# Can MT Output be Evaluated Through Eye Tracking?

**Stephen Doherty**
Centre for Next Generation Localisation
School of Applied Language and Intercultural Studies
Dublin City University
stephen.doherty2@mail.dcu.ie

**Sharon O'Brien**
Centre for Next Generation Localisation
School of Applied Language and Intercultural Studies
Dublin City University
sharon.obrien@dcu.ie

## Abstract

This paper reports on a preliminary study testing the use of eye tracking as a method for evaluating machine translation output. 50 French machine translated sentences, 25 rated as excellent and 25 rated as poor in an earlier human evaluation, were selected. 10 native speakers of French were instructed to read the MT sentences for comprehensibility. Their eye gaze data were recorded non-invasively using a Tobii 1750 eye tracker. They were also asked to record retrospective protocols while watching a replay of their eye gaze reading data. The average gaze time and fixation count were found to be significantly higher for the "bad" sentences, while average fixation duration was not significantly different. Evaluative comments uttered during the retrospective protocols were also found to agree to a satisfactory degree with previous human evaluation. Overall, we found that the eye tracking method correlates reasonably well with human evaluation of MT output.

**Key words:** MT Evaluation, user-based evaluation, eye tracking, gaze time, fixation count, fixation duration, retrospective protocols

## 1 Introduction

In this paper we report on a preliminary study of the suitability of eye tracking for measuring the ease with which machine translation output can be read. The focus of this paper lies firmly in the testing of methodology, rather than in the evaluation of specific MT outputs or systems.

Eye tracking is a method which records a subject's eye movements on screen as s/he is reading text, translating, or interacting with a software program. It has been used for many years in the investigation of, for example, reading patterns (Rayner, 1998) and translation processes (O'Brien, 2006, 2008; Göpferich et al 2008), among other cognitive processes. To the best of our knowledge, it has not yet been used in the evaluation of Machine Translation output.

Eye tracking offers an interesting method for evaluation of MT output because it enables the measurement of the cognitive effort involved in reading the target text. Cognitive demands and end-user evaluation are both largely ignored in MT research. It is generally assumed that when a human evaluates MT output as "good", that output will be easily read and understood by the end user. However, little empirical research has been done to demonstrate that this is true. We do not suggest that eye tracking should replace traditional human evaluation based on adequacy, fluency or other criteria. However, we hypothesize that it is a relatively objective method for measuring the effect that MT output has on a target language reader. It offers the additional advantages that the "evaluator" does not have to be bi-lingual and requires no training in evaluation techniques or criteria and this opens up the possibilities of including real end users in MT system evaluation. Eye tracking may very well be faster as an evaluation technique since it does not necessitate a comparison and evaluation or ranking of source and target sentences, but rather requires the evaluator to simply "read".

The main assumption behind eye tracking is the so-called "eye-mind hypothesis" (Ball et al 2006), which assumes that when the eye focuses on an object (e.g. a word) the brain is engaged in some kind of cognitive processing of that word. In eye tracking investigations of reading, researchers typically measure the number of "fixations" and the duration of these fixations to gauge how difficult the reading process is. "Fixations" are defined as "eye movements which stabilize the retina over a stationary object of interest" (Duchowski, 2003: 43). Fixations are usually measured in milliseconds and the more there are and the longer they are, the more difficulty the reader is assumed to be experiencing. Reading researchers are also interested in "saccades", i.e. "rapid eye movements used in repositioning the fovea to a new location in the visual environment" (Duchowski, 2003: 44). However, the measurement of saccades was beyond the scope of our study.

We set out with one question in mind: Would eye tracking data reflect the quality of MT output as rated by human evaluators? Section 2 explains our methodology and Section 3 presents and discusses the results. Section 4 summarises our conclusions and outlines further research we intend to concentrate on.

## 2 Methodology

A human evaluation was conducted on MT output from English to French for a previous study on Controlled Language and the acceptability of MT output (Roturier, 2006). In this evaluation, four human evaluators were asked to rate output on a scale of 1-4 where 4 signified "Excellent MT Output", 3 signified "Good", 2 "Medium" and 1 "Poor". A full description of the evaluation criteria for that study is available in Roturier (2006) and is beyond the scope of this paper. However, it is important to emphasize that the evaluation was carried out in a commercial context where the focus was on how much editing effort would be required to bring the output to a commercially acceptable quality level. Adequacy and fluency were not deemed appropriate evaluation measures in that context. 25 of the lowest rated ('poor') and 25 of the best rated ('excellent') sentences, according to four human evaluators, were selected from that corpus. We assumed that the highest rated sentences would be easier to read than the lowest rated ones. While this might seem like a trivial statement, little (if any) empirical research has been carried out to investigate it. Presumably, the ease with which sentences could be read and understood impacted on the scores given previously by the human evaluators.

10 native speakers of French were recruited to read the machine translated sentences (12 were recruited and two were dropped out due to poor quality data). The sentences came from the domain of documentation describing virus checking software. While this domain is quite obviously specialised, the data are taken from a successful commercial application of MT, and this was considered to be in keeping with our interests in user-focused evaluation. The participants were native speakers of the target language and, as such, counted as potential end users, but they were not experts in this domain and this was a deliberate choice on our part since prior knowledge of a domain has been shown to ease the reading experience (Kaakinen, Hyönä & Keenan 2003). By not having deep prior knowledge of the domain, we assumed that participants would have to make an effort to read and understand the sentences and that we would then be better able to differentiate between the sentences that were easy to read and those that posed more difficulty. Given that our focus was on testing the methodology, we gave higher priority to the "no prior knowledge" condition over "authentic" end-users. All participants were enrolled at the time of the study as full-time or exchange students in Dublin City University, some on translation programmes and others on business programmes.

The sentences were presented in a tool called Translog. Translog was originally developed for researching human translation processes (Jakobsen 1999), but has recently been altered to interface with a gaze-to-word mapping tool (GWM), developed within the EU-funded Eye-to-IT project (http://cogs.nbu.bg/eye-to-it/). Eye trackers can be somewhat inaccurate in the mapping of eye movements onto words and the GWM tool was developed to help remap words to fixations which were not successfully mapped by the eye tracker (Carl 2008). We do not present data from the GWM tool here, but have identified future research using this tool (see Conclusions).

The participants were first given a self-paced warm-up task, after which they were presented with sentences from our data set to read one by one. The sentences were presented in a random order (i.e. 'bad' and 'good' sentences were mixed, but presented in the same order for all participants) and participants were not aware that sentences had already been rated. They were asked to read the sentences for comprehension and, since motivation is an important factor in reading (Kaakinen et al. 2003), were informed that they would be asked some questions at the end to see if they had understood the text. They were told to press the "Return" key when they wanted to move to the next sentence and no time pressure was applied.

We used the Tobii 1750 eye tracker to monitor and record the participants' eye movements while reading. This eye tracker has inbuilt infra-read diodes which bounce light off the eyes. It records the position of the right and left eyes according to the X, Y coordinates of the monitor, as well as the length and number of fixations, gaze paths, and pupil dilations. During this study a fixation was defined as lasting at least 100 milliseconds. The Tobii 1750 is a non-invasive eye tracker (i.e. participants do not have to wear head-mounted equipment or use head rests or bite bars) and it compensates for head movement. While the non-invasive nature helps to relax participants and, presumably, allows them to behave more normally, the lack of control leads to some level of inaccuracy in the data. We attempted to compensate for this by using the retrospective think-aloud protocol method to supplement the eye tracking data.

The analysis software we used to analyse the eye tracking data was ClearView (version 2.6.3). ClearView also produces an AVI of the reading session, which displays the eye movements and fixations for each participant. This was played back to the participants immediately after the session in Camtasia Studio (screen recording software) and they were asked to comment on their reading behaviour. This commentary was recorded.

The retrospective protocols were transcribed and classified into "positive", "negative", "mixed" comments and two additional categories of "silent" and "N/A" were also applied. For example, "*It's ok*" (referring to the entire sentence) was classified as a positive comment. An example of a negative comment is "*Emm… this is kinda weird in French. 'Le premier est a l'aide des fichiers d'aide' emm… I don't really get the meaning*". A mixed comment contained some positive and some negative comments, e.g. "*It's ok. It should be 'les fichiers' and not 'des'. Yeah.*". These protocols were used to help us understand what the participants were actually thinking while they were (re-)reading. We were also interested in measuring how well the positive, negative, and mixed comments mapped onto the ratings of the human evaluators from the previous study.

To conclude this section, the measures we were interested in included average gaze time, fixation count and duration per sentence and per character, retrospective comments and their correlations with the previous human evaluation. Our results are presented in Section 3.

# 3 Results

## 3.1 Gaze Time

Gaze time is the period of time a participant spends gazing within an Area of Interest (henceforth AOI). For this study, the AOIs were defined around each sentence in order to allow for all possible data relating to the sentence to be captured (a minimum of 5cm radius around each letter/word) and to exclude unwanted data, e.g. looking at the toolbar or clock. The total gaze time per participant, given in minutes, is presented in Figure 1; the average total was 5.23 minutes (median = 5.06):

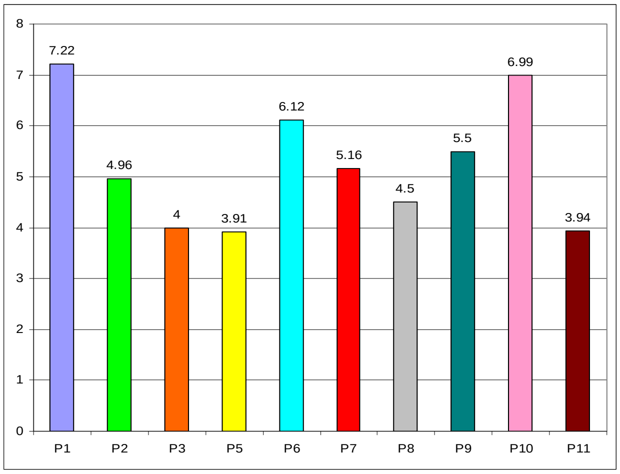

**Figure 1: Total Gaze Time for All Participants (in minutes)**

It is interesting to note the highest values (P1, P10, and P6 respectively) correspond to the three participants who had a strong language/translation background and who, according to the Think-Aloud data we present later, appear to have paid more attention to the text in terms of grammar, spelling, agreements etc. Analysis of the retrospective interview data supports this assumption in that these participants made several comments regarding their careful analysis of the segments. As an example of this detailed reading, P1 commented on carefully checking agreements of nouns with their corresponding adjectives, and observing the Passé Composé rule of French grammar.

Figure 2 shows the average gaze time per segment across all participants in milliseconds. As hypothesised, the 'bad' segments had longer gaze times than the 'good' segments.

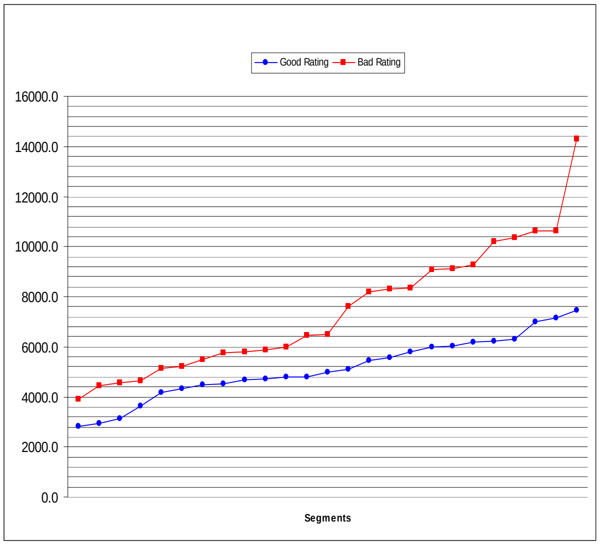

**Figure 2: Average Gaze Time for Good & Bad Segments for All Participants (in milliseconds)**

The average gaze time for good segments was 5124.7ms while that of the bad segments was higher at 7426.6ms. In other words, participants spent, on average, 45% more time looking at bad segments than good segments. Spearman's rho suggests a medium strength negative correlation between gaze time and sentence quality (ρ= -.46, p<0.01).

Obviously, some segments are longer than others. It therefore makes sense to examine the data according to the number of characters per segment. We first look at gaze time per character. As Figure 3 illustrates, a similar trend is evident in that the bad segments still had higher gaze time per character than the good segments. Additionally, when the average gaze time per character of all segments is taken into account (65.89 ms), we see that a majority of segments above this value were rated as bad (65% or 15 of 23).

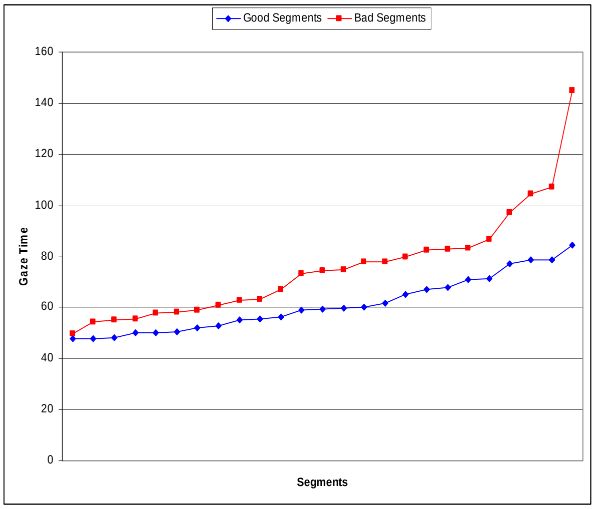

**Figure 3: Average Gaze Time for Good & Bad Segments per Character (in milliseconds across all segments)**

## 3.2 Fixation Count

Fixations occur when the eye focuses on a particular area of the screen. Fixations are defined according to the pixel radius and the minimum duration in milliseconds and the settings will vary depending on the object of study. In eye tracking studies of reading, in general, the pixel radius and minimum duration is lower than, for example, in usability research. However, there is no general agreement on how fixations should be defined. For our study, we used a fixation filter of 40 pixels x 100 milliseconds, which is the filter used in the Eye-to-IT project.

The fixation count shows the total number of fixations on a given sentence. Figure 4 shows the average fixation count per segment; a similar trend to that observed in the above figure of average gaze time per segment is evident, i.e. bad segments had, on average, more fixations than good segments. Spearman's rho suggests a medium strength negative correlation between fixation count and sentence quality (ρ= -.47, p<0.01).

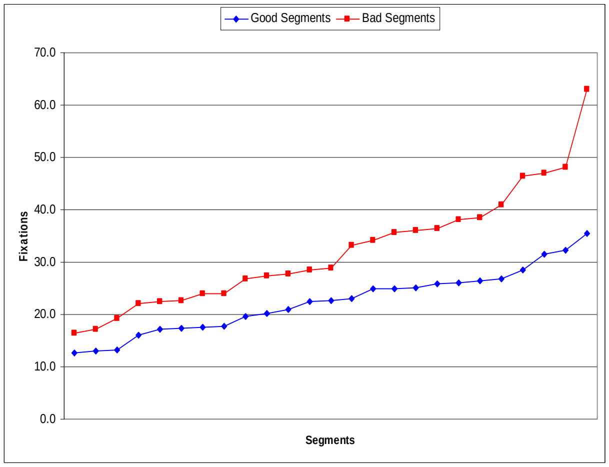

**Figure 4: Average Fixation Count per Segment**

When looking at the median (25.5) of the above average fixation count per segment we see that, out of the segments above the median, 8 segments were 'good', while 17 were 'bad'. As with total gaze time, the total number of fixations per participant is led by P1, P10, and P6 respectively. This may seem obvious as the longer the gaze time the more probable fixations are, but it is, nevertheless, worthy of note - Figure 5 illustrates:

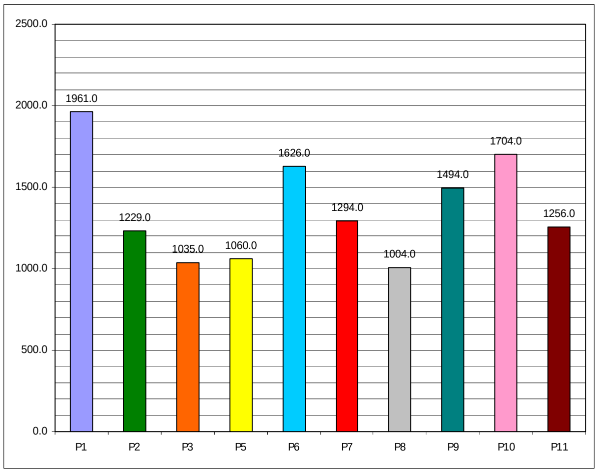

**Figure 5: Total Fixations for All Participants**

Moving on to fixation count per character, a similar and logical relationship to gaze time is observed. We see that, once again, the majority of the segments that had higher-than-average values were rated as bad (68% or 17 of 25). These results are shown in Figure 6:

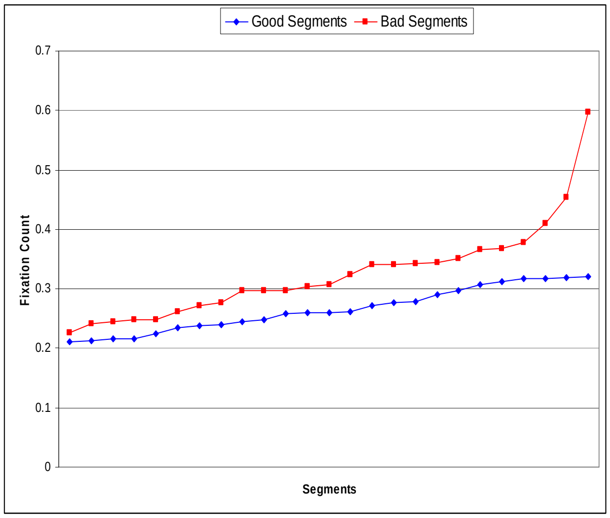

**Figure 6: Average Fixation Count for Good & Bad Segments per Character (in milliseconds across all segments)**

## 3.3 Average Fixation Duration

Average fixation duration has been used as an indicator of cognitive effort in many disciplines. When observing the average fixation duration across all segments and participants, it appears that the average fixation duration is quite similar in both good and bad segments, as Figure 7 illustrates:

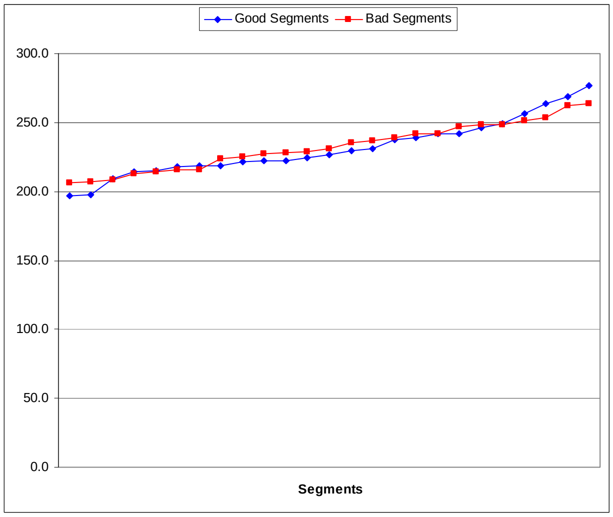

**Figure 7: Average Fixation Duration (milliseconds) for Good/Bad Segments for All Participants**

The presence of several good segments among the bad segments in the highest range of values for average fixation duration is surprising. An "acclimatisation effect" has been noted before in eye tracking studies (O'Brien 2006), where the initial cognitive effort is higher than for the rest of the task. In light of this, we omit the first five segments to see what effect it has on our Fixation Duration data. Figure 8 demonstrates the effect:

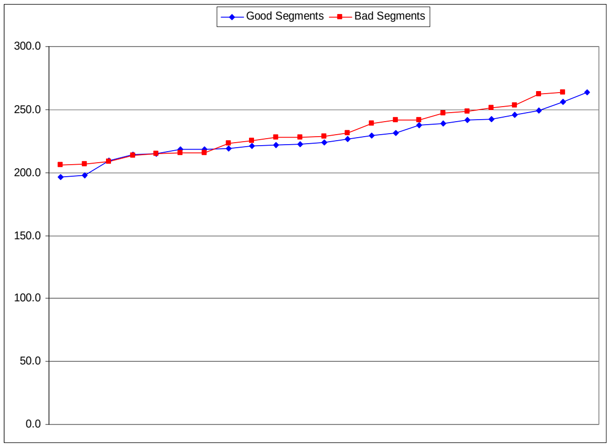

**Figure 8: Average Fixation Duration (ms) for All Participants from S6 to S50**

As we can see, the elimination of the first five segments has some effect on differentiating the good and bad segments, though the difference overall is still limited. Overall, it appears that the above measures correlate, for the most part, with the segment ratings. Figure 9 shows the fixation duration results per character for all fifty segments:

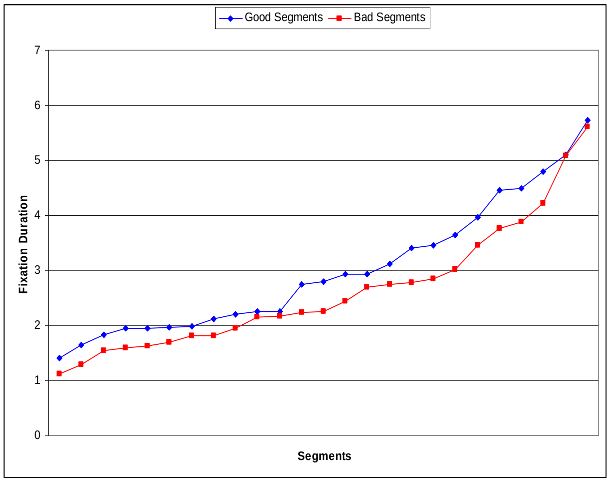

**Figure 9: Average Fixation Duration for Good & Bad Segments per Character (in milliseconds across all segments)**

While fixation duration per segment seems to be a reasonable indicator of good and bad MT output, when this measure is viewed per character, the trend is for bad segments to have shorter fixation durations than good ones and the differences were found to be non-significant.

The suitability of this measurement for predicting good and bad MT output therefore requires further investigation. This lack of differentiation in fixation duration reflects other studies. For example, O'Brien (forthcoming) found no significant difference in fixation duration for texts that had been edited using controlled language rules and versions that were uncontrolled. Jakobsen and Jensen (2009) also found insignificant differences in fixation duration across groups in translation process research. Additionally, Van Gog et al (2009: 328) suggest that while fixation duration is a useful measure of cognitive processing, it may reflect "different aspects of cognitive load".

## 3.4 Retrospective Data

A retrospective interview for each participant followed the completion of the main task. In this interview, participants were instructed to vocalise their thoughts on their reading patterns as presented to them through the Gaze Replay feature. The participants' comments were classified as follows: All Positive, All Negative, Mixed, Silence, and N/A. Mixed refers to a comment that had both good and bad reports and N/A was assigned when the participant made comments unrelated to the task.

In relation to good segments, we find that 47.2% were met with wholly positive comments. If we factor in the positive remarks in the "Mixed" comments then the participants agreed with the 'good' evaluation in 62.3% of cases. It should be noted here that at no point were participants aware of the original rating of the segments.

On examining the bad segments, we find that there was agreement with the initial evaluation in 54.5% of cases. If we factor in the mixed comments, as before, a value of 79.2% is reached.

Good and bad segments in relation to silence, i.e. no comment of any kind, give values of 15.3% and 9.6% respectively.

The initial human evaluation was a targeted evaluation where linguists were asked to rate MT output according to very specific criteria. In contrast, the "readers" in our study were not asked to rate the output, but to comment out loud on their thoughts as they were reading the sentences and as they viewed the gaze replays of their reading sessions. This type of task is much less targeted than the traditional human evaluation of MT output. Nonetheless, we feel that the transcriptions demonstrate a satisfactory level of correlation with the initial human evaluation.

## 4 Conclusions

Our initial question for this study was: Would eye tracking data reflect the quality of MT output as rated by human evaluators? We have shown that the gaze time and fixation count have correlated well with the previous evaluators' judgments on the segments in question. The differences in fixation duration results for both sentence types were less clear-cut, although this improves if we assume an acclimatisation effect and remove the first five initial segments.

Although the sample is small when number of sentences and participants is taken into account, we feel reassured that eye tracking methods for evaluating the readability and comprehensibility of MT data is worthy of further investigation. As mentioned in the Introduction, this would enable monolingual, objective, end-user evaluations, based on the cognitive effort associated with reading MT output. While we do not propose this as a replacement for traditional or automated human evaluation, nor as a faster, cheaper method, it nonetheless offers a new dimension in evaluating translations generated by MT, which gives insight into the cognitive effort involved on the part of genuine end users. Additionally, is it not unrealistic to imagine a system that could eventually learn to rank its own data based on the eye gaze data of human post-editors and end users.

It is our intention in the future to build on this research by increasing sample sizes, target languages, and domains.

The next step in this research will involve the aforementioned gaze-to-word mapping tool, where we will map fixations to specific words and then use a query language to investigate which words were fixated most and longest for both the good and bad sentences and which parts-of-speech were fixated and what differences, if any, exist between POS fixations and the good/bad sentences. We also intend to investigate correlations between automatic metrics, and eye tracking data, including pupil dilations, fixation durations and their suitability as measures of MT output quality.

## Acknowledgements

The authors wish to thank Dr. Johann Roturier, Principal Research Engineer at Symantec, Ireland, who gave permission to re-use human evaluated MT output. This research is 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.

## References

Linden J. Ball, Nichola Eger, Robert Stevens & Jon Dodd. 2006. Applying the Post-Experience Eye-Tracked Protocol (PEEP) Method in Usability Testing, *Interfaces*, 67:15-19.

Michael Carl. 2008. Framework of a probabilistic gaze mapping model for reading. Susanne Göpferich, Arnt L. Jakobsen, and Inger Mees (eds). *Looking at eyes – Eye Tracking Studies of Reading and Translation* Processing. Copenhagen Studies in Language 36, Copenhagen: Samfundslitteratur, 193-202.

Andrew Duchowski. 2003. *Eye-Tracking Methodology – Theory and Practice.* London: Springer-Verlag.

Susanne Göpferich, Arnt Lykke Jakobsen, Inger Mees (eds). 2009. *Looking at Eyes: Eye Tracking Studies of Reading and Translation Processing. Copenhagen Studies in Language 36, Copenhagen: Samfundslitteratur.*

Arnt Lykke Jakobsen and Kristian Jensen, 2009. Eye Movement Behaviour Across Four Different Types of Reading Task. Susanne Göpferich, Arnt L. Jakobsen, and Inger Mees (eds). *Looking at eyes – Eye Tracking Studies of Reading and Translation* Processing. Copenhagen Studies in Language 36, Copenhagen: Samfundslitteratur, 103-124.

Arnt Lykke Jakobsen, 1999. Logging Target Text Production with Translog. Gyde Hansen (ed). *Probing the Process in Translation: Methods and Results.* Copenhagen Studies in Language, 24. Copenhagen: Samfundslitteratur, 9-20.

Johanna Kaakinen, Jukka Hyönä & Janice Keenan. 2003. How Prior Knowledge, WMC, and Relevance of Information Affect Eye Fixations in Expository Text. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 29(3): 447-457.

Sharon O'Brien, (forthcoming). Controlled Language and Readability. Gregory Shreve and Erik Angelone (eds). *Translation and Cognition*, American Translators Association Scholarly Monograph Series, John Benjamins.

Sharon O'Brien, 2008. Processing Fuzzy Matches in Translation Memory Tools – an Eye-tracking Analysis. Susanne Göpferich, Arnt L. Jakobsen, and Inger Mees (eds). *Looking at eyes – Eye Tracking Studies of Reading and Translation* Processing. Copenhagen Studies in Language 36, Copenhagen: Samfundslitteratur, 79-102.

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

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

Johann Roturier, 2006. *An Investigation Into the Impact of Controlled English Rules on the Comprehensibility, Usefulness, and Acceptability of Machine-Translated Technical Documentation for French and German Users*. Unpublished PhD Dissertation. Dublin City University.

Tamara Van Gog, Liesbeth Kester, Fleurie Nievelstein, Bas Giesbers and Fred Paas, 2009. Uncovering Cognitive Processes: Different Techniques That Can Contribute to Cognitive Load Research and Instruction. *Computers in Human Behavior,* 25: 325-331.