Predicting Total Reading Time Using Romanian Eye-Tracking Data

Anamaria Hodivoianu[§] **Oleksandra Kuvshynova**[§]
Filip Popovici[⋆] **Adrian Luca**[⋆] **Sergiu Nisioi**[§*]

[§] Faculty of Mathematics and Computer Science
[⋆] Faculty of Psychology and Education Sciences
University of Bucharest
anamaria.hodivoianu@gmail.com
oleksandra.kuvshynova@unibuc.ro
sergiu.nisioi@unibuc.ro

Abstract

This work introduces the first Romanian eye-tracking dataset for reading and investigates methods for predicting word-level total reading times. We develop and compare a range of models, from traditional machine learning using handcrafted linguistic features to fine-tuned Romanian BERT architectures, demonstrating strong correlations between predicted and observed reading times. Additionally, we propose a lexical simplification pipeline that leverages these TRT predictions to identify and substitute complex words, enhancing text readability. Our approach is integrated into an interactive web tool, illustrating the practical benefits of combining cognitive signals with NLP techniques for Romanian, a language with limited resources in this area.

1 Introduction

Total Reading Time (TRT) refers to the cumulative duration a reader fixates on a given word, including all refixations. As an eye-tracking metric, TRT serves as a reliable indicator of the cognitive processing involved in both semantic and deep syntactic analysis during reading (Frazier and Rayner, 1982; Pickering et al., 2004). Unlike other reading-time metrics that may capture only initial attention, TRT reflects the full depth of engagement a word receives, offering valuable insight into processing difficulty.

The prediction of word-level reading times and their relationship to textual complexity have a long history of investigations. Previous studies demonstrate that models designed to estimate eye-tracking measures, such as first fixation duration and total reading time, can serve as effective indicators of text readability (González-Garduño and Søgaard, 2017). Furthermore, eye-tracking data has been used to improve neural network models; for example, Barrett et al. (2018) incorporate human attention patterns into recurrent neural networks, resulting in improved performance on a range of NLP tasks. More recently, research by Hollenstein et al. (2021) shows that large language models, including multilingual BERT, can approximate human reading behavior, supporting the integration of cognitive signals into language model development and evaluation. Additionally, it has been observed that transformer models inherently encode eye-tracking information during pre-training (Dini et al., 2025), and that intermediate fine-tuning with eye-tracking data does not negatively impact downstream task performance.

In this paper, we present a work-in-progress and several initial experiments on predicting word-level TRT using eye-tracking data collected from native Romanian speakers. Our work introduces the first dataset of Romanian eye tracking recordings collected in the framework of MultiplEYE[1] and we propose a variety of machine learning approaches to estimate TRT. All code is publicly available[2].

Accurate TRT prediction can inform a range of downstream applications, particularly in the development of cognitively informed tools such as lexical simplification systems and reading aids (Duffy et al., 1988).

2 Data

The dataset used in this study originates from the MultiplEYE project (Jakobi et al., 2025), and it represents the first eye-tracking corpus for reading in the Romanian language. It includes recordings from four participants, all of whom are native Romanian speakers.

[§] Corresponding authors.

[1] https://multipleye.eu/

[2] https://github.com/ana0101/eye-tracking

Proceedings of First International Workshop on Gaze Data and NLP associated with RANLP 2025, pages 71–75, Varna, Bulgaria, Sep 12, 2025.
https://doi.org/10.26615/978-954-452-104-2-009

The reading materials consist of 14 texts: 10 main texts, 2 practice texts, and 2 backup texts. These texts span a variety of genres, the majority being official Romanian translations from multiple source languages. Due to minor translation inconsistencies and updates across sessions, each participant read a slightly different version of the texts. The eye-tracking experiments were conducted using the EyeLink 1000 Plus system.

We process the raw gaze data using the *Pymovements* library (Krakowczyk et al., 2023) to extract fixations and their alignment to corresponding words in the text. For each word, the total reading time is computed as the average duration across all participants. To analyze how much the TRT varies between the participants, we calculate the coefficient of variation as the mean TRT divided by the standard deviation of the TRT. The coefficients are between 0 and 2, with a mean of 1.02 and a median of 0.98. The variation is quite high, which is expected given the small number of participants.

Figure 1 presents a histogram of the resulting, averaged TRT values. A significant number of words received a reading time of zero milliseconds, indicating that these words are skipped entirely during reading. This is a known and expected phenomenon, especially for short or high-frequency function words. Out of the 778 skipped words by all the participants, 689 are function words. At the opposite end of the spectrum, some examples of the words with the highest TRTs are *distorsiune* (distortion), *cosmodromică* (cosmodromic), *gravifice* (gravitational), and *premergătoare* (preliminary), which are all long, complex words. All TRT values were standardized to have a mean of 0 and a standard deviation of 1.

3 Results

We evaluate our models for predicting word-level TRT using several metrics: Mean Squared Error (MSE), R^2 score, Pearson and Spearman correlation coefficients, and Accuracy. Accuracy is defined as $100 - \text{MAE}$, where MAE is the Mean Absolute Error, with TRT values scaled to the $[0, 100]$ range, following established practices in eye-tracking prediction (Hollenstein et al., 2021).

We consider two primary modeling approaches: (1) traditional machine learning models trained on handcrafted features, and (2) fine-tuning pretrained BERT models.

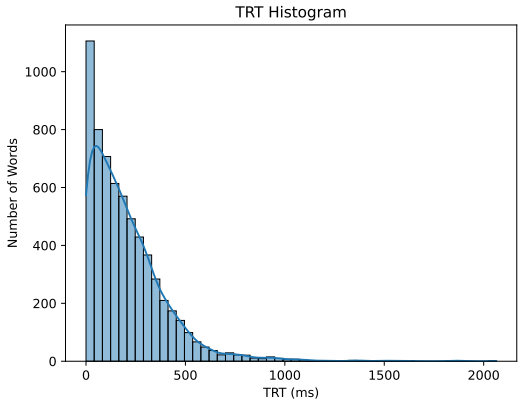

Figure 1: Histogram of TRT.

3.1 Feature Extraction and Analysis

For each word, we extract several features to aid in predicting the TRT. These include basic scalar attributes such as word length, the number of subword tokens generated by the Romanian BERT tokenizer (Dumitrescu et al., 2020), word frequency (obtained via the `wordfreq` library (Speer, 2022)), and the log probability of the word within its sentence context, estimated using a masked language modeling approach.

To calculate the log probability, we employ the pre-trained Romanian BERT model (Dumitrescu et al., 2020). The process involves first tokenizing the target word to determine its number of subword tokens. Then, these tokens are replaced in the sentence by an equal number of `[MASK]` tokens. The masked sentence is passed through the language model, which outputs probability distributions for each `[MASK]` token. The log probability for the word is taken as the negative logarithm of the probability assigned to the original first token by the model. While we also experiment with summing or averaging the log probabilities across all subword tokens, using only the first token's log probability yields better predictive performance.

In addition to scalar features, we derive contextual embeddings to capture semantic and syntactic information. We extract these embeddings from multiple layers of Romanian BERT: from the first, middle, last, and an average of all layers. Since BERT tokenizes words into subword units, we aggregate the embeddings of all tokens belonging to the same word by averaging them. This aggregation relies on character offset alignments to accurately map subword tokens back to their original words.

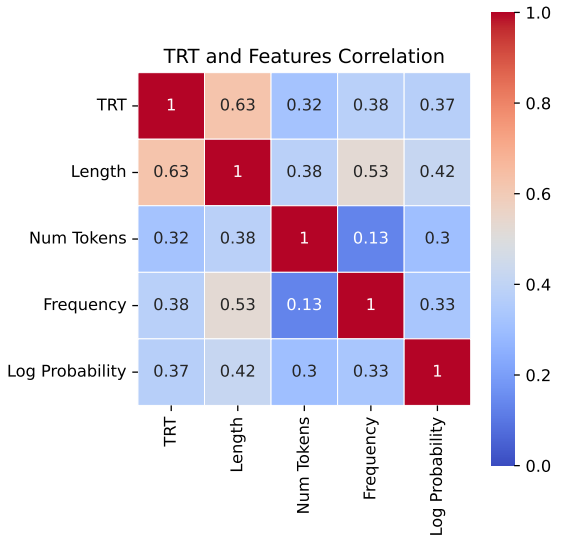

Figure 2: Pearson correlation between TRT and extracted features.

Figure 2 presents the Pearson correlation coefficients between TRT and the scalar features. Among these, word length shows the strongest correlation, followed by frequency, log probability, and number of tokens. Notably, the negative correlation between word length and frequency suggests that longer words tend to appear less frequently.

3.2 Traditional Regression Models

We train several regression models, including Linear Regression, Support Vector Regression, Random Forest, Gradient Boosting, Ridge Regression, Neural Networks, and more. Each model is trained on three feature sets: (1) only scalar features, (2) only embeddings, and (3) a combination of both. The data is split into train and test sets, with 80% of the data used for training and 20% for testing, which means 3796 words for training and 949 words for testing. When splitting the data, the words from the same sentence are not present in both the train and test sets. The features and reading times are standardized to have zero mean and unit variance.

All models achieve similar results: Pearson correlations between 0.6 and 0.7, accuracy between 70% and 95%, MSE between 0.4 and 0.7, Spearman correlation between 0.6 and 0.75, and R^2 scores ranging from 0.25 to 0.5.

Models trained on scalar features slightly outperform those trained on embeddings alone, although combining both types yields the best results overall. Among the embeddings, the average of all BERT layers generally performs best, so only the results with these embeddings were considered.

3.3 Fine-Tuning Pre-trained Language Models

We also fine-tune two Romanian BERT-based architectures:

- **BERT for Token Classification**: Modified to output a single regression value per token.
- **BERT with Regression Head**: Includes a linear layer, ReLU activation, layer normalization, dropout, and a final regression layer.

For token-level prediction, the TRT value of a word is assigned to each of its subtokens. During inference, token-level predictions are averaged to compute the word-level TRT.

The data is split into train, validation, and test sets, with 80% of the data used for training, 10% for validation, and 10% for testing. The train set contains 250 sentences, while the validation and test sets contain 25 sentences each. The reading times are standardized to have to have zero mean and unit variance.

Training is done in three phases using a gradual unfreezing strategy. For the first model, we unfreeze 4 additional layers every 8 epochs; for the second, we begin with only the regression head for 5 epochs and then unfreeze 6 layers every 10 epochs. Both models use the AdamW optimizer with a learning rate of 10^{-4}, weight decay of 10^{-4}, a cosine learning rate scheduler with warmup, batch size of 8, and dropout of 0.3. Padding tokens are ignored in the loss computation.

Table 1 summarizes the results. Both models perform comparably, achieving Pearson correlations around 0.7, Spearman correlations around 0.73, MSE near 0.5, and accuracy close to 90%, results that are similar to one of the best-performing traditional models, a neural network trained on all features.

4 Discussion

Our experiments demonstrate that predicting word-level reading times is feasible using both straightforward approaches, such as linear regression based on easily interpretable features like word length and frequency, as well as more sophisticated methods involving fine-tuning transformer-based language

Model	MSE	R^2	Pearson	Spearman	Accuracy
BERT for Token Classification	0.52	0.46	0.70	0.73	89.81
BERT with Regression Head	0.49	0.47	0.69	0.73	90.58
Neural network (all features)	0.41	0.44	0.69	0.74	90.43

Table 1: Performance of models.

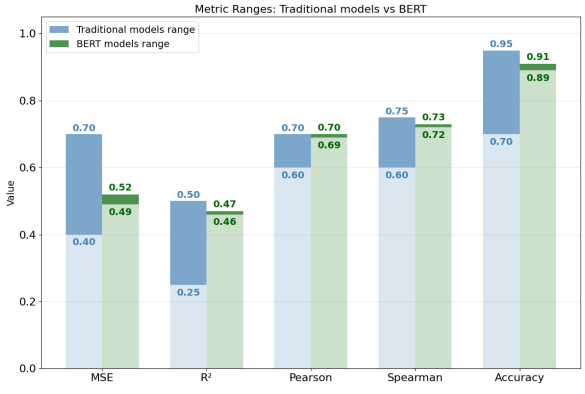

Figure 3: Metric ranges comparison between traditional models and BERT models.

models. The strong correlation between reading times and these features confirms their significance in capturing cognitive processing effort. Figure 3 illustrates the compared results of both methods.

One compelling application of accurate reading time prediction lies in lexical simplification. Since TRT effectively reflects the processing difficulty of words, it serves as a reliable indicator of lexical complexity. By identifying words with high TRT and substituting them with alternatives predicted to have lower TRT, we can enhance text readability and reduce overall reading effort.

To realize this, we implement a lexical simplification pipeline that first estimates the TRT for all words in a given text, selects candidates with elevated TRT, and generates potential replacements using the Romanian BERT masked language model (Dumitrescu et al., 2020). By masking the target word and leveraging the model's contextual predictions, we produce candidate substitutions. Inspired by Qiang et al. (2020), we experimented with concatenating the original and modified sentences in different orders to improve candidate quality, finding comparable improvements from both strategies. Before computing the predicted TRT for the candidates, we first make sure that the original word and the candidate are the same part of speech.

To make these capabilities accessible, we developed a user-friendly web interface called *Reading Time Estimator*. This tool enables users to input text, visualize predicted reading times on a word-by-word basis, and interactively simplify complex words by selecting suitable replacements with lower predicted TRT.

Overall, our work highlights the practical benefits of integrating cognitive signals such as eye-tracking data into NLP applications, particularly for languages like Romanian that have limited resources. A

5 Conclusions

In this paper, we introduced the first Romanian eye-tracking dataset focused on reading behavior, and demonstrated its utility in predicting word-level total reading time using both traditional machine learning and fine-tuned transformer-based models. Our experiments show that features such as word length and frequency are strong predictors of TRT, and that fine-tuned Romanian BERT models can achieve high predictive performance.

We also explored the practical implications of reading time prediction in the context of lexical simplification, proposing a pipeline that uses TRT estimates to identify and replace complex words. This system is implemented in an interactive web application that showcases the potential for user-centered NLP tools grounded in human reading behavior.

Our results highlight the value of eye-tracking data for advancing human-centered language technologies and pave the way for further work on Romanian and other low-resource languages in the domain of cognitive NLP.

Acknowledgments

This work was partially funded by the Romanian National Research Council (CNCS) through the Executive Agency for Higher Education, Research, Development and Innovation Funding (UEFISCDI) under grant PN-IV-P2-2.1-TE-2023-2007 (InstRead), and is supported by COST Action MultiplEYE, CA21131.

References

Maria Barrett, Joachim Bingel, Nora Hollenstein, Marek Rei, and Anders Søgaard. 2018. Sequence classification with human attention. In *Proceedings of the 22nd Conference on Computational Natural Language Learning*, pages 302–312, Brussels, Belgium. Association for Computational Linguistics.

Luca Dini, Lucia Domenichelli, Dominique Brunato, and Felice Dell'Orletta. 2025. From human reading to NLM understanding: Evaluating the role of eye-tracking data in encoder-based models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 17796–17813, Vienna, Austria. Association for Computational Linguistics.

Susan A Duffy, Robin K Morris, and Keith Rayner. 1988. Lexical ambiguity and fixation times in reading. *Journal of Memory and Language*, 27(4):429–446.

Stefan Dumitrescu, Andrei-Marius Avram, and Sampo Pyysalo. 2020. The birth of Romanian BERT. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4324–4328, Online. Association for Computational Linguistics.

Lyn Frazier and Keith Rayner. 1982. Making and correcting errors during sentence comprehension: Eye movements in the analysis of structurally ambiguous sentences. *Cognitive Psychology*, 14(2):178–210.

Ana Valeria González-Garduño and Anders Søgaard. 2017. Using gaze to predict text readability. In *Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 438–443, Copenhagen, Denmark. Association for Computational Linguistics.

Nora Hollenstein, Federico Pirovano, Ce Zhang, Lena Jäger, and Lisa Beinborn. 2021. Multilingual language models predict human reading behavior. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 106–123, Online. Association for Computational Linguistics.

Deborah Noemie Jakobi, Maja Stegenwallner-Schütz, Nora Hollenstein, Cui Ding, Ramune Kaspere, Ana Matić Škorić, Eva Pavlinusic Vilus, Stefan Frank, Marie-Luise Müller, Kristine M Jensen de López, Nik Kharlamov, Hanne B. Søndergaard Knudsen, Yevgeni Berzak, Ella Lion, Irina A. Sekerina, Cengiz Acarturk, Mohd Faizan Ansari, Katarzyna Harezlak, Pawel Kasprowski, Ana Bautista, Lisa Beinborn, Anna Bondar, Antonia Boznou, Leah Bradshaw, Jana Mara Hofmann, Thyra Krosness, Not Battesta Soliva, Anila Çepani, Kristina Cergol, Ana Došen, Marijan Palmovic, Adelina Çerpja, Dalí Chirino, Jan Chromý, Vera Demberg, Iza Škrjanec, Nazik Dinçtopal Deniz, Dr. Inmaculada Fajardo, Mariola Giménez-Salvador, Xavier Mínguez-López, Maroš Filip, Zigmunds Freibergs, Jéssica Gomes, Andreia Janeiro, Paula Luegi, João Veríssimo, Sasho Gramatikov, Jana Hasenäcker, Alba Haveriku, Nelda Kote, Muhammad M. Kamal, Hanna Kundefineddzierska, Dorota Klimek-Jankowska, Sara Kosutar, Daniel G. Krakowczyk, Izabela Krejtz, Marta Łockiewicz, Kaidi Lõo, Jurgita Motiejūnienundefined, Jamal A. Nasir, Johanne Sofie Krog Nedergård, Ayşegül Özkan, Mikuláš Preininger, Loredana Pungă, David Robert Reich, Chiara Tschirner, Špela Rot, Andreas Säuberli, Jordi Solé-Casals, Ekaterina Strati, Igor Svoboda, Evis Trandafili, Spyridoula Varlokosta, Mila Vulchanova, and Lena A. Jäger. 2025. Multipleye: Creating a multilingual eye-tracking-while-reading corpus. In *Proceedings of the 2025 Symposium on Eye Tracking Research and Applications*, ETRA '25, New York, NY, USA. Association for Computing Machinery.

Daniel G. Krakowczyk, David R. Reich, Jakob Chwastek, Deborah N. Jakobi, Paul Prasse, Assunta Süss, Oleksii Turuta, Paweł Kasprowski, and Lena A. Jäger. 2023. pymovements: A python package for processing eye movement data. In *2023 Symposium on Eye Tracking Research and Applications*, ETRA '23, New York, NY, USA. Association for Computing Machinery.

Martin J Pickering, Steven Frisson, Brian McElree, and Matthew J Traxler. 2004. Eye movements and semantic composition. In *The on-line study of sentence comprehension*, pages 33–50. Psychology Press.

Jipeng Qiang, Yun Li, Yi Zhu, Yunhao Yuan, and Xindong Wu. 2020. Lsbert: A simple framework for lexical simplification.

Robyn Speer. 2022. rspeer/wordfreq: v3.0.