Where Do Eyes Focus during Reading? A Context-Aware Model for Multilingual Prediction of Eye Movements

Xiangyu Lu[1], **Ben Yao**[1], **Xiaofan Sun**[1]

[1]MSc in IT and Cognition, University of Copenhagen
xilu@di.ku.dk, ben.yao@di.ku.dk, dcs354@alumni.ku.dk

Abstract

In psycholinguistics research, eye-tracking data are extensively used to investigate the brain's interpretation and processing of visual information in reading behavior. The accurate prediction of eye-tracking features is helpful in providing insightful perspectives for cognitive processes such as attention and information processing. The current study introduces a novel system that combines several pre-trained language models, a feature reassembler and a feature predictor to identify eye gaze features of tokens in six distinct languages. Trained separately on each of the six languages and four eye-tracking features, the model demonstrates notable capacity for this task. This is evidenced by achieving top 2 lowest MAE in 22 out of 24 the features of different languages compared to the baseline and other high-performing models. Regarding overall performance across the four features, the current model exhibits best results, as indicated by the lowest average MAE of 3.60.

1 Introduction

In the field of psycholinguistics, eye tracking is regarded as a highly reliable and useful method for examining information processing in human reading behavior [1]. As one of the human cognitive indicators, eye tracking data can also be leveraged in domains of machine learning, including computer vision (e.g., [2, 3]) and natural language processing (e.g., [4, 5, 6]). Accurate interpretation of eye movements holds the potential to enhance the human-like qualities of computational cognitive models and improves the efficiency of Natural Language Processing (NLP) models [7].

Eye movement can exhibit certain universal characteristics in reading behavior across diverse languages [8], and it also depends upon the given linguistic stimulus and is inherently language-specific. These different aspects are possibly connected to orthographies of different languages and linguistic features at both phonological and morphemic levels [9]. Therefore, the inclusion of a multilingual eye-tracking dataset in this study is important for a thorough analysis of human reading behavior, considering that reading behavior is realized through language representation.

Given the current multilingual eye-tracking dataset [10], the focus of this study is to predict four eye-tracking metrics for tokens in six different languages. To tackle this task, we develop a new model architecture[1] comprising multiple encoders, each tailored to a specific language. This system also includes a reassembler that reconstructs embeddings and a predictor focused on identifying features crucial for predicting First Fixation Duration (FFD) or Total Reading Time (TRT). The final prediction is evaluated using the average mean absolute error (MAE) metric. In comparison to the baseline and similar previous tasks, our model demonstrates exceptional effectiveness, achieving a comparably low MAE across more than half of the eye-tracking features in the six languages (see Table 4 and 5), with highly notable success in the TRT Avg feature.

[1] The source code of the project is available at $https://github.com/X1angyuLu/Reading_{p}atten_{p}rediction$.

2 Related Works

Eye movement, extensively used for studying cognitive processes during reading [1], corresponds with Just and Carpenter's eye-mind hypothesis, stating that the eye's fixation and the mind's processing occur simultaneously without significant delay [11]. This aligns with the assumption in [12] that the widespread use of eye tracking stems from its instant and measurable nature. To complete the cognitive process of reading, eye movements serve as the key method for encoding the visual information essential for further processing in brain [8]. Despite the close relationship between eye-movement behavior and human visual attention [13], eye gaze is not globally consistent but is influenced by a variety of factors. These factors include individual differences such as reading skills [14], along with semantic preview [15], lexical predictability [16], written forms and linguistic features of a certain language[8], and aspects like word frequency and complexity [11]. As a cognitive signal, eye gaze is significant in solving some NLP tasks, especially where the reader's mental engagement is crucial and text features alone are insufficient [17]. A notable example is the analysis of sarcastic texts [18], where the in-text inconsistency often results in prolonged fixations and regressions in eye traces, thereby making eye tracking a valuable modality for detecting sarcastic meanings.

In NLP tasks, particularly those similar to the current task, fine-tuning pre-trained models based on transformer architecture has been the predominant approach for addressing multilingual and cross-lingual comprehension since the introduction of the transformer architecture [19]. These pre-trained models are favored in such tasks due to their exceptional ability to generate contextual embeddings, which can capture semantic language features [20]. For experimenting with the multilingual dataset, various models such as multilingual BERT [21] and XLM-RoBERTa [22] are adopted, both of which are trained on large multilingual corpora over 100 languages and have demonstrated impressive task-specific performance across various languages [23]. However, one of core limitations of multilingual transformers has been noted: in the models with fixed capacity, parameter updates that benefit one language can negatively impact another when an extra language for pretraining is introduced [22]. To tackle this challenge in multilingual models, strategies involving language-specific sub-networks with shared global parameters have been proposed [24], proving beneficial for cross-lingual transfer in languages with limited training data.

Previous research has extensively developed fusion models that combine linguistic features with token representations for such tasks. A novel preprocessing method is introduced to standardize languages into a universally comparable International Phonetic Alphabet (IPA) format, generating features like word length and IPA vowel counts [25]. Contextual embeddings from three multilingual transformer models are employed in [26], augmented by hand-crafted input features such as word length, character counts, and relative length for enhanced uniformity and contextual awareness. Lexical features, including word position, length, frequency, capitalization, and syllable count, are incorporated in [27]. However, a single transformer model with task-specific feature engineering in multilingual contexts often lacks sufficient robustness, with certain features being inapplicable to some languages. Additionally, while previous research typically treats standard deviation and mean in isolation, we argue for the necessity of unified consideration in models due to their inherent connection.

3 Aim of the Study

The primary objective of this study is to fill the existing research gap highlighted in the literature by investigating how readers visually process words within sentences across various languages. The study is developed around two central hypotheses:

Hypothesis 1: Eye-tracking data in the multilingual dataset can uncover underlying cognitive patterns in human reading behavior, showcasing both common traits and language-dependent features.

Hypothesis 2: The details of human reading behavior can be detected by predicting eye traces through the novel deep learning model.

4 Data Analysis

4.1 Dataset Description

The dataset utilized in this research is obtained from the 2022 Cognitive Modeling and Computational Linguistics (CMCL) shared task [10], which collects data from multilingual corpora for the prediction of eye-tracking metrics during reading. In this dataset, reading is defined as the comprehension of sentences consisting of multiple tokens, rather than merely recognizing separate words. Distinguishing itself from similar datasets, this dataset's notable characteristic lies in its multilingual feature, extending beyond English to cover a diverse language realm.

In this dataset, human reading behavior is captured through four major eye-tracking metrics. Firstly, First Fixation Duration (FFD) measures the average duration of the initial gaze fixation on the given word. Secondly, Total Reading Time (TRT) represents the cumulative duration of all fixations on the specific word, inclusive of re-reading back to the word. The task of predicting these eye-tracking attributes and their variability across readers can be considered as a regression problem.

The dataset includes eye-tracking data from eight distinct corpora, representing six languages: Chinese, Dutch, English, German, Hindi, and Russian. These languages are markedly different, spanning different linguistic families, branches, and writing systems, which enriches the dataset's complexity and scope. Comprehensive data statistics are detailed in Table 1.

Language	Training		Validation		Test		Corpus
	Sentences	Tokens	Sentences	Tokens	Sentences	Tokens	
Chinese	120	1355	7	82	23	248	[28]
English	626	12790	38	724	119	2561	[29, 30, 31]
Dutch	640	7462	40	403	120	1475	[32]
German	80	1463	5	139	16	293	[33]
Hindi	122	2021	7	142	24	433	[34]
Russian	115	1140	7	59	22	218	[35]

Table 1. The numbers of sentences and tokens in the training, validation, and test sets for the subdatasets of six distinct languages, along with the source corpora for them.

4.2 Statistical Analysis

To investigate the common traits and language-specific characteristics of human reading behaviors, an illustrative statistical analysis is implemented.

Firstly, the means of the average value and standard deviations of TRT and FFD for six languages are computed using the original dataset (see sub-image A in Figure 1). Besides, we are also curious about how much information a reader receive at the first glance, therefore the ratio of FFD to TRT is calculated for each token. From these calculations, an aggregated average FFD/TRT ratio is determined for each of the six languages. The ratios shown in sub-figure B in Figure 1) illustrate the phenomenon of re-reading across different languages.

Subsequently, we focus on examining potential reading habits for specific parts of speech (POS) across these languages. This involves measuring the cumulative reading time allocated to each POS category in the six languages. Considering the diverse reading durations observed for tokens across languages, we normalize the average TRT for each token within each language. This normalization involves dividing the TRT average by the total TRT average computed for its respective language. After that, we opt for the StanfordNLP model [36] to assign universal POS tags to each token (see sub-image C in Figure 1).

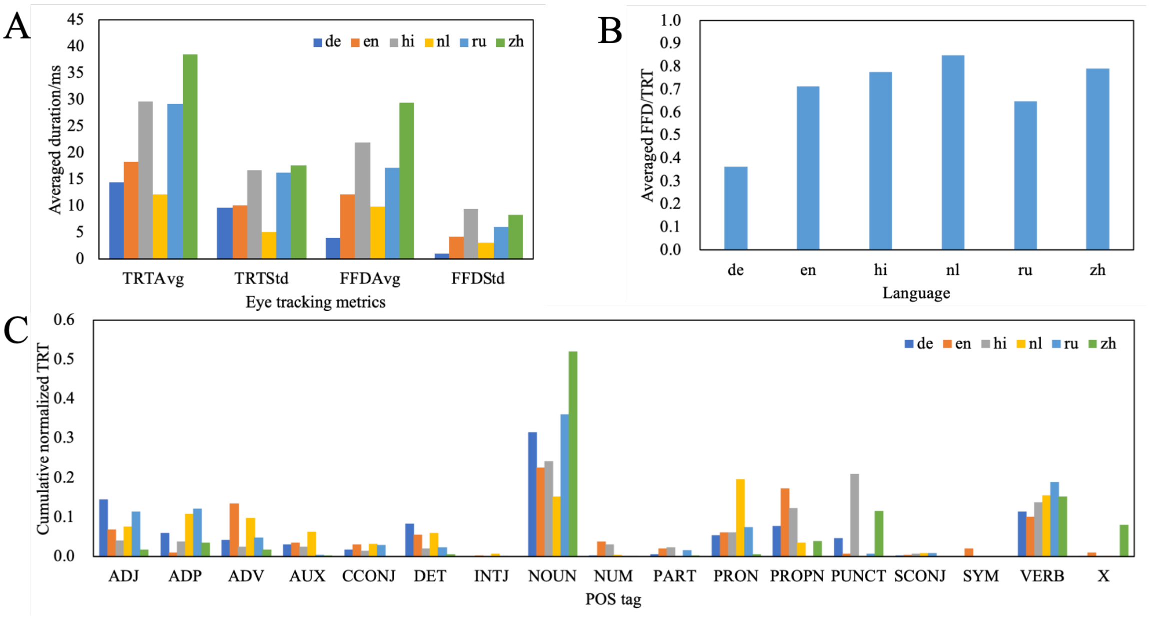

Figure 1. Data visualization of eye tracking data in six languages. A: The average eye-tracking data derived from the original dataset in six languages. B: The average ratio of FFD to TRT in six languages. A higher FFD/TRT value suggests the lower possibility of re-reading. C: The cumulative normalized TRT average across universal Part-of-Speech (POS) tags in six languages. A higher cumulative normalized TRT value implies the more time spent to read words with this specific POS tag.

5 Methodology

In this section, we initially provide a comprehensive overview of our proposed context-aware models designed for the analysis of reading patterns. Specifically, the model takes pre-segmented sequences of tokens as input, with the capability to predict the FFD and TRT average as well as the standard deviation of each individual token incurred by a collective readership when reading sentences in six distinct languages. Subsequently, the specific experimental details pertaining to the training of the model will be elaborated.

5.1 Model Details

Our model comprises three main components: an encoder responsible for encoding tokens, a feature reassembler reconstructing embeddings, and a predictor determining features relevant for predicting FFD or TRT. For each input token, the encoder initially transforms it into an embedding rich in semantic information conducive to subsequent prediction. Subsequently, based on the model's category (FFD feature or TRT feature), the specific token's embeddings and those of its neighboring tokens are integrated and reassembled by the feature reassembler before being fed into the predictor to obtain the final output. The overall architecture of our model is illustrated in Figure 2.

Considering the constraints imposed by the limited size of dataset, training a semantic encoder from scratch entails significant limitations. Consequently, we opted for leveraging pre-trained BERT models that have demonstrated exceptional performance on language processing tasks. It allows us to capitalize on the rich semantic representations embedded in these models that have been trained on extensive corpora, enhancing the overall robustness and effectiveness of our model.

It is worthy to note the distinctive characteristics of FFD and TRT, where FFD represents the fixation duration when a reader encounters a word for the first time, influenced solely by the word and its preceding tokens in a left-to-right reding sequence; and TRT represents the total time a reader spends fixating on a word throughout the reading of an entire sentence, more inclined to be

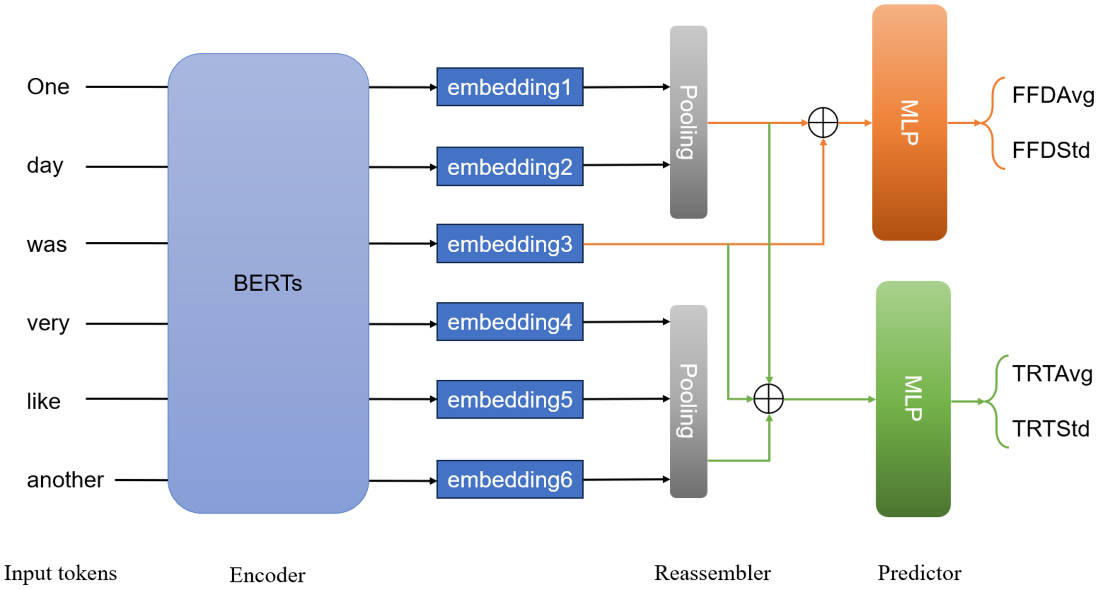

Figure 2. The overall structure of our proposed context-aware model, including an encoder based on pre-trained BERT, a predictor for forecasting eye-tracking features, and a reassembler dedicated to reconstructing embeddings for the input of the predictor.

influenced by every token in the entire sentence. Thus, our reassembler is designed as follows: for FFD prediction, the pooling of embeddings from all tokens preceding the target token is concatenated with the embedding of the target token and used as input to the predictor; for TRT prediction, the concatenation involves the pooling of embeddings from all tokens preceding the target token as well as the pooling of embeddings from all tokens succeeding the target token, combined with the embedding of the target token and provided as input to the predictor.

After integration, the embeddings are forwarded to the predictor, which constitutes a Multilayer Perceptron (MLP) composed of three fully connected layers. ReLU layers and dropout layers are interleaved between each pair of consecutive layers, contributing to the non-linearity of the model and aiding in regularization, respectively.

5.2 Experiments

Data Preprocessing. The text samples provided in the original dataset have undergone preliminary preprocessing. Specifically, for each complete sentence, tokenization has been performed by simple space-based segmentation, and each token has been annotated with four ground truth features: FFDAvg, FFDStd, TRTAvg, and TRTStd. However, the quality of this tokenization does not align with the requirements of the BERT input, as some punctuation marks are combined with words. The BERT tokenizer may potentially separate them again, resulting in a mismatch between the number of tokens and ground truth values. To address this issue, we conducted additional preprocessing. We removed non-alphabetic or non-numeric parts at the beginning and end of all pre-tokenized tokens (typically punctuation marks). However, for tokens composed entirely of non-alphabetic or non-numeric characters, no modifications were made. Additionally, the dataset contains some "dirty data" instances, such as tokens that failed to be properly segmented due to issues like space separation failures. Considering their limited quantity and negligible impact on the model's overall functionality, we chose not to address them in the preprocessing.

Training Settings. In our experimental setup, the training of our models was conducted on a NVIDIA RTX 3060 and a NVIDIA RTX 4090. We employed the Adam optimizer for all models, with an initial learning rate set at 0.00001. Following a warm-up period of 100 epochs, the learning rate gradually decayed to 0.000001 over the remaining 300 epochs. For various languages, distinct pre-trained BERT models were utilized as encoders, and the parameters of these BERT models remained frozen during the training process. Additionally, the size of the MLP in the predictor was adjusted based on the language, with specific configurations outlined in Table 2.

Language	**BERTs**	**MLP size**
Chinese	bert-base-chinese [21]	1024
English	bert-base-uncased [21]	2048
Dutch	GroNLP/bert-base-dutch-cased [37]	1024
German	bert-base-german-cased [38]	1024
Hindi	monsoon-nlp/hindi-bert [39]	1024
Russian	DeepPavlov/rubert-base-cased [40]	1024

Table 2. The names of the pre-trained BERT models utilized for each language and the dimensions of the fully connected layers in the Multilayer Perceptron.

Evaluation. To assess the performance of our models, we conducted a comparative analysis with the models presented in the two top-performing papers, referred to as [27] and [41], from the CMCL 2022 shared task. In order to ensure comparability, we adopted the same evaluation metric—MAE (1) —as the loss function and test metric, consistent with the referenced papers. Additionally, we incorporated the same baseline employed in CMCL 2022 where a naive predictor disregards input information and consistently predicts the average values of the four features across 6 languages.

$$MAE = \frac{\sum_{i=1}^{n} \mid y_i - x_i \mid}{n} \tag{1}$$

6 Results

Table 3 presents the overall prediction results on the four eye-tracking features, while Tables 4 and 5 depict the outcomes obtained by our models which are trained on the multiple languages. Overall, our models outperform more than half of the competing models in the four eye-tracking features across the six languages, with 22 out of 24 models ranking in the top two. This positions our models as the best performers in three out of the four features based on average MAE metrics and attains an overall MAE of 3.60, making it the best-performing model among all considered.

In separate evaluations for each language, our model demonstrates outstanding performance on large datasets, such as English, as well as on small datasets, such as Chinese. Moreover, for each language, our model achieves optimal predictions for at least one feature, indicating the robustness of our context-aware design.

7 Discussion

The results of the statistical analysis in the data analysis section support our first hypothesis. Concerning global traits, it is observed that readers prefer looking at content-rich words such as nouns and verbs. This preference can be attributed to the role of these words in forming semantic relationship and the cognitive focus on semantic processing during reading [12].

In each language, different reading patterns become prominent. Specifically, Chinese readers exhibit the longest durations for initial reading and re-reading, probably due to the absence of spacing between words in Chinese orthography. Since it is the words rather than single characters that provide the fundamental carriers of semantic information in Chinese reading [42], the lack of interword spacing complicates word boundary identification within a sentence [43] and further extends both durations. Besides, the lowest average FFD to TRT ratio in German, indicates faster initial processing but a greater tendency for re-reading. The unique feature of noun capitalisation in German might provide grammatical cues for identifying a word's syntactic category [44], potentially speeding up the

Model	FFD Avg	FFD Std	TRT Avg	TRT Std	MAE
Baseline	2.78	2.26	6.60	5.35	4.25
Salicchi et al. (2022) [27]	2.58	2.14	6.04	4.86	3.90
Takmaz (2022) [41]	**2.31**	2.11	5.82	4.77	3.75
Ours	2.35	**2.05**	**5.35**	**4.64**	**3.60**

Table 3. The MAE scores across all languages on the test set for baseline and other models. The bold numbers indicate that the models perform best on a specific feature among all models. The MAE term in the last column is the average of scores for four features. Our model surpasses both the baseline and the Salicchi et al.'s model model [27] across all the four features and also performs better than Takmaz's model [41] in the FFD Std, TRT Avg, and TRT Std metrics.

Model	FFD Avg						FFD Std					
	de	**en**	**hi**	**nl**	**ru**	**zh**	**de**	**en**	**hi**	**nl**	**ru**	**zh**
Baseline	0.50	5.04	2.80	1.88	2.79	3.69	0.46	2.10	2.82	1.56	2.92	3.71
Salicchi et al. (2022) [27]	0.45†	5.37	**2.31**	**1.72**	**2.45**	3.18	0.45	2.21	2.64	**1.47**	**2.43**	3.61†
Takmaz (2022) [41]	**0.43**	**3.24**	2.55	1.88	2.61†	3.16†	0.44	1.96	2.72	1.57	2.64	3.36
Ours	0.45†	3.56†	2.39†	1.86†	2.71	**3.15**	**0.42**	**1.84**	**2.59**	1.55†	2.57†	**3.34**

Table 4. The MAE scores for FFD Avg and FFD Std features in all languages on the test set for baseline and other models. The bold numbers indicate that the models perform best on a specific feature among all models while the numbers with '†' indicate that the models perform second best. Our model achieves top 2 lowest MAE in 11 out of 12 the features of different languages compared to the baseline and other high-performing models.

Model	TRT Avg						TRT Std					
	de	**en**	**hi**	**nl**	**ru**	**zh**	**de**	**en**	**hi**	**nl**	**ru**	**zh**
Baseline	6.05	8.30	6.42	3.29†	8.23†	7.35	4.46	5.66	6.25	2.94	5.67	7.11
Salicchi et al. (2022) [27]	**3.05**	8.28	5.32†	3.34†	9.71	6.52†	**2.57**	5.42	**5.23**	**2.78**	6.34	6.84
Takmaz (2022) [41]	3.78†	6.84†	5.75	3.35	8.36	6.82	2.89	4.95	5.57†	2.94	5.56†	6.70†
Ours	3.84	**6.33**	**5.24**	**3.27**	**7.83**	**5.60**	2.88†	**4.93**	5.36†	2.87†	**5.55**	**6.26**

Table 5. The MAE scores for TRT Avg and TRT Std features in all languages on the test set for baseline and other models. The bold numbers indicate that the models perform best on a specific feature among all models while the numbers with '†' indicate that the models perform second best. Our model achieves top 2 lowest average MAE in 11 out of 12 the features of different languages compared to the baseline and other high-performing models.

processing of initial fixation. As indicated in [11], the total time spent on a fixated word is negatively correlated with word frequency. This relationship can account for the extended total processing time in German, characterized by lengthy compound words that often necessitate decoupling the words for understanding its semantic meaning.

The satisfactory results in predicting eye-tracking traits affirm our second hypothesis, as evidenced by our model's superior performance compared to those employing feature engineering techniques, like the model in [27]. In this study, positional features extracted from texts are designed to collect sequential token information within sentences, yet they fall short in incorporating the semantic attributes of surrounding words. Additionally, the feature extraction method employed in previous works might lead to less optimal learning in some languages, suggesting that the features are not fully acquired by the model. In contrast, our model effectively captures contextual information by considering the words preceding and succeeding the target token, providing ample contextualized information for enhanced prediction precision.

When evaluating the language where our model excels the most, Chinese stands out. In comparison, models employing hand-crafted features extraction may not adequately align with Chinese [27], leading to less satisfying results for this language (see Table 4 and 5). Our model utilizes a BERT encoder pre-trained specifically for Chinese, enabling superior extraction of semantic information and syntactic structures. This tailored approach enhances the model performance in processing Chinese text. Regarding the specific eye-tracking features our model predicts most effectively, it can be observed that the model excels in TRT, particularly in its average value. This superior performance can be attributed to the model's strengthened generalization ability, stemming from the distinct encoders for each of the languages and our unique approach of considering preceding and upcoming words for richer contextual information. Consequently, it can be inferred that our model's improved prediction of TRT is linked to its ability to capture the contextual information governing TRT throughout the sentence.

8 Conclusion

In our study aiming at investigating human reading behavior through eye-tracking feature prediction, we introduce a context-aware approach with several language encoders. Trained on the four features of six languages in the CMCL 2022 dataset, our system exhibits impressive performance in detecting subtle eye movement patterns across six distinctly different languages. Compared to previous models in similar tasks, our model achieves superior results in FFD Std, TRT Avg, and TRT Std features, and delivers a comparable result in the FFD Avg feature on the test set. This proves that the current model architecture with language-specific BERT transformers can effectively capture cognitive aspects of human reading behavior, thereby enhancing our understanding of language processing in the human brain.

References

[1] K. Rayner, "Eye movements in reading and information processing: 20 years of research." *Psychological bulletin*, vol. 124, no. 3, p. 372, 1998.

[2] T. Judd, K. Ehinger, F. Durand, and A. Torralba, "Learning to predict where humans look," in *2009 IEEE 12th international conference on computer vision*. IEEE, 2009, pp. 2106–2113.

[3] N. Karessli, Z. Akata, B. Schiele, and A. Bulling, "Gaze embeddings for zero-shot image classification," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4525–4534.

[4] Y. Berzak, C. Nakamura, S. Flynn, and B. Katz, "Predicting native language from gaze," *arXiv preprint arXiv:1704.07398*, 2017.

[5] Y. Berzak, B. Katz, and R. Levy, "Assessing language proficiency from eye movements in reading," *arXiv preprint arXiv:1804.07329*, 2018.

[6] N. Hollenstein and C. Zhang, "Entity recognition at first sight: Improving ner with eye movement information," *arXiv preprint arXiv:1902.10068*, 2019.

[7] N. Hollenstein, M. Barrett, M. Troendle, F. Bigiolli, N. Langer, and C. Zhang, "Advancing nlp with cognitive language processing signals," *arXiv preprint arXiv:1904.02682*, 2019.

[8] S. P. Liversedge, D. Drieghe, X. Li, G. Yan, X. Bai, and J. Hyönä, "Universality in eye movements and reading: A trilingual investigation," *Cognition*, vol. 147, pp. 1–20, 2016.

[9] C. A. Perfetti, "The universal grammar of reading," *Scientific studies of reading*, vol. 7, no. 1, pp. 3–24, 2003.

[10] N. Hollenstein, E. Chersoni, C. L. Jacobs, Y. Oseki, L. Prévot, and E. Santus, "Cmcl 2022 shared task on multilingual and crosslingual prediction of human reading behavior," in *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, 2022, pp. 121–129.

[11] M. A. Just and P. A. Carpenter, "A theory of reading: from eye fixations to comprehension." *Psychological review*, vol. 87, no. 4, p. 329, 1980.

[12] S. Roussel, A. Rohr, E. Raufaste, and J.-L. Nespoulous, "Eye-movement analysis in reading content words and function words," *Cognition*, 2016.

[13] J. E. Hoffman and B. Subramaniam, "The role of visual attention in saccadic eye movements," *Perception & psychophysics*, vol. 57, no. 6, pp. 787–795, 1995.

[14] V. Kuperman and J. A. Van Dyke, "Effects of individual differences in verbal skills on eye-movement patterns during sentence reading," *Journal of memory and language*, vol. 65, no. 1, pp. 42–73, 2011.

[15] E. R. Schotter, M. Lee, M. Reiderman, and K. Rayner, "The effect of contextual constraint on parafoveal processing in reading," *Journal of memory and language*, vol. 83, pp. 118–139, 2015.

[16] A. Staub, "The effect of lexical predictability on eye movements in reading: Critical review and theoretical interpretation," *Language and Linguistics Compass*, vol. 9, no. 8, pp. 311–327, 2015.

[17] S. Mathias, D. Kanojia, A. Mishra, and P. Bhattacharyya, "A survey on using gaze behaviour for natural language processing," *arXiv preprint arXiv:2112.15471*, 2021.

[18] A. Mishra, D. Kanojia, and P. Bhattacharyya, "Predicting readers' sarcasm understandability by modeling gaze behavior," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 30, no. 1, 2016.

[19] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[20] N. F. Liu, M. Gardner, Y. Belinkov, M. E. Peters, and N. A. Smith, "Linguistic knowledge and transferability of contextual representations," *arXiv preprint arXiv:1903.08855*, 2019.

[21] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[22] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, and V. Stoyanov, "Unsupervised cross-lingual representation learning at scale," *arXiv preprint arXiv:1911.02116*, 2019.

[23] T. Pires, E. Schlinger, and D. Garrette, "How multilingual is multilingual bert?" *arXiv preprint arXiv:1906.01502*, 2019.

[24] R. Choenni, D. Garrette, and E. Shutova, "Cross-lingual transfer with language-specific subnetworks for low-resource dependency parsing," *Computational Linguistics*, vol. 49, no. 3, pp. 613–641, 2023.

[25] J. M. Imperial, "Nu hlt at cmcl 2022 shared task: Multilingual and crosslingual prediction of human reading behavior in universal language space," *arXiv preprint arXiv:2202.10855*, 2022.

[26] H. Srivastava, "Zero shot crosslingual eye-tracking data prediction using multilingual transformer models," *arXiv preprint arXiv:2203.16474*, 2022.

[27] L. Salicchi, R. Xiang, and Y.-Y. Hsu, "Hkamsters at cmcl 2022 shared task: Predicting eye-tracking data from a gradient boosting framework with linguistic features," in *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, 2022, pp. 114–120.

[28] J. Pan, M. Yan, E. M. Richter, H. Shu, and R. Kliegl, "The beijing sentence corpus: A chinese sentence corpus with eye movement data and predictability norms," *Behavior Research Methods*, pp. 1–12, 2021.

[29] N. Hollenstein, J. Rotsztejn, M. Troendle, A. Pedroni, C. Zhang, and N. Langer, "Zuco, a simultaneous eeg and eye-tracking resource for natural sentence reading," *Scientific data*, vol. 5, no. 1, pp. 1–13, 2018.

[30] N. Hollenstein, M. Troendle, C. Zhang, and N. Langer, "Zuco 2.0: A dataset of physiological recordings during natural reading and annotation," *arXiv preprint arXiv:1912.00903*, 2019.

[31] S. G. Luke and K. Christianson, "The provo corpus: A large eye-tracking corpus with predictability norms," *Behavior research methods*, vol. 50, pp. 826–833, 2018.

[32] U. Cop, N. Dirix, D. Drieghe, and W. Duyck, "Presenting geco: An eyetracking corpus of monolingual and bilingual sentence reading," *Behavior research methods*, vol. 49, pp. 602–615, 2017.

[33] L. A. Jäger, T. Kern, and P. Haller, "Potsdam textbook corpus (potec)," OSF, January 2021. [Online]. Available: https://doi.org/10.5167/uzh-212186

[34] S. Husain, S. Vasishth, and N. Srinivasan, "Integration and prediction difficulty in hindi sentence comprehension: Evidence from an eye-tracking corpus," *Journal of Eye Movement Research*, vol. 8, no. 2, 2015.

[35] A. K. Laurinavichyute, I. A. Sekerina, S. Alexeeva, K. Bagdasaryan, and R. Kliegl, "Russian sentence corpus: Benchmark measures of eye movements in reading in russian," *Behavior research methods*, vol. 51, pp. 1161–1178, 2019.

[36] P. Qi, T. Dozat, Y. Zhang, and C. D. Manning, "Universal dependency parsing from scratch," *arXiv preprint arXiv:1901.10457*, 2019.

[37] W. De Vries, A. van Cranenburgh, A. Bisazza, T. Caselli, G. van Noord, and M. Nissim, "Bertje: A dutch bert model," *arXiv preprint arXiv:1912.09582*, 2019.

[38] Deepset, "Framework for adapting representation models," https://github.com/deepset-ai/FARM, 2019.

[39] Nick Doiron, “hindi-bert (revision c54eb83),” 2023. [Online]. Available: https://huggingface.co/monsoon-nlp/hindi-bert

[40] Y. Kuratov and M. Arkhipov, “Adaptation of deep bidirectional multilingual transformers for russian language,” *arXiv preprint arXiv:1905.07213*, 2019.

[41] E. Takmaz, “Team dmg at cmcl 2022 shared task: Transformer adapters for the multi-and cross-lingual prediction of human reading behavior,” in *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, 2022, pp. 136–144.

[42] X. Bai, G. Yan, S. P. Liversedge, C. Zang, and K. Rayner, “Reading spaced and unspaced chinese text: evidence from eye movements.” *Journal of experimental psychology: Human perception and performance*, vol. 34, no. 5, p. 1277, 2008.

[43] C. Zang, F. Liang, X. Bai, G. Yan, and S. P. Liversedge, “Interword spacing and landing position effects during chinese reading in children and adults.” *Journal of Experimental Psychology: Human Perception and Performance*, vol. 39, no. 3, p. 720, 2013.

[44] M. Vogelzang, N. Fuhrhop, T. Mundhenk, and E. Ruigendijk, “Influence of capitalisation and presence of an article in noun phrase recognition in german: Evidence from eye-tracking,” *Journal of Research in Reading*, 2023.

Appendix

Section	Contributor(s)
Abstract	Ben
Introduction	Ben, Xiangyu, Xiaofan
Related Works	Xiaofan
Aim of the Study	Xiaofan
Data Analysis	Ben
Methodology	Xiangyu
Results	Ben, Xiangyu, Xiaofan
Discussion	Ben, Xiangyu, Xiaofan
Conclusion	Xiangyu

Table 6. Division of Labor. The joint part does not exceed 50% of the total work.