Automatic detection of dyslexia based on eye movements during reading in Russian

Anna Laurinavichyute
University of Potsdam
Department of Linguistics
Karl-Liebknecht-Straße 24-25
14476 Potsdam
laurinavichy@uni-potsdam.de

Anastasiya Lopukhina
Royal Holloway University of London
Department of Psychology
Egham TW20 0EX;
Center for Language and Brain, HSE University
Moscow, Krivokolenny pereulok, 3
Anastasiya.Lopukhina@rhul.ac.uk

David R. Reich
University of Potsdam
Department of Computer Science
An der Bahn 2, 14476 Potsdam;
University of Zurich
Dept of Computational Linguistics
reich@cl.uzh.ch

Abstract

Dyslexia, a common learning disability, requires an early diagnosis. However, current screening tests are very time- and resource-consuming. We present an LSTM that aims to automatically classify dyslexia based on eye movements recorded during natural reading combined with basic demographic information and linguistic features. The proposed model reaches an AUC of 0.93 and outperforms the state-of-the-art model by 7 %. We report several ablation studies demonstrating that the fixation features matter the most for classification.

1 Introduction

One of the most common learning disabilities is dyslexia, a difficulty that specifically affects reading and spelling in individuals with otherwise intact cognitive abilities. The prevalence of dyslexia is estimated to be between 9% and 12% (Katusic et al., 2001; Shaywitz et al., 1998). Early diagnosis is the key factor for getting the needed support and staying on track in the educational system (Glazzard, 2010; Torgesen, 2000; Vellutino et al., 2004).

There are various testing batteries for dyslexia, but most must be administered one-on-one by a trained specialist, who is not always present at school. Moreover, such batteries are still often evaluated using paper-and-pencil methods, which are time-consuming and error-prone. Without a cheap, fast, and reliable mass testing method, the only way to get proper support for a struggling reader is through the educator, who may notice reading difficulties and recommend additional testing.

For mass screening, data should be quick and affordable to obtain. Automatic classification based on (f)MRI and ERP recordings, while suitable from the machine learning perspective, does not fit this criterion. Eye tracking technology is more promising: It provides a rich signal, is unobtrusive, and is growing cheaper. Research on dyslexia and eye movements has a long history. Attempts to tie dyslexia to some characteristics of eye movements go at least as far back as 1980s (Olson et al., 1983; Pavlidis, 1981; Rayner, 1985). Since then, it has been established that the underlying deficit in dyslexia lies not in oculomotor control but rather in phonological decoding (International Dyslexia Association, 2024). Given that eye movements reflect phonological decoding in beginner and adult readers (Rayner et al., 1995, 1998; Blythe, 2014; Leinenger, 2019; Milledge and Blythe, 2019), inferring dyslexia from eye movements is strongly theoretically motivated.

Several machine-learning solutions based on eye movements have been proposed for the mass screening for dyslexia (Asvestopoulou et al., 2019; Nilsson Benfatto et al., 2016; Haller et al., 2022; Jothi Prabha and Bhargavi, 2022; Raatikainen et al., 2021; Rello and Ballesteros, 2015; Shalileh et al., 2023). Yet almost all of these models were trained on very modest samples of 61 (Asvestopoulou et al., 2019) to 185 participants (Nilsson Benfatto et al., 2016). This paper presents a comparison of two models that aim to automatically classify dyslexia on a large dataset comprising eye-movements while reading from 293 young readers of different ages.

1.1 Problem Setting

Inferring whether a child has dyslexia is a binary classification task, and the model's performance can be characterized by a false positive and a true positive rate. By altering the decision threshold, one can observe a receiver operator characteristic curve (ROC curve). The area under the ROC curve (AUC) provides an aggregated measure of performance for all possible classification thresholds.

Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pages 59–66

2 Experiments

2.1 Eye-movement data

The cross-sectional dataset contains eye movements while reading in 293 native speakers of Russian, from the 1st to the 6th grade (published by Shalileh et al. 2023). In Russia, grades 1 through 4 correspond to primary school, and grades 5 and 6 – to secondary school. Based either on a reading assessment or a speech therapist assessment, children were classified as typically developing ($N = 221$) or having developmental dyslexia ($N = 72$). Among children with dyslexia, 43 received their diagnosis based on reading assessment, and 29–based on therapist assessment. Reading assessment was based on the Standardized Assessment of Reading Skills test (SARS, Kornev and Ishimova 2010) and recent normative cut-offs reported by **?**. SARS requires a test-taker to read a short text aloud as quickly and as accurately as possible. The number of words read accurately in the first minute is taken as a measure of reading fluency. If a child scores at least 1.5 standard deviations below their corresponding age mean, a dyslexia label is assigned. Speech therapist assessment is based on a phonological test battery. Note that there are children with dyslexia (diagnosis based on a phonological assessment) whose reading speed is higher than that of some age-matched children in the control group.

All children had nonverbal intelligence scores within the normal range. The typically-developing children had age-appropriate reading fluency and comprehension. Their parents or primary caretakers reported no history of reading disorders. The detailed composition of both groups can be found in Table 1.

2.2 Reading materials

All children were asked to read the same set of 30 sentences comprising the Child Russian Sentence Corpus (Lopukhina et al. 2022). Sentence difficulty was at the level of 3rd to 4th grade, according to an automatic text difficulty measurement developed for Russian (Laposhina and Lebedeva, 2021), and estimated to be 7.42 on the Flesch-Kincaid scale adapted to Russian (Readability Test). Sentences were six to nine words long ($M = 7.6$, $SD = 0.85$). In total, children read 227 words, which contained 182 unique word forms (as words could be repeated across sentences). Individual words were on average 5.6 letters long (range 1–13), and had average lemma frequency of 50.29 items per million (median: 0.73, range: $0.0001 - 667$). The frequency was calculated from the sub-corpus of texts for children of the years 1920–2015 of the Russian National Corpus.

Since Russian is a morphologically rich language, and morphological composition of a word affects its reading time, the number of morphemes comprising each word was annotated. The annotation was first done by an automated parser and then reviewed by a trained linguist. Finally, each word's predictability was estimated using an online cumulative cloze task with 46 children who did not participate in the eye-tracking study. Predictability was measured as the number of correct guesses divided by the total number of guesses. Zero probabilities were replaced with $\frac{1}{2\times46}$, where 46 is the number of guesses for the word.

For a more detailed description of the dataset and reading materials, see Lopukhina et al. 2022.

2.3 Reference method

As a baseline, we use a state-of-the-art (SOTA) SVM-RFE with a linear kernel described and implemented by Haller et al. (2022). This approach was first proposed by Nilsson Benfatto et al. (2016), who reported 96% accuracy on a balanced dataset. As input, the SOTA model uses the means and standard deviations of 12 eye-movement features, such as first fixation duration, first-pass reading time, etc. (for the full list, refer to Haller et al. 2022).

Note that Haller et al. had a homogenous data set of age-matched readers, and did not include age as a predictor. Given that the present dataset includes readers of different ages, we report the performance of the SOTA model both without grade, for full comparability with Haller et al.'s results, and with grade, for a fairer comparison.

2.4 Proposed model

The model introduced by Haller et al. relied on aggregated reading measures. This aggregation, however, results in the loss of significant temporal and word-level information. We address this limitation by employing a sequential method—namely, LSTMs (Hochreiter and Schmidhuber, 1997)—to preserve and leverage temporal information. As demonstrated by previous studies (Ahn et al., 2020; Reich et al., 2022), this choice effectively captures the temporal dynamics necessary for eye-movement classification.

Table 1: Demographic and cognitive characteristics of both participant groups, organized by grade. Values before the slash ("/") represent the control group, while those after the slash correspond to participants with dyslexia.

Grade	**1** (N=50/8)	**2** (N=40/10)	**3** (N=37/20)	**4** (N=39/28)	**5** (N=31/6)	**6** (N=24/0)
Gender Female: N (%)	22 (44%) / 2 (25%)	24 (60%) / 2 (20%)	19 (51%) / 12 (60%)	18 (46%) / 9 (32%)	12 (39%) / 2 (33%)	10 (42%) / -
Age Mean ± SD	$7.32_{0.51}$ / $7.25_{0.46}$	$8.35_{0.48}$ / $8.40_{0.84}$	$9.30_{0.46}$ / $9.30_{0.57}$	$10.18_{0.56}$ / $10.25_{0.59}$	$11.29_{0.78}$ / $11.17_{0.41}$	$12.00_{0.59}$ / -
Nonverbal intelligence Mean ± SD	$29.88_{3.99}$ / $29.75_{4.74}$	$31.00_{3.23}$ / $29.00_{3.74}$	$31.24_{3.50}$ / $31.40_{5.75}$	$31.90_{3.59}$ / $32.14_{3.33}$	$32.81_{2.12}$ / $28.50_{4.85}$	$33.17_{2.39}$ / -
Reading speed (wpm) Mean ± SD	$63.80_{27.06}$ / $17.38_{8.52}$	$79.0_{17.54}$ / $30.70_{10.68}$	$95.57_{13.93}$ / $52.20_{20.48}$	$119.28_{20.67}$ / $57.50_{22.29}$	$122.48_{29.38}$ / $56.50_{16.60}$	$124.62_{23.50}$ / -

Our proposed model's input is a participant's fixation sequence on a sentence. Each input vector consists of basic demographic information, gaze-specific, and linguistic features:

(i) *Demographic features*: participant's age, grade, and gender. Age and grade are relevant for classification because reading skills in primary school are noticeably different between grades, and many reading evaluations are normed for a certain age (grade). Participant's gender is important from a clinical perspective. Boys are diagnosed with dyslexia more often than girls: the male-to-female sex ratio ranges from about 3:1 to 5:1 in self-identified samples, and from 1.5:1 to 3.3:1 in random samples. Arnett et al. (2017) claim that the difference in dyslexia rates between sexes is valid and is driven by lower and more variable processing speed in boys.
(ii) *Gaze-specific features*: fixation duration, fixation horizontal and vertical coordinates on the screen, landing position on the word, next fixation distance, next saccade amplitude, next saccade angle, next saccade velocity, and next saccade direction. We used all fixation features available through the eye movement recording device and reasoned that the importance of different features can be estimated through ablation studies.
(iii) *Linguistic features*: fixated word's length in letters, predictability and frequency (these features explain most of the eye-movement variance, see Kliegl et al. 2004; Shain 2019), as well as number of morphemes comprising the word.

We choose an BiLSTM-based architecture, where the mean of the hidden states is fed into two sequential linear layers, projecting it down to a single sigmoid output to represent the label prediction. Optimized hyperparameters and search space are reported in Appendix A.

3 Results

3.1 Model evaluation

The models are evaluated in two settings: prediction of the reader's status based on a single sentence data (sentence prediction setting) or based on all available reading data (reader prediction setting). In the reader prediction setting, predictions for individual sentences are aggregated to produce the final outcome. The motivation for introducing the sentence prediction setting was to identify whether some sentences serve as a better diagnostic material than others, and to evaluate model precision depending on the amount of input.

All models were evaluated and tuned using 10-fold nested cross-validation and random grid search (see Appendix A). Data from the same person is always constrained to one fold, so that the models always make predictions for unseen participants. The ratio of persons with/without dyslexia is balanced across all folds.[1]

3.2 Results

For all methods, we report AUC for reader- and sentence-level settings (see Table 2). A visual summary of ROC AUC performance can also be found in Figure 1. Classification performance in the reader-prediction setting was numerically higher than in the sentence-prediction setting. However, according to an unpaired one-tailed t-test, the difference between settings was not significant in any model (LSTM: $t(15.55) = 1.22$, $p = 0.12$; $\text{SOTA}_{\text{+Grade}}$: $t(16.21) = 0.81$, $p = 0.21$; $\text{SOTA}_{\text{-Grade}}$: $t(17.83) = 0.24$, $p = 0.41$). The SOTA model that included information on grade performed numerically better, but the difference was not significant (reader prediction setting: $t(17.96) = 1.03$, $p = 0.16$; sentence prediction setting: $t(16.64) = 0.71$, $p = 0.24$). Importantly, the proposed LSTM significantly outperformed

[1]All code is available online: https://github.com/annlaurin/Rus_dyslex_classification

the $\text{SOTA}_{\text{+Grade}}$ model in both reader-prediction ($t(12.146) = 2.12$, $p = 0.028$) and sentence-prediction settings ($t(17.92) = 2.20$, $p = 0.021$).

3.2.1 Ablation Studies

In the reader-prediction setting, we run additional ablation studies, assessing model performance without saccade-related measures (next fixation distance, next saccade amplitude, angle, velocity, and direction), without linguistic information (word length, frequency, predictability, and the number of morphemes), without demographic information (age, grade, and gender), and without all eye-movement features. In all ablation studies, AUC score was lower, but the decrease was not significant except for the model without all eye-movement features ($\text{LSTM}_{\text{-Saccade}}$: $t(14.70) = 0.95$, $p = 0.17$; $\text{LSTM}_{\text{-Ling}}$: $t(17.80) = 0.06$, $p = 0.47$; $\text{LSTM}_{\text{-Demographic}}$: $t(16.14) = 1.66$, $p = 0.058$; $\text{LSTM}_{\text{-All eye movement}}$: $t(17.93) = 3.25$, $p = 0.002$).

On average, children with dyslexia make more fixations than normally developing children. To evaluate whether the LSTM predictions are based mainly on the input length, we reduced the number of fixations to 29. The resulting AUC score did not differ from the score obtained on the whole input ($t(18) = 0.20$, $p = 0.42$).

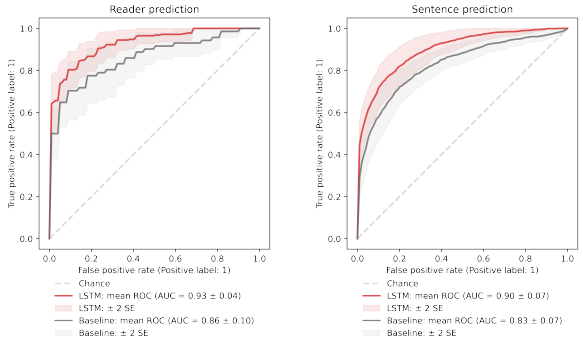

Figure 1: Summary of model performance. SOTA baseline model used grade information.

4 Discussion

The finding that the information on the grade of the reader did not significantly improve the performance of either model is rather surprising. In the present dataset, at least in some cases, dyslexia was diagnosed based on age-specific normative reading speed cut-offs (see Section 2.1). Consequently, information about reader's grade should be crucial for the classification performance. Grade-invariant performance might reflect that the model has captured some invariant property of the eye movements of readers with dyslexia. For the SOTA model trained exclusively on aggregated features, we consider this explanation unlikely. For the LSTM trained on a sequence of fixations, this explanation is more likely, but it is precisely the LSTM that shows a greater numerical decrease in performance without the grade information. In general, we believe that a successful model should be able to uncover the relationship between reading speed, grade, and dyslexia label.

		AUC
Reader	SOTA	$0.86_{\pm 0.10}$
	$\text{SOTA}_{\text{-Grade}}$	$0.81_{\pm 0.11}$
	LSTM	$0.93_{\pm 0.05}$
	$\text{LSTM}_{\text{-Ling}}$	$0.92_{\pm 0.05}$
	$\text{LSTM}_{\text{-Saccade}}$	$0.91_{\pm 0.07}$
	$\text{LSTM}_{\text{-Demographic}}$	$0.90_{\pm 0.06}$
	$\text{LSTM}_{\text{29 fixations}}$	$0.93_{\pm 0.04}$
	$\text{LSTM}_{\text{-All eye movement}}$	$0.87_{\pm 0.04}$
Sentence	SOTA	$0.83_{\pm 0.07}$
	$\text{SOTA}_{\text{-Grade}}$	$0.80_{\pm 0.10}$
	LSTM	$0.90_{\pm 0.07}$

Table 2: Summary of AUC ± standard error in the reader- and sentence-prediction settings.

Another surprising outcome is the lack of difference between the sentence- and reader-prediction settings. Given that the reader-prediction setting relies on $10\times$ to $30\times$ more data, we expected performance to be higher. The difference in performance may not be significant due to the relatively small size of the dataset and insufficient statistical power.

The finding that removing linguistic features did not significantly affect LSTM model's performance is less surprising. Arguably, the most crucial feature for dyslexia classification, word's orthographic transparency (Borleffs et al., 2017), was not available. Including a measure of word orthographic transparency might be a promising next step in improving model performance.

Most importantly, the proposed LSTM outperformed the SOTA model. We can confidently state that better performance is not a trivial result of hyperparameter tuning, as SOTA model was tuned for the same parameters within a similar search space. In the same vein, LSTM did not simply rely on the number of fixations made by children with

and without dyslexia for classification. Based on the outcomes of ablation experiments, we conclude that model performance increased due to the more detailed information about how the sequence of eye movements unfolds.

5 Conclusions

The model of automatic dyslexia detection proposed here has outperformed the SOTA model. Importantly, unlike most of the models proposed so far (Nilsson Benfatto et al., 2016; Haller et al., 2022; Asvestopoulou et al., 2019; Jothi Prabha and Bhargavi, 2022), the present LSTM was trained on an unbalanced dataset of eye movements of children of different ages, and might therefore be more robust and potentially more appropriate for the real-world applications.

Limitations

This decision to include information from participants who did not read all 30 sentences could potentially lead to data leakage: The model may learn that incomplete sessions are more likely to come from a child with dyslexia. We think that this is unlikely for two reasons: First, the proportions of incomplete sessions are not drastically different between the two groups. Second, this potential data leakage should only affect the reader-prediction setting (where the model expects to see 30 sentences), not the sentence-prediction setting (where the model expects to see one sentence). In the present case, there was no significant difference in performance between the reader prediction and the sentence prediction settings (see Section 3.2), so the reader-prediction setting is unlikely to have an unfair advantage.

Ethical considerations

Using demographic variables, such as age and gender, might lead to reproducing existing biases. For example, males are diagnosed with dyslexia more frequently, but at least part of the difference may be attributed to referral bias (Wadsworth et al., 1992). One way to control for bias is to withhold the potentially biasing feature. The ablation experiment that removed the demographic information performed on par with the full model. Therefore, we conclude that the model at least does not enhance the bias that might be present in the data set.

Funding statement

Anna Laurinavichyute was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – project ID 317633480 – SFB 1287. The contribution of Anastasiya Lopukhina was supported by the Basic Research Program at the National Research University Higher School of Economics (HSE University). The contribution of David R. Reich was partially funded by the Swiss National Science Foundation under grant IZCOZ0_220330 (EyeNLG).

References

Seoyoung Ahn, Conor Kelton, Aruna Balasubramanian, and Greg Zelinsky. 2020. Towards predicting reading comprehension from gaze behavior. In *ACM Symposium on Eye Tracking Research and Applications*, pages 1–5.

Anne B Arnett, Bruce F Pennington, Robin L Peterson, Erik G Willcutt, John C DeFries, and Richard K Olson. 2017. Explaining the sex difference in dyslexia. *Journal of Child Psychology and Psychiatry*, 58(6):719–727.

Thomais Asvestopoulou, Victoria Manousaki, Antonis Psistakis, Ioannis Smyrnakis, Vassilios Andreadakis, Ioannis M Aslanides, and Maria Papadopouli. 2019. Dyslexml: Screening tool for dyslexia using machine learning. *arXiv preprint arXiv:1903.06274*.

Hazel I Blythe. 2014. Developmental changes in eye movements and visual information encoding associated with learning to read. *Current directions in psychological science*, 23(3):201–207.

Elisabeth Borleffs, Ben AM Maassen, Heikki Lyytinen, and Frans Zwarts. 2017. Measuring orthographic transparency and morphological-syllabic complexity in alphabetic orthographies: a narrative review. *Reading and writing*, 30:1617–1638.

Jonathan Glazzard. 2010. The impact of dyslexia on pupils' self-esteem. *Support for learning*, 25(2):63–69.

Patrick Haller, Andreas Säuberli, Sarah Elisabeth Kiener, Jinger Pan, Ming Yan, and Lena Jäger. 2022. Eye-tracking based classification of Mandarin Chinese readers with and without dyslexia using neural sequence models. *arXiv preprint arXiv:2210.09819*.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Comput.*, 9(8):1735–1780.

International Dyslexia Association. 2024. Dyslexia basics. https://dyslexiaida.org/. Accessed: 2024-07-03.

A Jothi Prabha and R Bhargavi. 2022. Prediction of dyslexia from eye movements using machine learning. *IETE Journal of Research*, 68(2):814–823.

Slavica K Katusic, Robert C Colligan, William J Barbaresi, Daniel J Schaid, and Steven J Jacobsen. 2001. Incidence of reading disability in a population-based birth cohort, 1976–1982, rochester, minn. In *Mayo Clinic Proceedings*, volume 76, pages 1081–1092. Elsevier.

Reinhold Kliegl, Ellen Grabner, Martin Rolfs, and Ralf Engbert. 2004. Length, frequency, and predictability effects of words on eye movements in reading. *European journal of cognitive psychology*, 16(1-2):262–284.

AN Kornev and OA Ishimova. 2010. Metodika diagnostiki disleksii u detey [methods of diagnosis of dyslexia in children]. *St. Petersburg: Publishing house of the Polytechnic University*.

Antonina N Laposhina and Maria Yu Lebedeva. 2021. Textometr: An online tool for automated complexity level assessment of texts for russian language learners. *Russian Language Studies*, 19(3):331–345.

Mallorie Leinenger. 2019. Survival analyses reveal how early phonological processing affects eye movements during reading. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 45(7):1316.

Anastasiya Lopukhina, Nina Zdorova, Vladislava Staroverova, Nina Ladinskaya, Anastasiia Kaprielova, Sofya Goldina, Olga Vedenina, Ksenia Bartseva, and Olga Dragoy. 2022. Benchmark measures of eye movements during reading in russian children.

Sara V Milledge and Hazel I Blythe. 2019. The changing role of phonology in reading development. *Vision*, 3(2):23.

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.

Richard K Olson, Reinhold Kliegl, and Brian J Davidson. 1983. Dyslexic and normal readers' eye movements. *Journal of Experimental Psychology: Human Perception and Performance*, 9(5):816.

George Th Pavlidis. 1981. Do eye movements hold the key to dyslexia? *Neuropsychologia*, 19(1):57–64.

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.

Keith Rayner. 1985. Do faulty eye movements cause dyslexia? *Developmental Neuropsychology*, 1(1):3–15.

Keith Rayner, Alexander Pollatsek, and Katherine S Binder. 1998. Phonological codes and eye movements in reading. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 24(2):476.

Keith Rayner, Sara C Sereno, Mary F Lesch, and Alexander Pollatsek. 1995. Phonological codes are automatically activated during reading: Evidence from an eye movement priming paradigm. *Psychological Science*, 6(1):26–32.

Readability Test. Readability test for russian texts. Accessed: 2024-07-04.

David Robert 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*, pages 1–8.

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*, pages 1–8.

Russian National Corpus. Russian national corpus. Accessed: 2024-07-04.

Cory Shain. 2019. A large-scale study of the effects of word frequency and predictability in naturalistic reading. In *Proceedings of the 2019 conference of the north American chapter of the association for computational linguistics: Human language technologies, Volume 1 (Long and Short papers)*, pages 4086–4094.

Soroosh Shalileh, Dmitry Ignatov, Anastasiya Lopukhina, and Olga Dragoy. 2023. Identifying dyslexia in school pupils from eye movement and demographic data using artificial intelligence. *Plos one*, 18(11):e0292047.

Sally E Shaywitz, Bennett A Shaywitz, Kenneth R Pugh, Robert K Fulbright, R Todd Constable, W Einar Mencl, Donald P Shankweiler, Alvin M Liberman, Pawel Skudlarski, Jack M Fletcher, et al. 1998. Functional disruption in the organization of the brain for reading in dyslexia. *Proceedings of the National Academy of Sciences*, 95(5):2636–2641.

Joseph K Torgesen. 2000. Individual differences in response to early interventions in reading: The lingering problem of treatment resisters. *Learning disabilities research & practice*, 15(1):55–64.

Frank R Vellutino, Jack M Fletcher, Margaret J Snowling, and Donna M Scanlon. 2004. Specific reading disability (dyslexia): What have we learned in the past four decades? *Journal of child psychology and psychiatry*, 45(1):2–40.

Sally J Wadsworth, JC DeFries, Jim Stevenson, Jeffrey W Gilger, and BF Pennington. 1992. Gender ratios among reading-disabled children and their siblings as a function of parental impairment. *Journal of Child Psychology and Psychiatry*, 33(7):1229–1239.

A Model parameters

Model search space is summarized in Table 3.

Batch size	8, 16, 32, 64, 128
Learning rate	$15 \times \mathcal{U} \sim (1.0 \times 10^{-5}, 1.0 \times 10^{-1})$
LSTM hidden layer size	30, 40, 50, 60, 70

Table 3: Hyperparameter search space.

The optimal parameters can be found in Table 4.

Batch size	Learning rate	Hidden layer size
Reader prediction setting		
64	0.001	40
16	0.001	40
64	0.01	30
64	0.001	40
64	0.001	40
64	0.001	40
16	0.01	30
32	0.01	30
16	0.01	50
64	0.001	40
Sentence prediction setting		
8	7.07×10^{-5}	30
8	0.0003	50
128	4.21×10^{-5}	70
64	0.0025	50
8	7.07×10^{-5}	30
64	0.0025	50
128	0.0025	30
8	7.07×10^{-5}	30
32	5.34×10^{-5}	70
8	4.21×10^{-5}	50

Table 4: Resulting optimal parameters.