Research Master Thesis

Cross-lingual Transfer of Correlations between Linguistic Complexity and Human Reading Behaviour

Charlotte M. Pouw

a thesis submitted in partial fulfilment of the requirements for the degree of

MA Linguistics
(Human Language Technology)

Vrije Universiteit Amsterdam

Computational Lexicology and Terminology Lab
Department of Language and Communication
Faculty of Humanities

Supervised by: Lisa Beinborn
2^{nd} reader: Antske Fokkens

Submitted: July 1, 2022

Abstract

When humans read a text, their eye movements are influenced by linguistic characteristics of the input. For example, readers tend to fixate longer on infrequent and morphologically complex words, and regress to previous material if a syntactic structure is difficult to parse. Such effects have been observed in many different languages (e.g. Russian (Laurinavichyute et al., 2019), German (Kliegl et al., 2004), Finnish (Bertram and Hyönä, 2003)). Eye movement patterns of reading thus provide important clues about linguistic complexity. Recent studies have shown that transformer-based language models are remarkably good at predicting human reading behaviour, even for languages that are not seen during training (Hollenstein et al., 2021b). This suggests that such models are cognitively plausible and process linguistic complexity in a similar way as humans.

This thesis investigates if a multilingual transformer model (XLM-RoBERTa, (Conneau et al., 2020)) develops a sensitivity to linguistic complexity when it learns to predict patterns of human reading behaviour. After training the model on eye-tracking data of English readers, we find that it can accurately predict eye movement behaviour associated with 1) sentences that are more complex than those seen during training, and 2) languages that are not seen seen during training. These generalization abilities indicate that the model established a link between linguistic complexity and eye movement patterns, and that the learned correlations can be transferred to other languages. We provide further evidence for this by probing the linguistic knowledge that is encoded in the model's final-layer representations, both before and after fine-tuning on eye-tracking data. We find that features associated with the structural complexity of a sentence are better encoded after fine-tuning.

Declaration of Authorship

I, Charlotte Maria Pouw, declare that this thesis, titled *Cross-lingual Transfer of Correlations between Linguistic Complexity and Human Reading Behaviour* and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a degree at this University.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

Date:

Signed:

Acknowledgments

I want to express my sincere gratitude to dr. Lisa Beinborn, both for the dedicated supervision during this thesis project and for being my mentor throughout the entire master. Your feedback was always on point and I left every thesis meeting feeling encouraged and motivated to continue. I admire you both as a researcher and as a person, and I look forward to reading your future publications.

I also want to thank dr. Nora Hollestein for meeting with me during her visit in Amsterdam and for sharing her expertise knowledge about cognitively inspired natural language processing with me.

Thank you to Gabriele Sarti for writing well-documented code that I could adapt for this project, and for quickly answering my emails when I got stuck.

Many thanks to my peers Alessandra and Eliza for making the master so much more enjoyable, especially during the lockdowns.

Finally, I am incredibly grateful for all the support I received from my friends and family. Special thanks to Vincent for the endless encouragement, and of course for helping me make my plots more aesthetically pleasing.

List of Figures

List of Tables

Contents

Chapter 1

Introduction

Reading is a foundational skill for acquiring new information. Many sources of information are only available in written form, including educational material, news paper articles and letters from municipalities. Although many people learn how to read as a child, not everyone becomes equally skilled at it. In the Netherlands alone, more than 2.5 million people are low-literate, which means that they have trouble with reading or writing.[1] This hinders them from succeeding in education, applying to jobs, and staying informed about the news.

One way to alleviate this problem is to address text complexity. Texts that contain many infrequent words and long sentences are difficult to read, especially for low-literate people or language learners. By estimating text complexity, we can select texts that are sufficiently easy for a particular target audience. Natural Language Processing (NLP) technologies allow for the automatic assessment of text complexity. Given a corpus of texts that are annotated with their respective level of complexity, NLP models can learn correlations between textual characteristics and complexity levels. Such models can then be used to automatically assess the complexity of large amounts of unseen texts.

While this sounds promising, it is not straightforward to design a training corpus with an appropriate annotation scheme for text complexity. Many corpora divide their texts into discrete categories of complexity (often on a scale), but it is not straightforward to decide on the granularity of such categories. A large amount of categories makes it hard for annotators to determine the boundaries between them. Only using a binary distinction between easy and difficult texts, on the other hand, might not be fine-grained enough. A model trained to predict binary complexity labels might solely learn that easy texts are shorter than difficult texts, thus overlooking more sophisticated features of text complexity. Finally, the annotations need to be provided by trained annotators, which may be expensive and hard to find, especially for low-resource languages.

An alternative way of assessing which parts of a text cause reading difficulty is to examine how the eyes move during reading. Decades of psycholinguistic research have shown that people look longer at words that require more cognitive processing effort (e.g. infrequent or ambiguous words), and that the eyes regress to previous material if a grammatical structure is difficult to parse. As an illustration, Figure 1.1 shows how long English readers fixate on each word of the sentence *He lived at home while pursuing literary ambitions*, taken from the Ghent Eye-tracking Corpus (Cop et al., 2016). We

[1] `https://www.lezenenschrijven.nl/reading-and-writing-foundation`

see that the final words "pursuing literary ambitions" are fixated much longer than the previous words, and the preposition "at" is even entirely skipped. This shows that the cognitive effort it takes to process a word is reflected in fixation durations. Eye-tracking data recorded during reading can thus serve as a proxy for linguistic complexity.

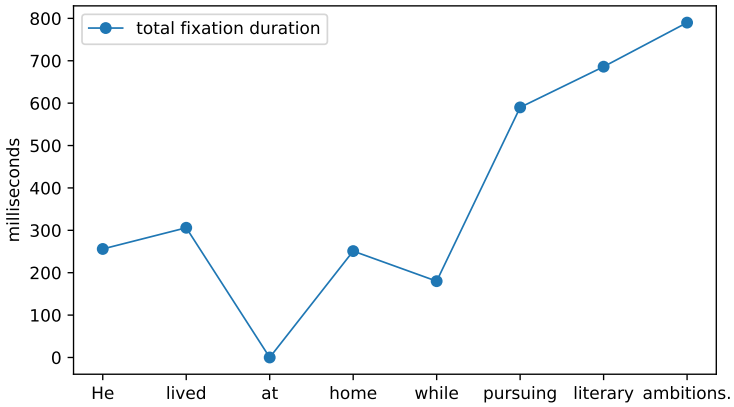

Figure 1.1: Eye movement pattern for an example sentence from the Ghent Eye-tracking Corpus. The fixation durations are averaged over readers (n=14).

Certain elements of linguistic complexity trigger consistent eye movement patterns across languages. For example, word length and word frequency influence fixation durations regardless of the language (Kliegl et al., 2004; Laurinavichyute et al., 2019). When we control for such factors, cross-lingual reading behaviour exhibits striking similarities. To illustrate, Figure 1.2 shows that native speakers of five typologically different languages fixate equally long on the sentence *The ancient Greeks had no equivalent to Janus, whom the Romans claimed as distinctively their own*, translated to their own language. Since the sentence has approximately the same content and linguistic complexity in each language, the amount of cognitive effort it takes to process the sentence remains stable. The fixation durations reflect this. Thus, speakers of different languages respond to linguistic complexity in a rather universal way.

Modern eye-trackers capture eye movements with a temporal accuracy of milliseconds. Eye-tracking data recorded during reading therefore provides a very fine-grained image of the linguistic properties that cause processing difficulty during reading. Such data can be obtained from native speakers directly and alleviates the need for expert annotators. In addition, low-cost eye-trackers are increasingly available and improving in quality. Collecting eye-tracking data might therefore become as easy as having a native speaker read a text from a mobile phone, tablet or laptop (Krafka et al., 2016).

By learning to predict how readers will move their eyes over a text, a computational model might develop a similar sensitivity to linguistic complexity as humans. In other words, eye movement data can adjust a model's *inductive bias* (i.e. the set of assumptions that the model relies on to map an input to an output) to become more like human language processing. When a model learns to predict eye movement patterns, it needs to learn correlations between linguistic characteristics and increased fixation durations or regressive eye movements. This relationship needs to be established for rather small linguistic units. Therefore, learning to predict eye movements should lead to more sophisticated knowledge about linguistic complexity than learning to predict single text complexity labels for long texts.

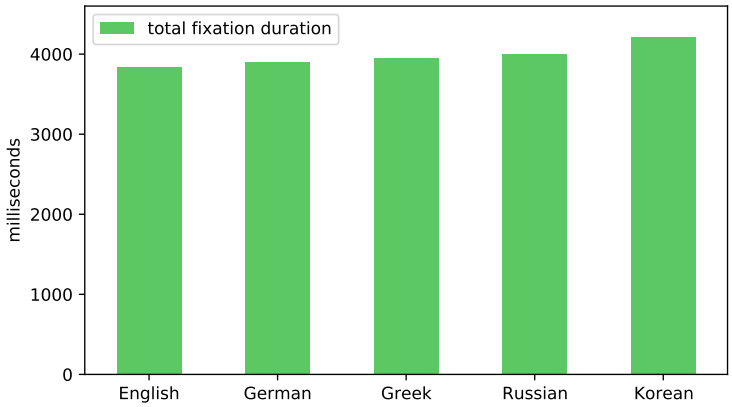

Figure 1.2: Total fixation durations for the sentence *The ancient Greeks had no equivalent to Janus, whom the Romans claimed as distinctively their own*, translated in five languages. The data is taken from the Multilingual Eye-tracking Corpus and the fixation durations are averaged over readers (32-46 per language).

In fact, there is evidence that learning eye movement patterns associated with reading facilitates the process of predicting text complexity. González-Garduño and Søgaard (2017) find that models that simultaneously learn to predict sentence-level eye-tracking metrics and complexity labels perform better than models that merely learn to predict complexity labels. Similar results have been found for other languages. Evaldo Leal et al. (2020) show that sequentially training the same model on complexity labels and fixation durations of Portuguese sentences leads to a better performance than when eye movement behaviour is not learned. Both results indicate that neural models can pick up fine-grained information about text complexity from eye-tracking data recorded during reading, and that this information cannot be learned from discrete complexity labels alone.

More recently, studies have found that transformer-based language models are remarkably good at predicting eye movements associated with reading, even outperforming linguistically motivated models that receive explicit features correlating with eye movement patterns as input (Hollenstein et al., 2021b; Sarti et al., 2021). This suggests that these models are *cognitively plausible*, i.e. they seem to process linguistic complexity in a similar way as humans. Multilingual transformer models are also able to accurately predict eye movements for unseen languages, possibly resulting from the fact that certain linguistic phenomena trigger consistent eye movement behaviour across languages. However, the exact linguistic knowledge that is picked up by multilingual language models as a result of learning to predict reading behaviour has not been analysed yet.

1.1 Research questions and objectives

In the current project, we investigate how well a state-of-the-art multilingual transformer model (XLM-RoBERTa, Conneau et al. (2020)) can predict eye movement patterns associated with reading across different domains and languages. We then analyse whether its performance can be explained by a newly acquired sensitivity to linguistic complexity. The following research questions are addressed:

1. **Cross-domain abilities**: Can eye movement patterns be predicted for sentences that come from a different domain and are linguistically more complex than those seen during training?

2. **Cross-lingual abilities**: Can eye movement patterns be predicted for languages that are not seen during training?

3. **Sensitivity to linguistic complexity**: Can high prediction accuracy for eye movement patterns be explained by an increased sensitivity to linguistic complexity?

1.2 Contributions

The project is novel in the following ways:

- We focus on the linguistic features underlying eye movement patterns during reading, and provide a thorough analysis of the relationship between linguistic complexity and several eye-tracking metrics in two different corpora. We show that the distribution of eye-tracking metrics depends on the **linguistic complexity** of the reading materials, and not on the **language** in which the reading materials were written, highlighting the universal tendencies of cross-lingual reading behaviour;

- We examine whether XLM-RoBERTa picks up the universality of cross-lingual reading behaviour by training it on English reading data and evaluating it on a range of typologically diverse languages from the newly released **Multilingual Eye-tracking Corpus** (Siegelman et al., 2022), which contains eye movement data for parallel texts in 13 different languages. By using the MECO corpus, we rule out the possibility that different prediction accuracy across languages is caused by differences in semantics of the reading materials;

- We contribute to the growing field of **interpretability** in NLP by analysing the linguistic knowledge that is encoded in XLM-RoBERTa's sentence representations, both before and after learning eye movement patterns of reading. In addition, we examine if the encoded linguistic information differs depending on the representational variants of the model.

1.3 Outline

Chapter 2 provides the reader with the necessary background information about eye-tracking research, and explains how eye-tracking data has been utilized for NLP. In addition, it discusses how linguistic knowledge can be analysed in large pre-trained language models. Chapter 3 provides an analysis of the eye-tracking corpora that we use for training and evaluating XLM-RoBERTa, focusing on the linguistic complexity of the reading materials and the corresponding distribution of several eye-tracking metrics. Chapter 4 discusses the experiments that were carried out to test XLM-RoBERTa's abilities to predict eye movement patterns across domains and languages. Chapter 5 analyses whether XLM-RoBERTa acquired a sensitivity to linguistic complexity as a result of learning to predict eye movement patterns (which should explain the results

reported in Chapter 4). Finally, the thesis ends with conclusions and suggestions for future work.

Chapter 2

Background and Related Work

Eye movement behaviour during reading has been studied for more than a century. The very first eye-tracking studies date back to the late 1800s. Back then, reading patterns were detected by placing a rubber tube on the eyelid and subsequently recording the sounds that were produced when the eyes moved forward and backward over a text (Hyönä et al., 1995). Nowadays, eye movements can be recorded with a temporal accuracy of milliseconds, which allows for the careful examination of textual characteristics and language processing during reading.

In recent years, NLP researchers have recognised that eye-tracking data recorded during reading provides important clues about human language processing. Such data can either be used to improve the performance of NLP models, or to compare the language processing strategies of computational models and humans. This chapter discusses the linguistic information that eye movement patterns of reading provide, and analyses studies that have incorporated such information in NLP models.

2.1 Basic notions in eye-tracking research

The *eye-mind hypothesis* states that "there is no appreciable lag between what is fixated by the eye and what is processed by the mind" (Carpenter and Just, 1983). This hypothesis suggests that fixation patterns on a text reveal which linguistic units are being processed. By recording eye movement patterns during reading and linking them to specific linguistic phenomena, we can develop theories about the processing of written language by the human brain. The following sections provide an overview of some the most foundational findings from eye-tracking research.

2.1.1 Eye movements during reading

When we read a piece of text, we might feel as though our eyes move smoothly from line to line. In reality, however, the eyes alternate between rapid, abrupt motions called *saccades* and relatively stable periods called *fixations*. Saccades typically last 20-40 ms, while fixations last 200-250 ms on average. Intake of visual input only happens during fixations, since the eyes move so fast during saccades that only a blur can be seen. However, we do not perceive this blur since the brain is still processing input that is available before and after the saccade. This phenomenon, where a reader is temporarily blind to visual input, is called *saccadic suppression* (Rayner, 1998; Hyönä and Kaakinen, 2019).

Saccades usually go in the normal reading direction, e.g. from left to right for English reading. Saccades can also go in the opposite direction and are known as *regressions*. Regressions typically occur when there is a misunderstanding or ambiguity at the current position of the text that can only be resolved with previous information. A regression can also correct a saccade that was launched too far ahead in the text (Rayner, 1998; Hyönä and Kaakinen, 2019).

There are several concrete eye-tracking metrics that are designed to capture language processing at different stages. These metrics are usually calculated at the word level, but can also cover larger linguistic units such as phrases or sentences. Early metrics target low-level processes such as word recognition, while late metrics aim to capture high-level processes such as syntactic integration (Hyönä and Kaakinen, 2019). For early processing, researchers often measure *first-pass duration*, which is the duration of the very first fixation on a target region. During late processing, target regions may be refixated. Researchers thus measure the *total fixation duration* (also known as *gaze duration*), which is the sum of all fixation durations on a target region. Late processing can also be measured by *fixation count*, defined as the total number of fixations on a target region.

To capture contextual effects, researchers often measure *go-past time*, which quite literally measures the time it takes to "go past" a target region. Concretely, this metric is the sum of all fixation and regression durations before the reader progresses to the right of the target region. A variation of this metric is *selective go-past time*, which excludes regressions that are launched from the target region. When subtracting selective go-past time from go-past time, one is left with *regression duration*.

Eye-tracking studies are usually conducted with multiple participants. This allows for transforming *absolute* eye-tracking metrics into *probabilities*. For example, one can measure the *fixation probability* for a target region by averaging boolean values (i.e. 1 if the region was fixated at least once, 0 if not) over participants. Eye-tracking metrics that are averaged across participants are more robust to individual differences between readers and thus capture reading behaviour in a more generalized fashion.

2.1.2 Linguistic properties affecting eye movements

As shown in Figure 1.1, fixation durations vary from word to word and some words are entirely skipped. This can be explained by the amount of cognitive effort it takes to process the linguistic input (Vasishth et al., 2013). Certain linguistic properties increase the processing effort required to understand a text. Psycholinguistic research has carefully examined the correlations between specific linguistic properties and specific eye movement patterns during reading.

At the word level, a well known factor that influences both early and late processing is **frequency**. Infrequent words are fixated longer and more often than frequent words (e.g. Hyönä et al. (1995); Kliegl et al. (2004)). (Rayner and Duffy, 1986). A similar effect is caused by **age of acquition**: words that are acquired earlier in life are read faster than words that are acquired later in life (Dirix and Duyck, 2017; Juhasz and Sheridan, 2019). These effects have been attributed to *lexical access*, which "refers to the retrieval of words from the mental lexicon, both in recognition and in production." (Taft, 2001). Words that are frequent or acquired at a young age can easily be retrieved from the lexicon and therefore take less time to read than words that are infrequent or acquired later in life.

Orthographic features can also affect fixation durations. An important factor is **word length**: longer words receive longer fixations than shorter words, simply because there are more letters to be processed. It should be noted that word length and word frequency often go hand in hand. For example, function words such as prepositions and determiners are skipped more often than content words (Carpenter and Just, 1983; Duffy et al., 1988) since they tend to be short, but also because they appear frequently in the text. Nonetheless, the length effect has also been observed when frequency remains constant (Liversedge et al., 2004) and vice versa (Kliegl et al., 2004).

Moving beyond the word level, contextual factors also affect eye movement behaviour. One such factor is **predictability**: words that are highly predictable from the context are read faster than unpredictable words and are often entirely skipped. The predictability of a word increases as a function of the constraints given by previous words. For example, the verb *eat* is most likely to be followed by an eadable object such as *cake*, while a verb such as *move* can be followed by a larger variety of objects (Altmann and Kamide, 2000). Relatedly, Morris (1994) finds that fixation durations decrease when previously read words are semantically related to the current word, because the preceding words increase the predictability of the current word. Finally, **lexical ambiguity** can also result in longer fixation durations, but polysemous words are only fixated longer when both meanings of the word are equally likely in the given context (Rayner and Duffy, 1986; Duffy et al., 1988).

Syntactic ambiguity can also trigger specific eye movement patterns. Interesting effects have been found for so-called "garden-path" sentences (Frazier and Rayner, 1982). Consider the famous example *The horse raced past the barn fell* (Bever, 1970). During first-pass reading, a reader initially adopts the simplest syntactic structure, where *raced* is the main verb and the prepositional phrase *past the barn* is attached to that verb. When encountering the final word *fell*, however, the reader realises that the initial analysis was wrong and that a re-analysis of the syntactic structure is needed. In eye-tracking research, this materializes as regressions towards the ambiguous region and an increase of fixations on that region (Clifton Jr and Staub, 2011).

Besides syntactic ambiguity, there are many other types of syntactic complexity that can cause processing difficulty. For example, Gordon et al. (2006) find that object-relative clauses such as *The banker that praised the barber climbed the mountain* are read more slowly and with more regressions than subject-relative clauses such as *The banker that the barber praised climbed the mountain*. Interestingly, they find that this effect is reduced when the object (*the barber*) is replaced by a name (*Sophie*). The authors speculate that this is caused by a reduced burden on a reader's **working memory**, since the subject and object are less similar in the latter case – *Sophie* is not similar to *the banker* and thus takes less effort to process than a highly similar object like *the barber*. Effects of "Chomskyan"[1] complexity, such as the number of nodes in a parse tree, have been less observed in eye-tracking studies. Nonetheless, sentences with a more complex tree structure are also less frequent and introduce more memory load, which does in fact influence eye movement patterns (Clifton Jr and Staub, 2011).

To summarise: linguistic features affecting text complexity, such as word length, frequency, predictability and ambiguity, as well as cognitive factors such as age of

[1]Chomskyan linguistics is a term that is commonly used to refer to the linguistic theory developed by Noam Chomsky, which is most famous for the idea that humans are born with an innate knowledge of linguistic structure, and that the grammar of all natural languages can be captured by formal mathematical descriptions.

acquisition and working memory constraints all affect cognitive processing effort, which in turn leads to variation in eye movement behaviour.

2.1.3 Cross-lingual differences

Effects of word length, frequency and predictability have been demonstrated in multiple languages (e.g. Russian (Laurinavichyute et al., 2019), German (Kliegl et al., 2004), Finnish (Bertram and Hyönä, 2003)). Only few studies, however, perform a systematic cross-lingual comparison, in which semantic variation is controlled. Liversedge et al. (2016) compare reading behaviour in English, Chinese and Finnish, which differ considerably in both linguistic and orthographic respects. Nonetheless, the authors find that sentences that are matched for content are read at a similar speed in all languages (which is in line with the data shown in Figure 1.2). This suggests that there is a common cognitive process that underlies constructing the meaning of a sentence, regardless of how that sentence is represented on paper.

Within sentences, the authors do find deviating eye movement behaviour for the three languages: Finnish readers make the most and shortest fixations, and launch the furthest outbound saccades. Chinese readers show the opposite behaviour, and English readers are inbetween. This reflects the average word length of the languages: Finnish has the longest words, followed by English, and then Chinese. Further, 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.

Very recently, Siegelman et al. (2022) add to these findings by comparing reading behaviour in a grand total of thirteen languages, using texts that are carefully matched for content. They find that fixation probability varies considerably across languages, which they also attribute to word length distributions in different languages (e.g. fixation probability is much higher for Dutch than for Korean). This suggests that universal eye movement patterns are more likely to be observed in sentence-level eye-tracking metrics. For the current study, this leads to the hypothesis that it will be easier for a multilingual language model to predict cross-lingual eye-tracking metrics at the sentence level than at the word level.

2.2 Using eye-tracking data for NLP

Eye movement data recorded during reading provides important clues about human language processing, which can potentially be picked up by NLP models. As mentioned in the introduction, it has been shown that learning eye movement patterns associated with reading facilitates the process of predicting text complexity (González-Garduño and Søgaard, 2017; Evaldo Leal et al., 2020). Similar results have been found for other NLP tasks, namely part-of-speech tagging, grammatical error detection, sentiment analysis and hate speech detection (Barrett et al., 2016a,b, 2018). In the following section, we describe these studies and analyse what their results imply about the specific linguistic information that eye movement data of reading provides. We then describe different modelling approaches that have been proposed for the direct prediction of eye movement patterns. Feature-based approaches provide insight in the linguistic features that are predictive of eye movement patterns, and neural approaches show which architectures are good at implicitly learning the features underlying eye

movement patterns.

2.2.1 Improving NLP models with eye-tracking data

Barrett et al. (2016a) augment a part-of-speech tagger with a range of eye-tracking features and find that this leads to a better performance than just using textual features. Interestingly, performance also improves when the textual features are completely left out and the tagger bases its predictions on the eye-tracking features alone. This indicates that human fixation patterns are influenced by parts of speech, and that these patterns can be used to classify words into part-of-speech categories. A follow-up study by Barrett et al. (2016b) finds that English eye-tracking features also improve a French part-of-speech tagger, which suggests that correlations between parts of speech and eye movements are consistent across languages. In the current study, we examine if this result extends to linguistic complexity.

A downside of incorporating eye-tracking data as features is that the eye-tracking features also need to be available at test time. An alternative approach is to employ multi-task learning, where eye movement behaviour is learnt as an auxiliary task. For example, in a study by Barrett et al. (2018), eye movement behaviour is learnt as an auxiliary task for three sequence labeling tasks: grammatical error detection, sentiment analysis, and hate speech detection. They train a bi-directional LSTM using an *alternating* training approach: either the model's parameters are updated in favor of a data point from the "main" corpus, or attention weights are altered in favor of a data point from an eye-tracking corpus. This way, the model's attention weights are *regularized* using eye movement behaviour. This approach significantly improves performance on all three sequence labeling tasks. Human fixation patterns can thus help a model to assess which parts of a text are important for a certain classification (e.g. humans look longer at the word *horrible* in the sentence *that movie was horrible*, which helps the model to assess that the sentence has negative sentiment).

2.2.2 Directly predicting eye movements

The CMCL 2021 Shared Task on Eye-Tracking Prediction (Hollenstein et al., 2021a) challenged researchers to predict five word-level eye-tracking metrics of the Zurich Cognitive Language Processing Corpus (ZuCo, (Hollenstein et al., 2018)), which contains eye-tracking data recorded from English native speakers during reading. For each word w, participating teams needed to predict: 1) the total number of fixations on w; 2) the duration of the first fixation on w; 3) the total fixation duration on w; 4) go-past time, i.e. the time it takes to move the eyes to the right of w; and 5) the proportion of participants that fixated w. For metrics 1-4, the average over all participants needed to be predicted. The submitted solutions were evaluated by calculating the Mean Absolute Error (MAE) (i.e. the absolute difference between predicted (y) and actual (x) values) for all five eye-tracking metrics. In addition, the solutions were compared to a mean baseline.

Interestingly, the best results were obtained using traditional machine learning algorithms trained with explicit features. Bestgen (2021), who submitted the best performing system, finds that surface features regarding the length and position of a word within a sentence are the most predictive for all eye-tracking features, followed by frequency features obtained from external corpora. In addition, bigram information (assumed to capture next word predictability), behavioural measures (e.g. reaction times during

lexical decision or naming tasks), and lexical features (capturing orthograpic and morphological information) were proven to be helpful for the task. The second best solution (Dary et al., 2021) also finds that surface and frequency features perform very well on their own, and that syntactic features (i.e. part-of-speech and dependency information) only add modest improvements. Nonetheless, their system outperforms Bestgen (2021) in predicting the total number of fixations, go-past time and total fixation duration per word, which indicates that the syntactic features are helpful for the prediction of eye movement behaviour during late processing.

Other teams approached the task with state-of-the-art language models based on the transformer architecture (Vaswani et al., 2017). Li and Rudzicz (2021) fine-tune RoBERTa (Liu et al., 2019) in two stages: first on data from an external English eye-tracking corpus (Provo, Luke and Christianson (2017)), and then on the target data from ZuCo. The fine-tuned representations are then fed into per-token regression heads, which simultaneously predict the five eye-tracking measures. This approach achieves third place in the shared task. Importantly, the authors also show that the RoBERTa-approach outperforms a simple linear regression baseline trained on four token-level surface and frequency features. This indicates that the fine-tuned RoBERTa representations encode more information than just simple surface and frequency cues.

Several lower-ranked submissions experiment with concatenating explicit features to transformer representations that are fine-tuned on the ZuCo data (Vickers et al., 2021; Oh, 2021; Choudhary et al., 2021; Yu et al., 2021). Interestingly, they all find that adding an explicit word length feature improves prediction accuracy, indicating that word length is not well encoded in the transformer representations after fine-tuning. This might be explained by the fact that the employed models split words into subwords, and that all teams used the representation of the first subword as input for the final prediction layer. In addition, Yu et al. (2021) find that concatenating a range of linguistic and behavioural features to fine-tuned BERT embeddings (Devlin et al., 2019) leads to more accurate predictions of the eye-tracking metrics than only using fine-tuned BERT embeddings. This indicates that the BERT embeddings do encode those features after fine-tuning on the ZuCo data. However, it remains unclear whether this is true for all of the concatenated features or only a subset of them.

As mentioned in the introduction of this thesis, Hollenstein et al. (2021b) show that transformer-based language models can successfully predict a range of eye-tracking metrics in four Indo-European languages: Dutch, English, German and Russian. To investigate the relationship between the complexity of the input sentences and the prediction accuracy for the eye-tracking metrics, they measure the correlation between the Flesch Reading Ease[2] (Flesch, 1948) and models' prediction accuracy before and after fine-tuning. They find that pre-trained models predict eye-tracking features more accurately for sentences with a lower Flesch score, and that this correlation disappears after fine-tuning. This suggests that the models learned a correlation between sentence complexity and eye movement patterns during fine-tuning. However, the authors note that the Flesch Reading Ease may not be a good proxy for complexity, since it only operates on word length and sentence length. Further research is needed to establish if sophisticated features of text complexity are captured by transformer models after fine-tuning on eye-tracking metrics (which is the goal of the current study).

[2]The Flesch Reading Ease is a measure of text complexity that captures average word length and average sentence length in a single number.

To summarise, the submissions for the Shared Task show that word-level eye-tracking metrics can be predicted using a combination of length, frequency, behavioural, lexical and syntactic features. While fine-tuned transformer models also yield high accuracy, they do not always outperform feature-based models and the addition of explicit features can improve their accuracy. Nonetheless, it has not been systematically analysed which linguistic information is (not) encoded in transformer representations after fine-tuning on eye-tracking metrics.

2.3 Linguistic knowledge in pre-trained language models

Current state-of-the-art language models are pre-trained on massive amounts of textual data. The idea is that such models implicitly learn about linguistic structure through pre-training, which would explain why they perform so well on many NLP tasks. As discussed in the previous sections, fine-tuning pre-trained language models on eye movement patterns of reading potentially leads to a better encoding of features associated with linguistic complexity. By analysing the linguistic knowledge that is acquired from pre-training alone, we can better understand the *additional* linguistic information that language models can pick up from eye-tracking data.

2.3.1 Probing implicit linguistic knowledge

A popular way of analysing the implicit linguistic knowledge that is encoded in language model representations is *probing*. When probing, a simple supervised model (often called a *diagnostic classifier* (Hupkes et al., 2018)) learns to predict a value for a given linguistic property based on a learned representation of a neural model. If the probe can accurately predict values for the linguistic property, the researcher may conclude that the property is encoded in the representation of the neural model.

More formally, probing can be defined as follows (Belinkov, 2022): Given a model $f : x \rightarrow \hat{y}$ that maps an input x to an output $\hat{y}$, we can denote intermediate representations of model f at layer l as $f_l(x)$. A probe takes such an intermediate representation as input and maps it to a particular property of interest $\hat{z}$, i.e. $g : f_l(x) \rightarrow \hat{z}$. If g can accurately predict property z given the representation $f_l(x)$, we may conclude that z is implicitly encoded in $f_l(x)$. Information-theoretically, probing is rather described as estimating the mutual information between representation $f_l(x)$ and property z (Pimentel et al., 2020). Examining such information is especially valuable for model developers, since it can reveal which linguistic knowledge a model is lacking and how it can be improved. In the current study, we use probing to investigate the linguistic information that is encoded before and after fine-tuning a multilingual transformer model on eye movement patterns of reading.

Observed linguistic phenomena Pimentel et al. (2020) investigate the encoding of part-of-speech information and dependency relations in the pre-trained representations of multilingual BERT. They obtain sentence representations of eleven typologically diverse languages from the BERT model and feed them into a Multi-Layer Perceptron (i.e. the probe), which then has to predict the part-of-speech tags and dependency labels of each sentence. The probe achieves 76 and 65 percent accuracy for part-of-speech tagging and dependency labelling respectively, suggesting that BERT encodes a decent amount of syntactic information after pre-training. However, the BERT inputs

only yield slight gains in probing accuracy as compared to baseline inputs (one-hot and fastText encodings (Bojanowski et al., 2017)), which unlike BERT do not capture contextual information and are therefore unable to know syntax.

Miaschi et al. (2020) probe 68 sentence-level linguistic features in pre-trained BERT representations. They use Support Vector Machines (SVMs) as probing models for each linguistic feature individually. The probes are best at predicting values for syntactic and morpho-syntactic features of English sentences, as represented by the [CLS] token of BERT. Although many of these features are related to sentence length (e.g. parse tree depth), the BERT representations always outperform a correlation baseline between sentence length and a given linguistic feature. This suggests that BERT's syntactic knowledge goes beyond surface-level information.

Hall Maudslay and Cotterell (2021) are sceptical about the syntactic knowledge encoded in several English pre-trained language models (BERT, RoBERTa and GPT-2 (Radford et al., 2019)). To test if the model knows about grammaticality as a distinct property, they construct a corpus of "Jabberwocky" sentences, i.e. sentences that are syntactically well-formed, but semantically nonsensical. They extract pre-trained representations both for the Jabberwocky sentences and normal sentences and feed them into several probes that predict the dependency relations in each sentence. They find that all probes perform worse on the Jabberwocky sentences as compared to the normal ones. In addition, they confirm the results found by Pimentel et al. (2020) and show that the pre-trained representations only yield slight improvements over uncontextualized baselines.

While the above studies all give some indication about the syntactic knowledge that is encoded in pre-trained language models, it clearly remains a challenge to estimate the exact amount of linguistic structure that is encapsulated in their representations. Therefore, our aim is not to measure the exact amount of linguistic knowledge that is acquired through learning eye movement patterns of reading. Rather, we investigate if there is an **improvement** in the encoding of certain linguistic properties as a result of fine-tuning on eye-tracking data, as compared to the pre-trained model.

Fine-tuning on eye-tracking data To the best of our knowledge, there is currently only one study that uses probing to investigate the linguistic knowledge that is acquired after fine-tuning on eye-tracking data of reading. Sarti et al. (2021) fine-tune ALBERT (a lightweight alternative to BERT (Lan et al., 2020)) on four sentence-level eye-tracking metrics: *first-pass duration*, *total fixation duration*, *fixation count* and *regression duration*. The authors extract these metrics from the Ghent Eye-Tracking Corpus (Cop et al., 2016), which contains eye-tracking data of English participants reading an entire novel. They then use the fine-tuned model to represent sentences from three English treebanks, and feed them into n probes to predict a value for one linguistic feature each. They repeat this with the pre-trained model. Since many of the probed linguistic features correlate with sentence length, the authors also probe ALBERT representations that are only fine-tuned on sentences containing 10 tokens. Interestingly, their results indicate that the fine-tuned representations, both with and without length-binning, have a better encoding of syntactic features (e.g. parse tree depth, number of prepositional chains) as compared to the pre-trained representations, while lexical and morpho-syntactic features (e.g. type-token ratio, lexical density) remain unaffected by the fine-tuning process. This suggests that sentence-level eye-tracking features contain information about syntactic complexity that can be picked

up by transformer-based language models. The current study investigates whether this result extends to other languages.

Chapter 3

Data Analysis

Reading materials used for eye-tracking corpora vary in terms of linguistic complexity. Together with individual differences between readers, this leads to divergent distributions of eye-tracking metrics across eye-tracking corpora. A language model predicting eye-tracking metrics, then, should know about the correlations between linguistic complexity and eye movement patterns of reading to be able to generalize to different distributions of eye-tracking metrics across corpora.

This chapter assesses which particular features of linguistic complexity account for the varying distributions of eye-tracking metrics in two corpora: The Ghent Eye-tracking Corpus (Cop et al., 2016) and the Multilingual Eye-tracking Corpus (Siegelman et al., 2022). After describing how the eye-tracking data were collected for each corpus, we examine the linguistic complexity of the reading materials and analyse how the distribution of several eye-tracking metrics relates to it. This will show which linguistic knowledge should be acquired by a language model in order to accurately predict eye-tracking values for sentences of varying linguistic complexity.

3.1 Ghent Eye-tracking Corpus

The Ghent Eye-tracking Corpus (henceforth GECO) contains eye movement recordings of Dutch bilinguals and English monolinguals reading an entire novel (*The Mysterious Affair at Styles* by Agatha Christie, or the Dutch translated version *De zaak Styles*). The English group read the entire novel in their native language, while the Dutch group read half of the novel in their first language and the other half in their second language (English). Thus, GECO allows for comparing monolingual reading behaviour in two different languages. It also allows for comparing monolingual and bilingual reading behaviour in a single language. In addition, it shows how readers process language while reading a longer narrative in a non-restricted, naturalistic setting.

For the current project, we focus on the monolingual parts of GECO. It contains reading data of fourteen English monolinguals and nineteen Dutch(L1)-English(L2) bilinguals reading in their first language. Each participant read the entire novel in four sessions of maximally 1.5 hours. After each session, participants had to answer a set of comprehension questions, ensuring that they understood the reading material. The novel was presented on a computer screen, in paragraphs of maximally 145 words at a time. Participants could read the passages at their own speed and proceed to the next passage by pressing a button. Their eye movements were recorded by a high quality eye-tracker. Further information about the size characteristics of the reading materials

are presented in Table 1.

Since the novel contains many dialogue-style sentences such as "Certainly, Aunt Emily" and exclamations such as "Oh, this fellow!", we remove sentences that are shorter than 5 words. This ensures that the training data contains a balanced amount of shorter and longer sentences, and that a model trained on this data sees an adequate amount of complex syntactic structures. This filtering resulted in the exclusion of 1259 English sentences and 1003 Dutch sentences.

Dataset	Language	#Words	#Sentences	Avg. sent. length	Avg. word length
GECO	English	52131	4041	12.90	4.60
	Dutch	56654	4187	13.53	4.74
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
	Norwegian	2106	116	18.16	5.62
	Greek	2082	99	21.03	5.67
	Spanish	2412	98	24.61	5.01
	Russian	1827	101	18.09	6.53
	Hebrew	1943	121	16.06	4.89
	Korean	1699	101	16.82	3.21
	Turkish	1696	104	16.31	6.92

Table 3.1: Size characteristics for the reading materials of GECO and MECO. GECO sentences which are shorter than five words are removed.

3.2 Multilingual Eye-tracking Corpus

While GECO allows for direct comparisons between English and Dutch reading behaviour, it does not allow for eye-tracking research across typologically diverse languages. The Multilingual Eye-tracking Corpus (henceforth MECO) fills this gap by providing eye movement data of reading in 13 languages, covering a wide range of typologies and language families.

The reading material consists of 12 short Wikipedia-style texts about various topics, which participants read in their native language. All texts were selected in English first, after which matching texts were collected for the other languages. Five of the 12 texts were directly translated from English to the other languages. The remaining 7 texts were carefully matched for topic, genre and complexity, but were not direct translations. Back-translations to English showed that the translated texts were more semantically similar to the English originals (mean cosine similarity = 0.88) than the matched texts (mean cosine similarity = 0.66). By including both direct translations and more loosely matched texts, MECO allows for analysing cross-lingual eye movement patterns across different levels of semantic control. Detailed information about the number of words and sentences in each individual text per language can be found in Siegelman et al. (2022), but averages are presented in Table 1.

The data was recorded in many different eye-tracking labs across the world. Therefore, the experimental setup was slightly different for each language (e.g. screen and

font size, number of lines used to present a text). Nonetheless, each of the 12 texts was always presented on a separate screen and in the same fixed order in all languages. The number of participants ranged from 29 to 54 per language (45 on average). They could read the texts at their own speed and proceed to the next text by pressing a button. After each text, participants had to answer four comprehension questions. Eye movements were recorded by a high quality eye-tracker.

3.3 Selected eye-tracking metrics

As discussed in Chapter 2, previous research suggests that universality in cross-lingual eye movement patterns is more likely to be observed at the sentence level than at the word level. Therefore, we hypothesize that cross-lingual transfer will work better for sentence-level eye-tracking metrics than for word-level metrics. Since the GECO annotations for all eye-tracking metrics are provided at the word level, we sum-aggregate them at the sentence level. MECO already provides annotations at the sentence level.

Following Sarti et al. (2021), we select four eye-tracking metrics that cover both early and late language processing. Given a sentence s, we measure:

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

Regression duration is not directly provided in the datasets, but is calculated by subtracting *selective go-past time* from *go-past time*. To obtain generalized eye movement patterns, we average all eye-tracking metrics over participants.

3.4 Distribution of eye-tracking metrics

When comparing the distribution of our selected eye-tracking metrics across datasets, we observe that MECO values are generally higher than GECO values. For example, Figure 3.1 shows that the majority of values for total fixation duration is higher in MECO as compared to GECO. We also observe that there is almost no difference in the distribution of total fixation duration in the Dutch and English parts of GECO, while MECO does exhibit some variation across languages. We attribute this to the fact that the reading materials of GECO are direct translations of each other, while MECO also contains more loosely matched texts. Thus, since the reading materials per language differ more in semantic content, the eye movement patterns in MECO may deviate more across languages. Another factor that may play a role is the average sentence length of each language. For example, Spanish has longer sentences than the other languages in MECO (see Table 3.1), which might explain why Spanish has higher fixation durations on average. The English and Dutch sentences in GECO on the other hand, are very similar in length. While the current study focuses on universals of cross-lingual reading behaviour, further research is needed to examine the factors that trigger differences in reading behaviour across languages.

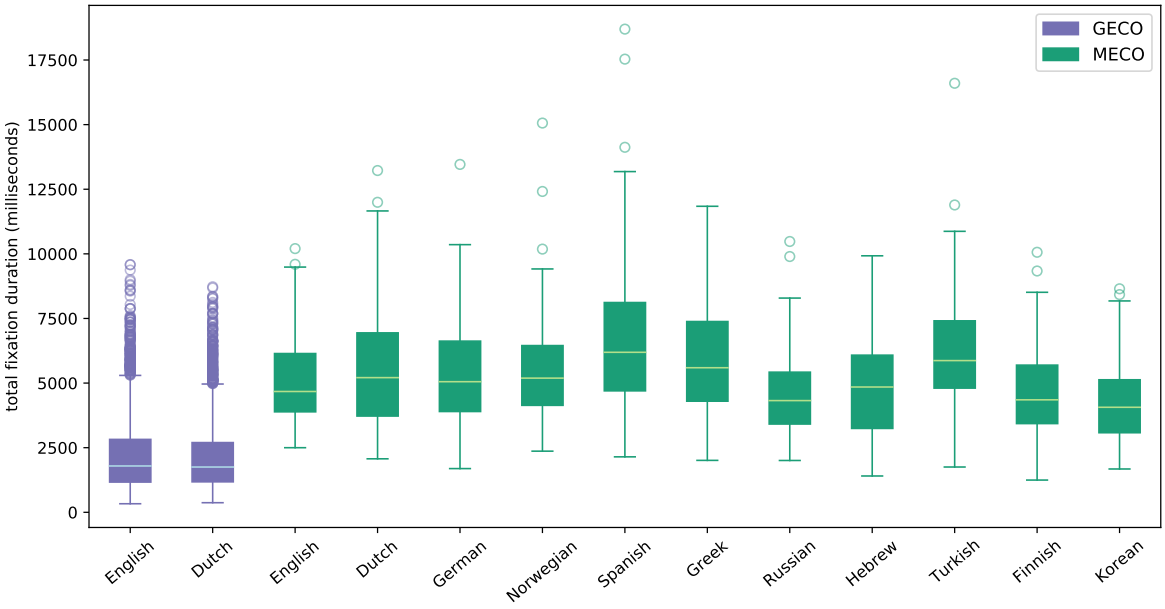

Figure 3.1: Comparison between GECO and MECO: Distribution of total fixation duration across languages, averaged over participants.

To make sense of the different distributions of the eye-tracking metrics in GECO and MECO, we compare the linguistic complexity of the reading materials of the two datasets. We limit this analysis to English since the reading materials are closely matched for linguistic complexity across languages in MECO. We consider four categories of sentence-level complexity features: length, frequency, morpho-syntactic, and syntactic.[1] The morpho-syntactic and syntactic features are computed using the Profiling-UD tool (Brunato et al., 2020), which allows for the extraction of more than a 100 linguistic features and can be applied to many different languages. Selected features are described below and summarized in Table 3.2.

Length features For each sentence, two length features are calculated: sentence length and average word length. Sentence length is measured in terms of tokens, and word length is measured in terms of characters. We do not count punctuation as tokens, since eye-tracking participants do not fixate punctuation separately from the attached word. For this same reason, we count attached punctuation as extra characters when calculating word length. This experimental choice is also employed in the study by Sarti et al. (2021). We already know from Table 3.1 that MECO words and sentences are longer on average than GECO words and sentences, but we also include these features here to analyse the overall distribution.

Frequency features We consider two sentence-level frequency features: average word frequency and number of low frequency words. Frequencies are obtained using the Python package wordfreq (Speer et al., 2018). The package is built on several

[1]It would also be interesting to consider cognitive features affecting linguistic complexity, such as age of acquisition. However, such features need to be extracted from psycholinguistic databases, which do not exist for all languages considered in this study. Therefore, we focus on linguistic features only.

frequency databases, including SUBTLEX lists (e.g. Brysbaert and New (2009)) and OpenSubtitles (Lison and Tiedemann, 2016).[2] We opt for *Zipf frequencies*, which have been standardized according to a logarithmic scale that ranges from 1 to 7. The advantage of Zipf frequencies is that they are not dependent on corpus size, like absolute frequency counts (van Heuven et al., 2014). Words with a Zipf frequency below 4 (i.e. the central value of the Zipf scale) are considered to be low-frequency words. Again, punctuation is excluded when calculating these features.

Morpho-syntactic features To quantify morpho-syntactic complexity, we calculate *lexical density*, which is defined as the ratio of content words (nouns, proper nouns, verbs, adjectives, adverbs) over the total number of words in a sentence. For example, the MECO sentence *A national flag is a flag which represents and symbolizes a country* has a lexical density of 0.5, since six out of twelve words are content words (i.e. *national*, two times *flag*, *represents*, *symbolizes*, *country*).

Syntactic features To quantify syntactic complexity, we obtain the dependency tree for each sentence and derive four features from it: parse tree depth, average dependency link length, maximum dependency link length, and number of verbal heads. Parse tree depth is defined as the longest path (in terms of dependency links) between the root of the dependency tree and some leaf. Dependency link length is defined as the number of tokens that occur linearly between a syntactic head and its dependent (excluding punctuation) – we calculate both the average length and the maximum length for each sentence. As a final measure of syntactic complexity, we count the number of verbal heads per sentence. Consider the following examples from MECO:

1. In ancient Roman religion and myth, Janus **is** the god of beginnings and gates.
2. He **has** a double nature and **is** usually depicted as **having** two faces, since he **looks** to the future and to the past.

Sentence 1 has one verbal head (*is*), while sentence 2 has four (*has*, *is*, *having*, *looks*). Sentence 2 therefore has a more complex syntactic structure than 1. Empirical evidence for this is reported in Brunato et al. (2018), who found that the number of verbal heads correlates with the perceived complexity of a sentence (even for sentences that have the exactly the same length).

Complexity distributions The distributions of the linguistic complexity features of the English GECO and MECO sentences are presented Figure 3.2. We see that MECO sentences are generally more complex with respect to all four categories of features. Compared to GECO sentences, they tend to have more and longer words, deeper parse trees, and longer dependency links. Their lexical density is also higher, and they contain more low-frequency words.

[2] A complete overview of all frequency databases included in wordfreq is given on the following page: `https://pypi.org/project/wordfreq/`

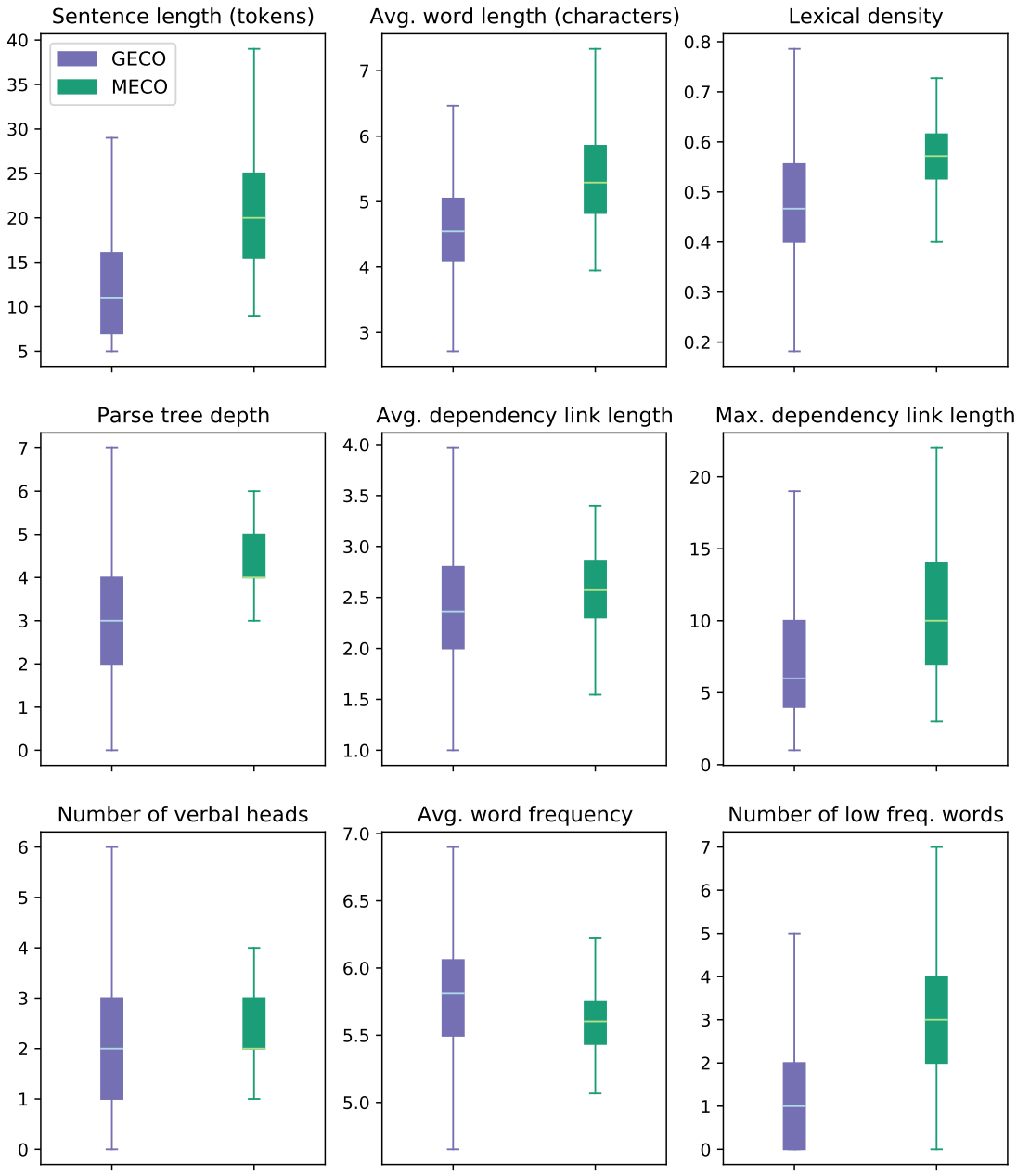

Figure 3.2: Distribution of linguistic complexity features in English sentences of GECO and MECO.

Category	Linguistic Feature
Length	Sentence length (tokens)
	Average word length (characters)
Frequency	Average word frequency
	Number of low frequency words (below Zipf=4)
Morpho-syntactic	Lexical density
Syntactic	Parse tree depth
	Average dependency link length
	Maximum dependency link length
	Number of verbal heads

Table 3.2: Sentence-level features capturing linguistic complexity.

We attribute the longer fixation durations measured for the MECO sentences to the fact that they are generally more complex than the GECO sentences. Nonetheless, we also note that there is much overlap between the complexity features of the two datasets. For example, lexical density ranges between 0.2 and 0.8 for GECO, but only between 0.4 and 0.7 for MECO. Thus, when a model is trained on GECO sentences, it will have seen example sentences of the same linguistic complexity as the MECO sentences. This should make it easier for the model to accurately predict eye-tracking values for MECO.

3.5 Predictors of eye-tracking metrics

As a final analysis of the relationship between linguistic complexity and eye movement patterns, we assess which linguistic features are the best predictors for our selected eye-tracking metrics. With this information, we can hypothesize which features will most likely be picked up by a neural model when it is trained on eye-tracking metrics. To this end, we calculate the Spearman[3] correlation between the nine linguistic complexity features presented in Table 3.2 and the four eye-tracking metrics described in Section 3.3. The results for the English GECO sentences (i.e. the training data) are shown in Figure 3.3.

We see that sentence length is the best predictor for all four eye-tracking metrics, followed by length-related syntactic features, i.e. maximum dependency link length, parse tree depth, and number of verbal heads. In contrast, features in which sentence length is factored out (average word length, lexical density and average word frequency) have a much weaker (yet significant) correlation with the eye-tracking metrics. It is therefore likely that a neural model will mostly rely on length-related information when learning to predict eye-tracking metrics.

With regard to the eye-tracking metrics individually, we see that first-pass duration, total fixation duration and fixation count are uniformly correlated with the nine linguistic features. Regression duration, on the other hand, exhibits less strong correlations with the linguistic features. Thus, not all variance in regression duration can be explained by the nine linguistic features presented here. Therefore, we hypothesize that regression duration will be harder to predict for a neural model than the other

[3]The Spearman correlation measures the monotonic relationship between two variables. We select Spearman rather than Pearson since these relationships might not be linear.

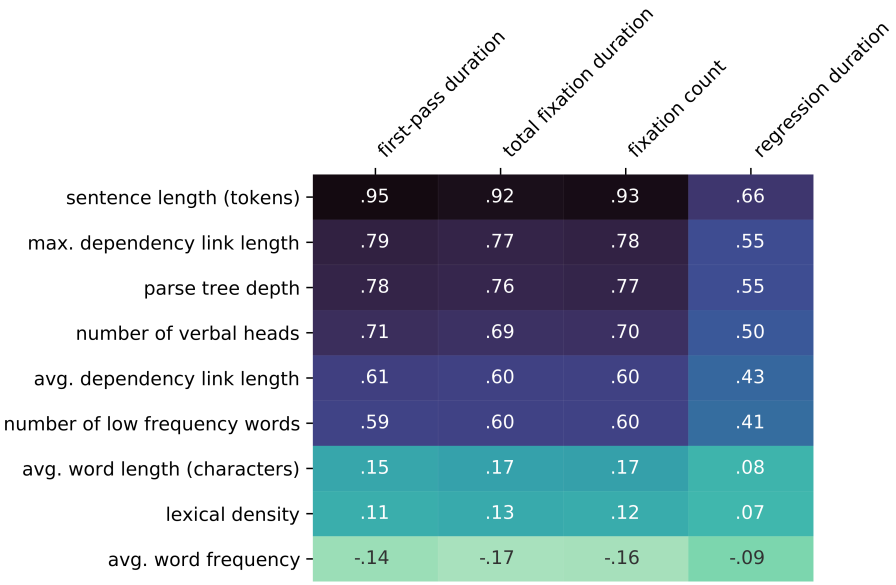

Figure 3.3: Spearman correlations between linguistic complexity features and eye-tracking metrics of English GECO sentences. All correlation coefficients have $p < 0.001$.

eye-tracking metrics, especially if the model learns to rely on features that are good predictors for the other three eye-tracking metrics.

We now examine the correlations between the linguistic complexity features and eye-tracking metrics for the English MECO sentences (i.e. the evaluation data). As shown in Figure 3.4, the correlations are generally weaker than those measured for the GECO data. This might be explained by the amount of sentences per dataset: GECO contains 4041 English sentences, while MECO contains only 99 English sentences. In addition, the GECO sentences were read by 14 participants, while the MECO sentences were read by 46 participants. Therefore, the MECO data contains more noise caused by individual differences between readers. The smaller amount of sentences and the larger amount of readers make it harder to observe strong correlations between linguistic complexity and eye movement patterns in MECO.

Another possible explanation for the different correlations measured for GECO and MECO is that some eye-tracking metrics might be domain-sensitive. Literary texts contain very different words than encyclopedic texts, which might influence fixation durations and trigger regressions that cannot solely be explained by linguistic complexity. Thus, a model that learns to rely on linguistic complexity might not be able to predict eye-tracking metrics across domains. We speculate that this will be the case for regression duration and first-pass duration especially, which exhibit the weakest correlations with the complexity features.

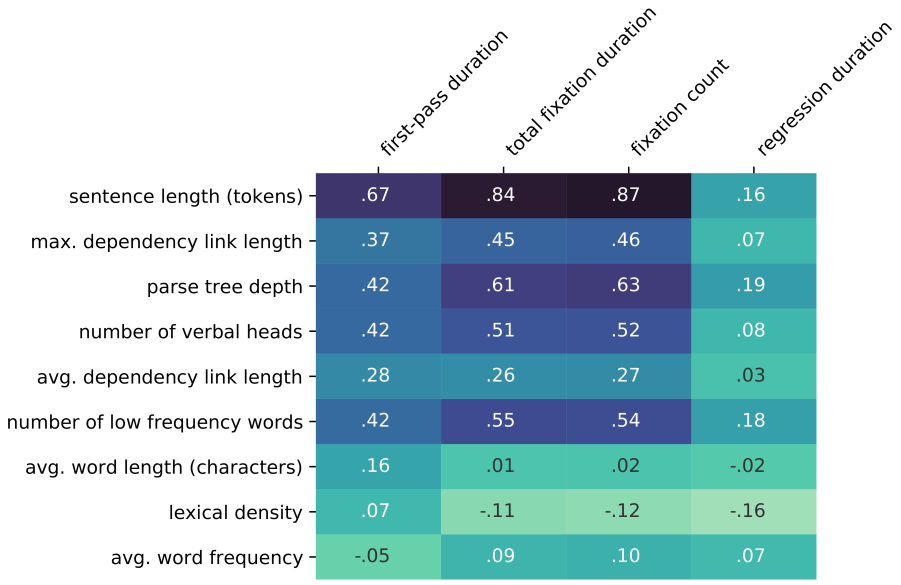

Figure 3.4: Spearman correlations between linguistic complexity features and eye-tracking metrics of English MECO sentences. All correlation coefficients have $p < 0.001$.

Chapter 4

Eye Movement Prediction

Previous research has shown that multilingual transformer models are capable of predicting eye movement patterns across domains and languages (Hollenstein et al., 2021b). This leads to the hypothesis that such models are cognitively plausible and rely on linguistic complexity to make their predictions. In this chapter, we examine if these results extend to another multilingual transformer model, other eye-tracking datasets, and other languages. We then try to form hypotheses about the linguistic characteristics that the model implicitly relies on, and examine if this is comparable to humans.

4.1 Experiments

Firstly, we examine eye movement prediction **across domains**, where the test data (English MECO) is linguistically more complex than the training data (English GECO) and is thus paired with larger eye-tracking values. If a multilingual transformer model is capable of linking linguistic complexity to larger eye-tracking values, it should be able to accurately predict eye movements associated with sentences that are more complex than the majority of training sentences. By evaluating the model on the same language it was trained on, we specifically target its cross-domain abilities. As a control, we also report the model's prediction accuracy on in-domain data.[1]

Secondly, we explore eye movement prediction **across languages**. Here, we again use the model trained on English GECO data, but evaluate it on the MECO data in other languages. This way, we can test if the model learned that certain linguistic features trigger universal eye movement patterns across languages. Previous research suggests that cross-lingual transfer works best for typologically similar languages (Pires et al., 2019). However, Figure 3.1 shows that the distribution of eye-tracking metrics in MECO is rather consistent across languages, even across typologically distant ones. If the transformer model picks up this universality of eye movement patterns, there should not be large differences in prediction accuracy depending on typology.

We also examine how well the MECO eye movement patterns can be predicted from explicit features. This will shed light on the features that are most informative for predicting eye movement behaviour, and which might be implicitly utilized by the transformer. To this end, we train several feature-based regression models on vectors of the sentence-level features presented in Table 3.2. We train four of these models for each eye-tracking metric respectively, using different subsets of features each time:

[1] We train all models described in this thesis on 90 percent of the GECO data. We use the additional 10 percent to evaluate XLM-RoBERTa's predictive qualities for in-domain data.

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. We expect the model trained on length features to perform especially well, since sentence length has the strongest correlation with all eye-tracking metrics.

4.1.1 Models

Multilingual Transformer We select XLM-RoBERTa (Conneau et al., 2020) as our multilingual language model. The model is based on the Transformer architecture (Vaswani et al., 2017) and was pretrained on 2.5TB CommonCrawl data containing 100 languages using the Masked Language Modelling objective. This means that, given a piece of text, the model randomly masks 15% of the tokens in the input, after which the entire piece of text (including masked tokens) is ran through the model.[2] The model then predicts which words belong in the masked positions. Since the model sees both the tokens preceding and following the masked token, it learns a *bidirectional* representation of the text. In other words, it learns to take context into account. At each iteration, the model is trained on batches of 64 texts sampled from one of the 100 languages. The texts are not parallel across languages.

The model splits the raw texts into tokens using subword tokenization (SentencePiece, Kudo and Richardson (2018)). It uses a shared vocabulary for all languages, consisting of 250k subwords. This means that a subword such as *normal* can be used both for the English word *normally* and for the German word *normalerweise*.

Feature-based Regressors We use Support Vector Machines (SVMs) as our feature-based regression models. We employ the SVR implementation from scikit-learn (Pedregosa et al., 2011) with all default parameters and a linear kernel.

4.1.2 Fine-tuning procedure

We select the Huggingface checkpoint *xlm-roberta-base* as our pretrained model and add a linear dense layer on top of it to predict four sentence-level eye-tracking metrics. We employ multi-task learning with hard parameter sharing to fine-tune the model on all eye-tracking metrics simultaneously.[3] This means that all model parameters are shared except for the task-specific regression heads in the final prediction layer. More specifically, one and the same sentence representation (encoded by XLM-RoBERTa) is fed into each of the four regression heads, which then predict their respective eye-tracking metric. Finally, a joint loss is computed by summing the individual *mean squared error* (MSE) loss scores of the four regression heads (i.e., the mean of the squared differences between the predicted values and the actual values). This is then used to optimize the model parameters jointly for all regression tasks. We 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., 2021b). The procedure is illustrated in Figure 4.1.

[2]XLM-RoBERTa received continuous streams of 256 tokens as input during pretraining, instead of sentence pairs as in the original training objective for BERT (Devlin et al., 2019).

[3]The code for multi-task learning was adapted from the FARM framework (Deepset, 2019) by Sarti et al. (2021) and can be found in the following repository: `https://github.com/gsarti/interpreting-complexity`

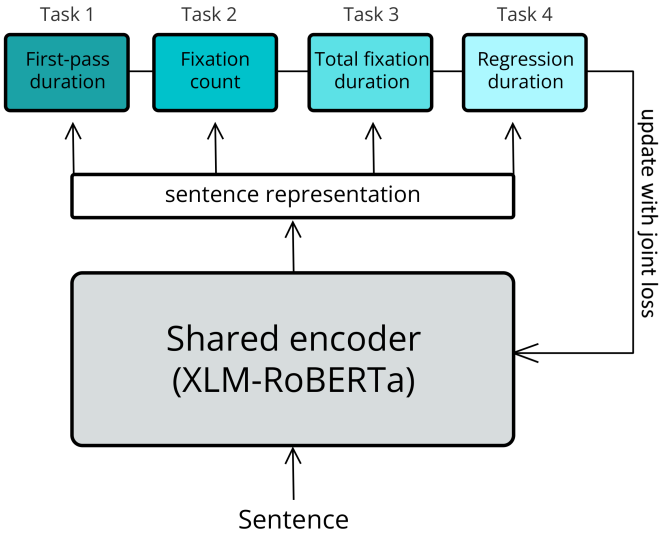

Figure 4.1: Multi-task learning with hard parameter sharing for prediction of four sentence-level eye-tracking metrics.

We train the model for 15 epochs with early stopping after 5 epochs without an improvement on 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. We set the maximum sequence length to 128 – shorter sentences are padded and longer sentences are truncated (although truncation never happened in our particular dataset). To obtain sentence representations from the model, we take the average of all token embeddings (i.e. mean pooling) as an alternative to the [CLS] token. This choice was motivated by the results reported in Mosbach et al. (2020), which indicate that RoBERTa-models encode more sentence-level information in the average token embedding as compared to the [CLS] token.

4.1.3 Evaluation

The *Mean Absolute Error* (MAE) measures the mean of the absolute differences between the predicted values x_i and actual values y_i, as shown in the following formula:

$$\frac{1}{n}\sum_{i=1}^{n}|x_i - y_i| \tag{4.1}$$

Since all eye-tracking metrics are scaled to the range 0-100, we can interpret the MAE as a percentage error (Hollenstein et al., 2021b). This allows us to derive prediction accuracy as 100–MAE. We use this metric to evaluate the accuracy for the four eye-tracking metrics individually. In addition, we calculate a mean baseline for each eye-tracking metric and report the accuracy improvement of the trained models relative to this baseline.

4.2 Results

This section discusses how well the fine-tuned XLM-RoBERTa model could predict the different eye-tracking metrics. We analyse the model's cross-domain and cross-lingual abilities, and try to form hypotheses about the linguistic features that the model implicitly relies on.

4.2.1 Cross-domain abilities

Table 4.1 shows how much XLM-RoBERTa improves over a mean baseline for each eye-tracking metric, both for the cross-domain data (English MECO) and the in-domain data (English GECO).[4] We see that the model yields consistent results across domains for two out of four eye-tracking features, i.e. fixation count and total fixation duration. For first-pass duration and regression duration, on the other hand, the in-domain results are more accurate than the cross-domain results.

	Cross-domain (MECO)	In-domain (GECO)
First-pass duration	5.52	10.18
Fixation count	8.68	8.94
Total fixation duration	9.78	9.45
Regression duration	1.48	4.04

Table 4.1: Improvement on prediction accuracy of XLM-RoBERTa relative to a mean baseline for cross-domain and in-domain English evaluation data.

Our results nicely fit the correlational analysis presented in Section 3.5. Firstly, we found that fixation count and total fixation duration strongly correlate with structural complexity features (i.e. length-related and syntactic features). These correlations are quite consistent across domains. Secondly, we found that first-pass duration and regression duration are domain-sensitive: these eye-tracking metrics strongly correlate with the structural complexity features in the in-domain data, but not in the cross-domain data. XLM-RoBERTa's behaviour is exactly in line with these two findings: it can accurately predict fixation count and total fixation duration for both domains, but it is much better at predicting first-pass duration and regression duration for the in-domain data as compared to the cross-domain data. Therefore, it is likely that the model learned to rely on the structural complexity of sentences for the prediction of all eye-tracking metrics.

4.2.2 Cross-lingual abilities

The cross-domain results reported in the previous section indicate that XLM-RoBERTa learned a correlation between the complexity of English text and eye movement behaviour of English readers. As discussed in Section 2.1.3, similar correlations have been observed in reading behaviour of other languages. Therefore, we tested whether XLM-RoBERTa would be able to apply the learned correlations to other languages. Figure 4.2 shows that this is the case: XLM-RoBERTa predicts total fixation duration

[4]The absolute accuracies can be found in the Appendix, i.e. Table A.1.

with 90 to 95 percent accuracy for all languages from MECO.[5]

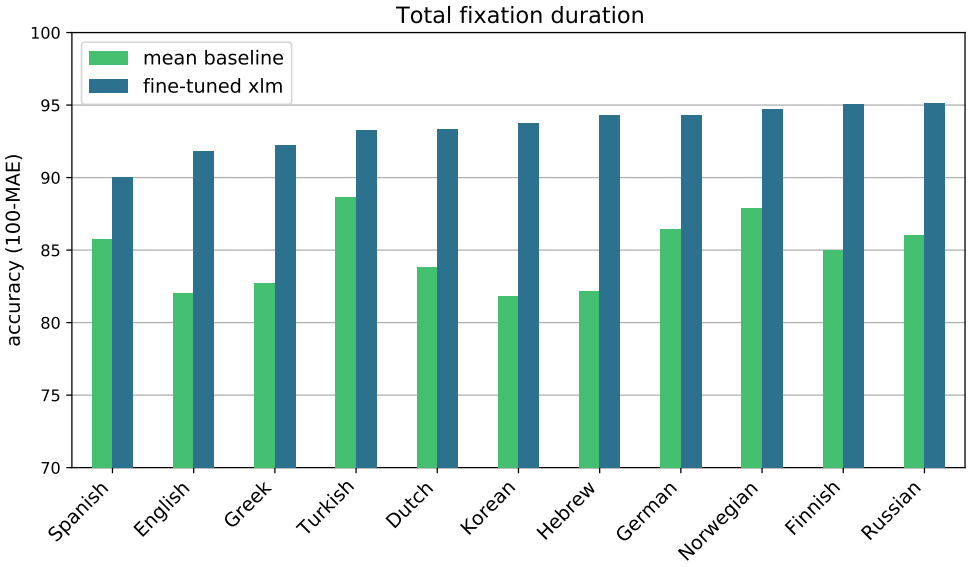

Figure 4.2: Prediction accuracy of fine-tuned XLM-RoBERTa and the mean baseline for total fixation duration for each language in MECO.

Interestingly, we see that XLM-RoBERTa performs slightly better for unseen languages than for English (the training language), even when they are typologically different or have different scripts. For example, the model reaches the best performance for Russian, which makes use of the Cyrillic script. This is in contrast to the results reported in Hollenstein et al. (2021b), who found that XLM models transfer better within scripts than across scripts. This could be a result of the fact that our model predicts sentence-level eye-tracking metrics, which are more consistent across languages than word-level metrics.

Another explanation is that Hollenstein et al. (2021a) did not use eye-tracking data associated with parallel texts for their cross-lingual evaluation. In fact, the reading materials of the corpora they used were quite different in terms of semantic content. The English corpus consisted of a combination of literary text, newspaper articles, movie reviews and Wikipedia articles; the Dutch corpus consisted of literary text; the German corpus consisted of college-level biology and physics textbooks; and the Russian corpus consisted of naturally occurring sentences from various sources. As discussed in Section 4.2.1, some eye-tracking metrics are domain-sensitive. Such eye-tracking metrics are difficult to predict across domains, regardless of the language. Thus, the cross-lingual results reported in Hollenstein et al. (2021a) might have been influenced by domain differences. It would be interesting to examine if the multilingual models tested in their study achieve better cross-lingual results on the parallel texts of MECO.

We now take a closer look at the predictions that were generated for Russian and Spanish, i.e., the languages for which XLM-RoBERTa was the most and the least accurate, respectively. Predictions versus true values for total fixation duration of these two languages are plotted in Figure 4.3. When comparing the mean of the true and

[5]The results for the other three eye-tracking metrics can be found in the Appendix, i.e. Figure A.1. The cross-lingual similarity in prediction accuracy that is shown for total fixation duration can be observed for the other eye-tracking metrics as well.

predicted values for each language, we see a closer correspondence for Russian than for Spanish. However, we also see that the predictions for Russian always stay very close to the mean. This shows that XLM-RoBERTa was not able to learn the entire range of values for total fixation duration, and that the prediction accuracy of 95 percent measured for Russian does not show the full picture of the model's performance.

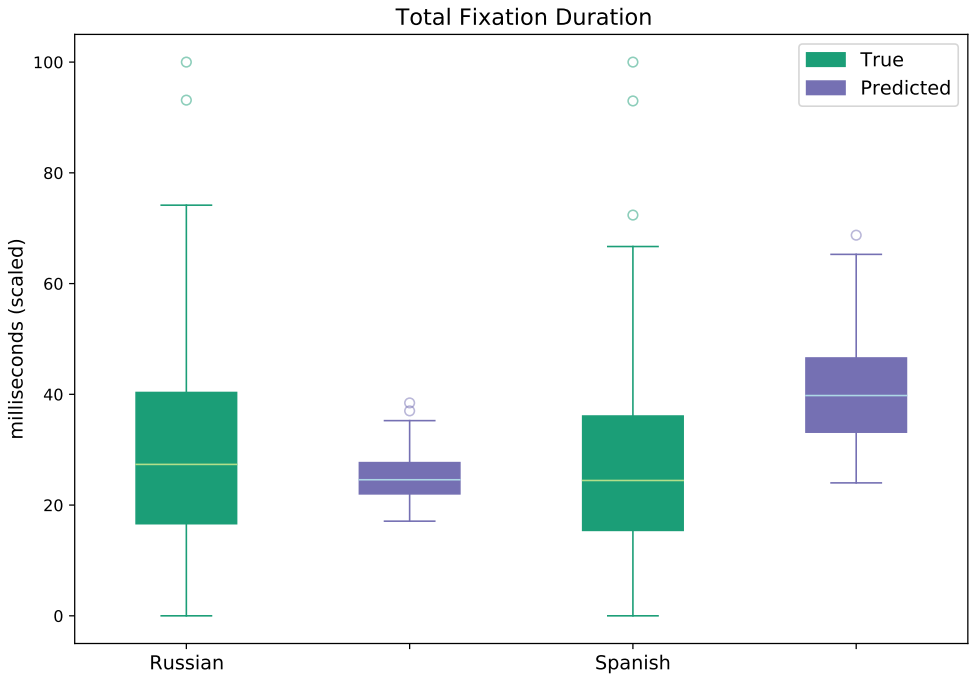

Figure 4.3: True versus predicted values for total fixation duration of the Spanish and Russian parts of MECO. Predictions are made by fine-tuned XLM-RoBERTa.

4.2.3 Implicit usage of complexity features

So far, our results indicate that XLM-RoBERTa learned a correlation between linguistic complexity and eye movement behaviour. However, the exact complexity features that the model implicitly relies on remain unclear. To form hypotheses about this, we examine how well the eye-tracking metrics can be predicted from different groups of explicit features.

Figure 4.4 shows how several feature-based SVM models perform on the English MECO data, as compared to the fine-tuned XLM-RoBERTa model. We find that all feature-based models improve over the mean baseline for all eye-tracking metrics, except for regression duration.[6] This shows that regression duration is hard to predict based on our selected complexity features (which is in line the finding that regression duration is only weakly correlated with the complexity features, see Figure 3.4). Nonetheless, XLM-RoBERTa *does* improve over the mean baseline for regression duration. Therefore, we speculate that it predicts regression duration based on a different combination of features than the combinations that were used to train the SVMs.

[6]The absolute accuracies can be found in the Appendix, i.e. Table A.2.

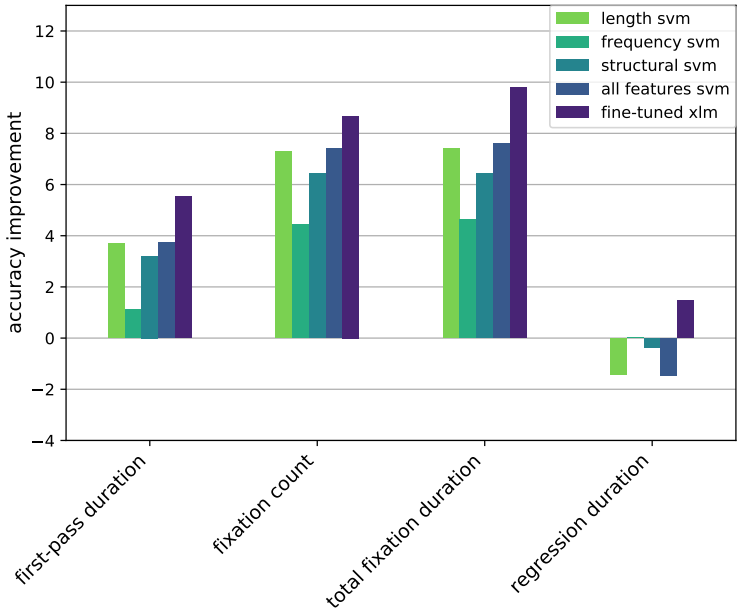

Figure 4.4: Improvement of prediction accuracy of the five different models relative to the mean baseline for each eye-tracking metric. The models are evaluated on the English part of MECO.

Another observation is that the length-based SVM performs almost identically to the SVM trained on *all* features from Table 3.2. Both of these models outperform the SVMs trained on frequency features and structural features. This highlights the fact that length is the best predictor for eye-tracking metrics, and suggests that structural and frequency features do not provide much additional information to the SVM models. We therefore hypothesize that XLM-RoBERTa also heavily relies on length to make its predictions.

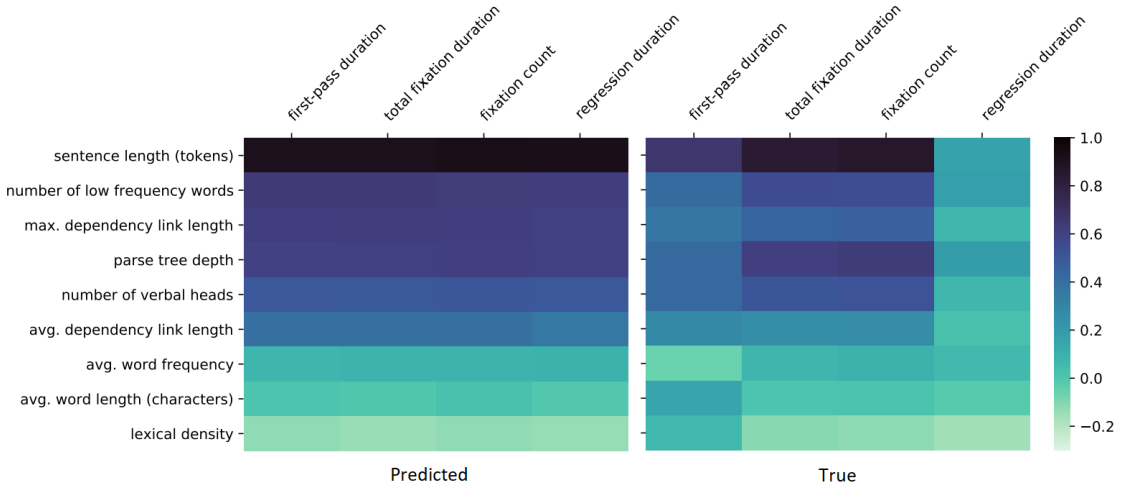

Figure 4.5: Correlations between complexity features and predicted versus true eye-tracking metrics for the English part of MECO. A darker color represents a stronger correlation.

To test this hypothesis, we calculate the Spearman correlation between XLM-RoBERTa's predictions and each complexity feature. Figure 4.5 shows the results. As hypothesized, we see that the predictions for all eye-tracking metrics strongly correlate with sentence length. The predictions also show moderate correlations with the syntactic features and the number of low frequency words, but this might be because those features are sensitive to sentence length as well.

Finally, we observe that the correlations between XLM-RoBERTa's predictions and the complexity features are nearly the same for all eye-tracking metrics. This highlights the effect of multi-task learning: since the loss is computed jointly over all tasks, accurate predictions for three out of four tasks already lead to a very small loss. Since first-pass duration, total fixation duration and fixation count can all be predicted from similar complexity features, the model learns to rely on those features and applies them to regression duration as well. This explains why the predictions for regression duration are inaccurate, because the true values of this metric are only weakly correlated with the complexity features. Further research is needed to better understand the linguistic features underlying regression duration.

4.2.4 Comparing length bins

Since XLM-RoBERTa seems to strongly rely on sentence length for the prediction of all eye-tracking metrics, we examine how it behaves on sentences of the same length. This will show if the model considers other features of linguistic complexity when length is not a differentiating factor anymore. Consider the following two sentences that both have 14 tokens (including attached punctuation):

1. *The most popular colours used for national flags are red, white, green, and blue.*
2. *During the late nineteenth century, the monocle was generally associated with wealthy, upper-class men.*

Intuitively, we can see that the first sentence is easier to read than the second one. This difference in reading ease is quantified in Table 4.2.

Category	**Linguistic Feature**	**Value of S1**	**Value of S2**
Length	Average word length (characters)	4.86	**6.43**
Frequency	Average word frequency	5.76	**5.20**
	Number of low frequency words	0	**3**
Morpho-syntactic	Lexical density	0.71	0.67
Syntactc	Parse tree depth	4	3
	Average dependency link length	2.54	**2.79**
	Maximum dependency link length	7	7
	Number of verbal heads	2	1

Table 4.2: Values of linguistic complexity features for S1: *The most popular colours used for national flags are red, white, green, and blue* and S2: *During the late nineteenth century, the monocle was generally associated with wealthy, upper-class men.*

We see that sentence 2 has longer words, less frequent words, and longer dependency links than sentence 1. XLM-RoBERTa reflects this difference in complexity by

predicting that readers will fixate longer on sentence 2 (31.17 ms) than on sentence 1 (24.47 ms), and that readers will spend more time regressing to previous material in sentence 2 (12.76 ms) than in sentence 1 (9.07 ms). This leads to the conclusion that XLM-RoBERTa's knowledge about linguistic complexity goes further than merely understanding that longer sentences lead to increased fixation durations.

To rule out the possibility that different predictions are caused by differences in semantic content, we compare predictions for translations of the same sentence across languages. We again select sentences that have the same number of words, so that the model cannot base its predictions on sentence length. Table 4.3 shows an example sentence in English, Finnish and Turkish, along with the true and predicted values for total fixation duration. We observe that the true eye-tracking values are consistent across languages. However, the model predictions deviate depending on the language. More specifically, XLM-RoBERTa predicts that Turkish readers will fixate longer on the sentence than Finnish readers. We hypothesize that the model generates these predictions based on subtle differences in linguistic complexity, since both the length and the semantic content of the sentence is the same.

Language	**Sentence**	**True**	**Predicted**
English	*In ancient Roman religion and myth, Janus is the god of beginnings and gates.*	38.53	29.52
Finnish	*Muinaisen roomalaisen mytologian mukaan Janus oli alkujen ja porttien jumala.*	38.67	20.96
Turkish	*Antik Roma inanı slarında ve mitlerinde, Janus ba slangı cların ve kapıların tanrısıdır.*	37.98	31.51

Table 4.3: True and predicted values for total fixation duration for the sentence *In ancient Roman religion and myth, Janus is the god of beginnings and gates* in three languages.

When comparing the Finnish and the Turkish sentence, we indeed see differences in terms of word length, word frequency and dependency link length. As we can see in Figure 4.6, the Turkish sentence has longer and less frequent words, but the Finnish sentence has longer dependency links. Nonetheless, XLM-RoBERTa predicts a much higher total fixation duration for Turkish than for Finnish. Therefore, it seems like the model relies on low-level complexity (i.e. word length and frequency) rather than syntactic complexity when it predicts eye-tracking metrics for sentences of the same length.

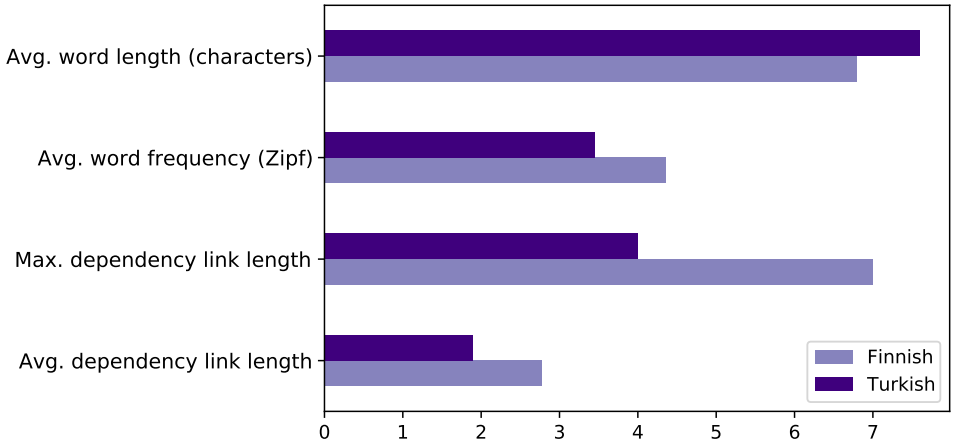

Figure 4.6: Linguistic complexity of the sentence *In ancient Roman religion and myth, Janus is the god of beginnings and gates* translated in Finnish and Turkish.

4.3 Summary of results

Our experiments indicate that XLM-RoBERTa was able to establish a link between linguistic complexity features and eye movement patterns of reading. We show that 1) the model can generalize across domains for eye-tracking metrics that strongly correlate with linguistic complexity; and that 2) the model can predict eye movement patterns for unseen languages, indicating that the model could abstract away from specific lexical items, and that the learned correlations are general enough to apply to all languages. Further analyses show that the model predictions strongly correlate with length-related and structural complexity features, but the model also seems to consider word length and frequency when predicting eye-tracking values for sentences of the same length.

Chapter 5

Probing Linguistic Knowledge

Based on the results of the previous chapter, we hypothesize that XLM-RoBERTa becomes sensitive to the structural complexity of a sentence when it learns to predict eye movement patterns of reading. To test this hypothesis, we **probe** the linguistic knowledge that is implicitly encoded in the model's final-layer representations, both before and after fine-tuning. This way, we get a better understanding of how a model's inner representations change as a result of fine-tuning on eye-tracking metrics. Finally, we analyse whether the linguistic knowledge that was acquired from English eye-tracking data transfers to other languages. This might explain why XLM-RoBERTa is able to accurately predict eye movement patterns for unseen languages.

5.1 Experiments

Our probing tasks consist of predicting a value for each of the nine linguistic complexity features presented in Table 3.2, which we will call $Z = z_1,...,z_9$. Let us denote the XLM-RoBERTa model as f and each probing regressor as g_i. Given an input sentence x, we obtain XLM-RoBERTa's final-layer representation $f_l(x)$ and feed it into each g_i, which then has to predict a value for its respective linguistic feature z_i. The prediction accuracy of g_i is an indication of how prominently the linguistic property z_i is encoded in $f_l(x)$. We analyse this both for the pre-trained and fine-tuned representations of XLM-RoBERTa. This allows us to measure the relative increase (or decrease) of the encoding of each z_i after fine-tuning on eye-tracking metrics.

There are two ways of extracting a sentence representation from a transformer model: 1) *mean pooling*, where we take the average of all token embeddings, and 2) *CLS-pooling*, where we take the embedding of the [CLS] token as a proxy for the entire sentence. Mosbach et al. (2020) find that mean pooling consistently produces better accuracy for three probing tasks as compared to CLS-pooling, possibly because the average of all token embeddings captures more sentence-level information than the [CLS] token. However, the authors also stress that probing accuracy is highly dependent on the specific probing task, model, and fine-tuning combination. To find out which representational variant works best for our particular combination, we first carry out the probing experiments with CLS-pooling, and repeat them with mean pooling. We then analyse whether the prediction accuracy of the probing regressors differs depending on the pooling method. We hypothesize that mean pooling will yield better results since we are predicting sentence-level linguistic features.

Since we fine-tune XLM-RoBERTa on English eye-tracking data, the model has

only seen examples of linguistic complexity in English during training. To test how well the model's knowledge about linguistic complexity transfers to other languages, the probing experiments are carried out [redacted] three typologically different languages: English, Korean and Turkish. Thus, we tr[redacted] 9 (linguistic features) x 3 (languages) x 2 (pre-trained versus fine-tuned) x 2 (mean pooling versus CLS-pooling) = 108 probing regressors in total.

5.1.1 Data

As input, we use the English, Korean and Turkish parts of the Parallel Universal Dependencies (PUD) treebanks, which were created for the CoNLL 2017 Shared Task on Multilingual Parsing (Zeman et al., 2017). For each language, there are 1000 parallel sentences, which were randomly selected from Wikipedia and news articles (usually only a few sentences per article). The first 750 sentences were originally English and translated to each of the other languages, and the last 250 sentences were originally German, French, Italian or Spanish, and were translated to the other languages via English. These parallel sentences are very suitable for our probing experiments, since they rule out the possibility that cross-lingual differences in probing accuracy are caused by differences in semantics. For each language, we use 800 sentences to train the probing regressors and the remaining 200 to test them.

5.1.2 Model

We use the same architecture as shown in Figure 4.1, but freeze the encoder model and only update the final regression layer during training. The final regression layer contains nine regression heads, one for each linguistic feature. We train these regression heads for 5 epochs without intermediate evaluation on a development set, and without early stopping. Other than that, the hyperparameters are the same as for the eye-tracking experiments (see Section 4.1.2).

5.1.3 Evaluation

For reliability, we perform 5-fold cross-validation and report the average result over all folds. We again use 100-MAE to measure probing accuracy (see Section 4.1.3). Since it is not straightforward to draw conclusions from probing accuracy in isolation, we only report *relative* probing accuracy, i.e. the gain (or loss) in prediction accuracy when the probing input is a fine-tuned representation versus some baseline – in our case, a pre-trained representation. This way, we can observe whether fine-tuning on eye-tracking metrics triggered a change in the encoding of linguistic features as compared to the original pre-trained model.

5.2 Results

In this section, we examine the probing accuracy for each linguistic complexity feature given different XLM-RoBERTa representations (i.e. pre-trained versus fine-tuned; CLS-pooling versus mean pooling) and different languages as input. We compare our results to those reported in Sarti et al. (2021), who probe the linguistic information encoded in the [CLS] token of a monolingual English transformer model (ALBERT, (Lan et al., 2020)), both before and after fine-tuning on the English part of GECO. This might reveal differences in the way that multilingual and monolingual transformer models pick up linguistic information from eye movement patterns of reading.

5.2.1 Probing accuracy across complexity features

Figure 5.1 shows the relative probing accuracy for each complexity feature. We see that length-related complexity features are easier to predict from a fine-tuned input than from a pre-trained input. We see that fine-tuning yields the largest improvements for sentence length, average dependency link length and number of low-frequency words, both using CLS-pooling and mean pooling. We also observe slight improvements for maximum dependency link length and parse tree depth, especially when using mean pooling. This confirms the findings reported in the previous sections: XLM-RoBERTa encodes information about the structural complexity of a sentence after learning to predict eye-tracking metrics.

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 complexity features for which sentence length is factored out, i.e. average word frequency, average word length and lexical density. This is somewhat contradictory to the results reported in Section 4.2.4, where the model seems to show a sensitivity to word length and word frequency. A possible explanation for this is that the sensitivity to word length and frequency was already present in the pre-trained model, and that it did not increase after fine-tuning on eye-tracking metrics.

Similar to our results, Sarti et al. (2021) find that features capturing structural sentence complexity (e.g. sentence length, dependency link length, parse tree depth, length of prepositional chains and subordinate clauses) are better encoded in ALBERT's [CLS] tokens after fine-tuning on English eye-tracking data. They show that these improvements remain present when only probing sentences of the same length. This provides strong evidence that the model picks up syntactic information from eye-tracking data, in addition to low-level length-related information. While we do not probe linguistic features for length-binned sentences in this study, the similarities of our results and those reported in Sarti et al. (2021) seem to indicate that monolingual and multilingual models pick up similar information from eye-tracking data recorded during reading.

5.2.2 CLS-pooling versus mean pooling

Figure 5.1 shows that mean pooling generally yields better probing accuracy than CLS-pooling. This difference is especially pronounced for features that are sensitive to sentence length, i.e. number of tokens, number of low frequency words, and maximum dependency link length. This indicates that the mean of all token embeddings captures more length-related information than the [CLS] token, which is in line with the

conclusions drawn by Mosbach et al. (2020).

Interestingly, average dependency link length can be predicted equally well from the [CLS] token and the average token embedding. The same is true for parse tree depth. This suggests that both representational variants capture a similar amount of syntactic information after fine-tuning on eye-tracking metrics.

5.2.3 Cross-lingual transfer of linguistic knowledge

Figure 5.1 shows that there are some minor differences in probing accuracy for individual complexity features of English, Korean and Turkish sentences. However, the general pattern is the same for all languages: features related to the structural complexity of sentence are more easily predicted after fine-tuning on eye-tracking metrics. This shows that XLM-RoBERTa is able to transfer linguistic knowledge acquired from English eye-tracking data to other languages.

As shown discussed in Section 4.2.2, XLM-RoBERTa predicts eye-tracking metrics with nearly the same accuracy across languages. In addition, our probing results show that 1) structural complexity features are better encoded after fine-tuning on eye-tracking metrics; and 2) this knowledge about structural complexity can be transferred to other languages. Taken together, these results indicate that the model learned a correlation between structural complexity and eye-tracking metrics during fine-tuning. Since this correlation is similar across languages, it can be learned from monolingual eye-tracking data and then be applied to other languages.

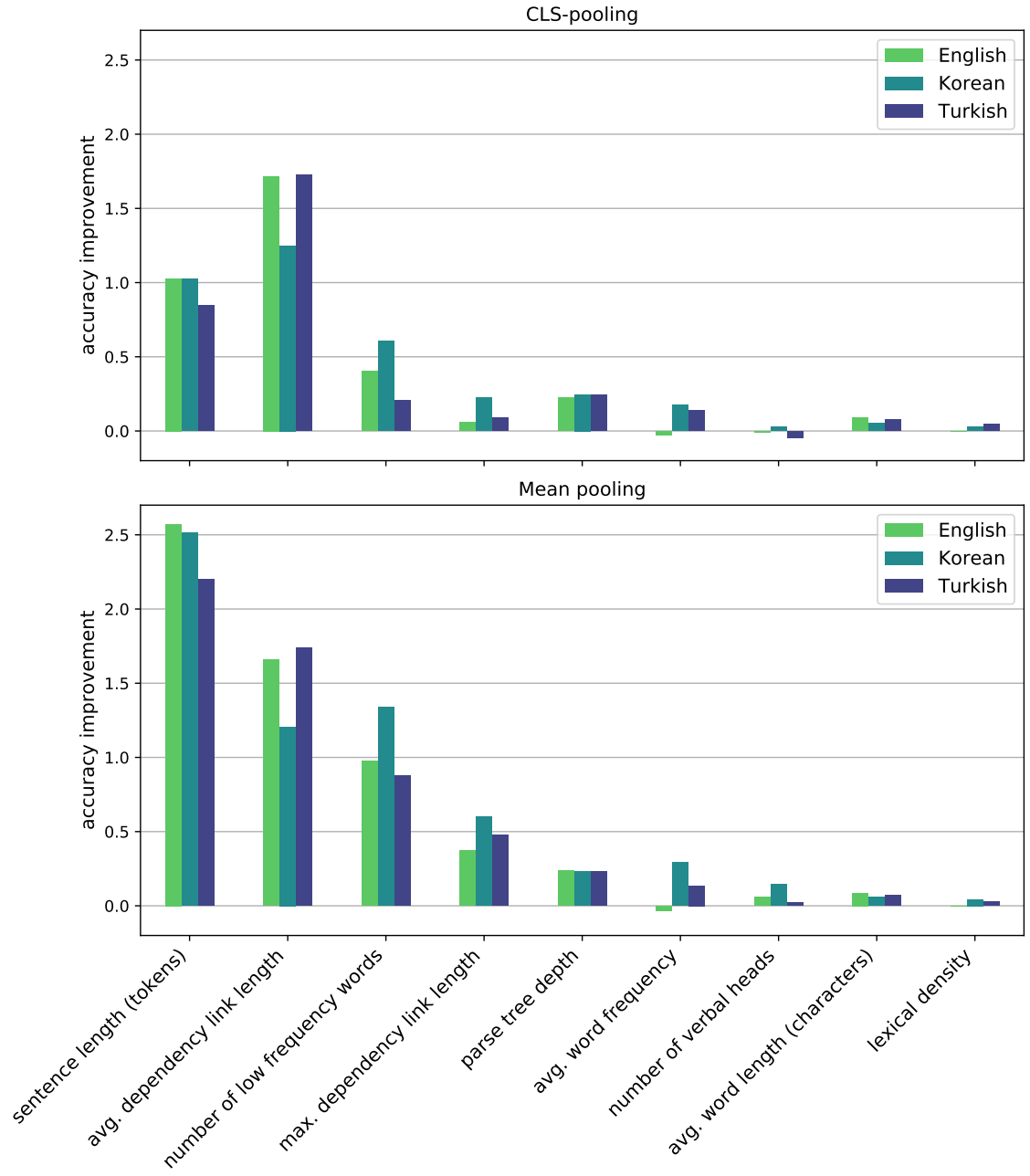

Figure 5.1: Improvements on probing accuracy for linguistic features of English, Korean and Turkish when using fine-tuned XLM-RoBERTa sentence representations as input to probing regressors, as compared to using pre-trained representations. The upper figure shows the results when the probing regressors receive the [CLS] token as input, and the bottom figure shows the results when they receive the average token embedding as input.

Chapter 6

Conclusion and Discussion

In this thesis, we aimed to answer the following research questions:

1. **Cross-domain abilities**: Can eye movement patterns be predicted for sentences that come from a different domain and are linguistically more complex than those seen during training?

2. **Cross-lingual abilities**: Can eye movement patterns be predicted for languages that are not seen during training?

3. **Sensitivity to linguistic complexity**: Can high prediction accuracy for eye movement patterns be explained by an increased sensitivity to linguistic complexity?

We investigated these three questions for one multilingual transformer model in particular: XLM-RoBERTa. Our results show that the answer to the first two questions is yes: 1) XLM-RoBERTa is capable of predicting the number of fixations and the total fixation duration for rather complex encyclopedic sentences, even though it was trained on easier sentences from the literary domain, and 2) XLM-RoBERTa is capable of predicting eye movement patterns for a range of typologically diverse languages, even though it was trained on eye-tracking data from English readers alone. These generalization abilities indicate that the model established a link between linguistic complexity and eye movement patterns, and that it could abstract away from specific words or languages.

To find evidence for this, we carried out several post-hoc analyses regarding the relationship between the model's predictions and the linguistic complexity of the input. We found that the model predictions where highly correlated with features capturing structural complexity, such as sentence length, dependency link length and parse tree depth. Furthermore, we probed the model's final-layer representations to see if features capturing linguistic complexity were better encoded after fine-tuning on eye-tracking metrics. We found that features capturing structural complexity are better encoded in fine-tuned representations than in pre-trained representations, confirming the results from the aforementioned correlational analysis. The increased encoding of structural complexity features was not only observed for English (i.e. the training language), but also for Turkish and Korean, which are very distant in typology. This shows that fine-tuning on eye-tracking data leads to a *general* understanding of linguistic complexity – the kind of complexity that triggers universal patterns in human reading behaviour.

6.1 Limitations and future work

The present study quantified linguistic complexity in terms of length, frequency, morpho-syntactic and syntactic features. However, linguistic complexity is also affected by lexical, semantic and cognitive factors. For example, ambiguity, concreteness and age of acquisition are known to affect fixation durations during reading (Gilhooly and Logie, 1980). A challenge, however, is that such information needs to be obtained from psycholinguistic databases, which are only available for a limited set of languages. Nonetheless, for the languages for which this is possible, it would be interesting to investigate whether such factors are picked up by transformer models when learning to predict eye movement patterns.

Another interesting direction of research would be to examine if so-called *spill-over* and *wrap-up* effects are captured by transformer models. The spill-over effect refers to the phenomenon that readers look longer at a word if the preceding word is difficult to process (i.e. the cognitive effort required to process word n "spills over" to word n+1 (Rayner and Duffy, 1986)). The wrap-up effect materializes as longer fixation durations towards the end of a sentence, which is assumed to reflect the process of relating sentences or clauses to each other (Carpenter and Just, 1983). There is already some evidence that spill-over information is beneficial for the prediction of eye movements associated with English texts. For example, Wiechmann et al. (2022) show that concatenating explicit linguistic features of the previous sentence to the current input improves the prediction accuracy of two monolingual English transformer models for a range of eye-tracking metrics. It would be interesting to examine if transformer models are capable of capturing such effects without explicit training.

The modelling approach for learning eye movement behaviour also needs further exploration. In this study, a single model learned four eye-tracking metrics *simultaneously*. Since two out of four metrics (total fixation duration and fixation count) could be predicted using the same linguistic features, the model started relying on those particular features. As a result, the linguistic information underlying the other two eye-tracking metrics (first-pass duration and regression duration) was disregarded. Thus, multi-task learning of eye-tracking metrics might not be the optimal approach for learning a wide range of linguistic complexity features. It would be interesting to examine the effect of *weighted learning*, where the model still learns all eye-tracking metrics simultaneously, but where the loss of certain eye-tracking metrics adds more to the joint loss than others. In our case, regression duration in particular would need a higher weight. That way, incorrect predictions for regression duration are penalized more than incorrect predictions for the other eye-tracking metrics. In turn, the model is forced to pay attention to the linguistic features underlying regression duration. It might also be interesting to train the model on the different eye-tracking metrics *sequentially*, or to train separate models for each eye-tracking metric individually. Such experiments will allow us better understand which linguistic information can be picked up from each individual eye-tracking metric.

Another choice with regard to the modelling approach is whether eye movement patterns are predicted at the sentence level of the token level. In this study, we opted for sentence-level eye movement prediction, which was motivated by the notion that cross-lingual universality is more likely to be observed at the sentence level than the token level. We find that this approach works well for learning about structural complexity, but that it is not optimal for learning about lexical complexity. This is because

sentence-level eye-tracking metrics can be predicted rather accurately using length-related linguistic features alone. As a result, the model does not need to pay attention to lexical complexity (except when it receives two sentence of exactly the same length is input, see Section 4.2.4). In a future experiment, one could train individual models for predicting token-level and sentence-level eye-tracking values associated with the same reading materials. This would allow us to carefully examine the linguistic knowledge that is acquired from eye movements associated with different linguistic units.

An important finding of our study is that XLM-RoBERTa was not able to learn variation within eye-tracking metrics. This could be a result of the fact that it was only trained on averaged eye-tracking metrics. Future studies should consider to not only predict the *average* of all readers, but also the *standard deviation* across readers Hollenstein et al. (2022). While averaging eye-tracking metrics over readers leads to a more robust indication of human reading behaviour, it also 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.

A final suggestion for future work is to extend the work by González-Garduño and Søgaard (2017) and Evaldo Leal et al. (2020), who improved readability classifiers by learning eye movement behaviour as an auxiliary task. The current study demonstrates that knowledge about linguistic complexity acquired from English eye-tracking data transfers to other languages. This is a promising finding for readability classification, because it implies that learning eye movement behaviour from a single language can improve readability classification for *all* languages.

Appendix A

Additional Tables and Figures

	MECO		**GECO**	
	Mean	XLM	Mean	XLM
First-pass duration	85.08	90.60	86.15	96.33
Fixation Count	82.99	91.67	86.6	95.54
Total fixation duration	82.03	91.81	86.08	95.53
Regression duration	89.38	90.86	88.38	92.42

Table A.1: Absolute prediction accuracy of XLM-RoBERTa and a mean baseline for eye-tracking metrics from the English parts of MECO and GECO. Accuracy is calculated as 100 minus the Mean Absolute Error.

	First-pass dur.	Fixation count	Total fixation dur.	Regression dur.
Mean baseline	85.08	82.99	82.03	89.38
Length SVM	88.77	90.28	89.43	87.95
Frequency SVM	86.21	87.42	86.66	89.39
Structural SVM	88.28	89.42	88.45	88.99
All features SVM	88.82	90.39	89.64	87.91
XLM-RoBERTa	91.67	90.60	91.81	90.86

Table A.2: Absolute prediction accuracy of XLM-RoBERTa, feature-based SVM models and a mean baseline for all eye-tracking metrics from the English part of MECO. Accuracy is calculated as 100 minus the Mean Absolute Error.

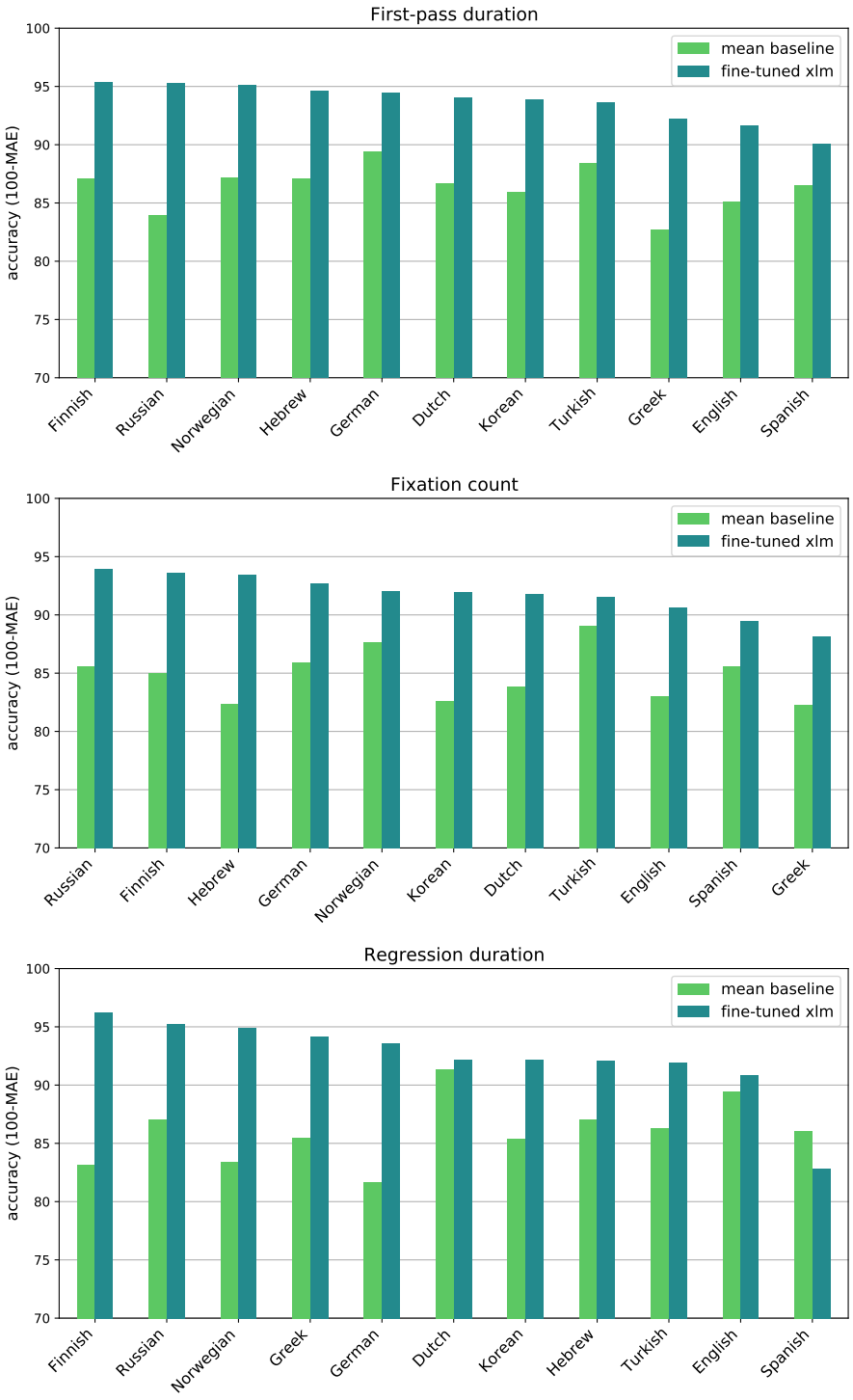

Figure A.1: Prediction accuracy of fine-tuned XLM-RoBERTa and the mean baseline for first-pass duration, fixation count, and regression duration for each language in MECO.

Bibliography

G. Altmann and Y. Kamide. Incremental interpretation at verbs: Restricting the domain of subsequent reference. *Cognition*, 73:247–264, 01 2000. doi: 10.1016/S0010-0277(99)00059-1.

M. Barrett, J. Bingel, F. Keller, and A. Søgaard. 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: Short Papers)*, pages 579–584, Berlin, Germany, Aug. 2016a. Association for Computational Linguistics. doi: 10.18653/v1/P16-2094. URL https://aclanthology.org/P16-2094.

M. Barrett, F. Keller, and A. Søgaard. 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, Dec. 2016b. The COLING 2016 Organizing Committee. URL https://aclanthology.org/C16-1126.

M. Barrett, J. Bingel, N. Hollenstein, M. Rei, and A. Søgaard. Sequence classification with human attention. In *Proceedings of the 22nd Conference on Computational Natural Language Learning*, pages 302–312, Brussels, Belgium, Oct. 2018. Association for Computational Linguistics. doi: 10.18653/v1/K18-1030. URL https://aclanthology.org/K18-1030.

Y. Belinkov. Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics*, 48(1):207–219, Mar. 2022. doi: 10.1162/coli_a_00422. URL https://aclanthology.org/2022.cl-1.7.

R. Bertram and J. Hyönä. The length of a complex word modifies the role of morphological structure: Evidence from eye movements when reading short and long finnish compounds. *Journal of Memory and Language*, 48(3):615–634, 2003. ISSN 0749-596X. doi: https://doi.org/10.1016/S0749-596X(02)00539-9. URL https://www.sciencedirect.com/science/article/pii/S0749596X02005399.

Y. Bestgen. LAST at CMCL 2021 shared task: Predicting gaze data during reading with a gradient boosting decision tree approach. In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, pages 90–96, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.cmcl-1.10. URL https://aclanthology.org/2021.cmcl-1.10.

T. Bever. *The Cognitive Basis for Linguistic Structures*, pages 279–352. Cognition and the Development of Language, 01 1970. ISBN 9780199677139. doi: 10.1093/acprof:oso/9780199677139.003.0001.

P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5(0):135–146, 2017. ISSN 2307-387X. URL https://transacl.org/ojs/index.php/tacl/article/view/999.

D. Brunato, L. De Mattei, F. Dell'Orletta, B. Iavarone, and G. Venturi. Is this sentence difficult? Do you agree? In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2690–2699, Brussels, Belgium, Oct.-Nov. 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1289. URL https://aclanthology.org/D18-1289.

D. Brunato, A. Cimino, F. Dell'Orletta, G. Venturi, and S. Montemagni. Profiling-UD: a tool for linguistic profiling of texts. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 7145–7151, Marseille, France, May 2020. European Language Resources Association. ISBN 979-10-95546-34-4. URL https://aclanthology.org/2020.lrec-1.883.

M. Brysbaert and B. New. 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, 11 2009. doi: 10.3758/BRM.41.4.977.

P. A. Carpenter and M. A. Just. What your eyes do while your mind is reading. In *Eye movements in reading*, pages 275–307. Elsevier, 1983.

S. Choudhary, K. Tandon, R. Agarwal, and N. Chatterjee. MTL782_IITD at CMCL 2021 shared task: Prediction of eye-tracking features using BERT embeddings and linguistic features. In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, pages 114–119, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.cmcl-1.14. URL https://aclanthology.org/2021.cmcl-1.14.

C. Clifton Jr and A. Staub. Syntactic influences on eye movements during reading. *Eye*, 3(2):897–909, 2011.

A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, and V. Stoyanov. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.747. URL https://aclanthology.org/2020.acl-main.747.

U. Cop, N. Dirix, D. Drieghe, and W. Duyck. Presenting GECO: An eyetracking corpus of monolingual and bilingual sentence reading. *Behavior Research Methods*, 49, 05 2016. doi: 10.3758/s13428-016-0734-0.

F. Dary, A. Nasr, and A. Fourtassi. TALEP at CMCL 2021 shared task: Non linear combination of low and high-level features for predicting eye-tracking data. In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, pages 108–113, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.cmcl-1.13. URL https://aclanthology.org/2021.cmcl-1.13.

Deepset. Farm: Framework for adapting representation models. 2019. URL `https://github.com/deepset-ai/FARM`.

J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. 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, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423. URL `https://aclanthology.org/N19-1423`.

N. Dirix and W. Duyck. An eye movement corpus study of the age of acquisition effect. *Psychonomic Bulletin and Review*, 24, 01 2017. doi: 10.3758/s13423-017-1233-8.

S. A. Duffy, R. K. Morris, and K. Rayner. Lexical ambiguity and fixation times in reading. *Journal of Memory and Language*, 27(4):429–446, 1988. ISSN 0749-596X. doi: https://doi.org/10.1016/0749-596X(88)90066-6. URL `https://www.sciencedirect.com/science/article/pii/0749596X88900666`.

S. Evaldo Leal, J. M. Munguba Vieira, E. dos Santos Rodrigues, E. Nogueira Teixeira, and S. Aluísio. 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), Dec. 2020. International Committee on Computational Linguistics. doi: 10.18653/v1/2020.coling-main.512. URL `https://aclanthology.org/2020.coling-main.512`.

R. F. Flesch. A new readability yardstick. *The Journal of applied psychology*, 32 3: 221–33, 1948.

L. Frazier and K. Rayner. Making and correcting errors during sentence comprehension: Eye movements in the analysis of structurally ambiguous sentences. *Cognitive psychology*, 14(2):178–210, 1982.

K. J. Gilhooly and R. H. Logie. Age-of-acquisition, imagery, concreteness, familiarity, and ambiguity measures for 1,944 words. *Behavior research methods & instrumentation*, 12(4):395–427, 1980.

A. V. González-Garduño and A. Søgaard. 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, Sept. 2017. Association for Computational Linguistics. doi: 10.18653/v1/W17-5050. URL `https://aclanthology.org/W17-5050`.

P. Gordon, R. Hendrick, M. Johnson, and Y. Lee. Similarity-based interference during language comprehension: Evidence from eye tracking during reading. *Journal of experimental psychology. Learning, memory, and cognition*, 32:1304–21, 12 2006. doi: 10.1037/0278-7393.32.6.1304.

R. Hall Maudslay and R. Cotterell. Do syntactic probes probe syntax? experiments with jabberwocky probing. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 124–131, Online, June 2021. Association for

Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.11. URL https://aclanthology.org/2021.naacl-main.11.

N. Hollenstein, N. Langer, A. Pedroni, M. Troendle, J. Rotsztejn, C. Zhang, C. Pfeiffer, and L. Muttenthaler. Zurich cognitive language processing corpus: A simultaneous eeg and eye-tracking resource for analyzing the human reading process. 2018.

N. Hollenstein, E. Chersoni, C. L. Jacobs, Y. Oseki, L. Prévot, and E. Santus. CMCL 2021 shared task on eye-tracking prediction. In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, pages 72–78, Online, June 2021a. Association for Computational Linguistics. doi: 10.18653/v1/2021.cmcl-1.7. URL https://aclanthology.org/2021.cmcl-1.7.

N. Hollenstein, F. Pirovano, C. Zhang, L. Jäger, and L. Beinborn. 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, June 2021b. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.10. URL https://aclanthology.org/2021.naacl-main.10.

N. Hollenstein, E. Chersoni, C. Jacobs, Y. Oseki, L. Prévot, and E. Santus. 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, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.cmcl-1.14. URL https://aclanthology.org/2022.cmcl-1.14.

D. Hupkes, S. Bouwmeester, and R. Fernández. Analysing the potential of seq-to-seq models for incremental interpretation in task-oriented dialogue. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 165–174, Brussels, Belgium, Nov. 2018. Association for Computational Linguistics. doi: 10.18653/v1/W18-5419. URL https://aclanthology.org/W18-5419.

J. Hyönä and J. K. Kaakinen. *Eye Movements During Reading*, pages 239–274. Springer International Publishing, Cham, 2019. ISBN 978-3-030-20085-5. doi: 10.1007/978-3-030-20085-5_7. URL https://doi.org/10.1007/978-3-030-20085-5_7.

J. Hyönä, R. Olson, J. Defries, D. Fulker, B. Pennington, and S. Smith. Eye fixation patterns among dyslexic and normal readers: Effects of word length and word frequency. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 21: 1430–1440, 12 1995. doi: 10.1037/0278-7393.21.6.1430.

B. J. Juhasz and H. Sheridan. The time course of age-of-acquisition effects on eye movements during reading: Evidence from survival analyses. *Memory & Cognition*, 48:83–95, 2019.

R. Kliegl, E. Grabner, M. Rolfs, and R. Engbert. Length, frequency, and predictability effects of words on eye movements in reading. *European Journal of Cognitive Psychology*, 16(1-2):262–284, 2004. doi: 10.1080/09541440340000213. URL https://doi.org/10.1080/09541440340000213.

K. Krafka, A. Khosla, P. Kellnhofer, H. Kannan, S. M. Bhandarkar, W. Matusik, and A. Torralba. Eye tracking for everyone. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2176–2184, 2016.

T. Kudo and J. Richardson. 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, Nov. 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-2012. URL https://aclanthology.org/D18-2012.

Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma, and R. Soricut. ALBERT: A lite BERT for self-supervised learning of language representations. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=H1eA7AEtvS.

A. K. Laurinavichyute, I. A. Sekerina, S. Alexeeva, K. Bagdasaryan, and R. Kliegl. Russian Sentence Corpus: Benchmark measures of eye movements in reading in Russian. *Behavior Research Methods*, 51:1161–1178, 2019.

B. Li and F. Rudzicz. TorontoCL at CMCL 2021 shared task: RoBERTa with multi-stage fine-tuning for eye-tracking prediction. In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, pages 85–89, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.cmcl-1.9. URL https://aclanthology.org/2021.cmcl-1.9.

P. Lison and J. Tiedemann. 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, May 2016. European Language Resources Association (ELRA). URL https://aclanthology.org/L16-1147.

Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov. Roberta: A robustly optimized bert pretraining approach, 2019. URL https://arxiv.org/abs/1907.11692.

S. P. Liversedge, K. Rayner, S. J. White, D. Vergilino-Perez, J. M. Findlay, and R. W. Kentridge. Eye movements when reading disappearing text: is there a gap effect in reading? *Vision Research*, 44(10):1013–1024, 2004. ISSN 0042-6989. doi: https://doi.org/10.1016/j.visres.2003.12.002. URL https://www.sciencedirect.com/science/article/pii/S0042698903007909.

S. P. Liversedge, D. Drieghe, X. Li, G. Yan, X. Bai, and J. Hyönä. Universality in eye movements and reading: A trilingual investigation. *Cognition*, 147:1–20, 2016. ISSN 0010-0277. doi: https://doi.org/10.1016/j.cognition.2015.10.013. URL https://www.sciencedirect.com/science/article/pii/S0010027715300913.

S. Luke and K. Christianson. The Provo corpus: A large eye-tracking corpus with predictability norms. *Behavior research methods*, 50, 05 2017. doi: 10.3758/s13428-017-0908-4.

A. Miaschi, D. Brunato, F. Dell'Orletta, and G. Venturi. Linguistic profiling of a neural language model. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 745–756, Barcelona, Spain (Online), Dec. 2020. International Committee on Computational Linguistics. doi: 10.18653/v1/2020.coling-main.65. URL https://aclanthology.org/2020.coling-main.65.

R. K. Morris. Lexical and message-level sentence context effects on fixation times in reading. *Journal of experimental psychology. Learning, memory, and cognition*, 20 1:92–103, 1994.

M. Mosbach, A. Khokhlova, M. A. Hedderich, and D. Klakow. On the interplay between bine-tuning and sentence-level probing for linguistic knowledge in pretrained transformers. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2502–2516, Online, Nov. 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.findings-emnlp.227. URL https://aclanthology.org/2020.findings-emnlp.227.

B.-D. Oh. Team Ohio State at CMCL 2021 shared task: Fine-tuned RoBERTa for eye-tracking data prediction. In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, pages 97–101, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.cmcl-1.11. URL https://aclanthology.org/2021.cmcl-1.11.

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. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

T. Pimentel, J. Valvoda, R. Hall Maudslay, R. Zmigrod, A. Williams, and R. Cotterell. Information-theoretic probing for linguistic structure. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4609–4622, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.420. URL https://aclanthology.org/2020.acl-main.420.

T. Pires, E. Schlinger, and D. Garrette. How multilingual is multilingual BERT? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4996–5001, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1493. URL https://aclanthology.org/P19-1493.

A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever. Language Models are Unsupervised Multitask Learners. 2019. URL https://openai.com/blog/better-language-models/.

K. Rayner. Eye movements in reading and information processing: 20 years of research. *Psychological Bulletin*, 124(3):372 – 422, 1998. ISSN 0033-2909. URL https://search-ebscohost-com.vu-nl.idm.oclc.org/login.aspx?direct=true&db=pdh&AN=1998-11174-004&site=ehost-live.

K. Rayner and S. A. Duffy. Lexical complexity and fixation times in reading: Effects of word frequency, verb complexity, and lexical ambiguity. *Memory & Cognition*, 14: 191–201, 1986.

G. Sarti, D. Brunato, and F. Dell'Orletta. 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, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.cmcl-1.5. URL https://aclanthology.org/2021.cmcl-1.5.

N. Siegelman, S. Schroeder, C. Acartürk, H.-D. Ahn, S. Alexeeva, S. Amenta, R. Bertram, R. Bonandrini, M. Brysbaert, D. Chernova, S. M. D. Fonseca, N. Dirix, W. Duyck, A. Fella, R. Frost, C. A. Gattei, A. Kalaitzi, N. Kwon, K. Lõo, M. Marelli, T. C. Papadopoulos, A. Protopapas, S. Savo, D. E. Shalom, N. Slioussar, R. Stein, L. Sui, A. Taboh, V. Tønnesen, K. A. Usal, and V. Kuperman. Expanding horizons of cross-linguistic research on reading: The multilingual eye-movement corpus (meco). *Behavior Research Methods*, page 1–21, 2022. doi: 10.3758/s13428-021-01772-6.

R. Speer, J. Chin, A. Lin, S. Jewett, and L. Nathan. Luminosoinsight/wordfreq: v2.2, Oct. 2018. URL https://doi.org/10.5281/zenodo.1443582.

M. Taft. Cognitive psychology of lexical access. In N. J. Smelser and P. B. Baltes, editors, *International Encyclopedia of the Social Behavioral Sciences*, pages 8743–8748. Pergamon, Oxford, 2001. ISBN 978-0-08-043076-8. doi: https://doi.org/10.1016/B0-08-043076-7/01538-2. URL https://www.sciencedirect.com/science/article/pii/B0080430767015382.

W. van Heuven, P. Mandera, E. Keuleers, and M. Brysbaert. SUBTLEX-UK: A new and improved word frequency database for British English. *Quarterly journal of experimental psychology (2006)*, 67, 01 2014. doi: 10.1080/17470218.2013.850521.

S. Vasishth, T. von der Malsburg, and F. Engelmann. What eye movements can tell us about sentence comprehension. *Wiley interdisciplinary reviews. Cognitive science*, 4 2:125–134, 2013.

A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

P. Vickers, R. Wainwright, H. Tayyar Madabushi, and A. Villavicencio. CogNLP-Sheffield at CMCL 2021 shared task: Blending cognitively inspired features with transformer-based language models for predicting eye tracking patterns. In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, pages 125–133, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.cmcl-1.16. URL https://aclanthology.org/2021.cmcl-1.16.

D. Wiechmann, Y. Qiao, E. Kerz, and J. Mattern. 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, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.362. URL https://aclanthology.org/2022.acl-long.362.

Q. Yu, A.-L. Kalouli, and D. Frassinelli. KonTra at CMCL 2021 shared task: Predicting eye movements by combining BERT with surface, linguistic and behavioral information. In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, pages 120–124, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.cmcl-1.15. URL https://aclanthology.org/2021.cmcl-1.15.

D. Zeman, M. Popel, M. Straka, J. Hajič, J. Nivre, F. Ginter, J. Luotolahti, S. Pyysalo, S. Petrov, M. Potthast, F. Tyers, E. Badmaeva, M. Gokirmak, A. Nedoluzhko, S. Cinková, J. Hajič jr., J. Hlaváčová, V. Kettnerová, Z. Urešová, J. Kanerva, S. Ojala, A. Missilä, C. D. Manning, S. Schuster, S. Reddy, D. Taji, N. Habash, H. Leung, M.-C. de Marneffe, M. Sanguinetti, M. Simi, H. Kanayama, V. de Paiva, K. Droganova, H. Martínez Alonso, Ç. Çöltekin, U. Sulubacak, H. Uszkoreit, V. Macketanz, A. Burchardt, K. Harris, K. Marheinecke, G. Rehm, T. Kayadelen, M. Attia, A. Elkahky, Z. Yu, E. Pitler, S. Lertpradit, M. Mandl, J. Kirchner, H. F. Alcalde, J. Strnadová, E. Banerjee, R. Manurung, A. Stella, A. Shimada, S. Kwak, G. Mendonça, T. Lando, R. Nitisaroj, and J. Li. 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, Aug. 2017. Association for Computational Linguistics. doi: 10.18653/v1/K17-3001. URL https://aclanthology.org/K17-3001.