Towards Predicting Post-Editing Productivity

Sharon O'Brien

School of Applied Language and Intercultural Studies

Centre for Translation and Textual Studies

Centre for Next Generation Localisation

Dublin City University

Ireland

sharon.obrien@dcu.ie

Abstract

Machine Translation (MT) quality is generally measured via automatic metrics, producing scores that have no meaning for translators who are required to post-edit MT output or for project managers who have to plan and budget for translation projects. This paper investigates correlations between two such automatic metrics (General Text Matcher and Translation Edit Rate) and post-editing productivity. For the purposes of this paper, productivity is measured via processing speed and cognitive measures of effort using eye tracking as a tool. Processing speed, average fixation time and count are found to correlate well with the scores for groups of segments. Segments with high GTM and TER scores require substantially less time and cognitive effort than medium or low-scoring segments. Future research involving score thresholds and confidence estimation is suggested.

Keywords

Post-editing; productivity; cognitive effort; automatic metrics for MT; eye tracking

Introduction

Post-editing is the correction of raw machine translated output by a human translator according to specific guidelines and quality criteria. Recent advances in machine translation technology have led to an increased implementation by organisations with large translation volumes and a broad range of target language requirements. Consequently, technical translators who work as in-house translators or, more commonly, as freelance translators for these organisations are increasingly asked to post-edit, as opposed to translate or revise human translations that are recycled through translation memory systems. Although already mentioned in the 1980s and 1990s, post-editing is still a relatively new task, which is different from translation and traditional revision (McElhaney and Vasconcellos 1988). Recent research and reports from industry indicate that it is possible to increase productivity by using MT and post-editing (O'Brien 2007; Takako et al. 2007; Guerberof 2009; Groves and Schmidtke 2009, Tatsumi 2009, Plitt and Masselot 2010, de Almeida and O'Brien 2010). However, it is not yet clear what productivity can be realistically expected from a post-editor (the term we will use here to refer to a translator who engages in a post-editing task). Organisations implementing MT are now searching for cost and productivity models for post-editing. While the downward pressure is impalatable for translators, it appears to be somewhat inevitable (Garcia 2009). Rather than turn a blind eye, the translation profession ought to engage with this development so that cost and productivity models are not unrealistic and, ultimately, technical translators can continue to make a living.

This paper examines two of the metrics used by the MT development and user communities to predict the "quality" of raw MT output and tests the correlations between these two measures and post-editing productivity. One cold economic explanation of the term "productivity" is the ratio of the quantity and quality of units produced to the labour required per unit of time (Fellbaum 1998 - Wordnet). While many traditional translation scholars would baulk at such a cold conceptualisation of translation, it is true that the speed with which translated material is produced and the subsequent quality is a major concern of many organisations who have translation needs (commercial, governmental and non-governmental) and it is therefore important to both clients and translators that productivity be discussed. Here we expand on the explanation of productivity given above and take post-editing productivity to mean not only the ratio of quantity and quality to time, but also to the cognitive effort expended. We understand "effort" to be inversely related to productivity. In other words, the higher the effort, the lower the productivity. Our focus in this paper is on temporal and cognitive effort. An analysis of the quality of the post-edited text is beyond the scope of the paper.

MT Developers and Quality Predictions

There is a vast body of literature on machine translation evaluation (e.g. King et al. 2003; White 2003; Callison-Burch et al. 2008 etc.), which we cannot discuss in detail here. In summary, the trend has moved away from human evaluation, which is time- and cost-intensive, towards automatic methods of evaluation.

Currently, the most common of such metrics are BLEU (Papineni et al. 2002), NIST (Doddington 2002), TER (Snover et al. 2006), METEOR (Banerjee and Lavie 2005) and GTM (General Text Matcher, Turian et al. 2003). However, metrics are being constantly refined and new proposals are made at a significant rate (cf. Lavie and Przybocki 2009), not least because of the annual competitions organised by, for example, the American National Institute of Standards and Technology (NIST 2010).

The limitations of these automatic metrics are acknowledged to some extent by the MT community, for example: "Automatic metrics have not yet been proved able to consistently predict the usefulness, adequacy, and reliability of MT technologies" (NIST 2010). Way (2010: 27) also acknowledges that "…while the introduction of automatic evaluation metrics in MT…has largely been beneficial, they have to a large extent taken on too much importance, especially since *real translation quality* is what we should be concerned with" (my emphasis). MT developers have taken some of the objections on board by, for example, enabling the metrics to consider more than one reference sentence (from its very first days the BLEU metric, for example, used multiple reference sentences), creating composite scores for texts, rather than isolated sentences, and implementing penalties for very short or long sentences or for word order etc.

A more recent development in this field is the estimation of confidence scores by MT systems (Blatz et al. 2004) . In this case, a score is generated by the MT system, without the need for one or more "reference sentences". The translation industry has an interest in this development as accurate scores would allow users to triage MT quality according to their own specific needs. However, how such confidence scores relate to actual post-editing effort has not yet been tested. We will return to this concept in our conclusions.

General Text Matcher (GTM) and Translation Edit Rate (TER)

For the research reported here, two automatic MT metrics were selected, i.e. GTM or "General Text Matcher" (Turian et al. 2003) and TER or "Translation Edit Rate" (Snover et al. 2006) . GTM is given priority in the analysis (the reasons for this will be explained under "Objectives and Methods") and TER is included as a secondary check of correlation with automatic metrics; here we will provide a brief overview of how both of these metrics work, starting with GTM.

The GTM metric measures similarity between the raw MT output (the "candidate" translation) and the reference sentence using measures of precision, recall and their composite F-measure (the harmonic mean). Precision measures the number of words generated by the MT system that match with words in the reference sentence out of the total number of words generated by the MT system for that segment. Recall measures the number of words generated by the MT system that match with words in the reference translation out of the total number of words in the reference translation. The GTM metric also rewards matching adjacent words. Turian et al. (ibid.) demonstrated that GTM correlates well with human judgements of "adequacy" and "fluency" – two concepts that are frequently used in human evaluation of MT output (Ma and Cieri 2006).

GTM gives a score on a scale of 0-1. The closer the score is to 1, the more similar the MT output is to the reference translation. To demonstrate, Example 1 scored 0.34 using the GTM metric, whereas Example 2 scored 0.91.

Example 1

Source Sentence
Click this to decompress, or expand, compressed files as they are backed up.

Raw MT Output
Cliquez sur cette option pour decompress ou développer, les fichiers compressés ils sont sauvegardés.

Reference Sentence
Permet de décompresser ou développer des fichiers compressés lors de leur sauvegarde.

Example 2

Source Sentence
If you delete the backup exec system logon account, you should create a new one that enables you to perform the specified operations and use the agent and applet.

Raw MT Output
Si vous supprimez le compte de connexion au système backup exec, vous devez en créer un nouveau qui vous permette d'effectuer les mêmes opérations et d'utiliser l'agent et l'applet.

Reference Sentence
Si vous supprimez le compte de connexion au système backupexec, vous devez en créer un nouveau qui vous permette d'effectuer les opérations spécifiées et d'utiliser l'agent et l'applet.

The reason for the low score in Example 1 is the substantial difference between the Raw MT Output and its Reference Sentence; of a total of 14 words in the Raw MT Output, only four are shared with the Reference Sentence, whereas ten words are unique to the Raw MT Output. In terms of "text matching" then, there is a low number of matches between the two. Of the thirty words in the Raw MT Output of Example 2, on the other hand, only four differ in the Reference Sentence, giving a higher text matching score of 0.91.

TER measures the number of edits required to change raw MT output into a reference sentence. One of the objectives of the TER developers was to seek higher correlations with human judges than was possible with other automatic metrics (Snover et al. 2006: 223). TER was selected here because, unlike other metrics such as BLEU, it does not require a large number of reference sentences in order to correlate with human judgements (ibid.). Additionally, it was conceived as a metric that measures the number of **edits** necessary to convert raw MT into a reference sentence and the effort required by such editing is the focus of our study too. TER and GTM are seen to be similar, with TER differing from GTM in that "TER assigns a lower cost to phrasal shifts […], and does not explicitly favor longer matching strings" (ibid : 224). The TER score is derived by dividing the number of edits (which can be insertions, deletions, substitutions or shifts in words or phrases) by the average number of reference words. The "best" TER score is 0 (meaning no edits were necessary to transform the raw MT output into the reference sentence) and there is no upper limit on the score. For Example 1 above (many edits required), the TER score is 92.31 while for Example 2 (minimal edits) the score is 14.71.

Objectives and Methods

Correlations between automatic metrics and post-editing productivity are often calculated on the basis of human annotation of *expected* post-editing effort, based on scales such as:

1. Requires complete translation
2. Post-editing quicker than retranslation
3. Little post-editing needed
4. Fit for purpose

(Specia et al. 2009a: 138)

The obvious weaknesses with this approach are that rating is subjective, it may be carried out by participants who have never post-edited before and who, therefore, have limited experiential knowledge from which to make judgements, each rating may be influenced by the previous rating, and fatigue or boredom may influence the motivation of raters. Even in cases where actual post-editing effort is captured this is sometimes done with participants who are not trained translators (see, for example, Koehn 2010) and so has limited application to the behaviour of professional translators. The objective of this exploratory study was to test for correlations between the specific automatic metrics mentioned above and *actual* post-editing productivity of *professional* translators. In the following section we explain the tools and methods employed to record productivity.

MT System and Editor

The research was carried out within the Centre for Next Generation Localisation (CNGL[i]) and the industrial partners included Alchemy (a localisation tool developer who provided the editing tool and funding) and VistaTEC (a localisation service provider who funded the professional post-editors). This set up meant that specific tools were availed of to measure post-editing productivity: the research-based, data-driven MT system developed within the CNGL was used as the MT engine (Du et al. 2009). A version of this system is also available as a free, open-source example-based MT system (Forcada 2010). The editing interface used was Alchemy Catalyst (version 8.0), which is commonly used by technical translators who work in the localisation field (see Figure 1).

Corpus and Automatic Metrics

A corpus of 55,000 sentence pairs of English-French translation was made available by a research collaborator in the IT domain. 45,000 sentence pairs were used to tune the MT system to the specific domain and 10,000 were reserved as a test set. This test set was randomly selected from the whole translation memory based on source (English) sentence length distribution, i.e. the sentence length in the 10,000 test set corpus has the same distribution as the whole 55,000 sentence corpus.

As previously mentioned, the automatic metrics GTM and TER were used. GTM was selected for a number of reasons: In previous research, GTM was found to have the highest correlation with post-editing speed when compared with other automatic metrics (Tatsumi 2009) and it has been found to have high correlation with human judgements of adequacy and fluency (Turian et al. 2003). In addition, it is in use by the company who donated the corpus as an *a posteriori* validator of the post-editing effort. By referring to GTM scores and correlating them with the actual post-editing effort recorded, we were able to test its accuracy as a validator of effort.

From the 10,000 sentence test set, 995 sentences were randomly selected based on the distribution of GTM scores, i.e. the distribution of GTM scores in this 995-sentence corpus is the same as that of the 10K-testset corpus. As GTM scores were already generated for the corpus, they were used as a basis for dividing the test set into the following categories:

- Sentences falling into a *GTM Low score* category. This was arbitrarily defined as those scores falling between 0 and 0.4.

- Sentences falling into a *GTM Medium score* category, i.e. less than 0.8 and greater than 0.41.
- Sentences falling into a *GTM High score* category, i.e. 0.81-1.

Of course, this categorisation is arbitrary, but to set up a controlled eye tracking experiment there was a need to categorise the segments in some way, using some logic. For the post-editing task, we needed to ensure that there was an equal representation of segments across the GTM scale to test for correlations with productivity. Also, as the post-editing task was time-limited, we had to ensure that the small number of segments represented the entire GTM scale. As GTM scores are based on a scale from 0, the lowest point, to 1, the highest point, it seemed logical to select an equal number of segments from a low, medium and high point on the scale. The difficulty here is in deciding what "low", "medium" and "high" is.[ii] For our purposes, we decided that a GTM score of 0.4 and below was a "low" score and 0.8-1 was a "high" score. The "medium" category is anything in-between. Clearly, this categorisation is open to revision, but it was useful for this exploratory study and for setting up the post-editing task in a relatively controlled manner.

The rationale for including the TER metric was given earlier. While GTM was used to categorise and select the segments to be included in the post-editing task, TER was only used in a post-task test for correlations with productivity and for comparison with GTM.

As funding for the payment of post-editors was limited, 20 segments were randomly selected from each of these categories, making 60 segments (or 782 source words) in total for the post-editing task.[iii] Nine additional segments (three from each category) were selected for a post-editing warm-up task which was used to familiarise participants with the experimental set-up.

A glossary of 38 terms was then created and added to the Catalyst environment so that the post-editors would know that any specialised terms encountered were reliable. The source segments were presented in random order for post-editing and no time constraints were imposed.

Post-editor Profile

Seven native speakers of French who work primarily as professional translators and/or reviewers were recruited through VistaTEC. Care was taken to ensure that they had as similar a profile as possible in terms of domain experience, familiarity with Catalyst, post-editing experience, professional qualifications etc. Ethics approval was applied for and granted through the relevant ethics committee and all participants signed a consent form. Anonymity was guaranteed and participants were paid the going rate for their time.

Procedure for Measuring Effort

It has been suggested in research into post-editing that "effort" should not just be measured by task time, but that cognitive effort (i.e. mental processing) and technical effort (i.e. keyboarding) should also be factored in (Krings 2001). This is typically where industry-led and academic-led research on the topic diverge. The former is primarily interested in how long the task takes (and how much it costs), whereas the latter is also interested in the cognitive load. For the research reported here, temporal effort in the form of processing speed (i.e. number of words post-edited per second) and cognitive effort in the form of eye-tracking data were included. The use of eye-tracking data as an indicator of cognitive load is well established (Rayner 1998, Radach et al. 2004) and is based on Just and Carpenter's (1980) eye-mind hypothesis, i.e. that there is no appreciable lag between where the eye is fixating and what is being processed, especially in information processing or visual search tasks.

In this instance, eye tracking involved the use of specialist non-intrusive equipment (the Tobii 1750 eye tracker), which looks just like a normal computer monitor. Post-editors sat facing the monitor and completed the post-editing task and, while doing so, the monitor tracked their eye movements, the number of fixations and length of time spent in fixations. Although the post-editors knew that their eye movements were being recorded, they could not see any trace of this on the screen as they worked.

"Fixations" are defined as: "eye movements which stabilize the retina over a stationary object of interest" (Duchowski 2003: 43) and the more there are and the longer the time spent in fixations, the more difficulty the reader or translator is assumed to be experiencing (for further information on eye tracking-related research in translation studies, see Mees et al. 2009, Göpferich et al. 2008). Number and length of time in fixations have been shown to correlate well with cognitive effort in different domains (Rayner 1998; Radach et al. 2004). However, what constitutes a "significant" fixation, i.e. one that really demonstrates cognitive processing, is not agreed. For example, Alves et al. (2009: 272) point out that studies in information processing during reading "report that average fixations in reading activities usually range from 200 to 250 ms" but that work done by Jensen et al. (2009) led the latter to suggest a lower threshold of 175 ms. Although translation involves more than just reading, the current proposed length from reading research of 200 ms was employed here based on the knowledge that eye-tracking studies of reading have a longer tradition than those of translation processes. The issue of what a valid fixation length is for measuring translation-related processes is one that cannot be solved in this paper.

The analysis of fixation data was carried out using the eye tracking software analysis tool, Tobii Studio (v. 2.0.4). Figure 1 shows the Catalyst user interface used in the experiment. Figure 1 also includes a superimposed heat map for all participants for the post-editing of *Medium GTM* segments (heat maps for *Low GTM* and *High GTM* scores were similar). The sections labelled as 'red' (indicating a 'hot' area for fixations) account for 524 fixations in total while those labelled 'yellow' ('warm') represent the next most fixated area followed by 'green' ('cool').[iv] It can be seen that most fixations (logically) occur on the Edit Window, where the source segment and raw MT target segment are displayed, along with the Translator Toolbar. The next highest number of fixations is for the List View (which displays the target translations in order of occurrence and their status symbol) as well as the Glossary Window.

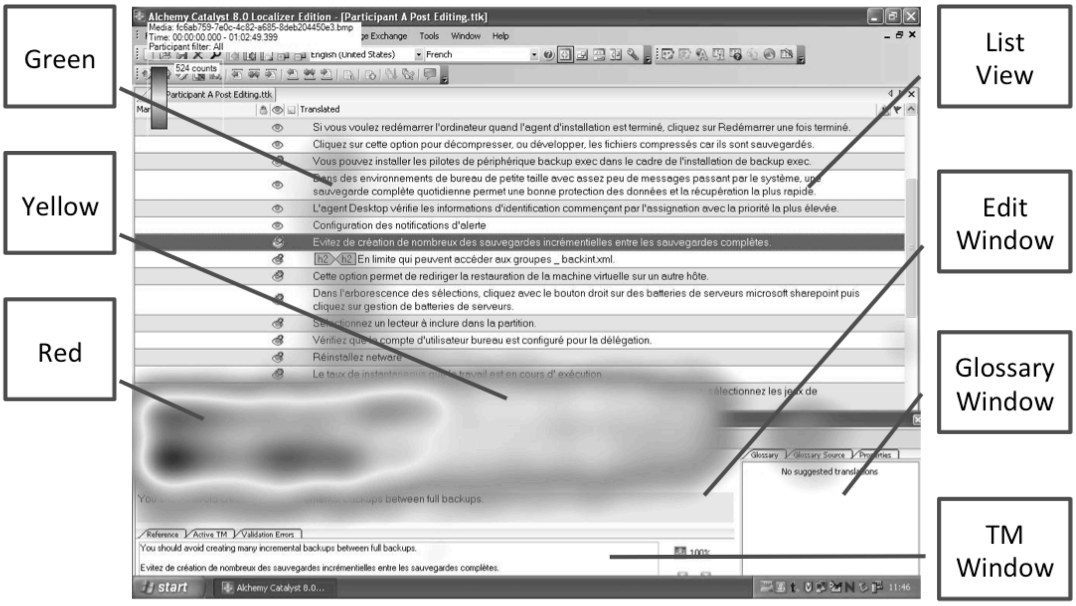

Fig. 1 Catalyst UI showing heat map for all participants for *Medium GTM* segments

Results

Processing Speed

Processing speed was measured by capturing the exact start and end times of each segment for each participant in Tobii Studio by replaying the AVI files recorded during the eye tracking session. The time in seconds was then divided by the number of words to get the measurement of words per second. While one of the most common ways of measuring translation or post-editing productivity is in words per day and not words per second, both funding limitations and the use of an eye tracker meant that a whole-day task could not be set. Also, the GTM and TER scores were assigned to each segment and we were interested in testing the correlations between these segment-level scores and post-editing productivity, as well as between groups of segments and productivity. For all of these reasons words per second was selected as the measure. From this measurement, words per hour or per day can be extrapolated, but this assumes a consistent level of throughput throughout the day, something that is questionable. One issue that has to be tackled is the fact that words are of varying lengths. We normalised the processing speed also by number of characters in order to reduce the effect of varying word length. The results per character confirmed the results per word and so only the latter are reported here.

Figure 2 shows average processing speed data (for all participants and all segments) along with the median value and standard deviations for each category of GTM score (low, medium, high).

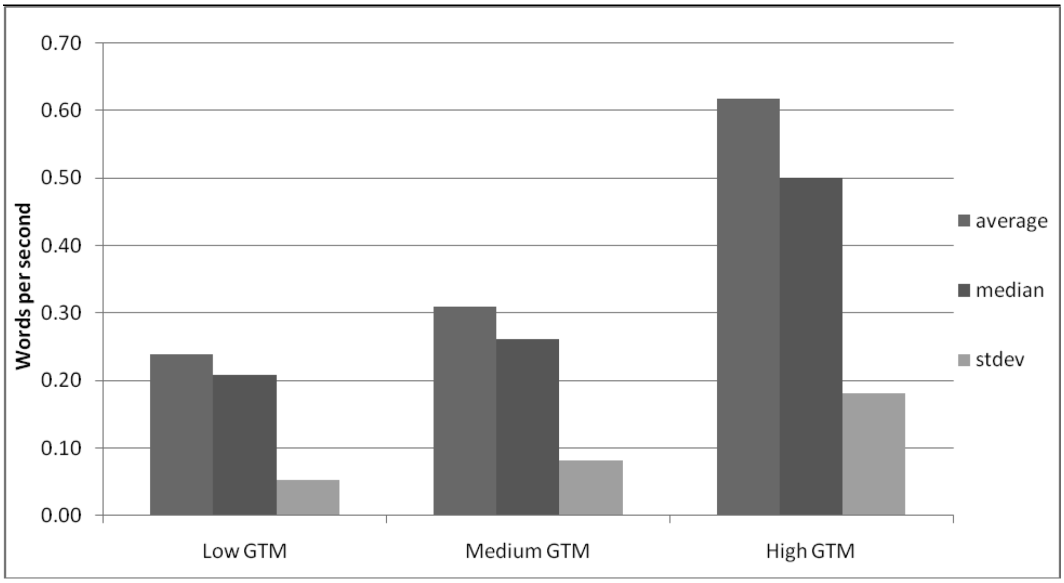

Fig. 2 Average processing speed (all participants and segments) for each category

The processing speed data suggest that there is a decreasing temporal post-editing effort from low, to medium, to high GTM scores. The differences between all GTM categories were found to be significant for processing speed.[v]

We also examined the processing speed data for correlations with TER scores. Readers are reminded that the best possible TER score is 0, so the higher the TER score, the more "edits" are required to convert the raw output to a reference sentence. As the TER score increases, then, we would expect average processing speed to decrease, if TER is a good predictor of post-editing effort measured in time. Figure 3 shows the trend for average TER score (converted to a % value for the sake of comparison with the words per second scale) and average processing speed for each group category. As can be seen, the higher the TER score, the lower the processing speed (in words per second) and vice versa.

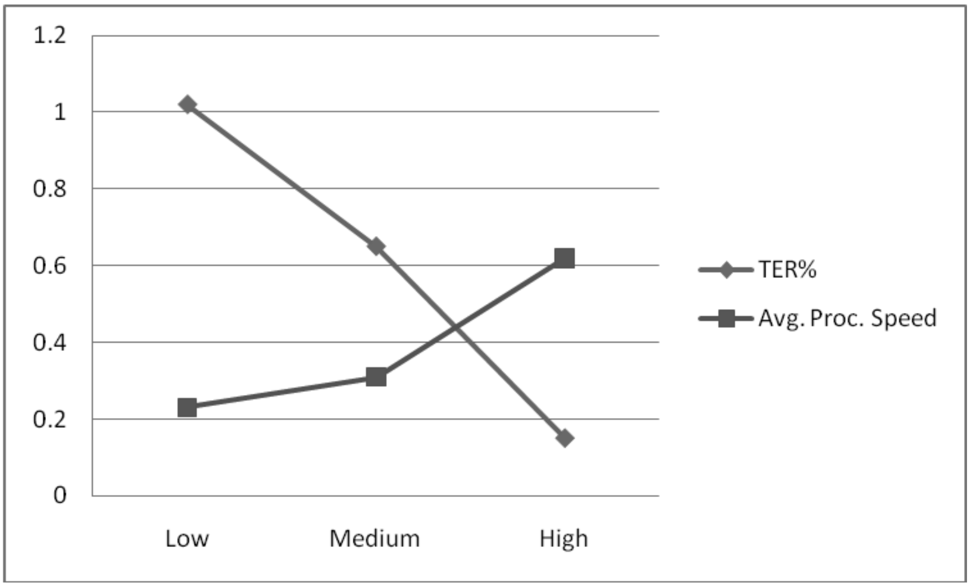

Fig. 3 Average TER score and processing speed (words per second) for each category

This result confirms the results using the GTM score categorisation. It is interesting to note that in both cases, the difference in average processing speed between the "low" and "medium" categories is relatively small whereas the difference between "medium" and "high" is substantial. We will return to this issue when analysing the eye tracking data.

Figures 2 and 3 present average processing speed results for the group of segments categorised as Low, Medium and High. It is useful to analyse the data on a segment basis to see if there are correlations between TER scores and average processing speed on an individual segment basis. Figures 4 to 6 show the TER scores and average words per second across participants for the individual Low, Medium and High segments.[vi]

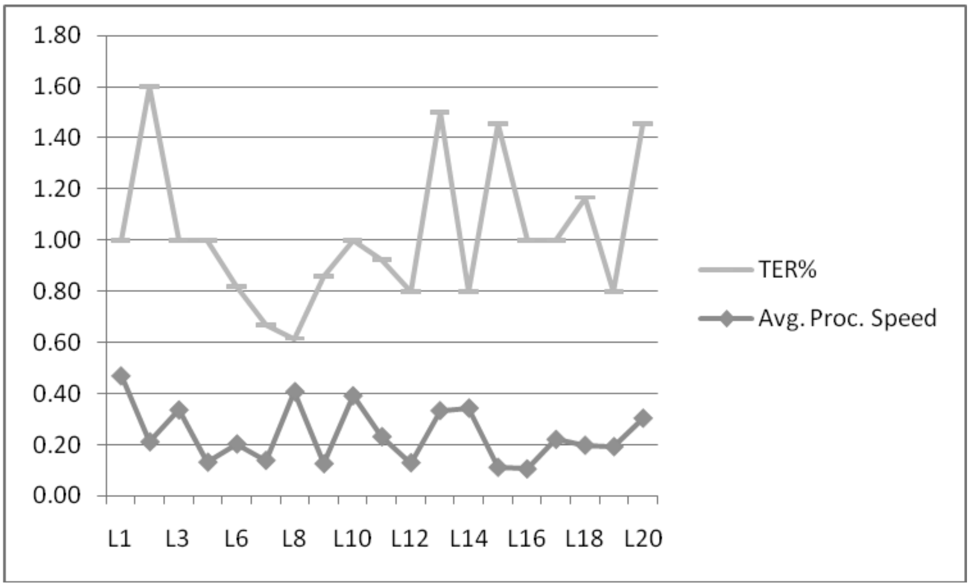

Fig. 4 Comparison between TER scores and average words per second for each segment classified as "Low" (L1-L20)

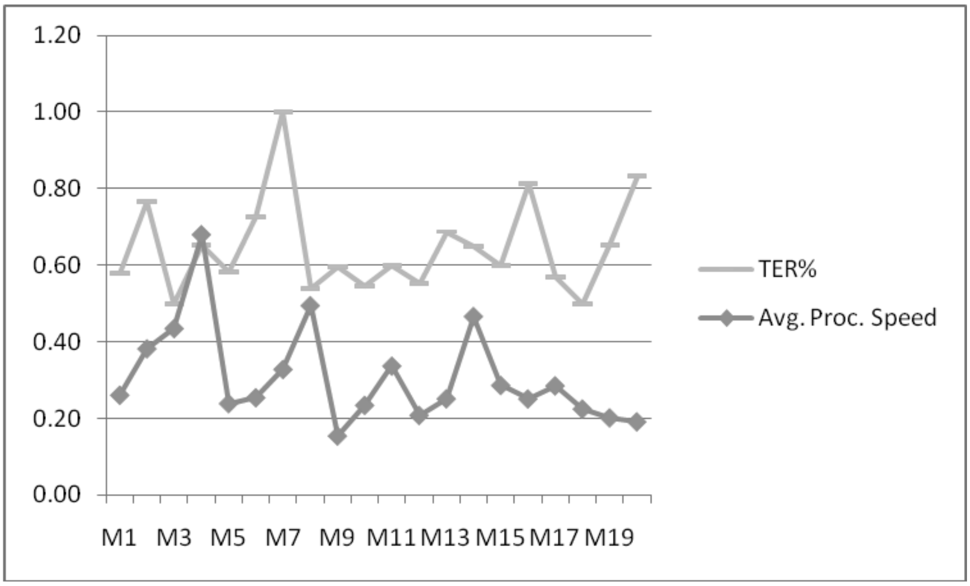

Fig. 5 Comparison between TER scores and average words per second for each segment classified as "Medium" (M1-M20)

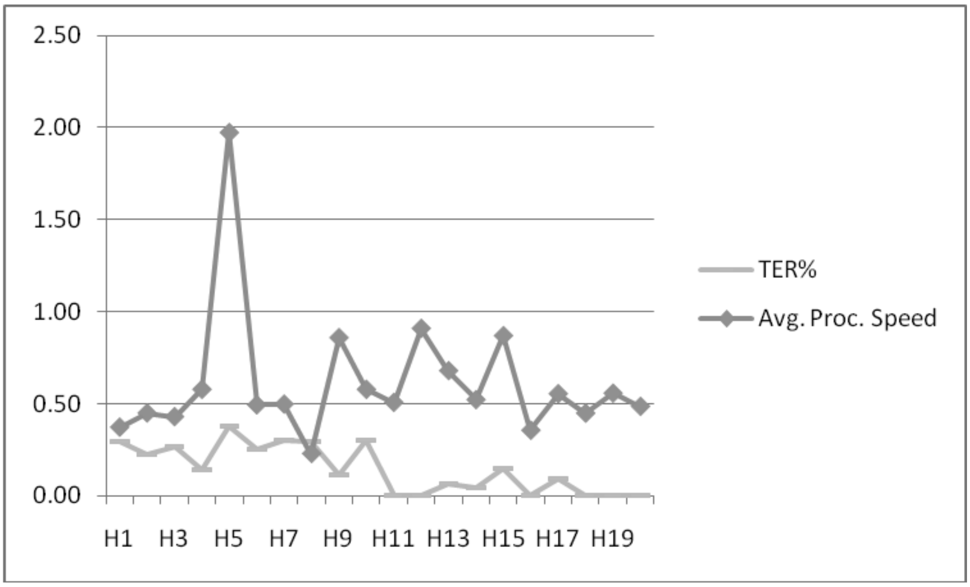

Fig. 6 Comparison between TER scores and average words per second for each segment classified as "High" (H1-H20)

In Figures 4 to 6, we would hope to see an opposite trend for processing speed vis-à-vis TER scores. This is evident for some individual segments (e.g. L2, L3, L8), where, when the TER score goes up the processing speed drops or vice versa, but this does not hold true for all segments. A Spearman's rho correlation[vii] confirms that there is little correlation on a segment level between TER scores and processing speed (-0.00 for "Low", -0.02 for "Medium", -0.01 for "High"). On the other hand, a Spearman's rho correlation for the group averages of TER scores and processing speed shows a high, negative correlation (-1.00, significant at the 0.01 level). A tentative conclusion from this is that automatic metrics like GTM and TER may be good predictors of post-editing speed for groups of segments lying within bands of scores, but they are not exact predictors on a segment-by-segment basis.

Average Fixation Time

Average Fixation Time is the average length of time spent fixating on an "area of interest" (AOI) within the Catalyst UI. For this study, the AOI was defined as the Edit, TM and Glossary Windows of the Alchemy Catalyst user interface (see Figure 1).

As mentioned earlier, time spent in fixations is seen to be a good indicator of cognitive effort. This measurement is used extensively in, for example, studies of reading effort (Rayner 1998; Radach et al. 2004). Generally, the longer the time spent in fixations, the higher the level of cognitive processing.[viii]

Before examining the results for Average Fixation Time, we first need to look at the average sentence length for each category. Table 1 reveals that there are some differences between sentence lengths:

	Low GTM	**Medium GTM**	**High GTM**
Average No. Words	6.55	17.85	19.2
Standard Deviation	5.14	9.13	15.64
Average No. of Characters (including spaces)	39.8	107.6	92.65
Standard Deviation	31.17	51.87	39.73

Table 1: Average word and character counts for Low, Medium and High GTM segments

Segments falling into the "Low GTM" category have a lower number of words on average than the other two.[ix] It is worth recalling here that previous research (especially in the field of Controlled Language) has shown that short segments (usually defined as four words or fewer) can be problematic for MT (cf. Gdaniec 1994; Underwood and Jongejan 2001; O'Brien 2003) and this might explain why the low-scoring segments have a relatively low average sentence length.

Given this difference in average sentence length across categories, we examine Average Fixation Time *per word* (Figure 7). The average time (in seconds) spent fixating in each category is divided by the total number of source words in that category (131 for "Low", 357 for "Medium" and 384 for "High").

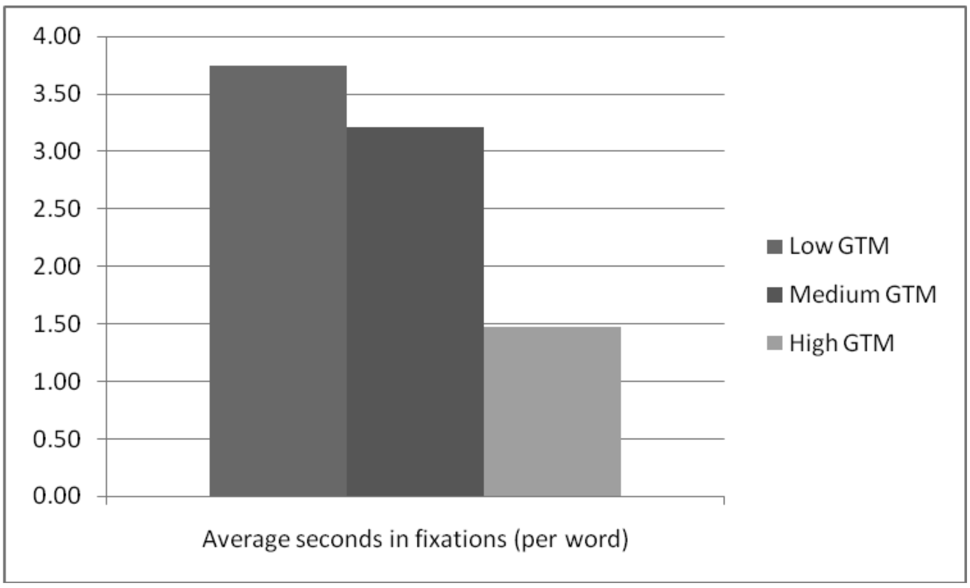

Fig. 7 Average fixation time per word

When normalised by sentence length, we can see that the average time spent in fixations is highest for segments in the "Low GTM" category, followed by "Medium" and then by "High". These results concur with those for processing speed.

Fixation Count

Data for fixation counts are presented in Table 2 (per participant) and Table 3 (averages across participants). In general, the more fixations there are, the more processing effort is assumed to be taking place.

	GTM – Low	GTM – Medium	GTM - High
Participant A	1050	2538	1270
Participant B	809	1940	1498
Participant C	594	1500	1400
Participant D	602	1400	846
Participant E	1043	2468	938
Participant F	984	2279	1163
Participant G	687	1630	1145

Table 2: Fixation counts per participant

	GTM – Low	GTM – Medium	GTM - High
Mean	824	1965	1088
Median	809	1940	1145
Standard Deviation	202	470	257

Table 3: Average fixation counts

While it would appear that the Medium category required more fixations than the other two, we again need to take into account the differences in sentence lengths across categories. Figure 8 presents average fixation counts normalised for sentence length.

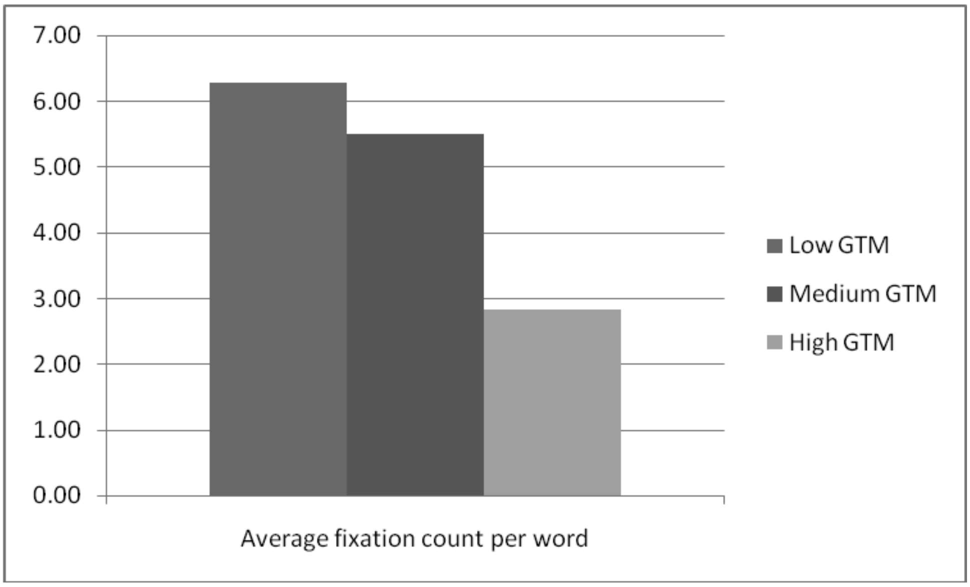

Fig. 8 Average fixation count per word for each category

As with the measurements for processing speed and average fixation time, we see that the "Low GTM" category requires the most fixations per word (and in theory then also the most effort) and "High" the least.

It is worth noting that for each of these measures of productivity, the "Medium" category was closer to the measures for "Low" than for "High". The possible reasons for this are explored below.

Summary and Discussion of Measurements

Average processing speed, fixation time and fixation count per word suggested significant correlations with the GTM categories of *Low*, *Medium* and *High*. In addition, a correlation was seen between TER scores and processing speed for each category. We could tentatively conclude from this that the GTM and TER scores are both fair predictors of post-editing productivity. However, as shown in Figures 4, 5, and 6, TER is a more reliable predictor for a band of scores than for individually scored segments. The implications here are that if MT users want to estimate productivity based on such automatic metric scores (or on confidence scores), they may have to do this on the level of groups of segments rather than on individual segment scores.

For average processing speed, fixation time and fixation count per word, we noted that the results for the "Medium" category were closer to that of "Low" than to that of "High". We can discount average sentence length as a factor here since the average for the “Medium” category (17.85) was closer to that of “High” (19.20), yet the productivity for the former was closer to that of the “Low” category (average 6.55). It is possible that the arbitrary GTM score band selected for the category “Medium” needs revision upwards and that this GTM score band cannot be considered, from a post-editing productivity viewpoint, to represent "Medium" effort. This needs further study. Another potential explanation is that segments that receive high GTM or TER scores (defined here as 0.81-1 for GTM or 0.2-0 for TER) are exponentially easier to post-edit than segments receiving lower scores.

Conclusions and Suggestions for Further Research

General Conclusions

Processing speed, average fixation time and fixation count per word were shown to correlate well with the GTM and TER bands of scores in this study. This provides some evidence that there are reasonable correlations between machine translation automatic metrics (at least GTM/TER) and *actual* post-editing productivity and, subject to further research and confirmation, means that we can rely on metrics like GTM/TER to reasonably reflect post-editing productivity. However, their accuracy on the level of the individual segment is open to question. There is some evidence that the post-editing effort required for high-scoring segments is exponentially lower compared to lower scoring segments.

Limitations

This study was exploratory in nature and the findings are consequently limited to one language pair and direction, one domain, one MT system and two automatic metrics. In addition, although seven professional translators with a similar profile were a very valuable asset for this study, the study “population” is quite limited. We would suggest that scaled up research is required before any generalisations can be made. We have not analysed the quality of the post-edited products for this project, but this analysis will be carried out and it is our intention to publish the results. Since eye tracking technology was used in this study, the task was controlled to avoid confounding variables (e.g. there was a relatively short task time, no second-pass revision time was factored in, see O’Brien 2009 for further details on restrictions associated with eye tracking studies). This would have had some impact on the nature of the task, though we feel that the impact would not be significant enough to invalidate the results obtained.

Further Research

Control for sentence length

While distribution of sentence length was controlled for when selecting the 10,000-word corpus, it was not specifically controlled for in the 995-segment or 60-segment corpus. It transpired that the segments scoring low GTM scores in both the 995- and 60-segment corpus had significantly shorter sentence lengths when compared with the other two categories. We have compensated for this by analysing measures on a per word or character basis. Future research could apply more control for sentence length. Additionally, the general relation between GTM/TER scores and sentence lengths also requires further investigation.

Moving from Reference Sentences to Confidence Scores

One of the primary drawbacks of automatic metrics is that they need a human-generated translation against which to compare the raw MT output. Therefore, their use is limited to cases where corpora of previous translations already exist. The metrics are, therefore, primarily used to measure advances in system development, to compare systems, or to tune systems (Way 2009: 27) rather than to predict the quality produced by the MT system and to estimate productivity. As mentioned in the Introduction, this issue is being tackled by commercial MT system developers, such as Asia Online, and by researchers whose MT systems are now producing "confidence scores" without the use of reference sentences (Specia et al. 2010, 2009a, 2009b; Bach et al. 2008, Blatz et al. 2004). This field is known as "Confidence Estimation" (Specia et al. 2009a). Confidence scores are generated by the MT system, based on information accumulated during the machine translation cycle. Examples of the features used to generate a confidence score are: source and target sentence lengths and their ratios; source and target sentence type/token ratios; source and target percentages of numbers, content words and non-content words (for a fuller list, see Specia et al. 2009a: 138). The score will indicate how “confident” the MT system is that the raw output is of “good” or “bad” quality. Assuming such confidence scores correlate well with post-editing productivity, they could be used to plan for the amount of time required to post-edit to a specified level of quality. This is where the vacuum exists: only a small body of research exists reporting on correlations between post-editing productivity and automatic measures such as BLEU, GTM, TER etc. and very little has to date been done on confidence estimation and actual post-editing productivity, where the trend has been to guess at how much effort might be required rather than to actually measure the effort. It is this vacuum that we hope to at least have partially filled here by examining correlations between an automatic metric and post-editing productivity. The next steps would be to scale this research up to cover other language pairs and MT engines, but also to test correlations between the MT-system generated confidence scores and actual post-editing effort.

Testing and Tuning Thresholds

We have speculated above that segments given high GTM or TER scores are exponentially easier to post-edit. This hypothesis needs further testing. It would be very useful to know what the threshold score is for segments that are exponentially easier to post-edit. This knowledge might be useful not only in the technical translation and localisation sectors, but could also be used by individual translators (whether professional or novice) to personalise (or tune) their own tolerance threshold for MT quality. Tuning may be relevant for taking account of post-editor experience levels, as suggested by Specia et al 2009a, where only very good quality output would actually improve the fastest translators’ productivity whereas even medium quality output might help a less experienced post-editor.
It is to be expected that not only level of expertise, but also translation context and quality expectations would all have an impact on the level of acceptability of the raw MT quality.

References

Alves F, Pagano A, da Silva I (2009) A new window on translators' cognitive activity, in Inger Mees, Fabio Alves and Susanne Göpferich (eds), *Methodology, technology and innovation in translation process research*, Copenhagen Studies in Language (38), Copenhagen: Samfundslitteratur, 267-291

Bach N, Gao Q, Vogel S (2008) Improving word alignment with language model based confidence scores, in *Proceedings of the Third Workshop on Statistical Machine Translation,* Columbus, Ohio, 19 June, New Jersey: Association for Computational Linguistics, 151-154

Banerjee S, Lavie A (2005) METEOR: an automatic metric for MT evaluation with improved correlation with human judgments, In *ACL-2005, Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, University of Michigan, Ann Arbor, 29 June, 65-72

Blatz, John, Fitzgerald E, Foster G, Gandrabur S, Goutte C, Kulesza A, Sanchis A, Ueffing N (2004) Confidence estimation for machine translation, *Coling 2004: 20th International Conference on Computational Linguistics*, 23-27 August, University of Geneva,Switzerland, Proceedings, pp.315-321

Callison-Burch C, Fordyce C, Koehn P, Monz C, Schroeder J (2008) Further meta-evaluation of machine translation, in *Proceedings of ACL-08: HLT. Third Workshop on Statistical Machine Translation,* June 19, The Ohio State University, Columbus, Ohio, USA (ACL WMT-08), 70-106

de Almeida G, O'Brien S (2010) Analysing post-editing performance: Correlations with years of translation experience, in *Proceedings of the 14th Annual Conference of the European Association for Machine Translation*, St. Raphaël, France, 27-28 May

Doddington, G (2002) Automatic evaluation of machine translation quality using n-gram co-occurrence statistics, In *Proceedings of the Second International Conference on Human Language Technology Research - HLT 2002*, March 24-27, San Diego, California,138-145

Du J, He Y, Penkale S, Way A (2009) MaTrEx: the DCU MT system for WMT 2009, in *Proceedings of the Fourth Workshop on Statistical Machine Translation*, Morristow, N.J.: Association for Computational Linguistics, 95-99

Duchowski, A (2003) *Eye tracking methodology: Theory and practice*, New York: Springer

Fellbaum C, (ed) (1998) *WordNet: An Electronic Lexical Database,* Cambridge, MA: MIT Press

Forcada M (2010) http://www.computing.dcu.ie/~mforcada/fosmt.html (accessed May 7, 2010)

Garcia I (2009) Beyond translation memory: Computers and the professional translator, *The Journal of Specialised Translation*, 12, [Online]

Gdaniec, Claudia (1994) The Logos Translatability Index, in *Proceedings of the First Conference of the Association for Machine Translation in the Americas*, 5th-8th October, Columbia, Maryland, USA, pp. 97-105

Göpferich S, Jakobsen A L, Mees I (eds) (2008) *Looking at eyes: Eye-tracking studies of reading and translation processing*, Copenhagen Studies in Language 36, Copenhagen: Samfundslitteratur

Groves , D, Schmidtke D (2009) Identification and analysis of post-editing patterns for MT, in *Proceedings of theTwelfth Machine Translation Summit*, August 26-30, Ottawa, Ontario, Canada, 429-436

Guerberof A, (2009) Productivity and quality in MT post-editing, *MT Summit XII - Workshop: Beyond Translation Memories: New Tools for Translators* August 29, Ottawa, Ontario, Canada, 8pp

Jensen K, Sjørup A, Winther Balling L (2009) Effects of L1 syntax on L2 translation, in Inger Mees, Fabio Alves and Susanne Göpferich (eds), *Methodology, technology and innovation in translation process research*, Copenhagen Studies in Language (38), Copenhagen: Samfundslitteratur, 319-336

Just M, Carpenter P (1980) A theory of reading: from eye fixation to comprehension, *Psychol Rev*, 87, 329–354

King M, Popescu-Belis A, Hovy E (2003) FEMTI – Creating and using a framework for MT evaluation, in *Proceedings of the Ninth Machine Translation Summit*, 23-27 September, New Orleans, USA, 224-231

Koehn, P (2010) Enabling monolingual translators: Post-editing vs. options, in *Proceedings of NAACL HLT 2010: Human Language Technologies – the 2010 annual conference of the North Americal Chapter of the Association for Computational Linguistics*, June 2-4, Los Angeles, California, 537-545

Krings H P (2001) *Repairing texts: Empirical investigations of machine translation post-editing processes*, Trans. G.S. Koby. Kent, Ohio: The Kent State University Press.

Lavie A, Przybocki M (eds) (2009) *Automated metrics for machine translation evaluation – Special issue of Machine Translation*, 23, 2/3, Amsterdam: Springer

Ma X, Cieri, C (2006) Corpus support for machine translation at LDC, in *Proceedings of the Fifth International Conference on Language Resources and Evaluation*, Genoa, Italy, 22-28 May, 859-864

McElhaney T, Vasconcellos M (1988) The translator and the postediting experience, in Vasconcellos, Muriel (ed.), *Technology as translation strategy*, American Translators Association Scholarly Monograph Series, Vol. II, State University of New York at Binghamton (SUNY), 140-148

Mees I, Alves F, Göpferich S (eds) (2009) *Methodology, technology and innovation in translation process research – A tribute to Arnt Lykke Jakobsen*, Copenhagen Studies in Language 38, Copenhagen: Samfundslitteratur

NIST (2010) *The NIST metrics for machine translation 2010 challenge (MetricsMATR10)*, National Institute of Standards and Technology (America), http://www.nist.gov/itl/iad/mig/upload/NISTMetricsMATR10EvalPlan.pdf (Accessed: 20/04/2010)

O'Brien S (2009) Eye tracking in translation-process research: methodological challenges and solutions, in Mees, I, Alves F, Göpferich S (eds), *Methodology, technology and innovation in translation process research – A tribute to Arnt Lykke Jakobsen*, Copenhagen Studies in Language 38, Copenhagen: Samfundslitteratur, 251-266

O'Brien S, (2007) An empirical investigation of temporal and technical post-editing effort. *Translation and Interpreting Studies (tis)*, II(I), 83-136

O'Brien, Sharon (2003) Controlling Controlled English - An Analysis of Several Controlled Language Rule Sets, in *Proceedings of the Joint Conference combining the 8th International Workshop of the European Association for Machine Translation and the 4th Controlled Language Applications Workshop* (CLAW 2003), 15th-17th May, Dublin City University, Ireland, Dublin: Ireland, 105-114

Papineni K, Roukos S, Ward T, Zhu W (2002) BLEU: a method for automatic evaluation of machine translation, in *Proceedings of ACL-2002*: *40th Annual meeting of the Association for Computational Linguistics*, Philadelphia, July 2002, 311-318

Plitt M, Masselot F (2010) A productivity test of statistical machine translation post-editing in a typical localization context, *The Prague Bulletin of Mathematical Linguistics*, 93, 7-16

Snover M, Dorr B, Schwartz R, Micciulla L, Makhoul J (2006) A study of translation edit rate with targeted human annotation, in *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas,* August 8-12, Cambridge, Massachusetts, USA, 223-231

Specia L, Raj D, Turchi M (2010) Machine translation evaluation versus quality estimation, *Machine Translation* 24(1), 39-50

Specia L, Saunders C, Turchi M, Wang Z, Shawe-Taylor J (2009a). Improving the confidence of machine translation quality estimates, in *Proceedings the Twelfth Machine Translation Summit*, August 26-30, Ottawa, Ontario, Canada, 136-143

Specia L, Cancedda N, Dymetman M, Turchi M, Cristianini N (2009b) Estimating the sentence-level quality of machine translation systems, in *Proceedings the Thirteenth Annual Conference of the European Association for Machine Translation*, May 14-15, Barcelona, Spain, 28-35

Radach R, Kennedy A, Rayner K (2004) *Eye movements and information processing during reading*. Hove: Psychology Press.

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

Takako A, Schwartz L, King R, Corston-Oliver M, Lozano M (2007) Impact of controlled language on translation quality and post-editing in a statistical machine translation environment, in *Proceedings of the Eleventh Machine Translation Summit*, 10-14 September, Copenhagen, Denmark, 1-7

Tatsumi M, (2009) Correlation between automatic evaluation scores, post-editing speed and some other factors, in *Proceedings of MT Summit XII*, Ottawa, Canada, 26-30 August 2009, 332-339

Turian J, Shen L, Melamed I.D. (2003) Evaluation of machine translation and its evaluation, in *Proceedings of the MT Summit IX*, New Orleans, USA, 23-27 September 2003, 386-393

Underwood, Nancy and B. Jongejan (2001) Translatability Checker: A Tool to Help Decide Whether to Use MT, in Maegaard, B. (ed), *Proceedings of the MT Summit VIII: Machine Translation in the Information Age*, 18-22 September, Santiago de Compostela, Spain, pp. 363-368

Way A, (2009) A critique of statistical machine translation, In Daelemans, Walter and Véronique Hoste (eds), *Evaluation of translation technology - Linguistica Antverpiensia new series, themes in translation studies*, 8, 17-41

White J, (2003) How to evaluate machine translation', in H. Somers (ed.) *Computers and Translation – A Translator's Guide*, Amsterdam, Philadelphia: John Benjamins, 211-244

Acknowledgements

This research was supported by the Science Foundation Ireland (Grant 07/CE/I1142) as part of the Centre for Next Generation Localisation (www.cngl.ie) at Dublin City University. Particular gratitude is extended to Dr. Yanjun Ma (previously of CNGL) and Ms. Giselle de Almeida (School of Applied Language and Intercultural Studies, DCU) who both participated in this research project.

---

[i] www.cngl.ie

[ii] Indeed, this is a difficulty in general with automatic metrics: one can positively say that one number on the scale is higher or lower than another, but few (if anyone) can say what the numbers *mean* in terms of post-editing effort, comprehensibility, acceptability etc.

[iii] The test sample was limited to 60 segments due to funding limitations (payment for translators) and due to the fact that the analysis of eye tracking data is labour intensive and time-consuming.

[iv] Heatmaps are normally produced in colour, so the red, yellow and green areas are easy to see. However, for publication purposes we converted this graphic to grey-scale.

[v] Using paired samples, two-tailed t-test.

[vi] We show only TER scores vs. processing speed here and do not present the graphs for GTM vs. processing speed for economy of space.

[vii] We use Spearman's rho instead of Pearson's product here as the data are not normally distributed.

[viii] We say "generally" here because research in information processing and reading tends to implement more complex measures of effort, including number of regressions (returns to a part of the sentence that has already been read) and saccade size (the measurement in degrees of the size of the jump from one fixation to the next). Also, effort in reading is known to be influenced by many factors, such as the frequency of the fixated word, how many meanings a word has, how familiar it is etc). Such a complex analysis is beyond the scope of this paper.

[ix] An examination of the average word counts for each category in the larger corpus of 995-segment corpus shows the following averages: Low=5.67 words; Medium=13.53; High=10.96. There are differences in average sentence length between this corpus and the 60-segment corpus. However, the trend whereby segments scoring Low GTM values are significantly shorter is mirrored.