Mach Translat (2010) 24:1–13
DOI 10.1007/s10590-010-9070-9

Eye tracking as an MT evaluation technique

Stephen Doherty · Sharon O'Brien · Michael Carl

Received: 9 May 2009 / Accepted: 9 February 2010 / Published online: 27 February 2010

Abstract Eye tracking has been used successfully as a technique for measuring cognitive load in reading, psycholinguistics, writing, language acquisition etc. for some time now. Its application as a technique for measuring the reading ease of MT output has not yet, to our knowledge, been tested. We report here on a preliminary study testing the use and validity of an eye tracking methodology as a means of semi-automatically 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. Ten 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. The average gaze time and fixation count were found to be higher for the "bad" sentences, while average fixation duration and pupil dilations were not found to be substantially different for output rated as good and output rated as bad. Comparisons between HTER scores and eye gaze data were also found to correlate well with gaze time and fixation count, but not with pupil dilation and fixation duration. We conclude that the eye tracking data, in particular gaze time and fixation count, correlate reasonably well with human evaluation of MT output but fixation duration and pupil dilation may be less reliable indicators of reading difficulty for MT output. We also conclude that eye tracking has promise as a semi-automatic MT evaluation technique, which does not require bi-lingual knowledge, and which can potentially tap into the end users' experience of machine translation output.

S. Doherty · S. O'Brien (✉)
Centre for Next Generation Localisation, School of Applied Language and Intercultural Studies, Dublin City University, Dublin, Ireland
e-mail: sharon.obrien@dcu.ie

M. Carl
Centre for Research and Innovation in Translation and Translation Technology, Copenhagen Business School, Frederiksberg, Denmark

Keywords MT evaluation · Eye tracking · Gaze time · Fixation count · Fixation duration · Pupil dilation · HTER

1 Introduction

In this paper we report on a preliminary study of the suitability of eye tracking methodologies for semi-automatically measuring the ease with which machine translation output can be read. Eye tracking is a method which records a person's eye movements across a screen as s/he is interacting with images or text on that screen. It has been used for many years to investigate different aspects of cognitive processing (e.g. reading, comprehension, bi-lingualism, cf. Rayner 1998), of cognitive load (e.g. in route planning and document editing tasks, cf. Iqbal et al. 2005), and for usability [e.g. in investigating the readability of online news as in the Stanford Poynter project (Stanford Poynter Project)]. More recently, it has been used as a supplementary method along with keyboard logging and think-aloud protocols to investigate human translation processes in general, and, more specifically, cognitive processing load when working with Translation Memory tools or interacting with sub-titled media (cf. O'Brien 2006, 2008, Caffrey 2008; Göpferich et al. 2009). To the best of our knowledge, it has not yet been used in the evaluation of Machine Translation output.

The main assumption behind eye tracking is the so-called "eye-mind hypothesis" (Ball Linden et al. 2006), which assumes that when the eye focuses on an object, for example a sentence, the brain is engaged in some kind of cognitive processing of that sentence.

In his extensive review of eye tracking research, Rayner (1998) summarises research findings that convincingly demonstrate that in complex information processing tasks, such as reading, there is a close link between gaze and attention.

In eye tracking investigations of reading (e.g. Kaakinen and Hyönä 2005; Hyönä and Nurminen 2006) researchers typically measure the reading time, 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, p. 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.

In addition to fixation measurements, cognitive load research typically also uses pupillometrics, i.e. measuring changes in the pupil diameter during task processing. Many studies have demonstrated reliable links between cognitive processing and changes in pupil dilation (e.g. Hess and Polt 1964; Nakayama et al. 2002; Iqbal et al. 2005). However, it is acknowledged that many factors can influence pupil dilation (e.g. lighting, sounds, caffeine, colour of eyes etc.) and pupil dilation has sometimes been found not to correlate well with other eye tracking measurements of cognitive processing (Schultheis and Jameson 2004; O'Brien 2008, Caffrey 2008).

Our primary research questions were: To what extent does eye tracking data reflect the quality of MT output as rated by human evaluators? And, related to this question, could eye tracking potentially be used as a tool for semi-automatically measuring MT quality? While this research cannot be said to involve fully automatic evaluation of

MT, as understood by the current research in the field of "automatic metrics" such as BLEU, NIST and other scores, it paves the way for the unobtrusive recording of MT reading effort, which could supplement or confirm automatic evaluation metrics. Section 2 explains our methodology and Sect. 3 presents and discusses the results. Section 4 summarises our conclusions and outlines further possible research.

2 Methodology

A human evaluation was conducted on rule-based MT output from English to French for a previous study on Controlled Language (CL) 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).[1] Twenty five of the lowest rated (denoted as 'bad' here) and 25 of the best-rated sentences (denoted as 'good' here), according to four qualified linguist evaluators, were selected from that corpus.

Since we had access to "gold" standards for the source text sentences, we calculated HTER scores with a view to testing correlations between these scores and the eye tracking data. It should be pointed out that the "gold" standards were not human translated, but post-edited versions of the raw MT output. While this differs from the usual approach in the research field, it is not unusual for *commercial* users of MT systems to use post-edited versions of MT output as "gold" standards since this is the quality level they wish to achieve in order to publish their machine translated material (Roturier 2009). Our corpus of sentences originated from one such commercial user.

The number of sentences was deliberately small since our main goal was to test eye tracking as an MT evaluation methodology and not to rate the MT output. We assumed that the highest rated sentences would be easier to read than the lowest rated ones. Likewise, we assumed that the ease with which sentences could be read and understood influenced the scores given previously by the human evaluators, even though they were not asked to pay attention specifically to "reading ease".

Ten 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. The participants were not experts in this domain and this was a deliberate choice since prior knowledge of a domain has been shown to ease the reading experience (Kaakinen et al. 2003). By not having deep prior knowledge of the domain, we assumed that participants would have to make an effort to construct an internal representation of the meaning of each sentence and that the effort to do so would be higher for the 'bad' sentences and this would, in turn, be reflected in our measurements.[2] All participants

[1] Roturier (2006) study did not use the common measurements of "adequacy" and "fluency" because the focus was on *how much post-editing effort* would be required to bring the MT output to a level acceptable by the commercial user of that specific MT system. As the evaluation focussed on perceived post-editing effort, the evaluation criteria outlined above were deemed to be more suitable.

[2] We draw here on Kintsch's (1998) Construction Integration (CI) theory which posits that the reader's background knowledge plays a crucial role in-text comprehension.

were enrolled at the time of the study as full-time or exchange students at Dublin City University, some on translation programmes and others on business and computer science programmes. Consequently, we assumed that our subjects had more or less equal reading ability. However, reading ability, reader type, prior knowledge, as well as working memory capacity may all influence reading behaviour (Daneman and Carpenter 1980; Kaakinen et al. 2003). It was beyond the scope of this study to measure the effects each of these variables might have on reading MT output. However, future studies could take some, if not all, of these factors into account.

The participants were first given a warm-up task. They were presented with five high quality sentences to read one by one. They were then presented with the test sentences 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 in a prior human evaluation task. 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 sentences. The sentences were presented in a tool called Translog. Translog was originally developed for researching human translation processes (Jakobsen 1999), but has recently been modified to interface with an eye-tracker and other tools developed within the EU-funded Eye-to-IT project (http://cogs.nbu.bg/eye-to-it/). The Translog tool allows text to be displayed in a window in a similar fashion to a text editor. The participants pressed the "Return" key when they wanted to move to the next sentence and no time pressure was applied. The sentences were read in isolation for two main reasons: (i) it is easier to measure fixation count, duration, pupil dilation etc. when only one sentence appears on the screen at any one time. This allowed us to increase measurement validity, but obviously reduced ecological validity since readers normally read "text" rather than isolated sentences; (ii) this scenario reflected the initial human evaluation where individual sentences (and not whole texts) were evaluated. As the focus here was on fluency, only the MT output was presented and not the reference translation, therefore, adequacy was not considered; this allows for monolingual MT evaluation (see Sect. 4).

We used the Tobii 1750 eye tracker to monitor and record the participants' eye movements while reading. This eye tracker has built-in 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). While the non-invasive nature increases the validity of the online reading experience, the lack of control leads to some level of inaccuracy in the data. We attempted to compensate for this by using a retrospective think-aloud protocol method, which provided useful supplementary data (a full report is beyond the scope of this paper). Experimental conditions such as distance from monitor, temperature, noise, and lighting were kept constant.

The analysis software we used to analyse the eye tracking data was ClearView (version 2.6.3). ClearView also produces an AVI (video file) of the reading session, which displays the eye movements and fixations for each participant overlaid on the text. This was played back to the participants immediately after the session in Camtasia

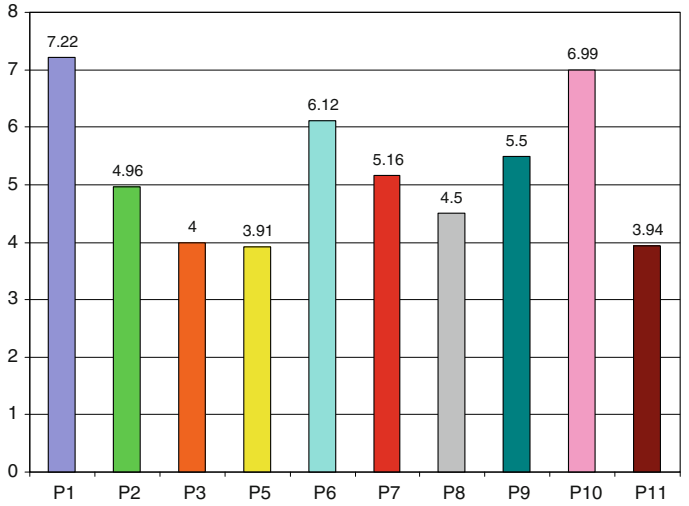

Fig. 1 Total gaze time for all participants (in minutes)

Studio (screen recording software) and they were asked to comment on their reading behaviour. This commentary was recorded.

To conclude this section, the measures we were interested in included average gaze time, fixation count and duration per sentence and per character for the two sets of sentences, average pupil dilation for both sentence types, HTER scores and their correlations with the eye tracking data. Our results are presented in Sect. 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 capture all possible data relating to the reading of the sentence. The total gaze time per participant, given in minutes, is presented in Fig. 1; the average was 5.23 min (median = 5.06):

Figure 2 shows the average gaze time per sentence across all participants in milliseconds. As hypothesised, the 'bad' sentences had longer gaze times than the 'good' sentences.[3]

The average gaze time for good sentences was 5124.7 ms while that of the bad sentences was higher at 7426.6 ms. In other words, participants spent, on average, 45% more time looking at bad sentences than good sentences. Spearman's rho

[3] Figures 2 to 8 are represented using box plots. The eye-tracking metric in question is shown on the y-axis and the good/bad sentence variables are shown on the x-axis. The antennae represent the range of values; the coloured/shaded box shows the standard deviation and also contains the mean.

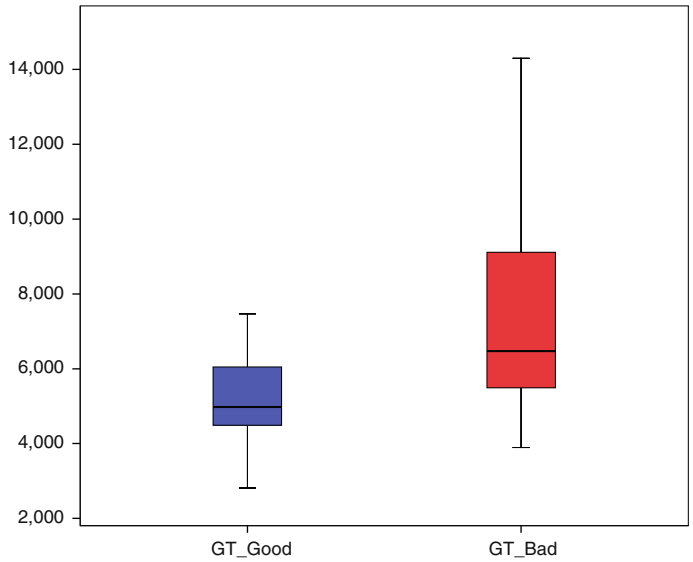

Fig. 2 Average gaze time for good and bad sentences for all participants (in milliseconds)

suggests a medium strength negative correlation between gaze time and sentence quality ($p = -0.46$, $p < 0.01$).

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

It is interesting to note that the average sentence length for good sentences was 85 characters (median = 78, SD = 28) and bad sentences had a value of 103 characters (median = 97, SD = 40). We therefore need to examine good and bad sentences of similar lengths: if we take the mean character length for good and bad sentences combined (94) and the standard deviation (36) and examine a group of 10 good and 10 bad sentences that fall within the standard deviation of the mean, we find that the sentences rated as bad still have a higher median gaze time (7256.9 ms vs. 5190.3 ms for good sentences) and a higher fixation count (see below) 89.3 vs. 88.1).

3.2 Fixation count

Fixations occur when the eye focuses on a particular area of the screen. Fixations are defined according to (i) pixel radius and (ii) the minimum duration in milliseconds and the settings will vary depending on the object of study. For our study, we used a fixation filter of 40 pixels × 100 milliseconds, which is the filter used in the Eye-to-IT project.

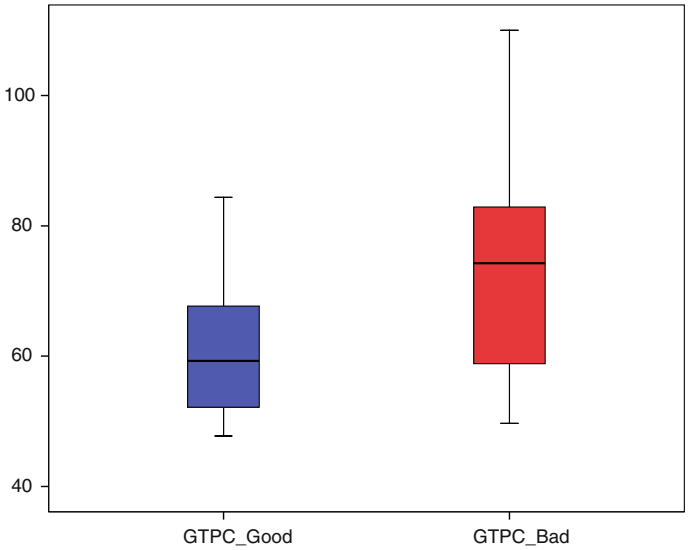

Fig. 3 Average gaze time for good and bad sentences per character (in milliseconds)

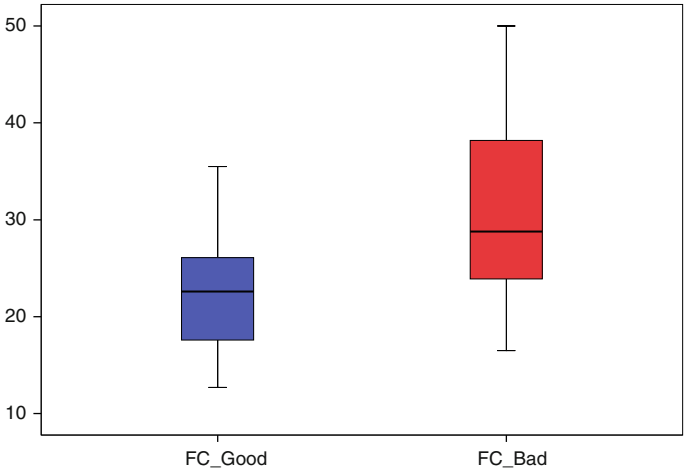

Fig. 4 Average fixation count per sentence (number of fixations)

The fixation count shows the total number of fixations on a given sentence. Figure 4 shows the average fixation count per sentence; a similar trend to that observed in the above figure of average gaze time per sentence is evident, i.e. bad sentences had, on average, more fixations than good sentences. Spearman's rho suggests a medium strength negative correlation between fixation count and sentence quality ($p = -0.47$, $p < 0.01$).

When looking at the median (25.5) of the above average fixation count per sentence we see that, out of the sentences above the median, eight sentences were 'good', while 17 were 'bad'.

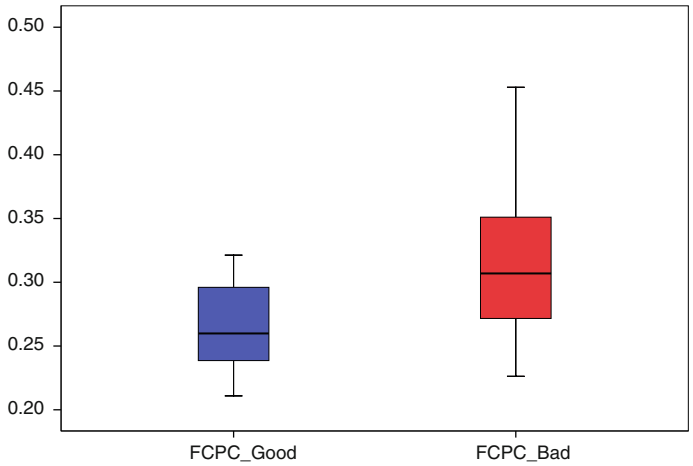

Fig. 5 Average fixation count for good & bad sentences per character (in milliseconds)

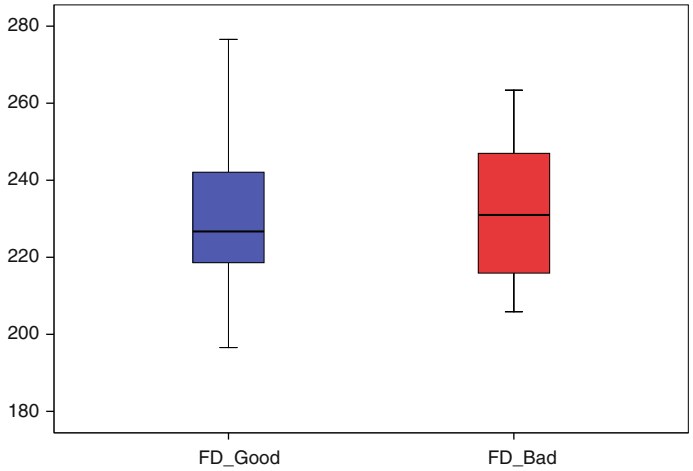

Fig. 6 Average fixation duration (milliseconds) for good/bad sentences for all participants

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 sentences that had higher-than-average values were rated as bad (68% or 17 of 25). These results are shown in Fig. 5.

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 sentences and participants, it appears that the average fixation duration is quite similar in both good and bad sentences, as Fig. 6 illustrates.

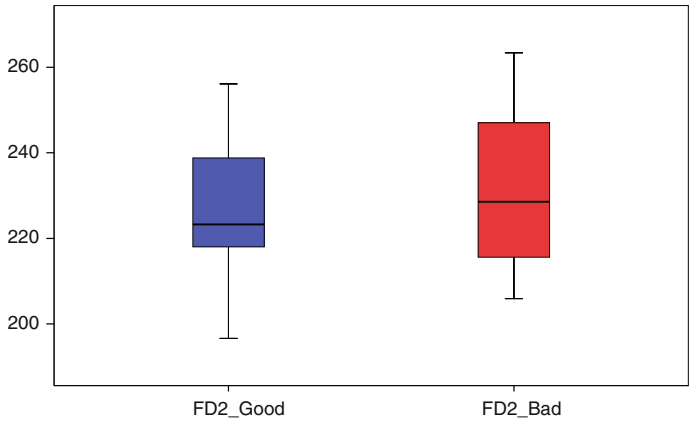

Fig. 7 Average fixation duration (ms) for all participants from S6 to S50

The presence of several good sentences among the bad sentences 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 sentences to see what effect it has on our Fixation Duration data. Figure 7 demonstrates the effect.

As we can see, the elimination of the first five sentences has some effect on differentiating the good and bad sentences, though the difference overall is still limited. When fixation duration is viewed per character, the trend is for bad sentences 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, p. 328) suggest that while fixation duration is a useful measure of cognitive processing, it may reflect "difference aspects of cognitive load".

3.4 Pupil dilations

A further measure used to establish a relationship between textual difficulties and cognitive effort is average pupil dilation. On examining the initial results for all sentences across all participants, hardly any difference in average dilation between bad and good sentences is observed (median = 3.83 mm and 3.82 mm respectively)—see Fig. 8.

Given the difficulty in establishing a clear trend in pupil dilation across all participants, we examine pupil dilation on an intra-subject level motivated by the fact that pupil dilation can vary considerably from person to person (Table 1).

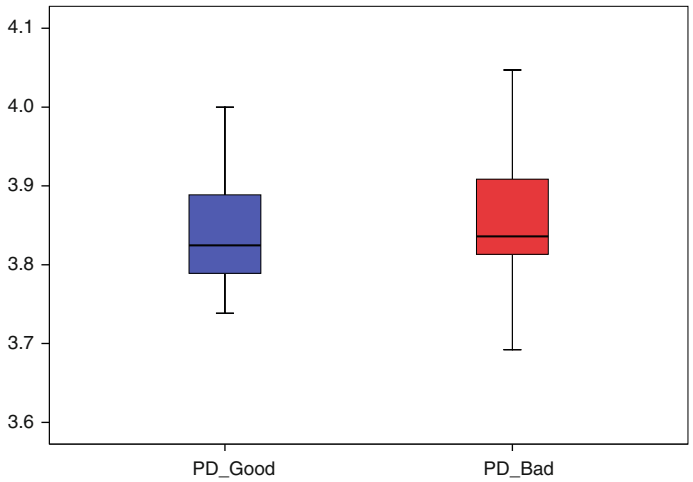

Fig. 8 Average pupil dilation for good and bad sentences (mm)

Table 1 Average pupil dilation (in mm) for each participant for good and bad sentences

Participant	Good sentences	Bad sentences
P1	3.61	3.61
P2	3.91	3.90
P3	3.70	3.66
P5	3.32	3.37
P6	2.93	2.95
P7	4.02	4.02
P8	3.58	3.61
P9	4.80	4.82
P10	3.75	3.70
P11	4.87	4.86

Table 1 illustrates that four of the participants had very slightly higher dilation values for bad sentences than good while six of them either had the same average dilation or had a higher dilation value for good sentences when compared with bad sentences.

Our first conclusion could be that the pupil dilation measurement does not adequately reflect the higher level of cognitive processing we anticipate for bad sentences. However, there are other plausible explanations. Perhaps the sentences were not differentiated enough on a "good/bad" axis for significant changes in pupil dilation to register for each sentence type? The results could also be explained by a latency effect in pupil dilation carrying over from bad to good sentences for example. Or, indeed, it could be that the data set is too small to display significant differences between the two sentence types. However, given that others have repeatedly demonstrated an effect on pupil dilation by increased cognitive load (Rayner 1998), we suggest that further study of pupil dilation as a machine translation evaluation metric is required before coming to any concrete conclusions.

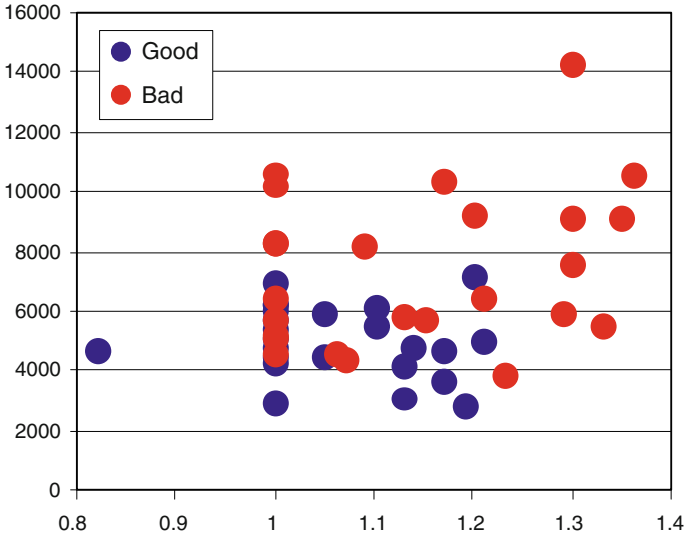

Fig. 9 Average gaze time (ms) and HTER score for good and bad sentences

3.5 Correlations with HTER

Firstly, we look at gaze time per sentence for good and bad sentences across all participants and find a trend where bad sentences had, on average, a higher HTER score and resulted in longer gaze times, whereas good sentences had lower scores and shorter gaze times; $p = 0.260$, $p < 0.05$ (Fig. 9).

The trend is echoed for average fixation count ($p = 0.256$, $p < 0; 0.05$) and supports our earlier findings of a correlation between gaze time, fixation count and reading difficulties.

On examining the correlations of HTER with average fixation duration and average pupil dilation we confirm our earlier findings that fixation duration and pupil dilation do not demonstrate a clear difference between good and bad sentences.

4 Conclusions

One of our initial questions for this study was: can eye tracking be used in MT evaluation and would the eye tracking data reflect the quality of MT output as rated by human evaluators? We have shown that gaze time and fixation count are two eye tracking measures which have medium strength correlation with the previous evaluators' judgments for the sentences used here. The differences in fixation duration results for both sentence types were smaller, although this increases if we assume an acclimatisation effect and remove the initial sentences in the reading task. When combined across subjects, the pupil dilation data do not show significant differences between good and bad sentences, although this is not altogether surprising given other reports of confounding results using pupil dilation, as mentioned above. When viewed as a

measure within subjects, average pupil dilation increases very slightly when reading bad sentences for some subjects, stays the same for others, and actually decreases when reading bad sentences for yet others. We conclude that further testing of this particular metric is required. The test for correlations with HTER scores suggest that gaze time and fixation count appear to have convincing correlations, in general, but pupil dilation and fixation duration do not.

Our second question in this study was: could eye tracking potentially be used as a tool for semi-automatically measuring MT quality? Although the sample is small, we are reassured that the use of eye tracking for semi-automatically evaluating the readability and comprehensibility of MT data is worthy of further investigation.

Using eye tracking requires human readers of text which, if they are employed in formal evaluation studies, is expensive. However, eye tracking could remove much of the subjectivity involved in human evaluation of machine translation quality as the processes it measures are largely unconscious. Eye tracking also opens up the possibility of involving end users in the evaluation of MT output, a development that would be welcomed by many. By recording the reading activity of real end users and how they *interact* with the MT output, MT developers could potentially accumulate data automatically on what the actual end user has difficulty with. This would expand the activity of MT evaluation into the field of user reception of MT output.[4]

Although the sample here is small when the number of sentences and participants is taken into account, our initial study reassures us that eye tracking methods for evaluating the readability and comprehensibility of MT data is worthy of further investigation. It is our intention in the future to build on this research by increasing sample sizes, target languages, MT engine types and domains. Additionally, the effect of controlled language rules and the relationship between different classes of errors and eye tracking metrics such as changes in pupil dilation will also be included in future work. As mentioned in the Introduction, the focus here was on testing and validating methodology. While we do not propose this as a replacement for traditional or automated MT 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 users.

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

Ball Linden J, Eger N, Stevens R, Dodd J (2006) Applying the post-experience eye-tracked protocol (PEEP) method in usability testing. Interfaces 67:15–19

Caffrey C (2008) Using pupillometric, fixation-based and subjective measures to measure the processing effort experienced when viewing subtitled TV anime with pop-up gloss. In: Göpferich S, Jakobsen A,

[4] Obviously users do not typically work with eye tracking monitors, which are still expensive. However, it is not unreasonable to predict that such functionality could eventually be "built-in" to computer monitors and even into hand-held devices.

Mees I (eds) Looking at eyes—eye tracking studies of reading and translation processing. Copenhagen Studies in Language 36. Samfundslitteratur, Copenhagen, pp 125–144
Daneman M, Carpenter P (1980) Individual differences in working memory and reading. J Verbal Learn Verbal Behav 19(4):450–466
Duchowski A (2003) Eye-tracking methodology—theory and practice. Springer-Verlag, London
Göpferich S, Jakobsen A, Mees I (eds) (2009) Looking at eyes—eye tracking studies of reading and translation processing. Copenhagen Studies in Language 36. Samfundslitteratur, Copenhagen
Hess E, Polt J (1964) Pupil size in relation to mental activity in simple problem solving. Science 143:1190–1192
Hyönä J, Nurminen AM (2006) Do adult readers know how they read? Evidence from eye movement patterns and verbal reports. Br J Psychol 97:31–50
Iqbal S, Adamzyck P, Zheng X, Bailey P (2005) Towards an index of opportunity: understanding changes in mental workload during task execution. In: Human factors in computing systems: proceedings of CHI'05. ACM Press, New York, pp 311–320
Jakobsen AL, Jensen K (2009) Eye movement behaviour across four different types of reading task. In: Göpferich S, Jakobsen A, Mees I (eds) Looking at eyes—eye tracking studies of reading and translation processing. Copenhagen Studies in Language 36. Samfundslitteratur, Copenhagen, pp 103–124
Jakobsen AL (1999) Logging target text production with Translog. In: Hansen G (ed) Probing the process in translation: methods and results. Copenhagen Studies in Language 24. Samfundslitteratur, Copenhagen, pp 9–20
Kaakinen JK, Hyönä J (2005) Perspective effects on expository text comprehension: evidence from think-aloud protocols, eyetracking, and recalls. Discourse Process 40:239–257
Kaakinen JK, Hyönä J, Keenan J (2003) How prior knowledge, WMC, and relevance of information affect eye fixations in expository text. J Exp Psychol 29(3):447–457
Kintsch W (1998) Comprehension: a paradigm for cognition, Cambridge, England: Cambridge University Press
Nakayama M, Koji T, Yasutaka S (2002) The act of task difficulty and eye-movement frequency for the oculo-motor indices. In: Proceedings of the symposium on eye tracking research and application, New Orleans, Louisiana, pp 37–42
O'Brien S (forthcoming) Controlled language and readability. In: Shreve G, Angelone E (eds) Translation and cognition. American Translators Association Scholarly Monograph Series. John Benjamins, Amsterdam
O'Brien S (2008) Processing fuzzy matches in translation memory tools—an eye-tracking analysis. In: Göpferich S, Jakobsen A, Mees I (eds) Looking at eyes—eye tracking studies of reading and translation processing. Copenhagen Studies in Language 36. Samfundslitteratur, Copenhagen, pp 79–102
O'Brien S (2006) Eye-tracking and translation memory matches. Perspectives 14(3):185–205
Rayner K (1998) Eye movements in reading and information processing: 20 years of research. Psychol Bull 124:372–422
Roturier J (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. PhD Dissertation, Dublin City University
Roturier J (2009) Deploying novel MT technology to raise the bar for quality: a review of key advantages and challenges. MT Summit XII: proceedings of the twelfth Machine Translation Summit, August 26–30, Ottawa, Ontario, Canada
Schultheis H, Jameson A (2004) Assessing cognitive load in adaptive hypermedia systems: Physiological and behavioural methods. In: Neijdlw W, de Bra P (eds) Adaptive hypermedia and adaptive web-based systems. Springer Verlag, Eindhoven, pp 18–24
Stanford Poynter Project: http://www.poynterextra.org/et/i.htm [Last accessed: 29/04/2009]
Van Gog T, Kester L, Nievelstein F, Giesbers B, Paas F (2009) Uncovering cognitive processes: different techniques that can contribute to cognitive load research and instruction. Comput Hum Behav 25: 325–331