Reading Does Not Equal Reading: Comparing, Simulating and Exploiting Reading Behavior Across Populations

David R. Reich **Shuwen Deng**
Marina Björnsdóttir **Lena A. Jäger** **Nora Hollenstein**
University of Potsdam Royal Danish Academy
University of Zurich Kaunas University of Technology
{david.reich,deng}@uni-potsdam.de, mbjo@kglakademi.dk, jaeger@cl.uzh.ch, nora.hollenstein@ktu.lt

Abstract

Eye-tracking-while-reading corpora play a crucial role in the study of human language processing, and, more recently, have been leveraged for cognitively enhancing neural language models. A critical limitation of existing corpora is that they often lack diversity, comprising primarily native speakers. In this study, we expand the eye-tracking-while-reading dataset CopCo, which initially included only Danish L1 readers with and without dyslexia, by incorporating a new dataset of non-native readers with diverse L1 backgrounds. Thus, the extended CopCo corpus constitutes the first eye-tracking-while-reading dataset encompassing neurotypical L1 and L1 readers with dyslexia as well as non-native readers, all reading the same materials. We first provide extensive descriptive statistics of the extended CopCo corpus. Second, we investigate how different degrees of diversity of the training data affect a state-of-the-art generative model of eye movements in reading. Finally, we use this scanpath generation model for gaze-augmented language modeling and investigate the impact of diversity in the training data on the model's performance on a range of NLP downstream tasks. The code can be found here: https://github.com/norahollenstein/copco-processing.

Keywords: eye-tracking, scanpaths, Danish, dyslexia, second language speaker, bias, reading, language processing, machine learning, generative models, cognitively enhanced NLP

1. Introduction & Related Work

Some readers will decipher this paper as effortlessly as a shopping list, others will take a few milliseconds longer to mentally translate certain words into their native language, and others still will struggle to make sense of the words because they lack the relevant background knowledge, or because letters appear to them in jumbled order. The cognitive processes of different types of readers can be captured with eye-tracking recordings. Human eye movements in reading are characterized by an alternating sequence of *fixations*, where the gaze remains relatively still and visual input is obtained, and *saccades*, which are rapid movements between fixations during which visual input is suppressed. In the following, a *sequence of fixations* is also referred to as *scanpath*.

The Copenhagen Corpus of Eye-Tracking Recordings from Natural Reading (CopCo, Hollenstein et al., 2022; Björnsdóttir et al., 2023) consists of eye movement data from a diverse set of participants reading naturally occurring Danish texts at their own pace to enable researchers from various subfields of linguistics to study reading behavior, and to leverage the data for computational models. CopCo contains eye-tracking recordings from adult native speakers (L1) with and without dyslexia. In this work, we extend CopCo by contributing an additional dataset of adult non-native speakers of Danish. In the following, for simplicity, we'll refer to non-native readers of Danish as L2 although Danish is the L3 or L4 for some of the participants. While most psycholinguistic studies have been focusing on highly homogeneous populations by defining precise inclusion criteria (e.g., only native speakers, dyslexic speakers without any comorbidities, L2 speakers with a specific L1), we contribute a naturalistic corpus read by a more diverse population to reflect the linguistic heterogeneity of contemporary Danish society. Our work adds to the collection of eye-tracking datasets where both L1 and L2 readers read from the same set of texts (Cop et al., 2017; Sui et al., 2022; Kuperman et al., 2023; Berzak et al., 2022). While in the vast majority of existing datasets, the L2 part consists of English L2 data recorded from speakers of the same (the respective local) L1, our dataset constitutes the first *non-English L2* eye-tracking-while-reading corpus from speakers of a *diverse set of native languages*. Gaze data scarcity is a persistent challenge that researchers have addressed by resorting to computational cognitive models that simulate eye movements in reading (Reichle et al., 2003; Engbert et al., 2005), or by developing non-explanatory machine learning models that are optimized to generate human-like eye movement patterns on a given text (Deng et al., 2023b; Bolliger et al., 2023). While these approaches have shown promising results for English, to date, none of these methods have been evaluated on a low-resource language like Danish,

LREC-COLING 2024, pages 13586–13594

nor on data from diverse populations.

A diverse reading dataset that approximates the contemporary socio-linguistic situation is not only useful for fundamental research of reading but might also be more beneficial for gaze-augmented natural language processing (NLP) since recent research in computational linguistics revealed that L2 reading behavior is better aligned with the representations learned by computational language models (Gonzalez-Garduno and Søgaard, 2018; Brandl and Hollenstein, 2022; Schneider et al., 2023). The ability to accurately model gaze features is crucial to advance our understanding of language processing (Hollenstein et al., 2021). Not only can it reveal the workings of the underlying cognitive processes of language understanding (Reichle et al., 2003; Engbert et al., 2005), but the performance of computational language models can also be improved if their inductive bias is adjusted using human cognitive signals (Hollenstein and Zhang, 2019; Sood et al., 2020; Deng et al., 2023a). Although it has been shown that augmenting language models with eye-tracking data is most beneficial in low-resource scenarios (Deng et al., 2023a), it has mostly been explored on English data (Barrett et al., 2018; McGuire and Tomuro, 2021; Mathias et al., 2020), but not on a lower-resource language such as Danish.

We first provide comprehensive descriptive statistics to compare eye movement patterns between native speakers, second-language speakers, and readers with dyslexia when reading Danish texts. We then investigate whether eye-tracking data from readers with diverse backgrounds is more suitable for training a state-of-the-art neural network model (Deng et al., 2023b) to generate ecologically valid human scanpaths. Subsequently, we examine how the training population of this scanpath generation model impacts the performance of a language model that is augmented with these synthetic data on a representative set of downstream NLU tasks. More precisely, we use a deep neural text-conditioned dual-sequence autoregressive model, Eyettention (Deng et al., 2023b), to synthesize gaze data of the three populations. We then apply the synthesized data to augment a Danish Foundation Model (DFM) that is tested on various Danish natural language understanding tasks.

2. The CopCo Corpus

In the first two iterations, the authors of CopCo collected data from typically developing L1 readers (Hollenstein et al., 2022) and L1 readers with dyslexia (Björnsdóttir et al., 2023).[1] Here, we introduce an extension of the CopCo dataset containing eye movements from non-native Danish readers. For simplicity, we'll refer to non-native readers of Danish as L2 although Danish is the L3 or L4 for some of the participants.

	L1 reader	L2 reader	Dyslexia
Demographic properties of the readers			
Number of readers	25	13	19
Self-reported gender	19 F / 6 M	9 F / 4 M	12 F / 7 M
Age	$30.20_{9.20}$	$29.00_{3.84}$	$36.05_{14.38}$
Descriptive statistics			
Number of texts read	$4.84_{1.71}$	$6.00_{3.55}$	$2.89_{1.02}$
Comprehension acc.	$0.84_{0.11}$	$0.76_{0.14}$	$0.79_{0.16}$
Reading time	$17.66_{3.79}$	$22.98_{6.25}$	$35.61_{17.96}$
#Fixations per para.	$4.34_{0.084}$	$5.25_{0.140}$	$6.54_{0.175}$
#Fixations per word	$0.92_{0.006}$	$1.31_{0.009}$	$1.57_{0.013}$
Progression length	$2.75_{0.024}$	$2.38_{0.037}$	$2.35_{0.025}$
Regression length	$6.26_{0.061}$	$6.18_{0.085}$	$4.59_{0.059}$
Regression ratio	$0.24_{0.002}$	$0.21_{0.003}$	$0.24_{0.003}$

Table 1: Overview of the extended CopCo dataset. The reading measures are defined in the appendix. Accuracy and paragraph are denoted by *acc.* and *para.*, respectively. Subscripts indicate the standard deviation for age, number of texts read, comprehension accuracy, and reading time, whereas the standard error pertains to all other reported statistics.

Data Collection Using the same materials and an identical experimental procedure and hardware set-up (EyeLink 1000 Plus) as for the collection of the L1 data (Hollenstein et al., 2022), we contribute L2 recordings to the CopCo dataset. We record data from 13 new participants (see Table 1 for details). The reading materials include 46 transcribed and proofread Danish speeches, accessed from the Danske Taler archive[2] and 12 articles from the Danish Wikipedia[3]. Each participant read a varying number of texts depending on their personal reading speed (see Hollenstein et al., 2022 for details).

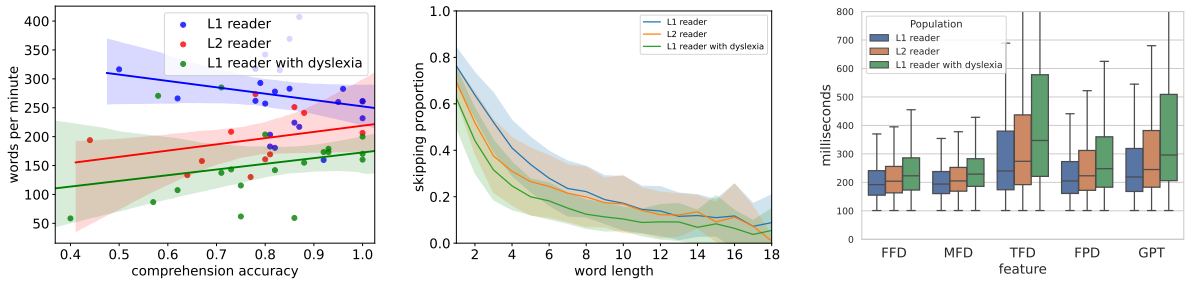
3. Comparing Reading Patterns Across Different Reader Groups

To the best of our knowledge, CopCo is the first dataset containing neurotypical L1 and L2 readers, as well as L1 readers with dyslexia, promoting inclusiveness across diverse reading profiles. Furthermore, the reader's age range distinguishes CopCo from other widely-used eye-tracking-while-reading datasets (Cop et al., 2017; Sui et al., 2022; Kuperman et al., 2023; Jakobi et al., 2024): Participants span around 30 years of age (Min: 20, Max: 64, Mean: 32.34, Median: 29) as opposed to the typical average age of 20 to 25 years. The level of

[1] https://osf.io/ud8s5/

[2] https://dansketaler.dk

[3] https://da.wikipedia.org/wiki/Wikipedia:Ugens_artikel



(a) (b) (c)

Figure 1: (a) Comprehension accuracy vs. words per minute read, (b) Skipping proportion with respect to word length (# characters), and (c) First-fixation duration (FFD), mean-fixation duration (MFD), total-fixation duration (TFD), first-pass duration (FPD), and go-past time (GPT) for each group of readers.

reading comprehension, operationalized in terms of response accuracy to comprehension questions randomly presented after 20% of all paragraphs, is very similar for each of the three groups, see Table 1. Figure 1a shows that in typically developing L1 readers, comprehension accuracy tends to be higher when reading speed is lower, while the opposite holds for the other two groups.

We present the shared characteristics and distinguishing features of the reading patterns observed in the three groups. First, we explore how word length affects skipping rates (i.e., the proportion of trials in which a word is not fixated during first-pass reading). Figure 1b illustrates the impact of word length on skipping rates for each of the groups. While the differences are not statistically significant, numerically the skipping rates for L1 readers with dyslexia are lower than those for typical L1 readers. The skipping rates of L2 readers lie between typical L1 readers and L1 readers with dyslexia. These trends align with findings from prior psycholinguistic research (Hawelka et al., 2010; Kuperman et al., 2023).

Figure 1c illustrates five commonly reported duration-based eye movement measures: first-fixation duration (FFD), mean-fixation duration (MFD), total-fixation duration (TFD), first-pass duration (FPD), and go-past time (GPT). Definitions can be found in the appendix. We observe that, across measures, the average values from typical L1 readers are lower than the ones from L2 readers, which, in turn, are lower than the ones from L1 readers with dyslexia.

For the count-based fixation measures presented in Table 1, we observe the following significant differences between groups: Typical L1 readers display the fewest and L1 readers with dyslexia the highest number of fixations per paragraph (including fixations between words). On a more fine-grained level, we extend the count-based analysis to fixations per word, where we see the same pattern as for the paragraph level. Finally, we analyze three saccadic measures: progressive saccade length (length of forward saccades measured in terms of characters); regressive saccade length (length of backward saccades in terms of characters); and regression rate (regression to total saccade count-ratio). For both progressive and regressive saccades, the same pattern is observed: Typical L1 readers produce longer saccades than the other groups; L1 readers with dyslexia display slightly shorter saccades compared to L2 readers. The regression ratio is somewhat higher for L1 readers in comparison to L2 readers.

4. Synthesizing Eye-Tracking Data

Despite the availability of a few non-English datasets like CopCo (Hollenstein et al., 2022), PoTeC (Jakobi et al., 2024), or the multilingual MECO data (Siegelman et al., 2022), generally eye-tracking-while-reading data in languages other than English remains very limited. One approach to overcome data scarcity is the simulation of human-like synthetic eye gaze data using a generative model. Notably, recent advancements in scanpath prediction (Deng et al., 2023b; Bolliger et al., 2023) are driven by the integration of contextualized embeddings from BERT-like language models (Devlin et al., 2019). However, the existing work in machine learning-based synthetic scanpath generation predominantly focuses on L1 readers, often neglecting diverse populations.

In our study, we aim to explore the extent to which models trained on L1 readers can generalize to L2 readers and to L1 readers diagnosed with dyslexia. We investigate whether the bias introduced by only recording L1 readers is detrimental to scanpath generation for readers with a different native language or non-typical readers, and examine whether the inclusion of L2 readers and L1 readers with dyslexia in the training data can enhance the overall quality of the generated

Model	Training data			NLL ↓ for test groups		
	L1	L2	Dys	L1	L2	Dys
Uniform				7.90	7.90	7.90
Et-Da	✓	×	×	$2.90_{0.09}$	$\mathit{2.67}_{\mathit{0.12}}$	$\mathit{2.63}_{\mathit{0.04}}$
Et-Da	✓	✓	×	$2.87_{0.09}$	$2.52_{0.12}$	$2.57_{0.05}$
Et-Da	✓	×	✓	$2.99_{0.07}$	$2.59_{0.14}$	$2.61_{0.04}$
Et-Da	×	✓	✓	$\mathit{3.00}_{\mathit{0.08}}$	$2.54_{0.15}$	$2.57_{0.06}$
Et-Da	✓	✓	✓	$\mathbf{2.73}_{\mathbf{0.08}}$	$\mathbf{2.44}_{\mathbf{0.12}}$	$\mathbf{2.48}_{\mathbf{0.05}}$

Table 2: Models are trained on 20 and tested on three randomly sampled users from each group. Both sets are resampled six times. Best results are indicated in **bold**, and worst results in *italic*. *Et-Da* refers to the Eyettention-Da model, and *Dys* to the group of readers with dyslexia. Subscripts denote the standard error.

scanpaths.

Eyettention-Da We train a state-of-the-art model for generating synthetic gaze data from each of the three populations and investigate the impact of biased training data. We train Eyettention (Deng et al., 2023b), a deep neural dual-sequence model that autoregressively predicts a fixation sequence, each fixation represented as a word index, for a given text. It consists of two encoders, one for the text and one for the fixation sequence, and a decoder that generates a probability distribution over next-fixation locations. To adapt the model for Danish texts (henceforth referred to as *Eyettention-Da*), we choose the state-of-the-art DFM encoder as the text encoder backbone. This choice is motivated by its recent success on the Danish part of the ScandEval benchmark (Nielsen, 2023). The DFM encoder is a fine-tuned version of the NB-BERT-large model (Kummervold et al., 2021), a BERT uncased architecture, which was pre-trained on a collection of Norwegian datasets using masked language modeling.

Evaluation Procedure & Results To evaluate the scenarios introduced above, we split the dataset such that the test set only contains eye-tracking recordings from readers who are held out from training. Given the limited number of readers, we conduct six random resamplings for both the training and test readers. All models are trained on 20 readers and evaluated on three readers from each reader group. For reference, we include a baseline model that samples the next fixation from a uniform distribution. We measure model performance in terms of negative log-likelihood (NLL). A lower NLL indicates better predictive performance.

We observe that Eyettention-Da outperforms the baseline model in all settings, independently from the training population(s) (see Table 2). However, it is worth noting that when Eyettention-Da is trained only on typical L1 readers, it performs poorer for both L2 readers and L1 readers diagnosed with dyslexia, and worse than any other combination of training populations. By contrast, the best performance on each of the test groups is observed when Eyettention-Da is trained on readers from each of the three groups.

5. Gaze-Augmented Language Modeling for ScandEval Tasks

We investigate the potential enhancement of scanpath-augmented language models when utilizing eye-tracking data from readers with diverse backgrounds. Augmenting high-performing models with gaze data for downstream NLP tasks typically requires a substantial volume of task-specific gaze recordings for both training and testing. However, obtaining such data has been challenging due to the resource-intensive endeavor of collecting gaze data. Utilizing (potentially large amounts of) synthetic data offers a promising solution, enabling the training of more robust models for various tasks and offering human-like gaze data even at application time (Deng et al., 2023a). To this end, we use synthesized gaze data from different populations to augment LMs, and evaluate the performance of these gaze-augmented language models on the Danish NLP tasks from the ScandEval benchmark (Nielsen, 2023), encompassing a comprehensive collection of four diverse natural language understanding (NLU) tasks in the Danish Language.

Gaze-Augmented Language Model In their seminal work, Deng et al. (2023a) employed the PLM-AS method proposed by Yang and Hollenstein (2023) to augment a language model's input text with synthetic gaze data. This approach alleviates the need for real human scanpaths, replacing them with simulated fixation sequences at inference time. Building upon this research, we take the method proposed by Deng et al. (2023a) as a starting point and replace the English scanpath generation model Eyettention with Eyettention-Da, and the backbone BERT model with the DFM encoder. We explore the impact that training Eyettention-Da on diverse reader populations will have on the performance of a PLM-AS gaze-augmented language model on various NLU tasks.

Evaluation Procedure & Results The Danish tasks from the ScandEval (Nielsen, 2023) benchmark[4] focus on low-resource challenges, with 1024, 256, and 2048 instances for each task's training, validation, and test set, respectively. These

[4] https://scandeval.github.io/

Table 3: Results on the ScandEval benchmark. PLM-AS with Eyettention-Da as scanpath generator is compared to the Danish Foundation Model (DFM), the state-of-the-art model for the Danish part of the ScandEval benchmark. Eyettention-Da (*Et-Da*) is pre-trained on 20 readers sampled from the respective group shown in the table (L1, L2, Dyslexia (*Dys*)). The evaluation protocol follows the ScandEval benchmark. Subscripts denote the standard error.

Model	Et-Da training data			ScaLA-da	Angry Tweets	DaNE	ScandiQA-da	DA
	L1	L2	Dys	MCC/F1	MCC/F1	F1/F1-MISC	EM/F1	Macro
DFM	×	×	×	$76.11_{1.17}$/$87.41_{0.67}$	$51.42_{2.30}$/$67.07_{1.97}$	$82.69_{\pm 0.85}$/$85.08_{0.77}$	$\mathbf{54.45_{1.65}}$/$\mathbf{59.52_{1.45}}$	$66.17_{1.49}$
PLM-AS+Et-Da	✓	×	×	$79.54_{0.63}$/$89.50_{0.38}$	$\mathbf{56.73_{1.09}}$/$\mathbf{71.33_{0.75}}$	$85.17_{1.07}$/$87.19_{1.23}$	$53.75_{1.94}$/$58.85_{0.92}$	$68.80_{2.57}$
PLM-AS+Et-Da	✓	✓	×	$61.34_{13.80}$/$77.00_{9.81}$	$55.60_{0.62}$/$70.44_{0.32}$	$79.37_{2.07}$/$77.29_{2.57}$	$48.91_{3.94}$/$53.65_{2.13}$	$61.31_{1.32}$
PLM-AS+Et-Da	✓	×	✓	$\mathbf{80.20_{1.30}}$/$\mathbf{89.72_{0.74}}$	$56.70_{0.71}$/$70.78_{0.32}$	$\mathbf{85.23_{0.98}}$/$\mathbf{87.37_{1.21}}$	$54.03_{1.96}$/$58.80_{0.91}$	$\mathbf{69.04_{1.69}}$
PLM-AS+Et-Da	✓	✓	✓	$62.67_{14.08}$/$77.45_{9.89}$	$55.26_{1.09}$/$70.15_{0.89}$	$81.28_{1.45}$/$83.37_{1.91}$	$49.03_{2.07}$/$52.57_{1.91}$	$62.06_{1.21}$

tasks cover assessments of linguistic acceptability (ScaLA-da), sentiment analysis (Angry Tweets), Named Entity Recognition (DaNE), and question answering (ScandiQA-da). We adhere to the ScandEval evaluation protocol. We report the results of the DFM encoder, the best-performing model on the Danish section of the ScandEval, as our baseline for assessment.

The results are summarized in Table 3. We find that on three of the four tasks, PLM-AS outperforms the current state-of-the-art DFM encoder. Only on one of the six metrics, the PLM-AS model, paired with the Eyettention-Da model trained exclusively on neurotypical L1 readers, achieves the best performance. However, for the remaining 5 metrics, a model trained on both populations of L1 readers establishes a new state-of-the-art. When considering the macro-average, which serves as the benchmark for the entire Danish portion of ScandEval, the model trained on L1 readers with and without dyslexia performs best. Finally, the addition of L2 readers in the pre-training of the scanpath generation model seems to induce noise for both ScandiQA-da and Scala-da, which leads to a subpar performance on both of these tasks as well as on the overall benchmark.

6. Discussion & Conclusion

We extended the CopCo corpus with an additional dataset with recordings from L2 readers of various native languages. We compare the data between populations contained in the extended CopCo dataset and subsequently train Eyettention-Da to synthesize population-specific scanpaths. The main finding is that training on a more diverse population yields more accurate scanpaths for all populations. This result suggests that population-specific differences in reading behavior lead to overfitting and thus should not be aggregated out of the data but embracing the diversity in the data results in more realistic, human-like scanpaths. Further, when evaluating the Eyettention-Da model as a backbone for a scanpath-augmented model, we achieve state-of-the-art results on the Danish part of the ScandEval benchmark. This shows that higher diversity in the gaze data not only improves the scanpath prediction but also enhances the performance of a gaze-augmented language model on downstream NLP tasks. We conclude that the inclusion of more diverse linguistic profiles is not only a desideratum for achieving inclusiveness in language modeling but can also have a regularizing effect which even increases the models' performance on a neurotypical population of native speakers.

7. Ethics Statement & Limitations

Working with human data necessitates ethical consideration. The data collection was approved by the relevant ethics committee. All data is anonymized. However, privacy risks associated with human gaze data collection, sharing, and processing are significant: recent research has revealed that individual identities can be potentially extracted from gaze data (Jäger et al., 2020; Lohr and Komogortsev, 2022). Other personal information such as gender and ethnicity may also become possible to extract in the future, posing a risk of personal information leakage. The proposed use of synthetic data during deployment significantly mitigates this privacy risk since synthetic eye movements do not (directly) reveal the reader's identity. Synthetic gaze data can also reduce the need for large-scale human experiments, though some real gaze data remains essential for training generative models. While generating synthetic gaze data helps mitigate the issue of gaze data scarcity, adopting this approach raises ethical concerns, as it opens the possibility of training models that could be used for various tasks, including those with malicious intent.

The ecological validity of psycholinguistic findings has been questioned, primarily due to the field focusing on English-speaking populations. Along the same lines, in eye-tracking based technological applications, the emphasis on native speakers of English has led to algorithmic bias, as demonstrated by Prasse et al. (2022). Our collection of

L2 Danish data, i.e., non-native readers of a low-resource language, helps to overcome limitations related to the availability and representativeness of reading data, fostering the development of more equitable and unbiased models.

One limitation of our work is that the scanpath generation model, Eyettention-Da, was (pre-)trained on eye-tracking data recorded from reading full paragraphs, not single sentences as in the original paper (Deng et al., 2023b), and the hyperparameters were not adjusted. Our experiments revealed that scanpath augmentation had a negative impact on the model's performance in question answering (ScandiQA-da). Thus, future work might explore pre-training the scanpath generation model on an eye-tracking corpus recorded from an information-seeking reading, rather than from natural reading. Finally, due to the nature of the recorded eye-tracking data, we only evaluated the Danish part of ScandEval (Nielsen, 2023), crucially missing the remaining Nordic languages.

8. Acknowledgements

The data for this work was collected while MB and NH were affiliated with the University of Copenhagen. This work was partially funded by the German Federal Ministry of Education and Research under grant 01|S20043 and was partially supported by the MultiplEYE COST Action (CA21131).

9. Bibliographical References

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

Yevgeni Berzak, Chie Nakamura, Amelia Smith, Emily Weng, Boris Katz, Suzanne Flynn, and Roger Levy. 2022. CELER: A 365-participant corpus of eye movements in L1 and L2 English reading. *Open Mind*, 6:41–50.

Marina Björnsdóttir, Nora Hollenstein, and Maria Barrett. 2023. Dyslexia prediction from natural reading of Danish texts. In *Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 60–70, Tórshavn, Faroe Islands.

Lena Bolliger, David Reich, Patrick Haller, Deborah Jakobi, Paul Prasse, and Lena Jäger. 2023. ScanDL: A diffusion model for generating synthetic scanpaths on texts. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 15513–15538, Singapore.

Stephanie Brandl and Nora Hollenstein. 2022. Every word counts: A multilingual analysis of individual human alignment with model attention. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing*, pages 72–77, Online only.

Uschi Cop, Nicolas Dirix, Denis Drieghe, and Wouter Duyck. 2017. Presenting GECO: An eyetracking corpus of monolingual and bilingual sentence reading. *Behavior Research Methods*, 49:602–615.

Shuwen Deng, Paul Prasse, David Reich, Tobias Scheffer, and Lena Jäger. 2023a. Pretrained language models augmented with synthetic scanpaths for natural language understanding. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6500–6507, Singapore.

Shuwen Deng, David Reich, Paul Prasse, Patrick Haller, Tobias Scheffer, and Lena Jäger. 2023b. Eyettention: An attention-based dual-sequence model for predicting human scanpaths during reading. *Proceedings of the ACM on Human-Computer Interaction*, 7(ETRA):1–24.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 4171–4186, Minneapolis, Minnesota.

Ralf Engbert, Antje Nuthmann, Eike Richter, and Reinhold Kliegl. 2005. SWIFT: A dynamical model of saccade generation during reading. *Psychological Review*, 112(4):777.

Ana Gonzalez-Garduno and Anders Søgaard. 2018. Learning to predict readability using eye-movement data from natives and learners. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, pages 5118–5124.

Stefan Hawelka, Benjamin Gagl, and Heinz Wimmer. 2010. A dual-route perspective on eye movements of dyslexic readers. *Cognition*, 115(3):367–379.

Nora Hollenstein, Maria Barrett, and Marina Björnsdóttir. 2022. The Copenhagen corpus of eye tracking recordings from natural reading of Danish texts. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference (LREC)*, pages 1712–1720, Marseille, France.

Nora Hollenstein, Emmanuele Chersoni, Cassandra L. Jacobs, Yohei Oseki, Laurent Prévot, and Enrico Santus. 2021. CMCL 2021 shared task on eye-tracking prediction. In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, pages 72–78, Online.

Nora Hollenstein and Ce Zhang. 2019. Entity recognition at first sight: Improving NER with eye movement information. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACT-HLT)*, pages 1–10, Minneapolis, Minnesota.

Lena Jäger, Silvia Makowski, Paul Prasse, Liehr Sascha, Maximilian Seidler, and Tobias Scheffer. 2020. Deep Eyedentification: Biometric identification using micro-movements of the eye. In *Machine Learning and Knowledge Discovery in Databases (ECML-PKDD)*, volume 11907 of *Lecture Notes in Computer Science*, pages 299–314, Cham, Switzerland.

Deborah Jakobi, Thomas Kern, David Reich, Patrick Haller, and Lena Jäger. 2024. PoTeC: A German naturalistic eye-tracking-while-reading corpus.

Per Kummervold, Javier De la Rosa, Freddy Wetjen, and Svein Arne Brygfjeld. 2021. Operationalizing a national digital library: The case for a Norwegian transformer model. In *Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 20–29, Reykjavik, Iceland (Online).

Victor Kuperman, Noam Siegelman, Sascha Schroeder, Cengiz Acartürk, Svetlana Alexeeva, Simona Amenta, Raymond Bertram, Rolando Bonandrini, Marc Brysbaert, Daria Chernova, and et al. 2023. Text reading in English as a second language: Evidence from the multilingual eye-movements corpus. *Studies in Second Language Acquisition*, 45(1):3–37.

Dillon Lohr and Oleg Komogortsev. 2022. Eye Know You Too: Toward viable end-to-end eye movement biometrics for user authentication. *IEEE Transactions on Information Forensics and Security*, 17:3151–3164.

Sandeep Mathias, Rudra Murthy, Diptesh Kanojia, Abhijit Mishra, and Pushpak Bhattacharyya. 2020. Happy are those who grade without seeing: A multi-task learning approach to grade essays using gaze behaviour. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pages 858–872, Suzhou, China.

Erik McGuire and Noriko Tomuro. 2021. Relation classification with cognitive attention supervision. In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, pages 222–232, Online.

Dan Nielsen. 2023. ScandEval: A benchmark for Scandinavian natural language processing. In *Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 185–201, Tórshavn, Faroe Islands.

Paul Prasse, David Reich, Silvia Makowski, Lena Jäger, and Tobias Scheffer. 2022. Fairness in oculomotoric biometric identification. In *Proceedings of the 2022 ACM Symposium on Eye-Tracking Research and Applications*, ETRA '22, Seattle, WA.

Erik Reichle, Keith Rayner, and Pollatsek Alexander. 2003. The E-Z reader model of eye-movement control in reading: Comparisons to other models. *The Behavioral and Brain Sciences*, 26:445–526.

Gerold Schneider, Beatrix Busse, Nina Dumrukcic, and Ingo Kleiber. 2023. Do non-native speakers

read differently? Predicting reading times with surprisal and language models of native and non-native eye tracking data. *Language and Linguistics in a Complex World*, 32:153–187.

Noam Siegelman, Sascha Schroeder, Cengiz Acartürk, Hee-Don Ahn, Svetlana Alexeeva, Simona Amenta, Raymond Bertram, Rolando Bonandrini, Marc Brysbaert, Daria Chernova, et al. 2022. Expanding horizons of cross-linguistic research on reading: The multilingual eye-movement corpus (MECO). *Behavior Research Methods*, 54(6):2843–2863.

Ekta Sood, Simon Tannert, Philipp Mueller, and Andreas Bulling. 2020. Improving natural language processing tasks with human gaze-guided neural attention. In *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pages 6327–6341.

Longjiao Sui, Nicolas Dirix, Evy Woumans, and Wouter Duyck. 2022. GECO-CN: Ghent eye-tracking corpus of sentence reading for Chinese-English bilinguals. *Behavior Research Methods*, 55:2743—-2763.

Duo Yang and Nora Hollenstein. 2023. PLM-AS: Pre-trained language models augmented with scanpaths for sentiment classification. In *Proceedings of the Northern Lights Deep Learning Workshop*, Tromsø, Norway.

A. Additional information about L2 cohort

The L2 readers' native languages encompass Spanish, Korean, Hungarian, Portuguese, Icelandic (2x), Czech (3x), German, Hebrew, English, and Basque. Their respective years of learning Danish range from 2.5 to 21 years.

B. Reading measure definitions

The *number of fixations per paragraph* is the mean number of fixations per paragraph normalized with respect to the number of words in the paragraph. The *number of fixations per word* is the mean number of fixations on each word. The *progression length* and *regression length* are the mean saccadic length (in degrees of visual angle) either to the bottom/right or top/left, respectively. The *regression ratio* is the ratio of regression given all saccades. The *first fixation duration* is the duration of the first fixation (in milliseconds) on a word, this might be zero if the word was first skipped. The *mean fixation duration* is the sum of fixation durations (in milliseconds) during the first encounter of the word divided by the number of fixations on the word. The mean *total fixation duration* is the sum of all fixation durations of any fixation on a given word divided by the number of fixations. The *first pass duration* is the summed duration (in milliseconds) of all fixations on the current word prior to progressing out of the current word (to the left or right). *Go-past time* is the sum duration (in milliseconds) of all fixations prior to progressing to the right of the current word, including regressions to previous words that originated from the current word.

C. Additional scanpath generation experiments

Using the same resampling methodology as in the experiments on scanpath generation in the main paper, we investigated models trained exclusively on a subset of 10 readers, accounting for scenarios where models are trained solely on L2 readers or L1 readers with dyslexia, see Table 4. The findings mirror the earlier results, with Eyettention trained on all three reader groups consistently outperforming all other models.

Table 4: Models are trained on 10 and evaluated on 3 randomly sampled users. Both training and test populations are resampled 6 times. **Bold** values indicate the best results, while *italic* values indicate the worst. ✓indicates that the reader group was included in the training, while × indicates the opposite.

	Training data			NLL ↓ for test groups		
Model	L1	L2	Dys	L1	L2	Dys
Eyett-da	✓	×	×	$3.00_{\pm 0.08}$	$2.65_{\pm 0.12}$	*$2.69_{\pm 0.03}$*
Eyett-da	×	✓	×	$3.03_{\pm 0.08}$	**$2.60_{\pm 0.14}$**	$2.67_{\pm 0.06}$
Eyett-da	×	×	✓	*$3.12_{\pm 0.08}$*	$2.64_{\pm 0.15}$	$2.65_{\pm 0.05}$
Eyett-da	✓	✓	×	$3.01_{\pm 0.09}$	$2.62_{\pm 0.11}$	$2.68_{\pm 0.05}$
Eyett-da	✓	×	✓	$3.04_{\pm 0.09}$	*$2.65_{\pm 0.13}$*	$2.67_{\pm 0.04}$
Eyett-da	×	✓	✓	$3.10_{\pm 0.08}$	$2.61_{\pm 0.12}$	$2.68_{\pm 0.06}$
Eyett-da	✓	✓	✓	**$2.94_{\pm 0.09}$**	**$2.60_{\pm 0.12}$**	**$2.65_{\pm 0.05}$**

D. ScandEval tasks

The ScandEval benchmark, as introduced by Nielsen (2023), encompasses a suite of natural language understanding (NLU) tasks for Scandinavian languages. It spans Danish, Faroese, Icelandic, Norwegian, and Swedish. However, our analysis focuses exclusively on this benchmark's Danish component. This component integrates four distinct tasks: linguistic acceptability (ScaLA-Da), question-answering (ScandiQA-Da), sentiment classification (Angry Tweets), and named entity recognition (DaNE), each designed for different facets of language understanding. ScaLA-Da is designed to evaluate the acceptability of Danish sentences, employing the Matthews Correla-

tion Coefficient (MCC) and the macro F1 score as its evaluation metrics. ScandiQA-Da introduces a classic question-answering format, requiring models to identify and tag the correct answer within a given context. The performance on this task is quantified through the Exact Match metric and the F1 score. Angry Tweets presents sentiment classification, where the objective is to predict the sentiment conveyed in a specific tweet accurately. The evaluation metrics are the MCC and the macro F1 score. Lastly, DaNE focuses on Danish named entity recognition, tasking models with the identification and classification of named entities within sentences. This is assessed using the micro F1 score, both with and without consideration of miscellaneous tags.