# Eye-tracking for MT evaluation

Natalia Glazyrina

September 20, 2023

## Abstract

Effective machine translation evaluation is vital for advancing translation technology. However, inherent subjectivity in human judgment poses challenges. Eye-tracking technology has emerged as a tool to explore cognitive processes in translation assessment. Previous studies have used standalone eye-tracking systems to predict preferred translations based on gaze patterns. This paper introduces a novel approach using an iPhone camera-based eye-tracker, enabling practical evaluation. Participants assess source sentences, select translations, and identify problematic words. Eye movement metrics are examined for correlation with translation choices. Additionally, an investigation is conducted to determine whether problematic words attract increased visual attention.

## 1 Introduction

Human evaluation of machine translation (MT) quality remains a crucial aspect in the advancement of translation technology. However, the subjectivity inherent in human judgment poses a challenge in achieving reliable and consistent assessment. In recent years, eye-tracking technology has emerged as a promising tool to delve into the cognitive processes underlying translation evaluation. By capturing individuals' reading patterns during the evaluation of translation options, eye-tracking provides insights into the linguistic cues that influence decision-making, thus offering an objective lens to complement traditional subjective evaluation methods.

Previous studies have showcased the potential of eye-tracking in predicting the preferred translation among multiple options. [1][3] These investigations have employed standalone eye-tracking systems to monitor participants' gaze movements, providing valuable insights into fixation points, gaze jumps, and reading duration. These studies have successfully correlated specific eye movement patterns with the selection of the most suitable translation, underscoring the relevance of visual attention in translation quality assessment.

In this paper, I contribute to the evolving landscape of MT evaluation using an iPhone camera-based eye-tracking approach. Unlike conventional standalone systems, this methodology offers a pragmatic alternative, avoiding the need for dedicated eye-tracking hardware and enabling broader accessibility. This approach embraces real-world scenarios, where users can employ their own devices for evaluation.

I implement the experimental design, where participants are presented with a source sentence in English and two target candidate translations in Russian. Participants are tasked with selecting the most suitable translation while also identifying problematic words within the suboptimal choice.

To evaluate the efficacy of the approach, I analyze the correlations between eye movement metrics (such as fixation count, jump frequency, and reading time) and participants' translation choices. I hypothesize that the utilization of an iPhone camera-based eye-tracker can be used to assess that correlation and to substantiate that problematic words within suboptimal translations are associated with a higher concentration of gaze fixations and gaze jumps.

## 2 Related Works

Stephen Doherty et al. [1] aimed to explore whether eye-tracking data can reflect the quality of MT output as rated by human evaluators and whether eye-tracking could be used as a semi-automated tool for evaluating MT quality. The "eye-mind hypothesis" is the underlying assumption of eye-tracking, suggesting a link between where the eyes focus and cognitive processing of the content. The study analyzed various eye-tracking metrics, including gaze time, fixation count, fixation duration, and pupil dilation. The results indicated correlations between eye-tracking metrics and the quality of MT output

as rated by evaluators. Specifically, "bad" sentences had longer gaze times and more fixations compared to "good" sentences. Fixation duration and pupil dilation showed less consistent correlations.

Ondřej Bojar et al. [2] aimed to understand the reasons behind inter-annotator disagreement, where different human evaluators may provide inconsistent rankings for MT systems. It was focused on the human annotation procedure used in evaluating submissions for the Shared Translation Task in the WMT workshop. Eight volunteers evaluated English-to-Czech MT systems participating in the WMT13 shared task. Eye movements were recorded using the EyeLink II eye-tracker, and participants viewed the screens from a fixed distance. Several observations were made from the eye-tracking data:

- Candidates with different types of errors led to disagreements. Candidates with both high fluency and low adequacy caused uncertainty.
- Participants spent less time looking at candidates positioned lower on the screen. However, this didn't significantly affect the agreement.
- Different strategies were observed in how annotators approached rankings, such as comparing source-reference pairs or sequentially evaluating candidates.
- Annotators spent more time looking at the source text than the reference, indicating the source's higher difficulty.

Hassan Sajjad et al.[3] discussed the use of eye-tracking data as a tool for evaluating the quality of MT systems. Traditional methods of MT evaluation, based on human judgments, have suffered from subjectivity and low inter-annotator agreements. The authors propose using eye-tracking data to analyze reading behavior and patterns of human evaluators to predict translation quality. It also employs the concept of the "eye-mind hypothesis". The article presents the results of using different sets of features to predict translation quality. It shows that reading patterns, such as the number of regressions and the time spent reading, can help distinguish between good and bad translations. The combination of eye-tracking features with BLEU (a commonly used MT evaluation metric) yields promising results in predicting translation quality, indicating that reading patterns capture more than just fluency.

# 3 Experimental Setup

## 3.1 Selection of Experimental Tool

The initial stage of configuring the experiment included the search and comparison of different webcam-based eye-tracking systems. After the exploration, two tools were selected for comparative evaluation: jsPsych [4] utilizing the WebGazer.js eye-tracker and Eyeware Beam [5].

The jsPsych library had limited flexibility, making it difficult to start its usage, and demonstrated a low accuracy during the calibration procedure. Furthermore, the coordinates of a gaze collected during the trial run posed challenges in interpretation.

Conversely, the Eyeware Beam tool has the option of using either an iPhone camera or a webcam for eye-tracking. It is convenient to use, and has elevated accuracy and interpretability of the collected data – it was possible to directly map them to the screenshots.

Hence, the Eyeware Beam tool, particularly its version utilizing the iPhone camera, was chosen due to its heightened precision in contrast to the webcam variant. Consideration was given to the idea of concurrently deploying both tracking mechanisms - iPhone, and webcam. However, this course of action presented difficulties in aligning participants' positioning to fulfill the requirements of both tools.

## 3.2 Experimental Environment

The experiment was conducted in the following way. There were 8 participants (4 male, 4 female), all native Russian speakers possessing a minimum English proficiency level of B2. In total 11 distinct screen sets were used, each including 10 screens with the layout depicted in Figure 1. Each screen presented an English source sentence at the top, accompanied by two candidate Russian translations underneath,

labeled as "1" and "2" respectively. Some sets included contextualized sentences, while others consisted of standalone sentences without any connection between them.

The introductory set was excluded from the analytical scope, serving solely to acquaint participants with the task and to give them the opportunity to ask for clarifications. Participants, using a laptop with a 14-inch display and a camera of iPhone 12 mini, were instructed to carefully read all sentences on each screen trying not to move their heads and subsequently make a choice of the best candidate translation. Additionally, participants were asked to click on words and phrases in the worst sentence that they identified as problematic for various reasons. This design element was inspired by Maja Popović research [6], which entailed an evaluation of MT through the identification of specific sentence segments presenting challenges, and distinguishing between different types of errors. However, for the sake of simplicity, there was no such distinction in the current experiment.

Throughout sentence comprehension and translation comparison, the eye-tracker recorded the gaze movement at the frequency of 90 Hz. However, gaze movement recordings were suspended during the selection of problematic words. In contrast with previous studies, the participants' head position was not fixed, and the exact screen-to-eye distance was not established. Nevertheless, adherence to the eye-tracker's requirements was ensured ( 50-60 cm screen-to-eye distance).

The set of test sentences was taken from the WMT Metric Task datasets for the years 2022 and 2021.

# 4 Analysis

## 4.1 Inter-annotator Agreement

To compute the inter-annotator agreement (IAA), the Fleiss Kappa for multiple annotators was used:

$$\kappa = \frac{p_o - p_e}{1 - p_e}, \tag{1}$$

where $p_o$ is the observed agreement of the raters and $p_e$ is the expected agreement of the raters.

The overall IAA $\kappa$ equals 0.36, which means "Fair agreement".

I analyzed the pairs of candidate translations characterized by notably low agreement and identified potential contributing factors as follows:

- In the contextualized sets, the annotator failed to perceive the context
- Certain sentences require specific knowledge for accurate translation evaluation
- One candidate translation exhibits greater fluency, while the other is closer to a word-to-word correspondence
- The two translations decide to resolve the common English ambiguity in gender differently, and the raters then don't have a clear opinion on which gender should be used
- Both sentences lack adequacy but in different places
- Within non-contextualized sets, the absence of context makes it challenging to decide which candidate is better
- One candidate contains a borrowed English word, while the other one employs a translated term

## 4.2 Feature Extraction

For the extraction of features for further analysis, a post-processing procedure was employed on the collected data. Notably, the collected traces exhibited a noticeable shift along the y-axis, possibly attributable to inaccuracies in the calibration process or head movement during the experiment. This phenomenon is illustrated in Figure 2. Consequently, a manual adjustment was required, using a constant addition to the y-coordinate across the entire trace for each screen. The modified, post-processed trace is shown in Figure 3.

It is worth noting that during the experiment a few times participants misclicked on the "Next" button and accidentally skipped a screen without noticing

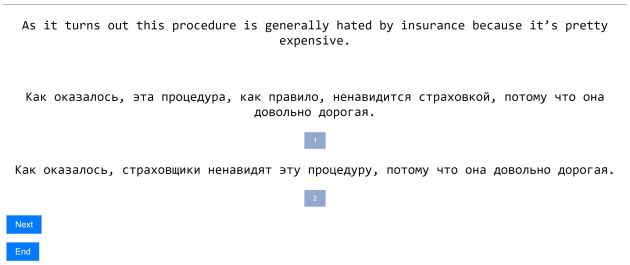


Figure 1: Example of the screen layout

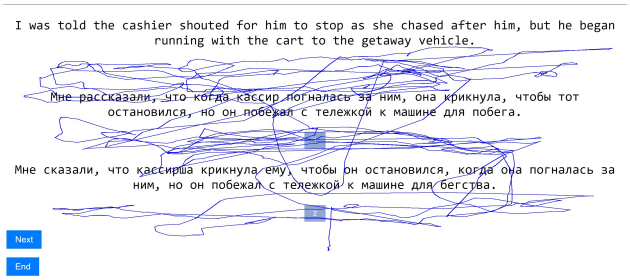


Figure 2: Mapping of originally collected trace of gaze

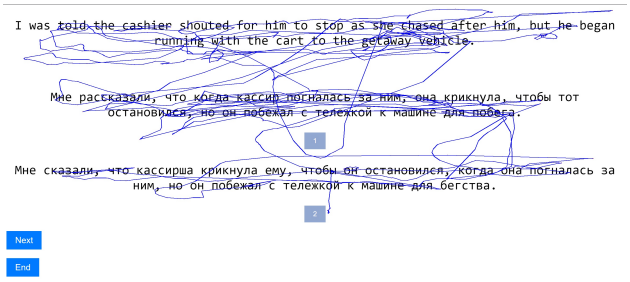


Figure 3: Mapping of shifted along y-axis trace of gaze

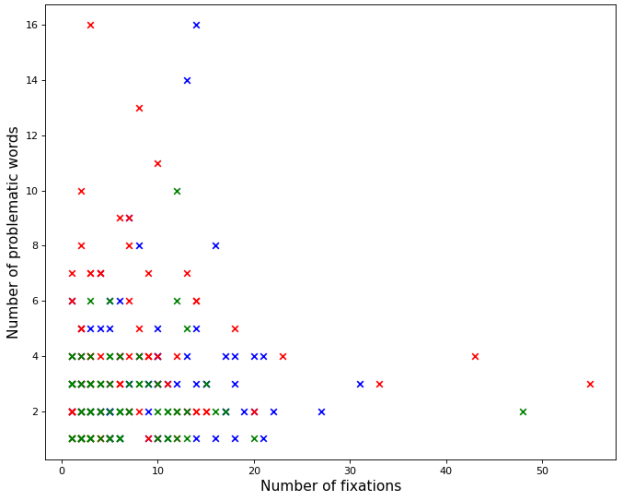


Figure 4: Number of problematic words vs. number of fixations for participants 1,2 and 3. Pearson correlation coefficient 0.037

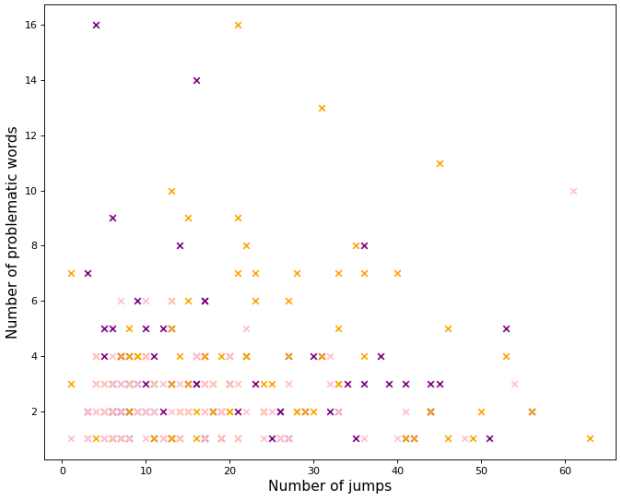


Figure 5: Number of problematic words vs. number of jumps for participants 1,2 and 3. Pearson correlation coefficient -0.024

it. These occurrences were infrequent and pointed to drawbacks of the technical implementation of the experiment.

Subsequently, the following features were derived: the count of gaze fixations per sentence, the count of gaze jumps per sentence, and the time allocated to each sentence. A gaze fixation was defined by two criteria: first, the dimensions of the rectangle encompassing the gaze's stationary region (224x70 pixels, with 224 representing the average word length), and second, the temporal duration for which the gaze remained within the rectangle without deviation – set at 500 milliseconds. Meanwhile, a gaze jump was established as a marked alteration of y-coordinates by more than 200 pixels.

Utilizing these features, a Logistic Regression model was constructed to explore the relationship between the values of these features and the choice of the best translation. However, the resulting p-value exceeded the threshold of statistical significance for prediction. Consequently, the hypothesis of the ability of the listed features to correlate with the choice of translation through the employment of the iPhone camera-based eye-tracker was rejected.

Hence, it emerges that the connection between gaze fixations, gaze jumps, and clicked problematic words is challenging to establish. The Figures 4,5 shows the scatter plot of number of problematic words vs. number of fixations and the scatter plot of number of problematic words vs. number of jumps, where each point corresponds to each screen for the participants 1, 2 and 3. The pattern of these plots is the same for all participants. It illustrates almost uniform density of data points, which means that there is no correlation between variables. The Pearson correlation coefficients are 0.037 for fixations vs. number of problematic words and -0.024 for jumps vs. number of problematic words.

# 5 Conclusion

In this study, I executed an experiment employing an iPhone camera-based eye-tracker to assess its viability for the task of MT evaluation. However, the obtained results underscore the unreliability of this approach within the tested settings. The findings reveal several plausible factors contributing to these outcomes:

- The utilization of a relatively compact screen (14 inches) may have constrained the discernment of fixations among distinct words.
- Suboptimal camera quality might have compromised tracking precision, suggesting potential improvement through a higher-grade camera, such as a newer iPhone model.
- The limitation of a maximum 90 Hz eye-tracking frequency could have impacted the robustness of the analysis.
- Manual data adjustments introduced for alignment purposes may not have achieved the required precision.

Future research could address the technical shortcomings by implementing refinements and upgrading hardware components. Potential improvements include considering alternative eye-tracking solutions, thereby elevating the depth and comprehensiveness of the current research. The project's source code and associated materials can be found on the GitHub repository: https://github.com/lack-of-purpose/et-mt-evaluation-project.

# References


[1] Michael Carl Stephen Doherty Sharon O'Brien. "Eye Tracking as an Automatic MT Evaluation Technique". In: (2010).

[2] Maria Zelenina Ondrej Bojar Filip Dechterenko. "A Pilot Eye-Tracking Study of WMT-Style Ranking Evaluation". In: *Proceedings of the LREC 2016 Workshop "Translation Evaluation - From Fragmented Tools and Data Sets to an Integrated Ecosystem"* (2016), pp. 20–26.

[3] et al. Hassan Sajjad Francisco Guzman. "Eyes Don't Lie: Predicting Machine Translation Quality Using Eye Movement". In: *Proceedings of NAACL-HLT* (2016), pp. 1082–1088.

[4] *jsPsych framework*. URL: `https://www.jspsych.org/7.3/`.

[5] *Eyeware Beam eye-tracker*. URL: `https://beam.eyeware.tech/`.

[6] Maja Popovic. "Informative Manual Evaluation of Machine Translation Output". In: *Proceedings of the 28th International Conference on Computational Linguistics* (2020), pp. 5059–5069.