Eye Tracking for Machine Translation Quality Evaluation

Natalia Glazyrina, Ondřej Bojar
Charles University, Faculty of Mathematics and Physics, Institute of Formal and Applied Linguistics
Prague, Czech Republic
nglazyrinav@gmail.com, bojar@ufal.mff.cuni.cz

Abstract

Eye tracking offers unique insights into cognitive processes, making it a promising tool for evaluating machine translation (MT). This study explores the feasibility of using an iPhone 12 camera-based eye tracker with a 14-inch laptop display for conducting translation evaluation in personal workspaces, offering a more accessible and cost-effective alternative to traditional setups. Participants evaluated source sentences, selected translations, and identified problematic words while their gaze metrics were recorded and analyzed. Our findings reveal statistically significant correlations between gaze patterns and preferred translations, as well as increased visual attention to problematic words. These results demonstrate that home-based eye tracking systems are technically sufficient for capturing gaze behavior accurately enough for MT evaluation purposes. A potential practical application is to speed up translation proof-reading using eye tracking technique to automatically mark portions of text that should be attended to and improved based on the gaze pattern during a quick reading.

Keywords: eye tracking, machine translation, correlation, fixation, saccade

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 a reliable and consistent assessment. In recent years, eye tracking technology has emerged as a promising tool to delve into the cognitive processes that underlie translation evaluation (Stymne et al., 2012; Sajjad et al., 2016). By capturing individuals' reading patterns during the evaluation of translation options, eye tracking provides insights into the linguistic cues that influence decision-making, such as incorrect word order, morphological disagreement, and semantic ambiguity, thus offering a more objective lens to complement traditional subjective and self-reported evaluation methods.

Previous studies have demonstrated the potential of eye tracking in predicting the preferred translation among multiple options (Sajjad et al., 2016; Doherty et al., 2010) or machine translation error analysis (Stymne et al., 2012). These investigations have employed standalone eye tracking systems to monitor participants' gaze movements, revealing that poorly translated text causes readers to frequently jump back while reading, which serves as a measurable marker of processing difficulty. Furthermore, research demonstrates that "bad" sentences result in significantly higher gaze times and fixation counts compared to high-quality ones.

In this paper, we contribute to the evolving landscape of MT evaluation using an iPhone camera-based eye tracking approach. This approach was chosen over webcam-based due to gaze tracing qualities revealed during the comparison of eye tracking systems. 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. By lowering technical and financial barriers, this method aims to democratize access to eye tracking technology, allowing researchers and practitioners to integrate cognitive insights into MT evaluation without relying on specialized equipment. Such an accessible solution has the potential to expand the reach of eye tracking research to diverse, including non-specialist, environments, making the evaluation process more inclusive and practical.

Despite the broad use and advances in automatic evaluation of machine translation, see e.g. Lavie et al. (2025), human evaluation remains the gold standard in the field of machine translation. Automatic metrics, while scalable, often struggle to capture semantic nuances, stylistic consistency, and the actual cognitive load experienced by a reader. This enduring importance is evidenced by the annual Conference on Machine Translation (WMT; Kocmi et al., 2025), where human judgment serves as the benchmark for validating the accuracy of automated systems.

However, traditional human evaluation often treats the translator's or rater's decision as a "black box," focusing on the final output rather than the process. By integrating eye tracking, we can move beyond simple preference scores to observe the cognitive effort involved in processing translation errors.

We implement an experimental design in which participants are presented with a source sentence

Proceedings of The Second International Workshop onEye-Tracking Resources and Evaluation for Human-Aligned NLP, pages 1–9
May 12, 2026.

in English and two target candidate translations in Russian. Participants are tasked with selecting a better option, while also identifying problematic words within the suboptimal choice.

To evaluate the efficacy of the approach, we analyze the correlations between eye movement metrics (such as fixation/saccade count and time spent on each sentence) and participants' translation choices. We 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 saccades.

2. Related Works

Although human judgment remains the gold standard, as seen in the annual Conference on Machine Translation (WMT),[1] human evaluation is not without flaws; it is resource-intensive and prone to high inter-annotator variability and subjectivity (Graham et al., 2013; Lommel et al., 2014).

To bridge the gap between automated scores and subjective human ratings, researchers have turned to eye tracking. The foundational assumption of eye tracking is the "eye-mind hypothesis" (Just and Carpenter, 1980), which suggests a link between gaze fixation and cognitive processing of linguistic content. Based on this theory, Doherty et al. (2010) 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. This 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. The duration of fixation and pupil dilation showed less consistent correlations.

Building on this, Sajjad et al. (2016) utilized eye tracking data to address the subjectivity and low inter-annotator agreement often found in traditional human judgments. The authors demonstrated that specific reading patterns, such as the number of regressions and the total reading time, effectively distinguish between high- and low-quality translations. They found that combining eye tracking features with BLEU scores (Papineni et al., 2002) yielded promising results in predicting translation quality, indicating that reading patterns capture more than just fluency. This suggests that gaze data capture cognitive nuances, such as semantic processing effort, that surface-level n-gram overlap metrics like BLEU inherently overlook.

Furthermore, Bojar et al. (2016) investigated the cognitive drivers of inter-annotator disagreement within the WMT Shared Translation Task. Using a high-precision EyeLink II tracker in a controlled laboratory setting, the authors found that inconsistent rankings often stemmed from specific error types – mainly in translations that displayed high fluency but low adequacy. Their gaze data revealed that these "deceptive" translations caused significant uncertainty and longer processing times. The study also highlighted the cognitive burden of the source text, noting that annotators focused more on source sentences than references, which was expected because the participants were native speakers of the target language but only second-language learners for the source.

Despite the established benefits of eye tracking metrics, their integration into large-scale MT evaluation has been hindered by a reliance on expensive, lab-bound hardware. Lately, several studies have compared webcam-based eye tracking systems with traditional in-lab systems, evaluating their viability across different research domains. In psycholinguistics, webcam-based systems have been used to study language processing in naturalistic environments, providing accessibility to diverse populations and geographically dispersed participants. For instance, Özsoy et al. (Özsoy et al., 2023) investigated heritage language processing using webcam-based eye tracking, demonstrating that data collected in such settings was largely consistent with in-lab systems. This approach facilitated the inclusion of heritage speakers who otherwise might not have access to laboratory facilities. Other studies have replicated psycholinguistic effects, such as the verb semantic constraint and lexical interference effects, using webcam-based tracking, confirming its ability to capture both robust and subtle phenomena (Prystauka et al., 2023). Similarly, a recent replication of a Visual World study on verb aspect processing showed that webcam-based eye tracking, even with off-the-shelf tools, can achieve comparable results to infrared systems, offering a cost-effective and accessible alternative (Vos et al., 2022). These findings suggest that while webcam-based systems can reliably replicate key effects, careful attention must be given to factors like calibration, lighting, and participant guidance to ensure data quality.

Our work contributes to this shift by exploring the efficacy of iPhone-based eye tracking specifically for MT quality evaluation. By moving the experimental environment from the controlled laboratory to a home-based setting, we aim to lower the financial and technical barriers to high-quality, "process-oriented" human evaluation. This approach not only

[1] https://www.statmt.org/wmt25/

democratizes access to cognitive data but also introduces a new layer of quality control, allowing researchers to filter human annotation based on real-time cognitive engagement and attentional focus.

3. Methodology

3.1. Tool Selection

The experimental configuration was finalized after a comparative pilot of webcam-based systems. We initially evaluated jsPsych[2] with the WebGazer.js library. While jsPsych is a well-established tool for behavioral experiments, it presented several limitations. To validate the accuracy of each system, we conducted a controlled reading task where the researcher read the stimulus text slowly and linearly, line-by-line. As shown in Figure 1, the resulting gaze trace for the webcam-based system was highly distorted and failed to follow the horizontal progression of the lines. Furthermore, the gaze coordinates collected during trial runs were challenging to interpret, complicating the analysis.

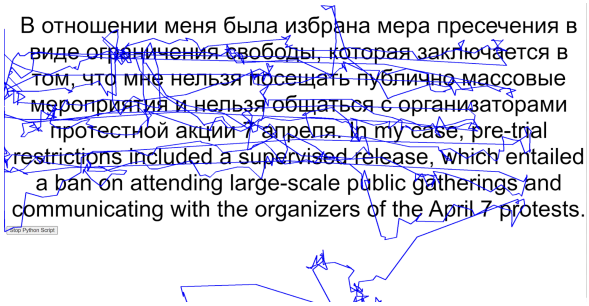

Figure 1: jsPsych eye trace mapped on the screen.

Consequently, we selected Eyeware Beam for data collection. This software supports high-precision tracking via an iPhone 12 mini camera (connected via USB) at a sampling rate of 90 Hz. The iPhone-based approach was chosen over standard webcams due to its far more accurate mapping, demonstrated in Figure 2. The camera was positioned horizontally at the base of a 14-inch laptop screen on the left side. Unlike previous studies, the participants' head positions were not fixed, and the exact screen-to-eye distance was not rigorously controlled. However, participants adhered to the eye tracker's recommended distance of approximately 50-60 cm.

3.2. Participants

Participants were recruited via convenience sampling from a pool of graduate-level volunteers within

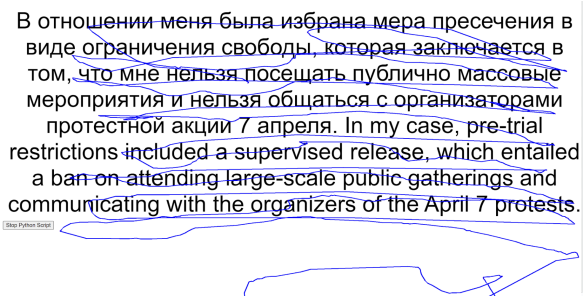

Figure 2: Eyeware Beam eye trace mapped on the screen.

a university environment. Participation was entirely voluntary, and no financial compensation was provided. All participants were informed of the study's objectives and the nature of the eye tracking data being collected prior to the start of the trial. For this pilot study we recruited 8 participants (4 male, 4 female) with the following profiles:

- **Language**: Native Russian speakers with B2+ English proficiency.
- **Age**: 25–30 years
- **Education**: Graduate-level or higher.
- **Vision**: Normal or corrected-to-normal vision (no glasses were worn during this specific trial to ensure maximum tracker stability).

The experimental protocol was designed following the principle of data minimization. The utilized software processes the camera feed locally in real-time to calculate gaze vectors. Crucially, no raw video or photographic data of participants was stored at any point during the study. The exported data consisted exclusively of numerical logs containing temporal markers (timestamps) and spatial gaze coordinates relative to the screen. Since no personally identifiable information (PII) was linked to the gaze logs, the dataset is inherently anonymized.

3.3. Research Materials

The test stimuli consisted of sentence pairs extracted from the WMT Metric Task (2021[3] and 2022[4]) datasets.

- **Structure**: 10 distinct screen sets, each containing 10 experimental screens.
- **Layout**: A standardized interface displaying one English source sentence at the top and

[2] https://www.jspsych.org/7.3/

[3] https://drive.google.com/drive/folders/1TNIeXirfNMa6WV7LlS3Z51UxNNCgGcmS

[4] https://drive.google.com/file/d/1I0O-NzOLCxrO6noub2pY81BtWxp42A46/view

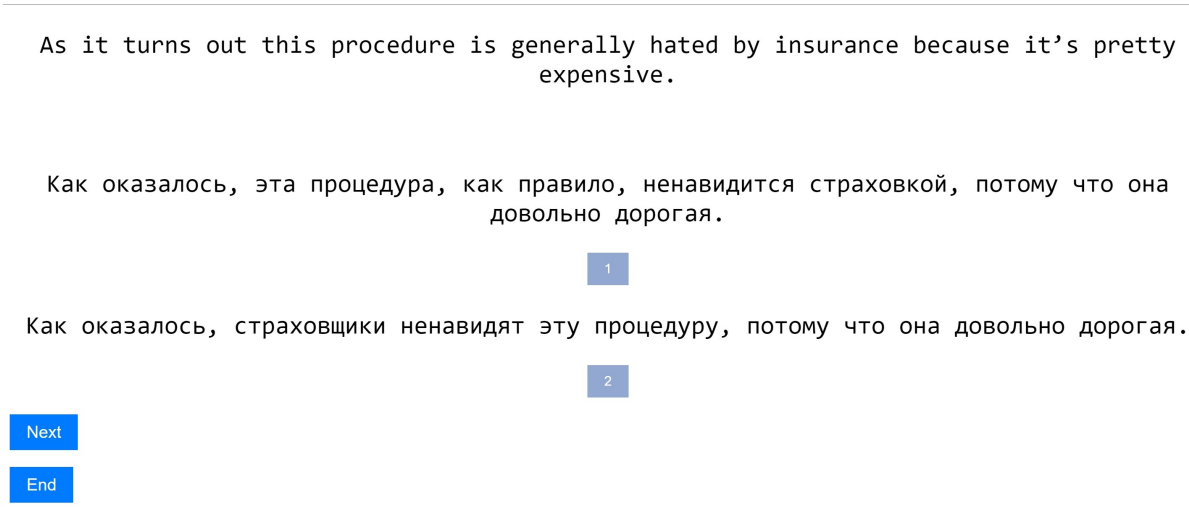

Figure 3: Example of the screen layout.

two candidate Russian translations (labeled "1" and "2") below (see Figure 3).

- **Calibration**: A warm-up set was provided to familiarize users with the interface, and calibration was verified at the start of each session and after breaks.

3.4. Experimental Task and Procedure

Participants were asked to perform a dual-stage evaluation task designed to capture both preference and cognitive load:

1. **Comparative Judgment**: Participants read the source and both translations, then selected the superior candidate by clicking a corresponding button. To prevent positional bias, the order of the translation candidates was randomized. Consequently, the 'better' translation appeared as either the first or second option with equal frequency throughout the experiment.

2. **Error Span Identification**: In the suboptimal translation, participants were instructed to click on specific words or phrases they perceived as problematic. This design follows Maja Popović's research (Popović, 2020) on identifying challenging sentence segments in machine translation, though no distinctions were made between different types of errors in this study.

To ensure data integrity, gaze data recorded during the clicking action (identification phase) was excluded from the cognitive load analysis. This allows us to isolate the uninterrupted reading process from the manual task of error marking.

3.5. Quality Control

A key contribution of this methodology is the use of gaze data as a quality control layer. By analyzing fixation density and saccadic movements, we can identify "inattentive" trials where the participant may have skimmed the text without full cognitive engagement. This enables the exclusion of unreliable human data that are typical for remote, home-based annotation tasks.

4. Analysis

4.1. Gaze Data Post-Processing

For the extraction of fixations and saccades 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 4. 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 5.

It is worth noting that during the experiment a few times participants misclicked on the "Next" button and accidentally skipped a screen without noticing it. These occurrences were infrequent (5 times) and pointed to drawbacks of the technical implementation of the experiment. Those 5 screens are skipped in the analysis.

To analyze gaze behavior, we extracted features related to gaze fixations and saccades using the Velocity-Threshold Identification algorithm (Salvucci et al., 2000), with a velocity threshold set to 100. This algorithm identifies fixations and saccades based on point-to-point velocity. Our analy-

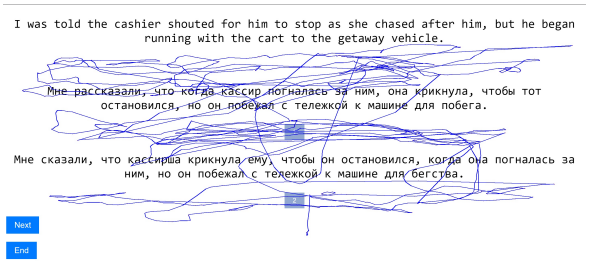

Figure 4: Mapping of originally collected trace of gaze.

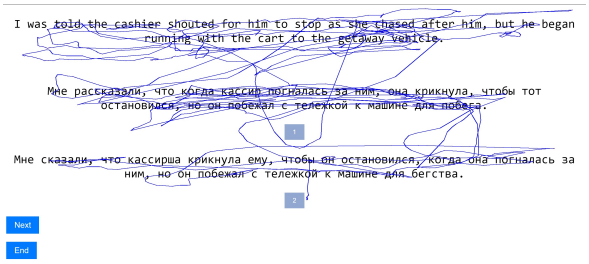

Figure 5: Mapping of shifted along y-axis trace of gaze.

sis includes two levels: word-level and sentence-level.

4.2. Word-level Analysis: The Cost of Errors

For the word-level analysis, we prepared two datasets, each containing three features, where each row represented data for a single screen and participant:

1. Dataset 1: Number of identified problematic words, number of fixations on these words, and total number of fixations on the screen.

2. Dataset 2: Number of identified problematic words, number of saccades on these words, and total number of saccades on the screen.

In both datasets, we conducted correlation analysis by calculating the Pearson correlation coefficient (PCC) between the number of words marked as problematic and the relative share of visual attention (fixations and saccades) those words received.

The scatter plots (Figures 6 and 7) illustrate these relationships. While there is a high density of points at low error counts, a clear upward trend is visible:

- **Fixation Proportion**: PCC=0.30 ($p < 0.001$)
- **Saccade Proportion**: PCC=0.44 ($p < 0.001$)

While the correlation coefficients indicate a low-to-moderate relationship, they are highly statistically significant. The higher correlation for saccades (0.44) suggests that problematic segments do not merely cause the eye to linger; they are more strongly associated with re-scanning behaviors as participants repeatedly glance back at the source sentence to check the original meaning whenever they run into a problematic translation. This finding implies that problematic words attract a disproportionate share of visual attention.

4.3. Sentence-level analysis: Predicting Preference

For the sentence-level analysis, we derived the following features: number of fixations per sentence, number of saccades per sentence, time spent on each sentence. Using these features, we modeled the participants' final translation choice (Sentence 1 vs. Sentence 2) using both a Logistic Regression (LR) model for statistical significance and a Decision Tree (DT) for behavioral interpretability.

4.3.1. Logistic Regression

The LR model was implemented using the statsmodels library (Seabold and Perktold, 2010) with default parameters, except for the maximum iteration parameter, which was set to 100.

The model summary revealed time spent and fixation counts on the second sentence as significant predictors ($p < 0.05$) with the following coefficient signs:

- Time spent on sentence 2: Negative
- Fixations/saccades on sentence 2: Positive

The coefficient signs reveal a "comparative pressure" effect: an increase in time spent on Sentence 1 significantly increases the probability of the user choosing Sentence 2. This suggests that the time metric captures the "struggle" to find meaning; when one candidate is difficult to parse, the user's preference shifts to the alternative. Additionally, a higher number of fixations or saccades on sentence 2 indicates tendency to choose that sentence.

4.3.2. Decision Tree Interpretation

To derive actionable thresholds for these behaviors, we trained a Decision Tree classifier (depth=3) using scikit-learn library (Pedregosa et al., 2011). Unlike the LR model, which provides a probability gradient, the DT identifies the exact points in the decision-making process.

The model's Feature Importance (Table 1) indicates that the decision-making process is primarily driven by metrics associated with the second

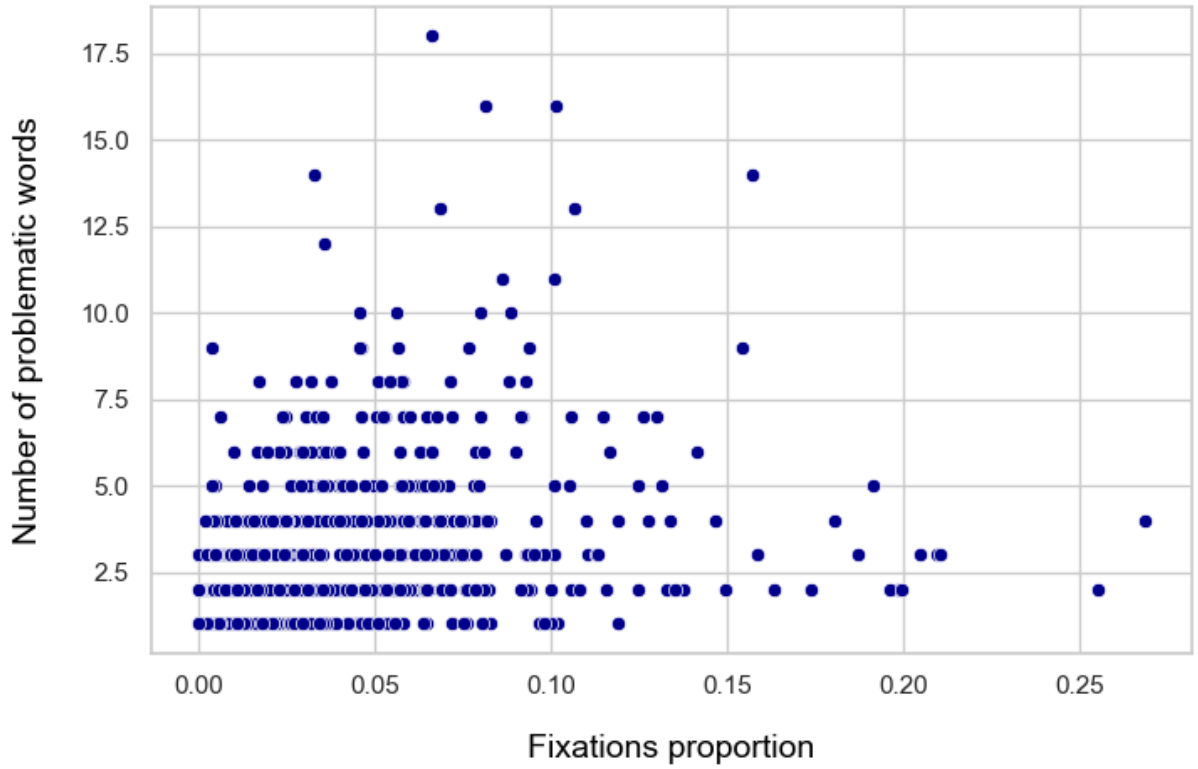

Figure 6: Number of problematic words vs. fixation proportion. Pearson correlation coefficient 0.30.

Feature	Importance
Time spent sentence 2	0.39
Saccades on sentence 2	0.32
Fixations on sentence 2	0.20
Saccades on sentence 1	0.08

Table 1: Feature Importance extracted from Decision Tree.

translation candidate. Specifically, time spent on sentence 2 (39.1%) and saccades on sentence 2 (32.3%) were the most influential factors, while metrics for sentence 1 provided significantly less predictive power.

The tree structure (Figure 8) revealed highly interpretable behavioral "thresholds." For instance, a specific path in the tree identified a high-certainty node (Gini = 0.188) where a low number of saccades (≤ 8.5) combined with a limited time investment (≤ 81 gaze units) on Sentence 2 led to a consistent selection of that candidate. This suggests that "fluency" – characterized by rapid, linear processing – is a stronger predictor of preference than simply the total amount of attention paid to a sentence.

5. Discussion

The results of this pilot study suggest that iPhone-based eye tracking is a viable, low-cost method for capturing cognitive effort in MT evaluation. The correlation found between visual attention – specifically gaze duration and fixation counts – and the final translation choice aligns with the "eye-mind hypothesis" (Just and Carpenter, 1980), suggesting that participants spend significantly more time processing suboptimal segments.

5.1. Asymmetry in Sentence Correlation

Interestingly, our analysis showed a stronger correlation between visual attention to the "second" translation candidate and the final choice than for the first. We hypothesize that this asymmetry does not reflect a lack of cognitive engagement with the first sentence, but rather a limitation in our current manual gaze-trace processing. Because the second sentence is often the final piece of information processed before a decision is made, the "recency effect" may make its associated gaze data more distinct.

5.2. Challenges in Home-Based Calibration

Our home-based approach offers an alternative to the traditional lab-based setup described in (Bojar

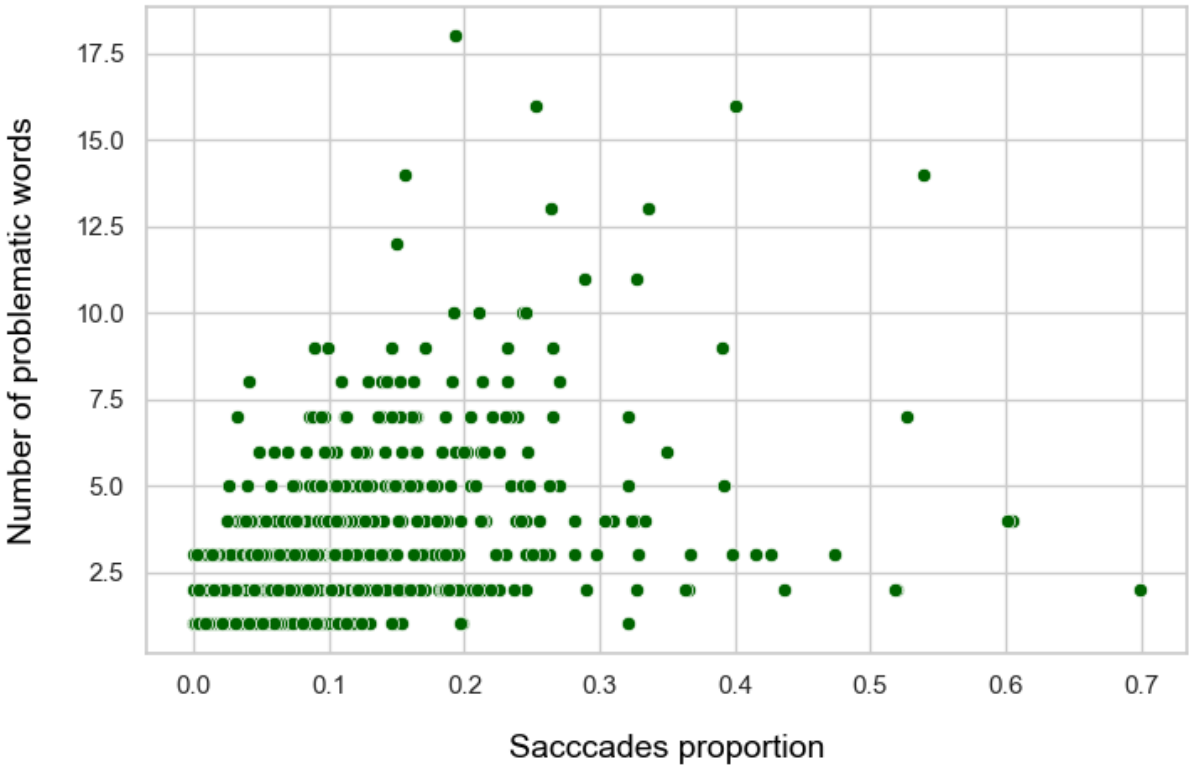

Figure 7: Number of problematic words vs. saccade proportion. Pearson correlation coefficient 0.44.

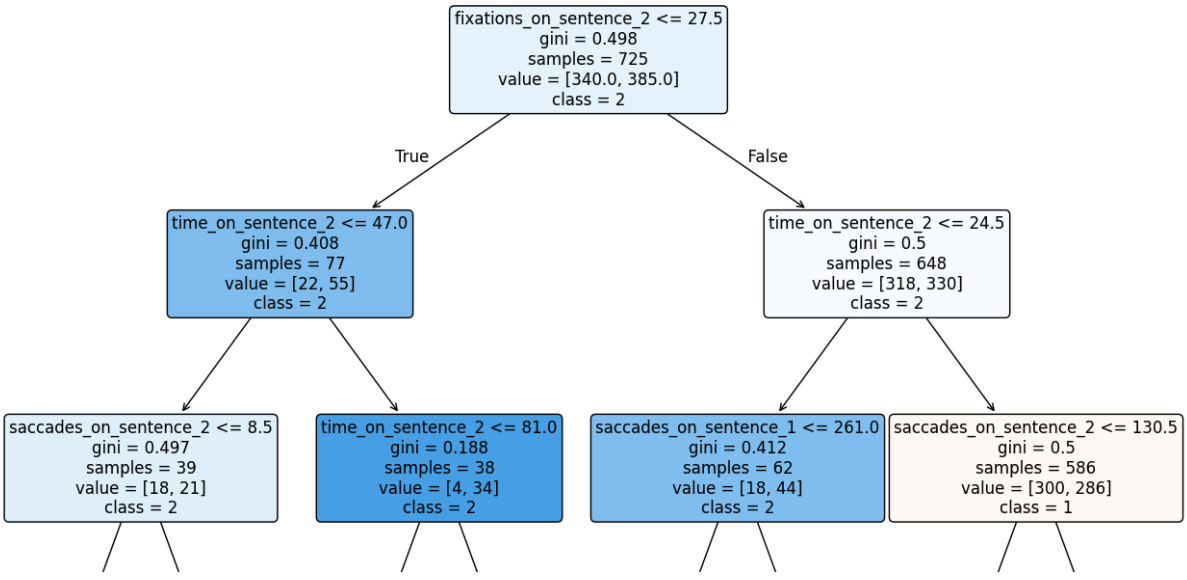

Figure 8: The decision tree analysis (depth=3) showing behavioral thresholds for MT selection.

et al., 2016), which utilized a high-precision EyeLink II tracker (250 Hz) and a chin rest to minimize noise. While the lab setup performed calibration before every screen and excluded data based on pupil size and blinks (removing 30 ms around each blink), our setup relied on an iPhone 12 mini (90 Hz) with calibration at session starts and breaks. Regarding trial exclusion, the lab study manually adjusted areas of interest due to non-linear distortions, while we excluded only 0.6% of screens due to technical misclicks and applied a constant y-axis shift to correct for calibration drift caused by the lack of head restraints. While the iPhone 12 mini provided sufficient precision for broad sentence-level analysis, the lack of head-restraints introduced “noise” during manual processing. This highlights a critical trade-

off: home-based environments offer higher accessibility but require more robust, automated post-processing scripts to handle natural head movements and slight calibration drifts.

6. Conclusion

While this study serves as a preliminary proof-of-concept, our findings suggest that eye tracking provides a 'process-oriented' layer that complements the 'black box' of traditional direct assessment. Unlike error span annotation (Kocmi et al., 2024), which only identifies the location of a flaw, gaze metrics - specifically saccade proportions - reveal the re-scanning behavior, when annotators double-check the source sentence when they encounter a translation that is hard to follow. This allows us to observe cognitive nuances, such as semantic processing effort, that traditional automatic metrics like BLEU or COMET inherently overlook. Furthermore, while the current requirement for manual alignment remains a technical bottleneck, the effort is justified by the potential to use gaze data as a quality control layer; this enables researchers to filter out 'inattentive' trials where participants may have skimmed the text without full cognitive engagement—a critical need for remote, home-based annotation.

6.1. Limitations and Future Work

While this study serves as a proof-of-concept for the technical viability of mobile-based tracking, we acknowledge that our participant pool was limited to a convenience sample of eight volunteers. This initial trial focused on demonstrating the workflow and technical feasibility rather than providing a large-scale demographic analysis.

Future research will focus on:

- **Scaling**: Expanding to a larger, more diverse group to validate the applicability of these metrics.
- **Linguistic Diversity and Cross-Family Pairs**: Our current study focused exclusively on an English–Russian language pair. Future iterations should expand to non-Indo-European languages, such as logographic systems (e.g., Chinese) or right-to-left scripts (e.g., Arabic). Investigating these diverse language pairs will help determine if the cognitive metrics identified here remain robust across different orthographies and reading directions.
- **Hardware**: Exploring higher-frequency sensors that could improve temporal resolution.

7. Acknowledgements

We would like to express our gratitude to Professor Krzysztof Krejtz for his invaluable comments and guidance throughout this study. His expertise in eye tracking research was instrumental in shaping our approach and ensuring the rigor of our analysis.

The work on this project was supported by the grant CZ.02.01.01/00/23_020/0008518 ("Jazykověda, umělá inteligence a jazykové a řečové technologie: od výzkumu k aplikacím").

8. Bibliographical References

Ondřej Bojar, Filip Děchterenko, and Maria Zelenina. 2016. A pilot eye-tracking study of wmt-style ranking evaluation. pages 20–26.

Stephen Doherty, Sharon O'Brien, and Michael Carl. 2010. Eye tracking as an automatic mt evaluation technique. *Machine Translation*, pages 1–13.

Yvette Graham, Timothy Baldwin, Alistair Moffat, and Justin Zobel. 2013. Continuous measurement scales in human evaluation of machine translation. pages 33–41.

M. A. Just and P. A. Carpenter. 1980. A theory of reading: From eye fixations to comprehension. *Psychological Review*, pages 329–354.

Tom Kocmi, Ekaterina Artemova, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Konstantin Dranch, Anton Dvorkovich, Sergey Dukanov, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Marzena Karpinska, Philipp Koehn, Howard Lakougna, Jessica Lundin, Christof Monz, Kenton Murray, Masaaki Nagata, Stefano Perrella, Lorenzo Proietti, Martin Popel, Maja Popović, Parker Riley, Mariya Shmatova, Steinþór Steingrímsson, Lisa Yankovskaya, and Vilém Zouhar. 2025. Findings of the wmt25 general machine translation shared task: Time to stop evaluating on easy test sets. In *Proceedings of the Tenth Conference on Machine Translation (WMT 2025)*, pages 355–413, Suzhou, China. Association for Computational Linguistics.

Tom Kocmi, Vilém Zouhar, Eleftherios Avramidis, Roman Grundkiewicz, Marzena Karpinska, Maja Popović, Mrinmaya Sachan, and Mariya Shmatova. 2024. Error span annotation: A balanced approach for human evaluation of machine translation. pages 1440–1453, Miami, Florida, USA.

Alon Lavie, Greg Hanneman, Sweta Agrawal, Diptesh Kanojia, Chi-kiu Lo, Vilém Zouhar,

Frederic Blain, Chrysoula Zerva, Eleftherios Avramidis, Sourabh Dattatray Deoghare, Archchana Sindhujan, Jiayi Wang, David Ifeoluwa Adelani, Brian Thompson, Tom Kocmi, Markus Freitag, and Daniel Deutsch. 2025. Findings of the wmt25 shared task on automated translation evaluation systems: Linguistic diversity is challenging and references still help. In *Proceedings of the Tenth Conference on Machine Translation (WMT 2025)*, pages 414–461, Suzhou, China. Association for Computational Linguistics.

Arle Lommel, Aljoscha Burchardt, Maja Popović, Kim Harris, Eleftherios Avramidis, and Hans Uszkoreit. 2014. Using a new analytic measure for the annotation and analysis of MT errors on real data. pages 165–172.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. pages 311–318.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, 12:2825–2830.

Maja Popović. 2020. Informative manual evaluation of machine translation output. pages 5059–5069.

Yanina Prystauka, Gerry T. M. Altmann, and Jason Rothman. 2023. Online eye tracking and real-time sentence processing: On opportunities and efficacy for capturing psycholinguistic effects of different magnitudes and diversity. *Behavior Research Methods*, 56:3504—-3522.

Hassan Sajjad, Francisco Guzmán, Nadir Durrani, Ahmed Abdelali, Houda Bouamor, Irina Temnikova, , and Stephan Vogel. 2016. Eyes don't lie: Predicting machine translation quality using eye movement. pages 1082–1088.

Salvucci, Dario D., and Joseph H. Goldberg. 2000. Identifying fixations and saccades in eye-tracking protocols. pages 71—-78.

Skipper Seabold and Josef Perktold. 2010. statsmodels: Econometric and statistical modeling with python.

Sara Stymne, Henrik Danielsson, Sofia Bremin, Hongzhan Hu, Johanna Karlsson, Anna Prytz Lillkull, and Martin Wester. 2012. Eye tracking as a tool for machine translation error analysis. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 1121–1126, Istanbul, Turkey. European Language Resources Association (ELRA).

Myrte Vos, Serge Minor, and Gillian Catriona Ramchand. 2022. Comparing infrared and webcam eye tracking in the visual world paradigm. *FGlossa Psycholinguistics*, 1.

Onur Özsoy, Büsra Çiçek, Zeynep Özal, Natalia Gagarina, and Irina A. Sekerina. 2023. Turkish-german heritage speakers' predictive use of case: webcam-based vs. in-lab eye-tracking. *Frontiers in Psychology*, 14.