Eye-tracking based classification of Mandarin Chinese readers with and without dyslexia using neural sequence models

Patrick Haller[1], **Andreas Säuberli**[1], **Sarah E. Kiener**[1]
Jinger Pan[3], **Ming Yan**[4], **Lena A. Jäger**[1,2]
[1]University of Zurich [2]University of Potsdam
[3]The Education University of Hong Kong [4]University of Macau
haller@cl.uzh.ch andreas@cl.uzh.ch sarahelisabeth.kiener@uzh.ch
jpan@eduhk.hk mingyan@um.edu.mo jaeger@cl.uzh.ch

Abstract

Eye movements are known to reflect cognitive processes in reading, and psychological reading research has shown that eye gaze patterns differ between readers with and without dyslexia. In recent years, researchers have attempted to classify readers with dyslexia based on their eye movements using Support Vector Machines (SVMs). However, these approaches (i) are based on highly aggregated features averaged over all words read by a participant, thus disregarding the sequential nature of the eye movements, and (ii) do not consider the linguistic stimulus and its interaction with the reader's eye movements. In the present work, we propose two simple sequence models that process eye movements on the entire stimulus without the need of aggregating features across the sentence. Additionally, we incorporate the linguistic stimulus into the model in two ways—contextualized word embeddings and manually extracted linguistic features. The models are evaluated on a Mandarin Chinese dataset containing eye movements from children with and without dyslexia. Our results show that (i) even for a logographic script such as Chinese, sequence models are able to classify dyslexia on eye gaze sequences, reaching state-of-the-art performance, and (ii) incorporating the linguistic stimulus does not help to improve classification performance.[1]

1 Introduction

Reading effortlessly constitutes a key skill in modern society. Individuals suffering from developmental dyslexia are characterized by specific and persistent reading problems. Global prevalence estimates range from 3 to 7% (Landerl et al., 2013; Peterson and Pennington, 2012). Previous research has consistently shown that early diagnosis and intervention is key to mitigate the resulting long-term consequences (Vaughn et al., 2010).

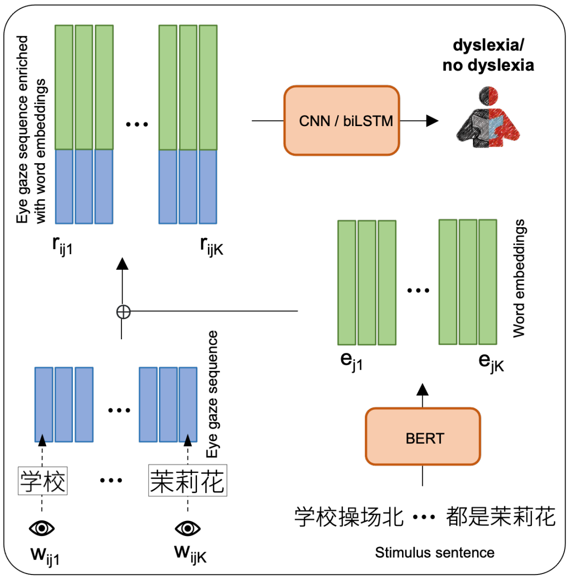

Figure 1: Proposed approach. Each eye-movement reading measure vector is concatenated with contextualized word embeddings and used as input for the sequence models to infer whether a reader suffers from dyslexia.

Psychological and clinical research on eye movement patterns has revealed that individuals with dyslexia exhibit gaze patterns that differ significantly from the patterns observed in individuals without dyslexia (Rayner, 1998; Pan et al., 2014). In particular, scanpaths of individuals with dyslexia are characterized by longer fixation durations, more fixations, decreased saccade durations and a higher proportion of regressions. In recent years, increasing effort has been spent on utilizing these findings and applying supervised classification methods such as SVMs and Random Forests on eye movement data (see Kaisar 2020 for an overview) to infer the presence or absence of dyslexia. There are several reasons why automatized approaches for assistance in dyslexia detection are desirable. Currently, paper-pencil diagnostic tools are conducted by trained speech therapists. These tools are time-intensive and are typically only considered after a suspected case has been reported by

[1] Model code is publicly available and can be found under https://github.com/hallerp/dyslexia-seqmod.

Proceedings of the Workshop on Text Simplification, Accessibility, and Readability (TSAR-2022), pages 111 - 118
December 8, 2022 ©2022 Association for Computational Linguistics

observant educational staff, leaving many cases overlooked. Eye-movement-based diagnostic tools have the potential to be deployed in schools in a relatively inexpensive manner and as part of a standard procedure aimed at early and comprehensive detection of dyslexia; making an important contribution to educational equity.

Although the aforementioned approaches provide promising results, they suffer from specific drawbacks: (i) The model input consists of eye movement features, aggregated for each subject over the presented stimulus material (text), thus disregarding the sequential nature of the eye movements; (ii) both the linguistic stimulus and its interaction with the reader's eye movements are not considered. For classification purposes, this does not pose a problem *per se*. However, it does not allow us to investigate questions such as: Which words (or, more specifically, what linguistic properties of the stimulus) are particularly informative to discriminate between individuals with and without dyslexia?

In the present work, we propose two neural sequence models, depicted in Figure 1, that process the eye movements on the entire stimulus without the necessity of feature aggregation over the sentence. To incorporate the linguistic stimulus into the model, we use pre-trained contextualized word embeddings. We evaluate our model on an eye-tracking-while-reading dataset from children with and without dyslexia reading Mandarin Chinese sentences by Pan et al. (2014).

2 Related Work

2.1 ML-based detection of dyslexia

To date, various data types and signals have been utilized to solve the task of automated detection of dyslexia such as text, MRI scans (Cui et al., 2016), EEG recordings (Frid and Breznitz, 2012), student engagement data (Abdul Hamid et al., 2018) as well as eye-tracking data (Rello and Ballesteros, 2015; Raatikainen et al., 2021; Benfatto et al., 2016). Benfatto et al. (2016) train a *Support Vector Machine with recursive feature elimination* (SVM-RFE) on 168 eye-tracking features obtained from an eye-tracking-while-reading dataset from 185 Swedish children (aged 9-10 years). Their best SVM-RFE model selected 48 features and achieved an accuracy score of 95.6% ± 4.5% (sic!) on a balanced dataset. We reimplement this method and use it as a reference method (cf. 4.1). Jothi Prabha and Bhargavi (2020), using the same dataset as Benfatto et al. (2016), experiment with various feature selection algorithms and machine learning models. They find that feature selection via Principle Component Analysis (PCA) in combination with a Particle Swarm Optimization based Hybrid Kernel SVM classifier yields the best accuracy.

Raatikainen et al. (2021) combine a Random Forest classifier for feature selection with an SVM, achieving an accuracy of 89.7%. They expand their feature space with transition matrices that represent the number of transitions between the different segments (question, answer selection) in a trial as well as the number of gaze shifts within one segment.

2.2 Modeling eye-tracking data with deep neural sequence models

Eye movement data for task inference. Deep neural sequence models have been deployed to solve inference tasks based on eye movements such as reader (Jäger et al., 2019) and viewer identification (Lohr et al., 2020; Makowski et al., 2020, 2021), ADHD detection (Deng et al., 2022) as well as the prediction of reading comprehension (Reich et al., 2022).

Integrating the linguistic stimulus. There has been growing interest in combining language and eye movement models to predict gaze patterns during naturalistic reading (Hollenstein et al., 2021; Merkx and Frank, 2021; Hollenstein et al., 2022). Wiechmann et al. (2022) investigate the role of general text features and their interaction with eye movement patterns in predicting human reading behavior and find that models incorporating the linguistic stimulus improves prediction accuracy.

3 Problem Setting

We investigate the two closely related tasks of classifying (i) whether a given eye gaze sequence on one sentence is from a reader with or without dyslexia and (ii) whether a given eye gaze sequence on a set of sentences is from a reader with or without dyslexia. Formally, our training data can be represented as a set $\mathcal{D} = \{(\mathbf{W}_{11}, y_1), \ldots, (\mathbf{W}_{NM}, y_N)\}$, where $\mathbf{W}_{ij} = \langle \mathbf{w}_{ij1} \ldots \mathbf{w}_{ijK} \rangle$ is a sequence of reading measure vectors[2] for each word $k \in 1 \ldots K_j$ obtained from subject i reading sentence j, where N is the number of participants, M is the number of stimulus sentences read by each of the participants and K_j

[2] Cf. the list of reading measures in Appendix B.

the number of words in a given sentence j. Each reading measure vector consists of R reading measures, i.e., $\mathbf{w}_{ijk} = (r_{ijk1} \dots r_{ijkR})$. The binary target label y_i denotes whether participant i is a reader with or without dyslexia. For (i), our goal is to train a binary classifier g_θ such that

$$\widehat{y_i} = \begin{cases} 1, & \text{if } g_\theta(\mathbf{W}_{ij}) \geq \delta \\ 0, & \text{else,} \end{cases}$$

where δ denotes the decision threshold and θ the set of hyperparameters. Accordingly, for (ii), $\widehat{y_i} = 1$, if $\frac{1}{M}\sum_{j=1}^{M} g_\theta(\mathbf{W}_{ij}) \geq \delta$.
The performance of a binary model can be characterized by a false-positive and a true-positive rate. By altering the decision threshold δ, a receiver operator characteristic (ROC) curve can be derived, with the area under the curve providing an aggregated measure for all possible values of δ.

4 Methods

4.1 Reference method

As a baseline method, we train an SVM-RFE, following the procedure described by Benfatto et al. (2016). We use the *scikit-learn* implementation (Pedregosa et al., 2011) of the SVM-RFE with a linear kernel. In the *subject-prediction* setting, we use eye movement features from each subject aggregated (mean and standard deviation) across trials and sentences as input vectors. In the *sentence-prediction* setting, we use aggregates of each sentence over all trials, yielding $2 \times 12 = 24$ features per instance in both settings.[3]

4.2 Proposed neural sequence models

Both models take as input an enriched reading measure vector $\mathbf{r}_{ij}$ (cf. Section 4.2.1) of a sentence j read by participant i, normalized for each train/test set separately, and predict a label y_i. We tune both models using random search.

LSTM. We implement a bidirectional recurrent neural network with LSTM cells. The mean of the hidden states is fed into a linear layer projecting it down to a single sigmoid output to represent the label prediction. Optimized hyperparameters and search space are reported in Appendix 2.

CNN. We implement a CNN that convolves the input accross the word sequence axis. It consists of two convolutional layers, each followed by a pooling layer, two dense layers, and a sigmoid output unit. Hyperparameters are listed in Appendix 2.

4.2.1 Incorporating the linguistic stimulus

Using contextualized word embeddings. To incorporate the linguistic stimulus (the words occurring in the current sentence), we first extract 768-dimensional BERT embeddings $\mathbf{e}_{jk}$ for each word w in a given sentence j, using the pre-trained BERT$_{\text{BASE}}$-embeddings, provided by Hugging Face (Wolf et al., 2020), and concatenate them with the reading measure vector $\mathbf{w}_{ijk}$, resulting in an enriched reading measure vector $\mathbf{r}_{ijk}$. Concatenating the full embedding to the feature vectors results in $768 + R$ dimensions, resulting in a substantial increase in parameters to be estimated. Given the small amount of available training data, we test two methods of dimensionality reduction: (i) We perform PCA on the word embeddings and use the first 20 principal components. (ii) *Mean-difference-encoding*: In order to capture domain-specific information from the word embeddings relating to differences in reading behaviour exhibited by individuals with and without dyslexia, we propose an alternative method, which we call *mean-difference-encoding*: We train a feed-forward neural network with one hidden layer of size 20 to predict differences between the mean values of each eye movement feature between the two groups for each word based on its original word embedding. The values of the hidden layer are a compressed representation of the original embedding that is optimized to encode information that discriminates between children with and without dyslexia. In order to avoid train-test data leakage, in each fold, the mean-difference-encoder is trained from scratch on the respective training set.

Using manually extracted features. As an alternative way to incorporate the linguistic stimulus, we add a range of *manually extracted linguistic features* for each token w_{jk} in sentence j: Surprisal, i.e., $-\log p(w_{jk} \mid \mathbf{w}_{j<k})$, estimated with GPT-2 (Radford et al., 2019), part of speech, dependency relation type, distance to syntactic head, extracted using spaCy (Honnibal et al., 2020), mean character frequency and lexical frequency extracted from SUBTLEX-CH (Cai and Brysbaert, 2010).

[3] We also experimented with training random forests as baseline, however, they were outperformed by the SVM-RFE.

5 Experiments

Data. We employ eye-tracking-while-reading data from 62 Mandarin Chinese children (33 with dyslexia) provided by Pan et al. (2014). Participants were instructed to read 60 sentences out loud while their eye movements were recorded. 40 sentences were selected from fifth grade textbooks and 20 additional control sentences were extracted from the Beijing Sentence Corpus (Pan et al., 2022). The dyslexia label had been assigned when a child scored at least 1.5 standard deviations below their corresponding age mean in standard character recognition test (Shu et al., 2003).

5.1 Evaluation procedure

We evaluate our models using 10-fold nested cross-validation in two settings. In the *sentence prediction* setting, we predict the label from a single sentence, read by a given subject. In the *subject prediction* setting, we average the sigmoid outputs from all sentences read by a given subject in order to obtain a subject-level prediction. In both settings, sentences are stratified over 10 folds, balanced by group. Data from the same subject is always constrained to one fold, thus, the model always makes predictions for unseen subjects.

Hyperparameter tuning. For each test fold, we iterate through 9 validation folds, training 50 LSTM and 100 CNN models using randomly sampled parameter combinations for each fold. We select the highest scoring parameter set over all 9 validation folds and train a final model using 8 training folds. We use one left-out fold for early stopping and evaluate it on the test fold.

5.2 Results

For all methods, we report AUC as well as accuracy, recall, precision and the harmonic precision-recall mean F_1 for a decision threshold of 0.5 on subject- and sentence-level. As can be seen in Table 1, our proposed models reach but do not outperform state-of-the art performance. While on subject-level, the CNN architecture enriched with PCA-reduced word embeddings achieves the highest AUC, on sentence-level, the best results are obtained by the LSTM that solely includes eye-movement features. Overall, we note that classification performance on subject-level is higher than on sentence-level and that adding the linguistic stimulus does not aid classification performance, neither as contextualized word embeddings nor as manually-extracted features. Furthermore, as can be seen in Figure 2, performance varies considerably with respect to different test sets. We also observe that the variance in AUC for models enriched with the linguistic stimulus is larger for LSTMs compared to CNNs. Lastly, our domain-specific dimensionality reduction method (cf. Section 4.2.1) has no advantage over PCA, although the former is explicitly trained on differences between the two groups.

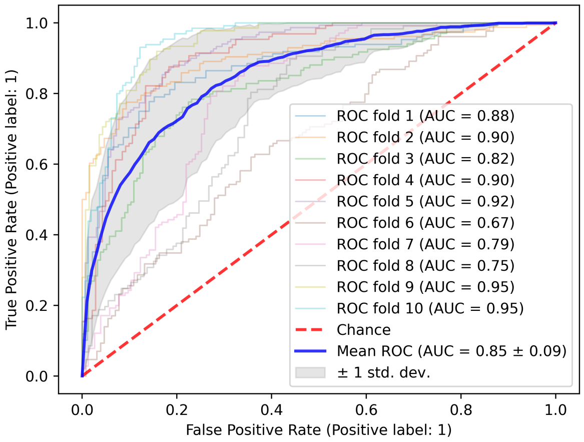

Figure 2: ROC curves over all test sets for best performing model (LSTM with no linguistic stimulus representation) on sentence-level.

6 Discussion

Our proposed neural sequence models reach state-of-the-art performance on solving the task of detecting dyslexia from eye gaze sequences, for the first time investigated for a logographic script such as Chinese. Our results suggest that for our dataset, (i) neural architectures processing eye-movement sequences along the sentence have no advantage over the parsimonious SVM-baseline where features are aggregated over the sentence, and (ii) enabling the interaction between stimulus input and eye movements does not improve classification performance. However, after having shown that our approach is able to reach SOTA performance, we aim to exploit its properties to investigate the informativeness of particular sentences, words, and other linguistic sub-units for dyslexia detection in the future.

Furthermore, for all investigated models, the overall performance appears to be driven by a small subset of individuals who presumably exhibit less typical reading behavior among their group and were more difficult to classify. Given that dyslexia is a spectrum disorder—not binary as it is often perceived—it is to be expected that individuals that are not located at the two extremes (clearly dyslexic

	Architecture		Evaluation Metrics				
	Model	Stimulus representation	AUC	Accuracy	Recall	Precision	F_1
SUBJECT-LEVEL	*Baseline*		**0.93** (±0.03)	**0.90** (±0.03)	0.87 (±0.04)	0.97 (±0.03)	0.91 (±0.03)
	LSTM	None	0.91 (±0.03)	**0.90** (±0.03)	**0.88** (±0.05)	0.98 (±0.02)	**0.92** (±0.03)
		BERT meandiff	0.88 (±0.03)	0.80 (±0.06)	0.78 (±0.06)	0.93 (±0.06)	0.83 (±0.06)
		BERT PCA	0.90 (±0.03)	0.83 (±0.05)	0.81 (±0.06)	0.97 (±0.03)	0.87 (±0.04)
		Manually extracted	0.87 (±0.04)	0.87 (±0.05)	0.84 (±0.05)	0.97 (±0.03)	0.89 (±0.04)
	CNN	None	0.91 (±0.04)	0.90 (±0.03)	0.86 (±0.04)	**1.00** (±0.00)	**0.92** (±0.02)
		BERT meandiff	0.91 (±0.03)	0.90 (±0.03)	0.88 (±0.04)	0.97 (±0.03)	0.91 (±0.02)
		BERT PCA	**0.93** (±0.03)	0.87 (±0.02)	0.86 (±0.04)	0.93 (±0.04)	0.88 (±0.02)
		Manually extracted	0.89 (±0.04)	0.83 (±0.04)	0.80 (±0.05)	0.97 (±0.03)	0.86 (±0.03)
SENTENCE-LEVEL	*Baseline*		**0.85** (±0.03)	**0.78** (±0.02)	**0.79** (±0.04)	0.76 (±0.02)	0.77 (±0.02)
	LSTM	None	**0.85** (±0.03)	0.77 (±0.03)	0.74 (±0.04)	0.83 (±0.03)	**0.78** (±0.03)
		BERT meandiff	0.81 (±0.04)	0.68 (±0.04)	0.65 (±0.04)	**0.86** (±0.05)	0.72 (±0.03)
		BERT PCA	0.79 (±0.04)	0.66 (±0.04)	0.64 (±0.04)	0.85 (±0.05)	0.71 (±0.03)
		Manually extracted	0.77 (±0.05)	0.71 (±0.03)	0.67 (±0.03)	0.85 (±0.05)	0.74 (±0.03)
	CNN	None	0.84 (±0.02)	0.76 (±0.02)	0.73 (±0.02)	0.83 (±0.04)	0.77 (±0.02)
		BERT meandiff	0.82 (±0.03)	0.75 (±0.02)	0.72 (±0.02)	0.82 (±0.04)	0.76 (±0.02)
		BERT PCA	0.82 (±0.03)	0.74 (±0.02)	0.70 (±0.02)	0.85 (±0.04)	0.76 (±0.02)
		Manually extracted	0.82 (±0.03)	0.74 (±0.02)	0.69 (±0.02)	**0.86** (±0.03)	0.76 (±0.02)

Table 1: Classification results using 10-fold cross validation on subject- and sentence-level. We report AUC, accuracy, recall, precision and $F1$ [results ± standard error]. The latter four were computed for a decision threshold of 0.5.

or clearly not dyslexic) are more difficult to classify in a binary environment.

Our study was able to show that an SVM-based approach, previously applied to alphabetic languages such as Swedish and Spanish, also works well on a logographic script such as Chinese. In future work, we would like to test our approach on alphabetic language data sets. This is particularly interesting given the fact that young Chinese readers are faced with different challenges, e.g., the absence of orthographic word boundaries, therefore requiring word segmentation, and the much larger number of characters required to be memorized.

Limitations. It should be noted that our dataset contained very little data. Considering that the number of parameters of our sequence models exceeded the one of the baseline model by orders of magnitude, it might be worth comparing the approaches again, once more data is available. The problem of data scarcity might be alleviated by pre-training on domain general eye-tracking datasets or with data augmentation methods[4]. Furthermore, we did not have access to the raw scores of the character recognition task. While our methods did not outperform the baseline in this binary environment, it would be interesting to assess their performance on a regression task.

7 Conclusion

For the first time, we deploy models to detect dyslexia from eye gaze sequences on data from Mandarin Chinese readers. We propose two sequence classification approaches that (i) take as input the full, non-aggregated linguistic stimulus and (ii) model the interaction of the stimulus with the eye movements. As a comparison, we adapt a previously proposed SVM-based approach for Mandarin Chinese. We find that all models reach SOTA performance for data based on a logographic script such as Chinese. In addition, we find that incorporating the linguistic stimulus does not improve the models' performance. Given that we reach SOTA performance on a very small dataset, our approach has proven worthwhile to be pursued, expanded, and further tested (e.g., on alphabetic language data sets). It has the potential to be successfully deployed in the context of automatized approaches for dyslexia detection with the final objective being the improvement of educational equity.

Acknowledgments

We thank David Reich for his continuous feedback on the model architectures as well as our anonymous reviewers for their invaluable feedback on the manuscript for this work. Lena Jäger was partially funded by the German Federal Ministry of Education and Research under grant 01|S20043.

[4] In a preliminary experiment, we pre-trained our models on the Beijing Sentence Corpus (Pan et al., 2022) and found that it did not increase classification performance.

References

Siti Suhaila Abdul Hamid, Novia Admodisastro, Noridayu Manshor, Azrina Kamaruddin, and Abdul Azim Abd Ghani. 2018. Dyslexia adaptive learning model: Student engagement prediction using machine learning approach. In *International Conference on Soft Computing and Data Mining*, pages 372–384. Springer.

Mattias Nilsson Benfatto, Gustaf Öqvist Seimyr, Jan Ygge, Tony Pansell, Agneta Rydberg, and Christer Jacobson. 2016. Screening for dyslexia using eye tracking during reading. *PLoS ONE*, 11(12):e0165508.

Qing Cai and Marc Brysbaert. 2010. SUBTLEX-CH: Chinese word and character frequencies based on film subtitles. *PLoS ONE*, 5(6):e10729.

Zaixu Cui, Zhichao Xia, Mengmeng Su, Hua Shu, and Gaolang Gong. 2016. Disrupted white matter connectivity underlying developmental dyslexia: A machine learning approach. *Human Brain Mapping*, 37(4):1443–1458.

Shuwen Deng, Paul Prasse, David R. Reich, Sabine Dziemian, Maja Stegenwallner-Schütz, Daniel Krakowczyk, Silvia Makowski, Nicolas Langer, Tobias Scheffer, and Lena A. Jäger. 2022. Detection of ADHD based on eye movements during natural viewing. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*.

Alex Frid and Zvia Breznitz. 2012. An SVM based algorithm for analysis and discrimination of dyslexic readers from regular readers using ERPs. In *2012 IEEE 27th Convention of Electrical and Electronics Engineers in Israel*, pages 1–4. Institute of Electrical and Electronics Engineers.

Nora Hollenstein, Itziar Gonzalez-Dios, Lisa Beinborn, and Lena A. Jäger. 2022. Patterns of text readability in human and predicted eye movements. In *Proceedings of the Workshop on the Cognitive Aspects of the Lexicon*. Association for Computational Linguistics.

Nora Hollenstein, Marius Tröndle, Martyna Plomecka, Samuel Kiegeland, Yilmazcan Özyurt, Lena A. Jäger, and Nicolas Langer. 2021. Reading task classification using EEG and eye-tracking data. *arXiv:2112.06310*.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spaCy: Industrial-strength natural language processing in Python. Zenodo. https://spacy.io/.

Lena A. Jäger, Silvia Makowski, Paul Prasse, Sascha Liehr, Maximilian Seidler, and Tobias Scheffer. 2019. Deep Eyedentification: Biometric identification using micro-movements of the eye. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, Lecture Notes in Computer Science, pages 299–314. Springer.

Appadurai Jothi Prabha and Renta Bhargavi. 2020. Predictive model for dyslexia from fixations and saccadic eye movement events. *Computer Methods and Programs in Biomedicine*, 195:105538.

Shahriar Kaisar. 2020. Developmental dyslexia detection using machine learning techniques: A survey. *ICT Express*, 6(3):181–184.

Karin Landerl, Franck Ramus, Kristina Moll, Heikki Lyytinen, Paavo H.T. Leppänen, Kaisa Lohvansuu, Michael O'Donovan, Julie Williams, Jürgen Bartling, Jennifer Bruder, Sarah Kunze, Nina Neuhoff, Dénes Tõth, Ferenc Honbolygõ, Valéria Csépe, Caroline Bogliotti, Stéphanie Iannuzzi, Yves Chaix, Jean François Démonet, Emilie Longeras, Sylviane Valdois, Camille Chabernaud, Florence Delteil-Pinton, Catherine Billard, Florence George, Johannes C. Ziegler, Isabelle Comte-Gervais, Isabelle Soares-Boucaud, Christophe Loïc Gérard, Leo Blomert, Anniek Vaessen, Patty Gerretsen, Michel Ekkebus, Daniel Brandeis, Urs Maurer, Enrico Schulz, Sanne Van Der Mark, Bertram Müller-Myhsok, and Gerd Schulte-Körne. 2013. Predictors of developmental dyslexia in European orthographies with varying complexity. *Journal of Child Psychology and Psychiatry and Allied Disciplines*, 54(6):686–694.

Dillon Lohr, Henry Griffith, Samantha Aziz, and Oleg Komogortsev. 2020. A metric learning approach to eye movement biometrics. In *2020 IEEE International Joint Conference on Biometrics*, pages 1–7. Institute of Electrical and Electronics Engineers.

Silvia Makowski, Lena A. Jäger, Paul Prasse, and Tobias Scheffer. 2020. Biometric identification and presentation-attack detection using micro- and macro-movements of the eyes. In *2020 IEEE International Joint Conference on Biometrics*, pages 1–10. Institute of Electrical and Electronics Engineers.

Silvia Makowski, Paul Prasse, David R. Reich, Daniel Krakowczyk, Lena A. Jäger, and Tobias Scheffer. 2021. DeepEyedentificationLive: Oculomotoric biometric identification and presentation-attack detection using deep neural networks. *IEEE Transactions on Biometrics, Behavior, and Identity Science*, 3(4):506–518.

Danny Merkx and Stefan L. Frank. 2021. Human sentence processing: Recurrence or attention? In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, pages 12–22. Association for Computational Linguistics.

Jinger Pan, Ming Yan, Jochen Laubrock, Hua Shu, and Reinhold Kliegl. 2014. Saccade-target selection of dyslexic children when reading Chinese. *Vision Research*, 97:24–30.

Jinger Pan, Ming Yan, Eike M. Richter, Hua Shu, and Reinhold Kliegl. 2022. The Beijing Sentence Corpus: A Chinese sentence corpus with eye movement data and predictability norms. *Behavior Research Methods*, 54(4):1989–2000.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Edouard Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Robin L. Peterson and Bruce F. Pennington. 2012. Developmental dyslexia. *The Lancet*, 379(9830):1997–2007.

Peter Raatikainen, Jarkko Hautala, Otto Loberg, Tommi Kärkkäinen, Paavo Leppänen, and Paavo Nieminen. 2021. Detection of developmental dyslexia with machine learning using eye movement data. *Array*, 12:100087.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI Blog*, 1(8):9.

Keith Rayner. 1998. Eye movements in reading and information processing: 20 years of research. *Psychological Bulletin*, 124(3):372–422.

David R. Reich, Paul Prasse, Chiara Tschirner, Patrick Haller, Frank Goldhammer, and Lena A. Jäger. 2022. Inferring native and non-native human reading comprehension and subjective text difficulty from scanpaths in reading. In *2022 Symposium on Eye Tracking Research and Applications*, ETRA '22, pages 1–8. Association for Computing Machinery.

Luz Rello and Miguel Ballesteros. 2015. Detecting readers with dyslexia using machine learning with eye tracking measures. In *Proceedings of the 12th International Web for All Conference*, W4A '15, pages 1–8. Association for Computing Machinery.

Hua Shu, Xi Chen, Richard C. Anderson, Ningning Wu, and Yue Xuan. 2003. Properties of school Chinese: Implications for learning to read. *Child Development*, 74(1):27–47.

Sharon Vaughn, Paul T. Cirino, Jeanne Wanzek, Jade Wexler, Jack M. Fletcher, Carolyn D. Denton, Amy Barth, Melissa Romain, and David J. Francis. 2010. Response to intervention for middle school students with reading difficulties: Effects of a primary and secondary intervention. *School Psychology Review*, 39(1):3–21.

Daniel Wiechmann, Yu Qiao, Elma Kerz, and Justus Mattern. 2022. Measuring the impact of (psycho-)linguistic and readability features and their spill over effects on the prediction of eye movement patterns. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, pages 5276–5290. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45. Association for Computational Linguistics.

A Pan et al.'s (2014) dataset

Each sentence was composed of seven to 13 words and each word consisted out of one to three characters, with 38 one-character words, 372 two-character words and 22 three-character words. Sentences in which a child blinked while reading a word, except the first and last one, are not included in the final dataset. The set therefore contains the data for between 24 up to 59 sentences for each child.

B Reading Measures

Word-level reading measures used as input for both the baseline models (aggregated over text or subject, respectively) and the neural models. All durations are in ms. Saccade distances refer to distances with respect to x/y-axis coordinates. Landing position refers to character index within a fixated word.

- Horizontal location of fixation on screen
- Total gaze duration (sum of all fixations landing on the word before moving away from it)
- Landing position of first fixation within the word
- Landing position of last fixation within the word
- Duration of first fixation
- Duration of outgoing saccade
- Horizontal distance of outgoing saccade
- Vertical distance of outgoing saccade
- Total distance of outgoing saccade
- Duration of incoming saccade
- Horizontal distance of incoming saccade
- Vertical distance of incoming saccade

C Hyperparameter tuning

Model	Hyperparameter	Range
Both	Batch size	$[8, 16, 32, 64, 128]$
	Learning rate	$15 \times \mathcal{U} \sim (1e^{-5}, 1e^{-1})$
	Decision boundary	$20 \times \mathcal{U} \sim (0.35, 0.65)$
LSTM	Hidden layer size	$[10, 20, \ldots, 70]$
CNN	C1 # channels	$[5, 10, \ldots, 30]$
	C1 kernel	$[3, 5]$
	C1 pooling	[average, max]
	C2 # channels	$[10, 20, \ldots, 50]$
	C2 kernel	$[3, 5]$
	C2 pooling	[average, max]
	L1 size	$[10, 20, \ldots, 60]$
	dropout	$[0.1, 0.2, \ldots 0.7]$

Table 2: Hyperparameter space for LSTMs and CNNs.