per segment of 19.53 (compared with 16.31 in Stage 1). The correlation between ratings and technical effort was slightly higher than with temporal effort, but still moderate, with r_s=0.518 with PEEI off and r_s=0.558 with PEEI on (p=0.000). As noted in Section 4.5, several segments required far more effort than the raters had estimated. Based on the average TER score, the green-rated Segment 9[7] in Test Set 2 required 137 edits with PEEI off and 156 edits with PEEI on, far greater than the average TER for all segments of 19.53. This segment was the most heavily-edited in the whole study, as may be seen in Figure 5. The tenth segment in Test Set 2 was also given a green PEEI, with a rating of 0.17, yet required an average of 76 edits with PEEI off and 86 edits with PEEI on. Discounting these and one other segment gives an average TER for green segments in Test Set 2 of 8.3, far lower than the average shown in Table 9.

	PEEI Off	PEEI On
TS1. Average TER/All Segments	17.33	16.42
TS2. Average TER/All Segments	16.28	19.79
TS1. Average Red Segment	24.89	24.42
TS1. Average Amber Segment	20.31	19.42
TS1. Average Green Segment	23.57	26.20
TS2. Average Red Segment	20.13	24.73
TS2. Average Amber Segment	19.62	24.08
TS2. Average Green Segment	10.74	12.92

Table 9. Technical effort (TER score) from Stage 3

There was no significant difference in technical effort when the PEEIs were on or off, again supporting the null hypothesis for the effect of adding the PEEIs. Table 9 shows the average TER scores for each test set per colour category. Figure 6 shows the relationship between PE effort estimation and actual technical effort with PEEI off and on.

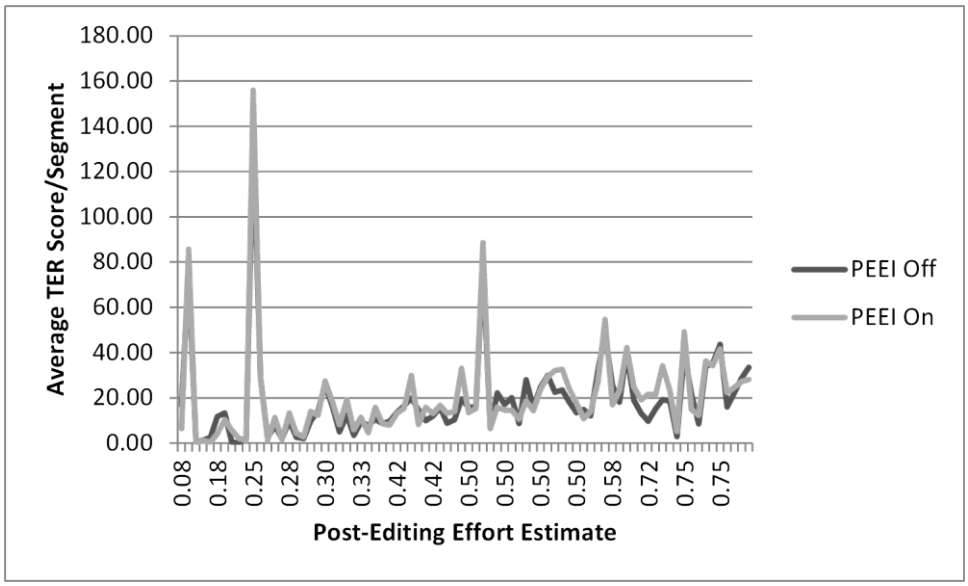

Figure 6. The relationship between Stage 1 estimated effort and average TER score in Stage 3

5 Concluding Remarks

5.1 Summary

We set out to answer two research questions in this study. Firstly, we tested whether human estimates of predicted post-editing effort were predictors of actual post-editing effort, by means of correlations with technical, temporal and cognitive measures. Secondly, we investigated whether the display of PE effort estimation indicators influenced PE behaviour. We tested this for English-Brazilian Portuguese, general domain text.

Our answers to these questions are limited by the small size and high variability of our Stage 1 participants. We noted a trend in the temporal data from Stage 2 and Stage 3 to suggest that as predicted PE effort ratings increased so too did the time required to post-edit. However, the correlations were only moderate, leading to a conclusion that human ratings of PE effort do not correlate strongly with the actual time required during post-editing. This could be due in part to the rating descriptions, which suggested technical effort by asking raters *how much* of the segment they estimated to require post-editing, although unbiased operationalisation of categories is difficult, and Category 2 did also specify that PE of the segment should be 'quicker than retranslation'. Nonetheless, the correlation between average Stage 1 ratings and Stage 3 technical effort, although

better, was still moderate. The average ratings showed a strong correlation with technical effort in Stage 2 of the study, although that did not necessarily equate to participants spending more time editing the segments. This suggests that, as found by Koponen (2012), some MT output edits are quicker to perform than others. However, her finding that participants tend to rate longer segments poorly, whether or not they require many edits, was not replicated here.

It was also found that the cognitive measures of PE effort did correlate with the general classifications in that as the classification moved through green, amber, and red, the fixation data suggested an increase in cognitive load. The correlations were, however, weak to moderate and there was no statistically significant difference between the red and amber segments.

Stage 3 of the study confirmed our hypothesis in Research Question 2 - that post-editors are not influenced by the score and will use their subjective judgement to decide whether the MT output needs little or extensive post-editing. There was no significant difference between measurements of PE effort for any of the rated categories of segments, despite their accurate reflection of predicted effort ratings and strong correlation with technical effort in Stage 2. This begs the question: are post-editors really interested in confidence scores? When looking through the screen recordings, it was possible to see moments where user behaviour was influenced by the presence of the indicators. For example, several users hovered with their mouse over the 'submit' button in the UI when they came across a segment that had the green indication, but this behaviour was not replicated across all users, nor did it have any impact on our measurements of PE effort.

5.2 Conclusion and future research

Although the sample is small and only applied to one language pair, our findings suggest that human ratings of predicted PE effort are not completely reliable indicators of actual effort. This may explain why post-editors in this study displayed no significant behavioural change whether or not indicators based on predicted effort were presented. Despite differences between expert and novice groups found in previous research (Moorkens and O'Brien 2015), there was a strong correlation between the technical and temporal effort of each group in this study, which suggests that actual effort, even from a different user group, would be a more reliable indicator of future effort. As such, an MT confidence estimation model based on PE effort, which Specia (2011) suggests can be built from a relatively small corpus of post-edited text, would, we suggest, be more reliable to one built using human-rated segments. How to generate confidence estimates that are reliable and are actually taken on board by post-editors is an open question.

In continuing this research, we intend to investigate what kind of linguistic problems are corrected more often during post-editing and see if there is a relationship with the colour of the indicator that was presented. Some linguistic problems can be more serious (requiring heavy editing) to some participants than to others (requiring little or no editing), and so this could explain the variation of estimated effort. We also would like to investigate how participants solved linguistic problems, and to identify problems of procedural and conceptual encoding, within the framework of relevance theory (Sperber and Wilson 1986). By comparing PE effort between problems of either type, this may help prioritise future research to reduce complexity of post-edits for the optimum reduction in human PE effort. Given the limitations of the research presented here in terms of number of languages, participants and words there is also room to build on the research by expanding on all of these dimensions.

Acknowledgement

This research is supported by Science Foundation Ireland (Grant 12/CE/I2267) as part of the CNGL (www.cngl.ie) at Dublin City University, Research Brazil Ireland, and by the FALCON Project (falcon-project.eu), funded by the European Commission through the Seventh Framework Programme (FP7) Grant Agreement No. 610879. The authors would like to place on record their thanks to Research Brazil Ireland, and to staff and students at the Faculdade de Letras at UFMG.

References

Alves F, Koglin A, Mesa-Lao B, García MM, Fonseca NBL, Sá AM, Gonçalves JL, Szpak KS, Sekino K, Aquino M (2015) Analysing the impact of interactive machine translation on post-editing effort. In Carl M,

Moorkens, J., O'Brien, S., Silva, I. A. L., Fonseca, N., Alves, F. 2015. Correlations of perceived post-editing effort with measurements of actual effort. Machine Translation 29(3).

Bangalore S, and Schaeffer M (eds) New Directions in Empirical Translation Process Research, Springer, New York, pp 77-95

Biçici E, Way A (2014) Referential translation machines for predicting translation quality. WMT 2014: ACL 2014 Ninth Workshop On Statistical Machine Translation, Baltimore

Blain F, Senellart J, Schwenk H, Plitt M, Roturier J (2011) Qualitative analysis of post-editing for high quality machine translation. In: Machine Translation Summit XIII, Asia-Pacific Association for Machine Translation (AAMT), Xiamen

Blatz J, Fitzgerald E, Foster G, Gandrabur S, Goutte C, Kulesza A, Sanchis A, Ueffing N (2004) Confidence estimation for machine translation. Proceedings of the 20th international conference on computational linguistics, 23–27 Aug 2004, Geneva, pp 315–321

Carl M, Dragsted B, Elming J, Hardt D, Jakobsen AL (2011) The process of post-editing: A pilot study. In: Proceedings of The 8th International Natural Language Processing and Cognitive Science Workshop, Frederiksberg, pp 131-142

Cooper A (2004) The inmates are running the asylum: Why hi-tech products drive us crazy and how to restore the sanity. Sams Publishing, Indianapolis

Da Silva IAL, Schmaltz M, Alves F, Pagano A, Wong D, Chao L, Leal ALV, Quaresma P, Garcia C (2015) Translating and Post-Editing in the Chinese-Portuguese Language Pair: Insights from an Exploratory Study of Key Logging and Eye Tracking. Translation Spaces, 4(1):145-169

De Almeida G, O'Brien S (2010) Analysing post-editing performance: Correlations with years of translation experience. In: Proceedings of the 14th Annual Conference of the European Association for Machine Translation, St. Raphaël

DePalma DA, Hegde V, Pielmeier H, Stewart RG (2013) The language services market: 2013. Common Sense Advisory, Boston

Doherty S (2012) Investigating the effects of controlled language on the reading and comprehension of machine translated texts: A mixed-methods approach. Dissertation, Dublin City University

Gaspari F, Toral A, Kumar NS, Groves D, Way A (2014) Perception vs reality: Measuring machine translation post-editing productivity. In: O'Brien S, Simard M, Specia L (eds) Proceedings of the 11th Conference of the Association for Machine Translation in the Americas: Workshop on Post-editing Technology and Practice (WPTP3), Vancouver, pp 60-72

Guerberof A (2009) Productivity and quality in the post-editing of outputs from translation memories and machine translation. Localisation Focus 7(1):11–21

Hokamp C, Liu C (2015) HandyCAT. In: Durgar El-Kahlout İ, Özkan M, Sánchez-Martínez F, Ramírez-Sánchez G, Hollowood F, Way A (eds) Proceedings of European Association for Machine Translation (EAMT) 2015, Antalya, pp 216

Koponen M (2012) Comparing human perceptions of post-editing effort with post-editing operations. In: Proceedings of the 7th workshop on statistical machine translation, Montreal, pp 181–190

Krings HP (2001) Repairing Texts. Kent State University Press, Ohio

Lacruz I, Shreve GM (2014) Pauses and Cognitive Effort in Post-Editing. In: O'Brien S, Balling LW, Carl M, Simard M, Specia L (eds), Post-editing of Machine Translation: Processes and Applications. Cambridge Scholars Publishing, Newcastle-Upon-Tyne, pp246-272

Läubli S, Fishel M, Massey G, Ehrensberger-Dow M, Volk M (2013) Assessing post-editing efficiency in a realistic translation environment. In: O'Brien S, Simard M, Specia L (eds) Proceedings of MT Summit XIV Workshop on Post-editing Technology and Practice, Nice, pp 83–91

Moorkens J, O'Brien S (2013) User attitudes to the post-editing interface. In: O'Brien S, Simard M, Specia L (eds) Proceedings of MT Summit XIV Workshop on Post-editing Technology and Practice, Nice, pp 19–25

Moorkens J, O'Brien S (2015) Post-editing evaluations: Trade-offs between novice and professional participants. In: Durgar El-Kahlout İ, Özkan M, Sánchez-Martínez F, Ramírez-Sánchez G, Hollowood F, Way A (eds) Proceedings of European Association for Machine Translation (EAMT) 2015, Antalya, pp 75–81

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

Moorkens, J., O'Brien, S., Silva, I. A. L., Fonseca, N., Alves, F. 2015. Correlations of perceived post-editing effort with measurements of actual effort. Machine Translation 29(3).

O'Brien S. (2010) Introduction to Post-Editing: Who, What, How and Where to Next? In: Proceedings of AMTA 2010, The Ninth Conference of the Association for Machine Translation in the Americas (online). Denver, CO

O'Brien S (2011) Towards predicting post-editing productivity. Machine Translation 25:197–215

Plitt M, Masselot F (2010) A productivity test of statistical machine translation post-editing in a typical localization context. Prague Bull Math Linguist 93:7–16

Quirk C (2004) Training a Sentence-Level Machine Translation Confidence Measure. In: Proceedings of the 4th Conference on Language Resources and Evaluation, Lisbon, pp 825–828

Rayner K (1998) Eye movements in reading and information processing: 20 years of research. Psychological Bulletin,124(3):372-422

Shah K, Specia L (2014) Quality estimation for translation selection. In: Tadic M, Koehn P, Roturier J, Way A (eds), Proceedings of the 17th Annual Conference of the European Association for Machine Translation (EAMT 2014), Dubrovnik, pp109-116

Snover M, Dorr B, Schwartz R, Micciulla L, Makhoul J (2006) A study of translation edit rate with targeted human annotation. In: Proceedings of the 7th conference of the Association for Machine Translation in the Americas (AMTA2006), "Visions for the Future of Machine Translation', Cambridge, MA, pp 223–231

Soricut R, Narsale S (2012) Combining quality prediction and system selection for improved automatic translation output. In: Proceedings of the ACL Seventh Workshop on Statistical Machine Translation (WMT-2012), Montreal, pp 163-170

Specia L, Cancedda N, Dymetman M, Turchi M, Cristianini N (2009) Estimating the sentence-level quality of machine translation systems. In: Proceedings of the 13th Annual Conference of the EAMT, Barcelona, pp 28–35

Specia L (2011) Exploiting objective annotations for measuring translation post-editing effort. In: Proceedings of the 15th conference of EAMT, Leuven, pp 73–80

Sperber D, Wilson D (1986) Relevance: Communication and Cognition. Blackwell, Oxford

Tatsumi M (2009) Correlation between automatic evaluation scores, post-editing speed and some other factors. In: Proceedings of MT Summit XII, Ottawa, pp 332–339

Tatsumi M, Roturier J (2010) Source text characteristics and technical and temporal post-editing effort: what is their relationship? In: Zhechev V (ed), Proceedings of the Second Joint EM+/CNGL Workshop "Bringing MT to the User: Research on Integrating MT in the Translation Industry" (JEC '10), Denver, CO, pp 43–51

Teixeira CSC (2014) Perceived vs. measured performance in the post-editing of suggestions from machine translation and translation memories. In: O'Brien S, Simard M, Specia L (eds) Proceedings of the 11th Conference of the Association for Machine Translation in the Americas: Workshop on Post-editing Technology and Practice (WPTP3), Vancouver, pp 45–59

Temnikova I (2010) Cognitive evaluation approach for a controlled language post-editing experiment. In: LREC 2010: proceedings of the seventh international conference on Language Resources and Evaluation, Valletta, pp 3485-3490

Turian J, Shen L, Melamed D (2003) Evaluation of machine translation and its evaluation. In: Proceedings of the MT Summit IX, New Orleans, pp 386–393

Vieira L, Specia L (2011) A review of translation tools from a post-editing perspective. In: Proceedings of the Third Joint EM+/CNGL Workshop Bringing MT to the Users: Research Meets Translators" (JEC '11), Luxembourg, pp 33-42

Vieira LN (2014) Indices of cognitive effort in machine translation post-editing. Machine Translation 28(3):187-216