Cognitive Demand and Cognitive Effort in Post-Editing

Isabel Lacruz ilacruz@kent.edu
Institute for Applied Linguistics, Kent State University, Kent OH 44240, U.S.A.

Michael Denkowski mdenkows@cs.cmu.edu
Language Technologies Institute, Carnegie Mellon University, Pittsburgh PA 15213, U.S.A.

Alon Lavie alavie@cs.cmu.edu
Language Technologies Institute, Carnegie Mellon University, Pittsburgh PA 15213, U.S.A.

Abstract

The pause to word ratio, the number of pauses per word in a post-edited MT segment, is an indicator of cognitive effort in post-editing (Lacruz and Shreve, 2014). We investigate how low the pause threshold can reasonably be taken, and we propose that 300 ms is a good choice, as pioneered by Schilperoord (1996). We then seek to identify a good measure of the cognitive *demand* imposed by MT output on the post-editor, as opposed to the cognitive *effort* actually exerted by the post-editor during post-editing. Measuring cognitive demand is closely related to measuring MT utility, the MT quality as perceived by the post-editor. HTER, an extrinsic edit to word ratio that does not necessarily correspond to actual edits per word performed by the post-editor, is a well-established measure of MT quality, but it does not comprehensively capture cognitive demand (Koponen, 2012). We investigate *intrinsic* measures of MT quality, and so of cognitive demand, through edited-error to word metrics. We find that the transfer-error to word ratio predicts cognitive effort better than mechanical-error to word ratio (Koby and Champe, 2013). We identify specific categories of cognitively challenging MT errors whose error to word ratios correlate well with cognitive effort.

1 Introduction

The task of the post-editor is to render machine translation output in a readily usable form in the target language. Anyone who has successfully struggled with strangely worded assembly instructions can attest that it is sometimes possible for a person with no knowledge of the source language to make good sense of an imperfect machine translation – provided there is sufficient context. However, most post-editing is carried out by professional translators.

Nevertheless, formal training in post-editing has only recently been introduced as a necessary part of translation training (e.g., O'Brien, 2002), and more work remains to be done to identify the critical competences that predict good post-editing performance (e.g. Almeida, 2013). Post-editing, the task of editing MT output in the target language while referring as needed to the source text in a different language, is very different from the task of translating directly from one language to another. Consequently, the cognitive processes involved in these two tasks will also differ. As a result, traditional translator training may not be ideal preparation for work as a post-editor. While translation process research has made considerable progress in recent years (see, for example, Muñoz Martín, 2014) our understanding of the post-editing process is more limited. It is now becoming important to gain a deeper understanding of the post-editing process, not only as an academic pursuit, but also as a tool to aid in the development of effective training for future translators who will work at least partly as post-editors.

Our objective in this paper is to contribute to post-editing process research by gaining more insight into effective measures of the cognitive demand an MT text imposes on the post-editor, and by investigating how that

demand relates to the cognitive effort expended by the post-editor. While this is important to understand from the theoretical and applied perspectives of post-editing process research, it is also relevant to the MT community. The amount of effort post-editors need to exert affects their productivity levels. Accordingly, a good understanding of what features of a machine translation result in higher post-editing effort levels will provide a valuable resource for machine translation researchers as they work to increase the utility of their systems. This is a different, but perhaps more effective focus than the traditional emphasis on improving adequacy compared to gold-standard reference translations (Denkowski and Lavie, 2012a).

Types of effort: Krings (2001) made significant early contributions to the study of effort in post-editing. He created a three-way categorization of different types of effort (temporal: time spent; cognitive: mental processin; and technical: physical action) and proposed that the combination of cognitive and technical effort gives rise to temporal effort. However, it is too simplistic to think that the time spent thinking without obvious action plus the time spent on keyboarding and mouse actions is the total time spent on the post-editing task. In particular, post-editors will be thinking as they type. Sometimes that thinking will not affect their typing, sometimes it will cause them to slow down slightly, and sometimes it will lead them to stop for a while. So, while it is possible to measure temporal and technical effort directly, the only options for assessing cognitive effort are indirect measures.

Technical effort, the effort required for the keyboarding and mouse actions made while editing MT output, can be measured using logging software. The software can classify, count, and time the post-editor's actions, including mouse clicks, insertions, deletions, substitutions, and shifts.

Measures of MT quality: From the utility-focused perspective of any individual post-editor, MT quality is highest when the effort required for post-editing carried out by that post-editor is least. Initially, MT quality was measured through subjective human judgments (King, 1996). It is important to note that human judgments are a measure of MT quality that is *extrinsic* to the post-editing process, since they are not made during the course of the post-editing process. They are the product of reflection and do not necessarily capture the complexities of subconscious processing during post-editing. Subsequently, a variety of automatic metrics - including TER (Snover et al., 2006), BLEU (Papineni et al., 2002), and METEOR (Lavie and Denkowski, 2009) – were developed to assess MT quality by measuring how well MT output matches one of a set of reference translations. Versions of these metrics (HTER, HBLEU, and HMETEOR) measure how well MT output matches the single post-edited version produced by an individual post-editor. Snover et al. (2006) report good correlations of the order of .6 between human judgments and each of HTER, HBLEU, and HMETEOR. These metrics are also extrinsic to the post-editing process. They do not measure the steps that were actually carried out by the post-editor. Instead, they measure the most efficient path from the MT output to the final post-edited product.

HTER can be viewed as a measure of *required* technical effort, rather than a measure of *actual* technical effort. It is computed as the ratio

$$\text{HTER} = \frac{\text{\# of required edits}}{\text{\# of reference words}},$$

where the number of edits refers to the least number of insertions, deletions, substitutions, and shifts required to convert the MT output to the final post-edited version, and the number of reference words is the number of words in the MT output. When the required technical effort for post-editing is low, HTER is also low, and MT quality is inferred to be high.

However, as observed for example by Koponen et al. (2012), HTER is not a perfect measure of actual technical effort exerted by the post editor. HTER measures the shortest route to the final product, but post-editors will often take a route that is not optimal. A simple example is where the post-editor begins to make a change in the MT output, but then reverses course and accepts the MT output without modification. The corresponding HTER will be zero. Nevertheless, the changes begun but then undone by the post-editor certainly constitute non-zero technical effort. Along with this technical effort, the post-editor will also have made cognitive effort through evaluating how to change the MT output and then deciding to abandon the change mid-stream. HTER also fails to fully capture cognitive effort (Koponen, 2012).

Pauses and cognitive effort: Overall processing rate is of great concern to businesses and translation professionals, and there are several promising studies that relate this to cognitive effort. See, for example, O'Brien (2011) and Koponen et. al. (2012). However, there are other parameters that also appear to give good insight into levels of cognitive effort during post-editing.

Previous work (Lacruz et al., 2012; Lacruz and Shreve, 2014; Lacruz and Muñoz, 2014; Green et al., 2013) has provided evidence that pauses in post-editing are indicators of cognitive effort, just as they are in other types of language production. Indeed, triangulation between keystroke logs and eye tracking data on fixations and gaze

duration demonstrate that pauses are associated with cognitive effort in monolingual language production (e.g., Schilperoord, 1996) and in translation and interpreting (e.g. Krings, 2001, Dragsted and Hansen, 2008, Shreve et al., 2011; Timarová et al., 2011). In post-editing, there is evidence that cognitively challenging edits give rise to clusters of short, possibly monitoring pauses (e.g. Lacruz et al., 2012). This motivated the consideration of Average Pause Ratio,

$$\text{APR} = \frac{\text{average time per pause}}{\text{average time per word}},$$

and Pause to Word Ratio,

$$\text{PWR} = \frac{\text{number of pauses}}{\text{number of words}}.$$

Lacruz and Shreve (2014) showed that both APR and PWR correlate well with cognitive effort, identified through detailed examination of keystroke logs. In particular, low APR and high PWR are associated with high levels of cognitive effort. This is consistent with clustering of short pauses in cognitively challenging segments. Since a short pause is not a major contributor to total pause time, the association between short pauses and cognitive difficulty also explains why O'Brien (2006) did not find an association between post-editing difficulty predicted by negative translatability indicators and Pause Ratio,

$$\text{PR} = \frac{\text{total pause time}}{\text{total time in segment}}.$$

From now, we will focus on the Pause to Word Ratio (PWR) as a measure of cognitive effort in post-editing.

2 Rationale

One issue that has not been investigated systematically is the question of what is an appropriate minimum threshold for pause length. See Green et al. (2013) for a recent discussion. It is apparent from Lacruz et al. (2012) and Lacruz and Shreve (2014) that pauses shorter than the frequently used 1000 ms or 2000 ms thresholds are important indicators of cognitive effort. However, the 500 ms threshold used in these papers was chosen somewhat arbitrarily. Others, such as Schilperoord (1996) and Green et al. (2013) have used an even lower 300 ms threshold. Keystroke logs show that a threshold below about 200 ms is not appropriate, since the time needed to routinely type consecutive characters is often in a range up to 150 ms or even more.

As the pause threshold decreases, the number of pauses will increase, and so PWR will also increase. A strong correlation between PWRs corresponding to different minimum pause thresholds would indicate that the integrity of PWR as a metric for cognitive effort during post-editing would not be compromised by some variability in the minimum pause thresholds. To investigate this issue,

- we will compute correlations between PWR values using 200 ms, 300 ms, 400 ms, and 500 ms minimum thresholds for pauses during post-editing.

The investigation of how MT quality correlates with cognitive effort in post-editing is quite recent. See, for example, Koponen (2012) and Koponen et al. (2012). In this paper, we will take a different approach by studying how well human judgments of MT quality and the automatic MT quality metric HTER, which are measures that are extrinsic to the actual post-editing process, correlate with PWR, a cognitive effort metric based on actual measurements of the post-editing process, and so intrinsic to the post-editing process. Although HTER is designed to estimate MT quality by quantifying the necessary technical effort to convert MT text into the post-edited version, it is likely that in many situations technical effort and cognitive effort will be related. Accordingly,

- we predict that increases in HTER (decreases in MT quality) will be associated with increases in PWR (increases in cognitive effort.)

The formal similarity, by which HTER measures optimal edits per word, while PWR measures actual pauses per word, and the expectation that the number of pauses will increase as the number of edits increases combine to reinforce the prediction of a positive correlation between HTER and PWR. However, since HTER estimates MT quality by measuring the distance between the MT output and the final post-edited product, but without taking into account the specific edit actions of the post-editor, there is no *a priori* guarantee that there will be a strong correlation between HTER and PWR. On the other hand, if a post-editor is asked to rate MT quality by judging how difficult an MT segment was to post-edit, their memory of actual edit actions will likely influence their judgment. It is plausible that segments that post-editors judge to be difficult will have required expenditure of higher levels of cognitive effort during the post-editing process. Accordingly,

- we predict that improvements in human MT quality ratings will be associated with decreases in PWR (decreases in cognitive effort in post-editing.)

As Snover et al. (2006) found that human ratings of MT quality correlate strongly with HTER, our two predictions are consistent with each other.

Cognitive demand imposed by MT output: The discussion so far has centered on how to measure cognitive effort expended by the post-editor in producing the final post-edited version; and how to relate that effort to extrinsic measures of the quality of the machine translation text. We have focused on PWR as a measure of cognitive effort in producing the final post-edited version. The main extrinsic measures of MT quality we discussed were HTER and human quality judgments; these did not rely on identifying specific features of the MT text.

Ultimately, we are interested in identifying intrinsic features of the source text that are associated with high levels of cognitive effort expended by the post-editor. In other words, we wish to determine which features of the source text are likely to give rise to MT output that imposes high levels of cognitive demand on the post-editor. This is a complex question, and it seems prudent to approach it step by step.

Post-editing involves three stable texts, the source text, the machine translation, and the final post-edited version. We begin by asking what intrinsic features of the machine translation, the text in the middle, place high levels of cognitive demand on the post-editor and so are associated with elevated cognitive effort on the part of the post-editor. Thus, we seek to determine *intrinsic* measures of MT quality such that increases in MT quality are associated with reductions in cognitive effort in post-editing, as measured by PWR.

This agenda was advocated by Lacruz and Muñoz (2014). Drawing on the work of Koponen (2012) and Koponen et al. (2012), they grounded their approach in the analysis of MT errors, categorized according to the linguistically based difficulty ranking proposed by Temnikova (2010) and later modified by Koponen et al. (2012). Temnikova classified MT errors into nine categories assumed to pose increasing cognitive difficulty for the post-editor. These categories are specified in Table 1.

Error ranking	Error Type
1	Correct word, incorrect form
2	Incorrect style synonym
3	Incorrect word
4	Extra word
5	Missing word
6	Idiomatic expression
7	Wrong punctuation
8	Missing punctuation
9	Word order at word level
10	Word order at phrase level

Table 1. Temnikova's MT error classification

Lacruz and Muñoz defined a cognitive demand metric for MT segments that they called Mental Load (ML). Each error in an MT segment was assigned a weight according to its type in the Temnikova classification. For

example a “Correct word, incorrect form” error was assigned weight 1; an “Idiomatic expression” error was assigned weight 6. ML for the segment was the sum of the weights for each error. Thus, a segment with three incorrect word errors, two idiomatic expression errors, and one missing punctuation error had an ML of $3\times3 + 2\times6 + 1\times8 = 29$. It was found that there was a significant strong correlation between ML and cognitive effort, as measured by Pause to Word Ratio, PWR. However, there are at least two difficulties with this analysis: Mental Load was not normalized for segment length; and Temnikova’s rankings provide order data, rather than interval or ratio data. It is not likely for example that a rank 9 error (word order at word level) is nine times more difficult to correct than a rank 1 error (correct word, incorrect form.) For the data analyzed, neither shortcoming was likely significant: the segments were mostly of very similar length, and most of the errors were low on Temnikova’s scale. In view of these two facts, it was not surprising that the total number of errors also correlated well with PWR.

In this paper, we work with a different error classification. Following the framework of the American Translators Association (ATA) grading rubric (Koby and Champe, 2013), we first classify MT errors into two categories, Mechanical (M) and Transfer (T). Mechanical errors are those that can routinely be fixed without reference to the source text. Consider, for example, an MT segment that contains the phrase *he drink the coffee*. If the machine translation is referring to a man and is consistently written in the present tense, it is clear - without reference to the source text - that this phrase contains a mechanical error and should be edited to become *he drinks the coffee.* Now consider a machine translation where the first segment is *Helen Monica helps.* This is a transfer error: without consulting the source text, it is impossible to know how to edit the segment to reflect the true meaning of the source.

- We hypothesize that the cognitive demand placed on post-editors by transfer errors is greater than the cognitive demand resulting from mechanical errors.

Error code	Error type
ILL	Illegibility
IND	Indecision, gave more than one option
MT	Mistranslation
MU	Misunderstanding of source text
A	Addition
O	Omission
T	Terminology, word choice
R	Register
F	Faithfulness
L	Literalness
FA	Faux ami
COH	Cohesion
AMB	Ambiguity
ST	Style
G	Grammar
SYN	Syntax
P	Punctuation
SP/CH	Spelling/Character
D	Diacritical marks/Accents
C	Capitalization
WF/PS	Word form/Part of speech
U	Usage

Table 2. ATA grading rubric

The American Translators Association uses a grading rubric, given in Table 2. Similarly to Angelone (2011), we construct a simplified version of the ATA rubric that we specify in Table 3. The objective is to provide a simple cognitively-based classification of MT errors that is more specific than the mechanical/transfer partition.

We combine ATA error types Mistranslation, Faux Ami, and Terminology into a single category of Mistranslation (MT); we combine error types of Addition and Omission into a single category of Omission or Addition (OA); we consider Syntax (SY) a single category; we combine error types Word Form, Grammar, and Spelling into a single category of Word Form (WF); we use a single category of Punctuation (P); and we omit the other error types related to style, since they are not relevant to the instructions given for the post-editing of machine translation output in the study in this paper. These umbrella error types do not necessarily divide cleanly into Mechanical or Transfer. In particular, Word Form errors may be either Mechanical or Transfer, depending on the context. For example, many errors of type OA will be transfer errors. However, if the text is about food and contains the phrase *fish chips*, there is no need to consult the source to realize that this should be edited to *fish and chips.*

Error code	Error type
MT	Mistranslation
OA	Omission or Addition
SY	Syntax
WF	Word Form
P	Punctuation

Table 3. Simplified error classification based on ATA rubric.

Note that we have measured cognitive effort by the intrinsic metric of pauses per word (PWR). On the other hand MT quality has been measured by extrinsic metrics, such as required edits per word (HTER). By analogy, we propose edited errors per word as a good candidate for cognitive demand, or intrinsic MT quality. Specifically, we define the Error to Word Ratio for an MT segment as

$$\text{EWR} = \frac{\#\text{ of edited errors}}{\#\text{ of words}}.$$

We will also be interested in errors of various special types. When we wish to work with errors of type X, we use the X-Error to Word Ratio,

$$\text{X-EWR} = \frac{\#\text{ of edited errors of type X}}{\#\text{ of words}}.$$

All these different EWRs can be thought of as intrinsic measures of MT quality. We investigate the extent to which it is reasonable to consider them to be measures of cognitive demand by determining how well they correlate with cognitive effort.

- We hypothesize that X-EWR will correlate more strongly with PWR when the error type X corresponds to errors that are more cognitively difficult.

In particular,

- we predict that EWR for transfer errors will correlate more strongly with PWR than will EWR for mechanical errors.

Also, in line with the general expectations of Temnikova's classification,

- we predict that EWR for transfer errors of type MT, OA, and SY will correlate more strongly with PWR than will EWR for errors of type P or WF.

3 Method

There were five participants in this study, all of whom were paid for their time. All participants had English as their first language (L1) and were highly proficient in Spanish as their second language (L2). Each participant was a student in a Master of Spanish Translation program at an American university. They had all completed a graduate level course that included instruction and practice in the process of post-editing and the use of translation memory systems as an aid in the translation process.

Source texts were extracts of Spanish language transcripts of TED talks on matters of general interest with little technical language. Four Spanish source texts were translated, each by two different adaptive machine translation systems. The adaptive MT systems learn in real time from each post-edited segment, which then impacts the translation that the MT system generates for the following segment. All participants post-edited a version of each of the four texts, two translated by one of the MT systems and two translated by the other system. Texts were divided into segments that roughly corresponded to sentences or stand-alone phrases that varied in length from 2 to 18 words, mean 9.3 words. Each participant became familiar with the set-up and procedure by post-editing a 10 segment practice text. Data corresponding to the practice text are not included in the analyses presented here. The remaining three texts, the experimental texts, contained 30 segments each. Analysis was thus carried out on 90 segments for each participant. All data was pooled since each participant post-edited potentially different MT segments. Participants post-edited the four texts in one session lasting less than two hours, although there were no time limits set for the task.

Data was collected remotely using TransCenter, a web-based translation interface that logs post-editing activity (Denkowski and Lavie, 2012b). The data used for this paper consisted of the keystroke log (for computing the number of pauses of different lengths), HTER ratings of MT quality, and user ratings of MT quality. Participants worked from their homes and were instructed to minimally post-edit. Specifically, they were asked to disregard issues of style and to focus on how well the machine translation conveyed the meaning of the source text. After participants logged in, the source segments appeared on the left of the screen and the machine translation for the first segment appeared on the right. Once the participant finished post-editing a segment, they were asked to rate that segment's suitability for post-editing on a scale from 1 to 5, as in Table 4. The scale was available for consultation at all times.

Rating	Criterion
1	Gibberish - The translation is totally incomprehensible
2	Non-usable - The translation has so many errors that it would clearly be faster to translate from scratch
3	Neutral - The translation has enough errors that it is unclear if it would be faster to edit or translate from scratch
4	Usable - The translation has some errors but is still useful for editing
5	Very good - The translation is correct or almost correct

Table 4. Criteria for user ratings of MT quality

Post-edited MT errors were classified independently by two experienced translation graders. Cases of disagreement were very limited (less than 5%). These cases were resolved through consultation between the graders.

4 Results and Discussion

Our results will be expressed in terms of correlations. We adopt Cohen's (1988) convention that a positive Pearson correlation is strong when r is at least .5, moderate when r is between .3 and .5, and weak when r is between .1 and .3. Similar conventions hold for negative r and for Spearman's ρ. We use Pearson's r for comparisons of ratio data, and Spearman's ρ for comparisons involving rank order data.

4.1 Pause Threshold

Our first objective was to assess the sensitivity of PWR to reductions in the pause threshold in 100 ms steps from 500 ms down to 200 ms. The highly significant correlations between PWR values at all of these thresholds were strong and positive, as shown in Table 5.

Pearson r	PWR-300	PWR-400	PWR-500
PWR-200	.95**	.93**	.90**
PWR-300		.98**	.96**
PWR-400			.98**

Table 5. Pearson correlations between PWRs for different pause thresholds. Significance: ** $p < .001$.

While the mean PWRs for all the pause thresholds were significantly different from each other, most differences were relatively small. However, as shown in Table 6, the difference was noticeably numerically larger for the transition from PWR-300 to PWR-200 than for the transitions from PWR-400 to PWR-300 or from PWR-500 to PWR-400. The same pattern is apparent for median values.

Center Measure	Median Value	Mean Value
PWR-200	0.50	0.71
PWR-300	0.43	0.58
PWR-400	0.40	0.50
PWR-500	0.38	0.43

Table 6. Median and mean values of PWR at different pause thresholds.

Correlations with PWR at the 200 ms threshold, while very strong and highly significant, were lower than for other comparisons. The 200 ms threshold was also dangerously close to typical typing latencies for some participants, so we took the evidence above to indicate possible contamination of pauses due to cognitive effort with pauses due to mechanical effort at this threshold. Although closer investigation would be necessary to draw firm conclusions, we chose to discard the 200 ms pause threshold for the purposes of our investigation of the relationship between utility based intrinsic measures of cognitive demand on post-editors (viewed also as a measure of MT quality) and cognitive effort in post-editing. Since the most pause information can be derived from smallest reasonable pause threshold, we will henceforth select 300 ms for the pause threshold used in computing PWR. The 300 ms choice has the benefit of conforming to some previous selections, as in Schilperoord (1996) and Green et al. (2013).

4.2 Cognitive Demand/MT Quality and Cognitive Effort

Correlations between HTER, User Ratings, and PWR: We predicted that increases in MT quality will be associated with decreases in cognitive effort. When we measure MT quality extrinsically by HTER (low HTER

corresponds to small minimal edit distance, so high MT quality) and cognitive effort by PWR, the prediction is equivalent to expecting increases in HTER to be associated with increases in PWR. This is borne out by the fact that $r = .75$, $p < .001$. In other words, as predicted, there is a highly significant strong positive correlation between PWR and HTER.

Our next prediction was a variant of the first: as MT quality improves, cognitive effort in post-editing decreases. We still measure cognitive effort by PWR, but this time we estimate MT quality by user ratings of quality – difficulty ratings made by post-editors after they complete their task. The ratings were on a scale of 1 to 5, with 1 being reserved for the most difficult segments. In these terms, decreases in user ratings were predicted to correspond to increases in PWR. This was confirmed: the Spearman correlation between user ratings and PWR was $\rho = -.71$, $p < .001$, a highly significant strong negative correlation.

We also confirmed that, as expected, there was a strong negative correlation ($\rho = -.77$, $p < .001$) between user ratings and HTER. This correlation was highly significant. Table 7 below summarizes the findings.

Correlation	HTER	User Rating
PWR-300	r = .75**	ρ = -.71**
HTER		ρ = -.77**

Table 7. Summary of correlations between HTER, User Rating, and PWR. Significance: ** $p < .001$.

Influence of Transfer and Mechanical Errors on Cognitive Effort: The next objective was to investigate how well Transfer-Error to Word Ratio (T-EWR) and Mechanical-Error to Word Ratio (M-EWR) serve as intrinsic measures of cognitive demand. Thus, in all cases, the errors considered were errors actually corrected by the post-editor.

The prediction was that transfer errors would generate more cognitive demand than mechanical errors, and so T-EWR would correlate more strongly than M-EWR with cognitive effort, measured by PWR. Likewise, since T-EWR is predicted to be a stronger intrinsic measure of cognitive demand, it should also correlate more strongly than M-EWR with extrinsic measures of MT quality, that is, extrinsic measures of cognitive demand. These predictions were confirmed by the analysis. Correlations between T-EWR and each of PWR, HTER, and User Rating were strong positive and highly significant. On the other hand, correlations between M-EWR and each of PWR, HTER, and User Rating were still highly significant but only moderate. See Table 8 for a summary.

Correlation	T-EWR	M-EWR
PWR-300	r = .56**	r = .43**
HTER	r = .60**	r = .41**
User Rating	ρ = -.61**	ρ = -.40**

Table 8. Summary of correlations of Transfer and Mechanical Error to Word Ratios with PWR, HTER, and User Rating. Significance: ** $p < .001$.

Influence of Errors in Simplified ATA Categories on Cognitive Effort: We examined correlations of Error to Word Ratios for the five error categories derived from the ATA grading rubric. In all cases, the errors considered were errors actually corrected by the post-editor. We had predicted that the more cognitively challenging error types (Mistranslation, Omission or Addition, Syntax) would be more reliable intrinsic measures of cognitive demand than Punctuation or Word Form, and so would correlate more strongly with cognitive effort (PWR) or extrinsic measures of cognitive demand (HTER or User Rating.) This was indeed the case. See the summary in Table 9 for precise details, but MT-EWR correlated strongly and very significantly and with all of PWR, HTER and User Rating, while the correlations for OA-EWR were moderate, but still highly significant. Surprisingly, correlations for SY-EWR, while highly significant, were only weak.

Other correlations were weak; those for P-EWR were highly significant, while those for WF-EWR had varied levels of significance. Accordingly, we see that EWRs for ATA categories of MT errors that were expected to be

cognitively challenging provided significant indications, albeit of variable strength, of cognitive demand in post-editing.

Correlation	PWR-300	HTER	User Rating
MT-EWR	r = .51**	r = .54**	ρ = -.58**
OA-EWR	r = .42**	r = .37**	ρ = -.39**
SY-EWR	r = .29**	r = .26**	ρ = -.28**
P-EWR	r = .17**	r = .22**	ρ = -.16**
WF-EWR	r = .05	r = .11*	ρ = -.17**

Table 9. Summary of correlations of Simplified ATA Error to Word Ratios with PWR, HTER, and User Rating. Significance: ** $p < .001$; * $p < .01$.

Influence of All Edited Errors on Cognitive Effort: However, the most reliable intrinsic measure of cognitive demand turned out to be the simple Error to Word Ratio (EWR), combining all error types. In all cases, the errors considered were errors actually corrected by the post-editor. There were strong and highly significant correlations between EWR and all of PWR ($r = .65$, $p < .001$), HTER ($r = .62$. $p < .001$), and User Rating ($\rho = -.68$, $p < .001$). This mirrors the finding in Lacruz and Muñoz (2014).

5 Conclusions and future directions

In this paper, we probed the sensitivity of the Pause to Word Ratio to changes in the pause threshold. We concluded that 300 ms is a good choice for pause threshold. It is not too short to be contaminated by normal typing activity, but is sufficiently short to capture much potentially informative pause activity.

We went on to compare PWR, an intrinsic measure of cognitive effort, with widely used metrics that have indirect relationships to cognitive effort and are often viewed as measures of MT quality. We found strong correlations between PWR and HTER, an edit to word ratio that estimates MT quality in terms of technical effort, and user ratings, that estimate MT quality in terms of perceived difficulty of post-editing.

Then we asked how we might measure cognitive demand on the post-editor. As a result of the cognitive demands placed on post-editors by features of the MT output, they must expend cognitive effort to complete the post-editing task. We chose to measure cognitive demand through Edited-Error to Word (EWR) metrics, formally analogous to the Pause to Word metric for cognitive effort and the Required-Edit to Word metric (HTER) for MT quality. Transfer errors require post-editors to review the source text to understand the meaning, while mechanical errors can reasonably be fixed without reference to the source text. The expectation is that transfer errors are more cognitively demanding to fix than are mechanical errors. This view is supported by the finding that EWR for transfer errors correlates more strongly with HTER and user ratings (MT quality measures; extrinsic measures of cognitive effort) or PWR (intrinsic measure of cognitive effort). Similarly, for other error classifications based on ATA rubrics, EWRs for those error types that were expected to be more cognitively demanding to fix correlated more strongly with PWR, HTER, and user ratings.

Results support the view that error to word ratios may be an effective way to gauge the cognitive demand imposed on post-editors by MT segments. However, these results must be viewed as preliminary, since they were generated from small samples of 90 source text segments and 5 post-editors.

Corroborating studies need to be carried out on a larger scale and supported by methodologies such as eye tracking or mouse tracking that allow direct observation of the focus of attention and have established metrics for assessing cognitive effort. It seems particularly interesting to study possible differences between the processing of transfer and mechanical errors. To gain maximum advantage, it would be worthwhile to undertake controlled experimental studies to filter out the noise of more ecological experiments. This would allow a closely focused investigation, which would potentially provide evidence to support hypotheses that could then be tested in a more natural setting.

The ultimate objective is to move beyond understanding what MT features are more or less cognitively demanding, and so require post-editors to expend more or less cognitive effort. The goal is to understand what features of the source text are associated with cognitively demanding errors in MT output. For this it may be

worthwhile to revisit the relationship between negative translatability indicators and pause data that was initiated by O'Brien (2006).

References

Almeida, Giselle. 2013. Translating the Post-Editor: An Investigation of Post-Editing Changes and Correlations with Professional Experience Across Two Romance Languages. Ph.D. Thesis, Dublin City University.

Angelone, Erik. 2010. Uncertainty, Uncertainty Management, and Metacognitive Problem Solving in the Translation Task. In Gregory M. Shreve and Erik Angelone (Eds.). *Translation and Cognition*, (pp. 17-40). Amsterdam/ Philadelphia: John Benjamins.

Denkowski, Michael and Alon Lavie. 2012a. Challenges in Predicting Machine Translation Utility for Human Post-Editors. Proceedings of the Tenth Biennial Conference of the Association for Machine Translation of the Americas, San Diego, California.

Denkowski, Michael and Alon Lavie. 2012b. TransCenter: Web-based translation research suite. In Workshop on Post-Editing Technology and Practice Demo Session, Tenth Biennial Conference of the Association for Machine Translation of the Americas, San Diego, California.

Carl, Michael, Barbara Dragsted, Jakob Elming, Daniel Hardt, and Arnt Lykke Jakobsen. 2011. The Process of Post-Editing: A Pilot Study. Proceedings of the 8th International NLPSC workshop. Special theme: Human machine interaction in translation. Copenhagen Studies in Language, 412. Frederiksberg: Samfundsliteratur.

Dragsted, Barbara and Inge Gorm Hansen. 2008. Comprehension and Production in Translation: a Pilot Study. Segmentation and the Coordination of Reading and Writing Processes. In Susanne Göpferich, Arnt Lykke Jacobsen,and Inger M. Mees (Eds.), *Looking at Eyes* (pp. 9–30). Copenhagen Studies in Language 36. Copenhagen: Samsfunds-litteratur.

Dragsted, Barbara and Inge Gorm Hansen. 2009. Exploring Translation and Interpreting Hybrids. The Case of Sight Translation. Meta: Translators' Journal, 54(3), 588-604.

Green, Spence, Jeffrey Heer, and Christopher D. Manning. 2013. The Efficacy of Human Post-Editing for Language translation. Conference on Human Factors in Computing Systems (CHI'13), April 27-May 2, 2013, Paris, France.

King, Margaret. 1996. Evaluating Natural Language Processing Systems. Communications of the Association for Computing Machinery, 29(1):73-79.

Koby. Geoffrey S. and Gertrud G. Champe. 2013. Welcome to the Ral World: Professional-Level Translator Certification. Translation & Interpreting, Vol 5(1), 156-173

Koponen, Maarit. 2012. Comparing Human Perceptions of Post-Editing Effort with Post-Editing Operations. Proceedings of the Seventh Workshop on Statistical Machine Translation (pp. 181-190), Montréal (Canada).

Koponen, Maarit, Wilker Aziz, Luciana Ramos, and Lucia Specia. 2012. Post-Editing Time as a Measure of Cognitive Effort. Proceedings of the Tenth Biennial Conference of the Association for Machine Translation of the Americas.Workshop on Post-Editing Technology and Practice (pp. 11-20), San Diego, California.

Krings, Hans P. 2001. Repairing Texts: Empirical Investigations of Machine Translation Post-Editing Processes. Geoffrey S. Koby (Ed.). Kent, Ohio: Kent State University Press.

Lacruz, Isabel, Gregory M. Shreve, and Erik Angelone. 2012. Average Pause Ratio as an Indicator of Cognitive Effort in Post-Editing: A Case Study. Proceedings of the Tenth Biennial Conference of the Association for Machine Translation of the Americas. Workshop on Post-Editing Technology and Practice (pp. 29-38), San Diego, California.

Lacruz, Isabel and Gregory M. Shreve. 2014. Pauses and Cognitive Effort in Post-Editing. In Sharon O'Brien, Laura Winther Balling, Michael Carl, Michel Simard, and Lucia Specia (Eds.), *Post-editing of Machine Translation: Processes and Applications*. Cambridge Scholars Publishing

Lacruz, Isabel and Ricardo Muñoz Martín. 2014. Pauses and Objective Measures of Cognitive Demand in Post-Editing. Paper presented at the American Translation and Interpreting Studies Association Conference, New York, April 2014.

Lavie, Alon and Michael Denkowski. 2009. The METEOR Metric for Automatic Evaluation of Machine Translation. Machine Translation Journal, 23 (2-3), 105-115.

Muñoz Martín, Ricardo. 2014. A Blurred Snapshot of Advances in Translation Process Research. In Muñoz Martín (Ed.), *Minding Translation*, MonTi, Special Issue 1, Universidad de Alicante, Spain.

O'Brien, Sharon. 2002. Teaching Post-Editing: A proposal for course content. Proceedings of the 6th European Association for Machine Translation Workshop "Teaching Machine Translation," 14-15 November, Centre for Computational Linguistics, UMIST, Manchester, England.

O'Brien, Sharon. 2005. Methodologies for Measuring the Correlations Between Post-Editing Effort and Machine Text Translatability. Machine Translation, 19(1): 37-58.

O'Brien, Sharon. 2006. Pauses as Indicators of Cognitive Effort in Post-editing Machine Translation Output. Across Languages and Cultures, 7(1), 1-21.

O'Brien, Sharon. 2011. Towards Predicting Post-Editing Productivity. Machine Translation, 25, 197-215.

Papineni, Kishore, Salim Roukos, Todd Ward, and Wei Jing Zhu. 2002. Bleu: a Method for Automatic Evaluation of Machine Translation. Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics.

Schilperoord, Joost. 1996. It's About Time: Temporal Aspects of Cognitive Processes in Text Production. Amsterdam: Rodopi.

Shreve, Gregory M., Isabel Lacruz, and Erik Angelone. 2011. Sight translation and Speech Disfluency: Performance Analysis as a Window to Cognitive Translation Processes. In Cecilia Alvstad, Adelina Hild, and Elisabet Tiselius (Eds.), *Methods and Strategies of Process Research* (pp. 121–146), Amsterdam/ Philadelphia: John Benjamins.

Snover, Matthew, Bonnie Dorr, Richard Schwartz, Linnea Miccuilla, and John Makhoul. 2006. A Study of Translation Edit Rate with Targeted Human Annotations. Proceedings of Association for Machine Translation in the Americas (pp. 223-231), August 8-12, 2006, Cambridge, Massachusetts, USA.

Specia, Lucia, Nicola Cancedda, Marco Dymetman, Marco Turchi, and Nello Cristianini. 2009. Estimating the Sentence-Level Quality of Machine Translation Systems. Proceedings of the 13th Annual Conference of the European Association for Machine Translation (pp. 28-35), Barcelona, Spain.

Temnikova, Irina. 2010. A Cognitive Evaluation Approach for a Controlled Language Post-Editing Experiment. International Conference Language Resources and Evaluation (LREC2010). Valletta, Malta, May 17-23.

Timarová, Sárka, Barbara Dragsted, and Inge Gorm Hansen. 2011. Time Lag in Translation and Interpreting. In Cecilia Alvstad, Adelina Hild, and Elisabet Tiselius (Eds.), *Methods and Strategies of Process Research* (pp. 121-146), Amsterdam/ Philadelphia: John Benjamins.