Measuring the Impact of (Psycho-)Linguistic and Readability Features and Their Spill Over Effects on the Prediction of Eye Movement Patterns

Daniel Wiechmann
University of Amsterdam
d.wiechmann@uva.nl

Yu Qiao
RWTH-Aachen University
yu.qiao@rwth-aachen.de

Elma Kerz
RWTH-Aachen University
elma.kerz@ifaar.rwth-aachen.de

Justus Mattern
RWTH-Aachen University
justus.mattern@rwth-aachen.de

Abstract

There is a growing interest in the combined use of NLP and machine learning methods to predict gaze patterns during naturalistic reading. While promising results have been obtained through the use of transformer-based language models, little work has been undertaken to relate the performance of such models to general text characteristics. In this paper we report on experiments with two eye-tracking corpora of naturalistic reading and two language models (BERT and GPT-2). In all experiments, we test effects of a broad spectrum of features for predicting human reading behavior that fall into five categories (syntactic complexity, lexical richness, register-based multiword combinations, readability and psycholinguistic word properties). Our experiments show that both the features included and the architecture of the transformer-based language models play a role in predicting multiple eye-tracking measures during naturalistic reading. We also report the results of experiments aimed at determining the relative importance of features from different groups using SP-LIME.

1 Introduction

Extensive studies using eye-trackers to observe gaze patterns have shown that humans read sentences efficiently by performing a series of fixations and saccades (for comprehensive overviews, see, e.g. Rayner et al. (2012), Seidenberg (2017), and Brysbaert (2019)). During a fixation, the eyes stay fixed on a word and remain fairly static for 200-250 milliseconds. Saccades are rapid jumps between fixations that typically last 20-40 ms and span 7-9 characters. In addition, when reading, humans do not fixate one word at a time, i.e. some saccades run in the opposite direction, and some words or word combinations are fixed more than once or skipped altogether. Much of the early work in this area was concerned with the careful construction of sentences to model human reading behavior and understand predictive language processing (Staub, 2015; Kuperberg and Jaeger, 2016). The use of isolated, decontextualized sentences in human language processing research has been questioned on ecological validity grounds. With the growing awareness of the importance of capturing naturalistic reading, new corpora of eye movement data over contiguous text segments have emerged. Such corpora serve as a valuable source of data for establishing the basic benchmarks of eye movements in reading and provide an essential testing ground for models of eye movements in reading, such as the E-Z Reader model (Reichle et al., 1998) and the SWIFT model (Engbert et al., 2005). They are also used to evaluate theories of human language processing in psycholinguistics: For example, the predictions of two theories of syntactic processing complexity (dependency locality theory and surprisal) were tested in the Dundee Corpus, which contains the eye-tracking record of 10 participants reading 51,000 words of newspaper text (Demberg and Keller, 2008). Subsequent work has presented accounts where the ability of a language model to predict reading times is a linear function of its perplexity (Goodkind and Bicknell, 2018). More recent work has employed transformer-based language models to directly predict human reading patterns across new datasets of eye-tracking and electroencephalogra-

phy during natural reading (Schrimpf et al., 2021; Hollenstein et al., 2021, for more details see the related work section below). While this work has made significant progress, there is limited work aimed at determining the role of general text properties in predicting eye movement patterns in corpora of naturalistic reading. To date, research has addressed this issue only peripherally (Lowder et al., 2018; Snell and Theeuwes, 2020; Hollenstein et al., 2021), examining the role of text features only on the basis of a small number of linguistic features.

In this paper, we conduct a systematic investigation of the effects of text properties on eye movement prediction: We determine the extent to which these properties affect the prediction accuracy of two transformer-based language models, BERT and GPT-2. The relationship between these properties and model performance is investigated in two ways: (a) building on the approaches in Lowder et al. (2018) and Hollenstein et al. (2021), by investigating the sensitivity of model predictions to a wide range of text features, and (b) by incorporating text features into the transformer-based language models. With respect to the latter, we examine the effects of the preceding sentence on gaze measurement within the sentence of interest. This was motivated by psycholinguistic literature that has demonstrated "spillover" effects, where the fixation duration on a word is affected by linguistic features of the preceding context (Pollatsek et al., 2008; Shvartsman et al., 2014, see also Barrett and Hollenstein (2020) for a reference to the utility of information about preceding input). Computational reading models have not addressed linguistic concepts beyond the level of the fixated word much, with a few exceptions, e.g. spillover effects related to previewing the next word n+1 during the current fixation on word n (Engbert et al., 2005). Here we extend the study of spillover effects to the effects of textual features of the preceding sentence. To our knowledge, this is the first systematic attempt to investigate the effects of textual features on the prediction of eye-tracking measures in a corpus of naturalistic reading by considering a large number of features spanning different levels of linguistic

analysis.

2 Related work

In this section, we provide a brief overview of the available literature that has used transformer-based language models to predict human reading patterns, as well as the literature that has investigated the role of text properties on word predictability during naturalistic reading.

Schrimpf et al. (2021) evaluated a broad range of language models on the match of their internal representations to three datasets of human neural activity (fMRI and ECoG) during reading. Their results indicated that transformer-based models perform better than recurrent networks or word-level embedding models. They also found that the models with the best match with human language processing were models with unidirectional attention transformer architectures: specifically the generative pretrained transformer (GPT-2) (Radford et al., 2019), consistently outperformed all other models in both fMRI and ECoG data from sentence-processing tasks. Hollenstein et al. (2021) presented the first study analyzing to what extent transformer language models are able to directly predict human gaze patterns during naturalistic reading. They compare the performance of language-specific and multilingual pretrained and fine-tuned BERT and XLM models to predict reading time measures of eye-tracking datasets in four languages (English, Dutch, German, and Russian). Their results show that both monolingual and multilingual transformer-based models achieve surprisingly high accuracy in predicting a range of eye-tracking features across all four languages. For the English GECO dataset, which is also used in the current study, the BERT and XLM models yielded prediction accuracies (100 - mean absolute error (MAE)) ranging between 91.15% (BERT-EN) and 93.89% (XLM-ENDE).

To our knowledge, the first study to investigate the role of textual characteristics on word predictability during naturalistic reading is an experimental study conducted by Lowder et al. (2018). This study implemented a large-scale cumulative cloze task to collect word-by-word predictability

data (surprisal and entropy reduction scores) for 40 text passages which were subsequently read by 32 participants while their eye movements were recorded. Lowder et al. (2018) found that surprisal scores were associated with increased reading times in all eye-tracking measures. They also observed a significant effect of text difficulty, measured by Flesch–Kincaid grade level of each paragraph (Kincaid et al., 1975), such that increases in text difficulty were associated with increased reading times. Crucially, their study yielded evidence of interactions between predictability (surprisal scores) and paragraph difficulty. In the above-mentioned computational study, Hollenstein et al. (2021) also investigated the influence of textual characteristics (word length, text readability) on model performance. Text readability was measured using Flesch Reading Ease scores (Flesch, 1948). Their results indicated that the models learned to reflect characteristics of human reading, such as sensitivity to word length. They also found that model accuracy was higher in more easily readable sentences.

3 Experiments

3.1 Datasets

We analyze eye movement data from two eye-tracking corpora of natural reading, the Ghent Eye-Tracking Corpus (GECO; (Cop et al., 2017)) and the Provo corpus (Luke and Christianson, 2018). In both corpora the participants read full sentences within longer spans of naturally occurring text at their own speed while their eye movements were recorded. The GECO corpus is large dataset of eye movement of a monolingual and bilingual readers who read a complete novel, Agatha Christie's 'The Mysterious Affair at Styles'. It contains eye-tracking data from 14 English native speakers and 19 bilingual speakers of Dutch and English, who read parts of the novel in its original English version and another part of its Dutch translation. In the present work, we focus on the analysis of the data from the monolingual English native speakers. These participants read a total of 5031 sentences amounting to a total of 54364 word tokens. The Provo Corpus

is a dataset of eye movements of skilled readers reading connected text. It consists of eye movement data from 84 native English-speaking participants from Brigham Young University, who read 55 short passages from a variety of sources, including online news articles, popular science magazines, and public-domain works of fiction. These passages were an average of 50 words long for a total of 2,689 word tokens.

3.2 Measurement of text properties

The texts from both datasets (GECO and PROVO) were automatically analyzed using CoCoGen (Ströbel et al., 2016), a computational tool that implements a sliding window technique to calculate sentence-level measurements that capture the within-text distributions of scores for a given language feature (for current applications of the tool in the context of text classification, see Kerz et al. (2020, 2021)). We extract a total of 107 features that fall into five categories: (1) measures of syntactic complexity (N=16), (2) measures of lexical richness (N=14), (3) register-based n-gram frequency measures (N=25), (4) readability measures (N=14), and (5) psycholinguistic measures (N=38). A concise overview of the features used in this study is provided in Table 5 in the appendix. Tokenization, sentence splitting, part-of-speech tagging, lemmatization and syntactic PCFG parsing were performed using Stanford CoreNLP (Manning et al., 2014). The syntactic complexity measures comprise (i) surface measures that concern the length of production units, such as the mean length of words, clauses and sentences, (ii) measures of the type and incidence of embeddings, such as dependent clauses per T-Unit or verb phrases per sentence or (iii) the frequency of particular types of particular structures, such as the number of complex nominal per clause. These features are implemented based on descriptions in Lu (2010) and using the Tregex tree pattern matching tool (Levy and Andrew, 2006) with syntactic parse trees for extracting specific patterns. Lexical richness measures fall into three distinct sub-types: (i) lexical density, such as the ratio of the number of lexical (as opposed to grammatical) words to the total number of words in a text,

(iii) lexical variation, i.e. the range of vocabulary as displayed in language use, captured by text-size corrected type-token ratio and (iii) lexical sophistication, i.e. the proportion of relatively unusual or advanced words in the learner's text, such as the number of New General Service List (Browne et al., 2013). The operationalizations of these measures follow those described in Lu (2012) and Ströbel (2014). The register-based n-gram frequency measures are derived from the five register sub-components of the Contemporary Corpus of American English (COCA, (Davies, 2008)): spoken, magazine, fiction, news and academic language[1]. These measures consider both the register-specific frequency rank and count:

$$Norm_{n,s,r} = \frac{|C_{n,s,r}| \cdot \log\left[\prod_{c \in |C_{n,s,r}|} freq_{n,r}(c)\right]}{|U_{n,s}|} \quad (1)$$

Let $A_{n,s}$ be the list of n-grams ($n \in [1,5]$) appearing within a sentence s, $B_{n,r}$ the list of n-gram appearing in the n-gram frequency list of register r ($r \in \{acad, fic, mag, news, spok\}$) and $C_{n,s,r} = A_{n,s} \cap B_{n,r}$ the list of n-grams appearing both in s and the n-gram frequency list of register r. $U_{n,s}$ is defined as the list of unique n-gram in s, and $freq_{n,r}(a)$ the frequency of n-gram a according to the n-gram frequency list of register r. The total of 25 measures results from the combination of (a) a 'reference list' containing the top 100k most frequent n-grams and their frequencies from one of five registers of the COCA corpus and (b) the size of the n-gram ($n \in [1,5]$). The readability measures combine a word familiarity variable defined by prespecified vocabulary resource to estimate semantic difficulty together with a syntactic variable, such as average sentence length. Examples of these measures are the Fry index (Fry, 1968) or the SMOG (McLaughlin, 1969). Finally, the psycholinguistic measures capture cognitive aspects of reading not directly addressed by the surface vocabulary and syntax features of traditional formulas. These measures include a word's average age-of-acquisition (Kuperman et al., 2012) or prevalence, which refers to the number of people knowing the word (Brysbaert et al., 2019; Johns et al., 2020).

3.3 Eye-tracking measures

We analyze data from eight word-level reading time measures, which were also investigated in Hollenstein et al. (2021). The measures include general word-level characteristics such as (1) the number of fixations (NFX), i.e. the number of times a subject fixates on a given word w, averaged over all participants, (2) mean fixation duration (MFD), the average fixation duration of all fixations made on w, averaged over all participants and (3) fixation proportion (FXP), the number of subjects that fixated w, divided by the total number of participants. 'Early processing' measures pertain to the early lexical and syntactic processing and are based on the first time a word is fixated. These features include: (4) first fixation duration (FFD), i.e. the duration of the first fixation on w (in milliseconds), averaged over all subjects and (5) first pass duration (FPD), i.e. the sum of all fixations on w from the first time a subject fixates w to the first time the subject fixates another token. 'Late processing' measures capture the late syntactic processing and are based on words which were fixated more than once. These measures comprise (6) total fixation duration (TFD), i.e. the sum of the duration of all fixations made on w, averaged over all subjects, (7) number of re-fixations (NRFX), the number of times w is fixated after the first fixation, i.e., the maximum between 0 and the NFIX-1, averaged over all subjects and (8) re-read proportion (RRDP), the number of subjects that fixated w more than once, divided by the total number of subjects. The means, standard deviations and observed ranges for all eye-tracking features are shown in Tables 1 and 2. Like in Hollenstein et al. (2021), before being entered into the models, all eye-tracking features were scaled between 0 and 100 so that the loss can be calculated uniformly over all features.

---

[1]The Contemporary Corpus of American English is the largest genre-balanced corpus of American English, which at the time the measures were derived comprised of 560 million words.

Feature	M	SD	Min	Max
NFX	0.81	0.45	0.00	7.50
MFD	128.41	58.98	0.00	350.92
FXP	0.61	0.25	0.00	1.00
FFD	129.28	60.06	0.00	371.31
FPD	143.25	77.49	0.00	1425.86
TFD	168.20	102.44	0.00	1804.00
NRFX	0.20	0.26	0.00	6.50
RRDP	0.15	0.16	0.00	1.00

Table 1: Descriptive statistics of eye-tracking measures for the GECO dataset.

Feature	M	SD	Min	Max
NFX	0.95	0.47	0.13	3.61
MFD	139.91	52.13	23.07	272.71
FXP	0.66	0.22	0.13	1.00
FFD	139.83	52.02	23.18	276.86
FPD	165.91	80.27	24.24	736.62
TFD	198.21	107.20	24.24	940.50
NRFX	0.28	0.29	0.00	2.62
RRDP	0.21	0.17	0.00	0.87

Table 2: Descriptive statistics of eye-tracking measures for the PROVO dataset.

4 Modeling approach

Deep neural transformer-based language models create contextualized word representations that are sensitive to the context in which the words appear. These models have yielded significant improvements on a diverse array of NLP tasks, ranging from question answering to coreference resolution. We compare two such models in terms of their ability to predict eye-tracking features: 'Bidirectional Encoder Representations from Transformers' (BERT) (Devlin et al., 2018) and 'Generative Pre-trained Transformer 2' (GPT-2) (Radford et al., 2019). BERT is an autoencoder model trained with a dual objective function of predicting masked words and the next sentence. It consists of stacked transformer encoder blocks and uses self-attention, where each token in an input sentence looks at the bidirectional context, i.e. tokens on left and right of the considered token. In contrast, GPT-2 is an autoregressive model consisting of stacked transformer decoder blocks trained with a language modelling objective, where the given sequence of tokens is used to predict the next token. While GPT-2 uses self-attention as well, it employs masking to prevent words from attending to following tokens, hereby processing language fully unidirectionally. BERT is trained on the BooksCorpus (800M words) and Wikipedia (2,500M words), whereas GPT-2 is trained on WebText, an 8-million documents subset of CommonCrawl amounting to 40 GB of text. We chose the BERT base model (cased) because it is most comparable to GPT-2 with respect to

number of layers and dimensionality (BERT base model (cased) has 110M trainable parameters, GPT-2 has 117M).

We evaluate the eye-tracking predictions of the models both on within-domain text, using an 80/10/10 split of the much larger GECO dataset (representing fiction language), as well as on out-of-domain text using the complete, much smaller PROVO dataset (comprising also online news and popular science magazine language). Furthermore, since overly aggressive fine-tuning may cause catastrophic forgetting (Howard and Ruder, 2018), we perform all experiments both with 'frozen' language models, where all the layers of the language model are frozen and only the attached neural network layers are trained, and also 'fully fine-tuned' language models, where the error is back-propagated through the entire architecture and the pretrained weights of the model are updated based on the GECO training set.

For all models we explored in this paper, we apply a dropout rate of 0.1 and a l2 regularization of 1×10^{-4}. We use AdamW as the optimizer and mean squared error as the loss function. We use a fixed learning rate with warmup. During warmup, the learning rates are linearly increased to the peak learning rates and then fixed. For BERT with a 'frozen' language model, the peak learning rate is 5×10^{-4} with 5 warmup steps and for GPT-2 with a 'frozen' language model, it is 0.001 also with 5 warmup steps. Models with 'fully fine-tuned' language models are trained with two phases. In the first phase, the weights of the lan-

guage models are frozen and only regression layers are trained. During this phase, peak learning rates of 3×10^{-4} for BERT and 0.001 for GPT-2 are used. For both models, the first phase is performed over 12 epochs with 5 warmup steps. In the second phase, we unfreeze the weights of language models and fine-tune the language models together with the regression layers. During this phase, the BERT-based model is trained with a peak learning rate of 5×10^{-5} while GPT-2-based model is trained with a peak learning rate of 5×10^{-4}. The number of warmup steps for training both models in this phase is 3. We adopted a two-phase training procedure since preliminary experiments showed that this procedure yields same results as training the entire models from the first epoch, yet it can speed up model convergence. All hyper-parameters are optimized through grid search.

4.1 Influence of text characteristics on model performance

To investigate the impact of the text properties listed in Section 3.2 on prediction accuracy, we partitioned the GECO testset into deciles according to each textual property, i.e. each of the 107 features. We then calculated the Pearson correlation coefficients between the decile of a given textual feature and the mean absolute error (MAE) of a given model. We expected to observe higher prediction accuracy (lower MAE) for sentences with higher readability, lower syntactic complexity, lower lexical richness, higher n-gram frequency and less demanding psycholinguistic properties, i.e. lower age-of-acquisition scores and higher prevalence scores.

4.2 Integration of text characteristics using a hybrid modeling approach

To determine whether eye movement patterns were affected by textual characteristics of the previous sentences (sentence spillover effects), a bidirectional LSTM (BLSTM) model was integrated into the predictive models (Figure 1). This BLSTM model reads 107 dimensional vectors of textual features $CM_{i-N}, \cdots, CM_{i-1}$ from N previous sentences[2] as its input, transforms them through 4 BLSTM layers of 512 hidden units each, and outputs a 1024 dimensional vector $[\overrightarrow{h}_{4N}|\overleftarrow{h}_{41}]$, that is a concatenation of the last hidden states of the 4th BLSTM layer in the forward and backward directions $\overrightarrow{h}_{4N}$, $\overleftarrow{h}_{41}$. A fully connected (FC) layer is added on top of the BLSTM layers to reduce the dimension of BLSTM model output to 256 (C_i). Meanwhile, another FC layer is added to the pre-trained language model (BERT or GPT-2) in order to reduce its logits to the same dimension ($E_{i1}, \cdots, E_{iM}$). The reduced BLSTM output is then added to each of the reduced language model logits. Finally, the 256-dimensional joint vectors are fed to a final regression layer to predict human reading behavior. The procedures used to train the 'hybrid' models with textual characteristics of the previous sentences was identical to those specified above. Grid search yielded the same optimized values for all hyperparameters, except for the peak learning rate of 'fully fine-tuned' model with GPT-2 in second training phase, which was 1×10^{-4}.

To assess the relative importance of the feature groups, we employed Submodular Pick Lime (SP-LIME; Ribeiro et al. (2016)), a method to construct a global explanation of a model by aggregating the weights of the linear models. We first construct local explanations using LIME with a linear local explanatory model, exponential kernel function with Hamming distance and a kernel width of $\sigma = 0.75\sqrt{d}$, where d is the number of feature groups. The global importance score of the SP-LIME for a given feature group j can then be derived by: $I_j = \sqrt{\sum_{i=1}^{n} |W_{ij}|}$, where W_{ij} is the jth coefficient of the fitted linear regression model to explain a data sample x_i.

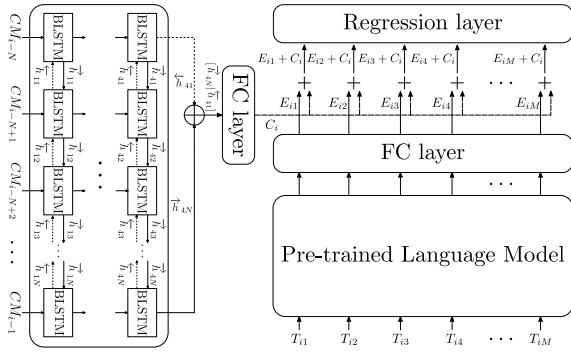[2] Experiments with $N \in [1, 5]$ were performed and $N = 1$ performed best.

Figure 1: Visualization of approach used to integrate information on complexity of preceding language input for sentence i.

5   Results & Discussion

We use sentence-level accuracy (100-MAE) and coefficients of determination ($R2$) as metrics to evaluate the performance of all models. Table 3 shows the evaluation results for all models averaged over all eye-tracking features. Table 3 shows that both BERT and GPT-2 models predicted the eye-tracking features of both datasets with more than 92% accuracy. The fine-tuned models performed consistently better than the pretrained-only ('frozen') models both on the within-domain text (GECO) and on the out-of-domain text (PROVO). This result indicates that the learned representations are general enough to be successfully applied both in the prediction of reading patterns of fiction texts as well as in the prediction of news and popular science texts. The BERT models consistently outperformed the GPT-2 models with a difference in $R2$ of as much as 10.54% on the within-domain data (GECO). This result stands in sharp contrast with those reported in Schrimpf et al. (2021) summarised in Section 2. In their interpretation of the success of GPT-2 in predicting neural activity during reading, Schrimpf et al. (2021) state that "GPT-2 is also arguably the most cognitively plausible of the transformer models (because it uses unidirectional, forward attention)". Especially in view of the remarkable margin by which the BERT models outperformed the GPT-2 models here, it appears that arguments that infer cognitive plausibility from prediction success should be viewed with caution (see also Merkx and Frank (2020) for

Table 3: Model performance across datasets.

Model	Dataset	R2(%)	MAE	Acc
BERT fr	GECO	42.14	7.01	92.99
	PROVO	42.19	6.93	93.61
BERT fr + com S-1	GECO	43.29	6.93	93.07
	PROVO	51.70	5.74	94.26
BERT ft	GECO	56.83	5.95	94.05
	PROVO	67.64	4.51	95.49
BERT ft + com S-1	GECO	**58.36**	**5.92**	**94.08**
	PROVO	**68.59**	**4.49**	**95.51**
GPT-2 fr	GECO	35.00	7.32	92.68
	PROVO	40.15	6.26	93.74
GPT-2 fr + com S-1	GECO	35.19	7.32	92.68
	PROVO	43.67	6.08	93.92
GPT-2 ft	GECO	46.29	6.48	93.52
	PROVO	55.73	5.06	94.94
GPT-2 ft + com S-1	GECO	47.53	6.38	93.62
	PROVO	56.77	5.08	94.92

Note: 'fr' = freeze all layers of language model; 'ft' = the entire model is fine-tuned; '+ com S-1' = including textual features of previous sentence

further intricacies of the issue). The most accurately predicted individual eye-tracking measures were fixation probability (FXP), mean fixation duration (MFD) and first fixation duration (FFD), indicating that prediction accuracy was generally better for early measures than for late measures. A detailed overview of the results for each eye-tracking measure across all models and datasets is provided in Table 7 in the appendix. This finding suggests that the accurate prediction of late measures – that are assumed to reflect higher order processes such as syntactic and semantic integration, revision, and ambiguity resolution – may benefit from the inclusion of contextual information beyond the current sentence.

5.1   Relationship of prediction accuracy and text characteristics

The correlation analyses of the textual features and the mean absolute error revealed that prediction accuracy was affected by the text characteristics of the sentence under consideration. Such effects were found across all eye-tracking met-

rics for both BERT and GPT-2 models in both their frozen and fully fine-tuned variants. For reasons of space, we focus our discussion on the predictions of the BERT frozen model of first pass durations on the GECO dataset (additional results for both frozen and fine-tuned BERT models for both first pass duration and total fixation duration are provided in Figure 3 in the appendix). Figure 2 visualizes the impact of all textual features that reached correlation coefficients $r > |0.2|$ along with the feature group they belong to. As is evident in Figure 2 the prediction accuracy of the BERT frozen model was impacted by features from all five feature groups with individual features affecting prediction accuracy in opposite ways. A strong impact ($r > |0.5|$) was observed for several features of the n-gram feature group: Fixation durations of sentences with higher scores on ngram-frequency features from the news, magazine and spoken registers were predicted more accurately than those with lower scores on these measures. The SMOG readability index, which estimates the years of education a person needs to understand a piece of writing, also has a strong impact: Predicted first pass durations were less accurate in sentences with higher SMOG scores. Several features from the lexical richness, syntactic complexity and readability groups had a moderate impact on prediction accuracy ($|0.3| < r < |0.5|$): For example, predictions of fixation durations were less accurate on sentences of with a more clausal embedding (ClausesPerSentence) and greater lexical sophistication (MeanLengthWord, Sophistication.ANC and Sophistication.BNC). A similar effect was also observed for the psycholinguistic age-of-acquisition features (AoA mean, AoA max), where predictions of fixations times were less accurate for later acquired words. Note that the finding that the correlation coefficients of the readability features have opposite signs is due to the fact that these are either defined to quantify ease of reading (e.g. Flesch Kincaid Reading Ease) or reading difficulty (e.g. SMOG index).
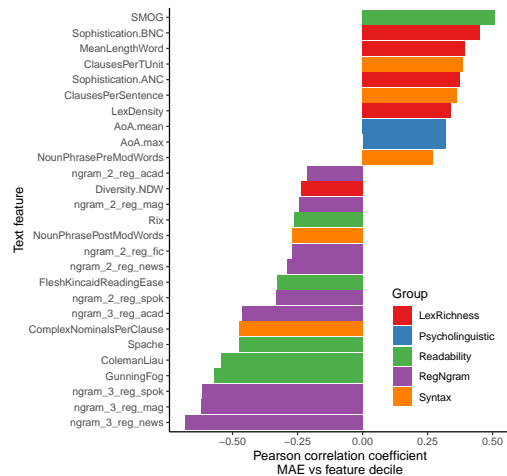

Figure 2: Pearson correlations between model performance (mean absolute error), and the deciles of the respective text characteristics. For measures with negative correlation coefficients, model performance increased with higher values of the text characteristics (Data from 'BERT frozen' predictions of First pass duration (FPD) on GECO testdata).

5.2 Prediction accuracy of hybrid models

Turning to the results of the hybrid models with integrated information on textual characteristics of the preceding sentence, we found that highest accuracy ($R2 = 58.36\%$) was achieved by the fine-tuned BERT model. This amounts to an increase in performance over a model trained without that information of 1.53%. This result demonstrates that future studies should take textual spillover effects into account. Our best-fitting model outperformed not only the best-performing BERT model in Hollenstein et al. (2021), BERT-BASE-MULTILINGUAL-CASED (Wolf et al., 2019) but also the overall best-performing transformer-based model, XLM-MLM-ENDE-1024 (Lample and Conneau, 2019) tested in that study. This result demonstrates that the claim put forth in Hollenstein et al. (2021) that multilingual models show an advantage over language specific ones and that multilingual models might provide cognitively more plausible representations in predicting reading needs to be viewed with caution.

The results of the feature ablation experiments revealed that the main sources of the greater prediction accuracy of the hybrid models was asso-

Table 4: Feature ablation of different models on PROVO dataset. Most important feature groups are bolded.

		Syn. complex	Lex. richness	Psych.	Reg. ngram	Read.
BERT	fr	**5.69**	5.34	3.22	5.44	4.74
BERT	ft	**2.48**	1.64	1.25	1.44	1.30
GPT-2	fr	9.06	9.62	7.91	**10.10**	9.85
GPT-2	ft	16.07	7.02	12.20	**18.82**	12.68

ciated with information concerning the syntactic complexity, lexical richness and n-gram frequency of the preceding sentence. An overview of the results is presented in Table 4. We focus here on the results on the out-of-domain testset (PROVO) for which improvements over models without the integrated textual information were more pronounced. As is evident in Table 4, the central role of the three feature groups listed above result was observed across models (BERT vs. GPT-2) and across training procedures (frozen vs. fine-tuning). However, Table 4 also demonstrates clear differences between the models: While the BERT models show greater sensitivity to syntactic complexity, the GPT-2 models mostly benefit from information concerning n-gram frequency. A possible interpretation of this finding is that a unidirectional model like GPT-2 relies more strongly on word sequencing than a bidirectional one. Future research is needed to examine this in more detail so that effects associated with differences in model architecture can be disentangled.

6 Conclusion

In this paper we conducted the first systematic investigation of the role of general text features in predicting human reading behavior using transformer-based language models (BERT & GPT-2). We have shown (1) that model accuracy is systematically linked to sentence-level text features spanning five measurement categories (syntax, complexity, lexical richness, register-specific N-gram frequency, readability, and psycholinguistic properties), and (2) that prediction accuracy can be improved by using hybrid models that con-

sider spillover effects from the previous sentence.

References

Maria Barrett and Nora Hollenstein. 2020. Sequence labelling and sequence classification with gaze: Novel uses of eye-tracking data for natural language processing. *Language and Linguistics Compass*, 14(11):1–16.

Charles Browne et al. 2013. The new general service list: Celebrating 60 years of vocabulary learning. *The Language Teacher*, 37(4):13–16.

Marc Brysbaert. 2019. How many words do we read per minute? a review and meta-analysis of reading rate. *Journal of Memory and Language*, 109:104047.

Marc Brysbaert, Paweł Mandera, Samantha F McCormick, and Emmanuel Keuleers. 2019. Word prevalence norms for 62,000 english lemmas. *Behavior research methods*, 51(2):467–479.

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(2):602–615.

Mark Davies. 2008. The Corpus of Contemporary American English (COCA): 560 million words, 1990-present.

Vera Demberg and Frank Keller. 2008. Data from eye-tracking corpora as evidence for theories of syntactic processing complexity. *Cognition*, 109(2):193–210.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Ralf Engbert, Antje Nuthmann, Eike M Richter, and Reinhold Kliegl. 2005. Swift: a dynamical model of saccade generation during reading. *Psychological review*, 112(4):777.

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

Edward Fry. 1968. A readability formula that saves time. *Journal of reading*, 11(7):513–578.

Adam Goodkind and Klinton Bicknell. 2018. Predictive power of word surprisal for reading times is a linear function of language model quality. In *Proceedings of the 8th workshop on cognitive modeling and computational linguistics (CMCL 2018)*, pages 10–18.

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.

Jeremy Howard and Sebastian Ruder. 2018. Universal language model fine-tuning for text classification. *arXiv preprint arXiv:1801.06146*.

Brendan T Johns, Melody Dye, and Michael N Jones. 2020. Estimating the prevalence and diversity of words in written language. *Quarterly Journal of Experimental Psychology*, 73(6):841–855.

Elma Kerz, Yu Qiao, and Daniel Wiechmann. 2021. Language that captivates the audience: Predicting affective ratings of TED talks in a multi-label classification task. In *Proceedings of the Eleventh Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 13–24, Online. Association for Computational Linguistics.

Elma Kerz, Yu Qiao, Daniel Wiechmann, and Marcus Ströbel. 2020. Becoming linguistically mature: Modeling english and german children's writing development across school grades. In *Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 65–74.

J Peter Kincaid, Robert P Fishburne Jr, Richard L Rogers, and Brad S Chissom. 1975. Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel. Technical report, Naval Technical Training Command Millington TN Research Branch.

Gina R Kuperberg and T Florian Jaeger. 2016. What do we mean by prediction in language comprehension? *Language, cognition and neuroscience*, 31(1):32–59.

Victor Kuperman, Hans Stadthagen-Gonzalez, and Marc Brysbaert. 2012. Age-of-acquisition ratings for 30,000 english words. *Behavior research methods*, 44(4):978–990.

Guillaume Lample and Alexis Conneau. 2019. Crosslingual language model pretraining. *arXiv preprint arXiv:1901.07291*.

Roger Levy and Galen Andrew. 2006. Tregex and tsurgeon: tools for querying and manipulating tree data structures. In *LREC*, pages 2231–2234. Citeseer.

Matthew W Lowder, Wonil Choi, Fernanda Ferreira, and John M Henderson. 2018. Lexical predictability during natural reading: Effects of surprisal and entropy reduction. *Cognitive science*, 42:1166–1183.

Xiaofei Lu. 2010. Automatic analysis of syntactic complexity in second language writing. *International journal of corpus linguistics*, 15(4):474–496.

Xiaofei Lu. 2012. The relationship of lexical richness to the quality of esl learners' oral narratives. *The Modern Language Journal*, 96(2):190–208.

Steven G Luke and Kiel Christianson. 2018. The provo corpus: A large eye-tracking corpus with predictability norms. *Behavior research methods*, 50(2):826–833.

Christopher Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven Bethard, and David McClosky. 2014. The stanford corenlp natural language processing toolkit. In *Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations*, pages 55–60.

G Harry McLaughlin. 1969. Clearing the smog. *Journal of Reading*.

Danny Merkx and Stefan L Frank. 2020. Human sentence processing: Recurrence or attention? *arXiv preprint arXiv:2005.09471*.

Alexander Pollatsek, Barbara J Juhasz, Erik D Reichle, Debra Machacek, and Keith Rayner. 2008. Immediate and delayed effects of word frequency and word length on eye movements in reading: A reversed delayed effect of word length. *Journal of experimental psychology: human perception and performance*, 34(3):726.

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

Keith Rayner, Alexander Pollatsek, Jane Ashby, and Charles Clifton Jr. 2012. Psychology of reading.

Erik D Reichle, Alexander Pollatsek, Donald L Fisher, and Keith Rayner. 1998. Toward a model of eye movement control in reading. *Psychological review*, 105(1):125.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. " why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144.

Martin Schrimpf, Idan Asher Blank, Greta Tuckute, Carina Kauf, Eghbal A. Hosseini, Nancy Kanwisher, Joshua B. Tenenbaum, and Evelina Fedorenko. 2021. The neural architecture of language: Integrative modeling converges on predictive processing. *Proceedings of the National Academy of Sciences*, 118(45).

Mark Seidenberg. 2017. *Language at the Speed of Sight: How We Read, Why So Many Can't, and What Can Be Done About It*. Basic Books.

Michael Shvartsman, Richard L Lewis, and Satinder Singh. 2014. Computationally rational saccadic control: An explanation of spillover effects based on sampling from noisy perception and memory. In *Proceedings of the Fifth Workshop on Cognitive Modeling and Computational Linguistics*, pages 1–9.

Joshua Snell and Jan Theeuwes. 2020. A story about statistical learning in a story: Regularities impact eye movements during book reading. *Journal of Memory and Language*, 113:104127.

Adrian Staub. 2015. The effect of lexical predictability on eye movements in reading: Critical review and theoretical interpretation. *Language and Linguistics Compass*, 9(8):311–327.

Marcus Ströbel. 2014. *Tracking complexity of l2 academic texts: A sliding-window approach*. Master thesis. RWTH Aachen University.

Marcus Ströbel, Elma Kerz, Daniel Wiechmann, and Stella Neumann. 2016. Cocogen-complexity contour generator: Automatic assessment of linguistic complexity using a sliding-window technique. In *Proceedings of the Workshop on Computational Linguistics for Linguistic Complexity (CL4LC)*, pages 23–31.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2019. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.

A Appendix

Table 5: Overview of the 107 features investigated in the work

| Feature group	Number of features	Features	Example/Description
Syntactic complexity	16	MLC	Mean length of clause (words)
		MLS	Mean length of sentence (words)
		MLT	Mean length of T-unit (words)
		C/S	Clauses per sentence
		C/T	Clauses per T-unit
		DepC/C	Dependent clauses per clause
		T/S	T-units per sentence
		CompT/T	Complex T-unit per T-unit
		DepC/T	Dependent Clause per T-unit
		CoordP/C	Coordinate phrases per clause
		CoordP/T	Coordinate phrases per T-unit
		NP.PostMod	NP post-mod (word)
		NP.PreMod	NP pre-mod (word)
		CompN/C	Complex nominals per clause
		CompN/T	Complex nominals per T-unit
		VP/T	Verb phrases per T-unit
Lexical richness	14	MLWc	Mean length per word (characters)
		MLWs	Mean length per word (sylables)
		LD	Lexical density
		NDW	Number of different words
		CNDW	NDW corrected by Number of words
		TTR	Type-Token Ration (TTR)
		cTTR	Corrected TTR
		rTTR	Root TTR
		AFL	Sequences Academic Formula List
		ANC	LS (ANC) (top 2000, inverted)
		BNC	LS (BNC) (top 2000, inverted)
		NAWL	LS New Academic Word List
		NGSL	LS (General Service List) (inverted)
		NonStopWordsRate	Ratio of words in NLTK non-stopword list
Register-based	25	Spoken ($n \in [1,5]$)	Frequencies of uni-, bi-
		Fiction ($n \in [1,5]$)	tri-, four-, five-grams
		Magazine ($n \in [1,5]$)	from the five sub-components
		News ($n \in [1,5]$)	(genres) of the COCA,
		Academic ($n \in [1,5]$)	see Davies (2008)

Table 6: Overview of the 107 features investigated in the work(Cont.

Feature group	Number of features	Features	Example/Description
Readability	14	ARI	Automated Readability Index
		ColemanLiau	Coleman-Liau Index
		DaleChall	Dale-Chall readability score
		FleshKincaidGradeLevel	Flesch-Kincaid Grade Level
		FleshKincaidReadingEase	Flesch Reading Ease score
		Fry-x	x coord. on Fry Readability Graph
		Fry-y	y coord. on Fry Readability Graph
		Lix	Lix readability score
		SMOG	Simple Measure of Gobbledygook
		GunningFog	Gunning Fog Index readability score
		DaleChallPSK	Powers-Sumner-Kearl Variation of the Dale and Chall Readability score
		FORCAST	FORCAST readability score
		Rix	Rix readability score
		Spache	Spache readability score
Psycholinguistic	38	WordPrevalence	See Brysbaert et al. (2019)
		Prevalence	Word prevalence list incl. 35 categories (Johns et al. (2020))
		AoA-mean	avg. age of acquisition (Kuperman et al. (2012))
		AoA-max	max. age of acquisition

Table 7: Model performance by eye-tracking feature across datasets

| model | dataset | R2(%) | | | | | | | | mean |
		NFX	FFD	FPD	TFD	MFD	FXP	NRFX	RRDP	R2(%)
BERT frozen	GECO dev	46.38	44.50	46.22	45.21	44.62	46.01	31.80	34.50	42.40
	GECO test	46.99	42.60	45.34	45.05	42.94	44.91	33.68	35.61	42.14
	PROVO	42.91	50.28	46.11	42.83	48.99	44.76	29.67	31.95	42.19
BERT frozen + complexity S-1	GECO dev	46.99	46.31	46.81	45.81	46.42	48.65	31.79	35.05	43.48
	GECO test	47.76	44.36	45.96	45.78	44.80	47.80	33.79	36.09	43.29
	PROVO	52.20	61.44	54.03	50.12	60.15	61.53	33.96	40.19	51.70
BERT fine-tuned	GECO dev	60.89	56.98	58.64	59.15	57.15	60.60	47.60	51.11	56.51
	GECO test	61.67	56.47	58.28	59.74	57.10	60.62	49.09	51.67	56.83
	PROVO	68.81	74.80	68.20	65.86	74.93	78.06	53.01	57.46	67.64
BERT fine-tuned + complexity S-1	GECO dev	62.50	57.89	60.85	60.47	58.09	61.14	47.61	50.03	57.32
	GECO test	**64.17**	**57.66**	**61.59**	**61.83**	**58.20**	**61.27**	**50.14**	**52.00**	**58.36**
	PROVO	**70.49**	**75.39**	**70.05**	**67.16**	**75.21**	**77.60**	**52.54**	**60.27**	**68.59**
GPT-2 frozen	GECO dev	41.06	40.69	41.26	39.54	40.83	42.73	25.94	29.30	37.67
	GECO test	38.01	38.08	38.69	36.08	38.19	40.55	23.43	26.98	35.00
	PROVO	38.40	51.14	43.17	38.36	50.06	47.19	23.43	29.41	40.15
GPT-2 frozen + complexity S-1	GECO dev	41.02	41.14	41.16	39.46	41.27	43.97	25.27	29.28	37.82
	GECO test	37.98	38.55	38.56	36.08	38.66	41.81	22.70	27.19	35.19
	PROVO	43.09	54.07	45.78	41.77	52.65	53.37	27.49	31.14	43.67
GPT-2 fine-tuned	GECO dev	52.17	51.63	51.83	49.97	51.79	55.68	33.36	37.70	48.02
	GECO test	50.65	49.69	49.71	47.86	49.97	54.13	32.24	36.09	46.29
	PROVO	55.02	67.48	56.59	52.43	66.64	68.82	35.56	43.27	55.73
GPT-2 fine-tuned + complexity S-1	GECO dev	54.46	53.34	53.91	52.21	53.69	57.20	35.11	39.63	49.94
	GECO test	**51.91**	**50.97**	**51.24**	**49.30**	**51.28**	**55.02**	**33.11**	**37.42**	**47.53**
	PROVO	**56.19**	**68.20**	**58.44**	**53.98**	**67.84**	**68.79**	**35.81**	**44.95**	**56.77**

Note: 'frozen' = all the layers of the language model are frozen and only the attached neural network layers are trained on the GECO dataset; the weights of only the attached layers will be updated during model training. 'fine-tuned' = the entire pretrained model is fine-tuned on the GECO training set; the error is back-propagated through the entire architecture and the pre-trained weights of the model are updated based on the GECO training set. Best-performing models on the two testsets (GECO test, PROVO) are highlighted in bold.
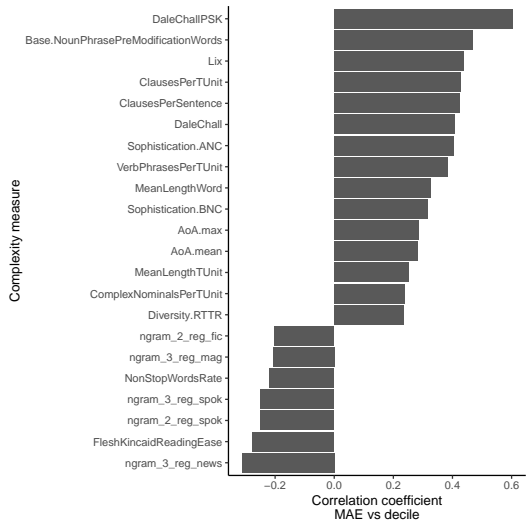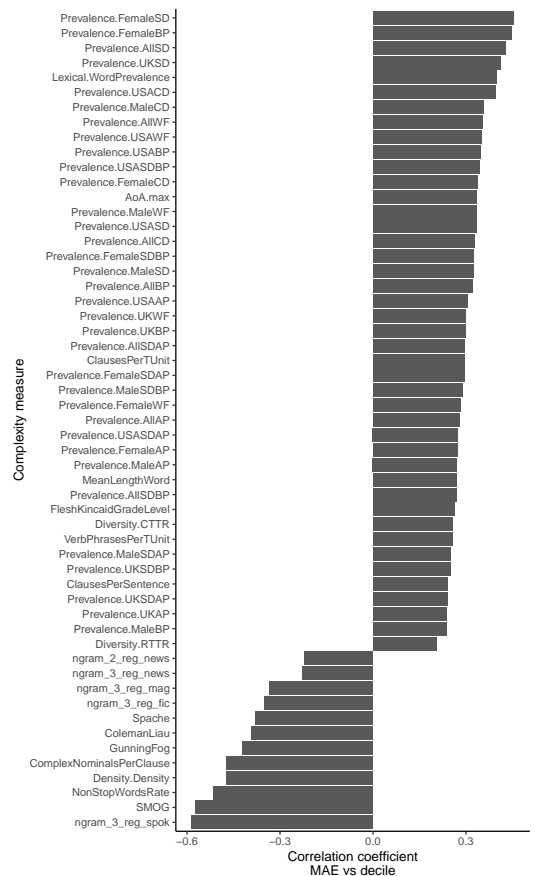
(a) Model: BERT frozen. Eye-tracking metric: Total fixation duration (TFD)
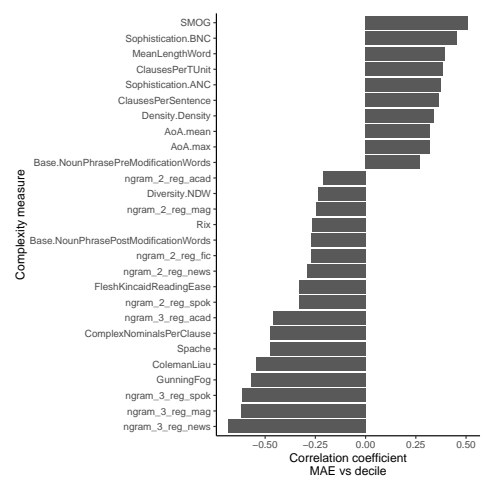
(b) Model: BERT frozen. Eye-tracking metric: First pass duration (FPD)
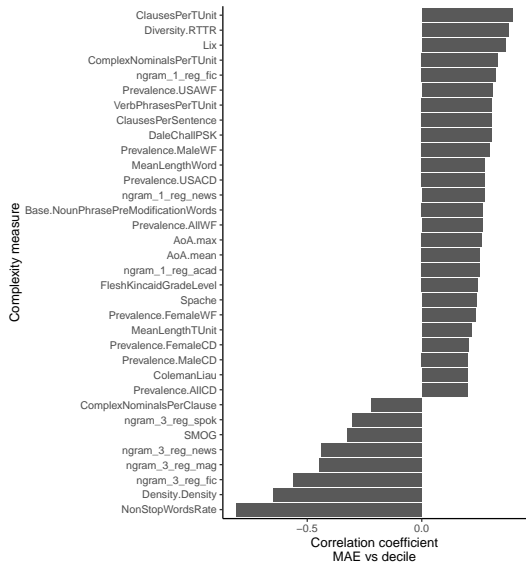
(c) Model: BERT fine-tuned. Eye-tracking metric: Total fixation duration (TFD)

(d) Model: BERT fine-tuned. Eye-tracking metric: First pass duration (FPD)

Figure 3: Pearson correlations between model performance (Mean Absolute Error), and the deciles of the respective text characteristics. For measures with negative correlation coefficients, model performance increased with higher values of the text characteristics.