Nunes Vieira, L. (2017). From process to product: links between post-editing effort and post-edited quality. In A. Lykke Jakobsen, & B. Mesa-Lao (Eds.), *Translation in Transition: Between Cognition, Computing and Technology* (pp. 162–186). (Benjamins Translation Library; Vol. 133). John Benjamins Publishing Company.
<https://doi.org/10.1075/btl.133>

Peer reviewed version

Link to published version (if available):
[10.1075/btl.133](https://doi.org/10.1075/btl.133)

[Link to publication record in Explore Bristol Research](#)
PDF-document

This is the author accepted manuscript (AAM). The final published version (version of record) is available online via John Benjamins Publishing Company at <https://benjamins.com/#catalog/books/btl.133/main>. Please refer to any applicable terms of use of the publisher.

University of Bristol - Explore Bristol Research

General rights

This document is made available in accordance with publisher policies. Please cite only the published version using the reference above. Full terms of use are available:
<http://www.bristol.ac.uk/red/research-policy/pure/user-guides/ebr-terms/>

AUTHOR'S ACCEPTED MANUSCRIPT

TO CITE THIS CHAPTER:

Vieira, L.N. From process to product: links between post-editing effort and post-edited quality. In: Jakobsen, A.L. and Mesa-Lao, B. (eds) *Translation in transition: between cognition, computing and technology*. Amsterdam: John Benjamins. 162-186. Doi: [10.1075/btl.133.06vie](https://doi.org/10.1075/btl.133.06vie)

This chapter is under copyright. the publisher should be contacted for permission to re-use or reprint the material in any form

CHAPTER 6

From process to product: links between post-editing effort and post-edited quality

Lucas Nunes Vieira

University of Bristol

Post-editing of machine translation (MT) is now increasingly implemented in the human translation workflow after studies in both industry and academia have demonstrated the efficacy of this practice. Post-editing still involves open questions, however, such as how best to train post-editors and how to estimate the effort required by post-editing tasks. In attempting to address some of these questions, many previous studies investigate the post-editing process, but less research has focused on the post-edited product. This chapter examines the link between the process and product of post-editing by checking to see how post-editing effort data related to the quality of post-edited texts, assessed in terms of fluency (linguistic quality) and adequacy (translation accuracy). A statistical analysis indicated that the association between editing operations and the fluency of post-edited texts is dependent on the quality of the raw MT output. Interestingly, a negative association was observed between the number of eye fixations on the text and the quality of the post-edited translations. The chapter shows empirical evidence supporting the distinction between the concepts of translation fluency and adequacy, and postulates that automatic processes play a central role in post-editing performance.

1. Introduction

The use of machine translation (MT) in the human translation workflow is now a common practice. A number of studies show evidence that confirms the efficacy of post-editing in terms of greater quality and speed (e.g. Plitt and Masselot 2010; Green, Heer, and Manning 2013), and a flurry of research has investigated different facets of post-editing (e.g. Carl et al. 2011; Koponen 2012; Koponen et al. 2012; de Almeida 2013). While much of this work has dealt with post-editing effort, little information is available regarding the connection between the post-editing process and the post-edited product. This chapter examines this connection by testing measures of post-editing effort as potential predictors of the quality of post-edited texts, assessed in terms of fluency (linguistic quality) and adequacy (translation accuracy) as per guidelines proposed by the Translation Automation User Society (TAUS).

Post-editing effort is normally regarded as a three-fold construct that comprises temporal, technical, and cognitive (or mental) effort (Krings 2001). Temporal effort consists of the amount of time required by the task. Technical effort concerns “purely technical operations” (Krings 2001:179) and is therefore distinguished from cognitive effort per se, which relates to “the type and extent of cognitive processes” (Krings 2001:179) involved in post-editing (i.e. the amount of mental processing that takes place). All of these effort dimensions were taken into account in the present study in empirical post-editing tasks. Post-editing effort was estimated here based on eye movements, post-editing time, number of editing operations, and subjective ratings on cognitive effort (henceforth “subjective cognitive effort”) (see section 3.1).

Three professional translators were hired to assess the post-edited texts so that the effort measures mentioned above could be contrasted with product quality. The results were analysed with mixed-effects regression. This method allowed post-editors’ individual characteristics such as working memory capacity (WMC), level of professional experience and source-language proficiency to be considered as additional factors (see section 3.2).

Details of the analysis are presented in the remainder of this chapter as follows: previous work is reviewed in section 2; the study’s methodology is outlined in section 3; results are reported and discussed in section 4; and conclusions are presented in section 5, together with suggestions for future research.

2. Related Work

Post-edited quality is assessed in previous research mainly as a way of comparing post-editing with traditional translation. For example, Plitt and Masselot (2010) showed that post-edited texts have fewer errors compared to texts translated from scratch. Carl et al. (2011) observed a quality gain for post-editing, but the difference in relation to traditional translation was not statistically significant. Green, Heer, and Manning (2013) showed that post-editing increases overall translation quality and improves translators’ productivity. Depraetere, De Sutter, and Tezcan (2014) observed a slight decrease in quality for the post-editing condition, but they did not consider the difference to be a significant one.

There have also been studies that looked at the impact of post-editors’ profiles on post-edited quality. Green, Heer, and Manning (2013) found significant effects of source-language skills in this respect. De Almeida (2013) observed that being able to follow guidelines and having a positive attitude towards MT are also factors to be taken into account. Interestingly, de Almeida found no correlation between post-editing time and participants’ level of professional experience – results similar to those observed by Guerberof (2014).

It is noteworthy that most of the studies described above involve either quantitative comparisons between traditional translation and post-editing or macro-level investigations of how post-editor profiles affect post-editing performance. Information on how

processual aspects of post-editing relate to post-edited quality is for the most part touched upon only tangentially. For example, Carl et al. (2011) found no correlation between post-edited quality and the number of changes implemented in the text, a relationship that is further tested in this chapter based on different levels of MT quality.

Green et al. (2014) examined the feasibility of an interactive translating interface where translations were automatically adapted on the fly according to post-editors' modifications. To this end, they used a statistical model where, together with the interactive condition and other factors, they tested post-editing time as a predictor of post-edited quality. They found a surprising negative link between time and quality, which was interpreted as an artefact of BLEU+1 (Lin and Och 2004), the quality score used in their study. This is a version of BLEU that is fit for sentence-level evaluation, but which similarly to other versions of the score is known to be influenced by sentence length.

Mitchell (2015) investigated the work of non-professionals in community post-editing. Among other factors, she contrasted attitude to MT, keystrokes, post-editing time and self-reported linguistic skills with post-edited fluency and adequacy. The effects she observed in these comparisons were mostly non-significant except for a negative effect of post-editing time on post-edited adequacy, and positive effects of a favourable attitude to MT and of source language knowledge. It should be noted, however, that Mitchell's analysis was based on correlation tests. With this method, it is not possible to statistically control for different factors or indeed to discriminate between different levels of MT quality in the same test.

The analysis presented in this chapter is capable of addressing these aspects, and can therefore shed light on a number of challenges left unsolved by previous research. In particular, the design adopted here distinguishes between individual traits and post-editing behaviour in searching for predictors of post-edited quality. Moreover, the quality of the raw MT output was controlled for, which allowed the post-editing process to be observed in a range of conditions, uncovering information that can influence professional practice and improve the current understanding of what constitutes effective post-editing behaviour.

3. Methods and Procedure

The tasks described below were conducted in the context of a larger project aimed at investigating cognitive effort in post-editing. For more information on the study design, see Vieira (2014;2016).

3.1. Study Materials and the Post-Editing Task

The post-editing task was based on the French-to-English language pair. The selected texts were extracts of two news articles taken from the 2013 corpus of the Workshop on Statistical Machine Translation (WMT). After a warm-up phase, participants edited

the texts in two sessions, in counterbalanced order. Participants subsequently took two tests aimed at measuring their proficiency in French and their working memory capacity (WMC), attributes that could have an impact on their editing performance (for a discussion on the potential relationship between editing and WMC, see Temnikova 2010; McCutchen 1996). Participants were also asked to rate their attitude towards MT from 1 (negative) to 5 (positive), as previous research has found this to be a potential factor in post-editing behaviour (de Almeida 2013; Mitchell 2015). The French proficiency task was an online vocabulary test largely used for placement purposes (Meara and Buxton 1987). The WMC test was an automatic reading span task developed by Unsworth et al. (2005). Participants took the French test and rated their attitude to MT in the context of a post-task questionnaire, which also had questions on their educational and professional background.

It seemed important to observe participants' behaviour in a range of conditions, so news articles that differed in translating difficulty and machine translations that differed in quality were selected for the study. Translating difficulty was estimated based on textual indicators used for this purpose in previous research, such as readability metrics, number of non-literal expressions, and word frequency (Hvelplund 2011; Jensen, Sjørup, and Balling 2009). The level of quality of the machine translations was estimated based on the Meteor automatic evaluation metric (Denkowski and Lavie 2011), which assesses the similarity between the raw MT output and human reference translations. The reference translations used to compute the Meteor scores were professional translations available in the WMT corpus.

The MT sentences selected for the study were taken from the WMT corpus and from a pool of additional systems.¹ The selection was performed in a way that maximised the range of Meteor scores by combining sentences from different systems in the same text. Meteor scores vary between 0 (no match between MT and reference translation) and 1 (perfect match). The MT sentences in the study sample had Meteor scores ranging between 0.14 ("low quality") and 1 ("high quality"), which means that the sample comprised nearly the entire Meteor range. The analysis was based on post-editing process data corresponding to a total of 41 sentences (844 source words).

Participants carried out the tasks in PET (post-editing tool) (Aziz, Castilho, and Specia 2012), on a computer connected to a Tobii X120 eye tracker. They were told to carry out the task in as little time as possible while aiming for final texts of high quality. The texts were edited on a sentence-by-sentence basis and with no backtracking, which was required to make sure that the eye-tracking data was reliable. Based on results from a less restrictive task involving the same materials and a participant sample with a similar profile, it was observed that these operating conditions did not have a large effect on the amount of effort required by the sentences, so these conditions are not considered here to be significantly harmful (see Vieira 2016:154-155).

Participants were instructed not to consult external sources during the task as again this could affect the quality of the eye-tracking data. However, before each post-editing session, participants read an informative description of the subject matter of the texts to be post-edited (namely, prostate cancer screening and the US voting system).

This served to mitigate the restriction on external sources and to iron out any potential discrepancies between participants in terms of subject-matter knowledge. After confirming each sentence, participants rated subjective cognitive effort using a scale largely employed for measuring mental effort in educational psychology. The scale ranges between 1 (“very, very low mental effort”) and 9 (“very, very high mental effort”) (Paas 1992). Participants selected points on the scale in PET’s interface in between post-editing the sentences.

3.2. *Post-Editors*

The sample of participants comprised professional translators, postgraduate translation students, final-year Modern Languages students who had received translation training, and non-professionals who had an educational background in translation or who were in the process of starting out professionally. The ethics committee at Newcastle University, where the post-editing tasks were conducted, reviewed the participant recruitment procedure. Participants with a diverse profile took part in the study. They were sampled from the University’s student population and from professional networks based in the local area. Their average age was 31.5 (SD = 14.5). Some of them had a low level of proficiency in French, so these participants were expected to post-edit the texts with less recourse to the ST. While the study could have had a more homogenous sample, participant variables were statistically controlled for in the analysis. This made it possible to investigate how source-language knowledge and other participant attributes were related to post-edited quality, which would not have been possible without participant variability. All participants were nevertheless native speakers of English; their profile is presented below in Table 1.

Table 1. Participants’ profile. UG: undergraduate students; PG: postgraduate students; NP: non-professionals; P: fully fledged professionals.

	WMC (0-75)	French (0-100)	Prof. Experience (in years)	Profile	Attitude to MT (1-5)
P01	25	79	1.5	PG	5
P02	42	95	-	UG	2
P03	56	13	-	UG	4
P04	14	88	-	NP	4
P05	36	95	-	NP	5
P06	56	50	< 1	NP	2
P07	44	97	< 1	UG	4

	WMC (0-75)	French (0-100)	Prof. Experience (in years)	Profile	Attitude to MT (1-5)
P08	68	95	-	UG	4
P09	61	97	-	UG	4
P10	38	89	4	PG	3
P11	61	18	-	UG	4
P12	32	60	13	P	2
P13	37	95	3	P	2
P14	51	3	-	NP	4
P15	46	93	-	NP	3
P16	33	93	5	PG	3
P17	35	63	< 1	NP	4
P18	50	97	4	P	2
P19	39	18	10	P	5

3.3. Assessing Post-Edited Quality

TAUS’s Dynamic Quality Evaluation Framework tool² was used in the evaluation task. Having quality scores for the post-edited texts as well as for the raw MT output was important for the analysis (see section 4), so the raw MT output was also assessed.

Each sentence was rated on adequacy and fluency as per TAUS’s 1-4 (low-high) rating scales. In the context of the present study, fluency measures the extent to which “the translation is well formed grammatically, contains correct spellings, adheres to common use of terms, titles and names, is intuitively acceptable and can be sensibly interpreted by a native speaker” (TAUS 2013; LDC 2005). Adequacy reflects “how much of the meaning expressed in the gold-standard translation or the source is also expressed in the target translation” (TAUS 2013; LDC 2005). The judges received an assessment brief that included an explanation of the scales with commented scoring examples. They also read the introductory texts given to post-editors with information on the subject matter.

The assessment task was designed in a way that allowed both the post-edited texts and the raw MT output to be blindly assessed. The target sentences were scrambled to make it more difficult for the judges to identify which version they were assessing (i.e. if the raw MT output or a post-edited sentence). The assessment was nonetheless carried out in document order, by combining sentences from different sources (i.e. human post-editors/MT) in the same document. Source and target sentences were presented on screen as shown in Figure 1.

[Home](#)

Batch_A

Source (French (Canada))

Previous De quoi faire hésiter encore plus les hommes déjà réticents à passer des tests de dépistage.

Current Passer le test ou non?

Next Nous avons demandé l'avis de deux spécialistes.

Target (English (United Kingdom))

Previous What makes people hesitate yet many men have already taken the prostate tests.

Current Take the test or not?

Next We have asked the opinion of two specialists.

Adequacy:

☐ None
 ☐ Little
 ☐ Most
 ☐ Everything

Fluency:

☐ Incomprehensible
 ☐ Disfluent
 ☐ Good
 ☐ Flawless

Comments

Characters left: 500

Figure 1. TAUS Dynamic Quality Evaluation interface

The fact that the target sentences were scrambled could have led to the presence of awkward cohesive links between the sentences, but scrambling the target versions seemed desirable as a way of avoiding the halo effect, whereby aspects of a single sentence could have influenced the assessment of all sentences in the same text. This could have happened because the judges could notice that they were rating machine translations and mark down all sentences in a text even when some of them were perfect, or because a single unsuccessfully post-edited sentence produced by a participant could have influenced the assessment of all of her/his other sentences.

The sentences were grouped into six batches.³ The judges were asked to rate each batch in a single sitting, working through two batches per day, on three consecutive days. The order of batches was counterbalanced between judges, who could take an unlimited break between batches that were supposed to be completed on the same day. The judges were instructed to assess fluency first, without looking at the French source, and only then to assess adequacy.

Initially, five translators were hired via the oDesk⁴ directory to act as quality judges in the study. However, only data provided by three judges was retained in the analysis, as the assessment carried out by two judges was not deemed reliable. Each batch had two mismatched sentence pairs where the content of the translation deliberately differed from that of the source. One judge was excluded from the study because top adequacy scores were given to such cases. The other excluded judge added comments to certain sentences identifying fluency errors, but failed to consider these errors in the

assessment, rating these cases with the maximum fluency score. This was also regarded as a sign of an unreliable assessment.

It should be noted that the judges spoke different varieties of English: they were from Canada (Judge 1), the United States (Judge 2), and the United Kingdom (Judge 3). Although this was not anticipated as a problem, in retrospect it would have been desirable for all judges to be speakers of British English, for consistency with the post-editors. However, any between-judge variability was statistically controlled for in the study with a view to reaching a wider consensus (see Section 4.3), so this was not deemed a significant issue. The judges were asked to complete the same French test given to post-editors. In addition, in the process of setting up their online profiles they had taken French-to-English translation skills tests made available to oDesk professionals as a way of displaying their skills to potential clients. The judges retained in the study had French scores of 93/100 (Judge 1), 100/100 (Judge 2) and 89/100 (Judge 3), and translation skills scores (i.e. obtained in the oDesk test) of 4.6/5 (Judge 1), 4.5/5 (Judge 2) and 4.9/5 (Judge 3). All judges had at least an undergraduate degree in French; Judge 2 also had a postgraduate translation degree.

3.4. *Post-Editing Effort Data*

Post-editing effort was estimated based on five specific measures: eye fixation⁵ count, average fixation duration, post-editing time, subjective cognitive effort, and Human-Targeted Translation Edit Rate (HTER), which is a score that estimates the minimum number of changes necessary to transform the raw MT output into the post-edited sentence (Snover et al. 2006).

The screen was recorded with Tobii Studio (v3.1). Tobii VT-I fixation filter (Tobii Technology 2012) was used to process the gaze data. The filter was set to discard fixations below 100 milliseconds (ms). The time post-editors spent on each sentence was identified as a scene in the screen recording. Participants' eye fixation count and duration were automatically extracted from each scene with Tobii Studio. The total duration of scenes was the total post-editing time for the sentences. To control for sentence length effects, fixation count and post-editing time for each sentence were normalised by the number of source-sentence characters. Average fixation duration was calculated by dividing the sentences' total fixation duration by the corresponding fixation count.

HTER provided an indication of the number of changes post-editors implemented. This score is computed by dividing the total number of estimated editing operations (insertions, deletions, substitutions, and shifts) by the number of tokens in the post-edited sentences. Higher HTER scores indicate larger differences between the machine-translated and the post-edited sentences. Zero is the lowest possible score, i.e. when no changes are implemented. The HTER scores provided in the PET log files were used in the analysis.

Because of accidental task errors (e.g. clicking on the wrong buttons in the editing interface), one sentence post-edited by P14 and four sentences post-edited by P19 were

excluded from the sample. Per-sentence total fixation time and average fixation duration were used to estimate the quality of the gaze data. Based on findings from previous research on the typical length of fixations (Rayner 1998), data points corresponding to sentences where average fixation duration was below 200 ms were removed from the dataset (8%). Additional data points where fixation time on screen was considered to take up too low a percentage of editing time (2.5 standard deviations or more from the sample mean) were also excluded (0.9%). After these exclusions, a total of 708 data points (post-edited sentences) remained.

The variables described above reflect different facets of effort in post-editing (see section 1). HTER can be seen as an estimate of technical effort; post-editing time is a direct measure of temporal effort; and subjective scores, average fixation duration and fixation count are estimates of cognitive effort (see Rayner 1998; Paas 1992; O'Brien 2011).

4. Results and Discussion

4.1. Fluency and Adequacy

The average of fluency and adequacy scores provided by all three judges for the sentences is presented in Figure 2. As can be seen, post-editing improved the MT output, but did not make it perfect. Average scores for the post-edited sentences were below 4 on both fluency and adequacy. Slightly higher scores are observed for adequacy, however, suggesting that MT systems as well as post-editors might perform slightly better on translation accuracy than on translation fluency.

Figure 2. Average quality scores (y-axis) for machine-translated and post-edited sentences (x-axis)

Table 2. Examples of post-edited fluency and adequacy errors and scores; ST: source text; REF: human reference; MT: raw MT output; PE: post-edited version

Comments and Scoring Examples (Judge 1)	
ST	On va leur suggérer de faire de la surveillance active et si la maladie progresse, on va leur offrir un traitement.
REF	Active monitoring will be suggested, and if the disease progresses, they will be offered treatment.
MT	We will suggest active monitoring and if the disease progresses, will provide a treatment.
PE	It is suggested to them to actively monitor the disease and if the disease progresses, treatment will be provided.
Comment: “awkward passive voice use” Fluency: 3; Adequacy: 4	
ST	Les dirigeants républicains justifiaient leur politique par la nécessité de lutter contre la fraude électorale.
REF	Republican leaders justified their policy by the need to combat electoral fraud.
MT	The Republican leaders justified their policies by the need to combat electoral fraud.
PE	The Republican leaders justified their policies by stressing the need to combat electoral fraud.
Comment: “‘stressing’ is added to the translation” Fluency: 4; Adequacy: 3	

Although the judges were not instructed to mark errors in the translations, they could leave any comments in a text box in the assessment interface. Judge 1 was a frequent user of this text box, so this judge’s comments are taken into account here to exemplify typical fluency and adequacy errors, shown in Table 2. Based on comments left by Judge 1, typical fluency errors in the post-edited sentences involved issues with prepositions, articles, verbs, word order, and punctuation. Typical adequacy errors occurred mainly because of verbal constructions deemed to alter the meaning of the ST, or problems with the omission or addition of concepts/words.

Scores of post-edited fluency and adequacy were found to be weakly to moderately correlated (Judge 1: $p(706) = 0.23$, $p < 0.001$; Judge 2: $p(706) = 0.32$, $p < 0.001$; Judge 3: $p(706) = 0.42$, $p < 0.001$).⁶ Callison-Burch et al. (2007) observed high correlations between fluency and adequacy scores applied to machine translations, which led them to suggest that the difference between these two constructs may not exist in practice. However, the correlations between fluency and adequacy observed here are moderate at best, which is consistent with the assumption that a translation may convey the source meaning adequately, but disfluently. It may be that in Callison-Burch et al. (2007) both these components were rated at the same time, which can inflate the halo

effect, whereby one assessment influences the other. In the present study, the fact that the judges were expressly instructed to assess fluency first and only then to look at the source might have lessened this effect. This is in line with the assessment strategy proposed by Daems, Macken, and Vandepitte (2013), constituting a procedure that should probably be adopted in future research, as the results obtained here show that fluency and adequacy constitute different constructs not only conceptually, but also empirically.

4.2. Rater Agreement

Krippendorff's alpha (α) was used to measure inter- and intra-rater agreement. This score allows nominal variables to be treated as ordinal. Traditional agreement scores that can be used with more than three raters, such as Fleiss's κ , only consider perfect agreement, i.e. they are not sensitive to disagreements with different levels of severity. Krippendorff's α takes into account the fact that a disagreement corresponding to a difference between 1 and 4 is more severe than a disagreement corresponding to a difference between 3 and 4, for example, so this score was chosen.

A small quantity of duplicate sentences (39) was used to calculate intra-rater agreement, i.e. how consistent the judges were in their own assessment. An α of 0 corresponds to chance agreement, and an α of 1 corresponds to perfect agreement. These results are shown in Table 3.

Inter-rater agreement was substantially higher for adequacy. The agreement level for fluency is marginally above chance, which suggests that assessing adequacy is more straightforward. This may be because adequacy is defined as the proportion of meaning overlap between the source and target sentences, which might make adequacy easier than fluency to quantify. Fluency seems more subjective, which means that the judges would need to go through a more extensive training phase if higher agreement was to be achieved. This is not a problem in the context of the present study, however, as discrepancies across judges are factored into the analysis.

Table 3. Intra- and inter-rater agreement

	Inter-Rater Agreement		Intra-Rater Agreement	
	Fluency	Adequacy	Fluency	Adequacy
α	0.171	0.519	0.762	0.636

4.3. *Post-Editing Effort and Post-Edited Quality*

To provide an initial idea of how effort and product quality interconnect in post-editing, the relationship between the post-editing effort measures described in section 3.4 and post-edited quality is presented in Figures 3 and 4.

The graphs' x-axes show the post-editing effort measures, and the y-axes show post-edited quality – fluency in Figure 3, and adequacy in Figure 4. Since the judges also assessed the raw MT output, these relationships are displayed for each level of MT quality, represented by the lines on the graphs. The lines were fitted with the non-parametric loess method (local polynomial regression), which is capable of showing non-linear relationships. The graphs are based on pooled data from all judges and post-editors.

Figure 3 suggests that subjective cognitive effort (top right pane) and average fixation duration (middle left pane) are not related to post-edited fluency. This is shown by the mostly flat lines for these variables, observed at all levels of raw MT fluency. Complex results are observed for HTER (top left pane), which has an overall positive association with post-edited fluency at MT fluency levels 1-3, but a negative association at MT fluency level 4 (i.e. the top level of MT quality). This suggests that when MT fluency was already high, modifying the MT output was not indifferent, but rather detrimental to post-edited quality. Fixation count (bottom left pane) and post-editing time (middle right pane) had mostly negative associations with post-edited fluency, which is observed particularly at MT levels 1 and 4.

Figure 3. Relationship between post-edited fluency (y-axes) and measures of post-editing effort (x-axes). Each line is a level of raw MT fluency. The line shades are 95% confidence intervals; wider shades correspond to fewer data points.

Figure 4. Relationship between post-edited adequacy (y-axes) and measures of post-editing effort (x-axes). Each line is a level of raw MT adequacy. The line shades are 95% confidence intervals; wider shades correspond to fewer data points.

Figure 4, which displays results for adequacy, shows patterns relatively similar to those observed for fluency, though with notable differences for HTER (top left pane). While HTER is positively related to post-edited adequacy at level 2 of MT adequacy, Figure 4 indicates an overall negative effect of HTER at MT adequacy levels 1 and 4, and a neutral effect at level 3. This is particularly interesting at MT adequacy level 1, where a non-linear effect is clearly observed, showing that only HTER scores below 0.5 (corresponding to occasions where the MT output is only slightly modified) were positively related to post-edited adequacy. When HTER gets higher, this effect becomes negative, suggesting that drastic changes in the MT output at this quality level introduced adequacy errors in the translations.

To factor into the analysis any differences between judges, post-editors, and sentences, the patterns illustrated above were further analysed with mixed-effects regression models. These models control for the effect of random variables sampled for the investigation. In the present study, these are ST-MT sentence pairs, post-editors, post-edited sentences, and the judges.⁷ Treating these variables as random effects allows the dependence between observations to be taken into account (i.e. the fact that the data has multiple observations corresponding to these elements).

The models were fit with the `clmm` R function of the `ordinal` package (Christensen 2010). Ordinal scores of fluency and adequacy were the outcome variables. The measures of post-editing effort described in section 3.4 were tested as predictors. Since the scale of subjective cognitive effort is a factor with a large number of levels (9), and is used as a predictor, rather than as a response variable, this score was treated as numeric.⁸ HTER, post-editing time and fixation count were positively skewed (i.e. the majority of the data concentrated on the low side of the scale), so these measures were transformed by taking their square root to avoid harmful outlier effects. Post-editor numeric variables (presented in Table 1) and the scores given by the judges to the raw MT output were also added to the models as predictors. All numeric predictors were standardised by subtracting the mean and dividing by one standard deviation. Moderate to high levels of correlation were observed between the post-editing effort measures, but multicollinearity was not found to be an issue. The overall collinearity of the measures had a condition number of 11, which is considered acceptable (see Baayen 2008:181ff).

The model fitting procedure consisted of adding new variables and removing non-significant ones one at a time. The impact of MT quality and post-editor variables was checked first. Once the effect of these variables had been taken into account, the impact of the post-editing effort measures was tested. Potential non-linear effects were also tested by including quadratic terms in the models – i.e. squared versions of the variables themselves. In addition, post-editing effort measures that remained in the models were allowed to interact with MT quality. An interaction occurs when the relationship between a predictor and the outcome variable is moderated by other predictors. It seemed plausible that the relationship between post-editing effort and post-edited quality could be moderated by the quality of the raw MT sentences, so interactions in this respect were tested.

The results obtained in the models are presented in Table 4, where “MT Fl” and “MT Ad” correspond to raw MT fluency and adequacy, respectively. “HTER2” is the quadratic term of HTER. “Fix. Count” corresponds to the normalised measure of fixation count, and “French” corresponds to participants’ scores in the French test.

Table 4. Mixed-effects results for fluency (left) and adequacy (right)

Fluency				Adequacy			
	sig.	β	z		sig.	β	z
MT Fl 2	••	1.5	3.23	MT Ad 2		0.34	1.57
MT Fl 3	•••	2.68	5.02	MT Ad 3	•	0.78	2.49
MT Fl 4	•••	3.68	6.21	MT Ad 4	•••	1.90	5.38
HTER	•••	3.81	4.49	French	••	0.41	3.12
HTER2	••	-1.51	3.12	Fix. Count	•••	-0.89	6.52
Fix. Count	•••	-0.72	5.32	-	-	-	-
MT Fl 2:HTER	•	-1.05	2.18	-	-	-	-
MT Fl 3:HTER	••	-1.83	3.06	-	-	-	-
MT Fl 4:HTER	•••	-4.02	6.16	-	-	-	-

MT Fl: MT fluency/ **MT Ad:** MT adequacy/ **French:** score in French vocabulary test/ **Fix. Count:** fixation count per ST character (square root) / **HTER:** HTER (square root)/ **HTER2:** quadratic term of HTER
Non-significant predictors removed from the models: professional experience, working memory capacity, attitude towards MT, post-editing time, subjective cognitive effort and average fixation duration
All numeric variables were z-standardised.
Observations: 2124/ Median fluency: 3/ Median adequacy: 4
••• p < 0.001 •• p < 0.01 • p < 0.05

Unsurprisingly, MT quality had a strong positive effect in the models. As raw MT fluency and adequacy are represented by ordinal scores, the coefficient (β) for these variables should be interpreted as the impact of each upper level (MT Fl/Ad 2-4) relative to level 1 (by default not displayed since it acts as the reference). As can be seen in Table 4, there are significant differences between all upper levels of MT fluency and level 1 (MT Fl 2: $\beta = 1.5$; MT Fl 3: $\beta = 2.68$; MT Fl 4: $\beta = 3.68$), which means that when MT fluency increases from level 1 to the other levels, post-edited fluency also significantly increases. A similar pattern is observed for adequacy, but the difference between MT adequacy levels 1 and 2 (MT Ad 2: $\beta = 0.34$) was not found to be significant. Another unsurprising effect is the positive impact of French knowledge on

post-edited adequacy ($\beta = 0.41$). This indicates that the post-edited texts produced by those with higher scores in the French test had fewer adequacy errors.

As for post-editing effort measures, two variables were found to be significant: HTER and fixation count. The effect of these variables is discussed below in detail.

HTER was found to be significant in the fluency model in an interaction with raw MT fluency. HTER was also found to have an overall non-linear effect (HTER2), which means that, irrespective of the MT quality level, higher HTER only corresponds to higher post-edited fluency up to a certain point; as HTER values carry on increasing (i.e. indicating more radical modifications in the MT output), post-edited fluency decreases. The model's main term for HTER ($\beta = 3.81$) holds for MT fluency at level 1. The effect of HTER at other MT levels can be obtained by subtracting the interaction coefficients from the main coefficient. Specifically, the effect of HTER at MT level 2 is $\beta = 2.76$ ($3.81 - 1.05$). At level 3, the effect gets smaller: $\beta = 1.98$ ($3.81 - 1.83$). At level 4, the effect is inverted: $\beta = -0.21$ ($3.81 - 4.02$). These results show that the relationship between editing operations and product quality depends on the level of quality of the raw MT output. This complex relationship may be one of the reasons why Carl et al. (2011) did not observe this association.

This interaction suggests that, for higher-quality MT, making a large number of changes in the raw MT output can be damaging, which is consistent with industry guidelines for post-editing (TAUS/CNGL 2010). Even in settings where higher levels of quality are required, these guidelines advise making use of as much of the MT output as possible. In most situations, the principle of not changing more than necessary would probably be expected to increase post-editors' productivity. Interestingly, these results indicate that this principle may also be linked to higher levels of post-edited quality. While the advice to keep changes to a minimum is not new (see e.g. Guzmán 2007), to the knowledge of the present author the results presented here for the first time show empirical evidence supporting this guideline while drawing a parallel with the quality of the raw MT output.

To investigate further the effect of HTER, individual data points were examined to check to see if any special cases or outliers could be driving the effect of this variable. Based on modal scores⁹ of MT fluency, it was noted that a single post-edited sentence remained at fluency level 1 after post-editing had been carried out. Excluding this sentence from the data did not significantly change the results in the fluency model apart from weakening the effect of HTER at MT level 2, so this data point was retained. Similarly, it was checked if excluding low-quality MT sentences from the materials (sentences scored by all judges with level 1 on fluency and/or adequacy) would significantly alter the results. This was checked for both models, and no significant differences were observed; trends were the same and all variables remained significant.

Again based on modal scores, it was noted that only 13 sentences went from level 4 of MT fluency down to level 3 after post-editing had been carried out, and just one sentence went from level 4 down to level 2. This indicates that most MT sentences deemed by the judges to be at the top level of fluency simply remained at this level after being post-edited. A potential outlier effect of the only sentence going down from

fluency level 4 to level 2 was tested by excluding it from the analysis, but results were not significantly altered. However, the small number of post-edited sentences where MT fluency was not kept at level 4 suggests that the occurrences supporting the negative effect of excessive edits for top-quality MT are relatively rare. In any case, this tendency can be observed across the data, since the impact of HTER becomes weaker as MT fluency increases.

To illustrate the decreases in fluency observed in the study, the only sentence going from level 4 of MT fluency down to level 2 and one of the 13 sentences that went from level 4 down to level 3 are presented below in examples (1) and (2), respectively.

- (1) **ST:** Passer le test ou non?
REF: Take the test or not?
MT: Take the test or not?
PE (P10): Carry out tests or not?
- (2) **ST:** D'ailleurs, les avocats républicains n'ont recensé que 300 cas de fraude électorale aux États-Unis en dix ans.
REF: Republican lawyers identified only 300 cases of electoral fraud in the United States in a decade.
MT: Indeed, Republican lawyers have identified only 300 cases of electoral fraud in the United States in a decade.
PE (P13): Also, Republican lawyers only identified 300 cases of electoral fraud in the United States within ten years.

The sentence in example (1) is quite short. While HTER is normalised by number of tokens, this could be reflecting a penalisation of extremely short sentences, where any modifications in the MT output will inevitably amount to a large part of the sentence in relative terms – an effect consistent with a pattern observed by O'Brien (2011) and Vieira (2014) for a similar score. However, excluding all occurrences of this sentence from the data did not change results, and the effect still holds for longer sentences, as seen in example (2). Interestingly, both of these sentences were post-edited by participants who had professional experience and a high level of French knowledge (see Table 1).

Fixation count was the only other post-editing effort variable found to have a significant effect on post-edited quality. This variable had negative effects in both models ($\beta = -0.72$ for fluency; $\beta = -0.89$ for adequacy).

With regard to results in the adequacy model, it is safe to assume that higher levels of post-edited adequacy presuppose an accurate understanding of the ST, as indicated by the positive effect that French proficiency had in this model. In view of this, it seemed relevant to check if fixations corresponding just to the ST would present a different relationship with post-edited adequacy. When examining this, however, results similar to those presented in Table 4 were obtained both in terms of absolute ST fixation count and in terms of a ratio of ST fixations to all fixations. This rules out the

possibility that the patterns observed for adequacy are different when only ST fixations are considered.

Overall, the negative effect of fixation count is quite surprising. An intuitive result in this regard would be for a larger number of fixations to indicate a larger amount of attention, and consequently a more accurate and fluent post-edited product. A large number of fixations with no positive effect on the product seems like a sign of cognitive overload, where the task proves too difficult for post-editors to achieve improvements in quality. It is however interesting to note that the positive effect of HTER (at MT fluency levels 1-3) and the negative effect of fixation count are both significant in the fluency model, which means that both these trends occur at the same time. This suggests that a dense editing behaviour, where edits take place amid few fixations, can have a positive effect on post-edited fluency. Fixation count is strongly correlated with post-editing time ($\rho(706) = 0.96, p < 0.001$), so in other words this indicates that “quick edits” (i.e. when modifications are implemented in short intervals of time) can have a positive effect on the quality of post-edited texts.

Of MT sentences scored with a modal score of 2 on fluency, for example, corresponding post-edited sentences that did not improve beyond MT quality and remained at level 2 have an average HTER of 0.22 and average fixation count of 1.23 per character. By contrast, post-edited versions corresponding to the same MT sentences (i.e. post-edited by other participants) which went up to post-edited level 4 have an average HTER of 0.29 and average fixation count of 0.95 per character (i.e. slightly higher HTER and yet lower fixation count compared to their less successful counterparts). This illustrates the effects observed in the fluency model, where HTER is for the most part positively associated with post-edited quality while fixation count presents a negative effect.

To illustrate this further, example (3) shows a machine translation with a modal score of 2 on fluency together with two corresponding post-edited sentences where fixation count and HTER values were particularly different. As can be seen, P05’s interventions in this MT sentence were slightly more radical and led to a higher level of post-edited fluency compared to P07’s. Interestingly, not only did P05 modify this sentence slightly more (e.g. by turning “disproportionate” into an adverb and bringing it to the beginning of the sentence), but she also did so in less time. Furthermore, it is interesting to note that these two participants have a similar profile (see Table 1), which suggests that these differences in behaviour are not explained by, for example, their knowledge of French or level of professional experience alone.

- (3) **ST:** Les nouvelles restrictions affectent de manière disproportionnée les jeunes, les minorités, et les personnes à faible revenu.

REF: The new restrictions disproportionately affect young people, minorities and people with low incomes.

MT: The new restrictions allocate the youths, the minorities, and the people in a disproportionate way to weak income.

PE (P05): The new restrictions disproportionately affect young people, minority groups and people on a low income. [HTER: 0.76; 81 fixations; fluency: 4; 53 seconds]

PE (P07): The new restrictions affect young people, minorities and people of low-income in a disproportionate way to the rest of citizens. [HTER: 0.5; 109 fixations; fluency: 3; 65 seconds]

It is worth noting that a negative link between post-editing time and post-edited quality has been observed in previous post-editing research, which is consistent with the pattern observed here. Green et al. (2014) suggested that the negative association between time and quality observed in their study resulted from the fact that the score they used to measure quality is known to penalise long sentences. Since longer sentences are expected to require more post-editing time, Green et al. argued that this negative association was an artefact of this score, rather than a true negative effect of time on quality. However, the correlations between post-editing time and sentence length they reported are at best moderate: 0.53 for French-English, and 0.43 for English-German. This suggests that editing time does not act as a perfect substitute for sentence length, so there might be more to the negative link between time and quality than just a methodological artefact. Similarly, Mitchell (2015) suggested that the negative correlation she observed between post-editing time and post-edited adequacy was an indirect effect of the low level of source-language proficiency of certain participants. While this is a plausible explanation for this result, source-language proficiency was controlled for in the present study. Even by considering only participants whose French scores were above the sample average (70.4) and refitting the models shown in Table 4, a significant negative link between fixation count and post-edited quality was observed here, which again suggests that there might be more to the negative link between time and quality than just a methodological issue.

Interestingly, previous research in cognitive psychology highlights that the incidence of a large number of automatic processes in a task may account for a dissociation between performance quality and cognitive effort (as measured by subjective ratings) (Gopher and Donchin 1986:24). This is consistent with the patterns presented in Figures 3 and 4, which show that subjective cognitive effort and average fixation duration, for example, are not linked to post-edited quality. A mental process is automatic when it does not rely on working memory and takes place without cognitive control (see Norman and Shallice 1986), i.e. when it is based on an intuitive reaction. As these processes are expected to be quicker (i.e. because they do not involve “thinking time”), it is hypothesised here that a high incidence of automatic processes could be one of the reasons behind the positive effect of a fast post-editing behaviour discussed above.

Indeed, research in decision-making proposes that automatic processes might play a more prominent role in certain tasks. It is usually assumed in this area that decisions are made based on a dual model that involves a deliberate and a tacit system (Hogarth 2005), i.e. systems based on analytical thought and on intuitive reactions, respectively. In the context of analytically complex tasks (assumed here to be the case of post-editing¹⁰), Hogarth suggests that when the individuals who carry out the task are qualified, and the chance that they will make errors in acting by intuition is between medium and low, the tacit system should be more effective than the deliberate one. Since all participants in the present study were native speakers of the target language, and had received prior training in translation, it seems reasonable to assume that their chance of making incorrect intuitive decisions is low, or perhaps medium in the case of participants with a low level of proficiency in French. This should be the case of most professional translators/post-editors, so it may be that having intuitive reactions to MT errors might constitute a more effective way of carrying out post-editing. In view of results presented in this chapter, this seems like a promising hypothesis to be explored in future post-editing research.

5. Conclusion

5.1. Research Findings

The main purpose of this chapter was to investigate potential connections between post-edited quality and the post-editing process. Post-edited quality was assessed in terms of fluency and adequacy, and, contrary to previous assumptions, the chapter provided evidence to suggest that these two constructs can be empirically distinguished. Furthermore, the analysis showed that in most circumstances editing operations are positively related to post-edited fluency. However, it was noted that for high-quality MT this trend gets weaker and even inverts, i.e. product fluency can decrease if too many edits are performed. In addition, it was observed that the number of fixations on the text (which correlates with post-editing time) had an overall negative association with both the fluency and adequacy of the post-edited product.

In practical terms, the results presented above also indicate that post-editing behaviour is a better predictor of post-edited quality than post-editors' individual attributes. The only aspect of post-editors' profile that had a significant effect on post-edited quality was source-language proficiency, which was unsurprisingly related to the adequacy of the post-edited sentences. This suggests that focusing on post-editors' actions, rather than their educational or professional background, might be a more promising approach to investigating post-editing performance.

As for the effects observed for HTER and fixation count, these results have two main implications. First, the number of edits implemented in the MT output should be kept to the absolute minimum required to improve the translation to the desired level, as failing to do so might compromise product quality. This was found to be the case

especially when edits were excessive and the quality of the raw MT output was already high. Second, edits carried out in short intervals of time were associated with high levels of efficacy in post-editing, which in the light of research in decision-making and cognitive psychology is interpreted here as a sign that more emphasis should be placed on the potential benefits of automatic processes. Hogarth (2005:80) argues that “people need to be more aware of how often they allow themselves to take decisions automatically”. In view of the results presented in this chapter, this is suggested here as possibly a key factor in further understanding the inner workings of post-editing behaviour.

5.2. *Suggestions for Future Research*

It would be interesting for future research to check to see how a whole-text assessment of translation quality can shed light on the relationship between post-editing effort and post-edited quality observed here. An analysis of this kind would allow aspects of textuality that could not be considered in this chapter to be taken into account, such as paragraph development and document consistency.

More generally, many previous studies have focused on simply quantifying differences between post-editing and traditional translation, which is perhaps justified by the commercial interests implied. However, with a few exceptions, conceptual comparisons of these and other activities are lacking in the field. Little empirical evidence is available to show how post-editing is different from translation or traditional revision (if at all) in terms of the thought processes these activities involve and the skill sets they require. Results presented here suggest that post-editing involves a large number of non-linear relationships, which seems to stem from the fact that too little and too much editing are both detrimental to product quality. Since translation involves editing the emerging text (see Shih 2006), it may be that some of these results apply also to translation self-revision and other types of translation editing, but these are questions left for future research to explore.

Notes

¹The additional systems that entered the study sample were SDL FreeTranslation.com (<http://www.freetranslation.com>), TransPerfect (<http://web.transperfect.com/free-translations/>) and Microsoft Translator, via MS Word (all of which obtained in October 2013).

² See <https://www.taus.net/evaluate/dqf-tools> [accessed 26 April 2017].

³ The batches included sentences produced in a separate task that is not dealt with here (see Vieira 2016).

⁴ This is a directory of freelance professionals, which is now called UpWork; see <https://www.upwork.com/> [accessed 01 August 2016].

⁵ Fixations are “eye movements that stabilize the retina over a stationary object of interest” (Duchowski 2007, 46), i.e. when the eyes focus on an object, allowing the mind to process it.

⁶ On some occasions, correlated outcome variables require the use of multivariate modelling (i.e. when more than one outcome variable is included in the same statistical model). However, the added power of these models is only appreciable when high correlations are observed (see Snijders and Bosker 1999:201), so following this approach was not deemed necessary (see section 4.3).

⁷ Each post-edited sentence received an individual label even when different post-editors happened to produce identical sentences. This is not problematic, as the model was not used to measure intra-rater agreement, which renders optional the use of this factor. This factor did not affect the nature or significance of the results.

⁸ This is statistically less complex than treating this score as ordered categorical ratings, and was not found to affect the results.

⁹ That is, the most frequent score among the judges; the average was taken instead when all three judges provided different scores.

¹⁰ As also assumed for translation (de Groot 2000).

Acknowledgments

The author would like to thank Dr Francis Jones, Dr Michael Jin, Dr Ya-Yun Chen, and Dr Peter Flom for their comments on an earlier draft of this chapter. Thanks also go to TAUS for enabling use of the Dynamic Quality Evaluation Framework tool in the study.

References

- Aziz, Wilker, Sheila Castilho, and Lucia Specia. 2012. “PET: a Tool for Post-Editing and Assessing Machine Translation.” In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC 2012)*, 21-27 May 2012, Istanbul, Turkey, ed. by Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Mehmet Uğur Doğan, Bente Maegaard, Joseph Mariani, Jan Odijk and Stelios Piperidis, 3982-3987.
- Baayen, R. Harald. 2008. *Analysing Linguistic Data: a Practical Introduction to Statistics*. Cambridge: Cambridge University Press.
- Callison-Burch, Chris, Cameron Fordyce, Philipp Koehn, Christof Monz, and Josh Schroeder. 2007. “(Meta-) Evaluation of Machine Translation.” In *Proceedings of the Second Workshop on Statistical Machine Translation*, Prague, Czech Republic, 23 June 2007, 136-158.
- Carl, Michael, Barbara Dragsted, Jakob Elming, Daniel Hardt and Arnt L. Jakobsen. 2011. “The Process of Post-Editing: a Pilot Study.” In *Proceedings of the 8th International NLPCS Workshop: Human-Machine Interaction in Translation*, ed. by Bernardette Sharp, Michael Zock, Michael Carl and Arnt L. Jakobsen, 131-142. Copenhagen: Samfundslitteratur.

- Christensen, Rune Haubo B. 2010. *ordinal: Regression Models for Ordinal Data*. R Package Version no. 22.
- Daems, Joke, Lieve Macken, and Sonia Vandepitte. 2013. "Quality as the Sum of its Parts: a Two-Step Approach for the Identification of Translation Problems and Translation Quality Assessment for HT and MT+PE." In *Proceedings of the MT Summit XIV Workshop on Post-Editing Technology and Practice (WPTP2)*, 2–6 September 2013, Nice, France, ed. by Sharon O'Brien, Michel Simard and Lucia Specia, 63-71.
- de Almeida, Giselle. 2013. *Translating the Post-Editor: an Investigation of Post-Editing Changes and Correlations with Professional Experience across Two Romance Languages*. PhD Thesis, Dublin City University.
- de Groot, Annette M. B. 2000. "A Complex-Skill Approach to Translation and Interpreting." In *Tapping and Mapping the Processes of Translation and Interpreting: Outlooks on Empirical Research*, ed. by Sonja Tirkkonen-Condit and Riitta Jääskeläinen, 53-68. Amsterdam: John Benjamins.
- Denkowski, Michael, and Alon Lavie. 2011. "Meteor 1.3: Automatic Metric for Reliable Optimization and Evaluation of Machine Translation Systems." In *Proceedings of the Sixth Workshop on Statistical Machine Translation (EMNLP 2011)*, 30-31 July 2011, Edinburgh, UK, 85-91.
- Depraetere, Ilse, Nathalie De Sutter, and Arda Tezcan. 2014. "Post-Edited Quality, Post-Editing Behaviour and Human Evaluation: a Case Study." In *Post-Editing of Machine Translation: Processes and Applications*, ed. by Sharon O'Brien, Laura Winther Balling, Michael Carl, Michel Simard and Lucia Specia, 78-108. Newcastle upon Tyne: Cambridge Scholars Publishing.
- Duchowski, Andrew T. 2007. *Eye Tracking Methodology: Theory and Practice*. London: Springer-Verlag.
- Green, Spence, Jason Chuang, Jeffrey Heer and Christopher D. Manning. 2014. "Predictive Translation Memory: a Mixed-Initiative System for Human Language Translation." In *Proceedings of the 27th Annual ACM Symposium on User Interface Software and Technology*, 05-08 October 2014, Honolulu, USA, 177-187.
- Green, Spence, Jeffrey Heer, and Christopher D. Manning. 2013. "The Efficacy of Human Post-Editing for Language Translation." In *ACM Human Factors in Computing Systems (CHI)*, 27 April - 2 May 2013, Paris, France, 439-448.
- Gopher, Daniel and Emanuel Donchin. 1986. "Workload: an Examination of the Concept." In *Handbook of Perception and Human Performance*, Vol. 2, ed. by Kenneth R. Boff, Lloyd Kaufman and James P. Thomas, 1-49. New York: John Wiley and Sons.
- Guerberof, Ana. 2014. "The Role of Professional Experience in Post-Editing from a Quality and Productivity Perspective." In *Post-Editing of Machine Translation: Processes and Applications*, ed. by Sharon O'Brien, Laura Winther Balling, Michael Carl, Michel Simard and Lucia Specia, 51-76. Newcastle upon Tyne: Cambridge Scholars Publishing.
- Guzmán, Rafael. 2007. "Manual MT Post-Editing: If It's Not Broken, Don't Fix It!" *Translation Journal* 11 (4). Accessed 26 June 2014. <http://www.translationjournal.net/journal/42mt.htm>.
- Hogarth, Robin M. 2005. "Deciding Analytically or Trusting Your Intuition? The Advantages and Disadvantages of Analytic and Intuitive Thought." In *The Routines of Decision Making*, ed. by Tilmann Betsch and Susanne Haberstroh, 67-82. Mahwah, NJ: Lawrence Erlbaum Associates.

- Hvelplund, Kristian Tangsgaard. 2011. *Allocation of Cognitive Resources in Translation: An Eye-Tracking and Key-Logging Study*. PhD Thesis, Copenhagen Business School.
- Jensen, K. T. Hvelplund, Annette Sjørup, and Laura Winther Balling. 2009. "Effects of L1 Syntax on L2 Translation." In *Methodology, Technology and Innovation in Translation Process Research*, ed. by Inger M. Mees, Fabio Alves and Susanne Göpferich, 319-339. Copenhagen: Samfundslitteratur.
- Koponen, Maarit. 2012. "Comparing Human Perceptions of Post-Editing Effort with Post-Editing Operations." In *Proceedings of the Seventh Workshop on Statistical Machine Translation* (NAACL 2012), 7-8 June 2012, Montreal, Canada, 181-190.
- Koponen, Maarit, Wilker Aziz, Luciana Ramos, and Lucia Specia. 2012. "Post-Editing Time as a Measure of Cognitive Effort." In *Proceedings of the AMTA 2012 Workshop on Post-Editing Technology and Practice* (WPTP 2012), 28 October 2012, San Diego, USA, ed. by Sharon O'Brien, Michel Simard and Lucia Specia, 10pp.
- Krings, Hans P. 2001. *Repairing Texts: Empirical Investigations of Machine Translation Post-Editing Processes*, ed. by Geoffrey Koby. Kent, Ohio: Kent State University Press.
- LDC. 2005. *Linguistic Data Annotation Specification: Assessment of Fluency and Adequacy in Translations*. Revision 1.5.
- Lin, Chin-Yew and Franz Josef Och. 2004. "ORANGE: a Method for Evaluating Automatic Evaluation Metrics for Machine Translation." In *Proceedings of the 20th International Conference on Computational Linguistics*, 23-27 August 2004, Geneva, Switzerland, 7pp.
- McCutchen, Deborah. 1996. "A Capacity Theory of Writing: Working Memory in Composition." *Educational Psychology Review* 8:299-325. doi: 10.1007/BF01464076.
- Meara, Paul, and Barbara Buxton. 1987. "An Alternative to Multiple Choice Vocabulary Tests." *Language Testing* 4 (2): 142-154.
- Mitchell, Linda. 2015. *Community Post-Editing of Machine-Translated User-Generated Content*. PhD Thesis, Dublin City University.
- Norman, Donald A. and Tim Shallice. 1986. "Attention to Action: Willed and Automatic Control of Behaviour." In *Consciousness and Self-Regulation: Advances in Research and Theory*, ed. by Richard J. Davidson, Gary E. Schwartz and David Shapiro, 1-18. New York: Plenum Press.
- O'Brien, Sharon. 2011. "Towards Predicting Post-Editing Productivity." *Machine Translation* 25 (3): 197-215. doi: 10.1007/s10590-011-9096-7.
- Paas, Fred G. 1992. "Training Strategies for Attaining Transfer of Problem-Solving Skill in Statistics: a Cognitive-Load Approach." *Journal of Educational Psychology* 84 (4): 429-434.
- Plitt, Mirko, and François Masselot. 2010. "A Productivity Test of Statistical Machine Translation Post-Editing in a Typical Localization Context." *Prague Bull Math Linguist* 93: 7-16.
- Rayner, Keith. 1998. "Eye Movements in Reading and Information Processing: 20 Years of Research." *Psychological Bulletin* 124 (3): 372-422.
- Shih, Claire. 2006. "Revision from Translators' Point of View: an Interview Study." *Target* 18 (2): 295-312.
- Snijders, Tom A. B., and Roel Bosker. 1999. *Multilevel Analysis: an Introduction to Basic and Advanced Multilevel Modeling*. London: Sage.
- Snover, Matthew, Bonnie Dorr, Richard Schwartz, Linnea Micciulla and John Makhoul. 2006. "A Study of Translation Edit Rate with Targeted Human Annotation." In *Proceedings of the 7th Conference of the Association for Machine Translation of the Americas – Visions for the Future of Machine Translation*, 8-12 August 2006, Cambridge, USA, 223-231.

- TAUS. 2013. *Adequacy/Fluency Guidelines*. Accessed 17 June 2015.
<https://www.taus.net/think-tank/best-practices/evaluate-best-practices/adequacy-fluency-guidelines>.
- TAUS/CNGL. 2010. *MT Post-Editing Guidelines*. Accessed 17 June 2015.
<https://www.taus.net/think-tank/best-practices/postedit-best-practices/machine-translation-post-editing-guidelines>.
- Temnikova, Irina. 2010. "A Cognitive Evaluation Approach for a Controlled Language Post-Editing Experiment." In *Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC 2010)*, 19-21 May 2010, Valetta, Malta, 3485-3490.
- Tobii Technology. 2012. *Determining the Tobii I-VT Fixation Filter's Default Values: Method Description and Results Discussion*. Accessed 20 December 2013. http://www.tobii.com/Global/Analysis/Training/WhitePapers/Tobii_WhitePaper_DeterminingtheTobiiI-VTFixationFilter'sDefaultValues.pdf.
- Unsworth, Nash, Richard P. Heitz, Josef C. Schrock, and Randall W. Engle. 2005. "An Automated Version of the Operation Span Task." *Behavior Research Methods* 37 (3): 498-505.
- Vieira, Lucas N. 2014. "Indices of Cognitive Effort in Machine Translation Post-Editing." *Machine Translation* 28 (3-4): 187-216. doi: 10.1007/s10590-014-9156-x
- Vieira, Lucas N. 2016. *Cognitive Effort in Post-Editing of Machine Translation: Evidence from Eye Movements, Subjective Ratings, and Think-Aloud Protocols*. PhD Thesis, Newcastle University.