Cross-Lingual Transfer of Cognitive Processing Complexity

Charlotte Pouw
ILLC, University of Amsterdam*
c.m.pouw@uva.nl

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

Lisa Beinborn
Vrije Universiteit Amsterdam
l.beinborn@vu.nl

Abstract

When humans read a text, their eye movements are influenced by the structural complexity of the input sentences. This cognitive phenomenon holds across languages and recent studies indicate that multilingual language models utilize structural similarities between languages to facilitate cross-lingual transfer. We use sentence-level eye-tracking patterns as a cognitive indicator for structural complexity and show that the multilingual model XLM-RoBERTa can successfully predict varied patterns for 13 typologically diverse languages, despite being fine-tuned only on English data. We quantify the sensitivity of the model to structural complexity and distinguish a range of complexity characteristics. Our results indicate that the model develops a meaningful bias towards sentence length but also integrates cross-lingual differences. We conduct a control experiment with randomized word order and find that the model seems to additionally capture more complex structural information.

1 Introduction

Approximately 7,000 languages are currently spoken in the world, exhibiting differences at almost every level of linguistic organization (Eberhard et al., 2022). Nonetheless, psycholinguistic theories are predominantly supported by evidence from a handful of Indo-European languages (Norcliffe et al., 2015). Only recently, researchers have started to explore cross-linguistic differences in the neural implementation of language, uncovering both striking similarities across languages and empirical differences that cannot be explained by a unitary account (Malik-Moraleda et al., 2022).

In natural language processing, multilingual language models are optimized for tasks such as machine translation or cross-lingual information retrieval (Conneau et al., 2020) and follow a linguistically naïve training regime. They are trained on dozens of languages simultaneously and do not account for typological differences between languages. Nevertheless, their cross-lingual transfer performance sets new records, even in zero-shot settings (Pires et al., 2019). The ability to transfer knowledge across languages has been attributed to the shared vocabulary that is used for all languages (Wu and Dredze, 2019) because it enables the reuse of common morphological roots for languages from the same family. However, recent studies indicate that vocabulary sharing is not a prerequisite for cross-lingual transfer (Artetxe et al., 2020) and that structural commonalities between languages play a more prevalent role in models (Karthikeyan et al., 2020).

Human sentence processing is sensitive to structural complexity. Eye movement data recorded during reading provide insights into cognitive processing patterns with a temporal accuracy of milliseconds (Winke, 2013). Structural processing difficulty materializes as regressions towards the complex region and an increase of fixations on that region (Clifton and Staub, 2011). For example, sentences with an object-relative structure trigger more regressions than sentences with more common subject-relative clauses (Gordon et al., 2006). A classical example of structural complexity are garden-path sentences which initially trigger a simplified interpretation that must be revised when reading the rest of the sentence (Bever, 1970).

On the surface level, eye movement patterns are language-specific since they are influenced by visual factors such as orthography and word length (Kliegl et al., 2004). For example, the Chinese script is much more visually dense than the alphabetic script, resulting in longer fixations and saccades that move to positions relatively close to the current word (Liversedge et al., 2016). On a deeper processing level, reading patterns seem to converge across languages. Predictability effects

*This research was developed when the first author was affiliated to Vrije Universiteit Amsterdam.

Findings of the Association for Computational Linguistics: EACL 2023, pages 655–669

have been demonstrated in multiple languages (Al-Jassmi et al., 2022; Laurinavichyute et al., 2019) and sentences that are matched for content are read at a similar speed in Chinese, English, and Finnish (Liversedge et al., 2016).

Sarti et al. (2021) find that the representations of an English pre-trained transformer-based language model encode structural complexity more prominently when they are fine-tuned to predict English eye-tracking patterns. Interestingly, Rama et al. (2020) claim that structural similarity between languages is only weakly represented in multilingual models. Nevertheless, Hollenstein et al. (2021) show that multilingual models are able to predict eye movement patterns of reading even for languages that are not seen during fine-tuning, which indicates a general learnability of the relationship between structural complexity and eye movement patterns. Their results are restricted to four languages (three of them are from the Germanic family), and it remains unclear which structural cues are leveraged for the cross-lingual prediction because the test sentences are not aligned across languages.

Contributions We examine whether the multilingual model XLM-RoBERTa (henceforth XLM-R) is sensitive to the structural complexity patterns that can be found in eye-tracking data. We use data from the newly released Multilingual Eye-tracking Corpus (Siegelman et al., 2022) to predict eye movement patterns for parallel texts in 13 typologically diverse languages. This allows us to specifically target the model's sensitivity towards structural information and rules out the possibility that the results are influenced by differences in semantics or dataset sizes.

We show that XLM-R can apply cross-lingual transfer to predict eye-tracking patterns for all 13 languages while being fine-tuned only on English eye-tracking data. Our results indicate that the model develops a meaningful bias towards sentence length, but also integrates cross-lingual differences. For a more detailed analysis of structural sensitivity, we probe the model's final layer for complexity features. Based on a control experiment with randomized word order, we conclude that the model seems to additionally capture more complex structural information. All our experimental code is publicly available at https://github.com/CharlottePouw/crosslingual-complexity-transfer.

2 Related Work

We introduce recent findings on the role of structural information for cross-lingual transfer in multilingual models and motivate the use of eye-tracking data as a proxy for cognitive processing complexity.

2.1 Cross-lingual Transfer in Multilingual Models

Massive multilingual language models such as mBERT (Devlin et al., 2019) and XLM-R (Conneau et al., 2020) are trained on more than a hundred languages simultaneously. Wu and Dredze (2019) show that this approach leads to surprisingly strong performances in cross-lingual transfer settings and attribute the improvements to the shared subword vocabulary. Pires et al. (2019) note that the model's ability to generalize "cannot be attributed solely to vocabulary memorization". Complementary, Artetxe et al. (2020) and Liu et al. (2020) find that a shared vocabulary is not necessary for cross-lingual transfer. Instead, the multilingual model seems to exploit structural similarity between the training and the target language to facilitate transfer (Karthikeyan et al., 2020).

Structural similarity is loosely defined as an overlap on a subset of typological characteristics which seem to be better reflected in multilingual language models explicitly optimizing for cross-lingual transfer (Beinborn and Choenni, 2020; Choenni and Shutova, 2022). In language-agnostic models such as mBERT and XLM-R, the multilingual representations of the input can be separated into language-specific and language-neutral components (Tanti et al., 2021; Libovický et al., 2020; Gonen et al., 2020). While Rama et al. (2020) find that structural similarity between languages is only weakly represented in these models, Bjerva et al. (2019) observe that structural similarity between languages correlates most with representational similarity. Experiments with artificial languages indicate that multilingual models are sensitive to hierarchical structure (De Varda and Zamparelli, 2022) and to word order (Chai et al., 2022; Deshpande et al., 2022). Ahmad et al. (2021) show that cross-lingual transfer can be improved by explicitly encoding structural information via an auxiliary syntactic objective and Guarasci et al. (2022) find that structural complexity knowledge can even be transferred across languages without explicit training.

2.2 Predicting Processing Complexity

Recent studies indicate that transformer-based language models are sensitive to structural characteristics of the input sentence when predicting eye-tracking patterns. Hollenstein et al. (2021) find a correlation between the Flesch reading ease score and eye-tracking prediction accuracy of pre-trained multilingual transformer models which disappears after fine-tuning. Wiechmann et al. (2022) detect similar correlations between the prediction accuracy of English transformer models and a wider range of readability features. Finally, Hollenstein et al. (2022b) find that eye-tracking metrics predicted by multilingual transformer models correlate in a similar way with readability features as eye-tracking metrics recorded from human readers.

Sensitivity to structural complexity also seems to increase when incorporating eye-tracking data in NLP models. Learning eye movement behavior as an auxiliary task has been shown to facilitate the prediction of text complexity in English and Portuguese (González-Garduño and Søgaard, 2017; Evaldo Leal et al., 2020). Barrett et al. (2016) show that English eye-tracking features improve the performance a French part-of-speech tagger, suggesting that information learned from monolingual eye-tracking data is transferable across languages.

In this work, we explicitly test for sensitivity to a range of structural characteristics in multilingual models and analyze if structural sensitivity increases by learning to predict eye-tracking patterns. We extend previous analyses to a much wider range of languages from five different families (Indo-European, Koreanic, Semitic, Turkic, and Uralic).

3 Methodology

We fine-tune a pre-trained multilingual transformer model to predict eye-tracking metrics in a setting of zero-shot cross-lingual transfer.

3.1 Data

We use the aligned multilingual eye-tracking corpus MECO for testing. As the multilingual data consists of only few samples, we use the larger monolingual English eye-tracking dataset GECO for training. Size statistics of both corpora can be found in the appendix in Table 3.

Multilingual Eye-tracking Corpus (MECO) The Multilingual Eye-tracking Corpus contains parallel eye-tracking data of reading in 13 different languages (Siegelman et al., 2022).[1] The reading material consists of 12 short Wikipedia-style texts about various topics, which participants read in their native language. The texts were either directly translated or carefully matched for topic, genre, and readability. Each of the 12 texts was presented on a single screen and in the same fixed order in all languages. The number of participants ranged from 29 to 54 per language (45 on average).

Ghent Eye-tracking Corpus (GECO) The Ghent Eye-tracking Corpus contains eye-tracking data from 14 monolingual English readers (Cop et al., 2016). They were reading the entire novel *The Mysterious Affair at Styles* by Agatha Christie which was presented on the screen one paragraph at a time.

3.2 Experimental Setup

We use multi-task learning for predicting four sentence-level eye-tracking metrics.

Sentence-Level Eye-Tracking Metrics Liversedge et al. (2016) find that eye movement patterns are more comparable across languages at the sentence level than at the word level. We select four sentence-level eye-tracking metrics that cover both early and late language processing in line with Sarti et al. (2021). For each sentence s, we consider:

1. *Fixation count*: number of fixations on s
2. *Total fixation duration*: total duration of all fixations on s
3. *First-pass duration*: duration of the first reading pass over s
4. *Regression duration*: total duration of all regressions within s.

Duration values are measured in milliseconds. To obtain generalized eye movement patterns, we average all eye-tracking metrics over participants and scale each eye-tracking feature to fall in the range 0–100, so that the loss can be calculated uniformly for durations and counts (Hollenstein et al., 2021). The distribution of the four metrics is shown in the appendix in Figure 7.

Model We use XLM-R (Conneau et al., 2020) as our multilingual transformer model since it achieved the best zero-shot results in the CMCL 2022 Shared Task on Multilingual and Crosslingual

[1] Dutch, English, Estonian, Finnish, German, Greek, Hebrew, Italian, Korean, Norwegian, Russian, Spanish, Turkish.

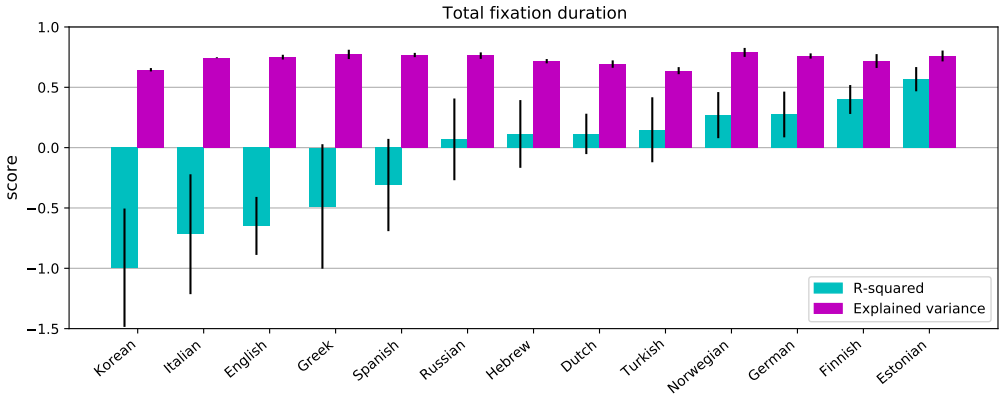

Figure 1: Cross-lingual transfer results for predicting cognitive processing complexity (i.e. sentence-level fixation duration). Prediction performance is evaluated with explained variance and R^2 for each language in MECO. The results are averaged over 5 folds; error bars denote the standard deviation over folds.

Prediction of Human Reading Behaviour (Srivastava, 2022; Hollenstein et al., 2022a). The model was pre-trained on 2.5TB CommonCrawl data containing 100 languages using the Masked Language Modelling objective and uses SentencePiece subword tokenization (Kudo and Richardson, 2018). We select the Huggingface checkpoint *xlm-roberta-base* and add a linear dense layer to predict four sentence-level eye-tracking metrics.

Multi-Task Learning We employ multi-task learning with hard parameter sharing to fine-tune the model on all eye-tracking metrics simultaneously in line with Sarti et al. (2021). This means that all model parameters are shared except for the task-specific regression heads in the final prediction layer. More specifically, the same sentence representation is fed into each of the four regression heads which predict their respective eye-tracking metric. The model parameters are optimized jointly for all regression tasks by summing the individual MSE losses in line with previous work (Hollenstein et al., 2021, 2022a; Wiechmann et al., 2022).

Training Parameters We fine-tune XLM-R for 15 epochs with early stopping after 5 epochs without an improvement in the validation accuracy. We use 10% of the training data as validation data and evaluate every 40 steps. We employ a batch size of 32 and a learning rate of 1e-5. The sentence representation is obtained by mean pooling over token representations. We train the model on the GECO data using 5-fold cross-validation and report the average over the folds for each language in MECO.

Evaluation We report explained variance and R-Squared (R^2) to capture the proportion of variance in the dependent variable that can be explained by our model in line with Sarti et al. (2021). Explained variance uses the biased variance to determine what fraction of the variance is explained. R^2 uses the raw sums of squares instead and provides complementary information about systematic offsets in the predictions. We report both metrics and evaluate the performance of the fine-tuned model individually for each of the four eye-tracking metrics.[2]

4 Cross-Lingual Transfer Results

Figure 1 shows the explained variance and R^2 scores of the fine-tuned model for total fixation duration across languages. In terms of explained variance, we see that the model achieves a similar performance across languages, i.e. it captures 60 to 80 percent of the variance in the original eye-tracking signal for all languages. The R^2 scores, on the other hand, vary much more depending on the language. Similar results were observed for two of the other eye-tracking metrics, i.e. fixation count and first-pass duration, but the model is worse at predicting regression duration (see Figure 8 in the appendix). To better control for spurious correlations, we ran the experiment on permuted input-output pairs, i.e., we paired input sentences with eye-tracking values corresponding to another random sentence and averaged the results over 5 folds. For this random baseline setup, both explained variance and R^2 are always strictly negative for all languages.

[2]In previous work on token-level eye-tracking prediction, the mean absolute error was reported instead but it is less informative for sentence-level predictions because sentence-level eye-tracking metrics are generally more centered around the mean.

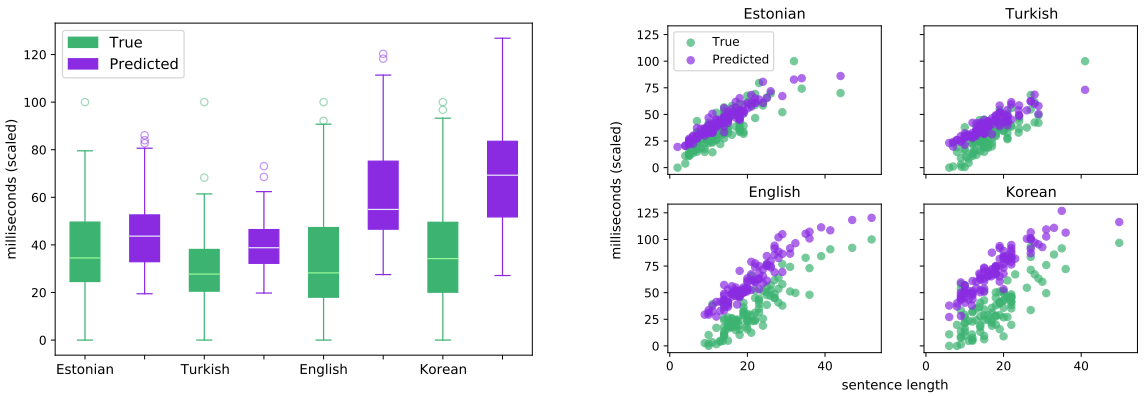

Figure 2: The left plot shows the distribution of true and predicted values for total fixation duration for Estonian, Turkish, English and Korean sentences in MECO. The right figure shows the distribution of values with respect to sentence length.

To better understand the varied R^2 scores for different languages, we show the distribution of the true and predicted values for total fixation duration for two languages with high R^2 (Estonian, Turkish) and two languages with low R^2 (English, Korean) in Figure 2. We see that the low R^2 for English and Korean is caused by predictions that are consistently too high. For Estonian and Turkish, the difference between true and predicted values is clearly smaller, resulting in a higher R^2. Nevertheless, the model is able to predict a significant amount of the variance in the eye-tracking signal of all languages, as expressed by the stable explained variance scores across languages.

Interestingly, the model performs slightly better for most zero-shot languages than for the fine-tuning language English. Recall that this performance difference cannot be attributed to cross-lingual differences in semantics, since all sentences are parallel with respect to content. On the right side of Figure 2, we analyze the predictions with respect to sentence length and find that both the model predictions and the true values for fixation duration correlate with sentence length in all languages. As sentence length is an indicator of structural complexity, we further dissect this phenomenon and conduct an analysis of a range of structural characteristics in the following section.

5 Sensitivity to Structural Complexity

We explore four categories of sentence-level complexity features: length, frequency, morpho-syntactic, and syntactic. Word frequencies are obtained as standardized Zipf frequencies using the Python package wordfreq (Speer et al., 2018). The package combines several frequency resources, including SUBTLEX lists (e.g. Brysbaert and New (2009)) and OpenSubtitles (Lison and Tiedemann, 2016). The morpho-syntactic and syntactic features are computed using the Profiling-UD tool (Brunato et al., 2020).

Cross-Lingual Differences We showcase an individual example sentence in Table 1 to compare the predicted fixation duration for English, Finnish and Turkish. We observe that the highest value is predicted for the English version. This is most likely caused by its length, as the sentence is less complex than the Finnish and Turkish versions in terms of all other linguistic features.

Interestingly, the model predicts that Finnish readers will fixate on the sentence longer than Turkish readers, even though both sentences have the same length. The Turkish sentence contains longer, less frequent words, and is lexically more dense, but the Finnish sentence contains longer dependency links. This indicates that the model is more sensitive to dependency structure than to low-level complexity (i.e. word length and frequency) when predicting eye-tracking values for sentences of the same length.

5.1 Sensitivity to Fine-Tuning Input

To analyze the model's sensitivity to the structural complexity of the fine-tuning data, we compare the performance of the fine-tuned model for in-domain data (English GECO) and cross-domain data (English MECO). Table 2 shows the explained variance and R^2 scores of the fine-tuned model predictions for each eye-tracking metric for both domains. We see that the model consistently yields

Example		Prediction
English	*In ancient Roman religion and myth, Janus is the god of beginnings and gates.*	**42.96**
Finnish	*Muinaisen roomalaisen mytologian mukaan Janus oli alkujen ja porttien jumala.*	38.91
Turkish	*Antik Roma inanışlarında ve mitlerinde, Janus başlangıçların ve kapıların tanrısıdır.*	32.28

Structural Complexity		English	Finnish	Turkish
Length	Sentence length (tokens)	**14**	10	10
	Avg. word length (characters)	4.57	6.80	**7.60**
Frequency	Avg. word frequency (Zipf)	5.63	4.36	**3.46**
	# low frequency words	2	**6**	**6**
Morpho-Syntactic	Lexical density	0.57	0.70	**0.73**
Syntactic	Parse tree depth	3	3	3
	Avg. dependency link length	2.15	**2.78**	1.90
	Max. dependency link length	**7**	**7**	4
	# verbal heads	1	1	1

Table 1: Predicted values for total fixation duration for the same example sentence in English, Finnish, and Turkish (top), and the respective values for the nine structural complexity features (bottom).

more accurate predictions for the in-domain data than for the cross-domain data.

	MECO		GECO	
	EV	R^2	EV	R^2
FC	.78 (.02)	-.63 (.35)	.93 (.00)	.93 (.01)
TFD	.75 (.02)	-.65 (.24)	.92 (.00)	.92 (.01)
FPD	.50 (.03)	-.87 (.27)	.95 (.00)	.95 (.01)
RD	-.28 (.14)	-.96 (.45)	.44 (.04)	.45 (.05)

Table 2: Explained variance (EV) and R^2-scores of the fine-tuned model predictions for four eye-tracking metrics from the English parts of MECO and GECO: fixation count (FC), total fixation duration (TFD), first-pass duration (FPD), and regression duration (RD). The results are averaged over 5 folds; standard deviations are indicated in parentheses.

To better understand why the model does not generalize well across domains for English, we visualize the Spearman correlation between complexity features and eye-tracking metrics for English GECO and MECO sentences in Figure 3. We see that the predicted values for the MECO sentences exhibit a similar correlation pattern with the complexity features as the GECO sentences. The true values of MECO are less consistent with this pattern. Literary texts contain very different words than encyclopedic texts, which might influence fixation durations and trigger regressions that cannot solely be explained by structural complexity. In addition, MECO is significantly smaller than GECO (99 vs 4,041 English sentences) and contains data from a higher number of participants (46 vs 14). The smaller amount of sentences and the larger amount of readers increase the effect of individual differences[3] which might obscure correlations between structural complexity and eye movement patterns. Directly applying the learned correlations from GECO to MECO might explain why the fine-tuned model fails to generalize across domains.

The average sentence length is considerably higher in GECO than in MECO (21 vs 13 words, see Table 3). As the model predictions strongly correlate with sentence length, we speculate that the model overestimates eye-tracking values for sentences that are longer than the majority of fine-tuning sentences which would explain the higher mean of the predictions in Figure 2.

Multi-Task Learning Effect Figure 3 further shows that regression duration is only weakly correlated with the complexity metrics in contrast to the other eye-tracking metrics. Nevertheless, the correlations between the model predictions and the complexity features are similar for all four metrics. This indicates a drawback of multi-task learning: since the loss is computed jointly over all tasks, accurate predictions for three out of four tasks already yield a small loss. The model seems to overfit to first-pass duration, total fixation duration and fixation count, which can all be predicted from similar complexity features, and does not learn the deviat-

[3] A higher number of participants leads to more diversity across readers with respect to individual factors that could influence reading strategies (e.g. age, education level). The GECO data came from 14 English readers who were all undergraduate students with an age range of 18-26. The MECO data came from 29 to 54 readers per language (45 on average), who had more diverse educational backgrounds and a wider age range (18-45). Based on these statistics, we assume that the increased heterogeneity of the MECO participants influences the correlations observed in Figure 3.

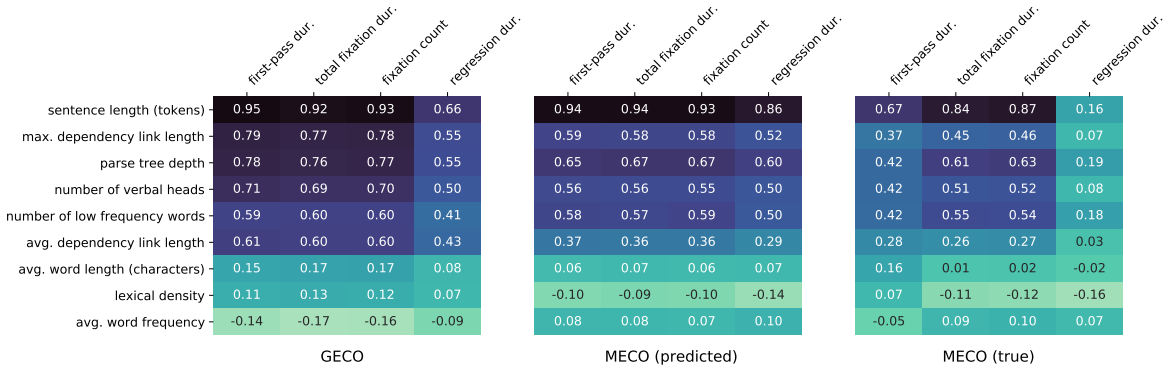

Figure 3: Spearman correlations between complexity features and eye-tracking metrics of GECO and the English part of MECO (predicted versus true). A darker color represents a stronger correlation. All GECO correlations are significant ($p < 0.001$); MECO correlations above 0.2 are significant ($p < 0.01$).

ing patterns to predict regression duration. Further research is needed to better understand the linguistic features underlying regression duration.

5.2 Feature-Based Prediction

To further establish which complexity features are good predictors for each individual eye-tracking metric, we examine the extent to which the four eye-tracking metrics can be predicted from explicit features. Since multi-task learning seems to have a negative impact on learning the structural features underlying each individual eye-tracking metric, we train a separate feature-based model for each eye-tracking metric individually. We use support vector machines (SVM) with a linear kernel as our feature-based regression models. We employ the SVR implementation from scikit-learn (Pedregosa et al., 2011) with all default parameters and use different subsets of features from Table 1: 1) only the two length features, 2) only the two frequency features, 3) only the five structural (i.e., morpho-syntactic and syntactic) features, and 4) all nine features.

As the SVM models predict a simpler problem (a single eye-tracking metric), it is not surprising that they outperform the fine-tuned multi-task model with respect to the absolute predictions (as measured by R^2, see appendix Figure 9). More interestingly, Figure 4 shows that the multi-task model is able to capture a similar amount of variance as the length-based SVM. Furthermore, we see that the length-based SVM performs almost identically to the SVM trained on *all* complexity features, outperforming the SVMs trained on frequency features and structural features. This shows that length is a strong predictor for sentence-level eye-tracking metrics, and suggests that structural and frequency features do not provide much additional information. We further investigate if length is the main factor affecting the predictions of the fine-tuned model in the following section.

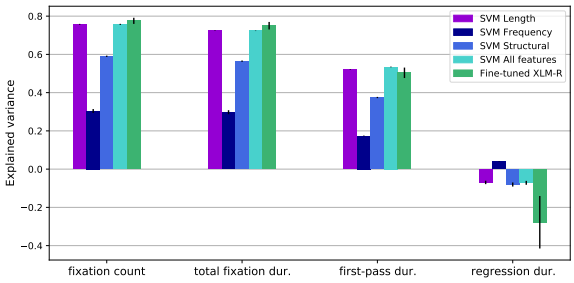

Figure 4: Explained variance of the four feature-based SVM models and the fine-tuned XLM-R model. The models are trained on GECO using 5-fold cross-validation and evaluated on the English part of MECO; error bars denote the standard deviation over folds.

6 The Role of Sentence Length

To test whether the fine-tuned XLM-R model captures more sophisticated structural information than sentence length, we conduct two additional experiments. First, we probe the final-layer representations of the model for the complexity features from Table 1, both before and after fine-tuning on eye-tracking data. Second, we compare the performance of the fine-tuned model to a control condition: we randomize the word order within each MECO sentence to analyze the prediction performance on scrambled input.

6.1 Probing Set-up

We train regressors g_i to predict a value for each of the nine latent factors of structural complexity

$Z = z_1,...,z_9$ using XLM-R's final-layer representation $\theta(x)$ of our input sentence x. The prediction accuracy of g_i is an indication of how prominently the linguistic property z_i is encoded in θ. We analyze this both for the pre-trained and fine-tuned representations of XLM-R to quantify the relative increase of sensitivity to z_i after fine-tuning on eye-tracking metrics.

We conduct the probing experiments for three typologically different languages to analyze if the structural sensitivity that was acquired from English eye-tracking data transfers to other languages. As input, we use 1,000 parallel sentences from the English, Korean and Turkish parts of the Parallel Universal Dependencies (PUD) treebanks which were randomly selected from Wikipedia and news articles (Zeman et al., 2017). We apply a 5-fold cross-validation setting with 800 sentences for training the probing regressors for each language and the remaining 200 for testing. We use the same architecture as described in Section 3.2, but freeze the encoder model and only update the final regression layer during training. The regression layer contains nine probing heads (one for each linguistic feature) and is trained for 5 epochs.[4]

6.2 Results

We report the results of the probing experiments and the model performance on scrambled inputs.

Probing Figure 5 shows the relative probing performance for each complexity feature. We see that fine-tuning yields the largest improvements for probing sentence length and average dependency link length. For the other complexity features, we see that the fine-tuned representations yield little to no improvement in probing accuracy compared to the pre-trained representations. This mostly concerns the features for which sentence length is factored out, i.e., average word frequency, average word length and lexical density. Sarti et al. (2021) report similar results and show that increased probing performance for dependency features persists for sentences of the same length. This provides additional evidence that structural information is learned in addition to low-level length information.

We observe only minor differences in probing accuracy for individual complexity features of English, Korean and Turkish sentences. The general pattern is consistent for all languages: features related to the structural complexity of sentences are more easily predicted after fine-tuning on eye-tracking metrics. This indicates that the fine-tuned model is able to transfer structural complexity knowledge acquired from English eye-tracking data to other languages.

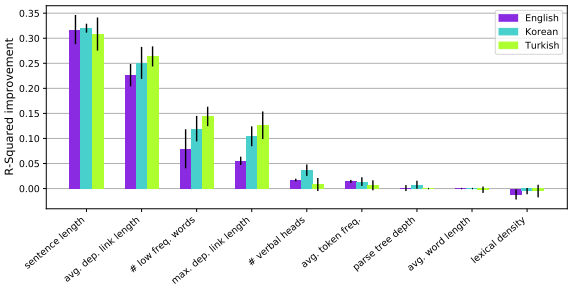

Figure 5: Relative improvement in R^2 for complexity features of English, Korean and Turkish sentences in fine-tuned XLM-R sentence representations over pre-trained representations. The results are calculated using probing regressors and averaged over 5 folds.

Influence of Word Order We compare the performance of the fine-tuned model on sentences with normal versus scrambled word order, both in terms of explained variance and R^2. We measure similar explained variance scores for both input types. This indicates that the model is able to account for a large portion of the variance in our eye-tracking data by merely considering sentence length. The R^2 scores, on the other hand, are consistently lower for scrambled inputs, as shown for total fixation duration in Figure 6 (see appendix Figure 10 for the other eye-tracking metrics). We conclude that the model is sensitive to word order and bases its eye-tracking predictions not only on sentence length but also on more complex structural characteristics.

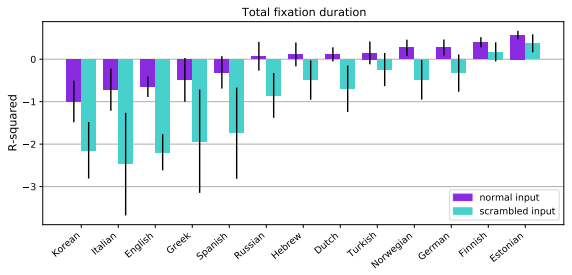

Figure 6: R^2 scores for total fixation duration for each language in MECO, both for sentences with normal and scrambled word order. The results are averaged over 5 folds; error bars denote the standard deviation.

[4]We report results for a multi-task set-up for probing in line with Sarti et al. (2021) and use the same hyperparameters as for the fine-tuning experiments but without intermediate evaluation on a development set. We also ran single-task probing as a sanity check and obtained similar results.

7 Conclusion

We find that XLM-R can apply cross-lingual transfer to predict cognitive processing difficulty with similar performance across 13 typologically diverse languages, despite being fine-tuned only on English data. We conducted a range of experiments to quantify the model's sensitivity to structural complexity and find that the fine-tuned model prominently encodes sentence length, but also considers more complex structural information such as dependency structure and word order for the prediction of eye-tracking metrics.

Our analyses suggest that domain differences in training and testing data have a greater impact on model performance than language differences within the same domain. More specifically, XLM-R performs better on in-domain GECO data than cross-domain MECO data, but within MECO, XLM-R shows similar performance across languages. This aligns with the findings of Morger et al. (2022), who show that the correlation between relative importance metrics and total fixation duration is influenced by text domain. Our study highlights the significance of controlling for text domain and size, as it allows to evaluate cross-lingual generalization that is independent of dataset characteristics.

In future work, we plan to better account for individual differences between readers (Brandl and Hollenstein, 2022) and spill-over effects across sentence boundaries (Wiechmann et al., 2022). The modeling approach for learning eye-tracking patterns also needs further exploration. We find that sentence-level prediction of eye-tracking patterns works well for learning about structural complexity, but that it is not optimal for capturing lexical complexity. Token-level measures, as predicted in Hollenstein et al. (2021), are more likely to be informative about lexical phenomena. A joint loss for sentence and token-level eye-tracking metrics might lead to sensitivity to a wider range of linguistic complexity features.

8 Limitations

The main limitation of our work is the use of relatively small datasets for testing our models due to limited availability of eye-tracking data in multiple languages. The dataset used for testing cross-lingual transfer (MECO) contains approximately 100 sentences per language. For probing structural complexity, we used a sample of 1,000 sentences per language.

As in related work, we averaged the eye-tracking metrics over readers to obtain a more robust indication of human reading behavior. This approach disregards the fact that reading is a highly individual process that is dependent on cognitive factors and experience. A computational model might develop a better sense of linguistic complexity when it learns about the linguistic properties that lead to variation across readers and we are working towards methods for integrating this information.

Acknowledgements

We thank the anonymous reviewers for their insightful feedback. L. Beinborn's research was supported by the Dutch National Science Organisation (NWO) through the projects CLARIAHPLUS (CP-W6-19-005) and VENI (Vl.Veni.211C.039).

References

Wasi Ahmad, Haoran Li, Kai-Wei Chang, and Yashar Mehdad. 2021. Syntax-augmented multilingual BERT for cross-lingual transfer. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4538–4554, Online. Association for Computational Linguistics.

Maryam AlJassmi, Kayleigh Warrington, Victoria McGowan, Sarah White, and Kevin Paterson. 2022. Effects of word predictability on eye movements during Arabic reading. *Attention, Perception, & Psychophysics*, 84(1):10–24.

Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. 2020. On the cross-lingual transferability of monolingual representations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4623–4637, Online. Association for Computational Linguistics.

Maria Barrett, Frank Keller, and Anders Søgaard. 2016. Cross-lingual transfer of correlations between parts of speech and gaze features. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1330–1339, Osaka, Japan. The COLING 2016 Organizing Committee.

Lisa Beinborn and Rochelle Choenni. 2020. Semantic drift in multilingual representations. *Computational Linguistics*, 46(3):571–603.

Thomas Bever. 1970. *The Cognitive Basis for Linguistic Structures*, pages 279–352. Cognition and the Development of Language.

Johannes Bjerva, Robert Östling, Maria Han Veiga, Jörg Tiedemann, and Isabelle Augenstein. 2019. What Do Language Representations Really Represent? *Computational Linguistics*, 45(2):381–389.

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

Dominique Brunato, Andrea Cimino, Felice Dell'Orletta, Giulia Venturi, and Simonetta Montemagni. 2020. Profiling-UD: a tool for linguistic profiling of texts. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 7145–7151, Marseille, France. European Language Resources Association.

Marc Brysbaert and Boris New. 2009. Moving beyond Kucera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior research methods*, 41:977–90.

Yuan Chai, Yaobo Liang, and Nan Duan. 2022. Cross-lingual ability of multilingual masked language models: A study of language structure. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4702–4712, Dublin, Ireland. Association for Computational Linguistics.

Rochelle Choenni and Ekaterina Shutova. 2022. Investigating language relationships in multilingual sentence encoders through the lens of linguistic typology. *Computational Linguistics*, 48(3):635–672.

Charles Clifton and Adrian Staub. 2011. Syntactic influences on eye movements during reading. In *The Oxford Handbook of Eye Movements*, pages 896–909. Oxford University Press.

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, Online. Association for Computational Linguistics.

Uschi Cop, Nicolas Dirix, Denis Drieghe, and Wouter Duyck. 2016. Presenting GECO: An eyetracking corpus of monolingual and bilingual sentence reading. *Behavior Research Methods*, 49.

Andrea De Varda and Roberto Zamparelli. 2022. Multilingualism encourages recursion: a transfer study with mBERT. In *Proceedings of the 4th Workshop on Research in Computational Linguistic Typology and Multilingual NLP*, pages 1–10, Seattle, Washington. Association for Computational Linguistics.

Ameet Deshpande, Partha Talukdar, and Karthik Narasimhan. 2022. When is BERT multilingual? isolating crucial ingredients for cross-lingual transfer. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3610–3623, Seattle, United States. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

David Eberhard, Gary Simons, and Charles Fennig (eds.). 2022. *Ethnologue: Languages of the World*, twenty-fifth edition. SIL International, Dallas, Texas.

Sidney Evaldo Leal, João Marcos Munguba Vieira, Erica dos Santos Rodrigues, Elisângela Nogueira Teixeira, and Sandra Aluísio. 2020. Using eye-tracking data to predict the readability of Brazilian Portuguese sentences in single-task, multi-task and sequential transfer learning approaches. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5821–5831, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Hila Gonen, Shauli Ravfogel, Yanai Elazar, and Yoav Goldberg. 2020. It's not Greek to mBERT: Inducing word-level translations from multilingual BERT. In *Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 45–56, Online. Association for Computational Linguistics.

Ana Valeria González-Garduño 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, Copenhagen, Denmark. Association for Computational Linguistics.

Peter Gordon, Randall Hendrick, Marcus Johnson, and Yoonhyoung Lee. 2006. Similarity-based interference during language comprehension: Evidence from eye tracking during reading. *Journal of experimental psychology. Learning, memory, and cognition*, 32:1304–21.

Raffaele Guarasci, Stefano Silvestri, Giuseppe De Pietro, Hamido Fujita, and Massimo Esposito. 2022. BERT syntactic transfer: A computational experiment on Italian, French and English languages. *Comput. Speech Lang.*, 71(C).

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

Nora Hollenstein, Itziar Gonzalez-Dios, Lisa Beinborn, and Lena Jäger. 2022b. Patterns of text readability in human and predicted eye movements. In *Proceedings of the Workshop on Cognitive Aspects of the Lexicon*, pages 1–15, Taipei, Taiwan. Association for Computational Linguistics.

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

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

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

Taku Kudo and John Richardson. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, Brussels, Belgium. Association for Computational Linguistics.

Anna K. Laurinavichyute, Irina A. Sekerina, Svetlana Alexeeva, Kristine Bagdasaryan, and Reinhold Kliegl. 2019. Russian Sentence Corpus: Benchmark measures of eye movements in reading in Russian. *Behavior Research Methods*, 51:1161–1178.

Jindřich Libovický, Rudolf Rosa, and Alexander Fraser. 2020. On the language neutrality of pre-trained multilingual representations. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1663–1674, Online. Association for Computational Linguistics.

Pierre Lison and Jörg Tiedemann. 2016. OpenSubtitles2016: Extracting large parallel corpora from movie and TV subtitles. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 923–929, Portorož, Slovenia. European Language Resources Association (ELRA).

Chi-Liang Liu, Tsung-Yuan Hsu, Yung-Sung Chuang, and Hung-Yi Lee. 2020. A study of cross-lingual ability and language-specific information in multilingual bert. *arXiv preprint arXiv:2004.09205*.

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.

Saima Malik-Moraleda, Dima Ayyash, Jeanne Gallée, Josef Affourtit, Malte Hoffmann, Zachary Mineroff, Olessia Jouravlev, and Evelina Fedorenko. 2022. An investigation across 45 languages and 12 language families reveals a universal language network. *Nature Neuroscience*, 25:1–6.

Felix Morger, Stephanie Brandl, Lisa Beinborn, and Nora Hollenstein. 2022. A cross-lingual comparison of human and model relative word importance. In *Proceedings of the 2022 CLASP Conference on (Dis)embodiment*, pages 11–23, Gothenburg, Sweden. Association for Computational Linguistics.

Elisabeth Norcliffe, Alice C. Harris, and T. Florian Jaeger. 2015. Cross-linguistic psycholinguistics and its critical role in theory development: early beginnings and recent advances. *Language, Cognition and Neuroscience*, 30(9):1009–1032.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

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, Florence, Italy. Association for Computational Linguistics.

Taraka Rama, Lisa Beinborn, and Steffen Eger. 2020. Probing multilingual BERT for genetic and typological signals. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 1214–1228, Barcelona, Spain (Online). International Committee on Computational Linguistics.

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

Noam Siegelman, Sascha Schroeder, Cengiz Acartürk, Hee-Don Ahn, Svetlana Alexeeva, Simona Amenta, Raymond Bertram, Rolando Bonandrini, Marc Brysbaert, Daria Chernova, Sara Maria Da Fonseca, Nicolas Dirix, Wouter Duyck, Argyro Fella, Ram Frost,

Carolina A Gattei, Areti Kalaitzi, Nayoung Kwon, Kaidi Lõo, Marco Marelli, Timothy C Papadopoulos, Athanassios Protopapas, Satu Savo, Diego E Shalom, Natalia Slioussar, Roni Stein, Longjiao Sui, Analí Taboh, Veronica Tønnesen, Kerem Alp Usal, and Victor Kuperman. 2022. Expanding horizons of cross-linguistic research on reading: The multilingual eye-movement corpus (meco). *Behavior Research Methods*, page 1–21.

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

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

Marc Tanti, Lonneke van der Plas, Claudia Borg, and Albert Gatt. 2021. On the language-specificity of multilingual BERT and the impact of fine-tuning. In *Proceedings of the Fourth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 214–227, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Daniel Wiechmann, Yu Qiao, Elma Kerz, and Justus Mattern. 2022. Measuring the impact of (psycho-)linguistic and readability features and their spill over effects on the prediction of eye movement patterns. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5276–5290, Dublin, Ireland. Association for Computational Linguistics.

Paula M. Winke. 2013. *Eye-Tracking Technology for Reading*, chapter 62. John Wiley & Sons, Ltd.

Shijie Wu and Mark Dredze. 2019. Beto, bentz, becas: The surprising cross-lingual effectiveness of BERT. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 833–844, Hong Kong, China. Association for Computational Linguistics.

Daniel Zeman, Martin Popel, Milan Straka, Jan Hajič, Joakim Nivre, Filip Ginter, Juhani Luotolahti, Sampo Pyysalo, Slav Petrov, Martin Potthast, Francis Tyers, Elena Badmaeva, Memduh Gokirmak, Anna Nedoluzhko, Silvie Cinková, Jan Hajič jr., Jaroslava Hlaváčová, Václava Kettnerová, Zdeňka Urešová, Jenna Kanerva, Stina Ojala, Anna Missilä, Christopher D. Manning, Sebastian Schuster, Siva Reddy, Dima Taji, Nizar Habash, Herman Leung, Marie-Catherine de Marneffe, Manuela Sanguinetti, Maria Simi, Hiroshi Kanayama, Valeria de Paiva, Kira Droganova, Héctor Martínez Alonso, Çağrı Çöltekin, Umut Sulubacak, Hans Uszkoreit, Vivien Macketanz, Aljoscha Burchardt, Kim Harris, Katrin Marheinecke, Georg Rehm, Tolga Kayadelen, Mohammed Attia, Ali Elkahky, Zhuoran Yu, Emily Pitler, Saran Lertpradit, Michael Mandl, Jesse Kirchner, Hector Fernandez Alcalde, Jana Strnadová, Esha Banerjee, Ruli Manurung, Antonio Stella, Atsuko Shimada, Sookyoung Kwak, Gustavo Mendonça, Tatiana Lando, Rattima Nitisaroj, and Josie Li. 2017. CoNLL 2017 shared task: Multilingual parsing from raw text to Universal Dependencies. In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 1–19, Vancouver, Canada. Association for Computational Linguistics.

A Additional Tables and Figures

Dataset	Language	#Words	#Sentences	Avg. sent. length	Avg. word length
GECO	English	52131	4041	12.90	4.60
MECO	English	2092	99	21.13	5.32
	Dutch	2226	112	19.88	5.54
	German	2019	115	17.56	6.38
	Finnish	1462	110	13.29	8.19
	Estonian	1542	112	13.77	7.35
	Norwegian	2106	116	18.16	5.62
	Italian	2111	90	23.46	5.70
	Spanish	2412	98	24.61	5.01
	Greek	2082	99	21.03	5.67
	Turkish	1696	104	16.31	6.92
	Russian	1827	101	18.09	6.53
	Hebrew	1943	121	16.06	4.89
	Korean	1699	101	16.82	3.21

Table 3: Size characteristics for the reading materials of GECO and MECO. GECO sentences which are shorter than five words are removed to ensure that the model sees an adequate amount of complex structures during training.

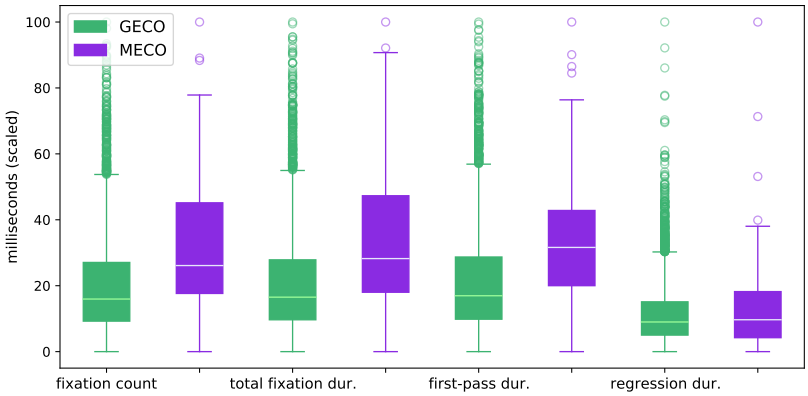

Figure 7: Distribution of four sentence-level eye-tracking metrics in English parts of GECO and MECO. All metrics are scaled between 0-100.

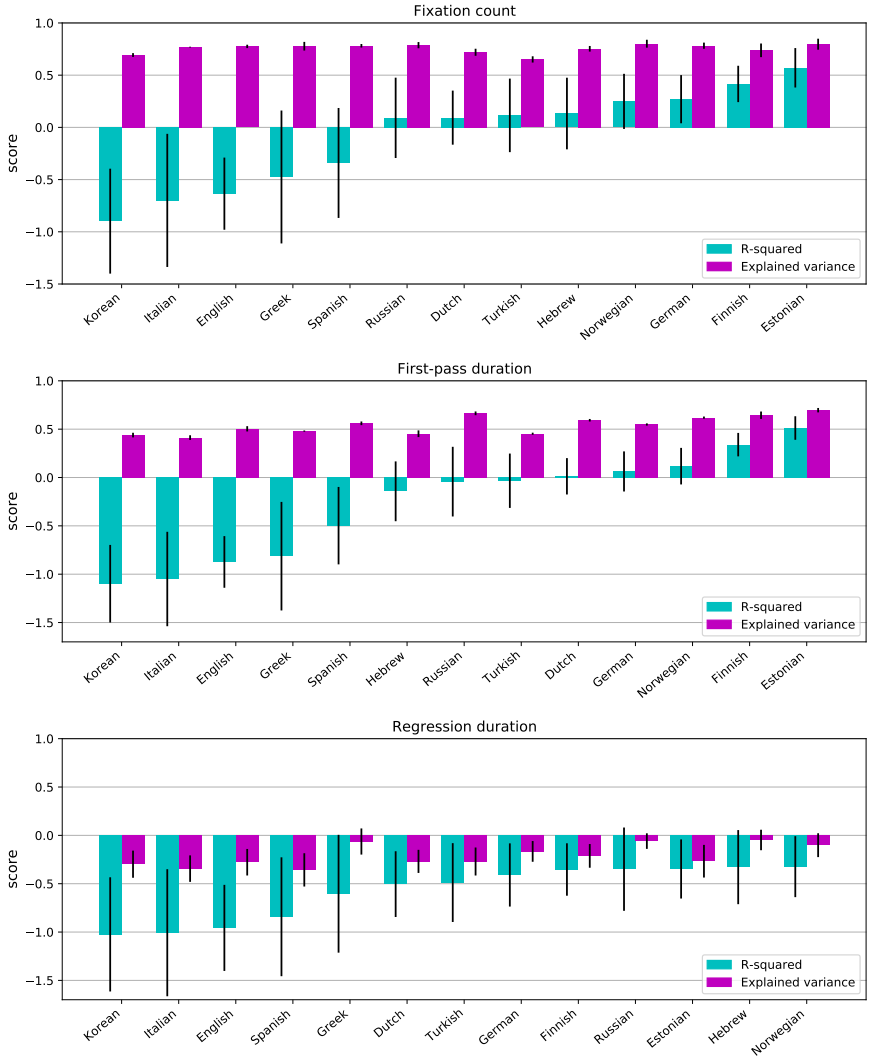

Figure 8: Cross-lingual transfer results for predicting cognitive processing complexity (i.e. fixation count, first-pass duration and regression duration). Prediction performance is evaluated with explained variance and R^2 for each language in MECO. The results are averaged over 5 folds; error bars denote the standard deviation over folds.

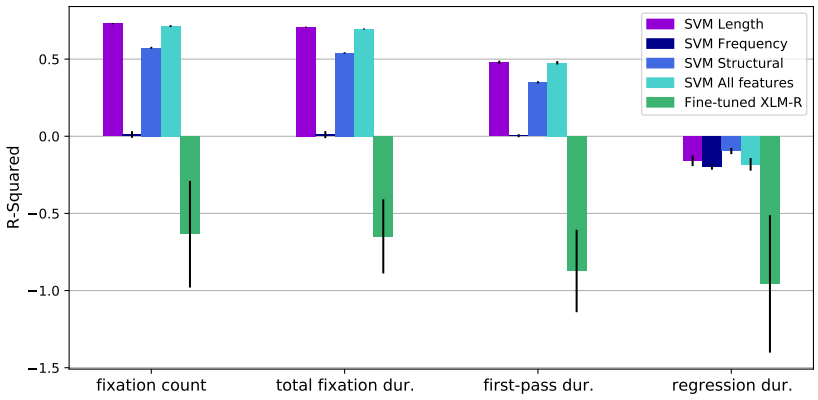

Figure 9: R^2 of the four feature-based SVM models and the fine-tuned XLM-R model. The models are trained on GECO using 5-fold cross-validation and evaluated on the English part of MECO; error bars denote the standard deviation over folds.

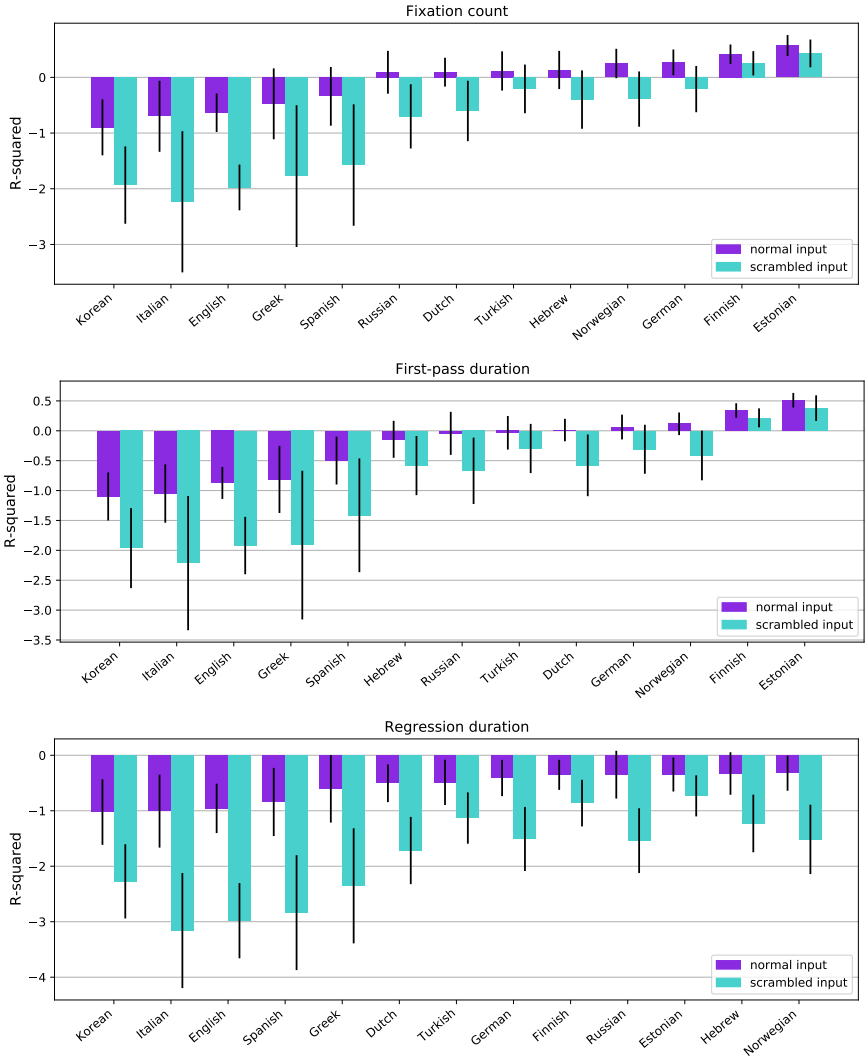

Figure 10: R^2 for fixation count, first-pass duration and regression duration for each language in MECO, both for sentences with normal and scrambled word order. The results are averaged over 5 folds; error bars denote the standard deviation.