Patterns of Text Readability in Human and Predicted Eye Movements

Nora Hollenstein
University of Copenhagen
nora.hollenstein@hum.ku.dk

Itziar Gonzalez-Dios
HiTZ Center - IXA (UPV/EHU)
itziar.gonzalezd@ehu.eus

Lisa Beinborn
CLTL Lab, VU Amsterdam
l.beinborn@vu.nl

Lena Jäger
University of Zurich; University of Potsdam
jaeger@cl.uzh.ch

Abstract

It has been shown that multilingual transformer models are able to predict human reading behavior when fine-tuned on small amounts of eye tracking data. As the cumulated prediction results do not provide insights into the linguistic cues that the model acquires to predict reading behavior, we conduct a deeper analysis of the predictions from the perspective of readability. We try to disentangle the three-fold relationship between human eye movements, the capability of language models to predict these eye movement patterns, and sentence-level readability measures for English. We compare a range of model configurations to multiple baselines. We show that the models exhibit difficulties with function words and that pre-training only provides limited advantages for linguistic generalization.

1 Introduction

Eye movement data of reading provides rich insights into cognitive processes of language understanding. The signal can be used to modulate the inductive bias of machine learning models towards more cognitively plausible processing which can increase model performance (Mathias et al., 2020; Hollenstein et al., 2019). It has been shown that large multilingual pre-trained language models are able to accurately predict eye tracking patterns when fine-tuned on small amounts of eye tracking data (Hollenstein et al., 2021; Takmaz, 2022; Salicchi et al., 2022).

Generally, transformer-based language models seem to be better at predicting cognitive signals of human language comprehension (e.g., self-paced reading times, eye movements, or brain activity) than language models based on other architectures (Merkx and Frank, 2021; Schrimpf et al., 2020). However, as *prediction is not explanation* (Demberg and Keller, 2019; Hale et al., 2022), we aim to dissect the predicted reading patterns and analyze them in more detail to gain clearer insights into the underlying representation of processing complexity. Eye tracking data can be very informative to evaluate sentence comprehension strategies, however, the interdependencies between the eye tracking measures need to be taken into account (Vasishth et al., 2013). We propose to use the relation between eye movements in reading and text readability in terms of linguistic complexity to better understand procedural patterns of English sentence comprehension in language models. We provide interpretable insights into the prediction errors to investigate the following two questions:

(1) What is the impact of pre-training on the performance of language models predicting human eye movements?

(2) Is the relationship between human reading patterns and English text readability preserved in the reading patterns predicted by the investigated language models?

We focus on multilingual pre-trained language models (mBERT and XLM), fine-tuned on a range of eye tracking features from reading in multiple languages (English, German, Dutch, and Russian). We build upon the approach by Hollenstein et al. (2021) and provide strong baselines and a series of model configurations to answer the first question in Section 2. Subsequently, we address the second question in Section 3, by performing an extensive readability analysis based on various aspects of English text complexity. We propose to evaluate the predicted gaze features by analyzing whether their correlation with a range of readability measures is similar to the correlation observed in human eye movement data.[1] These two contributions allow us to better interpret the ability of language models to predict human reading behaviour.

[1] Our code is available here: https://github.com/norahollenstein/readability-patterns

Proceedings of the Workshop on Cognitive Aspects of the Lexicon, pages 1–15
November 20, 2022. ©2022 Association for Computational Linguistics
https://doi.org/10.26615/978-954-452-056-4_001

Models	EN	NL	DE	RU	ALL
RANDOM BL	78.66 (0.06)	84.30 (0.11)	74.11 (4.6)	65.83 (2.55)	86.15 (0.43)
MEAN BL	89.94	90.15	84.98	85.35	92.54
mBERT					
$M_{\star}$	90.95 (0.11)	90.51 (0.31)	75.68 (3.99)	70.64 (2.38)	92.93 (0.13)
$M_{\heartsuit}$	93.73 (0.08)	91.91 (0.23)	77.41 (3.65)	77.30 (4.17)	94.68 (0.05)
$M_{\lightning}$	93.30 (0.03)	91.60 (0.36)	77.85 (2.85)	77.38 (1.85)	94.35 (0.13)
XLM-100					
$M_{\star}$	92.94 (0.05)	91.80 (0.40)	77.31 (2.75)	76.54 (1.92)	94.19 (0.10)
$M_{\heartsuit}$	**93.92** (0.07)	92.26 (0.33)	**86.38** (0.27)	**94.65** (0.88)	**94.89** (0.12)
$M_{\lightning}$	**93.92** (0.16)	**92.32** (0.36)	86.04 (0.28)	94.62 (0.84)	94.15 (1.20)

Table 1: Prediction accuracy aggregated across all eight eye tracking features (with standard deviation across three runs in parentheses). Fine-tuned models: last layer ($M_{\star}$), all layers ($M_{\heartsuit}$), all layers without pre-training ($M_{\lightning}$).

2 Multilingual Prediction of Eye Movements in Reading

Hollenstein et al. (2021) showed that language models can predict a range of eye tracking features in multiple languages. The prediction setup has been made widely available as a shared task to facilitate comparisons between models and the analysis of their inner workings (Hollenstein et al., 2022). In this work, we use a similar setup which we summarize below before we present the results of the eye tracking prediction.

2.1 Data

We use eye tracking corpora for sentences in four languages: English, Dutch, German, and Russian. Full sentences or longer naturally occurring text spans were read by multiple native speakers (see Appendix A.1 for detailed statistics) and tracked by high-precision eye trackers. The datasets report the following eye tracking features for each token of the stimulus text: (nFix), mean fixation duration (MFD), fixation proportion (fProp), first fixation duration (FFD), first pass duration (FPD), total reading time (TRT), number of re-fixations (nRefix), and re-read proportion (ReProp). All features are first computed for each subject separately by aggregating over the fixations, and then averaged over all subjects.

These features arguably reflect the complete reading process at the various stages of linguistic integration, from early lexical access for word recognition (e.g., FFD) up to subsequent syntactic integration taking into account regression movements (e.g., nRefix). For a review of which eye movement feature reflects which linguistic level, see Clifton et al. (2007).

For more detailed information about the data and the training procedure, see Hollenstein et al. (2021).

2.2 Model Configurations

The model is optimized to predict eye tracking features from reading as accurately as possible. For each token w in the input text, we predict a vector containing the eight eye tracking features listed above. We focus on the transformer-based models multilingual BERT (Devlin et al., 2019) and crosslingual XLM-100 (Lample and Conneau, 2019). We use pre-trained checkpoints from the HuggingFace repository.[2]

We propose the following baselines to benchmark model performance. First, we compare with a random baseline (RANDOM BL), which presents model predictions made from a randomly initialized regression layer. Second, we use a mean baseline averaged across all eye tracking features (MEAN BL), which calculates the mean value for each eye tracking feature from the training data and uses it as a prediction for all words in the test data.

Since one of our goals is to evaluate the gains from fine-tuning a pre-trained LM on eye tracking data, we also compare fine-tuning all layers (we call these models $M_{\heartsuit}$), to fine-tuning only the final regression layer ($M_{\star}$). Finally, we investigate the benefits of pre-training on large language corpora

[2] xlm-mlm-100-1280 and bert-base-multilingual-cased

by comparing the fully fine-tuned pre-trained language models ($M_{\heartsuit}$) to a model trained from scratch on randomly initialized weights ($M_{\lightning}$).

2.3 Results

Since we scale all gaze features to values between 0–100, we evaluate the models using the mean absolute error (MAE). For better readability, we report the results as *prediction accuracy*, defined as $100-$MAE. The results are presented in Table 1.

Baseline comparison The performance of the random baseline (RANDOM BL) is much lower than that of the mean baseline (MEAN BL), and therefore is not suitable for comparison. XLM-100 outperforms the mean baseline for all languages, but mBERT does not reach it for German (DE) and Russian (RU).

The impact of pre-training The results of the pre-trained and fine-tuned language models ($M_{\heartsuit}$) and the transformer models trained from scratch ($M_{\lightning}$) show a very similar performance. This demonstrates that the advantage of pre-training language models on large text corpora is only minimal for the task of predicting human eye movements. When fine-tuning only the regression layer ($M_{\medwhitestar}$), the models yield only modest (if any) improvements over the MEAN BL. However, when all layers of a model are fine-tuned ($M_{\heartsuit}$), the differences become more notable, especially for languages where less eye tracking data is available (DE and RU).

Generally, XLM-100 yields better results than mBERT for all languages, and especially for the ones with smaller datasets (DE and RU). Our results are in line with previous work showing that XLM models perform better at zero-shot eye tracking prediction for an unseen language than mBERT (Srivastava, 2022). Similarly, Hollenstein et al. (2021) find that mBERT is outperformed by monolingual models for languages with small eye tracking training datasets.

This indicates that the architecture and training objective of a model might be more important for eye tracking prediction than pre-training on large amounts of text. Transformer architectures are promising for predicting reading times, but the extensive pre-training on text input might be superfluous, as the models learn more from the fine-tuning on psychometric features. This could mean that not much linguistic knowledge is required for eye tracking prediction. Alternately, the choice of metric (MAE) might not be the most appropriate to capture the subtleties of the task. However, as we will see in Section 3, the pre-trained models show an advantage over randomly initialized models in their correlation with text readability measures.

Evaluation of individual eye tracking features The aggregated mean baseline across all eye tracking features can be misleading because it conceals the model's prediction performance for individual features. A model that yields a superior performance on the aggregated level does not necessarily outperform the mean baseline for all features. Therefore, we zoom in on individual eye tracking features and compare the performance of the fine-tuned mBERT and XLM-100 with the mean baseline in Figure 2. The results show that mBERT fails to predict MFD, FPROP and REPROP for German and Russian, while XLM-100 outperforms the aggregate mean baseline for all languages. Some features are more strongly affected by the large degree of individual variability in human eye movements (Kidd et al., 2018). We additionally visualize the feature ranges of the predicted eye tracking features compared to the real eye tracking data in Figure 1.

3 Readability Analysis

Eye movement patterns during reading are known to be influenced by the readability of texts (Rayner et al., 2006). Singh et al. (2016) assess text readability with automatically predicted eye tracking features. Although their readability assessment model was based only on predicted reading times, it yielded results comparable to models that use extensive syntactic features to compute linguistic complexity. Wiechmann et al. (2022) find that, for English, the accuracy of eye tracking prediction is systematically linked to sentence-level text features that approximate readability.

We try to disentangle the three-fold relationship between human eye movements, the capability of language models to predict these eye movement patterns, and sentence-level readability measures for English. We analyze the correlation between model predictions and readability measures to better understand the processing patterns that the model picks up.

3.1 Measuring Readability

The readability of a text is affected by variation at all levels of linguistic processing (Beinborn et al., 2012). Feng et al. (2009) introduce a large range of

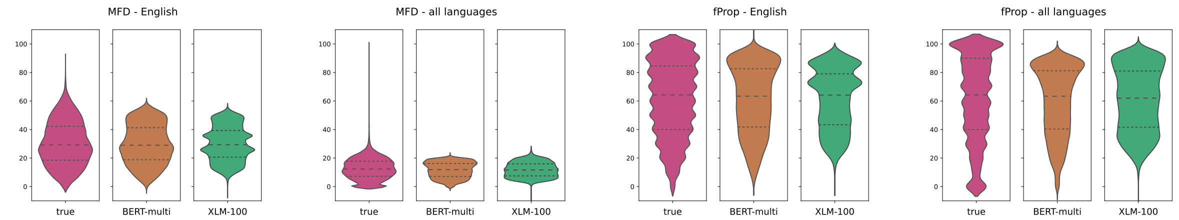

Figure 1: Feature ranges of the true eye tracking values for MFD and FPROP compared to the predicted eye tracking data (mBERT and XLM-100) for English and all four languages together.

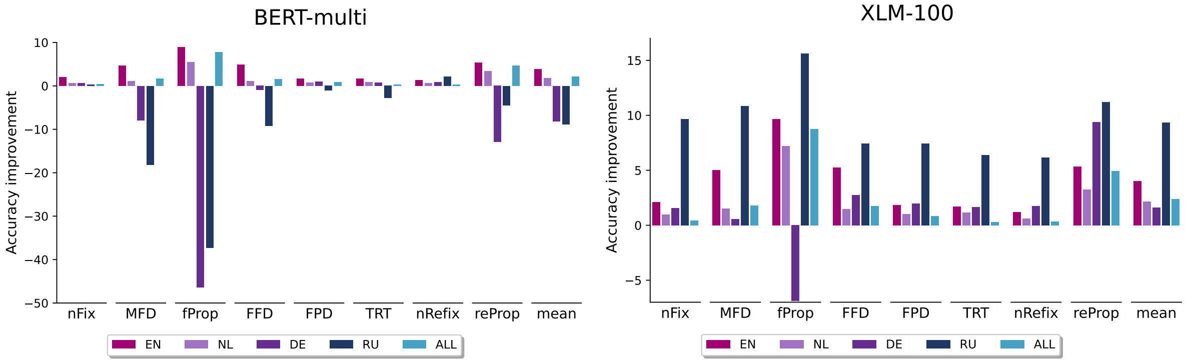

Figure 2: Improvements on prediction accuracy of the fine-tuned models mBERT and XLM-100 compared to the mean baseline across all four languages.

cognitively motivated readability measures that can be extracted using a standard natural language processing pipeline. Machine learning models trained on these measures can reliably predict the readability of texts in multiple languages (Vajjala Balakrishna, 2015). We explore a subset of 11 measures that are likely to affect eye movement patterns.

Flesch score (FLESCH): Flesch (1948) introduced the most renowned readability formula that takes the surface structure of a text into account, which is measured by the number of syllables, words and sentences. The Flesch reading ease score has been found to provide only a shallow readability estimation for English texts because it ignores deeper linguistic levels of text processing (Collins-Thompson, 2014; Bengoetxea and Gonzalez-Dios, 2021). Nahatame (2021) show that readability measures that quantify lexical and syntactic characteristics provide better approximations for predicting eye movement patterns than the Flesch score. We include it in our analysis mainly for the sake of comparison and completeness.

Word frequency (WF, ZIPF): The influence of lexical frequency on fixation duration is one of the most studied phenomena in psycholinguistic reading research. It is well established that readers tend to look longer at infrequent words (Rayner, 1977). We use the lexical frequency values provided by the `wordfreq` Python library (Speer et al., 2018) and its Zipfian variant on a logarithmic scale (Van Heuven et al., 2014).

Word length (WL): Longer words (measured in terms of number of characters) are generally fixed for longer periods. Nearly 70% of the variance in mean fixation duration can be explained by word length and word frequency (Just and Carpenter, 1980).

Sentence length (SL): We include sentence length, calculated as the number of tokens of each sentence, since the readability measures highlighted in our analysis are strongly related to the length of a sentence. Sarti et al. (2021) confirm that for all text complexity metrics, sentence length exhibits the highest correlation.

Distance to head (D2H): Sarti et al. (2021) find a strong correlation between readability measures related to dependency parsing (e.g., parse depth) and perceived complexity. Dependency features also correlate well with eye tracking patterns and can predict regressive eye movements (Lopopolo et al., 2019). In the opposite direction, Strzyz et al. (2019) show that eye tracking information can improve dependency parsing. We therefore measure

Feature	FLESCH	WL	WF	ZIPF	SL	D2H	AMB	AOA	FAM	CONC	IMAG
nFix	-0.55	0.94	-0.54	-0.82	0.96	-0.29	-0.28	0.29	-0.46	0.45	0.43
nRefix	-0.58	0.92	-0.55	-0.82	0.81	-0.23	-0.25	0.26	-0.35	0.33	0.32
MFD	-0.47	0.84	-0.44	-0.68	0.96	-0.26	-0.19	0.18	-0.42	0.42	0.40
FFD	-0.47	0.82	-0.43	-0.63	0.96	-0.26	-0.19	0.18	-0.42	0.42	0.41
FPD	-0.50	0.91	-0.50	-0.78	0.96	-0.28	-0.25	0.25	-0.44	0.43	0.42
TRT	-0.52	0.95	-0.54	-0.82	0.95	-0.28	-0.26	0.25	-0.44	0.44	0.42
fProp	-0.52	0.98	-0.48	-0.72	0.99	-0.30	-0.25	0.24	-0.45	0.45	0.44
reProp	-0.58	0.96	-0.52	-0.83	0.86	-0.23	-0.25	0.26	-0.36	0.33	0.32

Table 2: Spearman correlation coefficients of observed eye tracking features and readability measures. All correlations are significant ($p < 0.01$).

the distance to head as the number of words between the current word and its head according to the dependency tree. For example, in the sentence *She reads a mistery novel*, the distance from the word *novel* to its head *reads* is 2. We use the parser Stanza (Qi et al., 2020) for the dependency analysis.

Ambiguity level (AMB): The meaning of polysemous words can usually be disambiguated by processing the context. The effect of a high ambiguity level on eye movement patterns is therefore usually more pronounced for later processing measures such as nRefix than for early gaze metrics reflecting lexical access (Foraker and Murphy, 2012; Shen and Li, 2016). We calculate the ambiguity level for each open class word (nouns, adjectives, adverbs, and verbs) as the number of possible senses (synsets) that can be found in the NLTK implementation (Bird and Loper, 2004) of WordNet (Fellbaum). The minimum ambiguity level is 1, which means that there is only one sense for a given wordform. For example, the noun *car* has an ambiguity level of five because it appears in 5 nominal synsets. For this analysis, we exclude words that do not appear in WordNet.

Age of acquisition (AOA), familiarity (FAM), concreteness (CONC), and imageability (IMAG): We include four cognitively motivated features of word complexity that are likely to affect fixation durations in reading (Juhasz and Rayner, 2006). Ratings for age of acquisition, familiarity, concreteness, and imageability (the intensity with which a word evokes a clear mental image) in the MRC Psycholinguistic Database (Wilson, 1988) are strongly associated with each other and with other readability metrics.[3] Paetzold and Specia (2016) find that word frequencies correlate with familiarity and AoA, while the depth of a word in a thesaurus hierarchy correlates with both its concreteness and its imageability. We exclude words that do not appear in the database when calculating the correlations.

3.2 Readability and Eye Movement Patterns

We calculate the Spearman correlation coefficients between the recorded eye tracking data and the readability measures for English (Table 2). The strongest correlations can be found for sentence length, word length, and lexical frequency, which confirms three widely studied effects in reading research (Sarti et al., 2021).

Predictive Power In a second step, we analyze the correlations between four of the readability measures (FLESCH, WL, ZIPF frequency, SL) and the predictions of the different models for fixation proportion and mean fixation duration for English (see Table 3).[4] The results show that for fine-tuned mBERT, while yielding lower overall prediction accuracy when aggregating across all features, the correlation of the predicted eye movement values to word frequency and length is generally more similar to the correlation of real gaze features with word frequency and length than for the XLM-100 models. When comparing fine-tuned pre-trained models ($M_{\heartsuit}$) to models trained from scratch on eye tracking data ($M_{\lightning}$), the results presented in Table 3 also show that the correlation with word length and frequency is stronger in the predictions of the

[3] https://websites.psychology.uwa.edu.au/school/mrcdatabase/uwa_mrc.htm

[4] Correlations to other gaze features show the same trends. We analyze fProp because the models yield low prediction performance on this feature, compared to MFD, which yields high prediction results.

fine-tuned models and closer to the correlation of the real eye tracking features, showing that while pre-training might not strictly be needed for a high prediction accuracy, it does help the model to predict eye tracking features that are closer to human reading behavior in terms of text readability.

3.3 Prediction Errors

To systematically analyze the relationship between readability of the input and predictive power of the model, we focus on the prediction errors. We analyze a sample of 6,396 instances of the test set (20%) and calculate the percentage error (PE; Eq. 1) of the predictions compared to the observed scaled features.

$$PE = \frac{|Prediction - Observed|}{|Observed|} * 100 \qquad (1)$$

In Table 4, we present the correlations of the readability measures with the prediction errors for eye tracking features. Imageability, familiarity, concreteness, function words, and Zipf scale frequency values show a moderate correlation to all eye tracking features, which is slightly more notable in the case of the mBERT model. Interestingly, the strongest correlation can be found to fixation proportion. This indicates that the prediction of whether a word will be fixated or not is strongly linked to its imageability and concreteness. Similar tendencies are observed for the correlations between all investigated eye tracking measures and readability measures.

4 Word-Level Analysis

As lexical aspects seem to be highly relevant, we additionally analyze the influence of the word class on prediction errors. We focus on words that cause a prediction error ≥ the third quartile value for that feature. Figure 3 shows the aggregated results for all predicted eye movement features.

4.1 Word Classes

It can be seen that the large majority of prediction errors can be attributed to function words. This tendency is consistent across all eight gaze features.[5] Function words such as determiners, pronouns, prepositions and conjunctions, usually trigger low fixation duration and high skipping probability. It has been shown that distributional models are generally not well suited for representing function words (Bernardi et al., 2015) and that their representation in transformer-based models such as BERT is highly context-sensitive (Ethayarajh, 2019; Kim et al., 2019; Atanasova et al., 2020). Of the content words, nouns are most often mispredicted and responsible for around 10% of the errors. This is in line with Furtner et al. (2009), who indicated in a reading study that the noun is the most influential word class for facilitating the comprehension of other words.

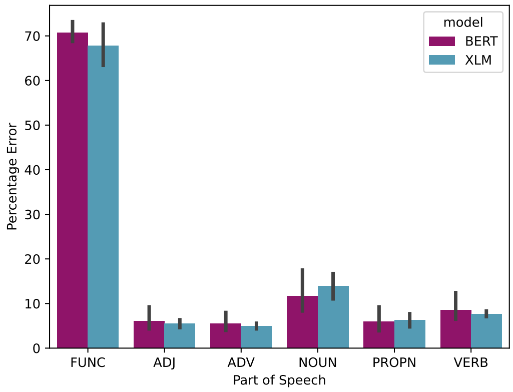

Figure 3: Percentage error per part-of-speech class, aggregated across all predicted gaze features.

4.2 A Closer Look at Function Words

We have seen that most prediction errors are caused by function words. Function or closed category words are words that are short, frequent, ambiguous, and subject to pragmatic effects in English. They are critical for language understanding.

Reading research has shown that short function words can be identified in reading without a direct fixation (Rayner et al., 1989). Similarly, Barrett and Søgaard (2015) show a negative correlation between function word frequency and fixation probability. Schmauder et al. (2000) found increased processing times in phrases immediately following a low-frequency function word. Function and content words are likely stored and accessed similarly (Diaz and McCarthy, 2009), but have different roles in text processing and constructing discourse representations. Function words show frequency effects in first fixation and first pass duration that are similar to those seen for content words. However, clear differences in reading patterns in the online processing of function and content words emerged in later processing measures (Schmauder

[5] Detailed results per feature can be found in the Appendix in Table 7.

	FPROP				MFD			
Eye tracking	FLESCH	ZIPF	WL	SL	FLESCH	ZIPF	WL	SL
Human	-0.52*	-0.72*	0.98*	0.99*	-0.47*	-0.68*	0.84*	0.96*
RANDOM BL	-0.03	-0.07	0.36	-0.16*	0.12*	-0.48	-0.20	-0.12*
MEAN BL	-0.61*	0.14*	-0.31	0.99*	-0.41*	-0.01	0.23	0.91*
$M_{\lightning}$ mBERT	-0.50*	-0.53*	0.72*	**0.99***	-0.49*	-0.52*	0.61*	**0.98***
$M_{\lightning}$ XLM-100	**-0.52***	-0.68*	0.67*	**0.99***	-0.49*	-0.65*	0.68*	0.97*
$M_{\heartsuit}$ mBERT	**-0.52***	-0.73*	**0.78***	**0.99***	**-0.48***	**-0.68***	**0.80***	0.97*
$M_{\heartsuit}$ XLM-100	-0.53*	**-0.72***	0.68*	**0.99***	-0.49*	-0.62*	0.62*	**0.98***

Table 3: Spearman correlation coefficients between real human eye tracking features or model predictions and word length, word frequency, and sentence length for **fixation proportion** (FPROP) on the left side and for **mean fixation duration** (MFD) on the right side. $M_{\lightning}$ stands for models trained from scratch and $M_{\heartsuit}$ for fine-tuned pre-trained models. Significant results are marked with * ($p < 0.01$) and results in bold are closest to human eye tracking features.

	IMAG		FAM		CONC		ZIPF		FUNCT	
Feature	BERT	XLM	BERT	XLM	BERT	XLM	BERT	XLM	BERT	XLM
NFIX	-0.18	-0.17	0.19	0.14	-0.19	-0.17	0.22	0.18	0.20	0.19
MFD	-0.18	-0.09	0.16	0.10	-0.19	-0.09	0.24	0.13	0.23	0.10
FPROP	-0.21	-0.17	0.24	0.20	-0.21	-0.17	0.34	0.26	0.29	0.22
FFD	-0.17	-0.08	0.16	0.09	-0.18	-0.09	0.23	0.12	0.22	0.10
FPD	-0.16	-0.06	0.15	0.07	-0.17	-0.06	0.19	0.07	0.18	0.06
TRT	-0.18	-0.13	0.16	0.12	-0.19	-0.14	0.21	0.16	0.19	0.16
NREFIX	-0.19	-0.14	0.22	0.16	-0.21	-0.15	0.27	0.17	0.23	0.16
REPROP	-0.14	-0.14	0.15	0.16	-0.16	-0.16	0.19	0.20	0.16	0.18

Table 4: Correlations between percentage error of the eye tracking predictions and the readability measures (imageability, familiarity, concreteness, function words and Zipf frequencies of words).

et al., 2000). These findings can be taken as evidence of the different roles the two word types have in sentence processing beyond the lexical level.

We analyze the variance in the observed eye tracking features for prepositions, determiners, and conjunctions. We calculate the standard deviation of each eye tracking feature for the six most frequent words of each category. Generally, there is a high variation in the data across all of the gaze features, and the highest deviation can be observed for mean fixation duration. To illustrate this, in Figure 4 we show the graphics of the six most common prepositions. Determiners and conjunctions are shown in Figures 5 and 6 in the Appendix (standard deviation values can also be found in Table 8 in the Appendix).

The lower fixation proportion on shorter words, together with the immense variation on the fixation duration of function words, is likely the cause for the difficulty in predicting reading times on this class accurately. The differences in skipping rate (that is, the ratio of words that are not fixated) have various origins. Skipping rate is regulated by word length (Drieghe et al., 2004), which therefore leads to differences across languages (see Fig. 1), and by proficiency, since highly proficient readers show a higher skipping rate (Eskenazi and Folk, 2015). We discuss these challenges in more detail in the next section.

5 Methodological Limitations

In this work, we make a few simplifying assumptions that are common in the field, but severely affect the interpretation of the results. We want to discuss these limitations explicitly and encourage methodological research to better address these open challenges.

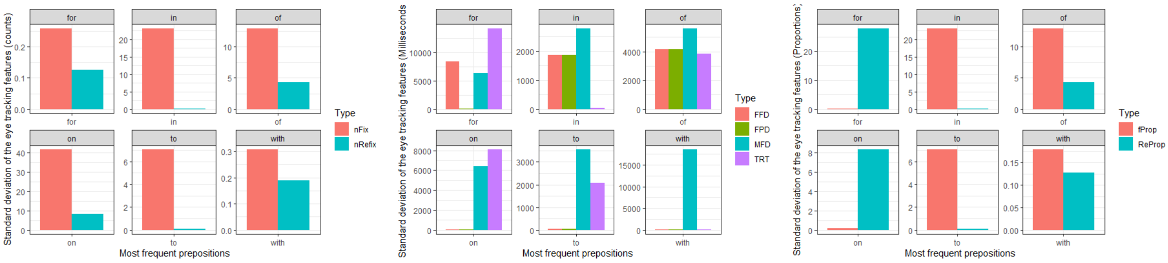

Figure 4: Standard deviations of the eye tracking features of the most frequent English prepositions. Note that the y-axis scales change in each subplot.

Token-level alignment Stimuli for eye tracking are usually pre-processed with high linguistic quality. Neural language models, on the other hand, are optimized for engineering objectives that sometimes compete with linguistic and cognitive plausibility. This discrepancy becomes obvious when inspecting the tokenization. The HuggingFace transformer-based models that we apply expect the use of subword tokenizers. Words such as close-knit! that are interpreted as a single unit in the eye tracking data are split into subword tokens ([‘close’, ‘##-’, ‘##knit’]) and it remains an open challenge how to align the different units. In our implementation, we assign the same gaze features to all subtokens and choose to compute the loss only with respect to the first subtoken. It is unclear, however, if this is the best strategy as it complicates fine-grained error analyses. For example, in the eye tracking data, punctuation is not separated from the preceding token as they are usually fixated jointly. If we apply part-of-speech tagging on words with attached punctuation signs they might get assigned incorrect tags leading to skewed results. In this work, we resolved this by manually aligning tokenized text with the eye tracking stimuli for the readability analysis. However, this is laborious and limits the size of the analyzed data.

Aggregating over participants The second matter we address is the variability between readers. The negligence of individual differences is a well-known issue in cognitive science, leading to a picture of an idealized human that is largely invariant across individuals (Levinson, 2012), and the resulting insights underestimate the extent to which human sentence processing is affected by individual differences (Kidd et al., 2018). Currently, the captured individual differences are merely treated as a source of variance that is controlled for through aggregation. As in this work, most often NLP researchers aggregate across all readers due to evidence showing that this leads to more robust results regarding model performance (Klerke and Plank, 2019). However, the high variability found in some features, such as FPROP, calls for more careful data preprocessing, possibly by considering additional cognitive tests performed during data collection and performing proper outlier detection. Alternatively, single-subject and cross-subject approaches should also be considered in eye tracking prediction for more practical applications.

Cross-lingual differences Finally, eye movements depend on the stimulus and therefore contain language-specific information (Liversedge et al., 2016). Reading patterns can be related to linguistic factors of the reader’s native language. Berzak et al. (2017) found evidence that similar languages have more similar reading patterns. Siegelman et al. (2022) found that readers of different languages vary considerably in their skipping rate and that this variability is explained by cross-lingual differences in word length distributions. It is unclear yet to what extent these differences affect the predictions of pre-trained language models. Therefore, more research is required on multilingual models that predict eye tracking in typologically more diverse languages.

6 Conclusion

Our results show that transformer models yield high accuracy in predicting cognitive language processing signals which confirms tendencies observed in previous work (e.g., Schrimpf et al., 2020; Michaelov et al., 2021). We go beyond aggregated performance metrics and provide a detailed analysis of the linguistic text complexity factors that underlie the prediction of eye movement patterns.

In a detailed analysis of fine-tuned language models that predict eye tracking features from reading, we found that the models learn more from the fine-tuning on psychometric features than from pre-training on textual input. However, the pre-trained models show an advantage over the randomly initialized models in their correlation to text readability measures. Although pre-training of large language models is not required to obtain high accuracy in predicting eye tracking feature from reading, it does contribute to a stronger correlation to text readability measures, making the predictions more similar to human reading behavior. The models struggle most to predict accurate eye tracking values for function words, which are exactly the class of words that exhibits large ranges in reading times and skipping rates, together with extreme variability between readers. The next step will be to extend the readability analysis to the other languages and to discern further between syntactic and semantic text difficulty and purely structural complexity.

This line of work does not only advance our understanding of language models and allows to compare their output to human language (Tuckute et al., 2022), but it also furthers research on new readability formulas supported by eye tracking data and machine learning methods (e.g., González-Garduño and Søgaard, 2018; Baazeem et al., 2021). We hope this work can serve as a stepping stone towards a more detailed evaluation setup for eye movement prediction from reading.

6.0.1 Acknowledgements

I. Gonzalez-Dios's research was partially funded by the Basque Government (Excellence Research Group, IT1570-22).
L. Beinborn's research is part of the VENI project Vl.Veni.211C.039, which is financed by the Dutch Research Council (NWO).
L. A. Jäger's research is partially funded by the German Federal Ministry of Education and Research (grant 01|S20043).

References

Pepa Atanasova, Jakob Grue Simonsen, Christina Lioma, and Isabelle Augenstein. 2020. A diagnostic study of explainability techniques for text classification. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3256–3274, Online. Association for Computational Linguistics.

Ibtehal Baazeem, Hend Al-Khalifa, and Abdulmalik Al-Salman. 2021. Cognitively driven Arabic text readability assessment using eye-tracking. *Applied Sciences*, 11(18):8607.

Maria Barrett and Anders Søgaard. 2015. Reading behavior predicts syntactic categories. In *Proceedings of the Nineteenth Conference on Computational Natural Language Learning*, pages 345–349, Beijing, China. Association for Computational Linguistics.

Lisa Beinborn, Torsten Zesch, and Iryna Gurevych. 2012. Towards fine-grained readability measures for self-directed language learning. In *Proceedings of the SLTC 2012 Workshop on NLP for CALL*, pages 11–19. Linköping University Electronic Press.

Kepa Bengoetxea and Itziar Gonzalez-Dios. 2021. Multiaztertest: A multilingual analyzer on multiple levels of language for readability assessment. *arXiv preprint arXiv:2109.04870*.

Raffaella Bernardi, Gemma Boleda, Raquel Fernández, and Denis Paperno. 2015. Distributional semantics in use. In *Proceedings of the First Workshop on Linking Computational Models of Lexical, Sentential and Discourse-level Semantics*, pages 95–101, Lisbon, Portugal. Association for Computational Linguistics.

Yevgeni Berzak, Chie Nakamura, Suzanne Flynn, and Boris Katz. 2017. Predicting native language from gaze. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 541–551, Vancouver, Canada. Association for Computational Linguistics.

Steven Bird and Edward Loper. 2004. NLTK: The natural language toolkit. In *Proceedings of the ACL Interactive Poster and Demonstration Sessions*, pages 214–217, Barcelona, Spain. Association for Computational Linguistics.

Charles Clifton, Adrian Staub, and Keith Rayner. 2007. Eye movements in reading words and sentences. In *Eye Movements:A Window on Mind and Brain*, pages 341–371. Elsevier.

Kevyn Collins-Thompson. 2014. Computational assessment of text readability: A survey of current and future research. *ITL-International Journal of Applied Linguistics*, 165(2):97–135.

Vera Demberg and Frank Keller. 2019. Cognitive models of syntax and sentence processing. In *Human Language: From Genes and Brains to Behavior*. MIT Press.

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, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Michele T. Diaz and Gregory McCarthy. 2009. A comparison of brain activity evoked by single content and function words: An fMRI investigation of implicit word processing. *Brain Research*, 1282:38–49.

Denis Drieghe, Marc Brysbaert, Timothy Desmet, and Constantijn De Baecke. 2004. Word skipping in reading: On the interplay of linguistic and visual factors. *European Journal of Cognitive Psychology*, 16(1-2):79–103.

Michael A. Eskenazi and Jocelyn R. Folk. 2015. Reading skill and word skipping: Implications for visual and linguistic accounts of word skipping. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 41(6):1923.

Kawin Ethayarajh. 2019. How contextual are contextualized word representations? Comparing the geometry of BERT, ELMo, and GPT-2 embeddings. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 55–65, Hong Kong, China. Association for Computational Linguistics.

Christiane Fellbaum, editor. *WordNet: An Electronic Lexical Database*. MIT Press, Cambridge, MA.

Lijun Feng, Noémie Elhadad, and Matt Huenerfauth. 2009. Cognitively motivated features for readability assessment. In *Proceedings of the 12th Conference of the European Chapter of the ACL (EACL 2009)*, pages 229–237, Athens, Greece. Association for Computational Linguistics.

Rudolph Flesch. 1948. A new readability yardstick. *Journal of Applied Psychology*, 32(3):221.

Stephani Foraker and Gregory . Murphy. 2012. Polysemy in sentence comprehension: Effects of meaning dominance. *Journal of Memory and Language*, 67(4):407–425.

Marco R Furtner, John F Rauthmann, and Pierre Sachse. 2009. Nomen est omen: Investigating the dominance of nouns in word comprehension with eye movement analyses. *Advances in Cognitive Psychology*, 5:91.

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

John T Hale, Luca Campanelli, Jixing Li, Shohini Bhattasali, Christophe Pallier, and Jonathan R Brennan. 2022. Neurocomputational models of language processing. *Annual Review of Linguistics*, 8:427–446.

Nora Hollenstein, Maria Barrett, Marius Troendle, Francesco Bigiolli, Nicolas Langer, and Ce Zhang. 2019. Advancing NLP with cognitive language processing signals. *arXiv preprint arXiv:1904.02682*.

Nora Hollenstein, Emmanuele Chersoni, Cassandra Jacobs, Yohei Oseki, Laurent Prévot, and Enrico Santus. 2022. CMCL 2022 shared task on multilingual and crosslingual prediction of human reading behavior. In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, pages 121–129, Dublin, Ireland. Association for Computational Linguistics.

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

Barbara J Juhasz and Keith Rayner. 2006. The role of age of acquisition and word frequency in reading: Evidence from eye fixation durations. *Visual Cognition*, 13(7-8):846–863.

Marcel A. Just and Patricia A. Carpenter. 1980. A theory of reading: From eye fixations to comprehension. *Psychological Review*, 87(4):329.

Evan Kidd, Seamus Donnelly, and Morten H Christiansen. 2018. Individual differences in language acquisition and processing. *Trends in Cognitive Science*, 22(2):154–169.

Najoung Kim, Roma Patel, Adam Poliak, Patrick Xia, Alex Wang, Tom McCoy, Ian Tenney, Alexis Ross, Tal Linzen, Benjamin Van Durme, Samuel R. Bowman, and Ellie Pavlick. 2019. Probing what different NLP tasks teach machines about function word comprehension. In *Proceedings of the Eighth Joint Conference on Lexical and Computational Semantics (*SEM 2019)*, pages 235–249, Minneapolis, Minnesota. Association for Computational Linguistics.

Sigrid Klerke and Barbara Plank. 2019. At a glance: The impact of gaze aggregation views on syntactic tagging. In *Proceedings of the Beyond Vision and LANguage: inTEgrating Real-world kNowledge (LANTERN)*, pages 51–61, Hong Kong, China. Association for Computational Linguistics.

Guillaume Lample and Alexis Conneau. 2019. Cross-lingual language model pretraining. In *Advances in Neural Information Processing Systems 32 (NeurIPS 2019)*.

Stephen C. Levinson. 2012. The original sin of cognitive science. *Topics in Cognitive Science*, 4(3):396–403.

Simon P Liversedge, Denis Drieghe, Xin Li, Guoli Yan, Xuejun Bai, and Jukka Hyönä. 2016. Universality in eye movements and reading: A trilingual investigation. *Cognition*, 147:1–20.

Alessandro Lopopolo, Stefan L. Frank, Antal van den Bosch, and Roel Willems. 2019. Dependency parsing with your eyes: Dependency structure predicts eye regressions during reading. In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, pages 77–85, Minneapolis, Minnesota. Association for Computational Linguistics.

Sandeep Mathias, Diptesh Kanojia, Abhijit Mishra, and Pushpak Bhattacharyya. 2020. A survey on using gaze behaviour for natural language processing. *Proceedings of IJCAI*.

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, Online. Association for Computational Linguistics.

James A Michaelov, Megan D Bardolph, Seana Coulson, and Benjamin Bergen. 2021. Different kinds of cognitive plausibility: Why are transformers better than RNNs at predicting N400 amplitude? In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 43.

Shingo Nahatame. 2021. Text readability and processing effort in second language reading: A computational and eye-tracking investigation. *Language learning*.

Gustavo Paetzold and Lucia Specia. 2016. Inferring psycholinguistic properties of words. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 435–440, San Diego, California. Association for Computational Linguistics.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 101–108, Online. Association for Computational Linguistics.

Keith Rayner. 1977. Visual attention in reading: Eye movements reflect cognitive processes. *Memory & Cognition*, 5(4):443–448.

Keith Rayner, Kathryn H Chace, Timothy J. Slattery, and Jane Ashby. 2006. Eye movements as reflections of comprehension processes in reading. *Scientific Studies of Reading*, 10(3):241–255.

Keith Rayner, Sara C Sereno, Robin K Morris, A Rene Schmauder, and Charles Clifton Jr. 1989. Eye movements and on-line language comprehension processes. *Language and Cognitive Processes*, 4(3-4):SI21–SI49.

Lavinia Salicchi, Rong Xiang, and Yu-Yin Hsu. 2022. HkAmsters at CMCL 2022 shared task: Predicting eye-tracking data from a gradient boosting framework with linguistic features. In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, pages 114–120, Dublin, Ireland. Association for Computational Linguistics.

Gabriele Sarti, Dominique Brunato, and Felice Dell'Orletta. 2021. That looks hard: Characterizing linguistic complexity in humans and language models. In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, pages 48–60, Online. Association for Computational Linguistics.

A René Schmauder, Robin K Morris, and David V Poynor. 2000. Lexical processing and text integration of function and content words: Evidence from priming and eye fixations. *Memory & Cognition*, 28(7):1098–1108.

Martin Schrimpf, Idan A Blank, Greta Tuckute, Carina Kauf, Eghbal A Hosseini, Nancy G Kanwisher, Joshua B Tenenbaum, and Evelina Fedorenko. 2020. Artificial neural networks accurately predict language processing in the brain. *BioRxiv*.

Wei Shen and Xingshan Li. 2016. Processing and representation of ambiguous words in Chinese reading: Evidence from eye movements. *Frontiers in Psychology*, 7:1713.

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*, pages 1–21.

Abhinav Deep Singh, Poojan Mehta, Samar Husain, and Rajkumar Rajakrishnan. 2016. Quantifying sentence complexity based on eye-tracking measures. In *Proceedings of the Workshop on Computational Linguistics for Linguistic Complexity (CL4LC)*, pages 202–212, Osaka, Japan. The COLING 2016 Organizing Committee.

Robyn Speer, Joshua Chin, Andrew Lin, Sara Jewett, and Lance Nathan. 2018. LuminosoInsight/wordfreq: v2.2.

Harshvardhan Srivastava. 2022. Poirot at CMCL 2022 shared task: Zero shot crosslingual eye-tracking data prediction using multilingual transformer models. In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, pages 102–107, Dublin, Ireland. Association for Computational Linguistics.

Michalina Strzyz, David Vilares, and Carlos Gómez-Rodríguez. 2019. Towards making a dependency parser see. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1500–1506, Hong Kong, China. Association for Computational Linguistics.

Ece Takmaz. 2022. Team DMG at CMCL 2022 shared task: Transformer adapters for the multi- and cross-lingual prediction of human reading behavior. In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, pages 136–144, Dublin, Ireland. Association for Computational Linguistics.

Greta Tuckute, Aalok Sathe, Mingye Wang, Harley Yoder, Cory Shain, and Evelina Fedorenko. 2022. SentSpace: Large-scale benchmarking and evaluation of text using cognitively motivated lexical, syntactic, and semantic features. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: System Demonstrations*, pages 99–113, Hybrid: Seattle, Washington + Online. Association for Computational Linguistics.

Sowmya Vajjala Balakrishna. 2015. *Analyzing Text Complexity and Text Simplification: Connecting Linguistics, Processing and Educational Applications*. Ph.D. thesis, Universität Tübingen.

Walter JB Van Heuven, Pawel Mandera, Emmanuel Keuleers, and Marc Brysbaert. 2014. SUBTLEX-UK: A new and improved word frequency database for British English. *Quarterly Journal of Experimental Psychology*, 67(6):1176–1190.

Shravan Vasishth, Titus von der Malsburg, and Felix Engelmann. 2013. What eye movements can tell us about sentence comprehension. *Wiley Interdisciplinary Reviews: Cognitive Science*, 4(2):125–134.

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 (Volume 1: Long Papers)*, pages 5276–5290, Dublin, Ireland. Association for Computational Linguistics.

Michael Wilson. 1988. MRC psycholinguistic database: Machine-usable dictionary, version 2.00. *Behavior Research Methods, Instruments, & Computers*, 20(1):6–10.

A Appendix

A.1 Eye Tracking Corpora

The details of the datasets used in this work are presented in Table 5.

Language	Corpus	Subjs.	Sents.	Sent. length	Tokens	Types	Word length	Flesch
English	Dundee	10	2,379	21.7 (1–87)	51,497	9,488	4.9 (1–20)	53.3
	GECO	14	5,373	10.5 (1–69)	56,410	5,916	4.6 (1–33)	77.4
	ZuCo	30	1,053	19.5 (1–68)	20,545	5,560	5.0 (1–29)	50.6
Dutch	GECO	19	5,190	11.64 (1–60)	59,716	5,575	4.5 (1–22)	57.5
German	PoTeC	30	97	19.5 (5–51)	1,895	847	6.5 (2–33)	36.4
Russian	RSC	103	144	9.4 (5–13)	1,357	993	5.7 (1–18)	64.7

Table 5: Descriptive statistics of all eye tracking datasets. Sentence length and word length are expressed as the mean with the min-max range in parentheses. The last column shows the Flesch Reading Ease score (Flesch, 1948) which ranges from 0 to 100 (higher score indicates easier to read). Adaptations of the Flesch score were used for Dutch (NL), German (DE) and Russian (RU).

A.2 Eye Movement Features

The values of the eye tracking features vary over different ranges (see Figure 1, left-most subplots). FFD, for example, is measured in milliseconds, and average values are around 200 ms, whereas REPROP is a proportional measure, and therefore assumes floating-point values between 0 and 1. We standardize all eye tracking features independently (range: 0–100), so that the loss can be calculated uniformly over all feature dimensions.

A.3 Additional Correlation Results

In Table 6, we present additional correlations of the readability measures in relation to the eye tracking prediction errors.

	AOA		D2H		AMB	
Feature	BERT	XLM	BERT	XLM	BERT	XLM
NFIX	0.00	0.02	-0.03	-0.05	0.04	0.01
MFD	-0.09	-0.04	0.00	-0.03	0.04	0.00
FPROP	-0.18	-0.12	-0.03	-0.04	0.13	0.05
FFD	-0.09	-0.04	-0.01	-0.03	0.02	-0.01
FPD	-0.06	-0.01	0.00	-0.02	0.00	-0.04
TRT	0.02	-0.05	-0.02	-0.04	0.01	0.01
NREFIX	-0.01	0.01	0.00	-0.03	0.14	0.09
REPROP	0.05	0.01	0.01	-0.01	0.13	0.12

Table 6: Correlations between percentage error and readability measures (age of acquisition, distance to head, and ambiguity level).

A.4 Error Rate by Part-of-Speech

Table 7 shows the percentage error for each part-of-speech class and each eye tracking feature.

	FUNC		ADJ		ADV		NOUN		PROPN		VERB	
Feature	Bert	XLM	Bert	XLM	Bert	XLM	Bert	XLM	Bert	XLM	Bert	XLM
nFix	68.54	66.10	4.69	5.44	4.63	4.82	9.38	15.76	6.38	8.26	6.38	7.88
MFD	68.86	62.16	5.32	6.32	4.57	5.94	9.63	17.39	4.69	7.50	6.94	8.19
fProp	72.17	69.17	4.57	4.75	4.50	4.44	8.13	14.82	4.38	6.82	6.25	6.82
FFD	67.98	62.48	5.25	6.25	4.63	5.88	10.44	16.89	4.75	7.13	6.94	8.51
FPD	68.36	58.91	5.25	7.32	4.44	6.19	9.82	18.89	5.19	8.51	6.94	8.69
TRT	68.36	66.60	16.60	6.57	14.03	5.19	31.03	13.07	16.40	6.94	21.94	8.57
nRefix	75.88	78.70	3.71	3.71	3.61	3.61	7.46	7.46	2.82	2.82	6.52	6.52
reProp	75.88	78.70	3.71	3.71	3.61	3.61	7.46	7.46	2.82	2.82	6.52	6.52

Table 7: Percentage error for each part-of-speech class and each eye tracking feature.

A.5 Standard Deviations of Function Words

Table 8 shows the standard deviations of the predictions of the most frequent function word classes, namely prepositions (PREP), determiners (DET), and conjunctions (CONJ).

	Lemma	nFix	FFD	FPD	TRT	MFD	fProp	nRefix	reProp
PREP	at	0.20	33.87	35.96	42.79	**4562.23**	0.15	10.10	10.10
	by	0.25	43.47	44.57	51.42	**5109.68**	0.19	0.09	0.07
	for	0.26	**8365.60**	42.11	**14199.50**	**6347.25**	0.19	0.13	27.87
	from	0.25	39.08	40.07	**12828.73**	38.53	0.18	0.11	35.43
	in	23.15	**1859.10**	**1859.05**	41.57	**2782.60**	23.15	0.07	0.06
	of	12.92	**4158.18**	**4158.11**	**3839.08**	**5614.73**	12.92	4.31	4.31
	on	41.64	35.31	36.95	**8120.68**	**6406.01**	0.16	8.33	8.33
	to	7.07	35.79	37.38	**2073.26**	**3562.01**	7.07	0.08	0.07
	up	0.23	46.17	48.59	52.53	46.42	0.20	0.07	0.06
	with	0.31	43.47	47.71	62.36	**18589.94**	0.18	0.19	0.13
DET	a	12.48	**1168.31**	**1168.30**	**1168.27**	**6726.54**	17.64	0.07	0.05
	all	0.24	43.44	47.06	57.88	**11703.01**	0.18	0.12	0.08
	another	0.44	56.89	70.55	98.66	54.99	0.23	0.29	0.18
	any	0.23	40.18	42.81	50.25	39.31	0.17	0.12	0.11
	every	0.33	40.36	46.56	48.93	**36800.71**	0.15	0.25	0.14
	no	0.25	39.11	40.87	50.50	38.53	0.18	0.10	0.08
	some	0.31	46.57	49.31	60.57	**26367.42**	0.18	0.20	0.11
	that	0.26	41.07	45.10	53.25	40.90	0.18	0.15	0.11
	the	9.06	**2060.29**	919.18	**2349.22**	**5173.09**	9.06	0.11	0.08
	this	0.32	47.54	53.64	67.83	**26854.66**	0.19	0.18	0.12
CONJ	and	0.27	37.11	41.45	48.52	**5318.58**	32.99	0.11	0.08
	as	0.22	38.20	39.10	44.08	**10602.03**	0.17	0.08	0.06
	because	0.32	38.95	42.76	53.55	38.80	0.15	0.24	0.12
	but	0.31	44.79	50.00	63.22	**5564.67**	0.20	0.15	0.09
	if	0.25	47.90	49.16	56.45	47.71	0.21	0.09	0.07
	of	0.25	44.95	51.17	56.85	45.99	0.20	0.07	0.07
	or	0.22	44.45	48.68	50.86	44.59	0.18	0.08	0.07
	that	0.26	41.08	44.46	56.21	**21359.14**	0.18	0.13	0.10
	when	113.83	35.28	44.65	63.01	**18228.19**	0.17	0.21	0.14
	where	0.39	48.09	57.04	68.93	47.88	0.19	0.25	0.16

Table 8: Standard deviation of the most frequent function words.

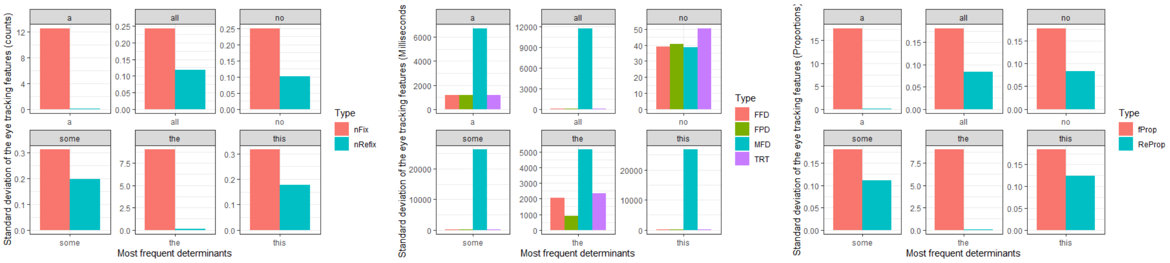

Figure 5: Standard deviations of the eye tracking features of the most frequent English **determiners**. Note that the y-axis scales change in each subplot.

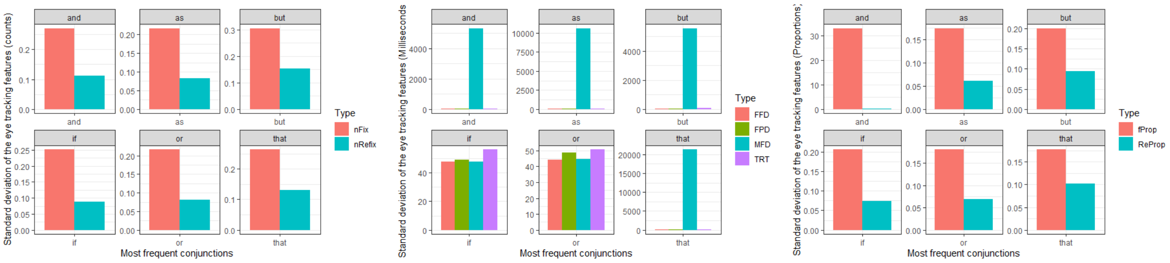

Figure 6: Standard deviations of the eye tracking features of the most frequent English **conjunctions**. Note that the y-axis scales change in each subplot.