Multilingual Language Models Predict Human Reading Behavior

Hollenstein, Nora; Pirovano, Federico; Zhang, Ce; Jäger, Lena; Beinborn, Lisa

published in
Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies

2021

DOI (link to publisher)
10.18653/v1/2021.naacl-main.10

document version
Peer reviewed version

Multilingual Language Models Predict Human Reading Behavior

Nora Hollenstein[1], Federico Pirovano[1], Ce Zhang[1], Lena Jäger[2,3], Lisa Beinborn[4]
[1] ETH Zurich
[2] University of Zurich
[3] University of Potsdam
[4] Vrije Universiteit Amsterdam
{noraho,fpirovan,ce.zhang}@inf.ethz.ch,
jaeger@cl.uzh.ch,l.beinborn@vu.nl

Abstract

We analyze if large language models are able to predict patterns of human reading behavior. We compare the performance of language-specific and multilingual pretrained transformer models to predict reading time measures reflecting natural human sentence processing on Dutch, English, German, and Russian texts. This results in accurate models of human reading behavior, which indicates that transformer models implicitly encode relative importance in language in a way that is comparable to human processing mechanisms. We find that BERT and XLM models successfully predict a range of eye tracking features. In a series of experiments, we analyze the cross-domain and cross-language abilities of these models and show how they reflect human sentence processing.

1 Introduction

When processing language, humans selectively attend longer to the most relevant elements of a sentence (Rayner, 1998). This ability to seamlessly evaluate relative importance is a key factor in human language understanding. It remains an open question how relative importance is encoded in computational language models. Recent analyses conclude that the cognitively motivated "attention" mechanism in neural models is not a good indicator for relative importance (Jain and Wallace, 2019). Alternative methods based on salience (Bastings and Filippova, 2020), vector normalization (Kobayashi et al., 2020), or subset erasure (De Cao et al., 2020) are being developed to increase the post-hoc interpretability of model predictions but the cognitive plausibility of the underlying representations remains unclear.

In human language processing, phenomena of relative importance can be approximated indirectly by tracking eye movements and measuring fixation duration (Rayner, 1977). It has been shown that fixation duration and relative importance of text segments are strongly correlated in natural reading, so that direct links can be established on the token level (Malmaud et al., 2020). In the example in Figure 1, the newly introduced entity *Mary French* is fixated twice and for a longer duration because it is relatively more important for the reader than the entity *Laurence*, which had been introduced in the previous sentence. Being able to reliably predict eye movement patterns from the language input would bring us one step closer to understand the cognitive plausibility of these models.

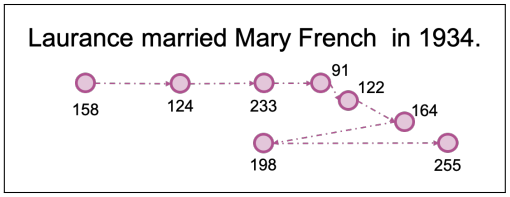

Figure 1: From the fixation times in milliseconds of a single subject in the ZuCo 1.0 dataset, the feature vector described in Section 3.2 for the wors "Mary" would be $[2, 233, 233, 431, 215.5, 1, 1, 1]$.

Contextualized neural language models are less interpretable than conceptually motivated psycholinguistic models but they achieve high performance in many language understanding tasks and can be fitted successfully to cognitive features such as self-paced reading times and N400 strength (Merkx and Frank, 2020). Moreover, approaches to directly predict cognitive signals (e.g., brain activity) indicate that neural representations implicitly encode similar information as humans (Wehbe et al., 2014; Abnar et al., 2019; Sood et al., 2020; Schrimpf et al., 2020). However, it has not been analyzed to which extent transformer language models are able to directly predict human behavioral metrics such as gaze patterns.

The performance of computational models can

be improved even further if their inductive bias is adjusted using human cognitive signals such as eye tracking, fMRI, or EEG data (Hollenstein et al., 2019; Toneva and Wehbe, 2019; Takmaz et al., 2020). While psycholinguistic work mainly focuses on very specific phenomena of human language processing that are typically tested in experimental settings with constructed stimuli (Hale, 2017), we focus on directly generating token-level predictions from natural reading.

We fine-tune transformer models on human eye movement data and analyze their ability to predict human reading behavior focusing on a range of reading features, datasets, and languages. We compare the performance of monolingual and multilingual transformer models. Multilingual models represent multiple languages in a joint space and aim at a more universal language understanding. As eye tracking patterns are consistent across languages for certain phenomena, we hypothesize that multilingual models might provide cognitively more plausible representations and outperform language-specific models in predicting reading measures. We test this hypothesis on 6 datasets of 4 Indo-European languages, namely English, German, Dutch and Russian.[1]

We find that pretrained transformer models are surprisingly accurate at predicting reading time measures in four Indo-European languages. Multilingual models show an advantage over language-specific models, especially when fine-tuned on smaller amounts of data. Compared to previous psycholinguistic reading models, the accuracy achieved by the transformer models is remarkable. Our results indicate that transformer models implicitly encode relative importance in language in a way that is comparable to human processing mechanisms. As a consequence, it should be possible to adjust the inductive bias of neural models towards more cognitively plausible outputs without having to resort to large-scale cognitive datasets.

2 Related Work

Using eye movement data to modify the inductive bias of language processing models has resulted in improvements for several NLP tasks (e.g., Barrett et al. 2016; Hollenstein and Zhang 2019). It has also been used as a supervisory signal in multi-task learning scenarios (Klerke et al., 2016; Gonzalez-Garduno and Søgaard, 2017) and as a method to fine-tune the attention mechanism (Barrett et al., 2018). We use eye tracking data to evaluate how well transformer language models predict human sentence processing. Therefore, in this section, we discuss previous work on probing transformers models as well as on modelling human sentence processing.

2.1 Probing Transformer Language Models

Contextualized neural language models have become increasingly popular, but our understanding of these black box algorithms is still rather limited (Gilpin et al., 2018). Current intrinsic evaluation methods do not capture the cognitive plausibility of language models (Manning et al., 2020; Gladkova and Drozd, 2016). In previous work of interpreting and probing language models, human behavioral data as well as neuroimaging recordings have been leveraged to understand the inner workings of the neural models. For instance, Ettinger (2020) explores the linguistic capacities of BERT with a set of psycholinguistic diagnostics. Toneva and Wehbe (2019) propose an interpretation approach by learning alignments between the models and brain activity recordings (MEG and fMRI). Hao et al. (2020) propose to evaluate language model quality based on the degree to which they exhibit human-like behavior such as predictability measures collected from human subjects. However, their metric does not reveal any details about the commonalities between the model and human sentence processing.

The benefits of multilingual models are controversial. Transformer models trained exclusively on a specific language often outperform multilingual models trained on various languages simultaneously, even after fine-tuning. This *curse of multilinguality* (Conneau et al., 2020; Vulić et al., 2020) has been shown for Spanish (Canete et al., 2020), Finnish (Virtanen et al., 2019) and Dutch (Vries et al., 2019). In this paper we investigate whether a similar effect can be observed when leveraging these models to predict human behavioral measures, or whether in that case the multilingual models provide more plausible representations of human reading due to the common eye tracking effects across languages.

2.2 Modelling Human Sentence Processing

Previous work of neural modelling of human sentence processing has focused on recurrent neural networks, since their architecture and learn-

[1] Code available on GitHub: https://github.com/DS3Lab/multilingual-gaze

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 1: Descriptive statistics of all eye tracking datasets.[2] 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) (see Appendix B).

ing mechanism appears to be cognitively plausible (Keller, 2010; Michaelov and Bergen, 2020). However, recent work suggests that transformers perform better at modelling certain aspects of the human language understanding process (Hawkins et al., 2020). While Merkx and Frank (2020) and Wilcox et al. (2020) show that the psychometric predictive power of transformers outperforms RNNs on eye tracking, self-paced reading times and N400 strength, they do not directly predict cognitive features. Schrimpf et al. (2020) show that contextualized monolingual English models accurately predict language processing in the brain.

Context effects are known to influence fixations times during reading (Morris, 1994). The notion of using contextual information to process language during reading has been well-established in psycholinguistics (e.g., Inhoff and Rayner 1986 and Jian et al. 2013). However, to the best of our knowledge, we are the first to study to which extent the representations learned by transformer language models entail these human reading patterns.

Compared to neural models of human sentence processing, we predict not only individual metrics but a range of eye tracking features covering the full reading process from early lexical access to late syntactic processing. By contrast, most models of reading focus on predicting skipping probability (Reichle et al., 1998; Matthies and Søgaard, 2013; Hahn and Keller, 2016). Sood et al. (2020) propose a text saliency model which predicts fixation durations that are then used to compute the attention scores in a transformer network.

3 Data

We predict eye tracking data only from naturalistic reading studies in which the participants read full sentences or longer spans of naturally occurring text in their own speed. The data from these studies exhibit higher ecological validity than studies which rely on artificially constructed sentences and paced presentation (Alday, 2019).

3.1 Corpora

To conduct a cross-lingual comparison, we use eye tracking data collected from native speakers of four languages (see Table 1 for details).

English The largest number of eye tracking data sources are available for English. We use eye tracking features from three English corpora: (1) The Dundee corpus (Kennedy et al., 2003) contains 20 newspaper articles from *The Independent*, which were presented to English native readers on a screen five lines at a time. (2) The GECO corpus (Cop et al., 2017) contains eye tracking data from English monolinguals reading the entire novel *The Mysterious Affair at Styles* by Agatha Christie. The text was presented on the screen in paragraphs. (3) The ZuCo corpus (Hollenstein et al., 2018, 2020) includes eye tracking data of full sentences from movie reviews and Wikipedia articles.[3]

Dutch The GECO corpus (Cop et al., 2017) additionally contains eye tracking data from Dutch readers, which were presented with the same novel in their native language.

German The Potsdam Textbook Corpus (PoTeC, Jäger et al. 2021) contains 12 short passages of 158 words on average from college-level biology and physics textbooks, which are read by expert and laymen German native speakers. The full passages were presented on multiple lines on the screen.

[2]Note that the exact numbers might differ slightly from the original publications due to different preprocessing methods.

[3]We use Tasks 1 and 2 from ZuCo 1.0 and Task 1 from ZuCo 2.0.

Short Name	Language	Model Checkpoint	Reference
BERT-NL	Dutch	WIETSEDV/BERT-BASE-DUTCH-CASED	Vries et al. (2019)
BERT-EN	English	BERT-BASE-UNCASED	Wolf et al. (2019)
BERT-DE	German	BERT-BASE-GERMAN-CASED	Chan et al. (2019)
BERT-RU	Russian	DEEPPAVLOV/RUBERT-BASE-CASED	Yu and Arkhipov (2019)
BERT-MULTI	104 languages	BERT-BASE-MULTILINGUAL-CASED	Wolf et al. (2019)
XLM-EN	English	XLM-MLM-EN-2048	Lample and Conneau (2019)
XLM-ENDE	English + German	XLM-MLM-ENDE-1024	Lample and Conneau (2019)
XLM-17	17 languages	XLM-MLM-17-1280	Lample and Conneau (2019)
XLM-100	100 languages	XLM-MLM-100-1280	Lample and Conneau (2019)

Table 2: Pretrained transformer language models analyzed in this work.

Russian The Russian Sentence Corpus (RSC, Laurinavichyute et al. 2019) contains 144 naturally occurring sentences extracted from the Russian National Corpus.[4] Full sentences were presented on the screen to monolingual Russian-speaking adults one at a time.

3.2 Eye Tracking Features

A fixation is defined as the period of time where the gaze of a reader is maintained on a single location. Fixations are mapped to words by delimiting the boundaries around the region on the screen belonging to each word *w*. A word can be fixated more than once. For each token *w* in the input text, we predict the following eight eye tracking features that encode the full reading process from early lexical access up to subsequent syntactic integration.

Word-level characteristics We extract basic features that encode *word-level* characteristics: (1) number of fixations (NFIX), the number of times a subject fixates *w*, averaged over all subjects; (2) mean fixation duration (MFD), the average fixation duration of all fixations made on *w*, averaged over all subjects; (3) fixation proportion (FPROP), the number of subjects that fixated *w*, divided by the total number of subjects.

Early processing We also include features to capture the *early* lexical and syntactic processing, based on the first time a word is fixated: (4) first fixation duration (FFD), the duration, in milliseconds, of the first fixation on *w*, averaged over all subjects; (5) first pass duration (FPD), the sum of all fixations on *w* from the first time a subject fixates *w* to the first time the subject fixates another token, averaged over all subjects.

Late processing Finally, we also use measures reflecting the *late* syntactic processing and general disambiguation, based on words which were fixated more than once: (6) total reading time (TRT), the sum of the duration of all fixations made on *w*, averaged over all subjects; (7) number of re-fixations (NREFIX), 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; (8) re-read proportion (REPROP), the number of subjects that fixated *w* more than once, divided by the total number of subjects.

The values of these eye tracking features vary over different ranges (see Appendix A). 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.

Eye movements depend on the stimulus and are therefore language-specific but there exist universal tendencies which remain stable across languages (Liversedge et al., 2016). For example, the average fixation duration in reading ranges from 220 to 250 ms independent of the language. Furthermore, word characteristics such as word length, frequency and predictability affect fixation duration similarly across languages but the effect size depends on the language and the script (Laurinavichyute et al., 2019; Bai et al., 2008). The word length effect, i.e., the fact that longer words are more likely to be fixated, can be observed across all four languages included in this work (see Appendix A).

4 Language Models

We compare the ability to predict eye tracking features in two models: BERT and XLM. Both models are trained on the transformer architecture (Vaswani et al., 2017) and yield state-of-the-

[4] https://ruscorpora.ru

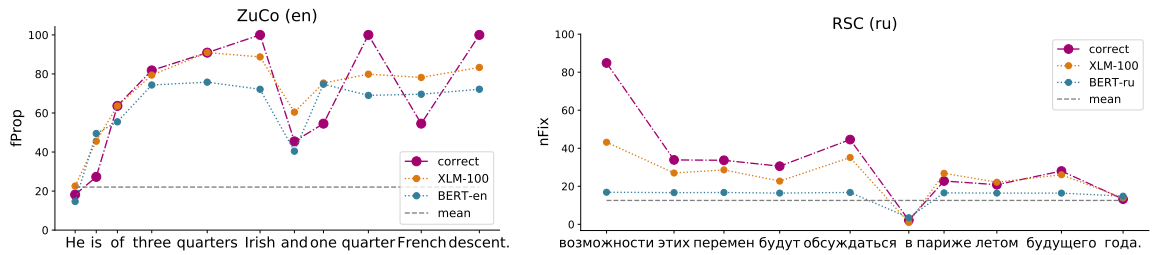

Figure 2: True and predicted feature values for two example sentences. On the left the fixation proportion (FPROP) values for an English sentence from the ZuCo dataset, and on the right the number of fixations (NFIX) values for a Russian sentence from the RSC dataset.

art results for a wide range of NLP tasks (Liang et al., 2020). The multilingual BERT model simply concatenates the Wikipedia input from 104 languages and is optimized by performing masked token and next sentence prediction as in the monolingual model (Devlin et al., 2019) without any cross-lingual constraints. In contrast, XLM adds a translation language modeling objective, by explicitly using parallel sentences in multiple languages as input to facilitate cross-lingual transfer (Lample and Conneau, 2019). Both BERT and XLM use subword tokenization methods to build shared vocabulary spaces across languages.

We use the pretrained checkpoints from the HuggingFace repository for monolingual and multilingual models (details in Table 2).[5]

5 Method

We fine-tune the models described above on the features extracted from the eye tracking datasets. The eye tracking prediction uses a model for token regression, i.e., the pretrained language models with a linear dense layer on top of it. The final dense layer is the same for all tokens, and performs a projection from the dimension of the hidden size of the model (e.g., 768 for BERT-EN or 1,280 for XLM-100) to the dimension of the eye tracking feature space (8, in our case). The model is trained for the regression task using the *mean squared error* (MSE) loss.

Training Details We split the data into 90% training data, 5% validation and 5% test data. We initially tuned the hyper-parameters manually and set the following values for all models: We use an AdamW optimizer (Loshchilov and Hutter, 2018) with a learning rate of 0.00005 and a weight decay of 0.01. The batch size varies depending on the model dimensions (see Appendix C.2). We employ a linear learning rate decay schedule over the total number of training steps. We clip all gradients exceeding the maximal value of 1. We train the models for 100 epochs, with early stopping after 7 epochs without an improvement on the validation accuracy.

Evaluation Procedure As the features have been standardized to the range 0–100, the *mean absolute error* (MAE) can be interpreted as a percentage error. For readability, we report the *prediction accuracy* as 100−MAE in all experiments. The results are averaged over batches and over 5 runs with varying random seeds. For a single batch of sentences, the overall MAE is calculated by concatenating the words in each sentence and the feature dimensions for each word, and padding to the maximum sentence length. The per-feature MAE is calculated by concatenating the words in each sentence. For example, for a batch of B sentences, each composed of L words, and G eye tracking features per word, the overall MAE is calculated over a vector of $B*L*G$ dimensions. In contrast, the MAE for each individual feature is calculated over a vector of $B*L$ dimensions.

6 Results & Discussion

Tables 3 and 4 show that all models predict the eye tracking features with more than 90% accuracy for English and Dutch. For English, the BERT models yield high performance on all three datasets with standard deviations below 0.15. The results for the XLM models are slightly better on average but exhibit much higher standard deviations. Similar to the results presented by Lample and Conneau (2019), we find that more training data from *multiple* languages improves prediction performance. For instance, the XLM-100 model achieves higher accuracy than the XLM-17 model in all cases. For

[5] https://huggingface.co/transformers/pretrained_models.html

Model	Dundee (en)	GECO (en)	ZuCo (en)	ALL (en)
BERT-EN	92.63 (0.05)	93.68 (0.14)	93.42 (0.02)	93.71 (0.06)
BERT-MULTI	92.73 (0.06)	93.73 (0.12)	93.74 (0.05)	93.74 (0.07)
XLM-EN	90.41 (2.16)	91.15 (1.42)	92.03 (2.11)	90.88 (1.50)
XLM-ENDE	92.79 (0.15)	**93.89** (0.12)	93.76 (0.15)	**93.96** (0.08)
XLM-17	92.11 (1.68)	91.79 (1.75)	92.05 (2.25)	93.80 (0.38)
XLM-100	**92.99** (0.05)	93.04 (1.40)	**93.97** (0.09)	**93.96** (0.06)

Table 3: Prediction accuracy over all eye tracking features for the English corpora, including the concatenated dataset. Standard deviation is reported in parentheses.

Model	GECO (nl)	PoTeC (de)	RSC (ru)	ALL-LANGS
BERT-NL	91.81 (0.23)	–	–	–
BERT-DE	–	78.38 (1.69)	–	–
BERT-RU	–	–	78.73 (1.38)	–
BERT-MULTI	91.90 (0.16)	76.86 (2.42)	76.54 (3.59)	94.72 (0.07)
XLM-ENDE	–	80.94 (0.88)	–	–
XLM-17	91.04 (0.70)	86.26 (1.31)	90.96 (3.96)	94.46 (0.83)
XLM-100	**92.31** (0.22)	**86.57** (0.54)	**94.70** (0.60)	**94.94** (0.11)

Table 4: Prediction accuracy over all eye tracking features for the Dutch, German and Russian corpora, and for all four languages combined in a single dataset. Standard deviation is reported in parentheses.

the smaller non-English datasets, PoTeC (de) and RSC (ru), the multilingual XLM models clearly outperform the monolingual models. For the English datasets, the differences are minor.

Size Effects More training data results in higher prediction accuracy even when the eye tracking data comes from various languages and was recorded in different reading studies by different devices (ALL-LANGS, fine-tuning on the data of all four languages together). However, merely adding more data from the same language (ALL (en), fine-tuning on the English data from Dundee, GECO and ZuCo together) does not result in higher performance.

To analyze this further, we perform an ablation study on varying amounts of training data. The results are shown in Figure 3 for Dutch and English. The performance of the XLM models remains stable even with a very small percentage of eye tracking data. The performance of the BERT models, however, drops drastically when fine-tuning on less than 20% of the data. Similar to Merkx and Frank (2020) and Hao et al. (2020) we find that the model architecture, along with the composition and size of the training corpus have a significant impact on the psycholinguistic modeling performance.

Eye Tracking Features The accuracy results are averaged over all eye tracking features. For a better understanding of the prediction output, we plot the true and the predicted values of two selected features (FPROP and NFIX) for two example sentence in Figure 2. In both examples, the model predictions strongly correlate with the true values. The difference to the mean baseline is more pronounced for the FIXPROPfeature.

Figure 4 presents the quantitative differences across models in predicting the individual eye tracking features.[6] Across all datasets, first pass duration (FPD) and number of re-fixations (NREFIX) are the most accurately predicted features. Proportions (FPROP and REPROP) are harder to predict because these features are even more dependent on subject-specific characteristics. Nevertheless, when comparing the prediction accuracy of each eye tracking feature to a baseline which always predicts the mean values, the predicted features FPROP and REPROP achieve the largest improvements relative to the mean baseline. See Figure 5 for a comparison between all features for the best performing model XLM-100 on all six datasets.

Performance of Pretrained Models To test the language models' abilities on predicting human reading behavior only from pretraining on textual input, we take the provided model checkpoints and use them to predict the eye tracking features without any fine-tuning. The detailed results are presented in Appendix D.1. The achieved accuracy aggregated over all eye tracking features lies between 75-78% for English. For Dutch, the models achieve

[6] Plots for the remaining datasets are in Appendix D.2

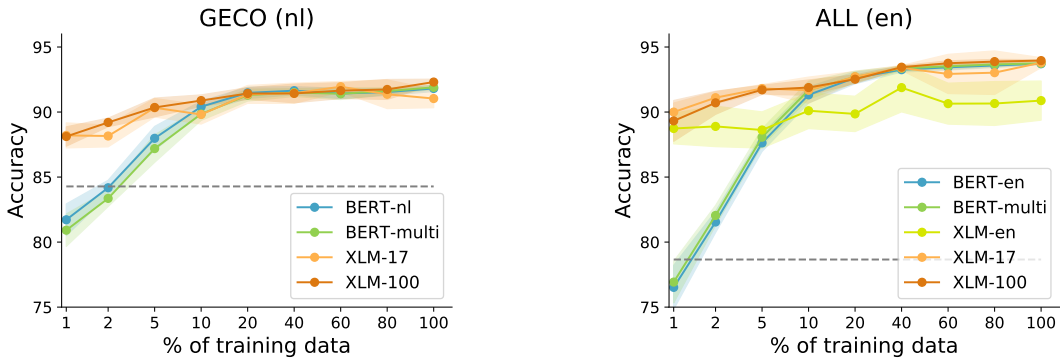

Figure 3: Data ablation study for Dutch and English. The results are aggregated over all eye tracking features. In addition to the mean across five runs, the shaded areas represent the standard deviation. The dashed line is the result of the pretrained BERT-MULTI model without fine-tuning. Results are aggregated over all eye tracking features.

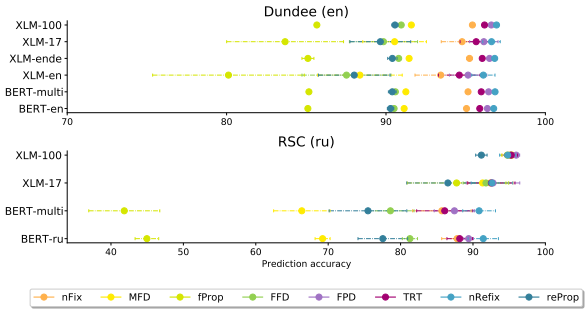

Figure 4: Results of individual eye tracking features for all models on the Dundee and RSC corpora.

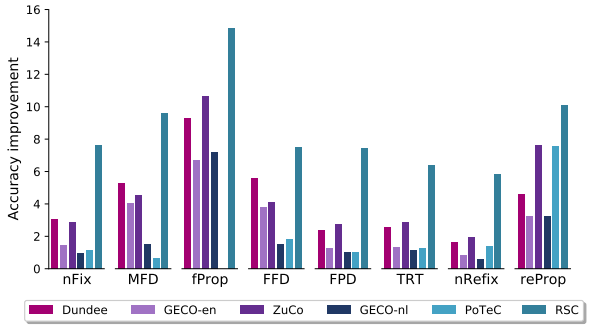

Figure 5: Improvement of prediction accuracy for the XLM-100 model relative to the mean baseline for each eye tracking feature.

84% accuracy but for Russian merely 65%. Across the same languages the results between the different language models are only minimal. However, on the individual eye tracking features, the pretrained models do not achieve any improvements over the mean baseline (see Appendix D.1).

7 Data Sensitivity

For the main experiment, we always tested the models on held-out data from the same dataset. In this section, we examine the influence of dataset properties (text domain and language) on the prediction accuracy. In a second step, we analyze the influence of more universal input characteristics (word length, text readability).

7.1 Cross-Domain Evaluation

Figure 6 shows the results when evaluating the eye tracking predictions on out-of-domain text for the English datasets. For instance, we fine-tune the model on the newspaper articles of the Dundee corpus and test on the literary novel of the GECO corpus. We can see that the overall prediction accuracy across all eye tracking features is constantly above. 90% in all combinations. This shows that our eye tracking prediction model is able to generalize across domains. We find that the cross-domain capabilities of BERT are slightly better than for XLM. BERT-EN performs best in the cross-domain evaluation, possibly because its training data is more domain-general since it includes text from Wikipedia and books.

7.2 Cross-Language Evaluation

Figure 7 shows the results for cross-language evaluation to probe the language transfer capabilities of the multilingual models. We test models fine-tuned on language A on the test set of language B. It can be seen that BERT-MULTI generalizes better across languages than the XLM models. This might be due to the fact that the multilingual BERT model is trained on one large vocabulary of many languages but the XLM models are trained with a cross-lingual objective and language information. Hence, during fine-tuning on eye tracking

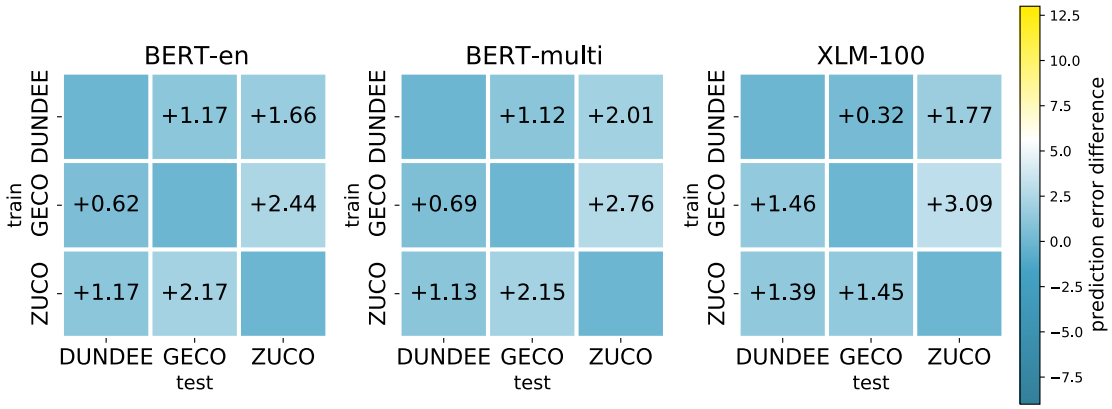

Figure 6: Cross-domain evaluation on pretrained English models. The results are expressed as the difference in the prediction error compared to the in-domain prediction. A smaller error (i.e., a color more similar to the color of the diagonal) represents better domain adaptation.

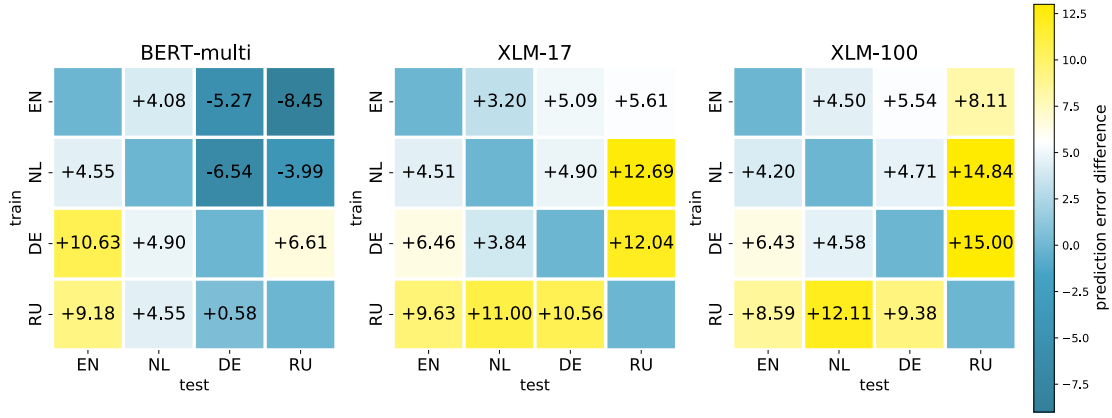

Figure 7: Cross-language evaluation on multilingual models across English, Dutch, German and Russian data. The results are expressed as the difference in the prediction error compared to the prediction on the same language. A smaller error (i.e., a color more similar to the color of the diagonal) represents better language transfer.

data from one language the XLM models lose some of their cross-lingual abilities. Our results are in line with Pires et al. (2019) and Karthikeyan et al. (2020), who showed that BERT learns multilingual representations in more than just a shared vocabulary space but also across scripts. When fine-tuning BERT-MULTI on English or Dutch data and testing on Russian, we see surprisingly high accuracy across scripts, even outperforming the in-language results. The XLM models, however, show the expected behavior where transferring within the same script (Dutch, English, German) works much better than transferring between the Latin and Cyrillic script (Russian).

7.3 Input Characteristics

Gaze patterns are strongly correlated with word length. Figure 8 shows that the models accurately learn to predict higher fixation proportions for longer words. We observe that the predictions of the XLM-100 model follow the trend in the original data most accurately. Similar patterns emerge for the other languages (see Appendix D.3). Notably, the pretrained models before fine-tuning do not reflect the word length effect.

On the sentence level, we hypothesize that eye tracking features are easier to predict for sentences with a higher readability. Figure 9 shows the accuracy for predicting the number of fixations (NFIX) in a sentence relative to the Flesch reading ease score. Interestingly, the pretrained models without fine-tuning conform to the expected behavior and show a consistent increase in accuracy for sentences with a higher reading ease score. After fine-tuning on eye tracking data, this behavior is not as visible anymore since the language models achieve constantly high accuracy independent of the readability of the sentences.

These results might be explained by the nature of the Flesch readability score, which is based only

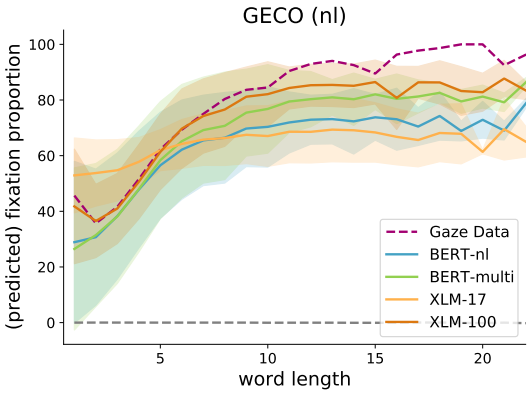

Figure 8: Prediction accuracy of FPROP with respect to word length. The gray dashed line is the result of the pretrained BERT-MULTI model without fine-tuning.

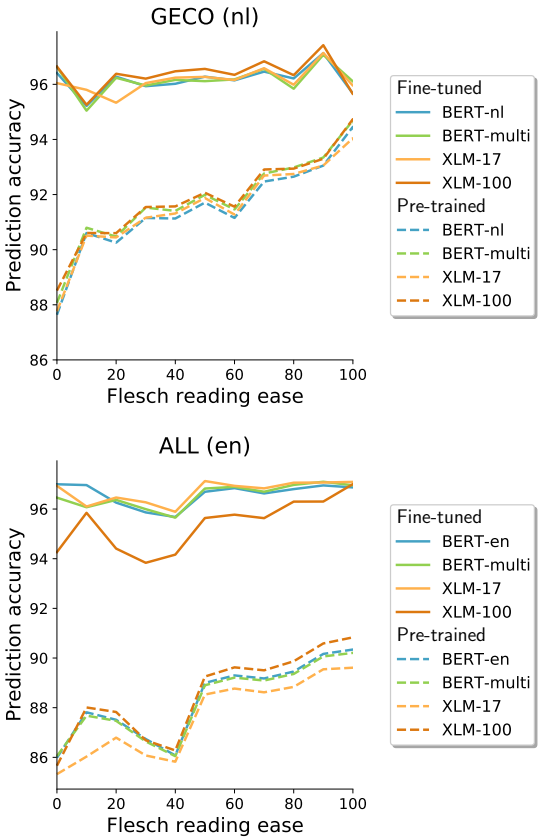

Figure 9: Prediction accuracy for NFIX relative to the Flesch reading ease score of the sentence. A higher Flesch score indicates that a sentence is easier to read. The dashed lines show the results of the pretrained language models without fine-tuning on eye tracking data.

on the structural complexity of the text (see Appendix B for a description of the Flesch Reading Ease score). Our results indicate that language models trained purely on textual input are more calibrated towards such structural characteristics, i.e., the number of syllables in a word and the number of words in a sentences. Hence, the Flesch reading ease score might not be a good approximation for text readability. In future work, comparing eye movement patterns and text difficulty should rely on readability measures that take into account lexical, semantic, syntactic, and discourse features. This might reveal deviating patterns between pretrained and fine-tuned models.

Our analyses indicate that the models learn to take properties of the input into account when predicting eye tracking patterns. These processing strategies are similar to those observed in humans. Nevertheless, the connection between readability and relative importance in text needs to be analysed in more detail to establish how well these properties are learned by the language models.

8 Conclusion

While the superior performance of pretrained transformer language models has been established, we have yet to understand to which extent these models are comparable to human language processing behavior. We take a step in this direction by fine-tuning language models on eye tracking data to predict human reading behavior.

We find that both monolingual and multilingual models achieve surprisingly high accuracy in predicting a range of eye tracking features across four languages. Compared to the XLM models, BERT-MULTI is more robust in its ability to generalize across languages, without being explicitly trained for it. In contrast, the XLM models perform better when fine-tuned on less eye tracking data. Generally, fixation duration features are predicted more accurately than fixation proportion, possibly because the latter show higher variance across subjects. We observe that the models learn to reflect characteristics of human reading such as the word length effect and higher accuracy in more easily readable sentences.

The ability of transformer models to achieve such high results in modelling reading behavior indicates that we can learn more about the commonalities between language models and human sentence processing. By predicting behavioral metrics such as eye tracking features we can investigate the cognitive plausibility within these models to adjust or intensify the human inductive biases.

Acknowledgements

Lena Jäger was partially funded by the German Federal Ministry of Education and Research under grant 01|S20043.

References

Samira Abnar, Lisa Beinborn, Rochelle Choenni, and Willem Zuidema. 2019. Blackbox meets blackbox: Representational similarity & stability analysis of neural language models and brains. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 191–203.

Phillip M Alday. 2019. M/EEG analysis of naturalistic stories: A review from speech to language processing. *Language, Cognition and Neuroscience*, 34(4):457–473.

Toni Amstad. 1978. Wie verständlich sind unsere Zeitungen? *Unpublished doctoral dissertation, University of Zürich, Switzerland.*

Xuejun Bai, Guoli Yan, Simon P Liversedge, Chuanli Zang, and Keith Rayner. 2008. Reading spaced and unspaced Chinese text: Evidence from eye movements. *Journal of Experimental Psychology: Human Perception and Performance*, 34(5):1277.

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

Maria Barrett, Joachim Bingel, Frank Keller, and Anders Søgaard. 2016. Weakly supervised part-of-speech tagging using eye-tracking data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, volume 2, pages 579–584.

Jasmijn Bastings and Katja Filippova. 2020. The elephant in the interpretability room: Why use attention as explanation when we have saliency methods? In *Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 149–155.

José Canete, Gabriel Chaperon, Rodrigo Fuentes, and Jorge Pérez. 2020. Spanish pre-trained BERT model and evaluation data. *PML4DC at ICLR*.

Branden Chan, Timo Möller, Malte Pietsch, and Tanay Soni. 2019. German BERT.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451.

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.

Nicola De Cao, Michael Sejr Schlichtkrull, Wilker Aziz, and Ivan Titov. 2020. How do decisions emerge across layers in neural models? interpretation with differentiable masking. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3243–3255.

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.

WH Douma. 1960. *De Leesbaarheid Van Landbouwbladen. Een Onderzoek Naar en Een Toepassing Van Leesbaarheidsformules.*

Allyson Ettinger. 2020. What BERT is not: Lessons from a new suite of psycholinguistic diagnostics for language models. *Transactions of the Association for Computational Linguistics*, 8:34–48.

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

Leilani H Gilpin, David Bau, Ben Z Yuan, Ayesha Bajwa, Michael Specter, and Lalana Kagal. 2018. Explaining explanations: An overview of interpretability of machine learning. In *2018 IEEE 5th International Conference on data science and advanced analytics (DSAA)*, pages 80–89. IEEE.

Anna Gladkova and Aleksandr Drozd. 2016. Intrinsic evaluations of word embeddings: What can we do better? In *Proceedings of the 1st Workshop on Evaluating Vector-Space Representations for NLP*, pages 36–42.

Ana Valeria Gonzalez-Garduno and Anders Søgaard. 2017. Using gaze to predict text readability. In *Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 438–443.

Michael Hahn and Frank Keller. 2016. Modeling human reading with neural attention. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*.

John Hale. 2017. Models of human sentence comprehension in computational psycholinguistics. In *Oxford Research Encyclopedia of Linguistics*.

Yiding Hao, Simon Mendelsohn, Rachel Sterneck, Randi Martinez, and Robert Frank. 2020. Probabilistic predictions of people perusing: Evaluating metrics of language model performance for psycholinguistic modeling. In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, pages 75–86.

Robert Hawkins, Takateru Yamakoshi, Thomas L Griffiths, and Adele Goldberg. 2020. Investigating representations of verb bias in neural language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4653–4663.

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, Jonathan Rotsztejn, Marius Troendle, Andreas Pedroni, Ce Zhang, and Nicolas Langer. 2018. ZuCo, a simultaneous EEG and eye-tracking resource for natural sentence reading. *Scientific Data*.

Nora Hollenstein, Marius Troendle, Ce Zhang, and Nicolas Langer. 2020. ZuCo 2.0: A dataset of physiological recordings during natural reading and annotation. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 138–146.

Nora Hollenstein and Ce Zhang. 2019. Entity recognition at first sight: Improving NER with eye movement information. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*.

Albrecht Werner Inhoff and Keith Rayner. 1986. Parafoveal word processing during eye fixations in reading: Effects of word frequency. *Perception & psychophysics*, 40(6):431–439.

Lena Jäger, Thomas Kern, and Patrick Haller. 2021. Potsdam Textbook Corpus (PoTeC): Eye tracking data from experts and non-experts reading scientific texts. available on OSF, DOI 10.17605/OSF.IO/DN5HP.

Sarthak Jain and Byron C Wallace. 2019. Attention is not explanation. 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 3543–3556.

Yu-Cin Jian, Ming-Lei Chen, and Hwa-wei Ko. 2013. Context effects in processing of Chinese academic words: An eye-tracking investigation. *Reading Research Quarterly*, 48(4):403–413.

Kaliyaperumal Karthikeyan, Zihan Wang, Stephen Mayhew, and Dan Roth. 2020. Cross-lingual ability of multilingual BERT: An empirical study. In *International Conference on Learning Representations*.

Frank Keller. 2010. Cognitively plausible models of human language processing. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics: Short Papers*, pages 60–67.

Alan Kennedy, Robin Hill, and Joël Pynte. 2003. The Dundee corpus. In *Proceedings of the 12th European Conference on Eye Movement*.

Sigrid Klerke, Yoav Goldberg, and Anders Søgaard. 2016. Improving sentence compression by learning to predict gaze. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1528–1533.

Goro Kobayashi, Tatsuki Kuribayashi, Sho Yokoi, and Kentaro Inui. 2020. Attention is not only a weight: Analyzing transformers with vector norms. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7057–7075.

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

AK Laurinavichyute, Irina A Sekerina, SV Alexeeva, and KA Bagdasaryan. 2019. Russian Sentence Corpus: Benchmark measures of eye movements in reading in Cyrillic. *Behavior research methods*, 51(3):1161–1178.

Yaobo Liang, Nan Duan, Yeyun Gong, Ning Wu, Fenfei Guo, Weizhen Qi, Ming Gong, Linjun Shou, Daxin Jiang, Guihong Cao, et al. 2020. XGLUE: A new benchmark dataset for cross-lingual pretraining, understanding and generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6008–6018.

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.

Ilya Loshchilov and Frank Hutter. 2018. Decoupled weight decay regularization. In *Proceedings of the International Conference on Learning Representations*.

Jonathan Malmaud, Roger Levy, and Yevgeni Berzak. 2020. Bridging information-seeking human gaze and machine reading comprehension. In *Proceedings of the 24th Conference on Computational Natural Language Learning*, pages 142–152.

Christopher D Manning, Kevin Clark, John Hewitt, Urvashi Khandelwal, and Omer Levy. 2020. Emergent linguistic structure in artificial neural networks trained by self-supervision. *Proceedings of the National Academy of Sciences*.

Franz Matthies and Anders Søgaard. 2013. With blinkers on: Robust prediction of eye movements across readers. *Proceedings of the 2013 Conference on empirical methods in natural language processing (EMNLP)*, pages 803–807.

Danny Merkx and Stefan L Frank. 2020. Comparing transformers and RNNs on predicting human sentence processing data. *arXiv preprint arXiv:2005.09471*.

James Michaelov and Benjamin Bergen. 2020. How well does surprisal explain N400 amplitude under different experimental conditions? In *Proceedings of the 24th Conference on Computational Natural Language Learning*, pages 652–663.

Robin K Morris. 1994. Lexical and message-level sentence context effects on fixation times in reading. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 20(1):92.

I Oborneva. 2006. Automatic assessment of the complexity of educational texts on the basis of statistical parameters.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual BERT? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4996–5001.

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

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

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.

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*.

Ekta Sood, Simon Tannert, Philipp Mueller, and Andreas Bulling. 2020. Improving natural language processing tasks with human gaze-guided neural attention. *Advances in Neural Information Processing Systems*, 33.

Ece Takmaz, Sandro Pezzelle, Lisa Beinborn, and Raquel Fernández. 2020. Generating image descriptions via sequential cross-modal alignment guided by human gaze. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4664–4677.

Mariya Toneva and Leila Wehbe. 2019. Interpreting and improving natural-language processing (in machines) with natural language-processing (in the brain). In *Advances in Neural Information Processing Systems*, pages 14928–14938.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 5998–6008.

Antti Virtanen, Jenna Kanerva, Rami Ilo, Jouni Luoma, Juhani Luotolahti, Tapio Salakoski, Filip Ginter, and Sampo Pyysalo. 2019. Multilingual is not enough: BERT for Finnish. *arXiv preprint arXiv:1912.07076*.

Wietse de Vries, Andreas van Cranenburgh, Arianna Bisazza, Tommaso Caselli, Gertjan van Noord, and Malvina Nissim. 2019. BERTje: A Dutch BERT Model. *arXiv preprint arXiv:1912.09582*.

Ivan Vulić, Edoardo Maria Ponti, Ira Leviant, Olga Majewska, Matt Malone, Roi Reichart, Simon Baker, Ulla Petti, Kelly Wing, Eden Bar, et al. 2020. Multi-SimLex: A large-scale evaluation of multilingual and cross-lingual lexical semantic similarity. *Computational Linguistics*, pages 1–73.

Leila Wehbe, Ashish Vaswani, Kevin Knight, and Tom Mitchell. 2014. Aligning context-based statistical models of language with brain activity during reading. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 233–243.

Ethan Gotlieb Wilcox, Jon Gauthier, Jennifer Hu, Peng Qian, and Roger Levy. 2020. On the predictive power of neural language models for human real-time comprehension behavior. *arXiv preprint arXiv:2006.01912*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2019. Huggingface's transformers: State-of-the-art natural language processing. *ArXiv*, abs/1910.03771.

Kuratov Yu and M Arkhipov. 2019. Adaptation of deep bidirectional multilingual transformers for Russian language. *Computational Linguistics and Intellectual Technologies*, (18):333–339.

A Eye Tracking Data

Table 6 presents information about the range of the eye tracking features.
Figure 10 shows the word length effect found in eye tracking data recorded during reading. i.e., the fact that longer words are more likely to be fixated. This effect is observable across all languages.

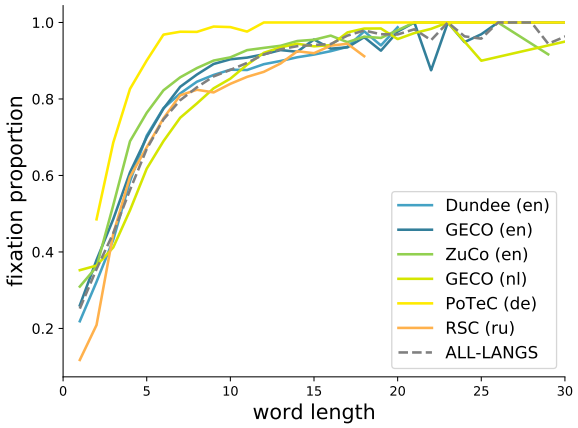

Figure 10: Word length effect on all datasets in all four languages.

Figure 11 shows the mean fixation duration (MFD) for adjectives, nouns, verbs, and adverbs for all six datasets. We use spacy[7] to perform part-of-speech tagging for our analyses. For Russian we load an externally trained model[8], for Dutch, English and German we use the provided pretrained models.
Figure 12 shows an additional analysis where we explore which parts-of-speech can be predicted more accurately by the language models.

B Readability Scores

We use the Flesch Reading Easy score (Flesch, 1948) to define the readability of the English text in the eye tracking corpora. This score indicates how difficult a text passage is to understand. Since this score relies on language-specific weighting factors, we apply the Flesch Douma adaptation for Dutch (Douma, 1960), the adaptation by Amstad (1978) for German, and the adaptation by Oborneva (2006) for Russian.

C Implementation Details

C.1 Tokenization

When using BERT or XLM for token classification or regression, a pressing implementation issue is represented by the subword tokenizers employed by the models. This tokenizer, in fact, handles unknown tokens by recursively splitting every word until all subtokens belong to its vocabulary. For example, the name of the Greek mythological hero "Philammon" is tokenized into the three subtokens "['phil', '##am', '##mon']". In this case, our models for token regression would produce an eight-dimensional output for all three subtokens, and we had the choice as to what to do in order to compute the loss, having only one target for the full word "Philammon". We chose to compute the loss only with respect to the first subtoken.

C.2 Training Setup

As described in the main paper, all experiments are run over 5 random seeds, which are $\{12, 79, 237, 549, 886\}$.
All models were fine-tuned on a single GPU Titan X with 12 GB memory. Due to memory restrictions of the GPUs and the dimensions of the language models, the batch size was adapted as needed. Table 5 shows the batch sizes for each model.

Model	Batch size
BERT-EN, BERT-NL, BERT-MULTI	16
BERT-DE, BERT-RU, XLM-ENDE, XLM-17, XLM-100	8
XLM-EN	2

Table 5: Batch sizes used for each of the language models.

On average the validation accuracy of BERT models stops improving after $\sim$ 50 epochs, while the XLM models only take $\sim$ 10 epochs. There is no noteworthy difference in training speed between monolingual and multilingual models.

D Detailed Results

In this section we present addition plots that strengthen the results shown in the main paper.

D.1 Pretrained Baseline

Tables 7 and 8 show the prediction accuracy of the pretrained models.
Moreover, Figure 13 shows the results of individual gaze features for all pretrained models (without

[7] spaCy.io
[8] https://github.com/buriy/spacy-ru

fine-tuning) on the Dundee (en) and RSC (ru) corpora.

Figure 14 presents the differences in prediction accuracy for the pretrained XML-100 model predictions relative to the mean baseline for each eye tracking feature. The pretrained models clearly cannot outperform the mean baseline for any language or dataset.

D.2 Individual Feature Results

Figure 15 shows the prediction accuracy of the fine-tuned language models for the individual eye tracking features for all datasets.

D.3 Word Length Effect

Figure 16 presents the comparison between models predictions and original word length effects for further languages.

Corpus	nFix	MFD	fProp	FFD	FPD	TRT	nRefix	reProp
Dundee (en)	0.8 (0.5)	119.5 (62.1)	0.6 (0.3)	120.7 (63.4)	140.6 (88.5)	156.1 (105.5)	0.2 (0.3)	0.2 (0.2)
GECO (en)	0.8 (0.5)	128.4 (59.0)	0.6 (0.2)	129.3 (60.1)	143.3 (77.5)	168.2 (102.4)	0.2 (0.3)	0.2 (0.2)
ZuCo (en)	1.1 (0.7)	78.4 (34.8)	0.7 (0.3)	77.3 (34.4)	92.3 (52.2)	129.8 (89.7)	0.4 (0.5)	0.3 (0.2)
GECO (nl)	0.8 (0.6)	121.3 (80.1)	0.6 (0.4)	121.8 (81.1)	134.1 (98.0)	158.1 (131.2)	0.2 (0.4)	0.1 (0.2)
PoTeC (de)	2.7 (2.9)	217.5 (117.3)	0.8 (0.4)	167.9 (157.4)	224.7 (264.2)	675.6 (727.0)	1.7 (2.2)	0.6 (0.5)
RSC (ru)	0.8 (0.4)	203.4 (115.1)	0.6 (0.3)	233.6 (49.5)	285.1 (101.9)	314.2 (179.8)	0.1 (0.1)	0.1 (0.1)

Table 6: Mean and standard deviation for all eye tracking features of the corpora used in this work.

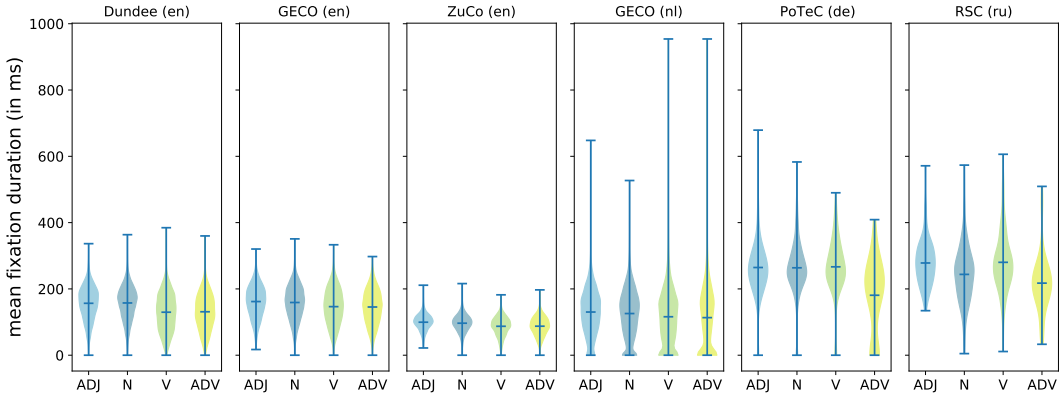

Figure 11: Mean fixation duration (MFD) for the most common parts of speech across all six datasets.

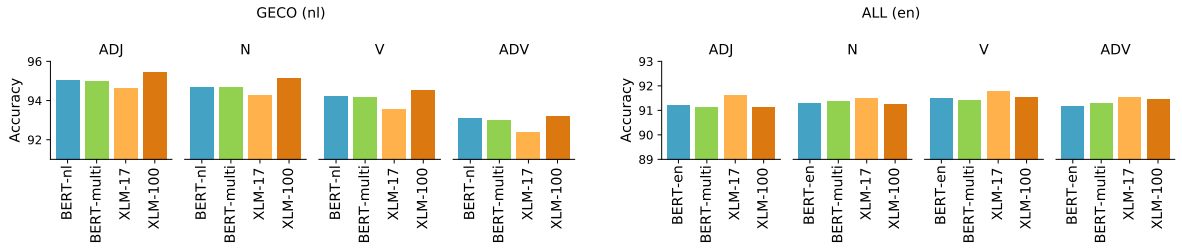

Figure 12: Accuracy of the language models predicting the mean fixation duration (MFD) across various parts of speech for Dutch (left) and English (right).

Model	**Dundee**	**GECO (en)**	**ZuCo (en)**	**ALL (en)**
BERT-EN	77.42 (0.21)	77.67 (0.13)	76.06 (0.38)	78.69 (0.09)
BERT-MULTI	77.41 (0.21)	77.68 (0.13)	76.07 (0.37)	78.66 (0.07)
XLM-EN	77.21 (0.29)	77.65 (0.24)	75.97 (0.60)	78.47 (0.11)
XLM-ENDE	77.40 (0.29)	77.67 (0.10)	76.10 (0.41)	78.66 (0.12)
XLM-17	77.31 (0.23)	77.66 (0.19)	75.99 (0.39)	78.39 (0.15)
XLM-100	77.35 (0.29)	77.63 (0.34)	75.93 (0.43)	78.49 (0.11)

Table 7: Prediction accuracy of the pretrained language models aggregated over all eye tracking features for the English corpora, including the concatenated dataset. Standard deviation is reported in parentheses.

Model	GECO (nl)	PoTeC (de)	RSC (ru)	ALL-LANGS
BERT-NL	84.20 (0.10)	-	-	-
BERT-DE	-	73.55 (3.07)	-	-
BERT-RU	-	-	64.83 (2.09)	-
BERT-MULTI	84.28 (0.10)	73.47 (3.01)	64.82 (2.11)	86.22 (0.29)
XLM-ENDE	-	73.49 (2.99)	-	-
XLM-17	83.93 (0.16)	73.17 (2.86)	65.02 (2.11)	85.84 (0.27)
XLM-100	83.94 (0.27)	73.28 (2.91)	64.67 (2.10)	85.94 (0.38)

Table 8: Prediction accuracy of the pretrained language models aggregated over all eye tracking features for the Dutch, German and Russian corpora, and for all four languages combined in a single dataset. Standard deviation is reported in parentheses.

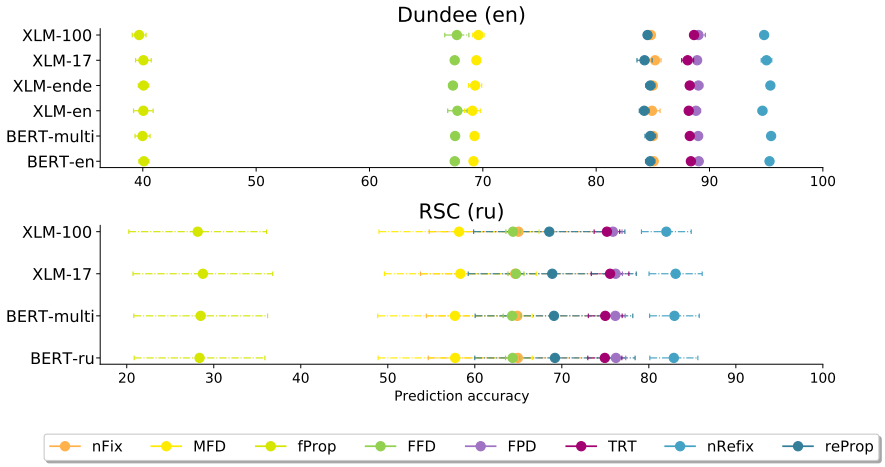

Figure 13: Results of individual gaze features for all pretrained models (without fine-tuning) on the Dundee (en) and RSC (ru) corpora.

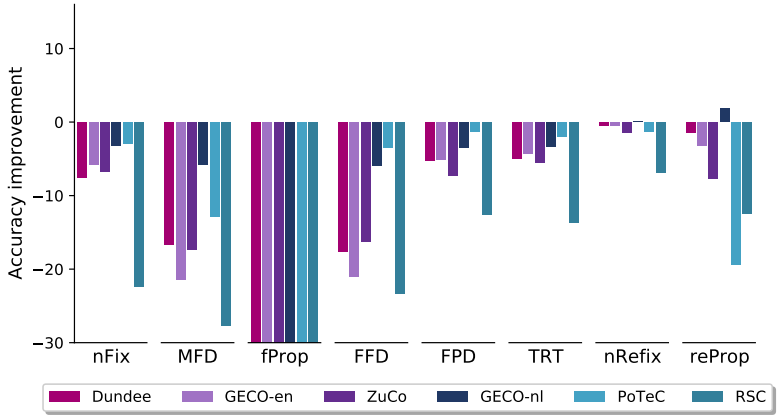

Figure 14: Differences in prediction accuracy for the pretrained XLM-100 model predictions (without fine-tuning on eye tracking data) relative to the mean baseline for each eye tracking feature.

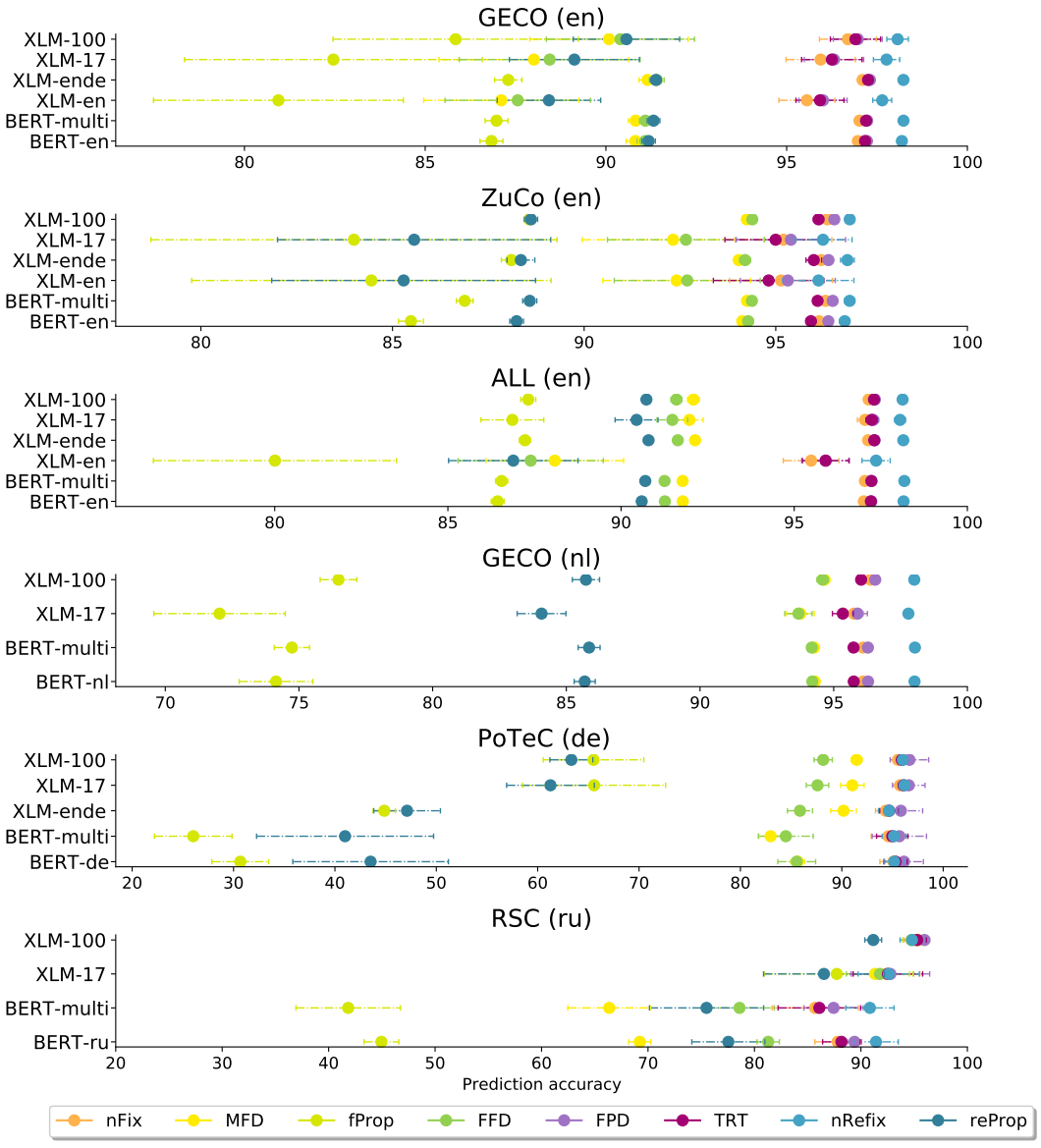

Figure 15: Results of individual eye tracking features for all fine-tuned models on all datasets not presented in the main paper.

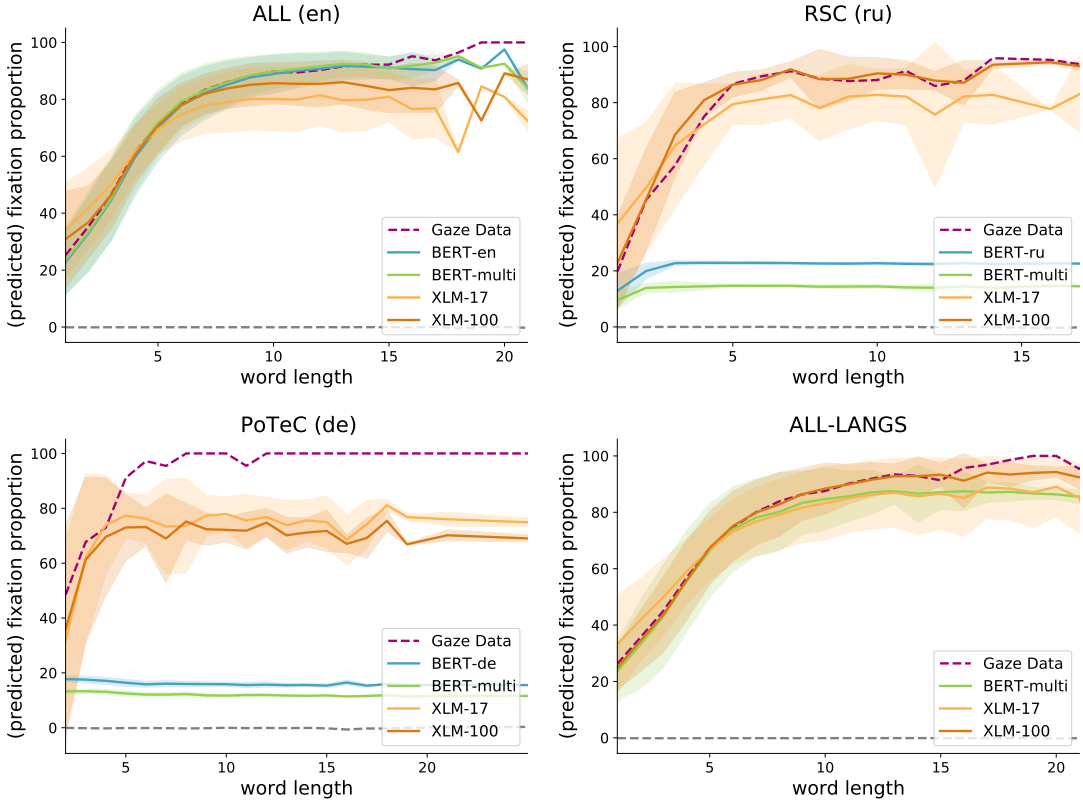

Figure 16: Word length versus predicted fixation probability for Russian, German and English. The gray dashed line is the result of the pretrained BERT-MULTI model without fine-tuning.